Pacific Journal of Mathematics

Volume 160 No. 1 September 1993

CONTENTS

Vol. 160, No. 1 September 1993

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

INEQUALITIES FOR QUASICONFORMAL MAPPINGS IN SPACE

G. D. ANDERSON, M. K. VAMANAMURTHY AND M. VUORINEN

A new lower bound for the conformal capacity of the Grötzsch ring and sharp bounds for the radial distortion of a quasiconformal automorphism of the unit ball are obtained in n-space, $n \geq 2$.

1. Introduction. The conformal capacities of the Grötzsch and Teichmüller extremal rings in $\mathbb{R}^n$, $n \geq 2$ (see §2), are denoted by

$$\gamma_n(s) = \operatorname{cap} R_{G,n}(s) \quad \text{and} \quad \tau_n(t) = \operatorname{cap} R_{T,n}(t), \tag{1.1}$$

respectively, where $s > 1$ and $t > 0$. The modulus $M_n(r)$ of the Grötzsch ring $R_{G,n}(1/r)$, $0 < r < 1$, is defined by

$$\gamma_n(1/r) = \omega_{n-1} M_n(r)^{1-n}, \tag{1.2}$$

where ω_{n-1} is the $(n-1)$-dimensional measure of the unit sphere S^{n-1} in $\mathbb{R}^n$. The capacities in (1.1) are related [**G**, §18] by

$$\gamma_n(s) = 2^{n-1}\tau_n(s^2-1), \qquad s > 1. \tag{1.3}$$

For $K > 0$ define increasing homeomorphisms $\varphi_{K,n}$ and $\psi_{K,n}$ from $(0,1)$ onto $(0,1)$ by

$$\begin{cases} \varphi_{K,n}(r) = 1/\gamma_n^{-1}(K\gamma_n(1/r)) = M_n^{-1}(\alpha M_n(r)), \\ \psi_{K,n}(r) = (1-\varphi_{1/K,n}^2(r'))^{1/2}, \end{cases} \tag{1.4}$$

where $r' = \sqrt{1-r^2}$ and $\alpha = K^{1/(1-n)}$. Given a domain D in $\mathbb{R}^n$, for $K \geq 1$ let $QC_K(D)$ and $QR_K(D)$ denote the class of all K-quasiconformal and K-quasiregular mappings, respectively, of D into itself [**V1**], [**Vu2**]. For $K \geq 1$, $0 < r < 1$, define [**AVV2**]

$$\varphi_{K,n}^*(r) = \sup\{|f(x)| : |x| = r,\, f(0) = 0,\, f \in QC_K(B^n)\}, \tag{1.5}$$

$$\varphi_{1/K,n}^*(r) = \inf\{|f(x)| : |x| = r,\, f(0) = 0,\, f \in QC_K(B^n),\, f(B^n) = B^n\}. \tag{1.6}$$

We extend the functions in (1.4), (1.5), (1.6) to [0,1] by defining them to be 0 at 0 and 1 at 1. For $n \geq 2$, $K \geq 1$, $0 < r < 1$, these distortion functions are related by the inequalities **[Vu1**, 3.5, 5.20], **[AVV2**, (1.4), Theorem 2.24]

$$(1.7)\qquad \begin{cases} \varphi^*_{K,n}(r) \le \min\{\varphi_{K,n}(r),\, \psi_{K,n}(r)\}, \\ \varphi^*_{1/K,n}(r) \ge \max\{\varphi_{1/K,n}(r),\, \psi_{1/K,n}(r)\}. \end{cases}$$

Each of these three functions is increasing from $[0, 1]$ onto $[0, 1]$ (see §3 below). For $n = 2$ the inequalities in (1.7) reduce to equalities.

It is well known that $M_n(r)+\log r$ is monotone decreasing on $(0, 1)$ **[T**, p. 632], **[G]**. The so-called Grötzsch ring constant λ_n defined by

$$(1.8)\qquad \log\lambda_n = \lim_{r\to 0+}(M_n(r) + \log r)$$

satisfies $\lambda_2 = 4$ **[T]**, **[LV**, (2.11), p. 62] and $\lambda_n \in (2e^{0.76(n-1)},\, 2e^{n-1})$ for $n \ge 3$ **[G]**, **[AVV4**, pp. 120–121], **[AF]**.

The main purpose of this paper is to show how one can translate information about the special function $M_n(r)$ into information about geometric properties of quasiconformal mappings. An important tool is the following result, which improves the above-mentioned monotone property of $M_n(r) + \log r$.

1.9. THEOREM. *For each* $n \ge 2$, *the function*

$$f_n(r) \equiv M_n(r) + \log\left(\frac{r}{1 + r'}\right)$$

is strictly decreasing from $(0, 1)$ *onto* $(0, \log(\lambda_n/2))$, *where* λ_n *is as in* (1.8) *and* $r' = (1 - r^2)^{1/2}$. *Moreover,* $f_2(r)$ *is strictly concave on* $(0, 1)$.

The next result is an immediate consequence of Theorem 1.9 (cf. **[G**, Lemma 8], **[AVV4**, Corollary 2.30]).

1.10. COROLLARY. *For each* $n \ge 2$, $0 < r < 1$, $r' = \sqrt{1 - r^2}$,

$$(1)\qquad M_n(r) < \log\left(\frac{\lambda_n}{2}\frac{1 + r'}{r}\right) < \log\frac{\lambda_n}{r},$$

$$(2)\qquad \frac{M_n(r)}{r\log r} \le \frac{d}{dr}M_n(r) \le -\frac{1}{rr'}.$$

The usual method of obtaining lower bounds for the capacity of a ring is to use spherical symmetrization **[G]** together with the extremal property of the Teichmüller ring and the fundamental inequality $M_n(r) \le \log(\lambda_n/r)$. By virtue of Corollary 1.10(1), all such earlier bounds can now be improved. A graphical comparison of some bounds for $\gamma_3(1/r)$ appears in Figure 1 in §2.

From Theorem 1.9 and Corollary 1.10(1) we shall derive several inequalities for the distortion functions in (1.4), (1.5), and (1.6). These functions are interesting not only for studying the radial behavior of mappings in $QC_K(B^n)$, but also as means of expressing other special functions.

1.11. THEOREM. *For* $n \geq 2$ *and* $K \geq 1$ *let* $\alpha = K^{1/(1-n)} = 1/\beta$, $L = (\lambda_n/2)^{1-\alpha}$, $l = (\lambda_n/2)^{1-\beta}$, $l_1 = 2l/(1+l^2)$. *Then*

$$(1) \qquad \varphi_{K,n}(r) \leq \tanh(2\operatorname{arctanh}(LA(r)^\alpha))$$

for $r \in (0, l_1)$, *and*

$$(2) \qquad \varphi_{1/K,n}(r) \geq \tanh(2\operatorname{arctanh}(lA(r)^\beta))$$

for all $r \in (0, 1)$, *where* $A(r) = r/(1+r')$, $r' = \sqrt{1-r^2}$. *For* $K = 1$, *both* (1) *and* (2) *reduce to equality.*

The following local Hölder continuity theorem simplifies and improves earlier results in **[G]**, **[R**, pp. 82–83], **[MRV]**, **[AVV2]**, **[Ca]**.

1.12. THEOREM. *For* $K \geq 1$, $n \geq 2$ *let* $f \in QR_K(B^n)$, $0 < r < 1$, $\alpha = K^{1/(1-n)}$, *and let* λ_n *be as in* (1.8).

(1) *If* $x, y \in \overline{B}^n(r)$ *then*

$$|f(x) - f(y)| \leq \lambda_n^{1-\alpha}(1-r^2)^{-\alpha}|x-y|^\alpha.$$

(2) *If* $x \in \overline{B}^n(r)$ *and* $y \in B^n$, *then*

$$|f(x) - f(y)| \leq 2^\alpha \lambda_n^{1-\alpha}(1-r)^{-\alpha}|x-y|^\alpha.$$

Next, the hyperbolic metric $\rho(x, y)$ on B^n is given by

$$(1.13) \qquad \tanh^2 \frac{\rho(x,y)}{2} = \frac{|x-y|^2}{|x-y|^2 + (1-|x|^2)(1-|y|^2)}$$

for $x, y \in B^n$ (cf. **[B**, p. 40], **[Vu2**, 2.47]). The following distortion theorem for the hyperbolic metric is a consequence of Theorem 1.9.

1.14. THEOREM. *For* $K \geq 1$, $n \geq 2$, *let* $f \in QR_K(B^n)$. *Then for all* $x, y \in B^n$,

$$\tanh\frac{\rho'}{4} \leq \left(\frac{\lambda_n}{2}\right)^{1-\alpha}\left(\tanh\frac{\rho}{4}\right)^\alpha \leq K\left(\tanh\frac{\rho}{4}\right)^\alpha \leq K\left(\tanh\frac{\rho}{4}\right)^{1/K},$$

where $\rho = \rho(x, y)$, $\rho' = \rho(f(x), f(y))$, *and* $\alpha = K^{1/(1-n)}$. *These estimates are sharp as* K *tends to* 1.

1.15. Corollary. *For* $K \geq 1$, $n \geq 2$, *let* $f \in QC_K(B^n)$. *Then for all* $x, y \in B^n$,

$$\tanh\frac{\rho'}{4} \leq \min\{2, K\}\left(\tanh\frac{\rho}{4}\right)^{1/K},$$

where $\rho = \rho(x, y)$ *and* $\rho' = \rho(f(x), f(y))$.

The results in §3 will show that for $n \geq 3$ the functions $\varphi_{K,n}(r)$, $\varphi^*_{K,n}(r)$, and $\psi_{K,n}(r)$ in (1.7) behave differently and that the behavior of $\varphi^*_{K,n}(r)$ differs drastically from that of $\varphi_{K,2}(r)$ (cf. 3.7(1) and 3.14).

In §4 we study the *linear dilatation*

$$(1.16)\quad H_n(K) = \sup\left\{\frac{|f(x)|}{|f(y)|} : |x| = |y| > 0,\right. \\ \left. f \in QC_K(\mathbb{R}^n),\, f(0) = 0\right\},$$

obtaining the following asymptotic estimate.

1.17. Theorem. *For each* $K > 1$, $\limsup_{n\to\infty} H_n(K) \leq K^4$.

An explicit expression for $H_n(K)$ is known only for $n = 2$, and

$$H_2(K) = \left(\frac{\mu^{-1}(\pi/(2K))}{\mu^{-1}(\pi K/2)}\right)^2$$

[**LV**, (6.4), p. 81] is also denoted by $\lambda(K)$. It follows from [**AVV5**, Theorems 1.1, 1.2] that

$$(1.18)\qquad K^{\pi\sqrt{K}} < e^{\pi(K-1)} < \lambda(K) = H_2(K) < e^{\pi(K-(1/K))}$$

for $K > 1$ and that $(\lambda(K))^{1/K}$ is strictly increasing to e^{π} as K tends to ∞. Thus 1.18 and Theorem 1.17 imply that

$$(1.19)\qquad H_n(K) < \left(1 + \frac{\log H_2(K)}{\pi}\right)^4 < (H_2(K))^{4/\pi}, \qquad K > 1,$$

for n sufficiently large. On the other hand by [**AVV1**, Theorem 1.14], $H_n(K) \geq \lambda(K^{1/(n-1)}) > 1$ for each $K > 1$ and $n \geq 2$. Whether the upper bound in Theorem 1.17 can be replaced by 1 remains an open problem (see 4.12(1) below). In any case, Theorem 1.17 shows that quasiconformal mappings become more rigid in high dimensions. If one wants to extend the theory of quasiconformal mappings to infinite-dimensional Banach spaces then the definition of a K-quasiconformal mapping must be chosen carefully, as e.g. in J. Väisälä **[V3]**.

1.20. NOTATION. For $n \geq 2$, $\mathbb{R}^n$ denotes the n-dimensional euclidean space, and $\overline{\mathbb{R}}^n$ its one-point compactification $\mathbb{R}^n \cup \{\infty\}$. Unit vectors along the coordinate axes in $\mathbb{R}^n$ are denoted by $e_1, e_2, \dots, e_n$. For x in $\mathbb{R}^n$ and $r > 0$ we let $B^n(x, r) = \{z \in \mathbb{R}^n : |x - z| < r\}$, $S^{n-1}(x, r)$ its boundary sphere, $B^n(r) = B^n(0, r)$, $S^{n-1}(r) = S^{n-1}(0, r)$, $B^n = B^n(1)$, and $S^{n-1} = S^{n-1}(1)$. For any $E \subset \overline{\mathbb{R}}^n$ we let $\overline{E}$ denote its closure. For $a, b \in \mathbb{R}^n$ we let $[a, b] = \{(1-t)a + tb : 0 \leq t \leq 1\}$ and for $a \in \mathbb{R}^n \setminus \{0\}$, $[a, \infty) = \{ta : t \geq 1\}$, $[a, \infty] = [a, \infty) \cup \{\infty\}$. Whenever $0 < r < 1$, by r' we shall mean $\sqrt{1 - r^2}$.

Given E, F, and G, subsets of $\overline{\mathbb{R}}^n$, we let $\Delta(E, F; G)$ denote the family of all curves joining E and F in G **[V1]**. The conformal modulus of a family of curves Γ in $\overline{\mathbb{R}}^n$ is defined by

$$M(\Gamma) = \inf_{\rho} \int_{\mathbb{R}^n} \rho^n \, dm,$$

where dm represents n-dimensional Lebesgue measure and the infimum is taken over all nonnegative Borel-measurable functions ρ satisfying $\int_{\gamma} \rho \, ds \geq 1$ for each locally rectifiable curve $\gamma \in \Gamma$ **[V1]**.

Let G be a domain (open connected set) in $\mathbb{R}^n$. A continuous function $f : G \to \mathbb{R}^n$ is said to be *K-quasiregular*, $1 \leq K < \infty$, if f is ACL^n, i.e. f is absolutely continuous on almost all lines parallel to the coordinate axes and the first partial derivatives are locally L^n-integrable, and if

$$(1.21) \quad \begin{cases} |f'(x)|^n \leq K J_f(x), & |f'(x)| \equiv \max_{|h|=1} |f'(x)h|, \\ J_f(x) \leq K l(f'(x))^n, & l(f'(x))^n \equiv \min_{|h|=1} |f'(x)h| \end{cases}$$

hold a.e. in G. Here $f'(x)$ denotes the formal derivative of f at x, i.e. $f'(x)e_i = \nabla f_i$, $i = 1, \dots, n$, $f = (f_1, f_2, \dots, f_n)$, and $J_f(x) = \det(f'(x))$. A K-quasiregular homeomorphism is called *K-quasiconformal.* A mapping f is called *quasiregular* or *quasiconformal* if it is K-quasiregular or K-quasiconformal, respectively, for some $K \in [1, \infty)$. For properties of these mappings the reader is referred to **[V1]**, **[V2]**, **[Vu2]**.

Acknowledgments. The research of the first two authors was supported in part by grants from the University of Auckland and the New Zealand Mathematical Society, and that of the third author by the Alexander von Humboldt Foundation and the Academy of Finland. The first author completed his research during his visit at Indiana University in the ITM/MUCIA Cooperative Program in Malaysia. The third author wishes to thank Professors Ch. Pommerenke and J. Becker for their kind help and hospitality during his visit in 1988-89 at the Technical University of Berlin as a Humboldt Fellow.

2. Modulus of the Grötzsch ring. A *ring* $R = R(C_0, C_1)$ in $\overline{\mathbb{R}}^n$ is a domain whose complement consists of two components C_0 and C_1, where $\infty \notin C_0$. Two extremal rings, the Grötzsch ring

$$R_{G,n}(s) = \overline{\mathbb{R}}^n \backslash (\overline{B}^n \cup [se_1, \infty]), \qquad s > 1,$$

and the Teichmüller ring

$$R_{T,n}(t) = \overline{\mathbb{R}}^n \backslash ([-e_1, 0] \cup [te_1, \infty]), \qquad t > 0,$$

are important in the study of distortion **[G]**, **[Vu1]**, **[Vu2]**, **[AVV1]**, **[AVV2]**, **[AVV5]**.

The *conformal capacity* cap R and *modulus* mod R of a ring $R = R(C_0, C_1)$ are given **[V1]** by

$$\operatorname{cap} R = M(\Gamma) = \omega_{n-1}(\operatorname{mod} R)^{1-n},$$

where $\Gamma = \Delta(C_0, C_1; R)$ and ω_{n-1} is the $(n-1)$-dimensional measure of S^{n-1}.

The function $M_2(r)$ defined in formula (1.2), usually denoted by $\mu(r)$, is given explicitly **[LV, (2.2), p. 60]** by

$$\mu(r) = \frac{\pi}{2}\frac{\mathscr{K}'(r)}{\mathscr{K}(r)}, \tag{2.1}$$

where

$$\mathscr{K}(r) = \int_0^{\pi/2} (1 - r^2 \sin^2 t)^{-1/2}\, dt, \quad \mathscr{K}'(r) = \mathscr{K}(r'), \quad r' = \sqrt{1 - r^2},$$

are complete elliptic integrals. We also use the dual integrals

$$\mathscr{E}(r) = \int_0^{\pi/2} (1 - r^2 \sin^2 t)^{1/2}\, dt, \quad \mathscr{E}'(r) = \mathscr{E}(r').$$

The next result is an analog of l'Hôpital's rule and will be useful in establishing monotoneity of a ratio of two functions.

2.2. Lemma. *For* $-\infty < a < b < \infty$ *let* $f, g\colon [a, b) \to \mathbb{R}$ *be differentiable functions such that* $g'(x) \neq 0$ *for* $x \in (a, b)$.

If $f'(x)/g'(x)$ *is increasing* (*decreasing*) *on* (a, b) *then so is* $(f(x)-f(a))/(g(x)-g(a))$.

Proof. We may assume that $g'(x) > 0$ for all $x \in (a, b)$ and that $f'(x)/g'(x)$ is increasing on (a, b). By the Cauchy mean value theorem, for $x \in (a, b)$ there exists $y \in (a, x)$ such that

$$\frac{f(x)-f(a)}{g(x)-g(a)} = \frac{f'(y)}{g'(y)} \leq \frac{f'(x)}{g'(x)},$$

which implies that $(f(x)-f(a))/(g(x)-g(a))$ has a positive derivative on (a, b). □

2.3. *Proof of Theorem* 1.9. The monotoneity appears in [**A1**, Theorem 5, p. 15], but for completeness we include the proof here. For $0 < a < b < 1$ let R be the ring whose boundary components are S^{n-1} and the segment $[-ae_1, ae_1]$. Then $S^{n-1}(a/b)$ separates R into two rings R_1 and R_2, where R_1 is conformally equivalent to $R_{G,n}((1+b^2)/2b)$ and R_2 is the spherical annulus $\{x\colon a/b < |x| < 1\}$. By superadditivity of the modulus [**F**] we get $\operatorname{mod} R \geq \operatorname{mod} R_1 + \operatorname{mod} R_2$, that is,

$$M_n\left(\frac{2a}{1+a^2}\right) \geq M_n\left(\frac{2b}{1+b^2}\right) + \log\frac{b}{a}.$$

Next, putting $r_1 = 2a/(1+a^2)$, $r_2 = 2b/(1+b^2)$, we conclude that f_n is decreasing. The limit as r tends to 0 follows from (1.8), while the limit at 1 is trivial. In [**A2**, §8] a different proof of this monotoneity is based on Hölder's inequality, which, by [**A2**, Lemma 4], cannot reduce to equality; hence the monotoneity is strict.

Next, let $n = 2$ and $f(r) = f_2(r)$. Then by [**AVV6**, Lemma 2.1(2)],

$$f'(r) = \frac{1}{rr'}\left(1 - \frac{\pi^2}{4r'\mathscr{K}(r)^2}\right),$$

which is negative for $0 < r < 1$ since $r'\mathscr{K}(r)^2$ is strictly decreasing on $(0, 1)$ [**AVV6**, Theorem 2.2(3)] and $\mathscr{K}(0) = \pi/2$. Thus f is strictly decreasing on $(0, 1)$. Next, $-f'(r) = g(r)h(r)/r$, where

$$g(r) = \frac{\frac{\pi}{2\sqrt{r'}\mathscr{K}(r)} + 1}{r'} \quad \text{and} \quad h(r) = \frac{\pi}{2\sqrt{r'}\mathscr{K}(r)} - 1.$$

Since $g(r)$ is increasing [**AVV6**, Theorem 2.1(3)], we need only prove that $h(r)/r$ is increasing, and by Lemma 2.2 it is sufficient to show that $h'(r)$ is increasing. First by [**BF**, 710.00]

$$\frac{h'(r)}{1+h(r)} = \frac{1}{2r'^2\mathscr{K}} \cdot \frac{H(r)}{r}, \quad \text{where } H(r) = (\mathscr{K} - \mathscr{E}) - (\mathscr{E} - r'^2\mathscr{K}(r)).$$

By [**AVV6**, Theorem 2.2(3)] the first factor is increasing; thus by Lemma 2.2 it is sufficient to show that $H'(r)$ is increasing. Now, by [**BF**, 710.04-.05], $H'(r) = (\mathscr{E} - r'^2\mathscr{K})rr'^{-2}$, which is clearly increasing. □

2.4. REMARK. For $0 < r < 1$, by [**A1**, Theorem 2], [**G**, Lemma 8] (cf. [**AVV4**, (1.10), (1.11)]), and 1.10(1) we have

$$\text{(2.5)} \qquad \begin{cases} 2^{n-1}c_n \log\frac{1+r}{1-r} \le \gamma_n(1/r) \le 2^{n-1}c_n\mu\left(\frac{1-r}{1+r}\right), \\ \omega_{n-1}\left(\log\frac{\lambda_n(1+r')}{2r}\right)^{1-n} \le \gamma_n(1/r) \le \omega_{n-1}\mu(r)^{1-n}, \end{cases}$$

where c_n is a constant [**V1**, p. 31]. Here equality holds on the right sides for $n = 2$. Let $f(r)$, $g(r)$ denote the minorants and $F(r)$, $G(r)$ the majorants in (2.5) and let $h(r) = \omega_{n-1}(\log(\lambda_n/r))^{1-n}$. Then, for $n = 3$, the graphs of these functions are in Figure 1, where the graph of $\gamma_3(1/r)$ lies in the shaded region.

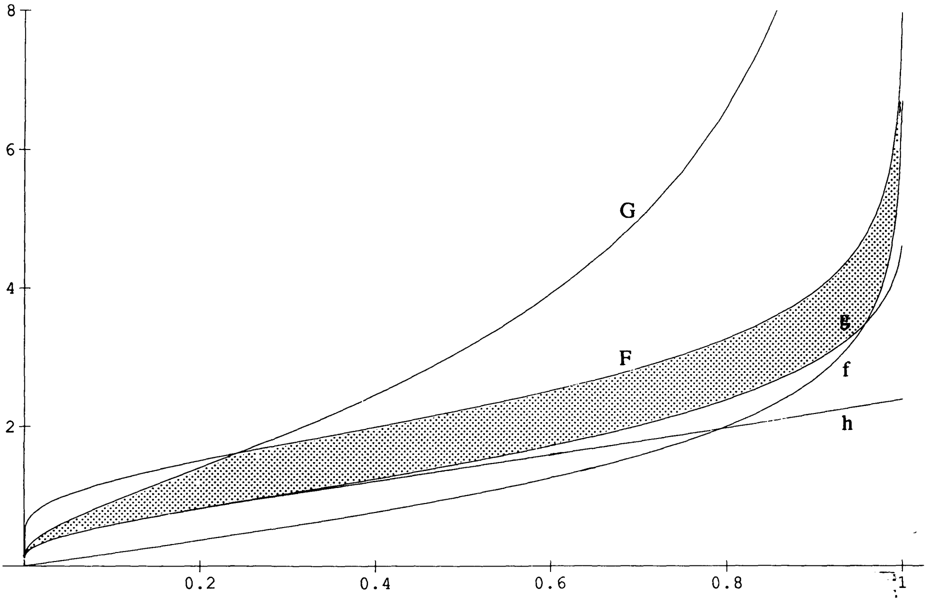

FIGURE 1. Bounds for $\gamma_3(1/r)$, $0 < r < 1$

Lower: $f(r) = 4c_3 \log\frac{1+r}{1-r}$, $g(r) = \frac{4\pi}{\log^2(4.9501(1+r')/r)}$, $h(r) = \frac{4\pi}{\log^2(9.9002/r)}$,

Upper: $F(r) = 4c_3\mu\left(\frac{1-r}{1+r}\right)$, $G(r) = \frac{4\pi}{\mu(r)^2}$.

From Figure 1 we see that the new lower bound $g(r)$ for $\gamma_3(1/r)$ given by Corollary 1.10(1) is the best of the present lower bounds when $0 < r < r_0$, $r_0 \approx 0.95$.

2.6. THEOREM. *For* $n \geq 2$ *the function*

$$f_n(r) \equiv M_n(r) \log((1+\sqrt{r})/(1-\sqrt{r}))$$

is strictly increasing from $(0, 1)$ *onto* I_n, *where* $I_2 = (0, \pi^2/2)$ *and* $I_n = (0, \infty)$ *for* $n \geq 3$.

Proof. First, by [**LV**, (2.3), p. 60], $f_2(r) = \mu(t)\log((1+t)/(1-t))$, where $t = 2\sqrt{r}/(1+r)$, and the result follows from [**AVV4**, Lemma 2.6(4)]. Hence for $n \geq 3$ the result follows from [**AVV4**, Lemma 2.6(5)]. □

As shown in [**AVV4**, 1.21], the function M_n satisfies several nonlinear functional inequalities. Next we shall prove additional results of this type.

2.7. THEOREM. *For each* $n \geq 2$ *and* $s, t \in (0, 1)$,

$$M_n\left(\frac{2s}{1+s^2}\right) + M_n\left(\frac{2t}{1+t^2}\right) \leq 2M_n\left(\frac{2\sqrt{st}}{1+st}\right).$$

There is equality when $s = t$.

Proof. The statement about equality is trivial. For the inequality, let R denote the ring in $\mathbb{R}^n$ whose complementary components are the sets $[-e_1/(st), \infty] \cup [e_1/(st), \infty]$ and $[-e_1, e_1]|$. Then R is conformally equivalent to the Teichmüller ring $R_{T,n}((1-st)^2/(4st))$, and the sphere $S^{n-1}(1/s)$ separates R into two rings R_1 and R_2 that are conformally equivalent to $R_{G,n}((1+s^2)/(2s))$ and $R_{G,n}((1+t^2)/(2t))$, respectively. The result now follows from (2.1) and the superadditivity of the modulus [**F**]. □

2.8. COROLLARY. *For* $n \geq 2$, $a, b \in (0, 1)$,

$$M_n(a) + M_n(b) \leq 2M_n(\sqrt{ab}),$$

where equality holds if and only if $a = b$.

Proof. For $a = b$, equality is clear. Next, for $a \neq b$, let $s = a/(1+a')$, $t = b/(1+b')$, where $a' = \sqrt{1-a^2}$, $b' = \sqrt{1-b^2}$. Then

$a = 2s/(1+s^2)$, $b = 2t/(1+t^2)$, and Theorem 2.7 gives

$$\begin{aligned} M_n(a) + M_n(b) &\le 2M_n\left(\frac{2\sqrt{st}}{1+st}\right) < 2M_n\left(\frac{2\sqrt{st}}{\sqrt{(1+s^2)(1+t^2)}}\right) \\ &= 2M_n(\sqrt{ab}), \end{aligned}$$

since M_n is strictly decreasing. □

2.9. Remark. Corollary 2.8 strengthens [**AVV4**, (1.22)].

2.10. Theorem. *For each* $n \ge 2$, *let*

$$f_n(r) \equiv \frac{M_n(r)}{\log((1+r')/r)} \quad \text{and} \quad g_n(r) \equiv \frac{M_n(r)}{\log(\lambda_n(1+r')/(2r))}.$$

Then $f_n(r)$ *is strictly increasing from* $(0, 1)$ *onto* $(1, \infty)$ *and* $g_n(r)$ *is decreasing from* $(0, 1)$ *onto* $(0, 1)$.

Proof. By [**LV**, (2.5), p. 60] we have

$$f_2(r) = \frac{\mu(r)}{\log\left(\frac{1+r'}{r}\right)} = \frac{2\mu(r)}{\log\left(\frac{1+r'}{1-r'}\right)} = \frac{\mu(t)}{\log\left(\frac{1}{t}\right)},$$

where $t = (1-r')/(1+r')$, and it follows from [**AVV6**, Theorem 4.3(4)] and $f_2(r)$ is increasing from $(0, 1)$ onto $(1, \infty)$. For $n \ge 3$ it follows from [**AVV4**, Lemma 2.6(5)] that the function

$$f_n(r) = f_2(r) \cdot \frac{M_n(r)}{\mu(r)}$$

has the desired properties.

Next,

$$1 - g_n(r) = \frac{\log\frac{\lambda_n}{2} - \left(M_n(r) + \log\frac{r}{1+r'}\right)}{\log\frac{\lambda_n(1+r')}{2r}},$$

which, by Theorem 1.9, is increasing from $(0, 1)$ onto itself. □

3. Distortion inequalities in $\mathbb{R}^n$. In this section we obtain estimates for the distortion functions $\varphi_{K,n}$, $\varphi^*_{K,n}$, and $\psi_{K,n}$ introduced in (1.4) and (1.5). In [**AVV3**, Theorem 2.2] it was shown that $\varphi^*_{K,n} \ne \varphi_{K,n}$ for $n \ge 3$, $K \ne 1$. We now show that $\varphi^*_{K,n} \ne \psi_{K,n}$ and also that $\varphi_{K,n}$ and $\psi_{K,n}$ are not comparable. In our first result we apply a method of O. Hübner [**H**] (cf. [**LV**, pp. 64, 65], [**AVV1**, p. 698]) along with our Theorem 1.9 to derive inequalities that improve earlier distortion estimates [**AVV1**, Theorem 4.10].

3.1. **Theorem.** *For* $n \geq 2$, $K \geq 1$, $0 < r < 1$, *let* $\alpha = K^{1/(1-n)}$, $\beta = 1/\alpha$, $r' = \sqrt{1-r^2}$, $A(r) = r/(1+r')$. *Then*

$$A(\varphi_{K,n}(r)) \leq \left(\frac{\lambda_n}{2}\right)^{1-\alpha} A(r)^\alpha \leq KA(r)^\alpha, \tag{1}$$

$$A(\varphi_{1/K,n}(r)) \geq \left(\frac{\lambda_n}{2}\right)^{1-\beta} A(r)^\beta \geq K^{-\beta}A(r)^\beta, \tag{2}$$

$$\varphi_{K,n}(r) \leq \left(\frac{\lambda_n}{2}\right)^{1-\alpha} (1+r')^{1-\alpha}r^\alpha \leq K(1+r')^{1-\alpha}r^\alpha, \tag{3}$$

$$\varphi_{1/K,n}(r) \geq \left(\frac{\lambda_n}{2}\right)^{1-\beta} (1+r')^{1-\beta}r^\beta \geq K^{-\beta}(1+r')^{1-\beta}r^\beta. \tag{4}$$

Proof. Let $s = \varphi_{K,n}(r)$. Then $M_n(s) = \alpha M_n(r)$, $s \geq r$. Hence by Theorem 1.9

$$\begin{aligned}\log A(s) &\leq (1-\alpha)(M_n(r) + \log A(r)) + \alpha \log A(r) \\ &\leq (1-\alpha)\log\left(\frac{\lambda_n}{2}\right) + \alpha\log A(r) \leq \log K + \alpha \log A(r),\end{aligned}$$

and (1) follows. Since $\varphi_{K,n}^{-1} = \varphi_{1/K,n}$, (2) follows from (1). Finally, (1) implies (3), and (2) implies (4). □

3.2. **Corollary.** For $n \geq 2$, $K_2 \geq K_1 > 0$, $\alpha = (K_2/K_1)^{1/(1-n)}$, $0 < r < 1$,

$$\varphi_{K_2,n}(r) \leq \left(\frac{\lambda_n}{2}\right)^{1-\alpha} \varphi_{K_1,n}^\alpha(r)\left(1 + \sqrt{1-\varphi_{K_1,n}^2(r)}\right)^{1-\alpha}.$$

Proof. Since $\varphi_{K_2,n}(r) = \varphi_{K_2/K_1,n}(\varphi_{K_1,n}(r))$, the result follows from 3.1(3). □

3.3. **Remark.** We can obtain analogous inequalities for $\varphi_{K,n}^*$ by combining (1.7) with Theorem 3.1.

3.4. *Proof of Theorem* 1.11. Part (1) follows from Theorem 3.1(1), since $2\operatorname{arctanh} A(r) = \operatorname{arctanh} r$ and $LA(r)^\alpha < 1$ if $0 < r < r_1 = \tanh(2\operatorname{arctanh} l)$. Part (2) follows similarly from Theorem 3.1(2). □

3.5. THEOREM. *For* $n \geq 2$, $K \geq 1$, $\alpha = K^{1/(1-n)}$, $\beta = 1/\alpha$, $0 < r < 1$, $r' = \sqrt{1-r^2}$,

$$\begin{aligned} (1)\qquad \tanh(\beta \operatorname{arctanh} r) &\leq \varphi^*_{K,n}(r) \\ &\leq \tanh(\operatorname{arctanh} r + (\beta - 1)M_n(r')) \\ &\leq \tanh(\beta \operatorname{arctanh} r + (\beta - 1)\log(\lambda_n/2)) \end{aligned}$$

$$\begin{aligned} (2)\qquad \tanh(\alpha \operatorname{arctanh} r) &\geq \varphi^*_{1/K,n}(r) \\ &\geq \tanh(\operatorname{arctanh} r + (\alpha - 1)M_n(r')) \\ &\geq \tanh(\alpha \operatorname{arctanh} r + (\alpha - 1)\log(\lambda_n/2)). \end{aligned}$$

The inequalities reduce to equality when $K = 1$. *Moreover,*

$$(3)\qquad \begin{aligned} &\liminf_{r\to 0} r^{-\alpha} \operatorname{arctanh} \varphi^*_{K,n}(r) \geq 4^{1-\alpha}, \\ &\limsup_{r\to 0} r^{-\beta} \operatorname{arctanh} \varphi^*_{1/K,n}(r) \leq 4^{1-\beta}, \end{aligned}$$

$$(4)\qquad \lim_{r\to 1} \frac{\operatorname{arctanh} \varphi^*_{K,n}(r)}{\operatorname{arctanh} r} = \beta, \quad \lim_{r\to 1} \frac{\operatorname{arctanh} \varphi^*_{1/K,n}(r)}{\operatorname{arctanh} r} = \alpha,$$

whereas

$$(5)\qquad \begin{aligned} &\lim_{r\to 0} r^{-\alpha} \operatorname{arctanh} \varphi_{K,n}(r) = \lambda_n^{1-\alpha}, \\ &\lim_{r\to 0} r^{-\beta} \operatorname{arctanh} \varphi_{1/K,n}(r) = \lambda_n^{1-\beta}, \end{aligned}$$

$$(6)\qquad \lim_{r\to 1} \frac{\operatorname{arctanh} \varphi_{K,n}(r)}{\operatorname{arctanh} r} = K, \quad \lim_{r\to 1} \frac{\operatorname{arctanh} \varphi_{1/K,n}(r)}{\operatorname{arctanh} r} = 1/K.$$

Proof. From [**AVV2**, 2.18] and [**AVV1**, 4.4]

$$\varphi^*_{K,n}(r) \geq \varphi_{\beta,2}(r) \geq \tanh(\beta \operatorname{arctanh} r)$$

and

$$\varphi^*_{1/K,n}(r) \leq \varphi_{\alpha,2}(r) \leq \tanh(\alpha \operatorname{arctanh} r).$$

Next, with $s = \varphi^*_{K,n}(r)$, $s' = \sqrt{1-s^2} \leq r' = \sqrt{1-r^2}$, Theorem 1.9 and (1.7) give

$$M_n(r') - \operatorname{arctanh} r \leq M_n(s') - \operatorname{arctanh} s \leq \beta M_n(r') - \operatorname{arctanh} s,$$

and the second inequality in (1) follows. Then Corollary 1.10(1) yields the third inequality in (1). The proof of the first and second inequalities in (2) is similar.

The first limit in (3) follows from [**AVV1**, Theorem 4.9] and [**LV**, p. 65]; and the second follows from the first by inversion. The limits in

(4) follow directly from (1) and (2); (5) follows from [**AVV4**, (3.7)]; and (6) is implied by [**AVV1**, Theorem 4.4]. □

Theorem 3.5 yields the following sharp quasiconformal distortion result for the hyperbolic metric in B^n.

3.6. COROLLARY. *If* $f \in QC_K(B^n)$ *and* $\beta = K^{1/(n-1)} = 1/\alpha$, *then*

$$(1)\qquad \begin{aligned}\rho(f(x), f(y)) &\le \rho(x, y) + 2(\beta - 1)M_n\left(\operatorname{sech}\frac{\rho(x, y)}{2}\right)\\ &\le \beta\rho(x, y) + 2(\beta - 1)\log\frac{\lambda_n}{2}\end{aligned}$$

for all $x, y \in B^n$. *Moreover, if* $f(B^n) = B^n$, *then we also get the reverse inequality,*

$$(2)\qquad \begin{aligned}&\alpha\rho(x, y) - 2(1 - \alpha)\log\frac{\lambda_n}{2}\\ &\le \rho(x, y) - 2(1 - \alpha)M_n\left(\operatorname{sech}\left(\frac{\rho(x, y)}{2}\right)\right)\\ &\le \rho(f(x), f(y)).\end{aligned}$$

Proof. Denoting $\tanh(\rho/2) = r$, $\tanh(\rho'/2) = s$, $\rho = \rho(x, y)$, $\rho' = \rho(f(x), f(y))$, then applying 3.5(1) and [**Vu2**, 11.2], [**MRV**], we get

$$\begin{aligned}\rho' &= 2\operatorname{arctanh} s \le 2\operatorname{arctanh}\varphi^*_{K,n}(r)\\ &\le \rho(x, y) + 2(\beta - 1)M_n\left(\operatorname{sech}\left(\frac{\rho(x, y)}{2}\right)\right)\\ &\le \beta\rho(x, y) + 2(\beta - 1)\log\frac{\lambda_n}{2},\end{aligned}$$

proving (1). Inequality (2) follows from (1) applied to f^{-1}. □

3.7. COROLLARY. *For* $K \ge 1$, $n \ge 2$, $\beta = K^{1/(n-1)} = 1/\alpha$, $0 < r < 1$,

$$(1)\qquad \left(\frac{1+r}{1-r}\right)^{\beta} \le \frac{1 + \varphi^*_{K,n}(r)}{1 - \varphi^*_{K,n}(r)} \le \left(\frac{\lambda_n}{2}\right)^{2(\beta-1)}\left(\frac{1+r}{1-r}\right)^{\beta},$$

$$(2)\qquad \left(\frac{\lambda_n}{2}\right)^{2(\alpha-1)}\left(\frac{1+r}{1-r}\right)^{\alpha} \le \frac{1 + \varphi^*_{1/K,n}(r)}{1 - \varphi^*_{1/K,n}(r)} \le \left(\frac{1+r}{1-r}\right)^{\alpha}.$$

Proof. Inequalities (1), (2) follow from Theorem 3.5(1), (2) respectively. □

In a recent paper D. Cooper [C] has shown that, for a class of homeomorphisms of B^n onto itself keeping 0 fixed, the image of $S^{n-1}(r)$ is close to $S^{n-1}(r)$ in a certain sense. (For an alternative proof see J. Väisälä [V3].) More precisely, Cooper showed, for a given mapping f in this class and for r close to 1 and $\varepsilon > 0$, that $S^{n-1}(r) = A \cup E$, where $A \cap E = \varnothing$ and

$$f(A) \subset \bigcup_{|z|=r} D(z, \varepsilon); \qquad D(z, \varepsilon) = \{w \in B^n : \rho(z, w) < \varepsilon\}$$

and the exceptional set E is small in the $(n-1)$-dimensional measure. Cooper's class of mappings includes quasiconformal mappings in particular. We shall now apply Corollary 3.7 to show that for every $\varepsilon > 0$ Cooper's exceptional set E is nonempty for r close to 1, even for K-quasiconformal mappings when $K > 1$. Indeed, with $s = \varphi^*_{K,n}(r)$ from Corollary 3.7 we obtain

$$\rho(S^{n-1}(r), S^{n-1}(s)) = \log\left(\frac{1+s}{1-s}\frac{1-r}{1+r}\right) \ge (\beta - 1) \cdot \log\frac{1+r}{1-r},$$

which tends to ∞ as $r \to 1$. Thus given $K > 1$ there exists a K-quasiconformal mapping f of B^n onto B^n keeping 0 fixed and for which there are points on $S^{n-1}(r)$ whose images under f are at hyperbolic distance at least $((\beta - 1)/2)\log((1+r)/(1-r))$ from $S^{n-1}(r)$. In other words, Cooper's exceptional set is nonempty for all $K > 1$ and for all r sufficiently close to 1.

3.8. *Proof of Theorem* 1.14. Let $r = \tanh(\rho/2)$, $s = \tanh(\rho'/2)$, $r' = \sqrt{1-r^2}$, $s' = \sqrt{1-s^2}$, and $\alpha = K^{1/(1-n)}$. Then by Theorem 3.1, [**AVV4**, (3.22)], and the Schwarz lemma [**MRV**, 3.1],

$$\begin{aligned}\tanh\frac{\rho'}{4} &= \frac{s}{1+s'} \le \frac{\varphi_{K,n}(r)}{1+\sqrt{1-\varphi_{K,n}(r)^2}} \le \left(\frac{\lambda_n}{2}\right)^{1-\alpha}\left(\frac{r}{1+r'}\right)^{\alpha} \\ &= \left(\frac{\lambda_n}{2}\right)^{1-\alpha}\left(\tanh\left(\frac{\rho}{4}\right)\right)^{\alpha} \le K\left(\tanh\frac{\rho}{4}\right)^{\alpha} \le K\left(\tanh\frac{\rho}{4}\right)^{1/K}.\end{aligned}$$

When $K = 1$ the inequality in the theorem reduces to $\rho' \le \rho$, which by conformal invariance is an equality if f is one-to-one and $f(B^n) = B^n$. □

3.9. *Proof of Corollary* 1.15. This follows immediately from Theorem 1.14 and [**AVV2**, (2.5)]. □

We shall require the following simple inequality (cf. e.g. **[Vu2**, (2.27)]):

$$|x-y| \le 2\tanh\frac{\rho(x,y)}{4} \le \frac{|x-y|}{\sqrt{(1-|x|^2)(1-|y|^2)}} \tag{3.10}$$

for $x, y \in B^n$.

3.11. *Proof of Theorem* 1.12. For (1), by Theorem 1.14 and (3.10) for $x, y \in \overline{B}^n$, with $\rho = \rho(x,y)$, $\rho' = \rho(f(x), f(y))$, we have

$$\begin{aligned}|f(x)-f(y)| &\le 2\tanh\frac{\rho'}{4} \le 2\left(\frac{\lambda_n}{2}\right)^{1-\alpha}\left(\tanh\frac{\rho}{4}\right)^{\alpha} \\ &\le \lambda_n^{1-\alpha}(1-r^2)^{-\alpha}|x-y|^{\alpha}.\end{aligned}$$

The proof of (2) is similar. □

The next result is elementary.

3.12. Lemma. *For $K \ge 1$ and $r \in (0,1)$, $r' = \sqrt{1-r^2}$, the following inequalities hold:*

$$\begin{aligned} r < \sqrt{1-r'^{2K}} &\le \tanh(K \operatorname{arctanh} r) \\ &\le \min\{Kr, \sqrt{1-4^{1-K}r'^{2K}}\}, \end{aligned} \tag{1}$$

$$\begin{aligned} \sqrt{\max\left\{\left(\frac{r}{K}\right)^2, 1-4^{1-1/K}r'^{2/K}\right\}} &\le \tanh\left(\frac{1}{K}\operatorname{arctanh} r\right) \\ &\le \sqrt{1-r'^{2/K}} < r. \end{aligned} \tag{2}$$

3.13. Theorem. *For $n \ge 2$, $0 \le r \le 1$, $1 \le K < \infty$, $\beta = K^{1/(n-1)}$, $r' = \sqrt{1-r^2}$,*

$$\sqrt{1-r'^{2\beta}} \le \varphi^*_{K,n}(r) \le \psi_{K,n}(r) \le \sqrt{1-\lambda_n^{2(1-\beta)}r'^{2\beta}}.$$

Equality holds for $K = 1$. The upper bound for $\psi_{K,n}(r)$ is asymptotically sharp as r tends to 1.

Proof. By Theorem 3.5(1) and Lemma 3.12 we obtain

$$\varphi^*_{K,n}(r) \ge \tanh(\beta \operatorname{arctanh} r) \ge \sqrt{1-r'^{2\beta}}.$$

The second inequality follows from (1.7). Next, by (1.4) and Theorem 3.1(4) we have

$$\psi^2_{K,n}(r) = 1-\varphi^2_{1/K,n}(r') \le 1-\lambda_n^{2(1-\beta)}r'^{2\beta},$$

proving the third inequality. The statement about equality is clear. Finally, the third inequality is asymptotically sharp as $r \to 1$ since

$$\lim_{r\to 1-} r'^{-2\beta}(1-\psi_{K,n}^2(r)) = \lim_{r\to 1-} r'^{-2\beta}\varphi_{1/K,n}^2(r') = \lambda_n^{2(1-\beta)}$$

by (1.4) and [**AVV1**, (4.12)]. □

3.14. Corollary. *Let* $n \geq 3$, $m \geq 2$, $1 < K < \infty$, *and* $0 < r < 1$.

(1) *If* r *is close to* 1 *then* $\varphi_{K,n}^*(r) < \varphi_{K,m}(r)$.

(2) *If* $m > n$ *and* r *is close to* 0, *then the above inequality is reversed.*

Proof. For (1) choose r such that $(1-\lambda_n^c)^{1/2} < r < 1$, where $c = 2(1-\beta)/(K-\beta)$, $\beta = K^{1/(n-1)}$. Then

$$1-\lambda_n^{2(1-\beta)}r'^{2\beta} < 1-r'^{2K}$$

and the result follows from Lemma 3.12 and Theorem 3.13.

By [**AVV1**, Theorem 4.10] and [**AVV2**, (2.18)] we get

$$\lim_{r\to 0}\varphi_{K,n}^*(r)/\varphi_{K,m}(r) = \infty,$$

and hence (2) follows. □

4. High dimensions.

In this section we prove Theorem 1.17, by using the following result from [**Vu3**].

4.1. Theorem. *The linear dilatation in* (1.16) *has the following majorant*:

$$H_n(K) \leq \inf_{0<t<1}\frac{A_{K,n}(t)}{B_{K,n}(t)},$$

where

$$A_{K,n}(t) = \frac{\varphi_{K,n}^2(\sqrt{t})}{1-\varphi_{K,n}^2(\sqrt{t})}, \quad B_{K,n}(t) = \varphi_{1/K,n}^2(\sqrt{t/(1+t)}).$$

4.2. *Proof of Theorem* 1.17. Fix $t \in (0, 4^{1-K}K^{-2})$. Then by Theorem 3.1(3), (4) and [**AVV4**, (3.23)],

$$\limsup_{n\to\infty} A_{K,n}(t) \leq \lim_{n\to\infty}\frac{\lambda_n^{2(1-\alpha)}t^\alpha}{1-\lambda_n^{2(1-\alpha)}t^\alpha} = \frac{K^2t}{1-K^2t},$$

$$\liminf_{n\to\infty} B_{K,n}(t) \geq \lim_{n\to\infty}\lambda_n^{2(1-\beta)}\left(\frac{t}{1+t}\right)^\beta = \frac{t}{K^2(1+t)}.$$

Hence

$$\limsup_{n\to\infty} \frac{A_{K,n}(t))}{B_{K,n}(t)} \le \frac{K^4(1+t)}{1-K^2t},$$

so that

$$\limsup_{n\to\infty} H_n(K) \le \limsup_{n\to\infty} \left(\inf_{0<t<1} \frac{A_{K,n}(t)}{B_{K,n}(t)} \right) \le K^4. \qquad \square$$

4.3. *Conjectures.* (1) $\lim_{n\to\infty} H_n(K) = 1$ for all $K > 1$.

(2) $\lim_{n\to\infty} \varphi^*_{K,n}(r) = r$ for all $K > 1$ and $0 < r < 1$.

(3) $M_n(r)$ can be replaced by $M_n^{n-1}(r)$ in Theorem 2.6.

4.4. Remark. We observe that [**AVV1**, (4.5), (4.11)] implies that

$$r < \liminf_{n\to\infty} \varphi_{K,n}(r) \le \limsup_{n\to\infty} \varphi_{K,n}(r) \le Kr$$

for all $K > 1$ and $0 < r < 1$, so that the analog of 4.3(2) is false for $\varphi_{K,n}(r)$ if $K > 1$.

References

[A1] G. D. Anderson, *Extremal rings in n-space for fixed and varying n*, Ann. Acad. Sci. Fenn. Ser. A I, **575** (1974), 1–21.

[A2] ——, *Derivatives of the conformal capacity of extremal rings*, Ann. Acad. Sci. Fenn. Ser. A I, **10** (1984), 29–46.

[AF] G. D. Anderson and J. S. Frame, *Numerical estimates for a Grötzsch ring constant*, Constr. Approx., **4** (1988), 223–242.

[AVV1] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Dimension-free quasiconformal distortion in n-space*, Trans. Amer. Math. Soc., **297** (1986), 687–706.

[AVV2] ——, *Sharp distortion theorems for quasiconformal mappings*, Trans. Amer. Math. Soc., **305** (1988), 95–111.

[AVV3] ——, *Inequalities for the extremal distortion function*, Proc. 13th Rolf Nevanlinna Colloquium (Joensuu, 1987), Lecture Notes in Math., Vol. 1351, pp. 1–11, Springer-Verlag, Berlin and Heidelberg, New York, 1988.

[AVV4] ——, *Special functions of quasiconformal theory*, Exposition. Math., **7** (1989), 97–138.

[AVV5] ——, *Distortion functions for plane quasiconformal mappings*, Israel J. Math., **62** (1988), 1–16.

[AVV6] ——, *Functional inequalities for complete elliptic integrals and their ratios*, SIAM J. Math. Anal., **21** (1990), 536–549.

[B] A. F. Beardon, *The Geometry of Discrete Groups*, Graduate Texts in Math. Vol. 91, Springer-Verlag, Berlin, Heidelberg and New York, 1983.

[BF] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Physicists*, Die Grundlehren der math. Wissenschaften, Vol. 57, Springer-Verlag, Berlin, Göttingen and Heidelberg, 1954.

[Ca] E. D. Callendar, *Hölder of n-dimensional quasiconformal mappings*, Pacific J. Math., **10** (1960), 499–515.

[C] D. Cooper, *Quasi-isometries of hyperbolic space are almost isometries*, Manuscript U.C.S.B., September 1988.

[F] B. Fuglede, *Extremal length and functional completion*, Acta Math., **98** (1957), 171–219.

[G] F. W. Gehring, *Symmetrization of rings in space*, Trans. Amer. Math. Soc., **101** (1961), 499–519.

[H] O. Hübner, *Remarks on a paper by Lawrynowicz on quasiconformal mappings*, Bull. de l'Acad. Polon. des Sci., **18** (1970), 183–186.

[LV] O. Lehto and K. I. Virtanen, *Quasiconformal Mappings in the Plane*, 2nd ed., Die Grundlehren der math. Wissenschaften, Band 126, Springer-Verlag, New York, Heidelberg and Berlin, 1973.

[MRV] O. Martio, S. Rickman and J. Väisälä, *Distortion and singularities of quasiregular mappings*, Ann. Acad. Sci. Fenn. Ser. AI, **465** (1970), 1–13.

[R] Yu. G. Reshetnyak, *Space Mappings With Bounded Distortion*, Transl. of Math. Monographs, Vol. 73, Amer. Math. Soc., Providence, RI, 1989.

[T] O. Teichmüller, *Untersuchungen über konforme und quasikonforme Abbildung*, Deutsche Math., **3** (1938), 621–678.

[V1] J. Väisälä, *Lectures on n-Dimensional Quasiconformal Mappings*, Lecture Notes in Math. Vol. 229, Springer-Verlag, Berlin, Heidelberg and New York, 1971.

[V2] ——, *A survey of quasiregular maps in* $\mathbb{R}^n$, Proc. Internat. Congr. Math. (Helsinki, 1978), Vol. 2, 685–691, Acad. Sci. Fennica, Helsinki, 1980.

[V3] ——, *Free quasiconformality in Banach spaces* II, Ann. Acad. Sci. Fenn. Ser. A. I. Math., **16** (1991), 255–310.

[Vu1] M. Vuorinen, *Conformal invariants and quasiregular mappings*, J. Analyse Math., **45** (1985), 69–115.

[Vu2] ——, *Conformal Geometry and Quasiregular Mappings*, Lecture Notes in Math., Vol. 1319, Springer-Verlag, Berlin, Heidelberg and New York, 1988.

[Vu3] ——, *Quadruples and spatial quasiconformal mappings*, Math. Z., **205** (1990), 617–628.

Received January 22, 1991 and in revised form January 27, 1992.

MICHIGAN STATE UNIVERSITY
EAST LANSING, MI 48824 U.S.A.

UNIVERSITY OF AUCKLAND
AUCKLAND, NEW ZEALAND

AND

UNIVERSITY OF HELSINKI
HELSINKI, FINLAND

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

A NONEXISTENCE RESULT FOR THE n-LAPLACIAN

TILAK BHATTACHARYA

Let P be a point in $\mathbb{R}^n$, $n \geq 2$; then the problem $\operatorname{div}(|\nabla u|^{n-2}\nabla u) = e^u$ **with** $u \in W^{1,n}_{\text{loc}} \cap L^\infty_{\text{loc}}$ **has no subsolutions in** $\mathbb{R}^n \backslash \{P\}$.

Introduction. Let $P = P(x_1, x_2, \dots, x_n)$ be a point in $\mathbb{R}^n$, $n \geq 2$, and $\Omega = \mathbb{R}^n \backslash \{P\}$. Without any loss of generality we will take P to be the origin. Consider the problem

$$(1.1) \qquad \begin{cases} L_p u = e^u & \text{in } \Omega, \\ u \in W^{1,p}_{\text{loc}}(\Omega) \cap L^\infty_{\text{loc}}(\Omega); & p > 1. \end{cases}$$

Here $L_p u \equiv \operatorname{div}(|\nabla u|^{p-2}\nabla u)$ is the p-Laplacian with $1 < p < \infty$. By a subsolution u of (1.1) we will mean that $u \in W^{1,p}_{\text{loc}}(\Omega) \cap L^\infty_{\text{loc}}(\Omega)$, and

$$\int_\Omega |\nabla u|^{p-2}\nabla u,\ \nabla\psi + \int_\Omega e^u \psi \leq 0, \quad \forall \psi \in C_0^\infty(\Omega) \text{ and } \psi \geq 0.$$

It is known that for $1 < p < n$, (1.1) has no subsolutions in the exterior of a compact set [AW]. However, for $p = n$ there exist radial subsolutions for large values of $|x|$. We show that (1.1) has no subsolutions in Ω, thus extending the results of [AW], namely

THEOREM 1. *The following problem*

$$L_n u = e^u \quad \text{in } \Omega, \ n \geq 2,$$

has no subsolutions in $W^{1,n}_{\text{loc}}(\Omega) \cap L^\infty_{\text{loc}}(\Omega)$.

The proof of Theorem 1 will be a consequence of a comparison principle and nonexistence of global radial solutions. The proof is presented in §4.

2. Preliminary results.

LEMMA 2.1. *Consider*

$$C(x) = \frac{(1+x)^{1/n}}{1+x^{1/n}} \quad \text{in } 0 \leq x \leq 1.$$

Then $C(x)$ *is decreasing on* $[0, 1]$.

Proof. Elementary computations show that

$$\frac{dC}{dx} = \frac{(1+x)^{1/n}(1-x^{(1-n)/n})}{n(1+x^{1/n})^2(1+x)} \le 0$$

in $0 \le x \le 1$. Furthermore, $C(0) = 1$ and $C(1) = 2^{1-n/n}$, and $C(x) \to 1$ as $x \to 0$. □

We now state an elementary inequality that is easy to prove

$$x^n - b^n \ge (x-b)^n, \quad \text{for } x \ge b \ge 0. \tag{2.1}$$

LEMMA 2.2. *Suppose $u(r) \in C^1$ satisfies the following differential inequality in (a, R),*

$$\dot{u} \ge A\left(e^{u/n} + \frac{B-b}{R-r}\right),$$

where $\dot{u}$ represents differentiation with respect to r, $0 < A < 1$, $0 < b < 1$, $0 < a < R$ and $B \ge \frac{n}{A} + b$. Then there is an $\bar{r}$ in (a, R) such that $u(r) \to \infty$ as $r \to \bar{r}$.

Proof. Setting $v = e^{-u/n}$, we obtain that

$$\dot{v} + \frac{c}{R-r}v \le -\frac{A}{n}, \quad a < r < R,$$

where $c = \frac{A(B-b)}{n}$. Using the integrating factor $\phi(r) = (\frac{1}{R-r})^c$ and setting $Z = v(r)\phi(r) - v(a)\phi(a)$, we obtain

$$Z \le \begin{cases} \left(-\dfrac{A}{n}\right)\ln\dfrac{R-a}{R-r}; & c = 1, \\ \left(-\dfrac{A}{n}\right)\left(\dfrac{1}{c-1}\right)\left\{\left(\dfrac{1}{R-r}\right)^{c-1} - \left(\dfrac{1}{R-a}\right)^{c-1}\right\}; & c > 1. \end{cases}$$

It is clear that for each $c \ge 1$, there is an $\bar{r} \in (a, R)$ such that $v(r) \to 0$ as $r \to \bar{r}$, and hence $u(r) \to \infty$ as $r \to \bar{r}$. □

We present a comparison lemma; please refer to [AW] for its proof.

LEMMA 2.3. *In a region $(\Omega) \subseteq R^n$, $n \ge 2$, suppose $u, v \in W^{1,p}_{\mathrm{loc}}(\Omega) \cap L^\infty_{\mathrm{loc}}(\Omega)$, and $(u-v)^+ \in W^{1,p}_0(\Omega)$. If g is a nondecreasing function and*

$$L_p u \ge g(u) \quad \text{in } D'(\Omega),$$
$$L_p v \le g(v) \quad \text{in } D'(\Omega),$$

then $u \le v$ a.e. in (Ω).

3. Nonexistence of radial subsolutions. Consider the following problem

$$(3.1)\qquad (n-1)|\dot{u}|^{n-2}\left(\ddot{u}+\frac{\dot{u}}{r}\right)=e^{u}, \qquad 0<r<\infty, \\ u(R)=a, \quad \text{and} \quad \dot{u}(R)=b; \qquad a, b\in R.$$

LEMMA 3.1. *For the problem in* (3.1), *there exists a* C^1 *radial solution* $u(r)$ *such that at least one of the following occurs.*

(i) *There is an* $\bar{r}$ *in* $(0, R)$ *such that* $u(r)\to\infty$ *as* $r\to\bar{r}$.

(ii) *There is an* $\bar{r}$ *in* (R, ∞) *such that* $u(r)\to\infty$ *as* $r\to\bar{r}$.

Furthermore, there are values of b *for which both* (i) *and* (ii) *occur.*

Proof. We divide the proof into three parts.

Case 1. Take $b=0$. Let $u(r)$ be the solution defined by

$$(3.2)\qquad u(r)=a+\int_R^r\frac{1}{t}\left\{\int_R^t s^{n-1}e^{u(s)}\,ds\right\}^{1/(n-1)}dt,$$

in $r>R$. The existence and uniqueness in a small interval follows from Picard's iteration. It can be shown by differentiating that u solves (3.1). From (3.2) it is clear that $r\dot{u}$ is increasing and thus $\dot{u}\geq 0$ in (R, r), and hence u is increasing. Continue u by (3.2). By differentiating (3.2) once,

$$\dot{u}(r)=\frac{1}{r}\left\{\int_R^r s^{n-1}e^{u(s)}\,ds\right\}^{1/(n-1)}.$$

Thus,

$$\frac{d}{dr}\left\{\frac{(\dot{u})^{n-1}}{r}\right\}=\frac{r^n e^{u(r)}-n\int_R^r s^{n-1}e^{u(s)}\,ds}{r^{n+1}} \\ \geq\frac{r^n e^{u(r)}-e^{u(r)}(r^n-R^n)}{r^{n+1}}\geq 0.$$

By simplifying the left side of the foregoing inequality,

$$(n-1)\ddot{u}\geq\frac{\dot{u}}{r}.$$

Note that u is C^2 except possibly where $\dot{u}=0$. Noting that $\dot{u}\geq 0$, (3.1) yields

$$n(n-1)(\dot{u})^{n-1}\ddot{u}\geq e^{u}, \qquad R<r<\infty.$$

Multiplying both sides by $\dot{u}$ and integrating once from R to r,

$$(\dot{u})^n \ge \frac{e^u - e^a}{n-1}. \tag{3.3}$$

For $\varepsilon > 0$, small enough, it follows from (3.2) and the fact that u is increasing that

$$u(r) > a + \int_{R+\varepsilon}^{r} \frac{1}{t}\left\{\int_R^{R+\varepsilon} s^{n-1} e^{u(s)}\, ds\right\}^{1/(n-1)} dt.$$

Hence for some appropriate constant $A > 0$,

$$u(r) > a + A \ln \frac{r}{R+\varepsilon}$$

implying that $u(r) \to \infty$ as r gets large. Thus in (3.3) we may take $r > R_1$, where R_1 is large enough so that $e^u/2 \le e^u - e^a$ for $r > R_1$. If $u(r) \to \infty$ as $r \to R_1$, then we are done. Otherwise, continue u using (3.2) past $r = R_1$. Hence

$$\dot{u} \ge C e^{u/n}, \quad \text{in } r > R_1,$$

for some $C > 0$. Integrating,

$$\int_{u(R_1)}^{u(r)} e^{-u/n}\, du \ge C(r - R_1).$$

It is clear that there exists an $\bar{r} > R$, such that $u(r) \to \infty$ as $r \to \bar{r}$. The case $b > 0$ follows similarly.

Case 2. Without any loss of generality, take $a = 0$. Take $b < 0$. Now $\dot{u}(r) < 0$ near $r = R$, so we obtain that $\dot{u}(r)$ satisfies

$$\dot{u}(r) = -\frac{1}{r}\left\{|bR|^{n-1} - \int_R^r t^{n-1} e^{u(t)}\, dt\right\}^{1/(n-1)}, \tag{3.4}$$

in $r > R$. We show that there is $\bar{b} < 0$ such that if $\bar{b} < b < 0$, there is an $\hat{r} > R$ such that $\dot{u}(r) \to 0$ as $r \to \hat{r}$. It follows from (3.4) that $r\dot{u}$ is increasing and thus

$$\frac{bR}{r} \le \dot{u} \le 0, \quad \text{for } r > R.$$

Set $c = bR$. Integrating, we find

$$e^u \ge r^c,$$

and so (3.4) yields

$$\dot{u}(r) \ge -\frac{1}{r}\left\{|c|^{n-1} - \int_R^r t^{n-1+c}\, dt\right\}^{1/(n-1)}.$$

Therefore,

$$\dot{u}(r) \geq \begin{cases} -\frac{1}{r}\left\{|c|^{n-1} - \dfrac{r^{n+c} - R^{n+c}}{n+c}\right\}^{1/(n-1)} ; & -n < c < 0, \\ -\frac{1}{r}\left\{|c|^{n-1} - \ln\frac{r}{R}\right\}^{1/(n-1)} ; & c = -n. \end{cases}$$

It is clear that there is an $\hat{r} > R$ for which $\dot{u}(r) \to 0$ as $r \to \hat{r}$. Now, take $c < -n$, satisfying

$$|c|^{n-1} - \frac{1}{|c|-n}\left(\frac{1}{R}\right)^{|c|-n} < n^{n-1}. \tag{3.5}$$

Now, (3.4) yields

$$\dot{u}(r) \geq -\frac{1}{r}\left[|c|^{n-1} - \frac{1}{|c|-n}\left\{\left(\frac{1}{R}\right)^{|c|-n} - \left(\frac{1}{r}\right)^{|c|-n}\right\}\right]^{1/(n-1)}.$$

Using (3.5), there is an $\tilde{r}$ such that $\dot{u}(r) \geq -\frac{n}{r}$ for $r > \tilde{r}$. If $\dot{u}(r) \to 0$ as $r \to \tilde{r}$, then we are done. Otherwise, continue u past $r = \tilde{r}$. Repeating the arguments preceding (3.5), we see that $\dot{u}(r) \to 0$ as $r \to \hat{r}$ for some $\hat{r} > R$. Continuing u past $r = \hat{r}$ using

$$u(r) = u(\hat{r}) + \int_{\hat{r}}^{r} \frac{1}{t}\left\{\int_{\hat{r}}^{t} s^{n-1}e^{u(s)}\,ds\right\}^{1/(n-1)} dt,$$

we may show, as in Case 1, that there is an $\bar{r} > R$ where u blows up.

Case 3. We may again take $a = 0$. Let $c < -n$, $t = R - r$, and $v(t) = u(r)$, where $0 < r \leq R$. Then $\dot{v}(t) = -\dot{u}(r)$, where $\dot{v}$ represents differentiation with respect to t. Then

$$(n-1)|\dot{v}|^{n-2}\left(\ddot{v} - \frac{\dot{v}}{R-t}\right) = e^{v}, \qquad 0 \leq t < R, \tag{3.6}$$
$$v(0) = 0 \quad \text{and} \quad \dot{v}(0) = -b.$$

A solution of (3.6) is given by

$$v(t) = \int_0^t \frac{1}{R-s}\left\{|c|^{n-1} + \int_0^s (R-w)^{n-1}e^{v(w)}\,dw\right\}^{1/(n-1)} ds.$$

Equation (3.6) yields that $\frac{d}{dt}\{(R-t)\dot{v}\} \geq 0$, thus $\dot{v} \geq 0$ in $t > 0$. Integrating this inequality from 0 to t, we obtain

$$\dot{v}(t) \geq \frac{|c|}{(R-t)}.$$

Hence,

$$e^{v(t)} \geq \left(\frac{1}{R-t}\right)^{|c|}. \tag{3.7}$$

Let $0 < \varepsilon_0 < 1$ be such that

$$|c| \geq n\left\{\frac{1+\varepsilon^{1/n}}{(1+\varepsilon)^{1/n}}\right\} + \varepsilon$$

for every ε in $(0, \varepsilon_0)$. It follows from (3.7) that there is a $t_1 < R$ such that

$$\left(\frac{|c|}{R-t}\right)^n e^{-v(t)} < \varepsilon_0,$$

for $t > t_1$. If $v(t) \to \infty$ as $t \to t_1$, then we are done; otherwise continue $v(t)$ past $t = t_1$. Furthermore, we may take t_1 such that $R - t_1 < \varepsilon_0$. Rearranging the terms in (3.6), and multiplying by $\dot{v}(t)$ yields

$$(n-1)(\dot{v})^{n-1}\ddot{v} = e^v\dot{v} + \frac{n-1}{R-t}(\dot{v})^n, \qquad 0 \leq t < R.$$

Integrating both sides from 0 to t, and noting that $\dot{v} \geq \frac{|c|}{R-t}$, we find

$$(\dot{v})^n \geq e^v - 1 + \left(\frac{|c|}{R-t}\right)^n, \qquad 0 \leq t < R.$$

By the definition of t_1, it follows that

$$(\dot{v})^n \geq e^v + \left(\frac{|c| - \varepsilon_0}{R-t}\right)^n, \qquad t_1 < t < R.$$

Setting

$$x = \left(\frac{|c| - \varepsilon_0}{R-t}\right)^n e^{-v},$$

the above may be rewritten as

$$(\dot{v})^n \geq e^v\{1 + x\}.$$

Hence,

$$\dot{v} \geq e^{v/n}\{1 + x\}^{1/n}.$$

Using Lemma 2.1 and the definition of t_1,

$$\dot{v} \geq C(\varepsilon_0)e^{v/n}\{1 + x^{1/n}\}.$$

Thus we obtain

$$\dot{v} \geq C(\varepsilon_0)\left\{e^{v/n} + \frac{|c| - \varepsilon_0}{R-t}\right\}, \qquad t_1 < t < R.$$

By Lemma 2.2, there is a $t_2 > t_1$ such that $v(t) \to \infty$ as $t \to t_2$. Hence there is an $\bar{r} \in (0, R)$ for which $u(r) \to \infty$ as $r \to \bar{r}$. Thus for every $c < -n$, we have a vertical asymptote in $(0, R)$. It is clear from (3.5) that there are values of b for which both (i) and (ii) happen. Call one such value to be b_R.

For the case $a \neq 0$, we introduce the following change of variables. Let $v(r) = u(r) - a$; then

$$(n-1)|\dot{v}|^{n-2}\left(\ddot{v} + \frac{n-1}{r}\dot{v}\right) = e^a e^v .$$

Setting $t = re^{a/n}$, and $w(t) = v(r)$, and differentiating with respect to t, we have

$$(n-1)|\dot{w}|^{n-2}\left(\ddot{w} + \frac{n-1}{t}\dot{w}\right) = e^w ,$$

$$w(\overline{R}) = 0 \quad \text{and} \quad \dot{w}(\overline{R}) = e^{-a/n} b ,$$

where $\overline{R} = e^{a/n}R$. There is a $b_{\overline{R}}$ so that the corresponding solution which we continue to call $w(t)$, blows up near zero and at a point past $\overline{R}$. Then $u(t) = a + w(e^{-a/n}t)$ is such a solution for the original problem. □

4. Proof of Theorem 1.

This follows easily from Lemma 2.3 and Lemma 3.1.

Proof of Theorem 1. Assume to the contrary. Let $U(x)$ be such a subsolution in (1.2). Let

$$a = \inf_{1/2 \le |x| \le 3/2} U(x) .$$

By Lemma 3.1, there is a radial solution $u(r)$ such that $u(1) = a - 1$, and $u(r)$ blows up at some $\underline{r} \in (0, 1)$ and $\bar{r} \in (1, \infty)$. Let

$$M = \sup_{\underline{r} \le |x| \le \bar{r}} U(x) ,$$

$\underline{\underline{r}} \in (\underline{r}, 1)$ and $\bar{\bar{r}} \in (1, \bar{r})$ be such that $u(\underline{\underline{r}})$, $u(\bar{\bar{r}}) \ge M + 1$. Using Lemma 2.3, $u(x) \ge U(x)$ in $\underline{\underline{r}} \le |x| \le \bar{\bar{r}}$, a contradiction. □

REMARK. In Theorem 1, $1 < p \le n$ is the best possible. For $p > n$, take $u = \ln(\frac{A}{r^p})$, where $0 < A \le (p-n)p^{p-1}$. Then

$$L_p u = \frac{(p-n)p^{p-1}}{r^p} \ge \frac{A}{r^p} .$$

REFERENCES

[AD] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, (1975).

[AW] P. Avilẽs and A. Weitsman, *On the singularities of certain non-linear partial differential equations*, Ann. Acad. Sc. Fenn. **7** series A. I. Mathematics, (1982), 147–152.

[BH1] T. Bhattacharya, *Radial symmetry of the first eigenfunction of the p-Laplacian in the ball*, Proc. Amer. Math. Soc., **104** no. 1 (1988).

[BH2] ——, *Some results concerning the eigenvalue problem for the p-Laplacian*, to appear in Ann. Acad. Sci. Fenn.

[S] J. Serrin, *Local behaviour of solutions of quasilinear equations*, Acta Math., **111** (1964), 247–302.

[VA] J. L. Vazquez, *An apriori interior estimate for the solutions of a nonlinear problem representing weak diffusions*, Nonlinear Anal., **5** (1981), 95–103.

[VAVE] J. L. Vazquez and L. Veron, *Removable singularities of some strongly nonlinear elliptic equations*, Manuscripta Math., **33** (1980), 129–144.

[VE1] L. Veron, *Global behaviour and symmetry properties of singular solutions of nonlinear elliptic equations*, Annales Fac. Sci. Toulous, **6** (1984), 1–31.

Received January 22, 1990.

INDIAN STATISTICAL INSTITUTE
NEW DELHI-110 016 INDIA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

BOURGAIN ALGEBRAS ON THE UNIT DISK

JOSEPH A. CIMA, KAREL STROETHOFF AND KEITH YALE

The Bourgain algebra of $H^\infty(\mathbb{D})$ relative to $L^\infty(\mathbb{D})$ is shown to be $H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) + V$, where V is an ideal of functions in $L^\infty(\mathbb{D})$ which vanish in an appropriate sense near the boundary of $\mathbb{D}$.

1. Introduction. Let $\mathscr{X}$ be a commutative Banach algebra with an identity and let $\mathscr{A}$ be a linear subspace of $\mathscr{X}$. J. Cima and R. Timoney [**6**] introduced the notion of the Bourgain algebra based on ideas of J. Bourgain [**3**]: the *Bourgain algebra* $\mathscr{A}_b$ consists of those f in $\mathscr{X}$ such that

$$\text{(1)} \qquad \text{if } f_n \to 0 \text{ weakly in } \mathscr{A}, \text{ then } \operatorname{dist}(f_n f, \mathscr{A}) \to 0 .$$

The distance $\operatorname{dist}(f_n f, \mathscr{A})$ between $f_n f$ and $\mathscr{A}$ is the quotient norm of the coset $f_n f + \mathscr{A}$ in the space $\mathscr{X}/\mathscr{A}$. The proof in [**6**] shows that $\mathscr{A}_b$ is a closed subalgebra of $\mathscr{X}$ and if $\mathscr{A}$ is an algebra then $\mathscr{A} \subseteq \mathscr{A}_b$. It is important to note that $\mathscr{A}_b$ depends upon the space $\mathscr{X}$ even though this is not reflected in the notation. For a brief survey of Bourgain algebras see K. Yale [**16**].

Let $H^\infty(\mathbb{D})$ be the algebra of bounded analytic functions on the open unit disk $\mathbb{D}$. There are at least three different natural spaces $\mathscr{X}$ containing $H^\infty(\mathbb{D})$. First we can let $\mathbb{T} = \partial\mathbb{D} = \{z \in \mathbb{C} : |z| = 1\}$ be the unit circle and consider the algebra $H^\infty(\mathbb{T})$ of boundary values of $H^\infty(\mathbb{D})$ functions as a subalgebra of $\mathscr{X} = L^\infty(\mathbb{T})$. In this context, J. Cima, S. Janson and K. Yale [**5**] showed that $H^\infty(\mathbb{T})_b = H^\infty(\mathbb{T}) + C(\mathbb{T})$. Another setting is to use the Gelfand map and regard $H^\infty(\mathbb{D})$ as a subalgebra of $\mathscr{X} = C(\mathscr{M})$, where $\mathscr{M}$ denotes the maximal ideal space of $H^\infty(\mathbb{D})$. In this context $H^\infty(\mathbb{D})_b$ has been determined by P. Ghatage, S. Sun and D. Zheng [**10**]. Yet another natural setting is to regard $H^\infty(\mathbb{D})$ as a subalgebra of $\mathscr{X} = L^\infty(\mathbb{D})$, where $L^\infty(\mathbb{D})$ is the usual space of equivalence classes of essentially bounded measurable functions on $\mathbb{D}$ with respect to area measure. The purpose of this paper is to determine $H^\infty(\mathbb{D})_b$ in the latter context. There is no Chang-Marshall theory for $L^\infty(\mathbb{D})$ in contrast to the well-known description of subalgebras between $H^\infty(\mathbb{T})$ and $L^\infty(\mathbb{T})$ which was used in [**5**] to determine $H^\infty(\mathbb{T})_b$. For a survey of the

Douglas algebra problem for $L^\infty(\mathbb{D})$, i.e., the problem of the determination of the algebras between $H^\infty(\mathbb{D})$ and $L^\infty(\mathbb{D})$, see P. Gorkin **[11]**. Ideas arising from the study of Bourgain algebras may shed some light on the Douglas algebra problem. K. Izuchi **[13]** has recently developed an abstract approach to the problem of determining Bourgain algebras in a variety of settings, which do not encompass our methods and results.

A substantial study of Bourgain algebras of Douglas subalgebras of $L^\infty(\mathbb{T})$ is contained in **[12]**, while a variety of interesting results concerning Bourgain algebras of some special subalgebras of $C(\mathscr{M})$ are found in **[17]**.

We show in §4 that $H^\infty(\mathbb{D})_b = H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) + V$, where V is an ideal of functions in $L^\infty(\mathbb{D})$ which vanish in an appropriate sense near the boundary of $\mathbb{D}$. Section 2 contains preliminaries as well as several examples. Functions in $H^\infty(\mathbb{D})_b$ have "boundary values" and this is proved in §3. The boundary value result enables one to reduce the determination of $H^\infty(\mathbb{D})_b$ to the known result for $H^\infty(\mathbb{T})_b$. In §5 we investigate the connection between $H^\infty(\mathbb{D})_b$ and the algebra AQ of bounded symbols of compact Hankel operators on the Bergman space. Let $A(\mathbb{D})$ denote the disk algebra, the algebra of continuous functions on $\overline{\mathbb{D}}$ which are analytic on $\mathbb{D}$. In §6 we study the boundary behavior of analytic functions in the Bourgain algebras $A(\mathbb{D})_b$ relative to $L^\infty(\mathbb{D})$. We end the paper with further remarks and questions in §7.

We wish to thank K. Izuchi and R. Mortini for their many helpful discussions concerning this work. We are especially grateful to the referee for pointing out a major error in an earlier version of this paper and for suggesting a better viewpoint.

2. Preliminaries. Several observations concerning weakly null sequences in $H^\infty(\mathbb{D})$ will be helpful in the determination of the Bourgain algebra $H^\infty(\mathbb{D})_b$. The mapping $f \mapsto f^*$, where $f^*(\zeta) = \lim_{r\to 1^-} f(r\zeta)$, for almost every $\zeta \in \mathbb{T}$, provides an isometric isomorphism from $H^\infty(\mathbb{D})$ onto $H^\infty(\mathbb{T})$. Thus $f_n \to 0$ weakly in $H^\infty(\mathbb{D})$ if and only if $f_n^* \to 0$ weakly in $H^\infty(\mathbb{T})$. If $f_n \to 0$ weakly in $H^\infty(\mathbb{D})$ then $\{f_n\}$ is uniformly bounded ($\|f_n\|_\infty \le M < \infty$ for all $n \ge 1$) and $f_n \to 0$ uniformly on compact subsets of $\mathbb{D}$.

Any sequence in $H^\infty(\mathbb{D})$ which is uniformly bounded and converges to 0 uniformly on compact subsets of $\mathbb{D}$ will be called *normal null*. Any weakly null sequence is certainly normal null, but not conversely.

For example, $f_n(z) = z^n$ is normal null but does not converge to 0 weakly.

For a subalgebra $\mathscr{A} \subseteq L^\infty(\mathbb{D})$ we define the *normal null algebra* $\mathscr{A}_N$ to be the set of those functions f in $L^\infty(\mathbb{D})$ such that

$$\text{(2)} \qquad \text{if } \{f_n\} \text{ is normal null in } \mathscr{A}\,, \text{ then } \operatorname{dist}(f_n f, \mathscr{A}) \to 0\,.$$

A minor modification of the argument in **[6]** shows that $\mathscr{A}_N$ is a closed subalgebra of $L^\infty(\mathbb{D})$ with $\mathscr{A} \subseteq \mathscr{A}_N$. Clearly $H^\infty(\mathbb{D})_N \subseteq H^\infty(\mathbb{D})_b$ since every weakly null sequence in $H^\infty(\mathbb{D})$ is normal null. For a closed subalgebra $\mathscr{A} \subseteq L^\infty(\mathbb{D})$ and $f \in L^\infty(\mathbb{D})$ define the Hankel type operator $S_f\colon \mathscr{A} \to L^\infty(\mathbb{D})/\mathscr{A}$ by $S_f h = fh + \mathscr{A}$, $h \in \mathscr{A}$. Note that the Bourgain algebra $\mathscr{A}_b$ consists of those f in $L^\infty(\mathbb{D})$ for which the operator S_f is completely continuous. The set $\mathscr{A}_{wc}$ of functions f in $L^\infty(\mathbb{D})$ for which S_f is weakly compact is also a closed subalgebra of $L^\infty(\mathbb{D})$ with $\mathscr{A} \subseteq \mathscr{A}_{wc}$ as is shown by the duality argument of B. Cole and T. Gamelin [7], Lemma 4.2. Note that $f \in \mathscr{A}_{wc}$ in case the operator $S_f : \mathscr{A} \to L^\infty(\mathbb{D})/\mathscr{A}$ defined by $S_f h = fh + \mathscr{A}$, $h \in \mathscr{A}$, is weakly compact while $f \in \mathscr{A}_b$ in case the operator S_f is completely continuous. Now weakly compact operators are completely continuous in any space which has the Dunford-Pettis property and so $H^\infty(\mathbb{D})_{wc} \subseteq H^\infty(\mathbb{D})_b$ follows immediately from the remarkable theorem of J. Bourgain [2] that $H^\infty(\mathbb{D})$ has the Dunford-Pettis property. If $f \in H^\infty(\mathbb{D})_N$ then a normal families argument shows that S_f is compact and hence weakly compact. In particular, we have the inclusion $H^\infty(\mathbb{D})_N \subseteq H^\infty(\mathbb{D})_{wc}$.

The space $H^\infty(\mathbb{D})$ is well supplied with weakly null sequences which do not converge to zero in norm. Such sequences arise from peak point and from interpolation type constructions; the following lemma is useful in this respect.

LEMMA 1. *Suppose that* $\{f_n\}$ *is a sequence of functions in* $H^\infty(\mathbb{D})$ *and* M *is a constant such that*

$$\text{(3)} \qquad \sum_{n=1}^{\infty} |f_n(z)| \le M\,, \qquad z \in \mathbb{D}\,.$$

Then $f_n \to 0$ *weakly in* $H^\infty(\mathbb{D})$.

Proof. See [5]. □

If $\{z_n\}$ is an interpolating sequence in the unit disk, by the P. Beurling Interpolation Theorem (see [9]) there exist functions f_n in

$H^\infty(\mathbb{D})$ such that $f_n(z_m) = \delta_{nm}$ and for which (3) holds. Consequently, $f_n \to 0$ weakly in $H^\infty(\mathbb{D})$.

Let $\{\zeta_n\}$ be a sequence of distinct points on the circle $\mathbb{T}$ which converges to a point ζ which is not equal to any of the ζ_n. Since each point of $\mathbb{T}$ is a peak point for the disk algebra $A(\mathbb{D})$, there exist functions $g_n \in A(\mathbb{D})$ such that $g_n(\zeta_n) = 1$ and $|g_n(z)| < 1$ if $z \in \overline{\mathbb{D}}$, but $z \neq \zeta_n$. Construct a disjoint collection $\{\mathscr{O}_n : n \geq 1\}$ of open subsets of $\mathbb{T}$ such that $\zeta_n \in \mathscr{O}_n$ and $\sup\{|g_n(\zeta)| : \zeta \in \mathbb{T}\backslash\mathscr{O}_n\} < 1$, and for each n choose a positive integer k_n so large that $f_n(z) = g_n(z)^{k_n}$ is small on $\mathbb{T}\backslash\mathscr{O}_n$: $|f_n(\zeta)| \leq 1/2^n$ for $\zeta \in \mathbb{T}\backslash\mathscr{O}_n$. Hence (3) holds with $M = 2$ and we conclude that $f_n \to 0$ weakly in $H^\infty(\mathbb{T})$ and hence in $H^\infty(\mathbb{D})$. Since the sequence $\{f_n\}$ is in $A(\mathbb{D})$ it is also weakly null in $A(\mathbb{D})$. Clearly $f_n(\zeta_n) = 1$ for all $n \geq 1$. A concrete example is given by

$$f_n(z) = \left(\frac{z + \zeta_n}{2}\right)^{k_n}.$$

A few simple examples and a proposition will serve to fix the ideas and to motivate the main theorems.

EXAMPLE 2. Let g be an arbitrary function in $L^\infty(\mathbb{D})$, $K \subset \mathbb{D}$ compact and put $f = \chi_K g$ where χ_K is the characteristic function of K. Then $f \in H^\infty(\mathbb{D})_N$ because $\operatorname{dist}(f_n f, H^\infty(\mathbb{D})) \leq \|f_n \chi_K g - 0\|_\infty \leq \|g\|_\infty \|f_n \chi_K\|_\infty \to 0$ for any normal null sequence $\{f_n\}$ in $H^\infty(\mathbb{D})$. □

PROPOSITION 3. $H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) \subseteq H^\infty(\mathbb{D})_N$.

Proof. By the Stone-Weierstrass Theorem it suffices to show that the function $\overline{z}$ belongs to $H^\infty(\mathbb{D})_N$. Let $\{f_n\}$ be a normal null sequence in $H^\infty(\mathbb{D})$, and for each positive integer n let $h_n \in H^\infty(\mathbb{D})$ be defined by

$$h_n(z) = \frac{f_n(z) - f_n(0)}{z}, \quad \text{for } z \in \mathbb{D} \setminus \{0\}.$$

It is easily seen that the sequence $\{h_n\}$ is normal null in $H^\infty(\mathbb{D})$. Let M be a finite positive number such that $\|h_n\|_\infty \leq M$ for each $n \geq 1$. Using $\overline{z} f_n(z) = |z|^2 h_n(z) + \overline{z} f_n(0)$ we see that

$$\begin{aligned}\operatorname{dist}(\overline{z} f_n, H^\infty(\mathbb{D})) \leq \|\overline{z} f_n - h_n\|_\infty &\leq (1 - r^2)M \\ &+ \max_{|z| \leq r} |h_n(z)| + |f_n(0)|,\end{aligned}$$

for each $0 < r < 1$. Taking the limit superior as $n \to \infty$, and subsequently letting $r \to 1^-$, it follows that $\operatorname{dist}(\overline{z} f_n, H^\infty(\mathbb{D})) \to 0$ as $n \to \infty$, proving that the function $\overline{z}$ belongs to $H^\infty(\mathbb{D})_N$. □

It follows from Example 2 and Proposition 3 that $H^\infty(\mathbb{D})_N$, and thus $H^\infty(\mathbb{D})_b$, contains all functions on the unit disk which are "$H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}})$ near the boundary," that is, coincide with a function from the algebra $H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}})$ on the complement of a compact subset of $\mathbb{D}$. We will actually prove that $H^\infty(\mathbb{D})_b$ is the sum of $H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}})$ and (equivalence classes of) functions that "vanish near the boundary." This statement will be made precise in Theorem 12 of §4.

EXAMPLE 4. Let $U = \{z \in \mathbb{D}\colon \operatorname{Im} z > 0\}$ be the upper open half disk and let $f = \chi_U$. Then $f \notin H^\infty(\mathbb{D})_b$. For if we suppose that $f \in H^\infty(\mathbb{D})_b$, then for every weakly null sequence $\{f_n\}$ in $H^\infty(\mathbb{D})$ there exists a sequence $\{g_n\}$ in $H^\infty(\mathbb{D})$, such that $\|f_n f - g_n\|_\infty \to 0$ as $n \to \infty$. Then $f_n - g_n \to 0$ uniformly on the closure of U relative to $\mathbb{D}$ while $g_n \to 0$ uniformly on the closure of $\mathbb{D}\setminus U$ relative to $\mathbb{D}$ since $f_n f - g_n = -g_n$ on $\mathbb{D}\setminus U$. Hence $f_n = (f_n - g_n) + g_n \to 0$ uniformly on the line segment $(-1, 1) = \overline{U} \cap \overline{\mathbb{D}\setminus U} \cap \mathbb{D}$. This is a contradiction because there are many weakly null sequences which do not converge uniformly to zero on the line segment $(-1, 1)$; in particular, if $\{z_n\}$ is an interpolating sequence in $(-1, 1)$ the P. Beurling sequence $\{f_n\}$ does not converge uniformly to zero on $(-1, 1)$ since $f_n(z_n) = 1$ for each $n \geq 1$. □

EXAMPLE 5. Let $\{r_n\}$ be a sequence of positive numbers increasing to 1 and let $S_n = \{z \in \mathbb{C}\colon r_{2n-1} < |z| < r_{2n}\}$. The values of the function $f = \chi_S$, where $S = \bigcup_{n=1}^\infty S_n$, oscillate as $|z| \to 1^-$, since S is the union of concentric annuli. Thus the radial limit, $f^*(\zeta)$, of f does not exist for any $\zeta \in \mathbb{T}$. The argument given in the previous example can be used to show $f \notin H^\infty(\mathbb{D})_b$. (The arbitrary weakly null sequence $\{f_n\}$ is forced to converge uniformly to 0 on the union of the circles $\bigcup_{n=1}^\infty \{z \in \mathbb{C}\colon |z| = r_n\} = \overline{S} \cap \overline{\mathbb{D}\setminus S}$ so that by the maximum modulus principle $\|f_n\|_\infty \to 0$.) □

These examples are typical. Thus f will belong to $H^\infty(\mathbb{D})_b$ if it vanishes near the boundary but the existence of boundary values in the absence of interior regularity is not sufficient to make f belong to $H^\infty(\mathbb{D})_b$. Bad boundary behavior of the function f will always force $f \notin H^\infty(\mathbb{D})_b$ as in Example 5. These statements are given precise form by the main theorems of §§3 and 4. The proofs of the main theorems require more delicate properties of weakly null sequences than were used in the examples.

Since $L^\infty(\mathbb{D})$ consists of *equivalence classes* of essentially bounded measurable functions, simple pointwise definitions of boundary values do not make sense because one can always modify a function on a set of area measure zero so that its limit does not exist as $|z| \to 1^-$. However, technically useful definitions can be made for essential non-tangential limits and oscillations and these notions have an important role in the proofs of our principal theorems.

For $\zeta \in \mathbb{T}$ and $0 < R < 1$ the open non-tangential cone $\Gamma_R(\zeta)$ at ζ is the interior of the convex hull of ζ and the disk $\{z \in \mathbb{C}: |z| \leq R\}$. We say that f *has essential non-tangential limit* L *at* ζ if

$$\operatorname{ess\,sup}\{|f(z) - L|: z \in \Gamma_R(\zeta),\ |z| > 1 - \delta\} \to 0 \quad \text{as } \delta \to 0^+$$

for all $0 < R < 1$, in which case we will write $f^*(\zeta)$ for L. We define BV to be the set of f in $L^\infty(\mathbb{D})$ such that an essential non-tangential limit $f^*(\zeta)$ exists for almost every $\zeta \in \mathbb{T}$. In order to give another description of BV we will introduce more notation. For $f \in L^\infty(\mathbb{D})$ and a nonempty set $E \subset \mathbb{D}$ define the *essential oscillation of* f *over* E to be

$$\omega(f, E) = \operatorname{ess\,sup}\{|f(z) - f(w)|: z, w \in E\}.$$

For each $z \in \overline{\mathbb{D}}$ and $0 < \delta < 1$ we will write $E(z, \delta) = \{w \in \mathbb{D}: |w - z| < \delta\}$. For $f \in L^\infty(\mathbb{D})$ define the *essential oscillation of* f *at* $z \in \overline{\mathbb{D}}$ to be

$$\omega(f, z) = \lim_{\delta \to 0^+} \omega(f, E(z, \delta)).$$

If we define the *essential non-tangential oscillation of* f *at* $\zeta \in \mathbb{T}$ *by*

$$\omega_R(f, \zeta) = \lim_{\delta \to 0^+} \omega(f, E(\zeta, \delta) \cap \Gamma_R(\zeta)),$$

then we have the following description of BV.

PROPOSITION 6.

$$BV = \{f \in L^\infty(\mathbb{D}): \omega_R(f, \zeta) = 0 \quad \textit{for all } 0 < R < 1 \textit{ and almost every } \zeta \in \mathbb{T}\}.$$

Proof. If for $\zeta \in \mathbb{T}$ the essential non-tangential limit $f^*(\zeta)$ exists, then a simple application of the triangle inequality shows that $\omega_R(f, \zeta) = 0$. Conversely, if $\omega_R(f, \zeta) = 0$, for all $0 < R < 1$, then it is possible to choose a suitable sequence $\{z_n(R)\}$ in $\mathbb{D}$, converging to ζ, for which $\{f(z_n(R))\}$ is Cauchy with limit, $f^*(\zeta)$, independent of R. A suitable sequence can be constructed by taking $z_n(R) \in E(\zeta, 1/n) \cap \Gamma_R(\zeta) = S$ for which $\operatorname{ess\,sup}\{|f(z) - f(z_n(R))|: z \in S\} \leq \omega(f, S)$. The details are left to the reader. □

Note that if f, $g \in L^\infty(\mathbb{D})$, and f has essential limit $f^*(\zeta)$ as z approaches $\zeta \in \mathbb{T}$ in the cone $\Gamma_R(\zeta)$, then

$$\omega_R(fg, \zeta) = |f^*(\zeta)|\omega_R(g, \zeta).$$

Also note that if $\omega_R(g, \zeta) = 0$, then $\omega_R(f, \zeta) = \omega_R(f - g, \zeta)$.

Since functions in $H^\infty(\mathbb{D})$ have non-tangential limits almost everywhere on the circle we have $H^\infty(\mathbb{D}) \subseteq BV$. From the main theorem in §3 it will follow that BV is closed and $(BV)_b = BV$.

3. Boundary Value Theorem. As we emphasized in §2 the algebra $H^\infty(\mathbb{D})_b$ is a subset of the Banach algebra $L^\infty(\mathbb{D})$ and as such its members are equivalence classes of functions. The following lemma on metric density is an important tool in the proof of the theorem of this section. Although it is valid in more general situations, we state it for the case of subsets of the real numbers $\mathbb{R}^1$. If E is a Lebesgue measurable subset of $\mathbb{R}^1$ the quantity

$$D(E, x) \equiv \lim_{\delta \to 0^+} \frac{|E \cap (x - \delta, x + \delta)|}{2\delta},$$

if it exists, is the metric density of E at the point x. It is well known that $D(E, x) = 1$ for almost all x in E. Let $D(E) = \{x \in E\colon D(E, x) = 1\}$, so that $|D(E)| = |E|$, and denote $J_n(x) = (x - \frac{1}{n}, x + \frac{1}{n})$. We identify $y \in J_n(x)$ with e^{iy} when working on the circle $\mathbb{T}$.

LEMMA 7. *Let E and Z be measurable subsets of the circle $\mathbb{T}$ with $|E| > 0$ and $|Z| = 0$. Then each $\zeta \in D(E)$ is the limit of a sequence $\{\zeta_n\}$ of distinct points in $D(E) \setminus Z$.*

Proof. Left to the reader. □

THEOREM 8. $H^\infty(\mathbb{D})_b \subseteq BV$.

Proof. Assume not and choose an f in $H^\infty(\mathbb{D})_b$ whose non-tangential limits fail to exist on a set $B \subseteq \mathbb{T}$ with $|B| > 0$. For each $\zeta \in B$ there is an $R(\zeta)$ in $(0, 1)$ such that $\omega_{R(\zeta)}(f, \zeta) \geq \delta(\zeta) > 0$. Setting $B_{jk} = \{\zeta \in B\colon \delta(\zeta) \geq \frac{1}{j}$ and $R(\zeta) \leq 1 - \frac{1}{k}\}$ for j, $k \in \mathbb{Z}^+$, we have $B = \bigcup\{B_{jk}\colon j, k \in \mathbb{Z}^+\}$. Since $|B| > 0$ there are j and k in $\mathbb{Z}^+$ for which $|B_{jk}| > 0$. We replace B with the set B_{jk} and henceforth we have

$$\omega_R(f, \zeta) \geq \delta > 0$$

for all $\zeta \in B$, where $0 < R < 1$.

For ζ fixed in $D(B)$ we apply the preceding lemma to obtain a sequence $\{\zeta_n\}$ in $D(B)$ which converges to ζ. Let $\{f_n\}$ in $A(\mathbb{D})$ be a weakly null sequence of peak point functions associated with the sequence $\{\zeta_n\}$. Since $f \in H^\infty(\mathbb{D})_b$ there is a sequence $\{h_n\}$ in $H^\infty(\mathbb{D})$ such that $\|ff_n - h_n\|_\infty = \varepsilon_n \to 0$ as $n \to \infty$. Then for almost every $\xi \in \mathbb{T}$,

$$|f_n(\xi)|\delta \le |f_n(\xi)|\omega_R(f, \xi) = \omega_R(ff_n, \xi) = \omega_R(ff_n - h_n, \xi) \le 2\varepsilon_n.$$

Since we here may choose $\xi \in B$ so close to ζ_n that $|f_n(\xi)| > 1/2$, the above inequality yields $\delta < 4\varepsilon_n$, a contradiction. □

REMARK 9. In the above proof only two properties of $H^\infty(\mathbb{D})$ were used. First, functions in $H^\infty(\mathbb{D})$ have non-tangential limits almost everywhere on the unit circle, i.e., $H^\infty(\mathbb{D}) \subseteq BV$. Secondly, $H^\infty(\mathbb{D})$ contains the particular weakly null sequence of peak point functions $\{f_n\}$ simply because $f_n \in A(\mathbb{D}) \subseteq H^\infty(\mathbb{D})$. Thus the proof just given actually shows more:

If $\mathscr{A}$ is a linear subspace of $L^\infty(\mathbb{D})$ which satisfies $A(\mathbb{D}) \subseteq \mathscr{A} \subseteq BV$, then $\mathscr{A}_b \subseteq BV$.

As an immediate consequence we have $BV_b = BV$.

The following theorem generalizes Example 4 of the previous section.

THEOREM 10. *If $f \in H^\infty(\mathbb{D})_b$, then $\omega(f, z) \to 0$ as $|z| \to 1^-$.*

Proof. Let $f \in H^\infty(\mathbb{D})_b$ and assume that $\{z_n\}$ in $\mathbb{D}$ with $|z_n| \to 1^-$. By going to a subsequence, which we will not relabel, we can furthermore assume that the sequence $\{z_n\}$ is interpolating. Choose a weakly null sequence $\{f_n\}$ of interpolating functions in $H^\infty(\mathbb{D})$ with $f_n(z_m) = \delta_{nm}$. Since $f \in H^\infty(\mathbb{D})_b$, there is a sequence $\{g_n\}$ in $H^\infty(\mathbb{D})$ such that $\|ff_n - g_n\|_\infty \to 0$ as $n \to \infty$. Then

$$\begin{aligned}\omega(f, z_n) &= |f_n(z_n)|\omega(f, z_n) = \omega(ff_n, z_n)\\ &= \omega(ff_n - g_n, z_n) \le 2\|ff_n - g_n\|_\infty,\end{aligned}$$

and so $\omega(f, z_n) \to 0$ as $n \to 0$. □

REMARK 11. A careful analysis of the above proof shows that more generally we have the following result:

Let $\mathscr{A}$ be a linear subspace of $L^\infty(\mathbb{D})$ such that $H^\infty(\mathbb{D}) \subseteq \mathscr{A} \subseteq C(\mathbb{D})$, and suppose that $f \in \mathscr{A}_b$. Then $\omega(f, z) \to 0$ as $|z| \to 1^-$.

4. The Structure Theorem. In this section we will give our description of the Bourgain algebra $H^\infty(\mathbb{D})_b$.

THEOREM 12. *$H^\infty(\mathbb{D})_b = H^\infty(\mathbb{D})_{wc} = H^\infty(\mathbb{D})_N = H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) + V$, where $V = \{g \in L^\infty(\mathbb{D}) \colon \|g\chi_{\mathbb{D}\setminus r\mathbb{D}}\|_\infty \to 0$ as $r \to 1^-\}$.*

Proof. In Proposition 3 we have already shown that $H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) \subseteq H^\infty(\mathbb{D})_N$. Noting that V is the closure in $L^\infty(\mathbb{D})$ of the set of all functions with compact support in $\mathbb{D}$, it follows from Example 2 and the fact that $H^\infty(\mathbb{D})_N$ is closed, that also $V \subseteq H^\infty(\mathbb{D})_N$. Thus we have $H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) + V \subseteq H^\infty(\mathbb{D})_N \subseteq H^\infty(\mathbb{D})_b$.

Now let $f \in H^\infty(\mathbb{D})_b$. By the Boundary Value Theorem f has non-tangential limits almost everywhere on $\mathbb{T}$. Let f^* denote the limit function on $\mathbb{T}$. We will first argue that $f^* \in H^\infty(\mathbb{T})_b$. Let $\{\varphi_n\}$ be weakly null in $H^\infty(\mathbb{T})$, and for each positive integer n choose $f_n \in H^\infty(\mathbb{D})$ such that $\varphi_n = f_n^*$. Because $f \in H^\infty(\mathbb{D})_b$ and $\{f_n\}$ is weakly null in $H^\infty(\mathbb{D})$, there exist $g_n \in H^\infty(\mathbb{D})$ such that $\varepsilon_n = \|f_n f - g_n\|_\infty \to 0$ as $n \to \infty$. Since the mapping $F \mapsto F^*$ is a contractive homomorphism of BV onto $L^\infty(\mathbb{T})$ we conclude that $\|f^* f_n^* - g_n^*\|_{L^\infty(\mathbb{T})} \le \varepsilon_n$. This proves our claim that $f^* \in H^\infty(\mathbb{T})_b$.

Denoting the Poisson integral of a function $\varphi \in L^\infty(\mathbb{T})$ by $\mathscr{P}[\varphi]$, we now put $g = f - \mathscr{P}[f^*]$. By [5], $H^\infty(\mathbb{T})_b = H^\infty(\mathbb{T}) + C(\mathbb{T})$, so $f^* \in H^\infty(\mathbb{T}) + C(\mathbb{T})$. As $\mathscr{P}$ maps $H^\infty(\mathbb{T}) + C(\mathbb{T})$ into $H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}})$, $\mathscr{P}[f^*] \in H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) \subseteq H^\infty(\mathbb{D})_b$, and we conclude that $g \in H^\infty(\mathbb{D})_b$. We claim that actually $g \in V$.

Assuming that $g \notin V$, there exists a sequence $\{r_n\}$ in the interval $(0, 1)$ tending to 1 and a positive number δ such that $\|g\chi_{\mathbb{D}\setminus r_n\mathbb{D}}\|_\infty \ge \delta$, for all $n \in \mathbb{Z}^+$. Put $A_n = \{z \in \mathbb{D} \colon |g\chi_{\mathbb{D}\setminus r_n\mathbb{D}}(z)| > \delta/2\}$. Then $|A_n| > 0$. Let z_n be a point of density of the set A_n. Noting that $A_n \subseteq \mathbb{D} \setminus r_n\mathbb{D}$ we see that $|z_n| \to 1$ as $n \to \infty$. By passing to a subsequence we may assume that $\{z_n\}$ is an interpolating sequence. Let $\{f_n\}$ be a sequence in $H^\infty(\mathbb{D})$ such that $f_n(z_n) = 1$ and $f_n \to 0$ weakly. For each $n \in \mathbb{Z}^+$, by the continuity of f_n at z_n, we can pick a positive $\delta_n < 1 - |z_n|$ such that $|f_n(z)| \ge 1/2$ whenever $|z - z_n| < \delta_n$. Because z_n is a point of density of the set A_n, the sets $B_n = A_n \cap \{z \in \mathbb{C} \colon |z - z_n| < \delta_n\}$ have positive measure and the property that $|f_n(z)| \ge 1/2$ for all $z \in B_n$. Using that $g \in H^\infty(\mathbb{D})_b$ and $f_n \to 0$ weakly in $H^\infty(\mathbb{D})$, there exist $h_n \in H^\infty(\mathbb{D})$ for which $\|f_n g - h_n\|_\infty \to 0$ as $n \to \infty$. As in the second paragraph of the proof this implies that $\|g^* f_n^* - h_n^*\|_{L^\infty(\mathbb{T})} \to 0$ as $n \to \infty$. Since $g^* = 0$ almost everywhere on $\mathbb{T}$, we get $\|h_n^*\|_{L^\infty(\mathbb{T})} \to 0$ as $n \to \infty$, and conclude that $\|h_n\|_\infty \to 0$ as $n \to \infty$. But then it follows that $\|f_n g\|_\infty \to 0$ as $n \to \infty$. However, for $z \in B_n$ we have $|f_n(z)g(z)| \geq (1/2)(\delta/2) = \delta/4$, so that $\|f_n g\|_\infty \geq \delta/4$ for all $n \in \mathbb{Z}^+$. This contradiction establishes our claim that $g \in V$, and hence $f = \mathscr{P}[f^*] + g \in H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) + V$.

So far we have shown that $H^\infty(\mathbb{D})_b = H^\infty(\mathbb{D})_N = H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}}) + V$. Recalling the inclusions $H^\infty(\mathbb{D})_N \subseteq H^\infty(\mathbb{D})_{wc}$ and $H^\infty(\mathbb{D})_{wc} \subseteq H^\infty(\mathbb{D})_b$ (the last inclusion because $H^\infty(\mathbb{D})$ has the Dunford-Pettis property) the theorem follows. □

REMARK 13. Theorem 12 can be used to give an alternative proof of Theorem 10: Clearly $\omega(f, z) = 0$ if $f \in H^\infty(\mathbb{D}) + C(\overline{\mathbb{D}})$. It is easily seen that $\omega(f, z) \leq 2\|f\chi_{\mathbb{D}\setminus r\mathbb{D}}\|_\infty$ for $r < |z| < 1$, implying that if $f \in V$, then $\omega(f, z) \to 0$ as $|z| \to 1^-$

5. Operator theoretic aspects. In this section we relate the Bourgain algebra of $H^\infty(\mathbb{D})$ to the algebra of bounded symbols for which the associated Hankel operators are compact operators on the Bergman space of the unit disk.

Let L^2_a denote the *Bergman space,* that is, the space of square integrable analytic functions on $\mathbb{D}$. Let dA denote Lebesgue area measure on $\mathbb{D}$, normalized so that $\mathbb{D}$ has measure 1. Since L^2_a is a closed subspace of $L^2(\mathbb{D}, dA)$ there is an orthogonal projection P of $L^2(\mathbb{D}, dA)$ onto L^2_a. For $f \in L^\infty(\mathbb{D})$ the *Hankel operator* with symbol f, denoted by H_f, is the operator from L^2_a into $L^2(\mathbb{D}, dA)$ defined by

$$H_f g = (I - P)(fg), \qquad g \in L^2_a.$$

Clearly H_f is a bounded operator with norm $\|H_f\| \leq \|f\|_\infty$. Let

$$AQ = \{f \in L^\infty(\mathbb{D}) : H_f \text{ is compact}\}.$$

It is known that AQ is a closed subalgebra of $L^\infty(\mathbb{D})$ (see page 475 of **[1]**).

In the context of the unit circle, Hankel operators on the Hardy space H^2 are defined similarly. In [5] the Bourgain algebra $H^\infty(\mathbb{T})_b$ is shown to coincide with the algebra of bounded measurable functions for which the associated Hankel operator on the Hardy space H^2 is compact. In the context of the unit disk this is no longer true as is shown in the following theorem.

THEOREM 14. $H^\infty(\mathbb{D})_b \subsetneq AQ$.

Proof. We will first show that $H^\infty(\mathbb{D})_b \subseteq AQ$. The inclusion $H^\infty(\mathbb{D}) \subseteq AQ$ is trivial, for if $f \in H^\infty(\mathbb{D})$, then H_f is the zero operator. Using Theorem 12 it is enough to show the inclusions

$C(\overline{\mathbb{D}}) \subseteq AQ$ and $V \subseteq AQ$. We will use a criterion for compactness of Hankel operators obtained in K. Stroethoff **[15]**. For $\lambda \in \mathbb{D}$ let φ_λ be the Möbius function defined by $\varphi_\lambda(z) = \frac{\lambda - z}{1-\bar{\lambda}z}$, $z \in \mathbb{D}$. Theorem 6 in **[15]**, then, states that for $f \in L^\infty(\mathbb{D}) : f \in AQ$ if and only if $\|f \circ \varphi_\lambda - P(f \circ \varphi_\lambda)\|_2 \to 0$ as $|\lambda| \to 1^-$.

Now, if $f \in C(\overline{\mathbb{D}})$, then $|f \circ \varphi_\lambda(z) - f(\lambda)| \to 0$ for each $z \in \mathbb{D}$ as $|\lambda| \to 1^-$ (because $|\varphi_\lambda(z) - \lambda| = (1-|\lambda|)|z|/|1-\bar{\lambda}z| \to 0$ as $|\lambda| \to 1^-$), and by the dominated convergence theorem, $\|f \circ \varphi_\lambda - f(\lambda)\|_2 \to 0$ as $|\lambda| \to 1^-$. Consequently,

$$\|f \circ \varphi_\lambda - P(f \circ \varphi_\lambda)\|_2 = \|(I-P)(f \circ \varphi_\lambda - f(\lambda))\|_2 \to 0$$

as $|\lambda| \to 1^-$, and it follows that $f \in AQ$. Thus $C(\overline{\mathbb{D}}) \subseteq AQ$.

For $0 < r < 1$ and $\lambda \in \mathbb{D}$ write $D(\lambda, r) = \varphi_\lambda(r\mathbb{D})$. Then $|D(\lambda, r)| = r^2(1-|\lambda|^2)^2/(1-r^2|\lambda|^2)^2$ (see, for example, **[9]**). If $f \in L^\infty(\mathbb{D})$, then for $0 < r < 1$ we have

$$\begin{aligned}\int_{\mathbb{D}} |f \circ \varphi_\lambda|^2 \, dA &= \int_{D(\lambda, r)} |f \circ \varphi_\lambda|^2 \, dA + \int_{\mathbb{D}\setminus D(\lambda, r)} |f \circ \varphi_\lambda|^2 \, dA \\ &\le \|f\|_\infty^2 |D(\lambda, r)| + \int_{\mathbb{D}\setminus r\mathbb{D}} |f(w)|^2 |\varphi_\lambda'(w)|^2 \, dA(w) \\ &\le \|f\|_\infty^2 |D(\lambda, r)| + \|f\chi_{\mathbb{D}\setminus r\mathbb{D}}\|_\infty^2 \int_{\mathbb{D}} |\varphi_\lambda'(w)|^2 \, dA(w) \\ &= \|f\|_\infty^2 |D(\lambda, r)| + \|f\chi_{\mathbb{D}\setminus r\mathbb{D}}\|_\infty^2 .\end{aligned}$$

Hence

$$\limsup_{|\lambda| \to 1^-} \|f \circ \varphi_\lambda\|_2 \le \|f\chi_{\mathbb{D}\setminus r\mathbb{D}}\|_\infty .$$

So if $f \in V$, then it follows that $\|f \circ \varphi_\lambda\|_2 \to 0$ as $|\lambda| \to 1^-$, so that $\|f \circ \varphi_\lambda - P(f \circ \varphi_\lambda)\|_2 \to 0$ as $|\lambda| \to 1^-$, and therefore $f \in AQ$.

It remains to exhibit a function f in AQ which is not in $H^\infty(\mathbb{D})_b$. Let f be as in Example 5. Then $f \notin H^\infty(\mathbb{D})_b$, and we will show that for appropriately chosen r_n, $f \in AQ$. We recall the following formula for H_f (see **[15]**):

$$(H_f g)(z) = \int_{\mathbb{D}} \frac{f(z) - f(w)}{(1 - z\bar{w})^2} g(w) \, dA(w),$$

for $g \in L_a^2$, $z \in \mathbb{D}$.

The operator H_f will be Hilbert-Schmidt, and thus compact, if we make sure that

$$\int_{\mathbb{D}} \int_{\mathbb{D}} \frac{|f(z) - f(w)|^2}{|1 - z\bar{w}|^4} \, dA(z) \, dA(w) < \infty .$$

We choose $r_{2n-1} = 1 - \frac{1}{n}$ and $r_{2n} = 1 - \frac{1}{n+2^{-n}}$ for $n \geq 1$. Using that $f = \chi_S$, we have

$$\begin{aligned}\int_{\mathbb{D}}\int_{\mathbb{D}} &\frac{|f(z)-f(w)|^2}{|1-z\overline{w}|^4}\,dA(z)\,dA(w)\\ &\leq 2\int_S\int_{\mathbb{D}}\frac{1}{|1-z\overline{w}|^4}\,dA(z)\,dA(w)\\ &= 2\int_S\frac{1}{(1-|w|^2)^2}\,dA(w) \leq 2\sum_{n=1}^{\infty}\int_{r_{2n-1}}^{r_{2n}}\frac{2}{(1-r)^2}\,dr\\ &= 4\sum_{n=1}^{\infty}\left(\frac{1}{1-r_{2n}} - \frac{1}{1-r_{2n-1}}\right) \leq 4\sum_{n=1}^{\infty}2^{-n} = 4.\end{aligned}$$

Thus $f \in AQ$. □

6. On the Bourgain algebra of the disk algebra. Let $A(\mathbb{D})_b$ be the Bourgain algebra of the disk algebra $A(\mathbb{D})$ relative to $L^\infty(\mathbb{D})$. We know that $A(\mathbb{D})_b \subseteq BV$ (by Remark 9) and that $A(\mathbb{D})$ contains many non-analytic functions since both $C(\overline{\mathbb{D}})$ and V are subsets of $A(\mathbb{D})_b$. We will show that to some extent the boundary values of analytic functions in $A(\mathbb{D})_b$ are restricted. For this purpose we set

$$AH = A(\mathbb{D})_b \cap H^\infty(\mathbb{D})$$

and note that AH is a Banach algebra between $A(\mathbb{D})$ and $H^\infty(\mathbb{D})$.

PROPOSITION 15. *If $f \in H^\infty(\mathbb{D})$ is continuous on $\overline{\mathbb{D}}\setminus\{\zeta_1, \dots, \zeta_k\}$, where $\zeta_1, \dots, \zeta_k \in \mathbb{T}$, then $f \in AH$.*

Proof. There is no loss of generality in assuming $k = 1$, $\zeta_1 = 1$ and $\|f\|_\infty \leq 1$. Let $\{f_n\}$ be a weakly null sequence in $A(\mathbb{D})$. Note that $a_n = |f_n(1)| + n^{-1} \to 0$ as $n \to \infty$. For each $n \in \mathbb{Z}^+$, we can choose $\delta_n > 0$ so that $|f_n(z)| < 2a_n$ if $z \in \mathbb{D}$ and $|z-1| < 2\delta_n$. Fix an $n \in \mathbb{Z}^+$, and choose $r_n \in (1-\delta_n, 1)$ so that

$$|f(z)f_n(z) - f(r_nz)f_n(r_nz)| < 4a_n$$

whenever $z \in \mathbb{D}$ and $|z-1| > \delta_n$. Note that then $|r_nz - 1| \leq (1-r_n)|z| + |z-1| < 2\delta_n$ whenever $z \in \mathbb{D}$ and $|z-1| < \delta_n$. Choosing $g_n(z) = f(r_nz)f_n(r_nz)$ in $A(\mathbb{D})$ we have $\|ff_n - g_n\|_\infty \leq 4a_n$. □

If $f \in L^\infty(\mathbb{D})$ and $\zeta \in \mathbb{T}$ the *essential cluster set of f at ζ* is

$$\mathscr{C}(f, \zeta) = \bigcap_{0<\delta<1} \overline{R_f(\zeta, \delta)},$$

where $R_f(\zeta, \delta)$ denotes the essential range of the function $f|E(\zeta, \delta)$. The diameter of a set $K \subseteq \mathbb{C}$ is written $\operatorname{diam}(K)$.

PROPOSITION 16. *If* $f \in A(\mathbb{D})_b$, *and* $Z \subseteq \mathbb{T}$ *is an infinite set, then* $\inf_{\zeta \in Z} \operatorname{diam} \mathscr{C}(f, \zeta) = 0$.

Proof. Assume there are distinct points ζ_n in $\mathbb{T}$ such that

$$\operatorname{diam} \mathscr{C}(f, \zeta_n) \geq \delta > 0, \qquad \text{for } n \in \mathbb{Z}^+.$$

Then it is clear that $\omega(f, \zeta_n) \geq \delta > 0$, for $n \in \mathbb{Z}^+$. By passing to a subsequence we may assume that $\zeta_n \to \zeta$ as $n \to \infty$. Choose $\{f_n\}$ in $A(\mathbb{D})$ peaking at ζ_n and tending to zero weakly in $A(\mathbb{D})$. Then there is a sequence $\{g_n\}$ in $A(\mathbb{D})$ such that $\varepsilon_n = \|ff_n - g_n\|_\infty \to 0$ as $n \to \infty$. Then for each $n \in \mathbb{Z}^+$ we have

$$\delta \leq \omega(f, \zeta_n) = \omega(ff_n, \zeta_n) = \omega(ff_n - g_n, \zeta_n) \leq 2\varepsilon_n,$$

a contradiction. □

REMARK 17. The above proof actually shows:

If $\mathscr{A}$ *is a subalgebra of* $L^\infty(\mathbb{D})$ *with* $A(\mathbb{D}) \subseteq \mathscr{A} \subseteq C(\overline{\mathbb{D}})$, *and if* $f \in \mathscr{A}_b$, *then* $\inf_{\zeta \in Z} \operatorname{diam} \mathscr{C}(f, \zeta) = 0$, *for any infinite set* $Z \subseteq \mathbb{T}$.

The following corollaries place restrictions on the boundary values of analytic functions in $A(\mathbb{D})_b$.

COROLLARY 18. *If* B *is a Blaschke product whose zero set has infinitely many accumulation points, then* $B \notin AH$.

Proof. If ζ is an accumulation point of the zeros of B, then it is easily verified that $\operatorname{diam} \mathscr{C}(B, \zeta) \geq 1$. □

The above corollary implies that $A(\mathbb{D})_b \neq H^\infty(\mathbb{D})_b$.

COROLLARY 19. *Let* $G \in H^\infty(\mathbb{D})$ *be nonconstant and suppose there is a set* $\{\zeta_k : k \in \mathbb{Z}^+\} \subseteq \mathbb{T}$ *such that* $G^*(\zeta_k) = \alpha$ *for all* $k \in \mathbb{Z}^+$ *and* $\overline{\{\zeta_k : k \in \mathbb{Z}^+\}} \cap \mathbb{T}$ *contains an arc. Then* $G \notin AH$.

Proof. For each ζ in $\overline{\{\zeta_k : k \in \mathbb{Z}^+\}} \cap \mathbb{T}$ one has $\operatorname{diam} \mathscr{C}(G, \zeta) \geq |G^*(\zeta) - \alpha|$. □

7. Remarks and open questions. For Douglas algebras $\mathscr{A}$ and $\mathscr{B}$ a monotonicity theorem is known ([12], Theorem 4): If $H^\infty(\mathbb{T}) \subseteq \mathscr{A} \subseteq \mathscr{B} \subseteq L^\infty(\mathbb{T})$, then $\mathscr{A}_b \subseteq \mathscr{B}_b$. This result fails if the Douglas algebra hypothesis is relaxed. For $\mathscr{A} = C(\mathbb{T})$ and $\mathscr{B} = H^\infty(\mathbb{T})_b$ we have $A(\mathbb{T}) \subseteq \mathscr{A} \subseteq \mathscr{B}$ and $\mathscr{A}_b \not\subseteq \mathscr{B}_b$. To see that $A_b \not\subseteq \mathscr{B}_b$ let $E \subseteq \mathbb{T}$ be a proper arc and consider the characteristic function χ_E. It is straightforward to verify that $\chi_E \in \mathscr{A}_b = C(\mathbb{T})_b$. On the other hand $\chi_E \notin \mathscr{B}_b = H^\infty(\mathbb{T})_{bb} = H^\infty(\mathbb{T})_b = \mathscr{B}$ ([**12**], Corollary 9) because the maximal ideal space of $H^\infty(\mathbb{T})_b$ $(= H^\infty(\mathbb{T}) + C(\mathbb{T}))$ is connected and so $H^\infty(\mathbb{T})_b$ can contain no non-trivial idempotent ([**8**], page 188; [**9**], Chapter IX, Theorem 2.2). Alternatively, one can directly verify that $\chi_E \notin VMO$. However, if $\chi_E \in H^\infty(\mathbb{T})_b$, then $\chi_E \in VMO$ by an argument similar to Theorem 2.3 in Chapter IX of [**9**].

For Bourgain algebras relative to $L^\infty(\mathbb{D})$ monotonicity does not hold in general: $V \subseteq H^\infty(\mathbb{D})_b$, but $V_b \not\subseteq H^\infty(\mathbb{D})_{bb}$. To see that $V_b \not\subseteq H^\infty(\mathbb{D})_{bb}$ note that V is an ideal in $L^\infty(\mathbb{D})$ so that $V_b = L^\infty(\mathbb{D})$, but $H^\infty(\mathbb{D})_{bb} \subseteq BV$. It would be interesting to know whether the monotonicity result of [**12**] holds in the $L^\infty(\mathbb{D})$ setting: if $\mathscr{A}$ and $\mathscr{B}$ are subalgebras of $L^\infty(\mathbb{D})$ with $H^\infty(\mathbb{D}) \subseteq \mathscr{A} \subseteq \mathscr{B}$, is it true that $\mathscr{A}_b \subseteq \mathscr{B}_b$?

For the polydisk ($X = \mathbb{D}^n$ or $X = \mathbb{T}^n$) the Bourgain algebras $A(X)_b = A(X)$ and $H^\infty(X)_b = H^\infty(X)$, have been determined by J. Cima and W. Wogen and also by K. Izuchi [**13**]. On the ball $\mathbb{B}^n$ the result $A(\partial\mathbb{B}^n)_b = C(\partial\mathbb{B}^n)$ (relative to $C(\partial\mathbb{B}^n)$) is implicit in J. Bourgain [**3**]. Concerning $H^\infty(\partial\mathbb{B}^n)_b$, a theorem of W. Rudin [**14**] says that $H^\infty(\partial\mathbb{B}^n) + C(\partial\mathbb{B}^n)$ is a closed subalgebra of $L^\infty(\partial\mathbb{B}^n)$. K. Izuchi [**13**] has recently shown that $H^\infty(\partial\mathbb{B}^n) + C(\partial\mathbb{B}^n) = H^\infty(\partial\mathbb{B}^n)_b$ (relative to $L^\infty(\partial\mathbb{B}^n)$).

REFERENCES

[1] S. Axler, J. B. Conway and G. McDonald, *Toeplitz operators on Bergman spaces*, Canad. J. Math., **34** (1982), 466–483.

[2] J. Bourgain, *New Banach space properties of the disc algebra and* H^∞, Acta Math., **152** (1984), 1–48.

[3] ——, *The Dunford-Pettis property for the ball algebras, the polydisc-algebra and the Sobolev spaces*, Studia Math., **77** (1984), 245–253.

[4] S.-Y. Chang and D. E. Marshall, *Some algebras of bounded analytic functions containing the disc algebra*, pp. 12–20 in *Banach Spaces of Analytic Functions*, Lecture Notes in Math., Vol. 604, Springer Verlag, Berlin and New York, 1977.

[5] J. Cima, S. Janson and K. Yale, *Completely continuous Hankel operators on* H^∞ *and Bourgain algebras*, Proc. Amer. Math. Soc., **105** (1989), 121–125.

[6] J. Cima and R. Timoney, *The Dunford-Pettis property for certain planar uniform algebras*, Michigan Math. J., **34** (1987), 99–104.

[7] B. Cole and T. W. Gamelin, *Tight uniform algebras*, J. Funct. Anal., **46** (1982), 158–220.

[8] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[9] J. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[10] P. Ghatage, S. Sun and D. Zheng, *A remark on Bourgain algebras on the disk*, Proc. Amer. Math. Soc., **114** (1992), 395–398.

[11] P. Gorkin, *Algebras of bounded functions on the disk*, pp. 155–167 in "Proceedings Conference on Function Spaces" (edited by K. Jarosz), Lecture Notes in Pure and Appl. Math., Vol. 136, Marcel Dekker, 1992.
[12] P. Gorkin, K. Izuchi and R. Mortini, *Bourgain algebras of Douglas algebras*, Canad. J. Math., **44** (1992), 797–804.
[13] K. Izuchi, *Bourgain algebras of the disk, polydisk and ball algebras*, Duke Math. J., **66** (1992), 503–519.
[14] W. Rudin, *Spaces of the type* $H^\infty + C$, Ann. Inst. Fourier Grenoble, **25** (1975), 99–125.
[15] K. Stroethoff, *Compact Hankel operators on the Bergman space*, Illinois J. Math., **34** (1990), 159–174.
[16] K. Yale, *Bourgain algebras*, pp. 413–422 in "Proceedings Conference on Function Spaces" (edited by K. Jarosz), Lecture Notes in Pure and Appl. Math., Vol. 136, Marcel Dekker, 1992.
[17] D. Zheng, *Bourgain algebras of some algebras on the disk*, preprint.

Received February 15, 1991 and in revised form April 20, 1992. The first author appreciated the hospitality of the University of Montana while this work was done; the second author acknowledges partial support of grants from the Montana University System and the University of Montana.

University of North Carolina
Chapel Hill, NC 27514

and

University of Montana
Missoula, MT 59812

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

LACUNARY STATISTICAL CONVERGENCE

J. A. FRIDY AND C. ORHAN

The sequence x is statistically convergent to L provided that for each $\varepsilon > 0$,

$$\lim_n n^{-1}\{\text{the number of } k \leq n\colon |x_k - L| \geq \varepsilon\} = 0.$$

In this paper we study a related concept of convergence in which the set $\{k\colon k \leq n\}$ is replaced by $\{k\colon k_{r-1} < k \leq k_r\}$, for some lacunary sequence $\{k_r\}$. The resulting summability method is compared to statistical convergence and other summability methods, and questions of uniqueness of the limit value are considered.

1. Introduction. A complex number sequence x is said to be *statistically convergent* to the number L if for every $\varepsilon > 0$,

$$\lim_n \frac{1}{n}|\{k \leq n\colon |x_k - LK| \geq \varepsilon\}| = 0, \tag{1}$$

where the vertical bars indicate the number of elements in the enclosed set. In this case we write $S\text{-}\lim x = L$ or $x_k \to L(S)$. We shall also use S to denote the set of all statistically convergent sequences. The idea of statistical convergence was introduced by Fast [**4**] and studied by several authors [**2**], [**3**], [**5**], [**6**], [**11**]. There is a natural relationship [**2**] between statistical convergence and strong Cesàro summability:

$$|\sigma_1| := \left\{x\colon \text{ for some } L,\ \lim_n \left(\frac{1}{n}\sum_{k=1}^{n}|x_k - L|\right) = 0\right\}.$$

By a *lacunary sequence* we mean an increasing integer sequence $\theta = \{k_r\}$ such that $k_0 = 0$ and $h_r := k_r - k_{r-1} \to \infty$ as $r \to \infty$. Throughout this paper the intervals determined by θ will be denoted by $I_r := (k_{r-1}, k_r]$, and the ratio k_r/k_{r-1} will be abbreviated by q_r. There is a strong connection [**7**] between $|\sigma_1|$ and the sequence space N_θ, which is defined by

$$N_\theta := \left\{x\colon \text{for some } L,\ \lim_r \left(\frac{1}{h_r}\sum_{k \in I_r}|x_k - L|\right) = 0\right\}.$$

The purpose of this paper is to introduce and study a concept of convergence that is related to statistical convergence (1) in the same way that N_θ is related to $|\sigma_1|$.

DEFINITION. Let θ be a lacunary sequence; the number sequence x is *S_θ-convergent* to L provided that for every $\varepsilon > 0$,

$$\lim_r \frac{1}{h_r} |\{k \in I_r \colon |x_k - L| \geq \varepsilon\}| = 0. \tag{2}$$

In this case we write $S_\theta\text{-}\lim x = L$ or $x_k \to L(S_\theta)$, and we define

$$S_\theta := \{x \colon \text{ for some } L,\ S_\theta\text{-}\lim x = L\}.$$

The limits in (1) and (2) can be expressed using matrix transformations of the characteristic function χ_K of the set

$$K = K(x, L, \varepsilon) := \{k \in \mathbb{N} \colon |x_k - L| \geq \varepsilon\}.$$

The limit in (1) is $\lim_n (C_1 \chi_K)_n = 0$, where C_1 is the Cesàro mean; the limit in (2) is $\lim_n (C_\theta \chi_K)_n = 0$, where C_θ is the matrix given by

$$C_\theta[n, k] := \begin{cases} \dfrac{1}{h_r}, & \text{if } k \in I_r, \\ 0, & \text{if } k \notin I_r. \end{cases}$$

In this form S_θ-convergence is seen to be a part of "A-density convergence" as defined in **[8]** and **[3]**.

In the next section we establish inclusion relations between S_θ and N_θ and also between S_θ and S. In §3 we show that the S_θ-limit of a given sequence x is not necessarily unique for different θ's, but different S_θ-limits cannot occur if $x \in S$. In the final section we get a relationship between S_θ-convergence and strong almost convergence, a concept introduced by Maddox **[10]** and (independently) by Freedman et al. **[7]**.

2. Inclusion theorems.

In this section we first give some inclusion relations between N_θ- and S_θ-convergence and show that they are equivalent for bounded sequences. We also study the inclusions $S \subseteq S_\theta$ and $S_\theta \subseteq S$ under certain restrictions on $\theta = \{k_r\}$.

THEOREM 1. *Let $\theta = \{k_r\}$ be a lacunary sequence; then*

(i) (a) $x_k \to L(N_\theta)$ *implies* $x_k \to L(S_\theta)$, *and*
   (b) N_θ *is a proper subset of* S_θ;
(ii) $x \in l_\infty$ *and* $x_k \to L(S_\theta)$ *imply* $x_k \to L(N_\theta)$;
(iii) $S_\theta \cap l_\infty = N_\theta \cap l_\infty$,

where l_∞ denotes the set of bounded sequences.

Before proving this theorem we remark that this result is included by Theorem 8 in **[3]**, where Connor bases the proof on the concept of ideals in l_∞; we give a direct proof.

Proof. (a) If $\varepsilon > 0$ and $x_k \to L(N_\theta)$ we can write

$$\sum_{k \in I_r} |x_k - L| \geq \sum_{\substack{k \in I_r \\ |x_k - L| \geq \varepsilon}} |x_k - L| \geq \varepsilon |\{k \in I_r \colon |x_k - L| \geq \varepsilon\}|,$$

which yields the result.

(b) In order to establish that the inclusion $N_\theta \subseteq S_\theta$ in (i) is proper, let θ be given and define x_k to be $1, 2, \ldots, [\sqrt{h_r}]$ at the first $[\sqrt{h_r}]$ integers in I_r, and $x_k = 0$ otherwise. Note that x is not bounded. We have, for every $\varepsilon > 0$,

$$\frac{1}{h_r} |\{k \in I_r \colon |x_k - 0| \geq \varepsilon\}| = \frac{[\sqrt{h_r}]}{h_r} \to 0 \qquad \text{as } r \to \infty,$$

i.e., $x_k \to 0(S_\theta)$. On the other hand,

$$\frac{1}{h_r} \sum_{k \in I_r} |x_k - 0| = \frac{1}{h_r} \frac{[\sqrt{h_r}]([\sqrt{h_r}] + 1)}{2} \to \frac{1}{2} \neq 0;$$

hence $x_k \not\to 0(N_\theta)$.

(ii) Suppose that $x_k \to L(S_\theta)$ and $x \in l_\infty$, say $|x_k - L| \leq M$ for all k. Given $\varepsilon > 0$, we get

$$\begin{aligned} \frac{1}{h_r} \sum_{k \in I_r} |x_k - L| &= \frac{1}{h_r} \sum_{\substack{k \in I_r \\ |x_k - L| \geq \varepsilon}} |x_k - L| + \frac{1}{h_r} \sum_{\substack{k \in I_r \\ |x_k - L| < \varepsilon}} |x_k - L| \\ &\leq \frac{M}{h_r} |\{k \in I_r \colon |x_k - L| \geq \varepsilon\}| + \varepsilon, \end{aligned}$$

from which the result follows.

We remark that the example given in (i) shows that the boundedness condition cannot be omitted from the hypothesis of Theorem 1 (ii).

(iii) This is an immediate consequence of (i) and (ii).

Since any N_θ-summable sequence is C_θ-summable, we conclude from Theorem 1 (ii) that any bounded S_θ-summable sequence is also C_θ-summable.

LEMMA 2. *For any lacunary sequence* θ, S-$\lim x = L$ *implies* S_θ-$\lim x = L$ *if and only if* $\liminf_r q_r > 1$. *If* $\liminf_r q_r = 1$, *then there exists a bounded* S_θ*-summable sequence that is not* S*-summable* (*to any limit*).

Proof. Suppose first that $\liminf_r q_r > 1$; then there exists a $\delta > 0$ such that $q_r \geq 1 + \delta$ for sufficiently large r, which implies that

$$\frac{h_r}{k_r} \geq \frac{\delta}{1+\delta}.$$

If $x_k \to L(S)$, then for every $\varepsilon > 0$ and for sufficiently large r, we have

$$\begin{aligned} \frac{1}{k_r}|\{k \leq k_r\colon |x_k - L| \geq \varepsilon\}| &\geq \frac{1}{k_r}|\{k \in I_r\colon |x_k - L| \geq \varepsilon\}| \\ &\geq \frac{\delta}{1+\delta} \cdot \frac{1}{h_r}|\{k \in I_r\colon |x_k - L| \geq \varepsilon\}|; \end{aligned}$$

this proves the sufficiency.

Conversely, suppose that $\liminf_r q_r = 1$. Proceeding as in [7; p. 510] we can select a subsequence $\{k_{r(j)}\}$ of the lacunary sequence θ such that

$$\frac{k_{r(j)}}{k_{r(j)-1}} < 1 + \frac{1}{j} \quad \text{and} \quad \frac{k_{r(j)-1}}{k_{r(j-1)}} > j, \quad \text{where } r(j) \geq r(j-1) + 2.$$

Now define a bounded sequence x by $x_i = 1$ if $i \in I_{r(j)}$ for some $j = 1, 2, \ldots$ and $x_i = 0$ otherwise. It is shown in [7; p. 510] that $x \notin N_\theta$ but $x \in |\sigma_1|$. The above Theorem 1 (ii) implies that $x \notin S_\theta$, but it follows from Theorem 2.1 of [2] that $x \in S$. Hence $S \not\subseteq S_\theta$, and the proof is complete.

LEMMA 3. *For any lacunary sequence θ, S-$\lim x = L$ implies S_θ-$\lim x = L$ if and only if $\limsup_r q_r < \infty$. If $\limsup_r q_r = \infty$, then there exists a bounded S-summable sequence that is not S_θ-summable (to any limit).*

Proof. If $\limsup_r q_r < \infty$, then there is an $H > 0$ such that $q_r < H$ for all r. Suppose that $x_k \to L(S_\theta)$, and let $N_r := |\{k \in I_r\colon |x_k - L| \geq \varepsilon\}|$. By (2), given $\varepsilon > 0$, there is an $r_0 \in \mathbb{N}$ such that

$$\frac{N_r}{h_r} < \varepsilon \quad \text{for all } r > r_0. \tag{3}$$

Now let $M := \max\{N_r\colon 1 \leq r \leq r_0\}$ and let n be any integer satisfying

$k_{r-1} < n \leq k_r$; then we can write

$$\begin{aligned}
\frac{1}{n}|\{k \leq n \colon |x_k - L| \geq \varepsilon\}| &\leq \frac{1}{k_{r-1}}|\{k \leq k_r \colon |x_k - L| \geq \varepsilon\}| \\
&= \frac{1}{k_{r-1}}\{N_1 + N_2 + \cdots + N_{r_0} + N_{r_0+1} + \cdots + N_r\} \\
&\leq \frac{M}{k_{r-1}} \cdot r_0 + \frac{1}{k_{r-1}}\left\{h_{r_0+1}\frac{N_{r_0+1}}{h_{r_0+1}} + \cdots + h_r\frac{N_r}{h_r}\right\} \\
&\leq \frac{r_0 \cdot M}{k_{r-1}} + \frac{1}{k_{r-1}}\left(\sup_{r>r_0}\frac{N_r}{h_r}\right)\{h_{r_0+1} + \cdots + h_r\} \\
&\leq \frac{r_0 \cdot M}{k_{r-1}} + \varepsilon \cdot \frac{k_r - k_{r_0}}{k_{r-1}}, \quad \text{by (3)}, \\
&\leq \frac{r_0 \cdot M}{k_{r-1}} + \varepsilon \cdot q_r \leq \frac{r_0 \cdot M}{k_{r-1}} + \varepsilon H,
\end{aligned}$$

and the sufficiency follows immediately.

Conversely, suppose that $\limsup_r q_r = \infty$. Following the idea in [7; p. 511] we can select a subsequence $\{k_{r(j)}\}$ of the lacunary sequence $\theta = \{k_r\}$ such that $q_{r(j)} > j$, and define a bounded sequence by $x_i = 1$ if $k_{r(j)-1} < i \leq 2k_{r(j)-1}$ for some $j = 1, 2, \ldots$, and $x_i = 0$ otherwise. It is shown in [7; p. 5.11] that $x \in N_\theta$ but $x \notin |\sigma_1|$. By Theorem 1 (i) we conclude that $x \in S_\theta$, but Theorem 2.1 of [2] implies that $x \notin S$. Hence, $S_\theta \not\subseteq S$.

Combining Lemma 2 and Lemma 3 we get

THEOREM 4. *Let* θ *be a lacunary sequence*; *then* $S = S_\theta$ *if and only if*

$$1 < \liminf_r q_r \leq \limsup_r q_r < \infty;$$

then $S\text{-}\lim x = L$ *implies* $S_\theta\text{-}\lim x = L$.

For an example of a lacunary sequence satisfying the conditions of Theorem 4, we can take $k_r = 2^r$ for $r > 0$, whence $S_{\{2^r\}} = S$. We remark that the examples given in Lemmas 2 and 3 illustrate the difference between S-convergence and S_θ-convergence.

We conclude this section with the following observation. Buck [**1**, Theorem 3.2] proved that if a real sequence is C_1-summable to its finite limit inferior, then the sequence "converges to that point for almost all n" (i.e., it is statistically convergent to its limit inferior [**2**]). Note that this result remains true if we replace limit inferior by limit superior. For each subset K of $\mathbb{N}$, define

$$D(K) := \lim_r (C_\theta \chi_K)_r = \lim_r \frac{|K \cap I_r|}{h_r};$$

then D is a density [**8**; p. 296], and it is not hard to get a result for S_θ-convergence that is analogous to Buck's. To be precise, the following result is such an analogue.

PROPOSITION 5. *If the real number sequence x is C_θ-summable to either its finite limit inferior or finite limit superior, then x is S_θ-convergent to that value.*

3. Uniqueness of S_θ-limit and lacunary refinements. It is easy to see that, for any fixed θ, the S_θ-limit is unique. It is possible, however, for a sequence—even a bounded one—to have different S_θ-limits for different θ's. This can be seen by applying Theorem 1 (i) to the sequence x given in [7, proof of Theorem 2.1] for which $N_{\theta_1}\text{-}\lim x = 0$ and $N_{\theta_2}\text{-}\lim x = 1$. The next theorem shows that this situation cannot occur if $x \in S$; in other words, every S_θ method is consistent with the S-method.

THEOREM 6. *If $x \in S \cap S_\theta$, then $S_\theta\text{-}\lim x = S\text{-}\lim x$.*

Proof. Suppose $S\text{-}\lim x = L$ and $S_\theta\text{-}\lim x = L'$, and $L \neq L'$. For $\varepsilon < \frac{1}{2}|L - L'|$ we get

$$\lim_n \frac{1}{n}|\{k \leq n \colon |x_k - L'| \geq \varepsilon\}| = 1.$$

Consider the k_mth term of the statistical limit expression $n^{-1}|\{k \leq n \colon |x_k - L'| \geq \varepsilon\}|$:

$$\text{(4)} \quad \frac{1}{k_m}\left|\left\{k \in \bigcup_{r=1}^{m} I_r \colon |x_k - L'| \geq \varepsilon\right\}\right|$$
$$= \frac{1}{k_m}\sum_{r=1}^{m} |\{k \in I_r \colon |x_k - L'| \geq \varepsilon\}| = \frac{1}{\sum_{r=1}^{m} h_r}\sum_{r=1}^{m} h_r t_r,$$

where $t_r = h_r^{-1}|\{k \in I_r \colon |x_k - L'| \geq \varepsilon\}| \to 0$ because $x_k \to L'(S_\theta)$. Since θ is a lacunary sequence, (4) is a regular weighted mean transform of t, and therefore it, too, tends to zero as $m \to \infty$. Also, since this is a subsequence of $\{n^{-1}|\{k \leq n \colon |x_k - L'| \geq \varepsilon\}|\}_{n=1}^{\infty}$, we infer that

$$\frac{1}{n}|\{k \leq n \colon |x_k - L'| \geq \varepsilon\}| \not\to 1,$$

and this contradiction shows that we cannot have $L \neq L'$.

We now consider the inclusion of $S_{\theta'}$ by S_θ, where θ' is a lacunary refinement of θ. Recall [7] that the lacunary sequence $\theta' = \{k'_r\}$ is called a *lacunary refinement* of the lacunary sequence $\theta = \{k_r\}$ if $\{k_r\} \subseteq \{k'_r\}$.

THEOREM 7. *If θ' is a lacunary refinement of θ and $x_k \to L(S_{\theta'})$, then $x_k \to L(S_\theta)$.*

Proof. Suppose each I_r of θ contains the points $\{k'_{r,i}\}_{i=1}^{\nu(r)}$ of θ' so that

$$k_{r-1} < k'_{r,1} < k'_{r,2} < \cdots < k'_{r,\nu(r)} = k_r, \quad \text{where } I'_{r,i} = (k'_{r,i-1}, k'_{r,i}].$$

Note that for all r, $\nu(r) \geq 1$ because $\{k_r\} \subseteq \{k'_r\}$. Let $\{I^*_j\}_{j=1}^{\infty}$ be the sequence of abutting intervals $\{I'_{r,i}\}$ ordered by increasing right end points. Since $x_k \to L(S_{\theta'})$, we get, for each $\varepsilon > 0$,

$$\lim_j \sum_{I^*_j \subset I_r} \frac{1}{h^*_r} |\{k \in I^*_j : |x_k - L| \geq \varepsilon\}| = 0. \tag{5}$$

As before we write, $h_r = k_r - k_{r-1}$, $h'_{r,i} = k'_{r,i} - k'_{r,i-1}$, and $h'_{r,1} = k'_{r,1} - k_{r-1}$. For each $\varepsilon > 0$ we have

$$\begin{aligned} &\frac{1}{h_r} |\{k \in I_r : |x_k - L| \geq \varepsilon\}| \\ &\quad = \frac{1}{h_r} \sum_{I^*_j \subseteq I_r} h^*_j \frac{1}{h^*_j} |\{k \in I^*_j : |x_k - L| \geq \varepsilon\}| \\ &\quad = \frac{1}{h_r} \sum_{I^*_j \subseteq I_r} h^*_j (C_{\theta'} \chi_K)_j, \end{aligned} \tag{6}$$

where χ_K is the characteristic function of the set $K := \{k \in \mathbb{N} : |x_k - L| \geq \varepsilon\}$. By (5), $C_{\theta'}\chi_K$ is a null sequence, and (6) is a regular weighted mean transform of $C_{\theta'}\chi_K$. Hence, the transform (6) also tends to zero as $r \to \infty$.

We conclude this section by observing that Theorem 7 establishes inclusion between two lacunary methods *only* when one sequence is a lacunary refinement of the other. The example cited at the beginning of this section shows that S_θ can be inconsistent with $S_{\theta'}$. A general description of inclusion between two arbitrary lacunary methods is left as an open problem.

4. Strong almost convergence and S_θ-convergence. The idea of almost convergence was introduced by Lorentz **[9]**: the sequence x is said to be *almost convergent* to L if

$$\lim_n \frac{1}{n} \sum_{i=m+1}^{m+n} (x_i - L) = 0, \quad \text{uniformly in } m.$$

Maddox **[10]** and (independently) Freedman et al. **[7]** introduced the notion of strong almost convergence: the sequence x is said to be *strongly almost convergent* to L if

$$\lim_n \frac{1}{n} \sum_{i=m+1}^{m+n} |x_i - L| = 0, \quad \text{uniformly in } m.$$

Let c, AC and $[AC]$, respectively, denote the sets of all convergent, almost convergent, and strongly almost convergent sequences. It is known **[10]** that

$$c \subsetneqq [AC] \subsetneqq AC \subsetneqq l_\infty. \tag{7}$$

THEOREM 8. *If $\mathscr{L}$ denotes the set of all lacunary sequences, then*

$$[AC] = l_\infty \cap \left(\bigcap_{\theta\in\mathscr{L}} S_\theta \right).$$

Proof. By [7, Theorem 3.1], the relations (7) and Theorem 1 (iii), we have

$$\begin{aligned} l_\infty \supset [AC] &= \bigcap_{\theta\in\mathscr{L}} N_\theta = l_\infty \cap \left(\bigcap_{\theta\in\mathscr{L}} N_\theta \right) \bigcap_{\theta\in\mathscr{L}} (l_\infty \cap N_\theta) \\ &= \bigcap_{\theta\in\mathscr{L}} (l_\infty \cap S_\theta) = l_\infty \cap \left(\bigcap_{\theta\in\mathscr{L}} S_\theta \right). \end{aligned}$$

Finally we remark that in contrast to [7, Theorem 3.1] where it was proved that $[AC] = \bigcap N_\theta$, the factor l_∞ cannot be omitted from Theorem 8. For, $\bigcap S_\theta \not\subseteq l_\infty$ and $\bigcap N_\theta = [AC]$ is a proper subset of $\bigcap S_\theta$. To see this consider the sequence x defined by $x_k = m$, if $k = m^2$ for $m = 1, 2, \ldots$, and $x_k = 0$ otherwise. Observe that x is not bounded, so it is not strongly almost convergent. On the other hand, for any lacunary sequence θ, we have

$$\frac{1}{h_r} |\{k \in I_r : x_k \neq 0\}| \leq \frac{\sqrt{h_r}}{h_r} \to 0, \quad \text{as } r \to \infty;$$

hence, $x_k \to O(S_\theta)$.

The authors wish to thank the referee for several very helpful suggestions that have improved the exposition of these results.

REFERENCES

[1] R. C. Buck, *Generalized asymptotic density,* Amer. J. Math., **75** (1953), 335-346.

[2] J. Connor, *The statistical and strong p-Cesàro convergence of sequences,* Analysis, **8** (1988), 47-63.

[3] ——, *On strong matrix summability with respect to a modulus and statistical convergence,* Canad. Math. Bull., **32** (1989), 194-198.

[4] H. Fast, *Sur la convergence statistique,* Colloq. Math., **2** (1951), 241-244.

[5] J. A. Fridy, *On statistical convergence,* Analysis, **5** (1985), 301-313.

[6] J. A Fridy and H. I. Miller, *A matrix characterization of statistical convergence,* Analysis, **11** (1991), 55-66.

[7] A. R. Freedman, J. J. Sember, and M. Raphael, *Some Cesàro type summability spaces,* Proc. London Math. Soc., **37** (1978), 508-520.

[8] A. R. Freedman and J. J. Sember, *Densities and summability,* Pacific J. Math., **95** (1981), 293-305.

[9] G. G. Lorentz, *A contribution to the theory of divergent sequences,* Acta Math., **80** (1948), 167-190.

[10] I. J. Maddox, *A new type of convergence,* Math. Proc. Cambridge Phil. Soc., **83** (1978), 61-64.

[11] I. J. Schoenberg, *The integrability of certain functions and related summability methods,* Amer. Math. Monthly, **66** (1959), 361-375.

Received November 11, 1990 and in revised form March 16, 1992. The second author's research was supported by the Scientific and Technical Research Council of Turkey.

KENT STATE UNIVERSITY
KENT, OH 44242
U.S.A.

AND

ANKARA UNIVERSITY
ANKARA, 06100
TURKEY

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

ON THE SHAPE OF FUNDAMENTAL DOMAINS IN GL(n, R)/O(n)

DOUGLAS GRENIER

We investigate parameters for the symmetric space $H = G/K$, $G = \mathrm{GL}(n, \mathbf{R})$, $K = \mathrm{O}(n)$, in the sense of positive definite quadratic forms. This leads to a description for the fundamental domain H/Γ where Γ is an arithmetic subgroup of G. We also see interesting relations with the Siegel sets. This enables us to explicitly describe Satake compactifcations of H/Γ. We will also consider the behavior at the "bottom" of the fundamental domains.

1. Introduction. The problem of reduction of quadratic forms is an old one. When the subject is positive definite quadratic forms, the first definition of reduction was achieved by Hermite [**8**]. However, it is Minkowski's reduction that is the most familiar, primarily because when we view the quadratic forms geometrically, Minkowski's reduction domain is easily seen to have a finite number of boundary components while this was not known for Hermite's. One may consult Cassels [**3**] or Terras [**14**] for more of the historical details and for a definition of these domains.

In modern language, positive definite quadratic forms may be considered as a symmetric space. Denote the space of positive quadratic forms by $\mathscr{P}_n$, and let $G = \mathrm{GL}(n, \mathbf{R})$, $K = \mathrm{O}(n)$. Then $\mathscr{P}_n$ may be identified with G/K by:

$$gK \mapsto {}^T gg$$

for any $g \in G$. We will be most interested in the space of quadratic forms of determinant one which will be denoted $\mathscr{SP}_n$. This may be obtained from G/K by modding out by the center of G. If Γ_n stands for $\mathrm{GL}(n, \mathbf{Z})/\{\pm I\}$, Γ_n acts discontinuously on $\mathscr{SP}_n$ by:

$$Z \mapsto Z[\gamma] = {}^T\gamma Z \gamma$$

where $Z \in \mathscr{SP}_n$ and $\gamma \in \Gamma_n$. We will always use this notation $Z[X] = {}^T XZX$ where $X \in \mathbf{R}^{n\times k}$ for any k (including $k = 1$ so that X is a vector). Then a fundamental domain in $\mathscr{SP}_n$ under the action of Γ_n is a subset of $\mathscr{SP}_n$ which may be identified with the quotient space $\mathscr{SP}_n/\Gamma_n$ and which represents the reduced forms. For studying arithmetic subgroups of $\mathrm{GL}(n, \mathbf{R})$ and automorphic forms an Hermite style fundamenal domain was preferable to Minkowski's. In **[5]** such a fundamental domain was defined, and it was shown that it was indeed possible to explicitly determine the (bottom) boundary components. This is also in **[6]**.

The object of this paper is to develop in more depth various properties of the fundamental domains of **[5]** and **[6]**. We will explicitly determine compactifications of these domains and hence the quotient groups. These will be Satake compactifications of $\mathscr{SP}_n/\Gamma_n$ modelled after those originally considered by Satake **[12]** for the Siegel space and generalized by Satake and others (see, for example, **[13, 1, 16]**) to many algebraic groups. We will describe the topologies of certain compactifications of the fundamental domains and the quotient groups, and see the relation between the so-called boundary components and the parabolic subgroups of G, as in the compactifications of Satake and Baily-Borel. We believe that these compactifications will be quite useful in the study of automorphic forms for $\mathrm{GL}(n, \mathbf{R})$, specifically towards generalizing the results of **[7]** that defined a map sending automorphic forms for $\mathrm{GL}(n, \mathbf{R})$ to forms for $\mathrm{GL}(n-1, \mathbf{R})$. The nature of the Satake topology defined in §4 indicates that there should be similar maps relative to each parabolic subgroup, sending an automorphic form for $\mathrm{GL}(n, \mathbf{R})$ to a product of lower rank automorphic forms. This in turn might lead to explicit Maass-Selberg relations for $\mathrm{GL}(n, \mathbf{R})$. The Siegel sets play an important role both in the definition of the fundamental domain and the compactification. The relation of the fundamental domain to the Siegel sets also has a bearing on many problems in automorphic forms and harmonic analysis on $\mathrm{GL}(n, \mathbf{R})$, for example as in Huntley **[9]** (also see **[14]**). Mostly we concentrate on the action of $\mathrm{GL}(n, \mathbf{Z})$ on $\mathscr{SP}_n$, but since an arithmetic subgroup of G is commensurable with $\mathrm{GL}(n, \mathbf{Z})$, these ideas are easily extended to fundamental domains for arbitrary arithmetic subgroups. Due to the celebrated arithmeticity theorems of Margulis (summarized in **[15]**), this last notion may deserve more attention than it is given here.

In the following, in §2 we will briefly summarize the results of **[5, 6]** concerning the fundamental domain we will consider here. Then these results will be modified and/or extended to suit the purposes of this paper. Section 3 will take a closer look at the Siegel sets (which are already preset in the definition of the fundamental domain). We will investigate the relations between the fundamental domain and the Siegel sets, including how they are used to establish a reduction algorithm, and in work of Huntley on the spectrum of the Laplacian. In §4 the fundamental domain will be compactified in the style of Satake **[12]**. This also gives a compactification of the quotient space $\mathcal{SP}_n/\Gamma_n$. The author would like to thank S. Zucker for several helpful discussions on these last matters.

2. Fundamental domains in $\mathcal{SP}_n$. This section is partly a summary of results from **[5, 6]**, although we have made various modifications and additions. Throughout we will use the following coordinates on $\mathcal{SP}_n$:

For any $Z \in \mathcal{SP}_n$, let

$$(1)\qquad Z = \begin{pmatrix} y^{-1} & O \\ O & y^{1/(n-1)}Z' \end{pmatrix} \begin{bmatrix} 1 & {}^T x \\ 0 & I \end{bmatrix}$$

where the square bracket notation is as before, and $y > 0$, $Z' \in \mathcal{SP}_{n-1}$, $x \in \mathbf{R}^{n-1}$. These may appear a little strange at first, but they have been chosen to point out parallels with the upper half-plane H (where x and y would just be the real and imaginary parts of $z \in H$), and so that $|Z|' = 1$, which enables us to define the coordinates recursively. More explicitly the identification of $\mathcal{SP}_2$ with H is given by:

$$H \to \mathcal{SP}_2,$$
$$x + iy = z \mapsto \begin{pmatrix} y^{-1} & 0 \\ 0 & y \end{pmatrix} \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix}.$$

The coordinates defined in (1) are called partial Iwasawa coordinates since if we repeat this decomposition for Z', and so on, we get the Iwasawa decomposition for $Z \in \mathcal{SP}_n$, which it is convenient to write

$$(2)\quad Z = y^{-1} \begin{pmatrix} 1 & & & & \\ & y_1^2 & & & \\ & & (y_1 y_2)^2 & & \\ & & & \ddots & \\ & & & & (y_1 y_2 \cdots y_{n-1})^2 \end{pmatrix} \begin{bmatrix} 1 & & x_{ij} \\ & \ddots & \\ & & 1 \end{bmatrix}.$$

Note that the y_i correspond to the simple roots of the Lie algebra of G, which in the Lie group setting are actually y_i^{-2} for $i = 1, 2, \ldots, n-1$. The partial Iwasawa coordinates demonstrate an embedding of the lower rank symmetric spaces $\mathcal{SP}_m$ into $\mathcal{SP}_n$.

Using the partial Iwasawa decomposition (1) we can now define a fundamental domain for Γ_n in $\mathcal{SP}_n$, corresponding to the quotient space $\mathcal{SP}_n/\Gamma_n$. Let $\mathcal{F}_n$ be the set of all $Z \in \mathcal{SP}_n$ satisfying:

(D1) For all $\left(\begin{smallmatrix} a & {}^T b \\ c & D \end{smallmatrix}\right) \in \Gamma_n$, $(a + {}^T x c)^2 + y^{n/(n-1)} Z'[c] \geq 1$ where $a \in \mathbf{Z}$, $b, c \in \mathbf{Z}^{n-1}$, and $D \in \mathbf{Z}^{n-1 \times n-1}$.

(D2) $Z' \in \mathcal{F}_{n-1}$.

(D3) $0 \leq x_{12} \leq 1/2$, $|x_{1j}| \leq 1/2$ for $j = 3, \ldots, n-1$.

This is slightly different from the definition in [**5**, **6**], as there the approach was more in keeping with the idea of quadratic forms (and not necessarily just those of determinant one). Here, we want to see a cusp as $y \to \infty$. As indicated in the introduction, this fundamental domain has the advantage of being defined recursively as in Hermite's original idea, but it was shown in [**5**, **6**] that only a finite number of combinations of $a \in \mathbf{Z}$ and $c \in \mathbf{Z}^{n-1}$ are necessary in condition (D1).

Since $\mathcal{SP}_2$ is identified with the upper half-plane H, the fundamental domain above for $n = 2$ is just half of the usual one in H, namely $\mathcal{F}_2$ is identified with $D \subset H$ where

$$D = \{x + iy = z \in H | x^2 + y^2 \geq 1, 0 \leq x \leq 1/2\}.$$

$\mathcal{F}_2$ is compactified by adding the point at infinity, i.e., $\mathcal{F}_2{}^* = \mathcal{F}_2 \cup \{\infty\}$ is a compact subset of $H^* = H \cup \{\infty\}$ corresponding to H^*/Γ_2. To see how to compactify $\mathcal{F}_n$ when $n > 2$ we will first consider the example of $\mathcal{F}_3$. This fundamental domain was pictured in [**4**]. For $n = 3$ the Iwasawa decomposition (2) becomes:

$$Z = y^{-1} \begin{pmatrix} 1 & & \\ & y_1^2 & \\ & & (y_1 y_2)^2 \end{pmatrix} \begin{bmatrix} 1 & x_{12} & x_{13} \\ & 1 & x_{23} \\ & & 1 \end{bmatrix}.$$

The fundamental domain $\mathcal{F}_3$ can be given as the set of all $Z \in \mathcal{SP}_3$ satisfying the following inequalities:

(i) $x_{12}^2 + y_1^2 \geq 1$,
(ii) $x_{12}^2 + y_1^2(x_{23}^2 + y_2^2) \geq 1$,
(iii) $(x_{12} - x_{13})^2 + y_1^2((1 - x_{23})^2 + y_2^2) \geq 1$,
(iv) $(1 - x_{12} + x_{13})^2 + y_1^2((1 - x_{23})^2 + y_2^2) \geq 1$,
(v) $x_{23}^2 + y_2^2 \geq 1$,
(vi) $0 \leq x_{12} \leq 1/2$,
(vii) $0 \leq x_{23} \leq 1/2$,
(viii) $|x_{13}| \leq 1/2$.

Inequalities (i)–(iv) come from condition (D1) in the definition, while (v) and (vii) come from (D2). For $y_1 \geq M$, for some sufficiently large M, the first four inequalities above are unnecessary to have $Z \in \mathcal{F}_3$. Thus we might say that as $y_1 \to \infty$ what was the point at infinity for $n = 2$ now becomes $\mathcal{F}_2$ for $n = 3$. As $y_2 \to \infty$ we get similar behavior except that it is inequalities (ii)–(v) that become unnecessary. However, we still see $\mathcal{F}_2$ as we approach the cusp determined by $y_2 \to \infty$.

To see how this would carry over to the general situation it will be convenient to write $Z \in \mathcal{F}_n$ as:

$$Z = y^{-1} \begin{pmatrix} 1 & O \\ O & Z_1 \end{pmatrix} \begin{bmatrix} 1 & {}^T x \\ 0 & I \end{bmatrix} \tag{3}$$

so $Z_1 = y^{(n/n-1)} Z'$ from the partial Iwasawa decomposition (1). Let us define $a_1(Z)$ to be the upper left corner entry of the matrix $Z \in \mathcal{P}_n$. Then

$$a_1(Z[\gamma]) = Z \begin{bmatrix} a \\ c \end{bmatrix}$$

where $\gamma = \left(\begin{smallmatrix} a & {}^T b \\ c & D \end{smallmatrix}\right)$ as usual. A simple computation gives

$$yZ \begin{bmatrix} a \\ c \end{bmatrix} = (a + {}^T x c)^2 + Z_1[c].$$

If $c = \left(\begin{smallmatrix} a' \\ c' \end{smallmatrix}\right)$, with $c' \in \mathbf{Z}^{n-2}$, then we can also decompose Z' as we did for Z above and compute $a_1(Z')$. Thus

$$y' Z'[c] = (a' + {}^T x' c')^2 + Z_2[c']$$

where x', y' and Z_2 would correspond to x, y and Z_1 for Z. We also know $y'Z' = y_1^{-2} Z_1$, so we can rewrite the above equation as:

$$y_1^{-2} Z_1[c] = (a' + {}^T x' c')^2 + Z_2[c']. \tag{4}$$

If we repeat the arguments above we eventually arrive at the following:

Lemma 1. *$Z_i[c] \geq y_i^2$ for all $c \in \mathbf{Z}^{n-i} - \{0\}$.*

Since for each i, $y_i \to \infty \Rightarrow y \to \infty$ (so $y^{-1} \to 0$), as $y_i \to \infty$, $\mathbf{Z}$ breaks into two blocks of sizes i and $n - i$, the first of which goes to 0, the second to ∞. We will write:

$$\text{As } y_i \to \infty,\ \mathcal{F}_n \to \mathcal{F}_i \times \mathcal{F}_{n-i}.$$

A variation of Lemma 1 will also be useful:

Lemma 2. *$Z_1[c] \geq y_1^2 \cdots y_k^2$ for all $c \in \mathbf{Z}^{n-1}$ which have $c_j \neq 0$ for all $j \geq k$.*

This may be proved by applying equation (4) and Lemma 1.

3. Siegel sets in $\mathcal{SP}_n$. In this section it will sometimes be convenient to consider the region $\mathcal{F}_n' = \bigcup_{\gamma \in D_n} \mathcal{F}_n[\gamma]$ where D_n is the subgroup of diagonal matrices

$$\begin{pmatrix} \pm 1 & & \\ & \ddots & \\ & & \pm 1 \end{pmatrix}.$$

We will denote this fundamental domain $\mathcal{F}_n'$. Condition (D3) would become for $\mathcal{F}_n'$:

(D3′) $|x_i| \leq 1/2$ for all i.

while (D2) would become accordingly:

(D2′) $Z' \in \mathcal{F}_{n-1}'$,

with (D1) remaining the same. Then $\mathcal{F}_n'$ is symmetric about $x_{ij} = 0$ for all i, j with $1 \leq i < j \leq n$.

The Iwasawa decomposition of G is written $G = KAN$, where $K = \mathrm{O}(n)$, A is the diagonal subgroup of G, and N is the nilpotent subgroup of upper triangular matrices with 1 on the diagonal. A Siegel set in G is a subset of the form $K \cdot A_t \cdot w$ where $A_t = \{a \in A | \alpha(a) \leq t\}$ for all simple roots α and w is a compact subgroup of N containing a neighborhood of I. More details may be found in [2].

Analogously, we may define Siegel sets for $\mathcal{SP}_n$. Since we have seen that the y_i in the decomposition (2) correspond to the simple roots of the Lie algebra of G and under the map $g \mapsto {}^T g g$, $g = kan$ is sent to $a^2[n]$, define the Siegel set $\mathcal{S}_{t,1/2}$ by

$$\mathcal{S}_{t,1/2} = \{Z \in \mathcal{SP}_n | y_i \geq t^{-1/2}, |x_{ij}| \leq 1/2\}.$$

Then we have the following:

THEOREM 1. $\mathcal{S}_{1,1/2} \subset \mathcal{F}_n' \subset \mathcal{S}_{4/3,1/2}$.

Proof. Clearly this is true for $n = 2$. One need only consider the standard picture of $\mathcal{F}_2'$ in the upper half-plane H. Using the partial Iwasawa decomposition (3), condition (D1) of the definition of $\mathcal{F}_n$ becomes:

(D1*) $(a + {}^T x c)^2 + Z_1[c] \geq 1$

for all a, c forming the first column of $\gamma \in \Gamma_n$. If we choose $a = 0$, $c = e_1$, the first standard unit vector in $\mathbf{R}^{n-1}$ we get $x_1^2 + y_1^2 \geq 1$ and since $|x_1| \leq 1/2$, $y_1 \geq 3/4$. This argument can be applied in turn to $Z' \in \mathcal{F}_{n-1}'$ to get $y_2 \geq 3/4$ and so on. Thus we have the second inclusion. To get the first we need to show that if $y_i \geq 1$ for all i, and $|x_{ij}| \leq 1/2$ for all i, j, then $Z \in \mathcal{F}_n'$. As mentioned above, we know this for $n = 2$, so we proceed by induction. Since Z_1 is positive, if λ is the least eigenvalue of Z_1, then $Z_1[c] \geq \lambda I[c]$. The eigenvalues of Z_1 are y_1^2, $(y_1y_2)^2$, etc., so $\lambda \geq 1$. If $c \neq 0$, $Z_1[c] \geq \lambda \geq 1$ so condition (D1*) is satisfied. If $c = 0$ then a must be ± 1 to have $\gamma \in \Gamma_n$ which again makes (D1*) satisfied. Condition (D2′) is met by the induction hypothesis and (D3′) by definition of $\mathcal{S}_{t,1/2}$.

Terras **[14]** says that $\mathcal{F}_n'$ has a "box shape" at infinity. For example, considering $n = 2$, the portion of $\mathcal{F}_2'$ with $y > 1$ is a semi-infinite strip of width one. It can be seen fairly easily that $\mathcal{S}_{1,1/2}$ is the largest Siegel set contained in $\mathcal{F}_n'$ and $\mathcal{S}_{4/3,1/2}$ is the smallest containing $\mathcal{F}_n'$. Again, this is well known in the case $n = 2$, where

$$\mathcal{S}_{t,1/2} = \{z \in H\colon |x| \leq 1/2, y \geq \sqrt{t^{-1}}\}.$$

It is also clear from that same picture that

$$\mathcal{S}_{4/3,1/2} \subset \mathcal{F}_2' \cup \mathcal{F}_2'[\mathcal{S}]$$

where $S = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Roelcke **[11]** uses the fact that it takes just two copies of the fundamental domain to cover the smallest Siegel set containing the fundamental domain to obtain a lower bound for the eigenvalues of the Laplacian on $L^2(H/\mathrm{SL}(2, \mathbf{Z}))$. In this case, $|\lambda| > 3\pi^2/2$. Huntley **[10]** has extended these ideas to $L^2(\mathcal{SP}_3/\Gamma_3)$ and shows that there $|\lambda| > 3\pi^2/10$. To do this one needs to know that it takes 10 copies of the fundamental domain to cover $\mathcal{S}_{4/3,1/2}$ in $\mathcal{SP}_3$. From well known properties of the Siegel sets (see for example **[2]**) it is clear that the number is finite, but we need the exact number. This is related to the reduction algorithm for $\mathcal{SP}_3$ discussed in **[4]** where it was necessary to form matrices $\gamma \in \Gamma_3$ having a and c as their first column for all the a and c necessary in (D1). However, there is an infinite number of choices for these matrices. For the application to Roelcke's method it is important to select these matrices more carefully.

Since we are using $\mathcal{F}_3'$ here, (D1) gives the inequalities:

(i) $x_{12}^2 + y_1^2 \geq 1$,
(ii) $x_{13}^2 + y_1^2(x_{23}^2 + y_2^2) \geq 1$,
(iii) $(x_{12} + x_{13})^2 + y_1^2((1 + x_{23})^2 + y_2^2) \geq 1$,
(iv) $(1 + x_{12} + x_{13})^2 + y_1^2((1 + x_{23})^2 + y_2^2) \geq 1$.

The difference in the inequalities here comes from the symmetry about $x_{12} = 0$ which $\mathcal{F}_3'$ has and $\mathcal{F}_3$ has not.

To get the smallest number of copies of $\mathcal{F}_3'$ needed to cover

$\mathscr{S}_{4/3,1/2}$, i.e., the number of γ_i needed so that

$$\mathscr{S}_{4/3,1/2} \subset \bigcup_{\gamma_i} \mathscr{F}_3'[\gamma_i]$$

we need to make sure that for any γ_i, $Z \in \mathscr{F}_3'[\gamma_i]$ satisfies $|x_{ij}| \le 1/2$ for all i, j where we use the Iwasawa decomposition (2). Take any $Z \in \mathscr{S}_{4/3,1/2}$ and use the reduction algorithm to move it to $\mathscr{F}_3'$. (See **[4]** for the construction of a reduction algorithm, although we will need different matrices S_i here.) First, if (using the decomposition (1)) $Z' \notin \mathscr{F}_2'$ then $Z[S']$ with

$$S' = \begin{pmatrix} 1 & O \\ O & S \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

has $Z' \in \mathscr{F}_2'$. At the same time this switches x_1 and x_2 where ${}^T x = (x_1, x_2)$, but this is no problem since each satisfies $|x_i| \le 1/2$. Thus, we may take $Z \in \mathscr{S}_{4/3,1/2}$ to satisfy conditions (D2′) and (D3′) for $\mathscr{F}_3'$, and we need to find $\gamma \in \Gamma_3'$ so that $Z[\gamma] \in \mathscr{F}_3'$. In **[6]** it is seen that for $Z[\gamma] \in \mathscr{F}_3'$ we have for all $\gamma \in \Gamma_3'$:

$$a_1(Z[\gamma]) \le a_1(Z[\gamma][M])$$

where $a_1(Z)$ denotes the upper left entry of Z as before. If we take $M = \gamma^{-1}$ we have $(aw + {}^T xc)^2 + y^{3/2} Z'[c] \le 1$. So, we need only consider γ with a and c satisfying this last inequality. Also based on results in **[6]** it can be shown that this inequalty can never be satisfed $(\operatorname{mod} D_3)$ unless:

$$\pm \begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

So we need to find matrices S_i wth the first columns above. Experimentation showed that the right choices of S_i are:

$$S_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad S_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

$$S_3 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad S_4 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix},$$

and $S_5 = I$. What is meant by calling these the "right choices" is that there are infinitely many choices of matrices with the same first columns, but in each case S_i is the only one with that first column to not affect (D2′) and (D3′). Also, each S_i fixes the piece of the boundary of $\mathscr{F}_n'$ from inequality (i) above. Note that for each i, $S_i^2 = I$. We will say more on this shortly.

Returning to the problem at hand, since $(S')^e S_i$, $i = 1, 2, 3, 4, 5$, $e = 0$ or 1 move Z into $\mathscr{F}_3'$, the inverses $S_i(S')^e$ are what we need to cover $\mathscr{S}_{4/3,1/2}$ with images of $\mathscr{F}_3'$, i.e.,

$$\mathscr{S}_{4/3,1/2} \subset \bigcup \mathscr{F}_3'[S_i(S')^e].$$

To check that these are sufficient, recall that only matrices of the form $(S')^e S$ with S having first column from the list above would move $Z \in \mathscr{S}_{4/3,1/2}$ into $\mathscr{F}_3'$. If we were to use a different S from one of the S_i, $S = S_i g$, $g \in \mathrm{GL}(3, \mathbf{Z})$. Specifically, $S = S_i\left(\begin{smallmatrix} 1 & {}^T q \\ O & R \end{smallmatrix}\right)$, with $q \in \mathbf{Z}^2$, $R \in \mathrm{GL}(2, \mathbf{Z})$, in other words, $S \in P(1, 2)$ where P(1, 2) is the parabolic subgroup of GL(3, **Z**) consisting of matrices of this form. Now, if $Z' = Z[(S')^e S_i] \in \mathscr{F}_3'$, then $Z'[g] \in \mathscr{F}_3' \Rightarrow g = I$ as long as Z' is not on the boundary of $\mathscr{F}_3'$. It is easily checked that they are all necessary. Thus, it takes 10 images of $\mathscr{F}_3'$ to cover $\mathscr{S}_{4/3,1/2}$. This may be simplified and generalized to any n by the following:

THEOREM 2. *If M_i are the minimal set such that*

$$\mathscr{S}_{4/3,1/2} \subset \bigcup \mathscr{F}_{n-1}'[M_i] \quad \textit{in } \mathscr{SP}_{n-1},$$

then in $\mathscr{SP}_n$, $\mathscr{S}_{4/3,1/2} \subset \bigcup \mathscr{F}_n'[S_j][M_i']$ where $M_i' = \left(\begin{smallmatrix} 1 & O \\ O & M_i \end{smallmatrix}\right)$ and the S_j are those uniquely determined matrices (mod $P(1, n-1)$) *with first columns given by the necessary and sufficient a and c in condition* (D1) *of Definition* 1, *which fix the corresponding pieces of the boundary of $\mathscr{F}_n'$. These satisfy $S_j^2 = I$.*

Most of this is clear from the preceding discussion, so we prove that we must have $S_j^2 = I$. The Riemannian metric on $\mathscr{SP}_n$ is defined by:

$$ds^2 = \mathrm{Tr}((dYY^{-1})^2).$$

Let Z_0 be on the boundary fixed by S_j but not on the other portions of the boundary, and let $Z_1 \in \mathscr{F}_n'$ (not on the boundary) be within a distance ε of Z_0 where ε is small enough so that no point on any other boundary piece is within ε of Z_0. Then, since the map

$Z \to Z[g]$ is continuous in Z, $Z_1[S_j]$ must be within ε_1 of Z_0, and $Z_1[S_j] \in \mathscr{F}_n'[S_j]$. Similarly, $Z_1[S_j][S_j] = Z_1[S_j^2]$ is within ε_2 of Z_0. By adjusting ε we can make sure ε_2 is small enough so that no other boundary portion is within ε_2 of Z_0. Thus $Z_1[S_j^2] \in \mathscr{F}_n'$. This gives $S_j^2 = I$.

4. Compactification of $\mathscr{F}_n$.

In this section we consider:

THEOREM 3. *Let* $\mathscr{F}_n^* = \bigcup_P \mathscr{F}_P$ *where if* P *is a partition of* n *written* $(n_1, n_2, \dots, n_k)$

$$\mathscr{F}_{(n_1, \dots, n_k)} = \mathscr{F}_{n_1} \times \cdots \times \mathscr{F}_{n_k}.$$

Then, with the topology defined below, $\mathscr{F}_n^*$ *is a Satake compactification of* $\mathscr{F}_n$, *i.e.* $\mathscr{F}_n^*$ *is a compact Hausdorff space whose topology is induced by the closure of* $\mathscr{F}_n$ *and this topology induces a compactification of* $\mathscr{SP}_n/\Gamma_n$.

Here $\mathscr{F}_1$ is just $\{\infty\}$, a single point, as in the compactification of $\mathscr{F}_2$, $\mathscr{F}_2^* = \mathscr{F}_2 \cup \{\infty\}$.

Recall that $a_1(Z)$ is defined to be the upper left entry of the matrix Z. Then for $(Z_r, Z_s) \in \mathscr{F}_r \times \mathscr{F}_s$ with $n = r+s$ and U a neighborhood of (Z_r, Z_s) in $\mathscr{F}_r \times \mathscr{F}_s$, define a neighborhood of (Z_r, Z_s) in $\mathscr{F}_n^*$ as follows: write Z_r and Z_s as

$$Z_r = a_1(Z_r) \begin{pmatrix} 1 & & & & \\ & y_1^2 & & & \\ & & (y_1 y_2)^2 & & \\ & & & \ddots & \\ & & & & (y_1 y_2 \cdots y_{r-1})^2 \end{pmatrix} \begin{bmatrix} 1 & & x_{ij} \\ & \ddots & \\ & & 1 \end{bmatrix}$$

$$Z_s = a_1(Z_s) \begin{pmatrix} 1 & & & & \\ & y_{r+1}^2 & & & \\ & & (y_{r+1} y_{r+2})^2 & & \\ & & & \ddots & \\ & & & & (y_{r+1} \cdots y_{n-1})^2 \end{pmatrix} \begin{bmatrix} 1 & & x_{ij} \\ & \ddots & \\ & & 1 \end{bmatrix}.$$

Then define y by

$$y^{n/2} = \prod_{j=1}^{n-1} y_j^{n-j}$$

and for $M > 0$ let

$$V(U, M) = \Big\{ Z \in \mathcal{SP}_n \,|\, Z = y^{-1} \begin{pmatrix} a_1(Z_r)^{-1}Z_r & O \\ O & a_1(Z_s)^{-1}(y_1 \dots y_r)^2 Z_s \end{pmatrix} \begin{bmatrix} I_r & X \\ O & I_s \end{bmatrix}$$
$$\text{with } y_r > M\,,\ Z_r \times Z_s \in U\,,$$
$$X = (x_{ij}) \text{ with } |x_{ij}| \leq 1/2\,,\ x_{r1} \geq 0 \Big\}.$$

If r or s is 1, we must take Z_r or Z_s to be 1. Now define a neighborhood of (Z_r, Z_s) in $\mathcal{F}_n^*$ by $V^*(U, M) = V(U, M) \cup U$. This definition can then be extended to define a neighborhood in $\mathcal{F}_n^*$ of $(Z_1, \dots, Z_k) \in \mathcal{F}_{n_1} \times \dots \times \mathcal{F}_{n_k}$. This sort of thing is usually called the Satake topology.

We need now show that for M sufficiently large, $V(U, M) \subset \mathcal{F}_n$. To prove condition (D1),

$$yZ \begin{bmatrix} a \\ c \end{bmatrix} = (a + {}^T xc)^2 + Z_1[c] \geq 1\,,$$

note that since Z is a quadratic form, then if $c_j = 0$ for $j \geq r$,

$$Z \begin{bmatrix} a \\ c \end{bmatrix} = Z_r[m]$$

where $m \in \mathbf{Z}^{n-r}$ and ${}^T m = (a, c_1, \dots, c_{r-1})$. But $Z_r \in \mathcal{F}_r$ so $yZ_r[m] \geq 1$. If $c_j \neq 0$ for some $j \geq r$, then we may apply Lemma 2. We have:

$$(a + {}^T xc)^2 + Z_1[c] \geq Z_1[c] \geq y_1^2 \cdots y_r^2 \geq (3/4)^{r-1} M^2.$$

Thus, if we choose $M > (4/3)^{(r-1)/2}$ we have $(a + {}^T xc)^2 + Z_1[c] \geq 1$. Now we must show that $Z' \in \mathcal{F}_{n-1}$. Since $y'Z' = y_1^{-2} Z_1$ and so on, this is equivalent to showing that $y_i^{-2} Z_i[c] \geq 1$ for all $0 \neq c \in \mathbf{Z}^{n-i}$ and $i = 1, \dots, n-1$. Since

$$y_i^{-2} Z_i[c] \geq Z_{i+1}[c']$$

this follows immediately from Lemma 2 for $i = 1, \dots, r-1$ using the same arguments as above. For $i = r$, $Z' \in \mathcal{F}_{n-1}$ follows from Lemma 1, and from $Z_s \in \mathcal{F}_s$ for $i > r$. Condition (D3) clearly holds by the definition of $V(U, M)$.

To show compactness of $\mathcal{F}_n^*$ it will suffice to show that any sequence $\{Z^{(\nu)}\}$ in $\mathcal{F}_n$ has a limit point in $\mathcal{F}_n^*$. Write the Iwasawa decomposition (2) of $Z^{(\nu)}$ with $y^{(\nu)}$, $y_i^{(\nu)}$ in place of the y, y_i. If all the sequences $\{y_i^{(\nu)}\}$ are bounded, then $Z^{(\nu)} \to Z \in \mathscr{F}_n$. If $\{y_r^{(\nu)}\}$ for some r is not bounded, let a_r be defined by

$$a_r^{r/2} = \prod_{j=1}^{r-1} (y_j^{(\nu)})^{n-j}$$

and write

$$Z^{(\nu)} = \begin{pmatrix} (y^{(\nu)})^{-1} a_r Z_r^{(\nu)} & O \\ O & (y^{(\nu)} a_r^{-1})^{n_1/n_2} Z_{n-r}^{(\nu)} \end{pmatrix} \begin{bmatrix} I_r & X^{(\nu)} \\ O & I_s \end{bmatrix}.$$

Then $Z_i^{(\nu)} \in \mathscr{F}_i$ for $i = r$ or $n - r$ and if no other $\{y_j^{(\nu)}\}$ is unbounded, $Z_i^{(\nu)}$ converges to some $Z_i \in \mathscr{F}_i$. If another $\{y_j^{(\nu)}\}$ is unbounded we can repeat for $Z_i^{(\nu)}$ for whichever i, r or $n - r$. Thus, we can establish the desired result inductively.

For some small n we can write out $\mathscr{F}_n^*$:

$$\mathscr{F}_3^* = \mathscr{F}_3 \cup \mathscr{F}_2 \cup \mathscr{F}_2 \cup \mathscr{F}_1$$

where we write $\mathscr{F}_2$ instead of $\mathscr{F}_2 \times \mathscr{F}_1$ since $\mathscr{F}_1$ is just one point. Note that this holds up very well with what we saw in §2. We also have

$$\mathscr{F}_4^* = \mathscr{F}_4 \cup \mathscr{F}_3 \cup \mathscr{F}_2 \times \mathscr{F}_2 \cup \mathscr{F}_3 \cup \mathscr{F}_2 \cup \mathscr{F}_2 \cup \mathscr{F}_2 \cup \mathscr{F}_1.$$

We can think of the first $\mathscr{F}_2$ as being a subspace of the first $\mathscr{F}_3$ as well as $\mathscr{F}_2 \times \mathscr{F}_2$ in the Satake topology while the second $\mathscr{F}_2$ would be a subspace of each $\mathscr{F}_3$. This follows the containments of the appropriate parabolic subgroups..

If π_n is the canonical projection of $\mathscr{SP}_n$ onto $\mathscr{SP}_n/\Gamma_n$, then π_n identifies $\mathscr{SP}_n/\Gamma_n$ with $\mathscr{F}_n$. Using this, and letting $V_n = \mathscr{SP}_n/\Gamma_n$, we get the Satake compactification $V_n^* = \bigcup_P V_P$ where if P is the partition of n into $(n_1, n_2, \dots, n_k)$ or the parabolic subgroup $P(n_1, n_2, \dots, n_k)$, then $V_P = V_{n_1} \times \cdots \times V_{n_k}$. Satake **[13]** actually defines several different compactifications of a quotient space (see also **[16]**); the one we have investigated here is the maximal Satake compactification. In **[6]** the compactification considered was not this maximal one. If we go all the way back to the beginning and restrict ourselves to the partial Iwasawa decompositions (1) we would have $\mathscr{F}_n^*$ defined recursively by:

$$\mathscr{F}_n^* = \mathscr{F}_n \cup \mathscr{F}_{n-1}^* = \mathscr{F}_n \cup \mathscr{F}_{n-1} \cup \cdots \cup \mathscr{F}_2 \cup \mathscr{F}_1$$

with the topology defined accordingly by taking a neighborhood of $Z' \in \mathscr{F}_{n-1}^*$, U, and defining its neighborhood in $\mathscr{F}_n^*$ to be:

$$\left\{ Z \in \mathscr{SP}_n | Z = \begin{pmatrix} y^{-1} & O \\ O & y^{1/(n-1)}Z' \end{pmatrix} \begin{bmatrix} 1 & {}^T x \\ O & I_{n-1} \end{bmatrix} \right.$$

$$\text{with } y > M,\ Z' \in U,\ {}^T x = (x_1, \ldots, x_{n-1})$$

$$\left. \text{with } |x_j| \leq 1/2,\ x_1 \geq 0 \right\}.$$

Thus we have:

THEOREM 4. *$\mathscr{F}_n^* = \mathscr{F}_n \cup \mathscr{F}_{n-1} \cup \cdots \cup \mathscr{F}_2 \cup \mathscr{F}_1$ is a Satake compactification of $\mathscr{F}_n$. If $V_n = \mathscr{SP}_n/\Gamma_n$ then $V_n^* = V_n \cup V_{n-1} \cup \cdots \cup V_1 \cup V_0 = V_n \cup V_{n-1}^*$ is a Satake compactification of V_n.*

This last compactification is analogous to Satake's original compactification of Siegel's quotient space **[12]**.

If Γ' is any arithmetic subgroup of G, we can similarly obtain the compactifications of $V' = \mathscr{SP}_n/\Gamma'$. If $\mathscr{F}_n'$ is the fundamental domain corresponding to $\mathscr{SP}_n/\Gamma'$ and $(\mathscr{F}_n')^*$ its compactification, Satake **[13]** shows that $\bigcup_{\gamma' \in \Gamma'} (\mathscr{F}_n')^*[\gamma'] = \bigcup_{\gamma \in \Gamma} \mathscr{F}_n^*[\gamma]$, and that the topologies on these spaces are the same.

REFERENCES

[1] W. Baily and A. Borel, *Compactification of arithmetic quotients of bounded symmetric domains*, Ann. of Math., **84** (1966), 442–528.

[2] A. Borel, *Introduction aux Groupes Arithmetiques*, Hermann, Paris, 1969.

[3] J. Cassels, *Rational Quadratic Forms*, Academic Press, London, 1978.

[4] D. Gordon, D. Grenier, and A. Terras, *Hecke operators and the fundamental domain for* SL(3, **Z**), Math. Comp., **48** (1987), 159–178.

[5] D. Grenier, *Fundamental Domains for* $\mathscr{P}_n$/ GL(n, **Z**) *and Applications in Number Theory*, Dissertation, UCSD, 1986.

[6] ——, *Fundamental domains for the general linear group*, Pacific J. Math., **132** (1988), 293–317.

[7] ——, *An analogue of Siegel's ϕ-operator for automorphic forms for* GL(n, **Z**), Trans. Amer. Math. Soc., to appear.

[8] C. Hermite, *Oeuvres* I, Gauthier-Villars, Paris, 1905.

[9] J. Huntley, *Multiplicity One Theorems for Automorphic Forms*, Ph.D. Thesis, Stanford University, 1987.

[10] J. Huntley, correspondence.

[11] W. Roelcke, *Über die Wellengleichung bei Grenzkreisgruppen erster Art*, Sitzber. Akad. Heidelberg, Math-naturwiss., 1953/55.

[12] I. Satake, *On the compactification of the Siegel space*, J. Indian Math. Soc., **20** (1956), 259–281.

[13] ——, *On compactifications of the quotient spaces for arithmetically defined discontinuous groups*, Ann. of Math., **72** (1960), 555–580.
[14] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications* II, Springer-Verlag, New York, 1988.
[15] R. Zimmer, *Ergodic Theory and Semisimple Groups*, Birkhäuser, Boston, 1984.
[16] S. Zucker, *L^p-cohomology and Satake compactifications*, preprint.

Received May 20, 1991 and in revised form August 21, 1991.

THE JOHNS HOPKINS UNIVERSITY
BALTIMORE, MD 21218

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

FIXED POINTS OF SURFACE DIFFEOMORPHISMS

BOJU JIANG AND JIANHAN GUO

We give a complete proof of the following theorem which was conjectured by Jakob Nielsen for closed oriented surfaces.

THEOREM. ***Let*** $f: M \to M$ ***be a homeomorphism of a compact surface. When*** M ***is closed, then*** f ***is isotopic to a diffeomorphism with*** $N(f)$ ***fixed points, where*** $N(f)$ ***is its Nielsen number. When*** M ***has boundary,*** $N(f)$ ***should be replaced by the relative Nielsen number*** $N(f; M, \partial M)$ ***defined by Schirmer.***

Another result is the inequality $|L(f) - \chi(M)| \leq N(f) - \chi(M)$ **when** $\chi(M) < 0$**, where** $L(f)$ **is the Lefschetz number and** $\chi(M)$ **is the Euler characteristic.**

Introduction. For a self-map f of a compact polyhedron X, the Nielsen number $N(f)$ is defined to be the number of essential fixed point classes. (See [**J3**] for an introduction to the Nielsen fixed point theory.) It is a classical theorem of Wecken [**W**] that $N(f)$ is a lower bound of the number of fixed points for all maps homotopic to f, and that if X is a manifold of dimension ≥ 3, this lower bound is always realizable (see also [**Br**], [**K**]). It is now known [**J4**] that when X is a surface with negative Euler characteristic, there exists a map $f: X \to X$ such that every map homotopic to f has more than $N(f)$ fixed points. The purpose of this paper is to show that for homeomorphisms of surfaces the Nielsen number is indeed the least number of fixed points in the isotopy class, as Nielsen himself conjectured (cf. [**N2**, §31]) in his study of oriented closed surfaces.

MAIN THEOREM. *Let* M *be a compact surface, closed or with boundary. Let* $f: M \to M$ *be a homeomorphism. Then* f *is isotopic to a smooth embedding which has* $N(f)$ *fixed points. If, in addition, no boundary component of* M *is mapped onto itself by* f *in an orientation-reversing manner, then* f *is isotopic to a diffeomorphism having* $N(f)$ *fixed points.*

This theorem was announced in [**J2**], here strengthened with smoothness considerations. An example in [**J1**] shows it is necessary to allow embeddings in order to get as few as $N(f)$ fixed points when there are orientation-reversing invariant boundary components. If we insist on diffeomorphisms, the least number of fixed points in the isotopy class turns out to be the relative Nielsen number $N(f; M, \partial M)$ defined by Schirmer [S]. See Theorem 5.1 below. Another result is an interesting inequality (Theorem 4.1) relating the Nielsen number with the Lefschetz number.

All the recent progress on Nielsen's fixed point conjecture originated in Thurston's theory of surface diffeomorphisms [T]. Thurston himself solved the important case of orientation-preserving pseudo-Anosov maps in Theorem 6 of [T]. See [**BK**] for the first published proof. The easy periodic case was treated in [**J1**]. The work [**I**] considered orientation-preserving maps of closed orientable surfaces and, besides pseudo-Anosov and periodic cases, hinted at the kind of analysis needed for reducible maps. In the present paper, we shall consider general diffeomorphisms of compact surfaces, and give complete proofs.

There are seven connected compact surfaces with positive or zero Euler characteristic. For them the truth of the Main Theorem can be checked case by case. The 2-sphere, the 2-disk and the annulus have respectively 2, 2 and 4 isotopy classes of self-homeomorphisms. The real projective plane has only one isotopy class, the Möbius band has two (see [**E**, Theorems 5.5 and 5.8]). All these isotopy classes have obvious simple representatives satisfying the requirements of the Main Theorem. The torus and the Klein bottle are more interesting. They are studied in Nielsen's classical paper [**N1**]. Both have the Euclidean plane E^2 as the universal covering space with isometric covering translations. Each isotopy class contains a linear representative, i.e. one that lifts to a linear map of E^2. Such a representative, slightly perturbed if necessary to remove inessential fixed point classes, minimizes the number of fixed points.

Henceforth we assume that M *is a compact surface such that each of its connected components has negative Euler characteristic.* Our proof of the Main Theorem is based on Thurston's classification of surface homeomorphisms.

THURSTON THEOREM ([T]). *Every homeomorphism* $f : M \to M$ *is isotopic to a "diffeomorphism"* φ *such that either*

(1) φ *is an isometry with respect to some hyperbolic metric on* M, *or equivalently,* φ *is a periodic map, i.e.* $\varphi^m = \mathrm{id}$; *or*

(2) φ *is a pseudo-Anosov map, i.e. there is a number* $\lambda > 1$ *and a pair of transverse measured foliations* $(\mathfrak{F}^s, \mu^s)$ *and* $(\mathfrak{F}^u, \mu^u)$ *such that* $\varphi(\mathfrak{F}^s, \mu^s) = (\mathfrak{F}^s, \frac{1}{\lambda}\mu^s)$ *and* $\varphi(\mathfrak{F}^u, \mu^u) = (\mathfrak{F}^u, \lambda\mu^u)$; *or*

(3) φ *is a reducible map, i.e. there is a system of disjoint simple closed curves* $\Gamma = \{\Gamma_1, \dots, \Gamma_n\}$ *in* $\operatorname{int} M$ *such that* Γ *is invariant by* φ (*but the* Γ_i*'s may be permuted*) *and* Γ *has a* φ*-invariant tubular neighborhood* $\mathcal{N}(\Gamma)$ *such that each component of* $M - \mathcal{N}(\Gamma)$ *has negative Euler characteristic and on each* (*not necessarily connected*) φ*-component of* $M - \mathcal{N}(\Gamma)$, φ *satisfies* (1) *or* (2).

REMARK. The pseudo-Anosov map in [T] is not an honest diffeomorphism in that it is not even C^1 at the singularities of $\mathfrak{F}^s$ and $\mathfrak{F}^u$. In the statement given in [T], Γ_i are two-sided simple closed curves on M, so some components of $M - \mathcal{N}(\Gamma)$ may be Möbius bands. We prefer to allow one-sided Γ_i in order to guarantee that every component of $M - \mathcal{N}(\Gamma)$ has negative Euler characteristic.

The proof of this theorem can be found in **[B]**, **[FLP]**, **[HT]** for the oriented surfaces, and in **[Wu]** for non-orientable surfaces.

The structure of the paper is as follows. In §§1–2 we develop standard forms for φ on the periodic pieces and the pseudo-Anosov pieces respectively. These are not meant to be representatives of the isotopy classes that have the minimal number of fixed points, rather they are designed to be building blocks that are ready to be glued together. In §3 these models are assembled into a standard form for a general φ, where we shall regard the case (3) in Thurston Theorem as the general case by regarding the first two cases as "reducible" with empty reducing curves ($\Gamma = \varnothing$). The rest of §3 is devoted to the detailed analysis of the fixed point classes of this standard form. This is a technical stepping-stone of the paper. The Main Theorem is then proved in §4 by shrinking these fixed point classes to single points via isotopy. The inequality relating the Nielsen number and the Lefschetz number is also proved there. A discussion on the relative Nielsen numbers is given in §5.

There is another obvious division of the problem into three cases:

(1) orientation preserving homeomorphisms on oriented surfaces,
(2) orientation reversing homeomorphisms on oriented surfaces,
(3) homeomorphisms on nonorientable surfaces.

Although we have chosen Thurston's structural trichotomy as the organizing principle of the paper, the above division is certainly the natural order of understanding. The orientation preserving and reversing cases involve complementary types of fixed point classes, while the nonorientable case combines both. So we suggest the reader to first focus on the orientation preserving case which contains the main ideas with only half of the technicalities, then the orientation reversing case and finally the nonorientable case. For the convenience of the reader, we systematically use the superscripts $+/-$ in the labels to indicate the relevance to orientation preserving/reversing cases.

The following notation and terminology will be used throughout the paper.

NOTATION. Let $f\colon X \to X$ be a map. Then $\operatorname{Fix} f$ denotes the fixed point set $\{x \in X \mid x = f(x)\}$. When X is a polyhedron and $A \subset X$ is such that $A \cap \operatorname{Fix} f$ is both open and closed in $\operatorname{Fix} f$, then $\operatorname{index}(f, A)$ denotes the fixed point index of $A \cap \operatorname{Fix} f$.

DEFINITION. Let $f\colon X \to X$ be a map. A subset $A \subset X$ is said to be *f-invariant* if $f(A) \subset A$. Two path-connected f-invariant subsets A_0, A_1 are said to be *f-related* if there is a path $c\colon I \to X$ such that $c \simeq f \circ c\colon I, 0, 1 \to X, A_0, A_1$.

In general this is not an equivalence relation among f-invariant subsets. When both A_0 and A_1 are single points, it reduces to the Nielsen equivalence relation between fixed points of f. Namely, two fixed points x and y are in the same fixed point class if there is a path $c\colon I \to X$ connecting them such that $c \simeq f \circ c$ rel endpoints.

1. The periodic case. Suppose $\varphi\colon M \to M$ is periodic, i.e. $\varphi^m = \mathrm{id}$ for some natural number m. It is well known that such a φ is an isometry with respect to some hyperbolic metric on M (of constant Gaussian curvature -1 and with totally geodesic boundary).

In this section, the connected components of $\operatorname{Fix} \varphi$ are explicitly described, and the fixed point classes of φ are identified with these components.

LEMMA 1.1. *Let A be a component of* $\operatorname{Fix} \varphi$. *Then one of the following is true:*

$(1)^+$ *A is a component of M and* $\operatorname{index}(\varphi, A) = \chi(A) < 0$.

$(2)^-$ *A is a closed geodesic, with a neighborhood diffeomorphic to $S^1 \times (0, 1)$ where φ acts as the reflection $(z, t) \mapsto (z, 1 - t)$;* $\operatorname{index}(\varphi, A) = 0$.

$(3)^-$ *A is a geodesic arc orthogonally connecting two (not necessarily distinct) components of ∂M, with a neighborhood diffeomor-*

phic to $I \times (0, 1)$ *where* φ *acts as the reflection* $(s, t) \mapsto (s, 1-t)$; index$(\varphi, A) = 1$.

$(4)^+$ A *is a point, with a neighborhood diffeomorphic to the open unit disc* $\operatorname{int} D = \{ z \in \mathbb{C} \mid |z| < 1 \}$ *where* φ *acts as a rotation*; index$(\varphi, A) = 1$.

Proof. It is clear that each component of the fixed point set of an isometry is a properly embedded totally geodesic submanifold. Hence the four possibilities. The index is easily calculated from the local description of φ. □

LEMMA 1.2. *Let* A_0, A_1 *be either a fixed point of* φ *or a* φ*-invariant component of* ∂M. *Suppose* A_0 *and* A_1 *are* φ*-related via a path* $c\colon I, 0, 1 \to M, A_0, A_1$. *Then there is a path* γ *in* $\operatorname{Fix}\varphi$ *such that* $\gamma \simeq c\colon I, 0, 1 \to M, A_0, A_1$.

Proof. Hyperbolic geometry guarantees a unique shortest geodesic in every homotopy class of paths $I, 0, 1 \to M, A_0, A_1$. (Recall that ∂M is totally geodesic.) Let γ be the shortest geodesic homotopic to c. Since φ is isometric, $\varphi \circ \gamma$ is the shortest geodesic homotopic to $\varphi \circ c$. But $c \simeq \varphi \circ c$ so $\gamma = \varphi \circ \gamma$. This means γ is in $\operatorname{Fix}\varphi$. □

COROLLARY 1.3. *Fixed point classes of* φ *are connected.* □

2. The pseudo-Anosov case. Suppose $\varphi\colon M \to M$ is a pseudo-Anosov map with stable and unstable measured foliations $(\mathfrak{F}^s, \mu^s)$ and $(\mathfrak{F}^u, \mu^u)$ respectively, and with expansion constant $\lambda > 1$.

The classical model for φ is described in §2.1. It is isotoped in §2.2 to a standard form $\bar{\varphi}$ which is smooth and moreover, in view of our later need in §3 of gluing periodic pieces and pseudo-Anosov pieces together, is required to be periodic on ∂M. In §2.3, the fixed point classes of $\bar{\varphi}$ are identified as the connected components of $\operatorname{Fix}\bar{\varphi}$.

2.1. *The classical model.* The following model is adapted from the description of Thurston's pseudo-Anosov map φ for closed surfaces (cf. [**GK**, pp. 176, 182]). The strategy to describe φ near ∂M is as follows: Collapsing each component of ∂M into a single point (called a puncture) we obtain a generalized pseudo-Anosov map $\hat{\varphi} : \widehat{M} \to \widehat{M}$ of a closed surface $\widehat{M}$ (cf. [**FLP**, pp.217,243]). The model for pseudo-Anosov maps works for generalized pseudo-Anosov maps as well by allowing the prong number $p \geq 1$. We then recover M and φ by "blowing up" the punctures of $\widehat{M}$.

We first introduce some notations. Let us write $\ell = \ln \lambda > 0$. Let p be a natural number.

Definition. On the complex plane $\mathbb{C} = \{ s = s_1 + is_2 \mid s_1, s_2 \in \mathbb{R} \}$, define the linear map

$$F\colon \mathbb{C} \to \mathbb{C}, \quad s_1 + is_2 \mapsto \lambda s_1 + is_2/\lambda.$$

It is the time-one map for the vector field V defined by $\dot{s}_1 = ls_1$, $\dot{s}_2 = -ls_2$, or

$$V(s) = l\bar{s}.$$

Definition. On the complex plane $\mathbb{C} = \{ z = \rho e^{i\theta} \mid \rho \geq 0,\ \theta \in \mathbb{R} \}$, consider the (multi-valued) map

$$\Phi = \Phi_p\colon \mathbb{C} \to \mathbb{C}, \quad z \mapsto z^{p/2}$$

and its inverse

$$\Phi^{-1} = \Phi_p^{-1}\colon \mathbb{C} \to \mathbb{C}, \quad s \mapsto s^{2/p}.$$

Consider the map $\Psi\colon \mathbb{C} - \operatorname{int} D, S^1 \to \mathbb{C}, 0$ defined by $\Psi(z) = z - z/|z|$. It restricts to a diffeomorphism $\Psi\colon \mathbb{C} - D \to \mathbb{C} - \{0\}$ with inverse

$$\Psi^{-1}\colon \mathbb{C} - \{0\} \to \mathbb{C} - D, \quad z \mapsto z + \frac{z}{|z|}.$$

Definition. Let $v = v_p$ be the vector field $\Phi_*^{-1} V$. A simple calculation gives

$$v(z) = \frac{2l}{p} z \left(\frac{z}{|z|}\right)^{-p} \quad \text{or} \quad v\left(\rho e^{i\theta}\right) = \frac{2l}{p} \rho e^{i(1-p)\theta}.$$

Let $f = f_p$ be the time-one map for $v = v_p$.

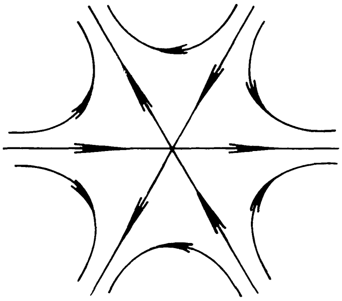

Figure 1. The flow of v and $\bar{v}$ ($p = 3$)

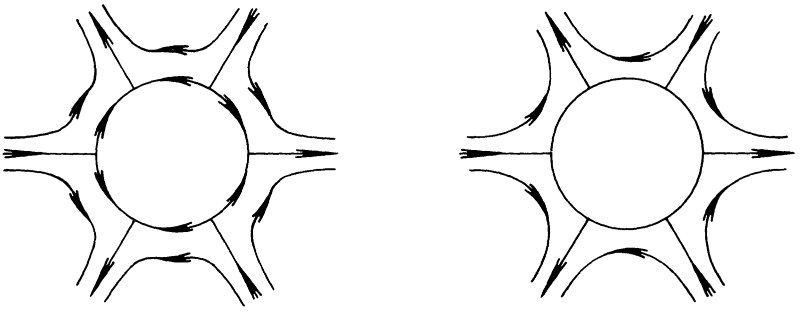

FIGURE 2. The flow of v' and $\bar{v}'$ $(p = 3)$

Define the vector field v' on $\mathbb{C} - D$ by $v' = \Psi_*^{-1} v$. The formula for v' is easily obtained:

$$v'(\rho e^{i\theta}) = \frac{2l}{p}\{(\rho - 1)e^{i(1-p)\theta} + e^{i(\theta-\pi/2)}\sin p\theta\},$$

so v' is smoothly extendable to $\mathbb{C} - \operatorname{int} D$ with $v'|_{\partial D}$ tangent to ∂D. Let f' be the time-one map for v'. We have a commutative diagram

$$\begin{tikzcd}
\mathbb{C} \arrow[d, "F"'] & \mathbb{C} \arrow[l, "\Phi"'] \arrow[d, "f"] & \mathbb{C} - \operatorname{int} D \arrow[l, "\Psi"'] \arrow[d, "f'"] \\
\mathbb{C} & \mathbb{C} \arrow[l, "\Phi"'] & \mathbb{C} - \operatorname{int} D \arrow[l, "\Psi"']
\end{tikzcd}$$

DEFINITION. Let $r = r_{(p,k)^\pm} \colon \mathbb{C} \to \mathbb{C}$ be the map

$$r(\rho e^{i\theta}) = \rho e^{\pm i(\theta + 2k\pi/p)}$$

where k is an integer, $0 \le k < p$. It is simply a rotation or a reflection according as the plus sign or the minus sign is chosen.

The atlas. A chart at a point $x \in \operatorname{int} M$ is an open neighborhood U_x and an embedding $u_x \colon U_x, x \to \mathbb{C}, 0$. A chart at a component A of ∂M is an open neighborhood U_A and an embedding $u_A \colon U_A, A \to \mathbb{C} - \operatorname{int} D, S^1$.

There is a finite smooth atlas $\mathcal{U}$ of M, consisting of one chart for each interior singularity, one chart for each boundary component, and some other charts at interior regular points, such that:

(1) The measures μ^s and μ^u on U_x are mapped by u_x to the measures $|\operatorname{Re} dz^{p/2}| = |\operatorname{Re} d\Phi_p(z)|$ and $|\operatorname{Im} dz^{p/2}| = |\operatorname{Im} d\Phi_p(z)|$ on $\mathbb{C}$ respectively, where $p = p_x \ge 3$ is the number of prongs of $\mathfrak{F}$ at

a singular point x, or $p_x = 2$ if x is a regular point; the leaves of $\mathfrak{F}^s$ and $\mathfrak{F}^u$ get mapped to the lines $\{\operatorname{Re}\Phi_p(z) = \text{constant}\}$ and $\{\operatorname{Im}\Phi_p(z) = \text{constant}\}$ respectively. (The prongs of $\mathfrak{F}^s$ are $\{z = \rho e^{i\theta} \mid \rho \geq 0,\ \theta = \frac{\text{odd}}{p}\pi\}$ and the prongs of $\mathfrak{F}^u$ are $\{z = \rho e^{i\theta} \mid \rho \geq 0,\ \theta = \frac{\text{even}}{p}\pi\}$.)

(2) The measures μ^s and μ^u on U_A are mapped by u_A to the measures $|\operatorname{Re} d\Phi_p \circ \Psi(z)|$ and $|\operatorname{Im} d\Phi_p \circ \Psi(z)|$ on $\mathbb{C} - \operatorname{int} D$ respectively, where $p = p_A \geq 1$ is the number of prongs of $\mathfrak{F}$ at A; the leaves of $\mathfrak{F}^s$ and $\mathfrak{F}^u$ get mapped to the lines $\{\operatorname{Re}\Phi_p \circ \Psi(z) = \text{constant}\}$ and $\{\operatorname{Im}\Phi_p \circ \Psi(z) = \text{constant}\}$ respectively.

We may further assume (change scale if necessary) that, in each chart at an interior singular point x, the closed unit disk D is contained in $u_x(U_x)$ and is disjoint from chart overlaps; similarly for the closed annulus $\{z \mid 1 \leq |z| \leq 2\}$ in each chart U_A.

The model for φ. In such an atlas $\mathscr{U}$, denoting $U'_x = U_x \cap \varphi^{-1} U_{\varphi(x)}$ and $U'_A = U_A \cap \varphi^{-1} U_{\varphi(A)}$, we have commutative diagrams

$$\begin{tikzcd}
U'_x, x \arrow[r, "\varphi"] \arrow[d, "u_x"'] & U_{\varphi(x)}, \varphi(x) \arrow[d, "u_{\varphi(x)}"] & U'_A, A \arrow[r, "\varphi"] \arrow[d, "u_A"'] & U_{\varphi(A)}, \varphi(A) \arrow[d, "u_{\varphi(A)}"] \\
\mathbb{C}, 0 \arrow[r, "r \circ f"] & \mathbb{C}, 0 & \mathbb{C} - \operatorname{int} D, S^1 \arrow[r, "r \circ f'"] & \mathbb{C} - \operatorname{int} D, S^1
\end{tikzcd}$$

in charts at interior singularities and boundary components, where $f = f_p$, $f = f'_p$ and $r = r_{(p,k)^\pm}$ for suitable p, k.

If such a singular point x is a fixed point of φ, it will be called a fixed point of type $(p, k)^\pm$. Similarly, if this A is φ-invariant, it is called a φ-invariant boundary component of type $(p, k)^\pm$. Here the sign $+/-$ indicates $\varphi|A : A \to A$ is orientation preserving or reversing.

Suppose $x \in \operatorname{int} M$ is a regular point and $\varphi(x) = x$. Then there exist local charts at x and $\varphi(x)$ (not necessarily in the atlas $\mathscr{U}$) with the above commutative diagram where p is taken to be 2. In this sense a regular point can be regarded as a "2-prong singularity", so that our discussion of singularities applies to regular points as well.

REMARK. In view of the rotation symmetry of v, v', f and f', for odd p every type $(p, k)^-$ is conjugate to, hence regarded the same as, the type $(p, 0)^-$; for even p, among the types $(p, k)^-$ there are two essentially different types: $(p, 0)^-$ and $(p, 1)^-$.

Note that in contrast to the non-smoothness of φ at an interior p-prong singularity ($p > 2$), this model of φ can be smoothly extended to the boundary ∂M. However, it fails to meet our later need in §3 that φ be periodic on ∂M.

2.2. *The standard form.* We shall modify the previous model to achieve two goals: to make φ smooth and to make $\varphi|\partial M$ periodic.

DEFINITION. Let $\alpha\colon \mathbb{R} \to I$ be a non-decreasing smooth function such that $\alpha(0) = 0$, $\alpha(1/\lambda^2) = 1$, and $\alpha(t) \geq 0$ if $t \geq 0$. Let $\overline{V}$ be the smooth vector field on $\mathbb{C}$ obtained by "slowing down" the vector field V, $\overline{V}(s) = \alpha(|s|)V(s)$. Then $\overline{V}$ and its differentials of any order vanish at $0 \in \mathbb{C}$. Let $\overline{F}\colon \mathbb{C} \to \mathbb{C}$ be the time-one map for $\overline{V}$. It is clear that $\overline{F}$ is a diffeomorphism and $\overline{F}(s) = F(s)$ when $|s| \geq 1$.

DEFINITION. Let $\bar{f} : \mathbb{C} - \{0\} \to \mathbb{C} - \{0\}$ be the time-one map for the vector field $\bar{v} := \Phi_*^{-1}\overline{V}$ on $\mathbb{C}-\{0\}$. Let $\bar{f}'\colon \mathbb{C}-D \to \mathbb{C}-D$ be the time-one map for the vector field $\bar{v}' := \Psi_*^{-1}\bar{v}$ on $\mathbb{C} - D$. It follows from the smoothness of $\overline{V}$ and its flatness at 0 that we can make $\bar{v}$ smooth on $\mathbb{C}$ by defining $\bar{v}(0) = 0$, and extend $\bar{v}'$ smoothly to $S^1 = \partial D$ by defining $\bar{v}'(z) = 0$ there. The diffeomorphisms $\bar{f}$ and $\bar{f}'$ are then extended accordingly. We have a commutative diagram

$$\begin{tikzcd}
\mathbb{C} \arrow[d, "\overline{F}"'] & \mathbb{C} \arrow[l, "\Phi"'] \arrow[d, "\bar{f}"] & \mathbb{C} - \operatorname{int} D \arrow[l, "\Psi"'] \arrow[d, "\bar{f}'"] \\
\mathbb{C} & \mathbb{C} \arrow[l, "\Phi"'] & \mathbb{C} - \operatorname{int} D \arrow[l, "\Psi"']
\end{tikzcd}$$

It is clear that $\bar{f} = f$ when $|z| \geq 1$, $\bar{f}' = f'$ when $|z| \geq 2$, and $\bar{f}'$ has the same jets (of any order) at S^1 as the identity map.

DEFINITION. In the atlas $\mathscr{U}$, let $\bar{\varphi}\colon M \to M$ be defined by the commutativity of the diagrams

$$\begin{tikzcd}
U'_x, x \arrow[r, "\bar{\varphi}"] \arrow[d, "u_x"'] & U_{\varphi(x)}, \varphi(x) \arrow[d, "u_{\varphi(x)}"] & U'_A, A \arrow[r, "\bar{\varphi}"] \arrow[d, "u_A"'] & U_{\varphi(A)}, \varphi(A) \arrow[d, "u_{\varphi(A)}"] \\
\mathbb{C}, 0 \arrow[r, "r\circ\bar{f}"] & \mathbb{C}, 0 & \mathbb{C} - \operatorname{int} D, S^1 \arrow[r, "r\circ\bar{f}'"] & \mathbb{C} - \operatorname{int} D, S^1
\end{tikzcd}$$

in the charts at interior singularities and boundary components, and be the same as φ elsewhere. Clearly, it is well-defined and smooth. Since $\bar{v}$ and $\bar{v}'$ are obtained from v and v' by "slowing down", $\bar{\varphi}$ is isotopic to φ. This isotopy does not change the interior fixed points and the invariant boundary components, but $\bar{\varphi}|\partial M$ becomes a periodic map. We shall call $\bar{\varphi}$ the *standard form* for the pseudo-Anosov map φ.

REMARK 1. We do not claim that $\bar{\varphi}$ is topologically conjugate to φ on $\operatorname{int} M$. Our concern here is only the isotopy, not the more difficult problem of conjugacy treated in **[GK]**.

For the standard form we have:

LEMMA 2.1. *Every interior fixed point of* $\bar{\varphi}$ *is isolated. Its index depends on its type*:

Type of x	$(p, 0)^+$	$(p, k)^+, p \nmid k$	$(p_{\text{even}}, 0)^-$	$(p_{\text{even}}, 1)^-$	$(p_{\text{odd}}, 0)^-$
$\operatorname{index}(\bar{\varphi}, x)$	$1 - p$	1	-1	1	0

Fixed points on an invariant boundary component A:

Type of A	$(p, 0)^+$	$(p, k)^+, p \nmid k$	$(p_{\text{even}}, 0)^-$	$(p_{\text{even}}, 1)^-$	$(p_{\text{odd}}, 0)^-$
$\operatorname{Fix} \bar{\varphi} \cap A$	A	$\varnothing$	2 *points*	2 *points*	2 *points*
$\operatorname{index}(\bar{\varphi}, A)$	$-p$	0	$0 + 0$	$1 + 1$	$1 + 0$

REMARK 2. Isolated fixed points of zero index are removable via local perturbation. There are two cases:

For an interior fixed point of type $(p, 0)^-$, p odd, locally $\bar{\varphi}$ switches the two sides of the reflection axis and moves every point of the axis (except the fixed point) in the same direction (cf. Fig. 1). This fixed point can be removed by composing $\bar{\varphi}$ with a slight push along the axis.

For a boundary fixed point on an unstable prong of a type $(p, 0)^-$ boundary component, locally φ switches the two sides of the unstable prong and moves every point of that prong (except the fixed point) inward (cf. Fig. 2). This fixed point is removed by pushing the unstable prong into $\operatorname{int} M$.

2.3. *Fixed point classes.*

LEMMA 2.2. *Let* A_0, A_1 *be either a fixed point of* φ *or a* φ*-invariant component of* ∂M. *Suppose* A_0 *and* A_1 *are* φ*-related via a path* $c\colon I, 0, 1 \to M, A_0, A_1$. *Then there is a path* $\gamma \simeq c\colon I, 0, 1 \to M, A_0, A_1$ *such that either*

(1) γ *is in* $\operatorname{Fix} \varphi$, *or*

(2) γ *is in a component of* ∂M *of type* $(p, k)^+$.

The same is true with φ *replaced by* $\bar{\varphi}$.

Proof. By assumption we have $c \simeq \varphi \circ c \colon I, 0, 1 \to M, A_0, A_1$. Deform c to an immersion γ quasi-transverse to $\mathfrak{F}^s$. Then $\gamma \simeq \varphi \circ \gamma \colon I, 0, 1 \to M, A_0, A_1$ and $\varphi \circ \gamma$ is also an immersion quasi-transverse to $\mathfrak{F}^s$. (Cf. [**FLP**, p.76] for the definition of quasi-transversals. The Propositions II.3 and II.6 of [**FLP**, Exposé 5] can be naturally generalized to homotopy classes of paths $I, 0, 1 \to M, A_0, A_1$.) Thus

$$\mu^s(\gamma) = \inf\{\mu^s(c') \mid c' \simeq \gamma\} = \inf\{\mu^s(c') \mid c' \simeq \varphi \circ \gamma\} = \mu^s(\varphi \circ \gamma).$$

But $\mu^s(\varphi \circ \gamma) = \lambda \mu^s(\gamma)$ and $\lambda > 1$, hence $\mu^s(\gamma) = 0$. This means γ runs along the leaves of $\mathfrak{F}^s$. Thus γ is quasi-transverse to $\mathfrak{F}^u$. Then a similar argument shows that γ also runs along the leaves of $\mathfrak{F}^u$. This can occur only if γ is a constant path in $\operatorname{int} M$ or γ is in ∂M. Hence the conclusion for φ.

The isotopy from φ to $\bar{\varphi}$, obtained by "gradually slowing down" the vector fields v and v' to $\bar{v}$ and $\bar{v}'$, does not change the fixed point set except on invariant boundary components of type $(p, 0)^+$. So the fixed point classes of φ and $\bar{\varphi}$ correspond in an obvious way. Hence the conclusion remains valid for $\bar{\varphi}$. □

COROLLARY 2.3. *For the standard form $\bar{\varphi}$, every fixed point class is connected.* □

3. The general case. In this section, a standard form for a general φ is introduced. Its restriction on the periodic pieces and the pseudo-Anosov pieces having been specified in the previous sections, it remains to specify its behavior in the neighborhood of the reducing curves. This is done in §3.1. We then concentrate on the fixed point classes of the standard form. Using a book-keeping scheme introduced in §3.2, we in §3.3 identify the fixed point classes with the connected components of the fixed point set. This enables us (in §3.4) to compile a complete list of possible types of fixed point classes.

Our standard form has some other useful features. For example, iterates of a standard φ are still standard. The standard form is also "equivariant" with respect to finite group actions on M. These will not be discussed in this paper.

Suppose $\varphi \colon M \to M$ is a homeomorphism of a compact surface M and $\Gamma = \Gamma_1 \cup \cdots \cup \Gamma_n$ $(n \geq 0)$ is a disjoint union of smooth simple closed curves $\Gamma_i \subset \operatorname{int} M$ such that Γ is φ-invariant. Let $\mathcal{N}(\Gamma)$ be a φ-invariant tubular neighborhood of Γ. The components $\{N_i\}$ of $\overline{\mathcal{N}(\Gamma)}$ are annuli or Möbius bands. The components $\{M_i\}$

of $M - \mathscr{N}(\Gamma)$ are subsurfaces with negative Euler characteristic. On these (not necessarily connected) φ-components, φ is either periodic or pseudo-Anosov.

Note that according to the Thurston Theorem this is the general case: If M has only one φ-component, the case $\Gamma = \varnothing$ $(n = 0)$ is either periodic or pseudo-Anosov, while the case $\Gamma \neq \varnothing$ $(n > 0)$ is reducible.

Suppose φ is given as in the Thurston Theorem, and on pseudo-Anosov pieces φ has already been isotoped into the standard form. Thus $\varphi|\partial\mathscr{N}(\Gamma)$ is a periodic map. We shall isotope $\varphi|\overline{\mathscr{N}(\Gamma)}$ rel $\partial\mathscr{N}(\Gamma)$ into a standard form.

3.1. *Standard form for* φ *on* $\mathscr{N}(\Gamma)$. We now consider the standard form of $\varphi|\overline{\mathscr{N}(\Gamma)}$ under isotopies relative to $\partial\mathscr{N}(\Gamma)$. A component of $\overline{\mathscr{N}(\Gamma)}$ is either an annulus or a Möbius band. Let N be a φ-invariant component of $\overline{\mathscr{N}(\Gamma)}$. Then $\varphi|N\colon N \to N$ is a diffeomorphism and $\varphi|\partial N$ is periodic.

LEMMA 3.1. *Let* N *be an annulus or a Möbius band,* $\varphi\colon N \to N$ *be a diffeomorphism such that* $\varphi|\partial N$ *is periodic. Then* φ *is isotopic* rel ∂N *to a diffeomorphism* $\bar{\varphi}$ *which is either periodic or a twist. More precisely*:

(A) *The annular case*: N *is the annulus* $S^1 \times I$. *Then* φ *is isotopic* rel $S^1 \times \partial I$ *to a diffeomorphism* $\bar{\varphi}$ *which is conjugate to one of the following standard maps* $\psi\colon S^1 \times I \to S^1 \times I$.

$(1)^{+}$ $\psi(z, t) = (ze^{2(a+bt)\pi i}, t)$, *where* a, b *are rational numbers. If* $b = 0$, Fix ψ *is either* $S^1 \times I$ *or empty*; *if* $b \neq 0$, *such a* ψ *will be called a* twist. Fix ψ *is finitely many parallel circles* $S^1 \times \{t\}$.

$(2)^{-}$ $\psi(z, t) = (ze^{2a\pi i}, 1-t)$ *where* a *is a rational number*; Fix ψ *is either* $S^1 \times \{\frac{1}{2}\}$ *or empty.*

$(3)^{-}$ $\psi(z, t) = (\bar{z}, t)$; Fix ψ *is two arcs* $\{1\} \times I$ *and* $\{-1\} \times I$.

$(4)^{+}$ ” $\psi(z, t) = (\bar{z}e^{a(1-2t)\pi i}, 1-t)$, *where* a *is rational. If* $a \neq 0$, *such a* ψ *will be called a* flip-twist. Fix ψ *is two points* $(1, \frac{1}{2})$ *and* $(-1, \frac{1}{2})$.

(B) *The Möbius band case*: N *is the Möbius band represented as* $S^1 \times I$ *modulo the identification* $(z, t) \sim (-z, 1-t)$. *Then* φ *is isotopic* rel ∂N *to a diffeomorphism* $\bar{\varphi}$ *which is conjugate to one of the following standard maps* $\psi : N \to N$.

(1) $\psi(z, t) = (ze^{2a\pi i}, t) \sim (-ze^{2a\pi i}, 1-t)$, *where* a *is a rational number*; Fix ψ *is either* N, *or empty, or the central circle represented by* $S^1 \times \{\frac{1}{2}\}$.

(2) $\psi(z, t) = (\bar{z}, t) \sim (-\bar{z}, 1-t)$; Fix ψ *consists of an arc* $\{1\} \times I \sim \{-1\} \times I$ *and another point* $(1, \frac{1}{2}) \sim (-1, \frac{1}{2})$.

Proof. The analysis is based on the following classical facts.

- Every periodic map of S^1 is either conjugate to a rotation $z \mapsto ze^{2a\pi i}$, where the rational number a (mod 1) is the Poincaré rotation number, or conjugate to a reflection $z \mapsto \bar{z}$.
- A homeomorphism of $S^1 \times I$ onto itself which is the identity on $S^1 \times \partial I$ is isotopic rel $S^1 \times \partial I$ to a Dehn twist $(z, t) \mapsto (ze^{2kt\pi i}, t)$, where k is an integer.
- A homeomorphism of the Möbius band onto itself which is the identity on the boundary is isotopic rel boundary to the identity (cf. [E, Theorem 3.4]).

These standard maps fall into two major types: periodic maps and twists. The characteristic property of the twists is that $S^1 \times \partial I$ can never be in the same fixed point class of any iterate of φ. We omit the details of the elementary but somewhat tedious arguments. □

Remark 1. Strictly speaking, in order to guarantee that $\varphi|\overline{\mathcal{N}(\Gamma)}$ matches smoothly along $\partial\mathcal{N}(\Gamma)$ with the standard form of $\varphi|M - \mathcal{N}(\Gamma)$, the standard formula for the twist should be $\psi(z, t) = (ze^{2(a+b\delta(t))\pi i}, t)$, where $\delta\colon I \to I$ is a smooth increasing function, $\delta(0) = 0$, $\delta(1) = 1$, and all the derivatives vanish at 0 and 1. Similarly for the flip-twist. This modification would not change the fixed point behavior that concerns us.

Remark 2. Interior fixed circles can be removed via isotopy rel boundary. There are two cases:

Interior fixed circles of type $(\mathrm{A1})^+$ can be removed by composing ψ with a diffeomorphism $S^1 \times I \to S^1 \times I$, $(z, t) \mapsto (z, \beta(t))$, where $\beta\colon \mathbb{R} \to \mathbb{R}$ is a diffeomorphism such that $\beta(t) \neq t$ iff $0 < t < 1$.

Fixed circles of types $(\mathrm{A2})^-$ and (B1) are removed by rotating the central circle $S^1 \times \{\frac{1}{2}\}$; e.g. composing ψ with a map $S^1 \times I \to S^1 \times I$, $(z, t) \mapsto (ze^{i\alpha(t)}, t)$ where $\alpha\colon \mathbb{R} \to I$ is a smooth function with $\alpha(0) = 0$, $\alpha(\frac{1}{2}) = 1$ and $\alpha(1-t) = \alpha(t)$.

Definition. A diffeomorphism φ is said to be in *standard form,* if its restriction to every periodic φ-component is periodic, its restriction to every pseudo-Anosov φ-component is in the standard form $\bar{\varphi}$ of §2.2, and its restriction to every φ-component of $\overline{\mathcal{N}(\Gamma)}$ is in the standard form $\bar{\varphi}$ of Lemma 3.1. (For simplicity, we shall omit the bar in the notation for the standard forms of §2.2 and Lemma 3.1.)

COROLLARY 3.2. *Every fixed point class of* $\varphi|\overline{\mathcal{N}(\Gamma)}$ *is connected. Moreover, if two fixed points* x, y *are joined by a path* c *such that* $c \simeq \varphi \circ c$, *then there is a path* γ *in* $\operatorname{Fix}\varphi$ *such that* $\gamma \simeq c$ rel *endpoints.* □

3.2. *Book-keeping in the universal cover.* Consider the following general setting which clearly applies to the situation of §3.3.

Suppose M is a connected smooth m-manifold, $S \subset M$ is a (not necessarily connected) proper $(m-1)$-submanifold (proper in the sense that $S \cap \partial M = \partial S$ and S has compact intersection with any compact subset of M). Let $p : \widetilde{M} \to M$ be the universal covering of M.

DEFINITION. The book-keeping graph $G(S)$ is defined as follows. Each connected component $\widetilde{U}_i$ of $\widetilde{M} - p^{-1}(S)$ gives rise to a vertex v_i. Each connected component $\widetilde{S}_j$ of $p^{-1}(S)$ gives rise to an edge e_j. A vertex v_i is incident to an edge e_j if and only if $\widetilde{S}_j$ is contained in the closure of $\widetilde{U}_i$.

DEFINITION. Suppose $\tilde{c}$ is a path in $\widetilde{M}$ from $\tilde{x}$ to $\tilde{y}$ transverse to $p^{-1}(S)$. Let $\tilde{z}_1, \dots, \tilde{z}_k$ be the successive points where $\tilde{c}$ crosses $p^{-1}(S)$. Let $\tilde{c}_0, \tilde{c}_1, \dots, \tilde{c}_k$ be the successive segments of $\tilde{c}$ cut by $p^{-1}(S)$. Define the *book-keeping path* $\beta(\tilde{c})$ for $\tilde{c}$ to be the edge-path $v_0 e_1 v_1 \cdots v_{k-1} e_k v_k$ on $G(S)$, where e_h corresponds to the connected component $\widetilde{S}_h$ of $p^{-1}(S)$ containing $\tilde{z}_h$, and v_h corresponds to the connected component $\widetilde{U}_h$ of $\widetilde{M} - p^{-1}(S)$ containing $\tilde{c}_h$.

LEMMA 3.3. *The book-keeping graph* $G(S)$ *is a tree. If every connected component* U *of* $M - S$ *is* π_1*-injective in the sense that* $\pi_1(U)$ *injects into* $\pi_1(M)$, *and if* $\tilde{c}$ *has minimal intersection with* $p^{-1}(S)$ *among paths from* $\tilde{x}$ *to* $\tilde{y}$, *then the book-keeping path* $\beta(\tilde{c})$ *is an arc (possibly degenerate to a single vertex).*

Proof. Since $\widetilde{M}$ is a simply connected m-manifold and $\widetilde{S}_j \subset \widetilde{M}$ is a connected proper $(m-1)$-submanifold, every $\widetilde{S}_j$ must separate $\widetilde{M}$ by homological reasons. Thus every edge e_j separates $G(S)$, hence the graph $G(S)$ is a tree.

For the second part, it suffices to show there is no spur in the edge-path $\beta(\tilde{c}) = v_0 e_1 v_1 \cdots v_{k-1} e_k v_k$. If otherwise $e_h = e_{h+1}$ for some h, the path $\tilde{c}_h$ in $\widetilde{U}_h$ would have both ends in the same $\widetilde{S}_h = \widetilde{S}_{h+1}$. Now $\widetilde{U}_h$ is simply connected by the π_1-injectivity assumption. So $\tilde{c}_h$ could be deformed into $\widetilde{S}_h = \widetilde{S}_{h+1}$, thus the product path $\tilde{c}_{h-1}\tilde{c}_h\tilde{c}_{h+1}$ could be further deformed into $\widetilde{U}_{h-1} = \widetilde{U}_{h+1}$, contradicting the minimality of $\tilde{c}$. □

3.3. *Connectedness of fixed point classes.* Suppose the diffeomorphism φ is in the standard form defined in §3.1.

LEMMA 3.4. *Let A_0, A_1 be either a fixed point of φ or a φ-invariant component of ∂M. Suppose A_0 and A_1 are φ-related via a path $c: I, 0, 1 \to M, A_0, A_1$. Then there is a path $\gamma \simeq c: I, 0, 1 \to M, A_0, A_1$ such that either*

(1) *γ is in* $\operatorname{Fix}\varphi$*, or*

(2) *γ is in $A_0 = A_1$ which is a boundary component containing no fixed point.*

Proof. (1) Without loss of generality we may assume the path c has the minimal number of intersections with $\partial\mathscr{N}(\Gamma)$ in its homotopy class $I, 0, 1 \to M, A_0, A_1$. Note that if c does not cross $\partial\mathscr{N}(\Gamma)$, the truth of the conclusion is already guaranteed by Lemmas 1.2, 2.2 and 3.2.

(2) We shall work on the universal cover, using another form of the notion of φ-relation defined at the end of the Introduction.

Alternative Definition. Path-connected φ-invariant subsets A_0, $A_1 \subset M$ are φ-related (via a path c) if and only if: there is a lifting $\tilde{\varphi}: \widetilde{M} \to \widetilde{M}$ of φ on the universal cover $\widetilde{M}$ of M, such that some connected component $\widetilde{A}_0$ of $p^{-1}(A_0)$ and some connected component $\widetilde{A}_1$ of $p^{-1}(A_1)$ (joined by a lifting $\tilde{c}$ of c) are $\tilde{\varphi}$-invariant.

(The "if" part is trivial. For the "only if" part, choose a lifting $\tilde{\varphi}$ of φ and a lifting $\tilde{c}$ of c such that the given homotopy $c \simeq \varphi \circ c$ lifts to a homotopy $\tilde{c} \simeq \tilde{\varphi} \circ \tilde{c}$. See [J3, Theorem I.1.10] for the case when both A_0, A_1 are single points.)

(3) Apply the book-keeping scheme of §3.2 to M with $m = 2$ and $S := \partial\mathscr{N}(\Gamma)$, and use the notation in the definition of $\beta(\tilde{c})$. The π_1-injectivity condition in Lemma 3.3 is clearly satisfied in our setting, and the minimality assumption on c guarantees the minimality of $\tilde{c}$ with respect to $p^{-1}(S)$. So $\beta(\tilde{c}) = v_0e_1v_1\cdots v_{k-1}e_kv_k$ is an arc in $G(S)$.

The diffeomorphism $\tilde{\varphi}$ leaves $p^{-1}(S)$ invariant, so the bookkeeping path $\beta(\tilde{\varphi} \circ \tilde{c}) = v_0' e_1' v_1' \cdots v_{k-1}' e_k' v_k'$ is also an arc.

Since $\widetilde{A}_0$ and $\widetilde{A}_1$ are $\tilde{\varphi}$-invariant, we have $v_0 = v_0'$ and $v_k = v_k'$. By the uniqueness of joining arcs in the tree $G(S)$, we conclude that $v_h = v_h'$ and $e_h = e_h'$ for all h. Thus every $\widetilde{U}_h$ and $\widetilde{S}_h$ is $\tilde{\varphi}$-invariant.

(4) Apply Definition (2) to the universal cover $p : \widetilde{\overline{U}}_h \to \overline{U}_h$, we see $S_h := p(\widetilde{S}_h)$ and $S_{h+1} = p(\widetilde{S}_{h+1})$ are φ-related on the subsurface $\overline{U}_h$, via $c_h := p \circ \tilde{c}_h$, the h-th segment of c.

(5) The subsurfaces $\overline{U}_0, \dots, \overline{U}_k$ are alternately of two different kinds: components of $M - \mathscr{N}(\Gamma)$ (the M_j's with $\chi < 0$), and components of $\overline{\mathscr{N}(\Gamma)}$ (the N_i's with $\chi = 0$).

(6) For every h of the first kind, Lemmas 1.2 and 2.2 say there is a path γ_h in $\operatorname{Fix}\varphi$ and paths τ_h, τ_{h+1} in S_h, S_{h+1} such that $c_h \simeq \tau_h \gamma_h \tau_{h+1}$ rel endpoints. Replacing the segment c_h with $\tau_h \gamma_h \tau_{h+1}$ and then slightly pushing the parts τ_h, τ_{h+1} into the neighboring regions U_{h-1}, U_{h+1} (of the second kind) respectively, we deform c into a new c with γ_h as the new c_h. This deformation does not affect the minimality of c with respect to S. Hence we may *assume from now on* that c_h is in $\operatorname{Fix}\varphi$ for every h of the first kind.

(7) Continue with h of the first kind. Now that $\varphi \circ c_h = c_h$, there exists a covering translation α_h such that $\tilde{\varphi} \circ \tilde{c}_h = \alpha_h \circ \tilde{c}_h$. We claim that $\alpha_h = 1$. Indeed, when A_0 or A_1 is a point, this is true for $h = 0$ or k because $\widetilde{A}_0$ or $\widetilde{A}_1$ is a point. In all other cases, apply the following geometric observation to the universal cover $p: \widetilde{\overline{U}}_h \to \overline{U}_h$.

Observation. Suppose $p: \widetilde{M} \to M$ is the universal cover of a connected compact surface M with boundary, $\chi(M) < 0$. Suppose α is a covering translation. If there is a path $\tilde{c}$ in $\widetilde{M}$ such that $\tilde{c}$ and $\alpha \circ \tilde{c}$ join the same pair of different connected components of $\partial \widetilde{M}$, then α must be the identity.

(Proof of this observation: Think of M as a hyperbolic surface with totally geodesic boundary, so that $\partial \widetilde{M}$ consists of hyperbolic straight lines. Via homotopy we may replace $\tilde{c}$ with the unique shortest geodesic joining that pair of components of $\partial \widetilde{M}$. Since α is an isometry, $\alpha \circ \tilde{c}$ is the same shortest geodesic. So $\alpha = 1$.)

(8) We have shown that for h of the first kind, $\tilde{c}_h$ is indeed a path in $\operatorname{Fix}\tilde{\varphi}$. In particular, $\tilde{z}_h, \tilde{z}_{h+1}$ are fixed points of $\tilde{\varphi}$. But every other h is of the first kind, so that *all* cut points $\tilde{z}_1, \dots, \tilde{z}_k$ of $\tilde{c}$ are fixed points of $\tilde{\varphi}$.

(9) Now we turn to those h of the second kind, $\overline{U}_h \subset \overline{\mathscr{N}(\Gamma)}$. By Definition (2), z_h, z_{h+1} are in the same fixed point class of $\varphi | \overline{U}_h$.

Corollary 3.2 tells us there is a path homotopy $c_h \simeq \gamma_h$ rel endpoints such that γ_h is in $\operatorname{Fix}\varphi$. Replacing every such segment c_h with γ_h, we obtain the desired path γ lying entirely in $\operatorname{Fix}\varphi$. □

COROLLARY 3.5. *Every fixed point class of φ is connected.* □

3.4. *Types of fixed point classes.* The last corollary enables us to identify the fixed point classes of φ. Each is a component of $\operatorname{Fix}\varphi$, hence a union of the fixed point components on the standard pieces described in Lemmas 1.1, 2.1 and 3.1. Putting together all the information there and paying attention to the consistency along $\partial\mathscr{N}(\Gamma)$, we can get

LEMMA 3.6. *The possible types of fixed point classes of φ are listed below, with a description of their local behavior.*

$(1)^{\pm}$ *Isolated fixed point x:*

(a)$^+$ $x \in \operatorname{int} M$, *$\varphi$ is conjugate to a rotation in a neighborhood of x;* $\operatorname{index}(\varphi, x) = 1$.

(b)$^+$ $x \in \operatorname{int} M$ *is a fixed point of an annular flip-twist;* $\operatorname{index}(\varphi, x) = 1$.

(c)$^+$ $x \in \operatorname{int} M$ *is a type $(p, k)^+$ interior fixed point of a pseudo-Anosov piece;* $\operatorname{index}(\varphi, x) = 1 - p$ *or* 1.

(d)$^-$ $x \in \operatorname{int} M$ *is a type $(p, k)^-$ interior fixed point of a pseudo-Anosov piece;* $\operatorname{index}(\varphi, x) = 1, -1$ *or* 0.

(e)$^-$ $x \in \partial M$ *and x is in a type $(p, k)^-$ invariant boundary component of some pseudo-Anosov piece;* $\operatorname{index}(\varphi, x) = 1$ *or* 0.

$(2)^{\pm}$ *Fixed circle C:*

(a)$^+$ $C \subset \operatorname{int} M$ *is a fixed circle of an annular twist;* $\operatorname{index}(\varphi, C) = 0$.

(b)$^-$ $C \subset \operatorname{int} M$ *and in a neighborhood of C, φ is conjugate to the reflection $(z, t) \mapsto (z, 1-t)$ on the annulus $S^1 \times I$ or the Möbius band $S^1 \times I/\sim$;* $\operatorname{index}(\varphi, C) = 0$.

(c)$^+$ $C \subset \operatorname{int} M$; *on one side C is a type $(p, 0)^+$ boundary component of some pseudo-Anosov piece, on the other side C is a boundary component of an annular twist;* $\operatorname{index}(\varphi, C) = -p$.

(d)$^+$ $C \subset \partial M$, *and C is a type $(p, 0)^+$ boundary component of some pseudo-Anosov piece;* $\operatorname{index}(\varphi, C) = -p$.

$(3)^-$ *Fixed arc A, contained in some subsurface B of M on which φ acts as an involution. Every endpoint x of A is either*

(a) $x \in \operatorname{int} M$, *on the outside of B x is in a type $(p, k)^-$ invariant boundary component of a pseudo-Anosov piece, or*

(b) $x \in \partial M$.

The possible values of index(φ, A) *are* $1, -1$ *or* 0.

$(4)^+$ *Fixed subsurface* B *of* M *with* $\chi(B) \leq 0$. *The possible forms for a component* C *of* ∂B:

(a) $C \subset \operatorname{int} M$, *on the outside of* B C *is a type* $(p_C, 0)^+$ *invariant boundary component of some pseudo-Anosov piece;*

(b) $C \subset \operatorname{int} M$, *on the outside of* B C *is a boundary component of an annular twist;*

(c) $C \subset \partial M$.

We have $\operatorname{index}(\varphi, B) = \chi(B) - \sum p_C < 0$ *where the summation is over the components* C *of* ∂B *of type* (a).

Proof. These are the only possible combinations of the fixed point sets of the standard models. The calculation of the index can be done using the well-known proposition below. □

PROPOSITION 3.7. *Let* $f: X \to X$ *be a self-map of a compact polyhedron. Suppose* X_0, X_1, X_2 *are subpolyhedra of* X *such that* $X = X_1 \cup X_2$, $X_0 = X_1 \cap X_2$. *We suppose* $f(X_i) \subset X_i$ *and write* $f_i: X_i \to X_i$ *for the restriction of* f, *for* $i = 0, 1, 2$. *Let* $A \subset \operatorname{Fix} f$ *be both open and closed in* $\operatorname{Fix} f$, *and let* $A_i = A \cap X_i$, $i = 0, 1, 2$. *Then*

$$\operatorname{index}(f, A) = \operatorname{index}(f_1, A_1) + \operatorname{index}(f_2, A_2) - \operatorname{index}(f_0, A_0). \quad \square$$

4. Proof of the Main Theorem. In this section we first prove the Main Theorem. Then we prove an inequality relating the Lefschetz number and the Nielsen number.

Proof of the Main Theorem. We are supposed to show that every essential fixed point class of φ (in the standard form) is shrinkable to a point, and every inessential one is removable, via a smooth isotopy (through diffeomorphisms or through embeddings). In Steps 1–4 below, we examine successively the various types of fixed point classes listed in Lemma 3.6.

Step 1^-. Isolated fixed points of zero index.

This can occur in types $(1d)^-$ and $(1e)^-$. They can be removed according to Remark 2 of §2.2.

Step $2^\pm$. Fixed circles.

Circles of types $(2a)^+$ and $(2b)^-$ can be removed according to Remark 2 of §3.1. Circles of types $(2c)^+$ and $(2d)^+$ will be treated later in Step 4.

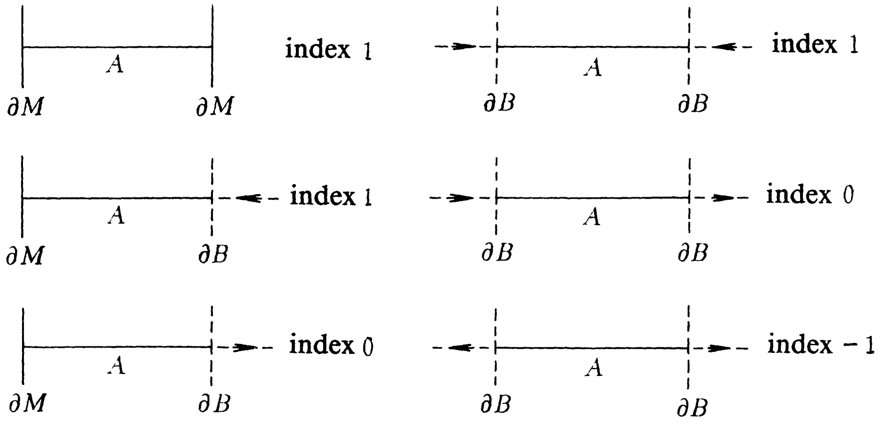

FIGURE 3

Step 3^-. Fixed arcs.

Arcs of type $(3)^-$ have six possible forms shown in Fig. 3, in which φ switches the two sides of the horizontal axis and moves points on the extension of A in the direction of the arrows. In the two forms with zero index, the fixed point set A can be removed by composing φ with a slight push along the axis in the direction of the arrow. In the forms with index ± 1 the fixed point class can be reduced to a point by composing φ with a slight contraction or stretch along the axis.

Step 4^+. Fixed point classes F of types $(2c)^+$, $(2d)^+$ and $(4)^+$ have some common features and will be given a unified treatment. They all have negative index so the task is to reduce F to a point via an isotopy. Let us fix a hyperbolic metric on M and let $\delta > 0$ be the minimal length of closed geodesics of M. Then every simple closed curve of length less than δ must bound a disk.

For all these three types, it is clear that there exists a neighborhood W of F and a smooth vector field w on W such that $w = 0$ on F, $|w| < \delta/2$ on W, φ coincides with the time-one map for w on a smaller neighborhood V of F. Perturb w to get a smooth vector field v on W so that $v(x) \neq w(x)$ only if x is in a sufficiently small neighborhood U of F, $|v| < \delta/2$ on W, v is tangent to ∂M on $\partial M \cap W$, and v has a unique singularity x_0 on W, x_0 being any pre-assigned point of F. The index of the singularity must equal to $\operatorname{index}(\varphi, F) < 0$. Let φ' be the time-one map for v. When U is small enough, φ' is well defined on V and coincides with φ near the boundary of V. So, letting $\varphi' = \varphi$ on $M - V$, we extend φ'

to a diffeomorphism $M \to M$. This φ' is clearly isotopic to φ. It remains to show φ' has no fixed points other than x_0.

The vector field v on W cannot have closed orbits shorter than δ, because otherwise on the disk bounded by this closed orbit the sum of indices of the singularities of v must add up to 1, contradicting the fact that v has only one singularity with negative index. But $|v| < \delta/2$, so the time-one map can have no fixed point other than the singularity x_0.

Thus all types have been treated and our goal is achieved.

Note that isotopies through embeddings are needed only in Step 1, type $(1e)^-$ and in Step 3, all involving orientation-reversing φ-invariant components of ∂M. □

As a by-product of the analysis leading to the Main Theorem, we have

THEOREM 4.1. *Let M be a compact surface with $\chi(M) < 0$, and $f: M \to M$ be a homeomorphism. Let $L(f)$ be the Lefschetz number of f. Then $N(f) \geq L(f) \geq -N(f) + 2\chi(M)$, or equivalently, $|L(f) - \chi(M)| \leq N(f) - \chi(M)$.*

Proof. Since both $L(f)$ and $N(f)$ are homotopy invariants of f, we may assume f is in the standard form φ. The fixed point classes of φ are as described in Lemma 3.6. $L(\varphi) \leq N(\varphi)$ simply because every fixed point class of φ has index ≤ 1. The other inequality needs a closer analysis. Clearly

$$L(\varphi) + N(\varphi) \geq \sum_A (\operatorname{index}(\varphi, A) + 1)$$

where the summation is taken over all fixed point classes A with $\operatorname{index}(\varphi, A) < -1$. A 0-dimensional fixed point class A in the last summation must be a point x which is an interior p_x-prong singularity of some pseudo-Anosov piece M_j and $\operatorname{index}(\varphi, A) + 1 = 2 - p_x$. A 1-dimensional fixed point class in this sum is a circle C which is a p_C-prong boundary component of some pseudo-Anosov piece of φ, $\operatorname{index}(\varphi, A) = -p_C$. For a 2-dimensional fixed point class A, $\operatorname{index}(\varphi, A) = \chi(A) - \sum p_C$ where the summation is over components C of ∂A which is at the same time a p_C-prong boundary component of a pseudo-Anosov piece M_j. Hence

$$L(\varphi) + N(\varphi) \geq \sum_1 \left\{ \sum_x (2 - p_x) + \sum_C (-p_C) \right\} + \sum_2 \chi(M_j)$$

where $\sum_1$ sums over the pseudo-Anosov pieces M_j of φ and $\sum_2$ sums over the periodic pieces M_j of φ, in the braces x runs over all interior singularities of $\varphi|M_j$ and C runs over the boundary components. By the Euler-Poincaré formula of **[FLP**, p.75] (applied to the stable foliation of $\varphi|M_j$), the sum in the braces equals $2\chi(M_j)$. Thus

$$L(\varphi)+N(\varphi)\geq \sum_1 2\chi(M_j)+\sum_2 \chi(M_j)\geq 2\chi(M)$$

since every $\chi(M_j)\leq 0$. □

REMARK. From the proof we see that the equality $L(f)+N(f)=2\chi(M)$ can occur only if f is isotopic to a pseudo-Anosov map of a closed surface.

Question. Is Theorem 4.1 true even for self-maps of M? Note that for compact surfaces with $\chi=0$, it is known that $|L(f)|\leq N(f)$ for every self-map f.

5. The relative Nielsen numbers. What is the best lower bound for the number of fixed points in an isotopy class of diffeomorphisms? The Main Theorem provides the answer only for the isotopy classes without orientation-reversing invariant boundary components. A candidate for a general answer is the relative Nielsen number introduced by Schirmer **[S]**.

Let (X, A) be a pair of compact polyhedra, $f\colon X, A\to X, A$ be a self-map of the pair. Let $f_A\colon A\to A$ be the restriction of f. Denote by $N(f, f_A)$ the number of essential fixed point classes of f that contain some essential fixed point class of f_A. Define

$$N(f; X, A):=N(f)+N(f_A)-N(f, f_A).$$

It is shown in **[S]** that for any map $g\simeq f\colon X, A\to X, A$, we always have $N(f; X, A)=N(g; X, A)$, therefore such a map g has at least $N(f; X, A)$ fixed points on X.

Another relative Nielsen number is introduced by Zhao **[Z]**. Define $N(f; X-A)$ to be the number of essential fixed point classes of f that are not f-related to the f-invariant set A. For any map $g\simeq f\colon X, A\to X, A$, we always have $N(f; X-A)=N(g; X-A)$, hence such a map g has at least $N(f; X-A)$ fixed points on the complement $X-A$.

Note that a map $f\colon X, A\to X, A$ has exactly $N(f; X, A)$ fixed points on X and exactly $N(f; X-A)$ fixed points on $X-A$ if and only if the following three conditions are satisfied:

- every fixed point class of f_A is a single point;
- every inessential fixed point class of f consists of essential fixed points of f_A;
- an essential fixed point class of f can have more than one point only if it consists of essential fixed points of f_A.

For homeomorphisms of surfaces we have:

THEOREM 5.1. *Let M be a compact surface and let $f: M \to M$ be a homeomorphism. Then f is isotopic to a diffeomorphism which has exactly $N(f; M, \partial M)$ fixed points on M and exactly $N(f; M-\partial M)$ fixed points in* $\operatorname{int} M$.

Proof. For the three connected compact surfaces M with $\chi(M) \geq 0$, the conclusion is obvious. Hence we assume that every component of M has negative Euler characteristic. By Thurston Theorem we may replace f with the diffeomorphism φ in the standard form, as at the beginning of §4. Now we isotope φ as in the proof of the Main Theorem, with the following modifications: In Step 1 and Step 3, when we push along the axis we leave the intersection of the axis with ∂M fixed. In Step 4, we choose the singularity x_0 of the vector field v to be in $F \cap \partial M$ when $F \cap \partial M \neq \varnothing$. The result φ' is then a diffeomorphism and satisfies the three conditions listed above. Hence φ' has exactly $N(\varphi'; M, \partial M)$ fixed points on M and exactly $N(\varphi'; M - \partial M)$ fixed points on $\operatorname{int} M$. □

Acknowledgment. We thank the referee for comments that led to improvement of exposition and to simplification of the analysis in §§3.2–3.

REFERENCES

[B] L. Bers, *An extremal problem for quasiconformal mapping and a theorem by Thurston*, Acta Math., **141** (1978), 73–98.

[BK] J. S. Birman and M. Kidwell, *Fixed points of pseudo-Anosov diffeomorphisms of surfaces*, Adv. in Math., **46** (1982), 217–220.

[Br] R. F. Brown, *The Lefschetz Fixed Point Theorem*, Scott-Foresman, Chicago, 1971.

[E] D. B. A. Epstein, *Curves on 2-manifolds and isotopies*, Acta Math., **115** (1966), 83–107.

[FLP] A. Fathi, F. Laudenbach, and V. Poénaru, *Travaux de Thurston sur les surfaces, Séminaire Orsay*, Astérisque vol. 66–67, Soc. Math. France, Paris, 1979.

[GK] M. Gerber and A. Katok, *Smooth models of Thurston's pseudo-Anosov maps*, Ann. Sci. Éc. Norm. Sup., **15** (1982), 173–204.

[HT] M. Handel and W. P. Thurston, *New proofs of some results of Nielsen*, Adv. in Math., **56** (1985), 173–191.

[I] N. V. Ivanov, *Nielsen numbers of self-maps of surfaces*, Studies in Topology IV, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov (LOMI), vol.122, Nauka, Leningrad, 1982, pp. 56–65, 163–164 (Russian); English transl., J. Sov. Math., **26** (1984), 1636–1641.

[J1] B. Jiang, *Fixed point classes from a differential viewpoint*, Fixed Point Theory (E. Fadell, G. Fournier, eds.), Lecture Notes in Math., vol. 886, Springer-Verlag, Berlin, Heidelberg, New York, 1981, pp. 163–170.

[J2] ——, *Fixed points of surface homeomorphisms*, Bull. Amer. Math. Soc., **5** (1981), 176–178.

[J3] ——, *Lectures on Nielsen Fixed Point Theory*, Contemp. Math., vol.14, Amer. Math. Soc., Providence, RI, 1983.

[J4] ——, *Fixed points and braids*, Invent. Math., **75** (1984), 69–74; II, Math. Ann., **272** (1985), 249–256.

[K] T.-H. Kiang, *The Theory of Fixed Point Classes*, Science Press, Beijing, 1979, 1986 (Chinese); English edition, 1989, Springer-Verlag, Berlin, Heidelberg, New York.

[N1] J. Nielsen, *Über die Minimalzahl der Fixpunkte bei den Abbildungstypen der Ringflächen*, Math. Ann., **82** (1921), 83–93.

[N2] ——, *Untersuchungen zur Topologie des geschlossenen zweiseitigen Flächen*, I, Acta Math., **50** (1927), 189–358; English transl., *Investigations in the topology of closed orientable surfaces*, I, Jakob Nielsen: Collected Mathematical Papers (V.L. Hansen, ed.), vol. 1, Birkhäuser, Boston, 1986, pp. 223–341.

[S] H. Schirmer, *A relative Nielsen number*, Pacific J. Math., **122** (1986), 459–473.

[T] W. P. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc., **19** (1988), 417–431.

[W] F. Wecken, *Fixpunktklassen*, I, Math. Ann., **117** (1941), 659–671; II, Math. Ann., **118** (1942), 216–234; III, Math. Ann., **118** (1942), 544–577.

[Wu] Y.-Q. Wu, *Canonical reducing curves of surface homeomorphisms*, Acta Math. Sinica, New Series, **3** (1987), 305–313.

[Z] X.-Z. Zhao, *A relative Nielsen number for the complement*, Topological Fixed Point Theory and Applications (B. Jiang, ed.), Lecture Notes in Math., vol. 1411, Springer-Verlag, Berlin, Heidelberg, New York, 1989, pp. 189–199.

Received October 17, 1990 and in revised form March 3, 1992. Both authors partially supported by NSFC grants.

PEKING UNIVERSITY
BEIJING 100871, CHINA

AND

HANGZHOU UNIVERSITY
HANGZHOU, ZHEJIANG 310028, CHINA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

THE MODULI OF RATIONAL WEIERSTRASS FIBRATIONS OVER $\mathbf{P}^1$: SINGULARITIES

PABLO LEJARRAGA

The Weierstrass equation $y^2 = x^3 + ax + b$, where a and b are rational functions of one variable, defines a fibration over $\mathbf{P}^1$, which we call a Weierstrass fibration. We consider the moduli space W of rational Weierstrass fibrations over $\mathbf{P}^1$. In this paper we determine the singular locus of W and we compute the general singularities. We work over $\mathbf{C}$, but it seems possible to generalize our methods to characteristic $p \neq 2, 3$.

Introduction. In **[Mi]** Miranda has constructed moduli spaces W_N, $N \geq 0$, for Weierstrass fibrations over $\mathbf{P}^1$ whose zero section has self intersection number $-N$ in the associated elliptic surface. Seiler has generalized and extended this work in **[Sei2]** and **[Sei3]**. For $N = 1$, we have the moduli space of rational fibrations $W = W_1$. The points of W parametrize isomorphism classes of rational Weierstrass fibrations over $\mathbf{P}^1$ with at most rational double point singularities whose associated elliptic surface (= minimal resolution of singularities) has only reduced fibers. By passing to the associated elliptic surface, W can be viewed as parametrizing isomorphism classes of relatively minimal elliptic surfaces over $\mathbf{P}^1$ admitting a section which have only reduced fibers. The basic definitions and constructions are reviewed in §1.

To determine the singular locus of W, we first find the locus S of Weierstrass fibrations that have non-negligible (= nontrivial) automorphisms. By means of the Weierstrass equation, this boils down to finding stable pairs of Weierstrass coefficients whose isotropy group with respect to the action of $G = \mathbf{GL}_2/\pm I$ is nontrivial. This work is the content of §2 and culminates in Theorem 1 where the 7 irreducible components of S are listed.

The general singularities turn out to be cyclic quotient singularities. We compute and classify them with the help of the slice theorem and work of Prill **[Pr]** in Theorem 2, §3.

This work is part of my Ph.D. thesis. I want to thank my advisor M. Artin and Rick Miranda for their help.

1. Generalities. All varieties we consider are defined over the field of complex numbers $\mathbf{C}$. Unless otherwise stated all topological notions refer to the Zariski topology. We refer the reader to **[Mi]**, **[Ka]** and **[M-S]** for proofs of the following facts in this section.

Let S be a variety. Let $p\colon Y \to S$ be a flat proper morphism of irreducible varieties whose fibers are of one of the following types:

(a) an elliptic curve,

(b) a rational curve with a node,

(c) a rational curve with a cusp.

Let σ be a section of p not touching the nodes and cusps of the fibers. The quadruple (Y, S, p, σ) is called a *Weierstrass fibration over* S. We usually denote Weierstrass fibrations by Y/S when there is no risk of confusion.

A morphism of a Weierstrass fibration (Y, S, p, σ) into a Weierstrass fibration (Y', S', p', σ') is given by a pair of morphisms $f\colon Y \to Y'$ and $\varphi\colon S \to S'$ such that $p' \circ f = \varphi \circ p$ and $f \circ \sigma = \sigma' \circ \varphi$.

When $S = C$ is a complete nonsingular connected curve, a Weierstrass fibration with nonsingular general fiber and only rational double point singularities is called a *Weierstrass model*. As is well known, a Weierstrass model Y/C can be described by a *Weierstrass equation* over C, i.e. there exists an invertible sheaf $\mathscr{L}$ over C and sections a of $\mathscr{L}^{\otimes 4}$ and b of $\mathscr{L}^{\otimes 6}$ such that Y is isomorphic to the hypersurface in $\mathbf{P}(O_C \otimes \mathscr{L}^{\otimes(-2)} \oplus \mathscr{L}^{\otimes(-3)})$ given by $y^2 = x^3 + ax + b$. The morphism $J = J(a, b) = 4a^3/(4a^3 + 27b^2)$ of C into $\mathbf{P}^1$ is called the J*-invariant.*

Let $S = \mathbf{P}^1$. Choose coordinates t, s such that $t = 1$, $s = 0$ is the point at infinity. Call V_n the set of homogeneous functions of degree n on $\mathbf{P}^1$ viewed as homogeneous forms of degree n in t, s. Call G the quotient group $\mathbf{GL}_2/(\pm I)$. We use the same notation for a matrix $\left(\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right)$ in $\mathbf{GL}_2$ and for its image in G. We also use the notation $\left(\begin{smallmatrix}\alpha & \\ & \delta\end{smallmatrix}\right)$ for diagonal matrices, $\alpha = \left(\begin{smallmatrix}\alpha & \\ & \alpha\end{smallmatrix}\right)$ for scalar matrices and $\left(\begin{smallmatrix} & \beta\\ \gamma & \end{smallmatrix}\right)$ for matrices with zeros in the main diagonal. Let $f(t, s) \in V_n$ and g be an element of G with matrix $\left(\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right)$. We define

$$(f \cdot g)(t, s) = f(\alpha t + \beta s, \gamma t + \delta s).$$

This defines a right action of G on V_n. The pair of coefficients (a, b) of a rational Weierstrass model over $\mathbf{P}^1$ can be interpreted as an element of $V_4 \times V_6$.

In this way we get an injection of the set of isomorphism classes of rational Weierstrass models over $\mathbf{P}^1$ into $(V_4 \times V_6)/G$, where G acts by means of its actions on V_4 and V_6. Denote by X the open set of $\mathbf{SL}_2$-*stable* (= finite stabilizer and closed orbit) elements of $V_4 \times V_6$ (to be called just *stable* from now on). The quotient algebraic variety $W = X/G$ is called the *moduli of rational Weierstrass fibrations over* $\mathbf{P}^1$. We denote by $\pi\colon X \to W$ the canonical map. Under the above injection points of W correspond to classes of Weierstrass models whose associated elliptic surface has reduced fibers. For $f \in V_n$ and $\tau \in \mathbf{P}^1$ denote by $v_\tau(f)$ the order of vanishing of f at τ. An element $(a, b) \in V_4 \times V_6$ is stable if and only if the following *numerical criterion* holds:

$$\min(3v_\tau(a), 2v_\tau(b)) < 6$$

for all $\tau \in \mathbf{P}^1$.

Let $x = (a, b) \in X$. Denote by Y_x the Weierstrass fibration with equation $\eta^2 = \xi^3 + a\xi + b$. Denote by $\operatorname{Stab} x$ the *isotropy group* (= stabilizer) of x with respect to the action of G. Denote by $\operatorname{Aut}_{WF}(Y_x/\mathbf{P}^1)$ the automorphism group of the Weierstrass fibration $Y_x/\mathbf{P}^1$ and by N the normal subgroup of *negligible* automorphisms, i.e., those of the form

$$\eta = \pm\eta', \quad \xi = \xi', \quad t = t', \quad s = s'.$$

Define $\operatorname{Aut}_{RWF}(Y_x/\mathbf{P}^1) = (\operatorname{Aut}_{WF}(Y_x/\mathbf{P}^1))/N$, the *reduced automorphism group of* $Y_x/\mathbf{P}^1$. Given $g \in \operatorname{Stab} x$ with matrix $\lambda\left(\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right)$, $\lambda \neq 0$, $\alpha\delta - \beta\gamma = 1$, the formulas

$$\eta = \lambda^{-3}\eta', \quad \xi = \lambda^{-2}\xi', \quad t = \alpha t' + \beta s', \quad s = \gamma t' + \delta s'$$

define an element of $\operatorname{Aut}_{RWF}(Y_x/\mathbf{P}^1)$ denoted by $\operatorname{Aut} g$. The following proposition follows from well known facts.

PROPOSITION 1. *The canonical group homomorphism* $\operatorname{Stab} x \to \operatorname{Aut}_{RWF}(Y_x/\mathbf{P}^1)$, $g \mapsto \operatorname{Aut} g$ *is bijective.*

We view the J-invariant $J(x) = J(a, b) = 4a^3/(4a^3 + 27b^2)$ as a morphism of $\mathbf{P}^1$ into $\mathbf{P}^1$. We denote by $\operatorname{Aut} J(x)$ the group of deck transformations of $J(x)\colon \mathbf{P}^1 \to \mathbf{P}^1$. For g an element of G with matrix $\left(\begin{smallmatrix}\alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right)$ we denote by Pg the linear fractional transformation $z \mapsto (\alpha z + \beta)/(\gamma z + \delta)$, viewed as an element of $\mathbf{PGL}_2 = \operatorname{Aut} \mathbf{P}^1$. The proof of the following easy corollary is left to the reader.

COROLLARY. *Suppose that* $x \in X$ *has nonconstant* J*-invariant. The canonical group homomorphism* $\operatorname{Stab} x \to \operatorname{Aut} J(x)$, $g \mapsto Pg$ *is injective.*

REMARK. In fact the homomorphism of the above corollary is *bijective*, but the proof is more involved.

2. Components of S. Recall that $\pi: X \to W$ is the canonical morphism. Define

$$S = \pi\{x \in X \mid \operatorname{Stab} x \neq 1\}.$$

By the corollary to Proposition 1, this set is the locus in moduli of Weierstrass fibrations with nontrivial automorphisms. In this section we determine the irreducible components of the closed set S.

Let the group Γ operate on the set E. Let H be a subgroup of Γ. We denote

$$E^H = \{x \in E \mid xg = x \text{ for all } g \in H\}.$$

For $g \in \Gamma$, define $E^g = E^{(g)}$, where (g) is the group generated by g; we remark that E^g is the set of x in E such that $xg = x$. When $E = X$, $\Gamma = G$, H subgroup of G, $g \in G$, we use the notations

$$\operatorname{Inv} H = X^H, \qquad \operatorname{Inv} g = X^g.$$

It is clear that

$$S = \bigcup\{\pi(\operatorname{Inv} g) \mid g \in G,\ \ g \neq 1,\ \ g \text{ of finite order}\}.$$

LEMMA 1. *Let $g \in G$ be of finite order. The sets* $\operatorname{Inv} g$ *and* $\pi(\operatorname{Inv} g)$ *are irreducible closed in X and W respectively.*

REMARK. It follows from Lemma 1 that the maximal elements among the $\pi(\operatorname{Inv} g)$ are the irreducible components of S. Since a Noetherian topological space has a finite number of irreducible components, the set S is closed.

Proof of Lemma 1. We have

$$\operatorname{Inv} g = (V_4 \times V_6)^g \cap X$$

where $(V_4 \times V_6)^g$ is a sub-vector space of $V_4 \times V_6$ and X is open in $V_4 \times V_6$. It follows that $\operatorname{Inv} g$ is irreducible and closed. Consequently $\pi(\operatorname{Inv} g)$ is irreducible. We have not used the fact that g is of finite order up to here.

Now let C be the conjugacy class of g. Since g is of finite order it follows from [**Bo**, pp. 227–228] that C is closed. Moreover G acts properly on X by [**GIT**, p. 41, Converse 1.13] and the fact that $\pi: X \to W = X/G$ is affine. Hence the morphism

$$X \times G \xrightarrow{\psi} X \times X,$$
$$(x, h) \mapsto (xh, x)$$

is proper.

Denote by Δ_X the diagonal morphism of X into $X \times X$. It follows that the set $\Delta_X^{-1}(\psi(X \times C))$ is closed. Since

$$\Delta_X^{-1}(\psi(X \times C)) = \{x \in X | xg = x \text{ for some } g \in C\}$$

is G-saturated, it is clear that

$$\pi(\operatorname{Inv} g) = \pi(\Delta_X^{-1}(\psi(X \times C)))$$

is closed. □

For any prime number p, let R_p be a system of representatives of the equivalence classes of elements of $\mathbf{F}_p^* - \{1\} = (\mathbf{Z}/p\mathbf{Z}) - \{0, 1\}$ with respect to the equivalence relation between elements u, v of $\mathbf{F}_p^* - \{1\}$ defined by the condition "$u = v$ or $u = v^{-1}$". Moreover we define $\zeta_n = e^{2\pi i/n}$.

LEMMA 2. *We have*

$$S = \bigcup \pi(\operatorname{Inv} g)$$

where g *runs over the following list*:

$$\begin{matrix} \begin{pmatrix} i & \\ & i \end{pmatrix}, & \begin{pmatrix} \zeta_3 & \\ & \zeta_3 \end{pmatrix}, & \\ \begin{pmatrix} \zeta_p & \\ & 1 \end{pmatrix}, & \begin{pmatrix} \zeta_p^l & \\ & \zeta_p \end{pmatrix}, & l \in R_p, \ p = 3, 5, 7, 11. \\ \begin{pmatrix} -1 & \\ & 1 \end{pmatrix}, & \begin{pmatrix} -i & \\ & i \end{pmatrix}, & \end{matrix}$$

The inclusion

$$\bigcup \pi(\operatorname{Inv} g) \subset S$$

is obvious. Now let $u \in S$. There are two cases:

(i) $J(x) = 0$ (resp. $J(x) = 1$) for all $x \in \pi^{-1}(u)$.
(ii) $J(x)$ is nonconstant for all $x \in \pi^{-1}(u)$.

Case (i). The conditions $J(x) = 0$ and $J(x) = 1$ are equivalent to $x \in \operatorname{Inv} \zeta_3$ and $x \in \operatorname{Inv} i$ respectively. We conclude in this case that

$$u \in \bigcup \pi(\operatorname{Inv} g)$$

where $g = i, \zeta_3$.

Case (ii). Since $J(x)$ is nonconstant, it follows from the rationality of the Weierstrass model determined by x, that $\deg J(x) \leq 12$, where $\deg J(x)$ denotes the degree of the cover $J(x)\colon \mathbf{P}^1 \to \mathbf{P}^1$. The following argument shows that every element φ of $\operatorname{Aut} J(x)$ has order $\leq d = \deg J(x) \leq 12$. Take a classical nonempty open set U in $\mathbf{P}^1$ such that $(J(x))^{-1}(U)$ is a disjoint union of d copies of U. Suppose that V is one of such copies. Then the sequence $V = \varphi^0(V), \varphi^1(V), \dots, \varphi^d(V)$ has a repetition, say

$$\varphi^i(V) = \varphi^j(V) \quad \text{for } 0 \leq i < j \leq d.$$

Thus

$$V = \varphi^{j-i}(V)$$

which implies, since φ is an analytic function, that φ has order $\leq j - i \leq d$. We conclude by the corollary to Proposition 1 that every element of $\operatorname{Stab} x$ has order ≤ 12. Now we notice the following facts.

(α) If $u \in \pi(\operatorname{Inv} g)$, there exists $x \in \pi^{-1}(u)$ such that $x \in \operatorname{Inv} g$. Thus $\operatorname{Stab} x \supset (g)$. It follows that g has order ≤ 12 by the above considerations.

(β) If $u \in \pi(\operatorname{Inv} g)$, there exists $x \in \pi^{-1}(u)$ such that $x \in \operatorname{Inv} g$. Since $J(x)$ is nonconstant, $x = (a, b)$ with $a \neq 0$, $b \neq 0$. Suppose g were scalar with matrix $\left(\begin{smallmatrix}\lambda & \\ & \lambda\end{smallmatrix}\right)$. It follows that

$$ag = \lambda^4 a = a, \qquad bg = \lambda^6 b = b$$

which implies $\lambda^4 = \lambda^6 = 1$. Thus $\lambda^2 = 1$, which contradicts the fact that $g \neq 1$ in G. Consequently g is nonscalar.

(γ) Given g of finite order there exists $g' \in (g)$ of prime order such that

$$\pi(\operatorname{Inv} g) \subset \pi(\operatorname{Inv} g').$$

(δ) Given g of finite order there exists a diagonal element g' conjugate to g such that

$$\pi(\operatorname{Inv} g) = \pi(\operatorname{Inv} g').$$

(ε) If (g) is conjugate to (g'), then

$$\pi(\operatorname{Inv} g) = \pi(\operatorname{Inv} g').$$

We conclude from (α) to (ε) that

$$u \in \bigcup \pi(\operatorname{Inv} g),$$

where g runs through a system of representatives of the equivalence classes of nonscalar diagonal elements of G of prime order ≤ 12 with respect to the equivalence relation between elements g, g' of

G defined as follows. We say that g is *equivalent* to g' if (g) is conjugate to (g').

Let g be of prime order $p \leq 12$ with matrix $\begin{pmatrix}\lambda_1 & \\ & \lambda_2\end{pmatrix}$, $\lambda_1 \neq \lambda_2$. In case $p = 2$, we have either $\lambda_1^2 = \lambda_2^2 = 1$ or $\lambda_1^2 = \lambda_2^2 = -1$. Thus g is equivalent to one of $\begin{pmatrix}-1 & \\ & 1\end{pmatrix}$, $\begin{pmatrix}-i & \\ & i\end{pmatrix}$. In case p is odd, suppose first that $\lambda_1^p = \lambda_2^p = 1$. If $\lambda_2 = 1$, then $\lambda_1 \neq 1$. There exists an integer μ such that

$$\begin{pmatrix}\lambda_1 & \\ & \lambda_2\end{pmatrix}^{\mu} = \begin{pmatrix}\zeta_p & \\ & 1\end{pmatrix}.$$

Thus g is equivalent to $\begin{pmatrix}\zeta_p & \\ & 1\end{pmatrix}$. The case $\lambda_1 = 1$ reduces to the previous one by conjugation with the matrix

$$\begin{pmatrix} & 1 \\ 1 & \end{pmatrix}.$$

If $\lambda_1 \neq 1$, $\lambda_2 \neq 1$, there exists an integer μ such that

$$\begin{pmatrix}\lambda_1 & \\ & \lambda_2\end{pmatrix}^{\mu} = \begin{pmatrix}\lambda_1^{\mu} & \\ & \zeta_p\end{pmatrix}.$$

For some integer $l \neq 0, 1$

$$\begin{pmatrix}\lambda_1 & \\ & \lambda_2\end{pmatrix}^{\mu} = \begin{pmatrix}\zeta_p^l & \\ & \zeta_p\end{pmatrix}.$$

Thus g is equivalent to

$$\begin{pmatrix}\zeta_p^l & \\ & \zeta_p\end{pmatrix}, \qquad l \neq 0, 1.$$

When $\lambda_1^p = \lambda_2^p = -1$, set $\lambda_i' = -\lambda_i$, $i = 1, 2$ and reduce to the previous case.

The proof of Lemma 2 is finished by the observation that whenever $m \cdot l = 1 \pmod{p}$,

$$\begin{pmatrix}\zeta_p^m & \\ & \zeta_p\end{pmatrix} \text{ is equivalent to } \begin{pmatrix}\zeta_p^l & \\ & \zeta_p\end{pmatrix}. \qquad \square$$

For $g \in G$, we have

$$\operatorname{Inv} g = (V_4 \times V_6)^g \cap X = (V_4^g \times V_6^g) \cap X.$$

Let g be diagonal. The g-invariant monomials of V_n form a vector basis of V_n^g. Thus a general element of V_n^g is given by a linear combination with general coefficients of elements from such a basis. A general element of $V_4^g \times V_6^g$ is just a pair of general elements of V_4^g and V_6^g. Such a general element is also a general element of $\operatorname{Inv} g$ since X is open. It is stable if some specialization is stable.

Now we choose the following R_p for $p = 3, 5, 7, 11$:

$$\begin{aligned} R_3 &= \{2\}, \\ R_5 &= \{2, 4\}, \\ R_7 &= \{2, 3, 6\}, \\ R_{11} &= \{2, 3, 5, 7, 10\}. \end{aligned}$$

In Table 1 we give bases of g-invariant monomials of V_4^g and V_6^g for the different values of g that appear in Lemma 2 subject to the above choice of R_p's, except for the cases $g = i$, ζ_3 which are trivial. We also indicate for which values of g the set $\operatorname{Inv} g$ is nonempty.

TABLE 1. *Invariant Monomials.* We list all g-invariant monomials of degrees 4 and 6

p	g	degree 4	degree 6	g-invariant pairs
2	$\begin{pmatrix}-1 & \\ & 1\end{pmatrix}$	t^4, t^2s^2, s^4	t^6, t^4s^2, t^2s^4, s^6	some stable
	$\begin{pmatrix}-i & \\ & i\end{pmatrix}$	t^4, t^2s^2, s^4	t^5s, t^3s^3, ts^5	some stable
3	$\begin{pmatrix}\zeta_3 & \\ & 1\end{pmatrix}$	t^3s, s^4	t^6, t^3s^3, s^6	some stable
	$\begin{pmatrix}\zeta_3^2 & \\ & \zeta_3\end{pmatrix}$	t^2s^2	t^6, t^3s^3, s^6	some stable
5	$\begin{pmatrix}\zeta_5 & \\ & 1\end{pmatrix}$	s^4	t^5s, s^6	some stable
	$\begin{pmatrix}\zeta_5^2 & \\ & \zeta_5\end{pmatrix}$	ts^3	t^4s^2	some stable
	$\begin{pmatrix}\zeta_5^4 & \\ & \zeta_5\end{pmatrix}$	s^4	s^6	all unstable
7	$\begin{pmatrix}\zeta_7 & \\ & 1\end{pmatrix}$	s^4	s^6	all unstable
	$\begin{pmatrix}\zeta_7^2 & \\ & \zeta_7\end{pmatrix}$	t^3s	ts^5	some stable
	$\begin{pmatrix}\zeta_7^3 & \\ & \zeta_7\end{pmatrix}$	No solutions	No solutions	
	$\begin{pmatrix}\zeta_7^6 & \\ & \zeta_7\end{pmatrix}$	t^2s^2	t^3s^3	all semistable
11	$\begin{pmatrix}\zeta_{11}^{10} & \\ & \zeta_{11}\end{pmatrix}$	t^2s^2	t^3s^3	all semistable
No other solutions for $p = 11$.				

TABLE 2. *Components of* S. Here Γ is the component $\pi(\operatorname{Inv} g)$, $x = (a, b)$ is an element of the general orbit over Γ

Γ	g	$x = (a, b)$	$\operatorname{Stab} x$	$\deg J(x)$	$\dim \Gamma$
A	$\begin{pmatrix} -1 & \\ & 1 \end{pmatrix}$	$\begin{cases} a = (t^2 - s^2)(t^2 - k^2s^2) \\ b = c(t^2 - m^2s^2)(t^2 - n^2s^2)(t^2 - p^2s^2) \end{cases}$	C_2	12	5
B	$\begin{pmatrix} -i & \\ & i \end{pmatrix}$	$\begin{cases} a = (t^2 - s^2)(t^2 - k^2s^2) \\ b = cts(t^2 - m^2s^2)(t^2 - n^2s^2) \end{cases}$	C_2	12	4
C	$\begin{pmatrix} \zeta_3 & \\ & 1 \end{pmatrix}$	$\begin{cases} a = (t^3 - s^3)s \\ b = c(t^3 - m^3s^3)(t^3 - n^3s^3) \end{cases}$	C_3	12	3
D	$\begin{pmatrix} \zeta_5 & \\ & 1 \end{pmatrix}$	$\begin{cases} a = s^4 \\ b = c(t^5 - s^5)s \end{cases}$	C_5	10	1
E	$\begin{pmatrix} \zeta_5^2 & \\ & \zeta_5 \end{pmatrix}$	$\begin{cases} a = ts^3 \\ b = t^4s^2 \end{cases}$	C_5	5	0
F	$\begin{pmatrix} \zeta_7^2 & \\ & \zeta_7 \end{pmatrix}$	$\begin{cases} a = t^3s \\ b = ts^5 \end{cases}$	C_7	7	0
G	$\begin{pmatrix} \zeta_3 & \\ & \zeta_3 \end{pmatrix}$	$\begin{cases} a = 0 \\ b = t(t-s)s(t-ms)(t-ns)(t-ps) \end{cases}$	C_3	$J \equiv 0$	3

THEOREM 1. *The irreducible components of* S *are listed in Table* 2. *Suppose* g *and* x *are entries in a row of Table* 2 *with* x *an element of the general orbit over* $\Gamma = \pi(\operatorname{Inv} g)$. *Then* $\operatorname{Stab} x = (g)$.

In Table 2 the element $x = (a, b)$ is obtained by taking a general element of $\operatorname{Inv} g$ constructed from Table 1 and eliminating parameters redundant with respect to the action of G. The resulting parameters are chosen in such a way as to make explicit the zeros of a and b.

Keeping in mind the remark after Lemma 1, we first prove that the sets $\pi(\operatorname{Inv} g)$ for g in Table 2 are an irredundant decomposition of S. Among the $\pi(\operatorname{Inv} g)$ in Lemma 2 the following inclusions hold:

$$(\alpha) \qquad \pi \operatorname{Inv}\begin{pmatrix} i & \\ & i \end{pmatrix} \subset \pi \operatorname{Inv}\begin{pmatrix} -1 & \\ & 1 \end{pmatrix},$$

$$(\beta) \qquad \pi \operatorname{Inv}\begin{pmatrix} \zeta_3^2 & \\ & \zeta_3 \end{pmatrix} \subset \pi \operatorname{Inv}\begin{pmatrix} -1 & \\ & 1 \end{pmatrix}.$$

(α) By putting $c = 0$ in the element of the general orbit over $\pi \operatorname{Inv}\left(\begin{smallmatrix} -1 & \\ & 1 \end{smallmatrix}\right)$ we get

$$\begin{cases} a = (t^2 - s^2)(t^2 - k^2s^2), \\ b = 0. \end{cases}$$

By dimension considerations we get that the locus $J \equiv 1$ which equals $\pi \operatorname{Inv}\left(\begin{smallmatrix} i & \\ & i \end{smallmatrix}\right)$ is contained in $\pi \operatorname{Inv}\left(\begin{smallmatrix} -1 & \\ & 1 \end{smallmatrix}\right)$.

(β) Since $\left(\begin{smallmatrix} & 1\\ 1 & \end{smallmatrix}\right)$ is conjugate to $\left(\begin{smallmatrix} -1 & \\ & 1\end{smallmatrix}\right)$ we have

$$\pi \operatorname{Inv}\begin{pmatrix} & 1\\ 1 & \end{pmatrix} = \pi \operatorname{Inv}\begin{pmatrix} -1 & \\ & 1\end{pmatrix}.$$

But the general element of $\operatorname{Inv}\left(\begin{smallmatrix} \zeta_3^2 & \\ & \zeta_3\end{smallmatrix}\right)$ is invariant under $\left(\begin{smallmatrix} & 1\\ 1 & \end{smallmatrix}\right)$.

To check that there are no other inclusions we use systems of eigenvalues. More precisely, let R be the equivalence relation on $\mathbf{C}^2$ generated by the relations "$(x', y') = (y, x)$" and "$(x', y') = (-x, -y)$" both between the elements (x, y) and (x', y') of $\mathbf{C}^2$. Thus the class of (x, y) consists of the elements (x, y), (y, x), $(-x, -y)$ and $(-y, -x)$. The system of eigenvalues of an element of finite order $g \in G$ will be considered as an element of $\mathbf{C}^2/R$.

For a subgroup of finite order H of G we denote by $\operatorname{Eigenval}(H)$ the set of systems of eigenvalues of elements of H. Clearly $\operatorname{Eigenval}(H)$ depends only on the conjugacy class of H.

Let g_1, g_2 appear in Table 2 and suppose that $\pi(\operatorname{Inv} g_1) \subset \pi(\operatorname{Inv} g_2)$. Suppose x_1 is the element of the general orbit over $\pi(\operatorname{Inv} g_1)$, given in Table 2. Then $\pi(x_1) \in \pi(\operatorname{Inv} g_2)$, which implies that $x_1 \in \operatorname{Inv} h^{-1} g_2 h$ for some h. By (ii) it follows that

$$(g_1) = \operatorname{Stab} x_1 \supset (h^{-1} g_2 h).$$

Thus $\operatorname{Eigenval}(g_1) \supset \operatorname{Eigenval}(g_2)$. The reader can easily check case by case that this can only happen when $g_1 = g_2$.

Now it remains to prove that $\operatorname{Stab} x \subset (g)$, for g, x satisfying the conditions of the theorem. The inclusion $(g) \subset \operatorname{Stab} x$ is obvious. In cases D, E, F, suppose $h = \left(\begin{smallmatrix} \alpha & \beta\\ \gamma & \delta\end{smallmatrix}\right) \in \operatorname{Stab} x$, $x = (a, b)$. By comparing coefficients in the equations $ah = a$ and $bh = b$, one concludes that $h \in (g)$. The remaining cases depend on a series of lemmas.

As usual, we identify $\mathbf{P}^1$ with $\mathbf{C} \cup \{\infty\}$ and automorphisms of $\mathbf{P}^1$ with linear fractional transformations. Moreover, given a set E of $n \geq 3$ distinct points of $\mathbf{P}^1$, every automorphism of $\mathbf{P}^1$ stabilizing the set E is determined by the induced permutation of E. We indicate such automorphisms by giving only the induced permutation. We omit the proof of the following well-known lemmas.

LEMMA 3. *Every automorphism of* $\mathbf{P}^1$ *that permutes the points* 0, 1, ∞, m, n, p, *where* m, n, p *are in general position, is the identity.*

LEMMA 4. *The group of automorphisms of* $\mathbf{P}^1$ *that permute the points* 0, 1, ∞, k, *for general* k *is Klein's four-group consisting of* $(0\ 1)(\infty\ k)$, $(0\ \infty)(1\ k)$, $(1\ \infty)(0\ k)$ *and the identity* e.

LEMMA 5. *The group of automorphisms of* $\mathbf{P}^1$ *that permute the points* 1, ζ_3, ζ_3^2, ∞ *is the tetrahedral group.*

LEMMA 6. *The group of automorphisms of* $\mathbf{P}^1$ *that permute the points* 1, -1, k, $-k$, *for general* k *is Klein's four-group consisting of* $(1\ -1)(k\ -k)$, $(1\ k)(-1\ -k)$, $(1\ -k)(-1\ k)$ *and the identity* e.

Now we return to the proof of Theorem 1. We omit Case B because it is similar to case A. We treat case G first.

Case G. Suppose $bh = b$ for $b = t(t-s)s(t-ms)(t-ns)(t-ps)$, m, n, p in general position and $h \in G$. By Lemma 3, we infer that h has the matrix $\left(\begin{smallmatrix}\lambda & \\ & \lambda\end{smallmatrix}\right)$ since the automorphism of $\mathbf{P}^1$ induced by h permutes the zeros of b. Thus $\lambda^6 = 1$.

Case A. Suppose $ah = a$ for $a = (t^2 - s^2)(t^2 - k^2s^2)$, k general, $h \in G$. By Lemma 5 the linear fractional transformation Ph is one of $(1\ -1)(k\ -k)$, $(1\ k)(-1\ -k)$, $(1\ -k)(-1\ k)$ or the identity. But $(1\ k)(-1\ -k)$ and $(1\ -k)(-1\ k)$ cannot stabilize the set $\{m, -m, n, -n, p, -p\}$ for m, n, p in general position.

Case C. Suppose $ah = a$ for $h \in G$. By Lemma 6, Ph belongs to the tetrahedral group permuting the points 1, ζ_3, ζ_3^2, ∞, which is isomorphic to the alternating group of the set $\{1, \zeta_3, \zeta_3^2, \infty\}$. Taking into account the form of the elements of this group ([**Se**], p. 41), for m, n in general position the subgroup stabilizing the set of zeros of b is $((1\ \zeta_3\ \zeta_3^2))$. □

3. Singularities. In this section we prove that S is the singular locus of W and we determine the general singularities.

All the representations we consider in the following are finite dimensional linear representations over $\mathbf{C}$ of finite groups.

We need the notion of isomorphism of two representations $\rho\colon H \to \mathbf{GL}(V)$ and $\rho'\colon H' \to \mathbf{GL}(V')$ of not necessarily identical groups H, H'. The definition is obvious. The representation $\rho\colon H \to \mathbf{GL}(V)$ is called *small* if no element in the image of ρ has 1 as eigenvalue of multiplicity $\dim V - 1$. We gather in the following proposition the results we need from [**Pr**].

Proposition 2. *Two small faithful representations* $\rho\colon H \to \mathbf{GL}(V)$ *and* $\rho'\colon H' \to \mathbf{GL}(V')$ *are isomorphic if and only if the germs of analytic space* $(V/H, 0)$ *and* $(V'/H', 0)$ *are isomorphic.*

A small faithful representation $\rho\colon H \to \mathbf{GL}(V)$ *is identically equal to the identity if and only if* $(V/H, 0)$ *is nonsingular.*

Now let u be a point of W and x an element of X such that $u = \pi(x)$. Put $H = \operatorname{Stab} x$ and let N be an H-invariant complement to $T_x(xG)$ in $T_x(X)$. By the slice theorem ([**Sch**], p. 56) and the fact that π is affine, there exists an isomorphism of germs of analytic space $(W, u) \xrightarrow{\sim} (N/H, 0)$. Call $\rho = \rho_{x,N}$ the representation of H defined by its action on N and $\rho^f = \rho^f_{x,N}$ the faithful representation of $H/\operatorname{Ker}\rho$ induced by ρ. The isomorphism class of the germ (W, u) depends only on the isomorphism class of the representation ρ. We say that the representation $\rho = \rho_{x,N}$ is *associated* to the point $u = \pi(x)$.

Theorem 2. *The set* S *is the singular locus of* W. *Representations associated to the general singularities, which are given in Table* 3, *are faithful and small.*

The following corollary is immediate by Proposition 2.

Corollary. *The isomorphism classes of the associated representations classify the general singularities up to isomorphism.*

Table 3. *Associated Representations.* Here Γ is the component of $\pi(\operatorname{Inv} g)$, x is the element of the general orbit over Γ such that $\operatorname{Stab} x = (g)$, $\rho = \rho_{x,N}$ is a representation of $\operatorname{Stab} x$ associated to $u = \pi(x)$

Γ	$\operatorname{Stab} x$	Eigenvalues of $\rho(g)$							
A	C_2	$1,$	$1,$	$1,$	$1,$	$1,$	$-1,$	$-1,$	-1
B	C_2	$1,$	$1,$	$1,$	$1,$	$-1,$	$-1,$	$-1,$	-1
C	C_3	$1,$	$1,$	$1,$	$\zeta_3,$	$\zeta_3,$	$\zeta_3,$	$\zeta_3^2,$	ζ_3^2
D	C_5	$\zeta_5,$	$\zeta_5,$	$\zeta_5^2,$	$\zeta_5^2,$	$\zeta_5^2,$	$\zeta_5^3,$	$\zeta_5^3,$	ζ_5^4
E	C_5	$1,$	$\zeta_5,$	$\zeta_5,$	$\zeta_5^2,$	$\zeta_5^2,$	$\zeta_5^3,$	$\zeta_5^3,$	ζ_5^4
F	C_7	$\zeta_7,$	$\zeta_7^2,$	$\zeta_7^2,$	$\zeta_7^3,$	$\zeta_7^4,$	$\zeta_7^5,$	$\zeta_7^6,$	ζ_7^6
G	C_3	$1,$	$1,$	$1,$	$\zeta_3,$	$\zeta_3,$	$\zeta_3,$	$\zeta_3,$	ζ_3

Proof of Theorem 2. It is clear that the representations in Table 3 are faithful and small. We have to prove that they are associated to the general points of the components of S.

First of all we recall some generalities on infinitesimals of first order. Let X be an analytic space, $x \in X$. Let $\mathbf{C}[\varepsilon]$ be the algebra of dual numbers. Let $\operatorname{Specan}\mathbf{C}[\varepsilon]$ be the analytic space with only one point o and with local ring $\mathbf{C}[\varepsilon]$ at that point. We use the notation

$$\begin{aligned} X(\mathbf{C}[\varepsilon])_x &= \operatorname{Hom}((\operatorname{Specan}\mathbf{C}[\varepsilon], o), (X, x)) \\ &= \operatorname{Hom}_{\mathbf{C}-\mathrm{alg\,loc}}(\mathscr{O}_{X,x}, \mathbf{C}[\varepsilon]) \end{aligned}$$

for the set of $\mathbf{C}[\varepsilon]$-valued points of X at x.

The map

$$\begin{gathered} \operatorname{Der}_{\mathbf{C}}(\mathscr{O}_{X,x}, \mathbf{C}) \to \operatorname{Hom}_{\mathbf{C}-\mathrm{alg\,loc}}(\mathscr{O}_{X,x}, \mathbf{C}[\varepsilon]), \\ t \to u = x + \varepsilon t \end{gathered}$$

establishes a bijection of $T_x(X)$ onto $X(\mathbf{C}[\varepsilon])_x$.

Now denote by X the set of stable elements of $V_4 \times V_6$ and by G the group $\mathbf{GL}_2/(\pm I)$ as before. Let $x \in X$. The orbital map $\rho\colon G \to X$, $g \mapsto xg$ is étale because $\operatorname{Stab} x$ is a finite set. We get the following commutative diagram:

$$\begin{tikzcd} G(\mathbf{C}[\varepsilon])_1 \arrow[r, "\rho"] & X(\mathbf{C}[\varepsilon])_x \\ \mathbf{M}_2(\mathbf{C}) \arrow[u, "\varphi"] \arrow[r, "d\rho_1"'] & V_4 \times V_6 \arrow[u, "\psi"'] \end{tikzcd}$$

where we identify $T_1(G) = T_I(\mathbf{GL}_2) = \mathbf{M}_2(\mathbf{C}) = 2 \times 2$ matrices, $T_x(X) = T_x(V_4 \times V_6) = V_4 \times V_6$, $T_x(G) = \operatorname{Im} d\rho_1$ and φ and ψ are the bijections described above. Note that ρ and $d\rho_1$ are injective. We have a canonical basis $t_1, \dots, t_4$ of $\operatorname{Im} d\rho_1$ namely the image of the canonical basis

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad E_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

of $\mathbf{M}_2(\mathbf{C})$. It can be computed explicitly from the equation

$$x \cdot (I + \varepsilon E_i) = x + \varepsilon t_i$$

which follows from

$$(\rho \circ \varphi)(e_i) = x \cdot (I + \varepsilon E_i)$$

and

$$(\psi \circ d\rho_1)(e_i) = x + \varepsilon t_i .$$

Now let us explain how to choose N. Let $\theta\colon V_4 \times V_6 \xrightarrow{\sim} \mathbf{C}^{12}$ be the isomorphism defined by the canonical basis, $(t^4, 0), (t^3 s, 0), \ldots,$ $(s^4, 0), (0, t^6), \ldots,$ of $V_4 \times V_6$. Let $t_i = (\sum \alpha^i_{kl} t^k s^l)$, $(\sum \beta^i_{mn} t^m s^n)$. Let A be the matrix

$$\left(\begin{array}{c|c} \alpha^1_{4,0} \cdots \alpha^1_{0,4} & \beta^1_{6,0} \cdots \beta^1_{0,6} \\ \alpha^4_{0,4} \cdots \alpha^4_{0,4} & \beta^4_{6,0} \cdots \beta^4_{0,6} \end{array}\right) .$$

The row space of A is $\operatorname{Im} d\rho_1$. We choose a square submatrix B of A such that $\det B \neq 0$. The submatrix B is gotten from A by deleting a row $(\alpha^i_{k,l})$ (resp. a row $(\beta^i_{m,n})$) if and only if $(k, l) \in D$ (resp. $(m, n) \in E$) for well determined sets D, E. The subspace $N = N_B$ generated by $(t^k s^l, 0)$, $(k, l) \in D$ and $(0, t^m s^n)$, $(m, n) \in E$ is a complement of $\operatorname{Im} d\rho_1$. This is obvious by considering their images under θ.

To calculate the matrix A we use the following formulas, where $f = \sum a f_{ij} t^i s^j$ is an element of V_n.

$$\begin{aligned} f \cdot (I + \varepsilon E_1) &= f + \varepsilon \sum f_{ij} i t^i s^j , \\ f \cdot (I + \varepsilon E_2) &= f + \varepsilon \sum f_{ij} i t^{i-1} s^{j+1} , \\ f \cdot (I + \varepsilon E_3) &= f + \varepsilon \sum f_{ij} j t^{i+1} s^{j-1} , \\ f \cdot (I + \varepsilon E_4) &= f + \varepsilon \sum f_{ij} j t^i s^j . \end{aligned}$$

We indicate the explicit choice of the square submatrix B in each case by underlining the corresponding columns of A. The reader should keep in mind Table 2.

Case A. By setting

$$\begin{aligned} \kappa &= -(1 + k^2), \\ \lambda &= k^2, \\ \mu &= -(m^2 + n^2 + p^2), \\ \nu &= m^2 n^2 + m^2 p^2 + n^2 p^2, \\ \pi &= -m^2 n^2 p^2, \end{aligned}$$

the general element x can be written

$$\begin{cases} a = t^4 + \kappa t^2 s^2 + \lambda s^4 , \\ b = c(t^6 + \mu t^4 s^2 + \nu t^2 s^4 + \pi s^6) . \end{cases}$$

The matrix A has the form

$$\left[\begin{array}{ccccc|ccccccc} 4 & 0 & 2\kappa & 0 & 0 & 6c & 0 & 4\mu c & 0 & 2\nu c & 0 & 0 \\ 0 & 4 & 0 & 2\kappa & 0 & 0 & 6c & 0 & 4\mu c & 0 & 4\nu c & 0 \\ 0 & 2\kappa & 0 & 4\lambda & 0 & 0 & 2\mu c & 0 & 4\nu c & 0 & 6\pi c & 0 \\ \underline{0} & \underline{0} & \underline{2\kappa} & \underline{0} & 4\lambda & 0 & 0 & 2\mu c & 0 & 4\nu c & 0 & 6\pi c \end{array}\right]$$

and the submatrix B consists of the underlined columns. It is clear that $\det B \neq 0$ for k, m, n, p general enough.

The action of $g = \left(\begin{smallmatrix} -1 & \\ & 1 \end{smallmatrix}\right)$ on N_B is given by

$$\begin{array}{l} s^4;\ t^6,\ t^5s,\ t^4s^2,\ t^3s^3,\ t^2s^4,\ ts^5,\ s^6 \\ +1,\ +1,\ -1,\ +1,\ -1,\ +1,\ -1,\ +1 \end{array}$$

where the first line is the canonical basis of N_B which consists of eigenvectors of g and the second line are the corresponding eigenvalues.

In the following cases we just indicate the matrices A, B.

Case B. Here

$$\kappa = -(1+k^2), \quad \lambda = k^2, \quad \mu = -(m^2+n^2), \quad \nu = m^2n^2.$$

$$\left[\begin{array}{ccccc|ccccccc} 4 & 0 & 2\kappa & 0 & 0 & 0 & 6c & 0 & 3\mu c & 0 & \nu c & 0 \\ 0 & 4 & 0 & 2\kappa & 0 & 0 & 0 & 6c & 0 & 3\mu c & 0 & \nu c \\ 0 & 2\kappa & 0 & 4 & 0 & c & 0 & 3\mu c & 0 & 5\nu c & 0 & 0 \\ \underline{0} & 0 & 2\kappa & 0 & \underline{4} & \underline{0} & c & 0 & 3\mu c & 0 & 5\nu c & \underline{0} \end{array}\right].$$

Case C. Here

$$\mu = (m^3+n^3), \qquad \nu = m^3n^3.$$

$$\left[\begin{array}{ccccc|ccccccc} 0 & 1 & 0 & 0 & -4 & 0 & 0 & 0 & 3\mu c & 0 & 0 & 6\nu c \\ 0 & 3 & 0 & 0 & 0 & 6c & 0 & 0 & 3\mu c & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 6c & 0 & 0 & 3\mu c & 0 & 0 \\ 1 & 0 & 0 & -4 & 0 & \underline{0} & \underline{0} & \underline{3\mu c} & 0 & 0 & 6\nu c & \underline{0} \end{array}\right].$$

Case D.

$$\left[\begin{array}{ccccc|ccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 5c & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 5c & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & c & 0 & 0 & 0 & 0 & -6c & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & \underline{c} & \underline{0} & 0 & 0 & \underline{0} & -\underline{6c} \end{array}\right].$$

Case E.

$$\left[\begin{array}{ccccc|ccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \underline{0} & \underline{3} & \underline{0} & 0 & 0 & \underline{2} & 0 & 0 & 0 & 0 \end{array}\right].$$

Case F.

$$\left[\begin{array}{ccccc|ccccccc} 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 & 0 & 0 \\ \underline{0} & \underline{1} & \underline{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \underline{5} & 0 \end{array}\right].$$

Case G. Here

$$\begin{aligned} \mu &= -1 - m - n - p, \\ \nu &= m + n + p + mn + mp + np, \\ \pi &= pmn - mp - np - mnp, \\ \rho &= mnp, \end{aligned}$$

$$\left[\begin{array}{c|ccccccc} & 0 & 5 & 4\mu & 3\nu & 2\pi & \rho & 0 \\ 0 & 0 & 0 & 5 & 4\mu & 3\nu & 2\pi & \rho \\ & 0 & \mu & 2\nu & 3\pi & 4\rho & 0 & 0 \\ & 0 & \underline{0} & \mu & 2\nu & \underline{3\mu} & \underline{4\rho} & \underline{0} \end{array}\right].$$

The reader can check by specialization that the underlined matrix B has $\det B \neq 0$ for m, n, p general enough.

REFERENCES

[Bo] A. Borel, *Linear Algebraic Groups*, Benjamin, 1969.

[Br] E. Brieskorn, *Rationale Singularitäten komplexer Flächen*, Invent. Math., **4** (1968), 336–358.

[Fu] W. Fulton, *Hurwitz schemes and irreducibility of moduli of algebraic curves*, Ann. of Math., **90** (1969), 542–575.

[H-M] J. Harris and D. Mumford, *On the Kodaira dimension of the moduli space of curves*, Invent. Math., **67** (1982), 23–86.

[Ho] H. Holman, *Quotienten komplexer Räumen*, Math. Ann., **142** (1961), 407–440.

[Hu] A. Hurwitz, *Über Riemansche Flächen mit gegebenen Verzweigungspunkten*, Math. Ann., **39** (1981), 1–61.

[Ka] A. Kas, *Weierstrass normal forms and invariants of elliptic surfaces*, Trans. Amer. Math. Soc., **225** (1977), 259–266.

[Kau] B. Kaup, *Äquivalenzrelationen auf allgemeinen komplexen Räumen*, Schriftenreihe Math. Institut Univ. Münster, Heft **39** (1968).

[Ko1] K. Kodaira, *On compact analytic surfaces*, II Ann. of Math., **77** (1963), 563–626.

[Ko2] ——, *On compact analytic surfaces*, III Ann. of Math., **78** (1963), 1–4.

[Le1] P. Lejarraga, *The moduli of elliptic surfaces*, thesis, Brandeis, 1985.

[Le2] ——, *The moduli of rational Weierstrass fibrations over* $\mathbf{P}^1$: *stratification*, Comm. Algebra, **20** (6), (1992), 1557–1590.

[Le3] ——, *The moduli of Weierstrass fibrations over* $\mathbf{P}^1$: *rationality*, to appear in The Rocky Mountain J. Math.

[Mi] R. Miranda, *The moduli of Weierstrass fibrations over* $\mathbf{P}^1$, Math. Ann., **255** (1981), 379–394.

[GIT] D. Mumford, *Geometric Invariant Theory*, Springer, 1963.

[M-S] D. Mumford and K. Suominen, *Introduction to the theory of moduli*, 5th Nordic Summer School in Mathematics, Oslo, 1970.

[Ne] A. Néron, *Modèles minimaux de variétés abéliennes sur les corps locaux et globaux*, Pub. Math. IHES, 1964.

[Pr] D. Prill, *Local classification of quotients of complex manifolds by discontinuous groups*, Duke Math. J., **34** (1967), 375–386.

[Sch] G. Schwartz, *Lifting Smooth Homotopies of Orbit Spaces*, Publ. Math. IHES, No. 51, (1980), 37–135.

[Sei1] W. K. Seiler, *Moduli elliptischer Flächen mit Schnitt*, thesis, Universität Karlsruhe, 1982.

[Sei2] ——, *Global moduli for elliptic surfaces with a section*, Compositio Math., **62** (1987), 169–185.

[Sei3] ——, *Global moduli for polarized elliptic surfaces*, Compositio Math., **62** (1987), 187–213.

[Sei4] ——, *Deformations of Weierstrass elliptic surfaces*, Math. Ann., **281** (1988), 263–278.

[Se] J. P. Serre, *Linear Representations of Finite Groups*, Springer, 1977.

Received February 12, 1991.

St. John's College
Annapolis, MD 21404
E-mail address: pablo@math2.sma.usna.navy.mil

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

ON DISCRETE ISOMETRY GROUPS OF NEGATIVE CURVATURE

GAVEN J. MARTIN

In this paper we extend well-known results concerning the algebraic limits and deformations of groups of hyperbolic isometries of hyperbolic 3-space, $\mathbf{H}^3$, to *negatively curved groups*. For us these will be groups of isometries of variable negative curvature metrics satisfying a pinching condition and in particular will include the $\mathbb{R}$-rank one Lie groups. We accomplish these goals, as in the hyperbolic case, by producing a version of Jørgensen's inequality for such groups. Using an appropriate normalisation we can consider algebraic limits and deformations of such groups in the homeomorphism group of the n-ball, $\operatorname{Hom}(\mathbf{B}^n)$. We ask that the generators of each group move continuously or some sequence of generators have limits in $\operatorname{Hom}(\mathbf{B}^n)$, but there is no such restriction on the associated negatively curved metrics. We then recover many of the standard results for groups of hyperbolic isometries of $\mathbf{H}^3$ in this more general setting under mild and usually necessary restrictions, such things as the limits being discrete, or the deformations are algebraically trivial and so forth.

We point out, as a warning, that some authors use the term negatively curved groups to mean hyperbolic groups in the sense of Gromov [**G**].

The pinching condition we assume may be relaxed if the curvature is nonpositive, bounded below and the associated Hadamard manifold is a visibility manifold [**EO**]. The version of Jørgensen's inequality [**J**] that we produce follows more or less directly as in [**M**], where the n-dimensional hyperbolic case is considered, from Gromov's generalization of the convergence of iterated commutators, see Buser and Karcher [**BK**, §2]. The results of this paper can be viewed as further applications of that fundamental result. Other aspects of the theory of isometry groups of negative curvature can be found in the book of Ballman, Gromov and Schroeder [**BGS**] which we use as a general reference.

In §§3, 4 we discuss the algebraic limits of negatively curved groups (in the appropriate homeomorphism group) and show that under mild (and necessary) hypotheses they are discrete. We also obtain that a group of isometries of a negatively curved metric is discrete if and only if all its two generator subgroups are discrete under either the algebraic assumption that the order of the torsion is bounded, or the geometric assumption that the fixed points of hyperbolics are in general position. It is not difficult to see that some assumption is necessary. Our result includes those negatively curved groups which, as abstract groups, are hyperbolic in the sense of Gromov.

In §5 we show that continuous deformations of most (for instance those that are virtually torsion free) discrete negatively curved groups are algebraically trivial; that is, the deformation consists entirely of isomorphic groups. (Here we vary the generators continuously in the homeomorphism group, the underlying metrics need not change continuously.) From this it turns out that a continuous deformation of a cocompact torsion free group is topologically trivial; that is, all the groups are topologically conjugate (including the action on the sphere at infinity).

There are three main reasons why our results (except in the $\mathbb{R}$-rank one case where more precise statements can be made) are not as good as the hyperbolic case:

First, in this general setting there is no Selberg Lemma [**S**] asserting the existence of torsion free subgroups of finite index in finitely generated matrix groups. Indeed there is an example, due to Gromov [**G**, §4.5C], of a finitely generated infinite torsion group acting on a space of nonpositive curvature (this space is not a manifold). He remarks, page 79, that it might be possible to achieve a uniform bound on the order of the torsion and thus produce a geometric Burnside group. It is true however that a purely torsion negatively curved group is finite. This is a simple consequence of [**G**, Lemma 8.1 A].

Secondly, there is no known compact core theorem, as for instance Scott's Core Theorem for 3-manifolds [**Sc**]. Thus we cannot assume that a finitely generated group is finitely presented. Indeed, there are finitely generated discrete hyperbolic groups (in dimension $n \geq 4$) which are not finitely presented [**KP**]. We point out that there is a useful version of the core theorem if one assumes pinched curvature and the injectivity radius goes to zero [**BGS**, Theorem 10.5].

And thirdly, in [**GM**] we construct an infinite parabolic convergence group which contains free groups of arbitrarily large rank and so is not virtually nilpotent. The Heisenberg group is a purely parabolic group of isometries of complex hyperbolic space which is not virtually abelian. Consequently, parabolicity is not as easy a condition to deal with as it is in the constant curvature case. It should be noted, however, that a finitely generated discrete parabolic negatively curved group is virtually nilpotent; see for instance **[B]** who shows that discreteness implies finite generation in this case.

Finally we remark that it is an interesting feature of our work that the underlying metrics do not play a crucial role in the discreteness of limits. The dependence is more on the topological action of such groups on the sphere at infinity (as convergence groups in the sense of **[GM]**) which enables us to produce the appropriate version of Jørgensen's inequality. The algebraic limits of discrete negatively groups may only be groups of homeomorphisms, but our results imply they are discrete and have reasonable geometric structure.

1. Notation and Definitions. We denote by $\mathbf{B}^n$ the closed unit ball of euclidean n-space $\mathbb{R}^n$. A *Hadamard manifold* M is a complete simply connected manifold all of whose Riemannian sectional curvatures $K(M) \leq 0$. Every n-dimensional Hadamard manifold is diffeomorphic to $\mathrm{int}(\mathbf{B}^n)$ via the usual exponential mapping (Cartan-Hadamard Theorem). The universal covering of any manifold whose sectional curvatures are nonpositive is a Hadamard manifold. Two unit speed geodesics c_1, c_2 in M are *asymptotic* if there is some constant b such that $d(c_1(t), c_2(t)) \leq b$ for all $t \geq 0$. The equivalence classes of asymptotic geodesics are called *points at infinity* and the collection of all such points is denoted S_∞. There is a natural topology on $M \cup S_\infty$ which makes it homeomorphic to $\mathbf{B}^n$ **[EO]**. If the sectional curvatures are strictly negative, $K(M) \leq -a^2 < 0$, then M is a *visibility manifold* in the sense of **[EO]**. Then for distinct x and y in S_∞ there is a unique geodesic $c\colon \mathbb{R} \to M$ such that $c(+\infty) = x$ and $c(-\infty) = y$. We say that a Hadamard manifold M has *pinched curvature* if all of the sectional curvatures of M satisfy

$$-1 \leq K(M) \leq -a^2,$$

where $a \neq 0$. The choice of -1 on the left-hand side of the above inequality is a normalization which can be achieved by scalar multiplication of the Riemannian metric as soon as the sectional curvatures satisfy $-B^2 \leq K(M) \leq -b^2$. It is easy to see that any isometry of a pinched Hadamard manifold M extends to a homeomorphism of $M \cup S_\infty$ via its action on geodesics.

It is clear from the above discussion that the study of pinched Hadamard manifolds and their isometries is equivalent to the study of Riemannian metrics g defined on the open unit ball of $\mathbb{R}^n$ and for which all the sectional curvatures satisfy $-1 \leq K(g) \leq -a^2$. For simplicity, this is the framework in which we will work. We say a group Γ of homeomorphisms of $\mathbf{B}^n$ is a *negatively curved group* if there is some metric g satisfying the above pinching condition and for which Γ acts as a group of isometries.

We say that a negatively curved group Γ is *discrete* if the identity is isolated in Γ (the topology here is essentially irrelevant). Discrete negatively curved groups arise naturally in the study of manifolds with pinched curvature. If (M, g) is a Riemannian manifold with pinched sectional curvature, then the fundamental group $\pi_1(M)$ is a negatively curved discrete group. Discrete subgroups of $\mathbb{R}$-rank one Lie groups are negatively curved groups. Let γ be an isometry of a pinched curvature metric g with associated distance function $d(\cdot, \cdot)$. For $x \in \operatorname{int}(\mathbf{B}^n)$ we define $d_\gamma(x) = d(\gamma(x), x)$ and $d_\gamma = \inf\{d_\gamma(x) \colon x \in \operatorname{int}(\mathbf{B}^n)\}$; d_γ is called the *translation length* of γ. The isometry γ is classified by its translation length. We say γ is *elliptic* if $d_\gamma = 0$ and this infimum is attained, γ is *hyperbolic* if $d_\gamma > 0$ and this infimum is attained, and γ is *parabolic* if the infimum is not obtained. If γ is elliptic, then the fixed point set of γ in $\operatorname{int}(\mathbf{B}^n)$ is a nonempty complete totally geodesic subspace of codimension at least 1 (actually 2 if γ is orientation preserving). If γ is hyperbolic, then γ has two fixed points on the sphere at infinity. The unique geodesic connecting these two points is completely invariant and γ translates along the geodesic by the distance d_γ. If γ is parabolic, then γ has a single fixed point on the sphere at infinity [**BGS**]. This characterization implies the notion of hyperbolicity, parabolicity and ellipticity are well defined for negatively curved groups. That is, they are independent of the underlying metric (of which there may be many).

In [**MS**] we prove that a discrete negatively curved group is a *convergence group* in the sense of [**GM**]. These are groups of homeomorphisms with the compactness properties of quasiconformal (and hence conformal) mappings. More precisely a group Γ of homeomorphisms of $\mathbf{B}^n$ is called a convergence group if every infinite family of elements of Γ contains a sequence $\{\gamma_i\}_{i \geq 1}$ for which there are x and y in $\mathbf{S}^{n-1}$ such that

$$\gamma_i \to x \text{ locally uniformly in } \mathbf{B}^n - \{y\}$$

and

$$\gamma_i^{-1} \to y \text{ locally uniformally in } \mathbf{B}^n - \{x\},$$

the possibility $x = y$ may occur. This definition is actually that of a discrete convergence group in our reference [**GM**], but we adopt this terminology for simplicity. Note too that in accordance with the usual terminology for isometries of negatively curved spaces we use the term hyperbolic for loxodromic elements.

The *limit set*, $L(\Gamma)$, of a convergence group Γ is the set of accumulation points of $\Gamma(0) = \{\gamma(0) \colon \gamma \in \Gamma\}$. We say that Γ is *elementary* if the limit set contains fewer than three points. Otherwise the limit set is a perfect set, and the fixed points of hyperbolic elements are pairwise dense in $L(\Gamma) \times L(\Gamma)$. In **[GM]** we classify the elementary convergence groups. The elements in a convergence group are also classified according to their fixed point data and their order. In the case of a discrete group of isometries of a pinched metric, the above implies that the two classifications correspond.

Let g be a negatively curved Riemannian metric on $\operatorname{int}(\mathbf{B}^n)$, $d(\cdot, \cdot)$ the associated distance function, γ an isometry and $x \in \operatorname{int}(\mathbf{B}^n)$. Then the *rotation* of γ at x is the angle

$$r_\gamma(x) = \max\{\angle(w, P_\gamma(\gamma_{*x}(w))) \colon w \in T_x\mathbf{B}^n \text{ and } |w| = 1\}.$$

Here $\gamma_{*x} \colon T_x\mathbf{B}^n \to T_{\gamma(x)}\mathbf{B}^n$ is the differential at x and $P_\gamma \colon T_{\gamma(x)}\mathbf{B}^n \to T_x\mathbf{B}^n$ is parallel transportation along the geodesic from x to $\gamma(x)$. We then define the *norm* of γ at x as

$$n_\gamma(x) = \max\{r_\gamma(x), 8\, d_\gamma(x)\}.$$

One of the most important results we use in the following consequence of the generalization (due to Gromov) of the convergence of iterated commutators in Lie groups, see Corollary 2.4.4 of **[BK]**. This result can be found in **[BGS**, Corollary p. 106].

1.1. THEOREM. *Let g be a pinched Riemannian metric, $x \in \operatorname{int}(\mathbf{B}^n)$ and N a discrete group of isometries generated by elements α with $n_\alpha(x) \leq 0.49$. Then N is nilpotent.*

It is interesting to note that the constant 0.49 is independent of dimension. In the three-dimensional hyperbolic case, a related result can be proved purely in terms of the translation length **[GM2**, §5].

It is easy to prove that a nonelementary convergence group contains a subgroup isomorphic to the free group on two generators and is therefore not nilpotent. Thus in Theorem 1.1 we could have arranged the (weaker) conclusion that N is elementary and so in particular the limit set consists of at most two points.

2. Jørgensen's inequality for negative curvature. In this section we prove a generalization of Jørgensen's inequality for Kleinian groups

[J]. This result can be seen as a generalization of the classical Margulis lemma, or the existence of Zassenhaus neighbourhoods. First we need a couple of lemmas. We assume that γ is an isometry of a negatively curved metric. A *line* is a complete geodesic. We fix $x \in \mathrm{int}(\mathbf{B}^n)$.

2.1. LEMMA. *If γ interchanges the endpoints of a line, then γ is elliptic and $n_\gamma(x) \geq \pi/2$.*

Proof. If γ interchanges the endpoints of a line, then γ^2 has two additional fixed points and hence is elliptic. Thus γ too is elliptic. Next, γ has a fixed point on this line and so the set V of perpendicular geodesics to this line and passing through the fixed point is invariant and separates $\mathbf{B}^n$ into two components (if there were more components we would find conjugate points). One of these components V^+ contains x, the other V^- contains $\gamma(x)$. Let C be the geodesic joining x to $\gamma(x)$ and y the point of intersection of C and V. Let w be the unit tangent vector at x whose parallel transport along C to y is perpendicular to V and points into V^+. Then $\gamma *_x (w)$ points in the direction of V^- as does its parallel transport to y (note that $\gamma^2(x) \in V^+$). Thus the angle between the parallel transports of w and $\gamma *_x (w)$ to y is at least $\pi/2$. This establishes the lemma.

The following two lemmas are proved in a manner similar to Lemmas 4.1 and 4.2 of [M]. We illustrate the proof of the second lemma.

2.2. LEMMA. *Suppose α and β are two isometries generating a discrete negatively curved group. If α is parabolic or hyperbolic and if the group $\langle \alpha, \beta\alpha\beta^{-1} \rangle$ is elementary, then the group $\langle \alpha, \beta \rangle$ is also elementary.*

2.3. LEMMA. *Let α and β be two isometries generating a discrete negatively curved group. Suppose that α is elliptic and that p is the dimension of the fixed point set of α. If the group*

$$\Gamma = \langle \beta^i \alpha \beta^{-i} \colon i = 0, 1, 2, \ldots, p+1 \rangle$$

is elementary, then either $\langle \alpha, \beta \rangle$ is elementary or $n_\alpha(x) \geq \pi/2$.

Proof. Let $V = \mathrm{fix}(\alpha)$. As we have noted earlier V is a nonempty complete totally geodesic subspace. Given a collection of points x_0, $x_1, \ldots, x_m$ we denote by $\mathrm{sp}(x_0, x_1, \ldots, x_m)$ the smallest totally geodesic complete subset containing these points. The curvature assumption easily implies that $\mathrm{sp}(x_0, \dots, x_m)$ is diffeomorphic (for instance via the exponential mapping) to the interior of a ball of some dimension at most n and that the closure of this set is topologically a ball. As with the hyperbolic case we need to consider three cases.

(1) Suppose Γ is finite. Then there is a nonempty complete totally geodesic subspace F pointwise fixed by every element of Γ. Let $w \in F$. Then $\beta^i\alpha\beta^{-i}(w) = w$ for all $i = 0, 1, 2, \dots, p+1$ and so $x_i = \beta^{-i}(w) \in V$ for all $i = 1, \dots, p+1$. Let j be the smallest integer such that $\mathrm{sp}(x_0, x_1, \dots, x_j) = \mathrm{sp}(x_0, x_1, \dots, x_j, x_{j+1})$. Such a j exists by general position as the $\{x_i\}$ are a collection of $p+2$ points lying in the p-dimension subspace V. As β is an isometry we have

$$\begin{aligned}\beta(\mathrm{sp}(x_0, x_1, \dots, x_j)) &= \mathrm{sp}(\beta(x_0), \beta(x_1), \dots, \beta(x_j)) \\ &= \mathrm{sp}(x_1, x_2, \dots, x_{j+1}).\end{aligned}$$

But $\mathrm{sp}(x_1, x_2, \dots, x_{j+1})$ is a subspace of $\mathrm{sp}(x_0, x_1, x_2, \dots, x_{j+1})$ and a dimension count implies

$$\beta(\mathrm{sp}(x_0, x_1, \dots, x_j)) = \mathrm{sp}(x_0, x_1, \dots, x_j).$$

Since the closure of the set is topologically a ball, β has a fixed point in $\mathrm{sp}(x_0, x_1, \dots, x_j)$, a subset of V, and therefore $\langle\alpha, \beta\rangle$ fixes a point and must be an elementary group **[GM]**.

(2) Suppose $L(\Gamma) = \{x_0\}$. Since the limit set is Γ invariant we have $\alpha(x_0) = x_0$ and for all $i = 1, 2, \dots, p+1$, $\beta^i\alpha\beta^{-i}(x_0) = x_0$. The above argument now applies.

(3) Suppose $L(\Gamma) = \{x_0, y_0\}$. Again, the limit set is Γ invariant and so every element either fixes or interchanges the set $\{x_0, y_0\}$. If every element of Γ fixes this set, the argument of part one applies to give the result. Otherwise there is an element of Γ which interchanges these points and therefore the line between them. Consequently one of the generators must have this property. Since they are all conjugates of α, α must have this property. Then Lemma 2.1 implies $n_\alpha(x) \geq \pi/2$.

The following is the necessary generalization of Jørgensen's inequality.

2.4. THEOREM. *Suppose that α and β generate a discrete nonelementary negatively curved group. If α is hyperbolic or parabolic, then for each $x \in \mathrm{int}(\mathbf{B}^n)$*

$$\begin{aligned}&\max\{n_\alpha(x),\ n_{[\alpha,\beta]}(x)\} \geq 0.49 \quad \textit{and} \\ &\max\{n_\alpha(x),\ n_{\beta\alpha\beta^{-1}}(x)\} \geq 0.49.\end{aligned}$$

If α *is elliptic and* $p = \dim(\operatorname{fix}(\alpha))$, *then*

$$\max\{n_\alpha(x), n_{[\alpha, \beta^i]}(x) \colon i = 1, 2, \dots, p+1\} \geq 0.49 \quad \textit{and}$$
$$\max\{n_{\beta^i \alpha \beta^{-i}}(x) \colon i = 0, 1, 2, \dots, p+1\} \geq 0.49.$$

Proof. Suppose that the inequality does not hold. In the parabolic or hyperbolic case we find from Theorem 1.1 that the group $\langle \alpha, \beta\alpha\beta^{-1} \rangle$ is nilpotent and hence elementary. The result in this case follows by Lemma 2.2. In the elliptic case, we again obtain from Theorem 1.1 that the group $\langle \beta^i \alpha \beta^{-i} \colon i = 0, 1, \dots, p+1 \rangle$ is elementary. The result then follows by Lemma 2.3.

It is the following principle, first espoused by Jørgensen, that makes Theorem 2.4 such a valuable tool.

If two isometries α, β *of a negatively pinched metric generate a nonelementary discrete group, then given* α, β *cannot be too close to the identity.*

Jørgensen's inequality is conjugacy invariant, while the above are not (because of the dependence on x). However a simple corollary is

2.5. COROLLARY. *Suppose that* α *and* β *generate a discrete nonelementary negatively curved group. If* α *is hyperbolic or parabolic,*

$$\inf\{\max\{n_\alpha(x), n_{[\alpha, \beta]}(x)\} \colon x \in \operatorname{int}(\mathbf{B}^n)\} \geq 0.49.$$

There is of course a corresponding result if α is elliptic.

3. A limit theorem. We recall here that even in the hyperbolic case, for dimensions greater than three, the limit of finitely generated discrete nonelementary groups may not be discrete (even if the limit is assumed to be nonelementary) [**M**, §5]. The problem arises as there may be elliptics of high order converging to an irrational rotation about a codimension two or larger subspace stabilized by a nonelementary discrete subgroup. Thus in [**M**] we had to introduce the notion of uniformly bounded torsion. We must of course make a similar assumption in this more general case. To begin with we need the following lemma, essentially due to Newman [**N**], which we leave the reader to verify. It can be proved using Newman's result and a compactness argument for near the fixed point set the rotation angle is proportional to the reciprocal of the order.

3.1. LEMMA. *There is a positive constant* $\delta = \delta(m, n)$ *such that if* $x \in \operatorname{int}(\mathbf{B}^n)$ *and* α *is a periodic isometry of period less than or equal* *to* m *(which is not the identity), then*

$$n_\alpha(x) \geq \delta(m).$$

3.2. Definition. Let $\{\Gamma_i\}_{i\geq 0}$ be a family of groups. We say that $\{\Gamma_i\}$ has *uniformly bounded torsion* if there is an integer N such that for all i and $\gamma \in \Gamma_i$, either $\operatorname{ord}(\gamma) = \infty$ or $\operatorname{ord}(\gamma) \leq N$. Here $\operatorname{ord}(\gamma)$ is the minimal $m \neq 0$ such that $\gamma^m = \text{identity}$.

If Γ is group of homeomorphisms of $\mathbf{B}^n$ (possibly not discrete) we say that Γ is *nonelementary* if there are two elements α and β for which

(a) α^2 and β^2 have disjoint fixed point sets (if one has finite order) or

(b) α and β are of infinite order and α^2 and β^2 have different fixed point sets.

By different we mean that they do not coincide exactly. Otherwise Γ is *elementary*. This apparently weaker definition of nonelementary is equivalent to the usual definition for finitely generated Kleinian groups. We make this definition so as to include infinite torsion groups without a common fixed point amongst the nonelementary groups. We point out that if in addition Γ is a convergence group, then (a) and (b) follow from the definition of nonelementary given earlier. As (a) and (b) are essentially all that is needed in the limit groups for our proofs we shall use this definition henceforth.

Here is the first convergence theorem. It is analogous to [**J**, Proposition 1].

3.3. Theorem. *Let* Γ *be a nonelementary group of homeomorphisms of* $\mathbf{B}^n$. *For each* $m \geq 0$ *let* ψ_m *be a mapping* $\psi_m\colon \Gamma \to \operatorname{Isom}(\mathbf{B}^n, g_m)$ *into the isometry group of a Riemannian metric* g_m *whose sectional curvatures satisfy* $-1 \leq K(g_m) \leq -a_m^2 < 0$ *and which has discrete image. Suppose that* $\{\psi_m(\Gamma)\}_{m\geq 0}$ *has uniformly bounded torsion and that for each* $\gamma \in \Gamma$

$$\psi_m(\gamma) \to \gamma$$

uniformly in $\mathbf{B}^n$ *as* $m \to \infty$. *Then* Γ *is discrete.*

Proof. Suppose that Γ is not discrete. Then there is a sequence of elements $\{\alpha_i\}_{i\geq 0}$ converging (uniformly) to the identity in Γ. From our hypothesis there are two elements β_1 and β_2 satisfying (3.2) (a) or (b). As the fixed point sets are closed, they are uniformly separated.

Let $x \in \operatorname{int}(\mathbf{B}^n)$. It follows by continuity that for $j = 1, 2$ and sufficiently large i and m that

$$n_{\psi_m(\alpha_i)}(x) + n_{[\psi_m(\alpha_i), \psi_m(\beta_j)]}(x) \leq \min\{\delta(N), 0.49\}.$$

Where N is the bound on the maximal order of the cyclic subgroups and $\delta(N)$ is the number from Lemma 3.1. Thus $\psi_m(\alpha_i)$ is not elliptic for m and i sufficiently large and so from Theorem 2.4 the groups $\langle \psi_m(\alpha_i), \psi_m(\beta_j)\psi_m(\alpha_i)\psi_m(\beta_j)^{-1}\rangle$, $j = 1, 2$, are elementary and have nonempty limit set. If the limit set consists of one point, then these two elements have a common fixed point which is stabilized by $\psi_m(\beta_j)$, $j = 1, 2$. If the limit set contains two points, then $\psi_m(\beta_j)$ fixes or interchanges this set, $j = 1, 2$ and the squares have a common fixed point. Now letting $m \to \infty$ we obtain the desired contradiction.

REMARK. Of course in the hyperbolic case the assumption of uniform convergence implies that the limits are again hyperbolic isometries. Notice too that in the above situation we make no assumption about the convergence of the metrics $\{g_m\}_{m \geq 0}$, and that in the limit Γ may not even be a group of diffeomorphisms. Presumably however there is some underlying metric structure in the limit.

3.4. COROLLARY. *Let Γ be a nonelementary group of isometries of a Riemannian metric all of whose sectional curvatures satisfy $-A^2 \leq K(g) \leq -a^2$, which has bounded torsion. Then Γ is discrete if and only if every two generator subgroup is discrete.*

Proof. By multiplying the metric by a suitable constant we may assume the curvature satisfies $-1 \leq K(g) \leq -c^2 < 0$. Suppose that Γ is not discrete and let $\gamma_i \to$ identity in Γ. Since Γ has bounded torsion we may assume γ_i is not elliptic. By hypothesis there are two elements β_j, $j = 1, 2$, which are parabolic or hyperbolic and whose fixed point sets do not coincide exactly, or the fixed point sets are disjoint. Also by hypothesis the sequence of groups $\langle \gamma_i, \beta_j\rangle$, $j = 1, 2$, is discrete. Finally an argument similar to Theorem 3.3 implies this sequence of groups is elementary for sufficiently large i. We therefore reach a contradiction as above and this establishes the result.

3.5. COROLLARY. *Let Γ be a group of isometries of a Riemannian metric all of whose sectional curvatures satisfy $-A^2 \leq K(g) \leq -a^2$. Suppose that the fixed points of hyperbolic elements of Γ are in general position (that is, the fixed points of hyperbolic elements do not lie in a* *totally geodesic codimension* 1 *set*). *Then* Γ *is discrete if and only if every two generator subgroup is discrete.*

Proof. Let β_j, $j = 1, 2, \ldots, k$ be hyperbolic elements whose fixed point sets are in general position. Let $\{\gamma_i\}$ be a sequence as in Corollary 3.4. If this sequence (or some subsequence) is not eventually elliptic we can argue as above and conclude the desired result. Thus we assume γ_i are all elliptic. Then for all j and for all i sufficiently large we have (as above) $\langle \gamma_i, \beta_j \rangle$ is a discrete elementary group and so for all j, $\operatorname{fix}(\beta_j)$ lies in $\operatorname{fix}(\gamma_i)$ as β_j is hyperbolic. This is impossible as $\operatorname{fix}(\gamma_i)$ is totally geodesic and codimension at least 1.

In the above two results it is clear that some additional hypothesis, such as bounded torsion or the constraint on the fixed points of hyperbolic elements, is necessary (even in the hyperbolic case). We might for instance encounter a negatively curved group isomorphic to $\Gamma \times \mathbf{S}$, with the circle action stabilizing a lower dimensional hyperbolic space on which Γ acts discretely. If $\mathbf{Q}$ is that subgroup of $\mathbf{S}$ containing all the elements of finite order, then $\Gamma \times \mathbf{Q}$ has the property that all its two generator subgroups are discrete. Of course $\Gamma \times \mathbf{Q}$ is not discrete. However, it is quite possible that the additional hypotheses are redundant in the case that the group is finitely generated (this of course implies bounded torsion by Selberg's lemma in the symmetric cases). Thus we ask the question: *is it true that finitely generated groups of isometries of a metric of pinched negative curvature have bounded torsion*? G. Mess showed me an argument which implies this is the case if the abstract group is hyperbolic in the sense of Gromov [**G**].

4. Algebraic convergence.

4.1. DEFINITION. Let $\{\Gamma_i\}_{i \geq 0}$ be a sequence of negatively curved groups each with the same finite number of generators $\{\gamma_{i,1}, \gamma_{i,2}, \ldots, \gamma_{i,m}\}$. If for each $j = 1, 2, \ldots, m$, there is a self homeomorphism γ_j of $\mathbf{B}^n$ such that

$$\gamma_{i,j} \to \gamma_j \quad \text{as } i \to \infty,$$

then we say that the groups Γ_i *converge algebraically* to the group $\Gamma = \langle \gamma_1, \gamma_2, \ldots, \gamma_m \rangle$.

REMARK. Again we note that there is no assumption on the convergence of the underlying Riemannian metrics of negative curvature, and moreover no assumption that their curvature is uniformly bounded above by a negative constant. Thus the limit could possibly be a group of isometries of a flat metric (and therefore a negatively curved group). It is a consequence of our results that this is not the case.

4.2. THEOREM. *Let* Γ *be a group of homeomorphisms of* $\mathbf{B}^n$ *which is the algebraic limit of a sequence* $\{\Gamma_i\}_{i\geq 0}$ *of nonelementary negatively curved groups with uniformly bounded torsion. Then* Γ *is infinite and discrete. Moreover, if* Γ *is a negatively curved group, then* Γ *is nonelementary.*

In the three-dimensional hyperbolic case the assumption of bounded torsion is unnecessary and also that Γ is nonelementary is a consequence of algebraic convergence [**JK**]. For the higher-dimensional hyperbolic case, the assumption of bounded torsion (or a related assumption) is necessary and also implies that Γ is nonelementary. Notice in Theorem 4.2 we have skirted the delicate issue of when the limit of negatively curved groups is again negatively curved. We hope to return to this at a later date.

Proof. Let

$$\Gamma_i = \langle \gamma_{i,1}, \gamma_{i,2}, \dots, \gamma_{i,m}\rangle \quad \text{with } \gamma_{i,j} \to \gamma_i$$

be a sequence as in Theorem 4.2 with $\Gamma = \langle \gamma_1, \gamma_2, \dots, \gamma_m\rangle$. We proceed by a series of lemmas.

4.3. LEMMA. Γ *is not finite.*

Proof. Suppose Γ is finite. Then there is a $0 \leq k < \infty$ such that every element of Γ can be expressed as a word in the generators $\{\gamma_i\}$ of word length at most k. Since Γ_i is nonelementary, there are words of all lengths. For each i choose a word w_i of length $k+1$ in Γ_i (in terms of the generators given). Since there are finitely many such words, passing to subsequence we may assume that the sequence of words converges to a word w in Γ. This limit word can be expressed in a word w' of length k, and then let w_i' be the corresponding word in Γ_i. Then the sequence of words $v_i = w_i w_i'^{-1}$ is nontrivial, has word length at most $2k+1$ and converges to the identity. The assumption of bounded torsion implies that v_i is not elliptic for i sufficiently large by Lemma 3.1. Thus v_i is parabolic or hyperbolic. Then there is a generator $\gamma_{i,j}$ which does not stabilize the set $\operatorname{fix}(v_i)$. Hence the group $\langle v_i, \gamma_{i,k}\rangle$ is nonelementary. But as $v_i \to$ identity and $\gamma_{ij} \to \gamma_j$, we easily obtain a contradiction to Corollary 2.5.

Hence Γ is finitely generated and infinite. In the classical case (or whenever Γ is isomorphic to a subgroup of $\mathrm{GL}(m, \mathbb{C})$ for some m) we could now use Selberg's Lemma [S] to assert the existence of elements of infinite order. This would simplify the following arguments.

4.4. LEMMA. *Γ is discrete.*

Proof. If not choose $\alpha \in \Gamma$, so close to the identity (but $\alpha \neq$ identity) so that if α_i is any approximant to α in Γ_i, then α_i is parabolic or hyperbolic and also that for some $x \in \mathrm{int}(\mathbf{B}^n)$

$$n_{\alpha_i}(x) + n_{\gamma_{i,j}\alpha_i\gamma_{i,j}^{-1}}(x) < 0.49, \qquad j = 1, 2, \dots, m.$$

We can of course do this by appealing to continuity and Lemma 3.1. Then the group $\langle \alpha_i, \gamma_{i,j}\alpha_i\gamma_{i,j}^{-1} \rangle$ is elementary by Corollary 2.5. Thus each generator stabilizes $\mathrm{fix}(\alpha_i)$ and the group they generate is elementary. This is contrary to our hypothesis.

We suppose henceforth that Γ is a negatively curved group.

4.5. LEMMA. *Γ is nonelementary.*

Proof. As Γ is infinite and negatively curved, there is a hyperbolic or parabolic α in Γ ([**G**, Lemma 8.1 A] implies the group is not purely torsion). Since Γ_i is nonelementary, some generator $\gamma_{i,k}$ does not setwise fix the fixed points of α_i. Passing to a subsequence we may assume that $k = 1$ and $\gamma_{i,1} = \gamma_i$. If γ and α do not have a common fixed point, then we are done and so we suppose otherwise. Now suppose $\langle \alpha, \gamma \rangle$ is discrete and elementary. If α is hyperbolic, then by [**GM**] the group $\langle \alpha, \gamma \rangle$ is virtually cyclic. This easily leads to a contradiction. We are left to consider the case that every element of Γ of infinite order is parabolic with the same fixed point [**GM**]. That is, Γ is a finitely generated negatively curved parabolic group and so is virtually nilpotent [**B**]. Then there are integers p and q such that $a = \alpha^p$ and $b = \gamma\alpha^q\gamma^{-1}$ generate a nilpotent group. Let $x \in \mathrm{int}(\mathbf{B}^n)$. There is a d such that any d-fold commutator $[\ ,\ ]_d$ of a and b is trivial. Thus, summing over all the finitely many d-fold commutators involving only a and b

$$\sum_{[\ ,\ ]_d} n_{[\ ,\ ]_d}(x) = 0.$$

Then by continuity and Theorem 1.1, for sufficiently large i, the group generated by the d-fold commutators of the approximants a_i and b_i

will be nilpotent. Thus there is a d' so that all d'-fold commutators involving (only) the approximants a_i and b_i are trivial. Then the commutator identity $[a, bc] = [a, b][b, [a, c]][a, c]$ implies that all d'-fold commutators are trivial as they can be written as (much longer) products of d'-fold commutators in a_i and b_i, see for instance [**BK**, p. 28]. Thus $\langle a_i, b_i \rangle$ is eventually nilpotent and hence elementary. But this is easily seen not to be the case. The contradiction establishes the lemma.

Actually, the above argument only needs that Γ is a convergence group which is not purely torsion and whose discrete parabolic subgroups are virtually nilpotent. If this is actually the case, we can further restrict the possibilities for Γ with the following

4.6. LEMMA. *If Γ is infinite and purely torsion convergence group, then Γ is discrete and has bounded torsion.*

Proof. Γ is discrete. If there is no bound on the order of torsion elements, then the approximants to all elements of sufficiently high order will have to be parabolic or hyperbolic. Powers of these elements will again be parabolic or hyperbolic and close to the identity. The argument of Lemma 4.4 again produces a contradiction.

4.7. THEOREM. *Let Γ be a finitely generated abstract group with uniformly bounded torsion. For each $m \geq 0$ let*

$$\psi_m \colon \Gamma \to \operatorname{Isom}(\mathbf{B}^n, g_m), \qquad -1 \leq K(g_m) \leq -a_m^2 < 0,$$

be an isomorphism such that the images $\psi_m(\Gamma) = \Gamma_m$ are discrete and nonelementary and converge algebraically to Γ_∞. Then Γ_∞ is discrete and the correspondence of generators induces an isomorphism $\psi_\infty \colon \Gamma_\infty \to \Gamma$.

Proof. As ψ_m is an isomorphism, the sequence $\{\Gamma_m\}_{m \geq 0}$ has uniformly bounded torsion. Thus from Theorem 4.2 Γ_∞ is discrete. It is clear that the correspondence of generators $\psi_\infty \colon \Gamma_\infty \to \Gamma \cong \Gamma_m$ induces a homomorphism onto. Suppose $\alpha \in \Gamma_\infty - \{\text{Identity}\}$ and $\psi_\infty(\alpha) = \text{Identity}$. Uniformly bounded torsion implies that $\psi_m^{-1}(\alpha)$ is parabolic or hyperbolic for all m sufficiently large as $\psi_m(\gamma) \to \text{Identity}$ by Lemma 3.1. Now apply the argument of Lemma 4.4.

Indeed the above argument easily implies (see [**M**, Theorem 6.1]).

4.8. Theorem. *Let* $\{\Gamma_i\}_{i\geq 0}$ *be a sequence of discrete nonelementary negatively curved groups with uniformly bounded torsion converging algebraically to a finitely presented group of homeomorphisms* Γ. *Then* Γ *is discrete and the correspondence of generators* $\Gamma \rightarrow \Gamma_i$ *induces a homomorphism for all* i *sufficiently large.*

If each Γ_i *has a finite presentation, each relation of which has uniformly bounded wordlength, then the correspondence eventually induces an isomorphism.*

From [**BGS**, Theorem 10.2] if Γ is a negatively curved torsion free group with the injectivity radius of the quotient manifold $\mathbf{B}^n/\Gamma$ going to zero at infinity (for instance cofinite volume), then this quotient is diffeomorphic to the interior of a compact manifold (with boundary) and so the fundamental group Γ is finitely presented. The conclusion of Theorem 4.7 would then hold and all groups sufficiently close to Γ, in the topology of algebraic convergence, would be factors of Γ.

The symmetric Riemannian $\mathbb{R}$-rank one spaces of negative curvature are real hyperbolic space $\mathrm{SO}(1, n)/\mathrm{SO}(n)$, complex hyperbolic space $\mathrm{SU}(1, n)/\mathrm{U}(n)$, quaternionic hyperbolic space $\mathrm{Sp}(1, n)/\mathrm{Sp}(n)$ and hyperbolic space over the Cayley numbers $F_4/\mathrm{Spin}\,9$. The isometry groups are the Lie groups $\mathrm{SO}(1, n)$, $\mathrm{SU}(1, n)$, $\mathrm{Sp}(1, n)$ and F_4 respectively. The metric is the canonical left invariant metric with $-1 \leq K \leq -\frac{1}{4}$, see [**Mo**] and [**H**].

Let Γ be a discrete subgroup of one of the isometry groups above and let $\mathbf{H}(\Gamma)$ be the geodesically convex hull of the limit set $L(\Gamma)$ of Γ. We say that Γ is of *compact type* if $\mathbf{H}(\Gamma)/\Gamma$ is compact. Clearly compact type groups are finitely generated and finitely presented. Therefore and application of Theorem 4.8 yields

4.9. Corollary. *Suppose* Γ *is a compact type discrete subgroup of the isometry group of an* $\mathbb{R}$*-rank one space of negative curvature. Then there is a neighbourhood of the generators of* Γ *in* $\mathfrak{G} \times \mathfrak{G} \times \cdots \times \mathfrak{G}$ *such that every discrete group whose generators lie in this neighbourhood is a factor of* Γ.

5. Continuous deformations. In this section we shall show that a continuous deformation of a negatively curved group, through discrete negatively curved groups, is algebraically trivial, Theorem 5.3. That is, all the groups have the same isomorphism type. It then follows that in the torsion free cocompact case that such a deformation is topologically trivial as well. That is, all the associated quotients are homeomorphic and the deformation is induced by a topological conjugacy on the whole (closed) ball.

5.1. DEFINITION. We say a group Γ is *virtually torsion free*, if Γ has a torsion free subgroup of finite index. Let $D(k, n)$ denote the space of all virtually torsion free k generator discrete nonelementary negatively curved groups. We give $D(k, n)$ the topology of algebraic convergence. In what follows we fix $x \in \operatorname{int}(\mathbf{B}^n)$ and define the norm of α as $n(\alpha) = n_\alpha(x)$.

The proof of the following lemma is more or less implicit in what we have done in §4, see for instance [**M**, Lemma 6.2]

5.1. LEMMA. *Let* $\Gamma \in D(k, n)$. *Then there is a neighbourhood* $N_\varepsilon(\Gamma)$ *of* Γ *such that if* $\alpha \in N_\varepsilon(\Gamma)$ *and* $n(\alpha) < \varepsilon$, *then* $\alpha = \mathit{Identity}$.

The following is a generalization of [**J**, Theorem 3] and [**M**, Theorem 6.3]. The proof given here follows the latter reference.

5.2. THEOREM. *Let* E *be a connected compact subset of* $D(k, n)$. *Then* E *consists entirely of isomorphic groups.*

Proof. Let E_m denote that subset of E for which the maximal order of a finite cyclic subgroup is exactly m. Then $\{E_m\}_{m\geq 0}$ is a disjoint collection, E_m is compact by Theorem 4.2 and as each element of $D(k, n)$ is virtually torsion free $E = \bigcup_{m\geq 0} E_m$. Let $\Gamma \in E_m$ and R a relation in Γ, and F_m that subset of E_m with the relation R. Lemma 5.1 implies that F_m is relatively open; clearly it is also closed and so is a union of components of E_m. Thus each component of E_m consists entirely of isomorphic groups. As E is compact, Sierpinski's Theorem [**K**, §47III, Theorem 6] (which states that a compact connected set cannot be the countable union of closed disjoint subsets) implies that $E = E_m$ for some m and we are done.

It is interesting to note that each component of E_0 (the torsion free groups) consists entirely of isomorphic groups. Notice too that in the proof we did not use the full hypothesis that every group is virtually torsion free. Only that each group has bounded torsion. Furthermore, although Sierpinski's Theorem is not true for arbitrary closed sets, it is true for the real line. We can then obtain the following slightly different result. Here, by a *continuous deformation*, we mean that a fixed finite set of generators is being continuously deformed in $\operatorname{Hom}(\mathbf{B}^n)$. The underlying metrics (of negative curvature) need not change continuously.

5.3. THEOREM. *Let* Γ *be a finitely generated nonelementary negatively curved group with bounded torsion. Let* $\{\Gamma_t \colon t \in \mathbb{R}\}$ *be a continuous deformation of* $\Gamma = \Gamma_0$ *through discrete nonelementary negatively curved groups. Then* $\{\Gamma_t \colon t \in \mathbb{R}\}$, *and its closure in the topology of algebraic convergence in* $\mathrm{Hom}(\mathbf{B}^n)$, *consists entirely of isomorphic groups.*

Proof. Since the elliptics have integer-valued order, one cannot continuously change this without the order becoming infinite. Since each group in the deformation is assumed discrete, if the order of an elliptic is not preserved by the deformation it must be perturbed to a parabolic or hyperbolic element. But then some power of these elements will be arbitrarily close to the identity (by continuity) and still parabolic or hyperbolic. Pairing this power with its conjugate by each generator will imply by Corollary 2.5, as we have seen before in Lemma 4.4, that the groups generators stabilize a set containing at most two points and therefore the group they generate is elementary, contrary to our hypothesis. Thus the deformation preserves the orders of elliptics. As Γ has bounded torsion so does every element of the deformation. Then the proof of Theorem 5.2 implies the deformation is through isomorphic groups.

We say a negatively curved group is *cocompact* if the orbit space $\mathrm{int}(\mathbf{B}^n)/\Gamma$ is compact. Notice that cocompact implies nonelementary.

5.5. THEOREM. *Let* $\{\Gamma_t \colon t \in \mathbb{R}\}$ *be a continuous deformation of a torsion free group* Γ_0. *If each* Γ_t *is a discrete cocompact negatively curved group, then there is a continuous family of homeomorphisms* $f_t \colon \mathbf{B}^n \to \mathbf{B}^n$ *such that* $f_t \Gamma_0 f_t^{-1} = \Gamma_t$. *That is, the deformation is topologically trivial.*

Sketch of Proof. For each t the quotient is a compact negatively pinched manifold, each with the same isomorphic fundamental group Γ. The obvious action of this group on the space $\mathrm{int}(\mathbf{B}^n) \times [0, 1]$ (here $[0, 1]$ is the parameter space) is proper. The orbit space is a manifold foliated by the codimension 1, two-sided compact manifolds (leaves) $\mathrm{int}(\mathbf{B}^n)/\Gamma_t$. By Reeb stability, all the leaves are homeomorphic and the orbit space is a trivial fibration. Thus there is a conjugacy $f_t \colon \mathrm{int}(\mathbf{B}^n) \to \mathrm{int}(\mathbf{B}^n)$. Since each group is cocompact, it is uniform and therefore contains no parabolics [**BGS**, Lemma 8.2]. The usual Mostow-Margulis construction, see [**Mo**] and [**T**], shows that f_t is a pseudo-isometry, that is, a continuous map which is a bounded distance from an isometry (the minimal number of fundamental domains between two points is preserved by f as it is automorphic). In particular, the image under f of a geodesic line will be a bounded distance from another geodesic line. (Here the curvature assumptions and co-compactness simplify matters greatly. The details are not trivial, see **[Mo]** and compare with the $\mathbb{R}$-rank one lattice case there. Alternatively the argument given by Thurston **[T]** works in this general setting). Such maps as f extend homeomorphically to the boundary via their action on geodesic lines.

We remark that the much deeper results of Farrell and Jones **[FJ]** also imply that the leaves are homeomorphic (for $n \neq 3, 4$).

REFERENCES

[B] B. H. Bowditch *Discrete parabolic groups*, I.H.E.S. Preprint (1990).

[BK] P. Buser and H. Karcher, *Gromov's almost flat manifolds*, Astérique, **81** (1981).

[BGS] W. Ballman, M. Gromov and V. Schroeder, *Manifolds of nonpositive curvature*, Progress in Math., **61** (1985) Birkhäuser.

[EO] P. Eberlin and B. O'Neill, *Visibility manifolds*, Pacific J. Math., **46** (1973), 45–110.

[FJ] F. T. Farrell and L. E. Jones, *compact negatively curved manifolds* (*of* dim $\neq$ 3, 4) *are topologically rigid*, Proc. Nat. Acad. Sci. U.S.A., **86** No. 10 (1989), 3461–3463.

[GM] F. W. Gehring and G. J. Martin, *Discrete quasiconformal groups*, Proc. London Math. Soc., (3) **55** (1987), 331–358.

[GM2] ——, *Inequalities for Möbius transformations and discrete groups*, J. Reine Angew. Math., **418** (1991), 31–76.

[G] M. Gromov, *Hyperbolic Groups*, Essays in group theory. Edited by S. M. Gersten. Springer Verlag (1987), 75–265.

[H] S. Helgason, *Differential Geometry and Symmetric Spaces*, Academic Press, (1962).

[J] T. Jørgensen, *On discrete groups of Möbius transformations*, Amer. J. Math., **98**, 3 (1976), 739–749.

[JK] T. Jørgensen and P. Klein, *Algebraic convergence of finitely generated Kleinian groups*, Quart. J. Math. Oxford (2), **33** (1982), 325–332.

[K] K. Kuratowski, *Topology*, Academic Press, (1966).

[KP] M. Kapovich and L. Potyagailo, *On the absence of Ahlfor's finiteness theorem for Kleinian groups in dimension* 3, (to appear) Topology Appls.

[M] G. J. Martin, *On discrete Möbius groups in all dimensions*, Acta Math., **163** (1989), 253–289.

[MS] G. J. Martin and R. Skora, *Group actions of the two sphere*, Amer. J. Math., **111** (1989), 387–402.

[Mo] G. D. Mostow, *The strong rigidity of locally symmetric spaces*, Annals of Math. Studies, **78**. Princeton Univ. Press, (1973).

[N] M. H. A. Newman, *A theorem on periodic transformations of spaces*, Quart. J. Math., (2) (1931), 1–8.

[P] P. Pansu, *Métriques de Carnot-Carathéodory et quasi-isometries des espaces symétrique de rang un*, Annals. of Math., **129** (1989), 1–60.

[S] A. Selberg, *On discontinuous groups in higher dimensional symmetric spaces*, Contributions to function theory, Bombay (1960), 147–164.

[Sc] G. P. Scott, *Finitely generated 3-manifold groups are finitely presented*, J. London Math. Soc., **6** (1973), 437–440.

[T] W. P. Thurston, *The geometry and topology of 3-manifolds*, Princeton (1978).

[Tu] P. Tukia, *On isomorphisms of geometrically finite Möbius groups*, Publ. I.H.E.S., **61** (1985), 171–214.

Received October 11, 1990 and in revised form January 18, 1992.

THE UNIVERSITY OF AUCKLAND
AUCKLAND, NEW ZEALAND
E-mail address: martin@mat.auckland.ac.nz

AND

INSTITUTE MITTAG-LEFFLER
AURAVÄGEN 17
DJURSHOLM, SWEDEN

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

ADJOINT LINEAR SYSTEMS ON A SURFACE OF GENERAL TYPE IN POSITIVE CHARACTERISTIC

TOHRU NAKASHIMA

Let X be a minimal surface of general type defined over an algebraically closed field of positive characteristic p. For a given divisor D, we consider the spannedness properties of adjoint linear systems $|K+D|$ on X. Under some numerical conditions on p and D, the failure of spannedness of $|K+D|$ implies the existence of divisors with special properties. This leads to the following result: Let L be an ample line bundle and assume $p \geq 5$. Then $|m(K+L)|$ is base point free for $m \geq 2$ and very ample for $m \geq 3$. Our proof is based on a technique of Shepherd-Barron using unstable vector bundles.

1. Introduction. After Reider introduced a new method ([**R**]), many results have been obtained concerning adjoint linear systems on algebraic surfaces defined over an algebraically closed field of characteristic 0. Recently, Shepherd-Barron ([**SB**]) treated the positive characteristic case and obtained results on pluricanonical systems, improving the work of Ekedahl ([**E**]). He also showed Reider's analysis holds for surfaces of special type except the quasi-elliptic ones. His method is, as in [**R**], based on the theory of unstable vector bundles in the sense of Bogomolov.

In the present note we shall consider adjoint linear systems on a minimal surface of general type in characteristic p and prove some results of Reider's type.

Let X be a minimal surface of general type defined over an algebraically closed field k of $\operatorname{char} k = p > 0$ and let D be a nef divisor such that $D - K$ is nef and big. We shall prove the following

THEOREM 1. *Let X and D be as above and let $d := D^2$.*
(i) *Suppose that one of the following conditions holds:*
(1) $p \geq 2$, $d \geq 5$ *and X is not uniruled,*
(2) $p = 3$ *and* $d \geq 12$,
(3) $p = 5$ *and* $d \geq 6$,
(4) $p \geq 7$ *and* $d \geq 5$.
If $|K+D|$ has a base point, then there exists an effective divisor Δ such

that

$$\text{either} \quad D\cdot\Delta = 1, \quad \Delta^2 = 0,$$
$$\text{or} \quad D\cdot\Delta = 0, \quad \Delta^2 = -1.$$

(ii) *Suppose that one of the following conditions holds*:
(1′) $p \geq 2$, $d \geq 10$ *and* X *is not uniruled,*
(2′) $p = 3$ *and* $d \geq 30$,
(3′) $p = 5$ *and* $d \geq 13$,
(4′) $p = 7$ *and* $d \geq 11$,
(5′) $p \geq 11$ *and* $d \geq 10$.
If $|K + D|$ *is not very ample, then there exists an effective divisor* Δ *such that*

$$\text{either} \quad D\cdot\Delta = 0, \quad \Delta^2 = -1, -2,$$
$$\text{or} \quad D\cdot\Delta = 1, \quad \Delta^2 = -1, 0,$$
$$\text{or} \quad D\cdot\Delta = 2, \quad \Delta^2 = 0.$$

As a corollary of the above theorem, we obtain the following result on pluri-adjoint systems:

COROLLARY 2. *Let* L *be an ample divisor on* X.
(i) *Suppose that one of the following conditions holds*:
(1) $p \geq 2$, $m \geq 2$ *and* X *is not uniruled,*
(2) $p = 3$ *and* $m \geq 3$,
(3) $p \geq 5$ *and* $m \geq 2$.
Then $|m(K + L)|$ *is base point free.*
(ii) *Suppose that one of the following conditions holds*:
(1′) $p \geq 2$, $m \geq 3$ *and* X *is not uniruled,*
(2′) $p = 3$ *and* $m \geq 4$,
(3′) $p \geq 5$ *and* $m \geq 3$.
Then $|m(K + L)|$ *is very ample.*

Proof. Apply the theorem to $D = (m - 1)K + mL$. □

2. Proof of the theorem. Let p be a base point of $|K + D|$ in (i) (resp. p, q be the points not separated by $|K + D|$ in (ii)) and let $\pi\colon \widetilde{X} \to X$ be the blowing up at p in (i) (resp. at p and q in (ii)). Put $l := \pi^{-1}(p)$, $m := \pi^{-1}(q)$, and $\widetilde{D} := \pi^*D - 2l$ in (i) (resp. $\widetilde{D} := \pi^*D - 2(l + m)$ in (ii)).

Since we have $H^1(\widetilde{X}, \mathcal{O}_{\widetilde{X}}(-\widetilde{D})) \neq 0$, we have a nonsplit sequence on $\widetilde{X}$:

$$0 \to \mathcal{O}_{\widetilde{X}} \to E \to \mathcal{O}_{\widetilde{X}}(\widetilde{D}) \to 0.$$

Since E satisfies the inequality $c_1(E)^2 > 4c_2(E)$, by Theorem 1 in **[SB]**, there exists a Frobenius map $F^e\colon \widetilde{X} \to \widetilde{X}$ and an exact sequence

$$0 \to \mathcal{O}_{\widetilde{X}}(p^e\widetilde{D} - \Delta_1) \to \widetilde{E} \to \mathcal{J}_Z \otimes \mathcal{O}_{\widetilde{X}}(\Delta_1) \to 0.$$

Here $\widetilde{E} := (F^e)^*E$, Z is a 0 cycle and Δ_1 is some effective divisor such that $p^e\widetilde{D} - 2\Delta_1$ is contained in the positive cone of $\widetilde{X}$. We write $\Delta_1 = \pi^*\Delta + rl$ in (i) (resp: $\Delta_1 = \pi^*\Delta + rl + sm$ in (ii)) where r, s are some integers and Δ is an effective divisor on X.

If $p^e = 1$, Reider's argument shows Δ satisfies the properties stated in the theorem (cf. **[R]**). We shall show that the case $p^e \geq p$ never occurs under our assumption.

Suppose $p^e \geq p$. Then there is a purely inseparable covering $\rho\colon Y \to \widetilde{X}$ of $\deg \rho = p^e$ and we have the following estimates (cf. **[SB]**).

Lemma 3. *Assume* $d := D^2 \geq 5$ *in* (i) (*resp.* $d \geq 10$ *in* (ii)). *Then*

$$D \cdot \Delta \leq \left(\frac{d}{2} - \sqrt{\frac{d^2}{4} - d}\right) p^e$$

and

$$\chi(\mathcal{O}_Y) \geq \left(\chi(\mathcal{O}_X) + \frac{p^e - 1}{12}[(2p^e - 1)\widetilde{D}^2 - 3\widetilde{D} \cdot K_{\widetilde{X}}]\right) p^e.$$

Since both D and $D - K$ are nef and big, the Hodge index theorem yields $K \cdot D \leq d - 3$ and $K^2 \leq d - 5$ in (i) (resp. $K^2 \leq d - 6$ in (ii)). By these estimates, we have

$$\begin{aligned}
\omega_Y \cdot \rho^*\pi^*D &= 2(p^e - 1)D \cdot \Delta + p^e(K \cdot D - (p^e - 1)D^2) \\
&\leq 2(p^e - 1)\left(\frac{d}{2} - \sqrt{\frac{d^2}{4} - d}\right) + p^e(d - 3 - (p^e - 1)d) \\
&= (d - 3 - (p^e - 1)\sqrt{d^2 - 4d})p^e.
\end{aligned}$$

Thus we obtain $\omega \cdot \rho^*\pi^*D < 0$. Therefore Y is ruled and hence X is uniruled. Let $q(X)$ be the irregularity of X. Then by Lemma 34 in **[SB]**, we have $\chi(\mathcal{O}_Y) \leq 1 - q(X) \leq 1$. Assume we are in the case (i). Then Corollary 30 and Proposition 35 in the same paper yield

$$\begin{aligned}\chi(\mathscr{O}_Y) &\geq \left(\chi(\mathscr{O}_X) + \frac{p^e-1}{12}[(2p^e-1)\tilde{D}^2 - 3\tilde{D}\cdot K_{\tilde{X}}]\right)p^e \\ &\geq \left(-\frac{K^2}{10} + \frac{p^e-1}{12}[(2p^e-1)(D^2-4) - 3(D\cdot K+2)]\right)p^e \\ &\geq \left(-\frac{d-5}{10} + \frac{p^e-1}{12}[(2p^e-1)(d-4) - 3(d-3) - 6]\right)p^e.\end{aligned}$$

Similarly in the case (ii) we obtain

$$\chi(\mathscr{O}_Y) \geq \left(-\frac{d-6}{10} + \frac{p^e-1}{12}[(2p^e-1)(d-8) - 3(d-3) - 12]\right)p^e.$$

However, under the assumption that one of (1) to (4) (resp. (1′) to (5′)) holds, we have $\chi(\mathscr{O}_Y) > 1$. This is a contradiction and hence the theorem is proved. □

REFERENCES

[E] T. Ekedahl, *Canonical models of surfaces of general type in positive characteristic*, Publ. Math. I.H.E.S., **67** (1988), 97–144.

[R] I. Reider, *Vector bundles of rank 2 and linear systems on algebraic surfaces*, Ann. of Math., **127** (1988), 309–316.

[SB] N. I. Shepherd-Barron, *Unstable vector bundles and linear systems on surfaces in characteristic p*, preprint.

Received March 18, 1991.

TOKYO METROPOLITAN UNIVERSITY
TOKYO 192-03, JAPAN

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

A HOMOTOPY TRANSFER FOR FINITE GROUP ACTIONS

BILL RALPH

We obtain a transfer for group actions on spaces for which the orbit map admits a section. This transfer exists for sets of homotopy classes as well as for any generalized homology theory.

Introduction. Intuitively, one feels that there should exist special relationships between the homotopy invariants of a space Y and its quotient by the action of some finite group G. The main result of this paper is the construction, for an arbitrary homology theory, of a version of the transfer that exists for ordinary homology. Recall that this is a homomorphism $\tau\colon H_n(Y/G) \to H_n(Y)$ with the following properties:

(a) $\tau \circ \rho(z) = \sum_{g\in G} g * z$ for all $z \in H_n(Y)$,
(b) $\rho \circ \tau(v) = |G|v$ for all $v \in H_n(Y/G)$

where $\rho\colon H_n(Y) \to H_n(Y/G)$ is the projection. An account of this can be found in **[Br]**.

Unfortunately, the existence of a transfer map satisfying (a) and (b), or their duals in cohomology, seems to be a special property of the ordinary homology and cohomology functors which is closely tied to the fact that Eilenberg-Mac Lane spaces have the homotopy type of abelian monoids. In view of this, it is not surprising that in general there is no transfer for covariant functors F such as π_n and those associated with generalized homology theories.

In this paper we will recover a version of transfer for many functors F including generalized homology theories. In order to deal with the fact that the H-spaces that arise are not, in general, of the homotopy type of abelian monoids, we will have to multiply equation (a) by a number $c(G)$, that I have been calling the coherence number of the group G. This number depends only on the group G and is currently under intense investigation. For cyclic groups this number is 1 and hence the transfer equations will have their usual form in this case. It is not yet known whether this number is always finite, so there may be groups to which our transfer cannot be applied, although our feeling is that this is not the case. **[R3]** contains all the current information on this number. The reason the coherence number is not required for abelian monoids can be understood in general form **[R1]** and, for ordinary homology in particular, from **[R2]**.

1. Statement of results. Throughout this paper, we will assume that all spaces have basepoints, that G is a finite group acting on a space Y, that G fixes the basepoint of Y and that the orbit map admits a section. We will let $\rho\colon Y \to Y/G$ denote the orbit map and let $s\colon Y/G \to Y$ denote a section of it. The symbol $c(G)$ which occurs in the following theorems denotes a positive integer or infinity and is a combinatorial invariant of the group that will be defined later.

The main results of this paper are the following:

THEOREM 1. *Let Y have the homotopy type of a G-C.W. complex and suppose that $c(G) < \infty$. Let $\tilde{h}_*$ be the (reduced) homology theory associated with some spectrum and let $\rho\colon \tilde{h}_n(Y) \to \tilde{h}_n(Y/G)$ denote the obvious homomorphism. Then there exists a transfer homomorphism $\tau\colon \tilde{h}_n(Y/G) \to \tilde{h}_n(Y)$ satisfying the following:*

(1) $c(G)(\tau \circ \rho(z) - \sum_{g\in G} g * z) = 0$ *for all* $z \in \tilde{h}_n(Y)$,

(2) $\rho \circ \tau(v) = |G|v$ *for all* $v \in \tilde{h}_n(Y/G)$. □

We also give a version of Theorem 1 for homotopy. See Theorem 6 below.

The following theorems give the flavour of the kind of constraints imposed by the transfer.

THEOREM 2. *Suppose that $c(G) < \infty$ and there is some homology theory such that $\tilde{h}_n(Y) = \mathbb{Z}$ and $\tilde{h}_n(Y/G) = 0$. Then G must have a subgroup of index 2.*

THEOREM 3. *Suppose that $\mathbb{Z}_4$ acts on W and induces an effective action on $\tilde{h}_n(W) \cong \mathbb{Z}_{16}$ for some homology theory. Then the orbit map $W \to W/\mathbb{Z}_4$ does not admit a section.* □

Here is the definition of the transfer map in its most general setting.

DEFINITION 1. Let X and Y be spaces. Let W be an H-space for which there exists a topological group $(L, 1)$ and a map $\varepsilon\colon (W, w_0) \to (L, 1)$ inducing an isomorphism of groups $\bar{\varepsilon}\colon [X, W] \to [X, L]$. Let α be a map from Y to W and let $\bar{\alpha}$ be the induced map from $[X, Y]$ to $[X, W]$. Let $\bar{\rho}\colon [X, Y] \to [X, Y/G]$ be the map induced by the projection, let $\bar{s}\colon [X, Y/G] \to [X, Y]$ be the map induced by the section and let $\beta\colon [X, W] \to [X, W]^{\mathrm{ab}}$ be the abelianizing map. The transfer map is the function

$$\tau_\alpha\colon [X, Y/G] \to [X, W]^{\mathrm{ab}}$$

defined by

$$\tau_\alpha(v) = \sum_{g\in G} \beta \circ \bar{\alpha}(g * \bar{s}(v)). \qquad \square$$

2. Coherence numbers. As mentioned in the introduction, the development of this transfer requires a new idea, namely the "Coherence Number" of a finite group. We first define the coherence number of a set of permutations on a finite set and then specialize this definition to finite groups. The reader should refer to [**R3**] for details and proofs. Let $P(S)$ denote the set of all permutations on the set S.

DEFINITION 2. Let S be a finite set and let $A \subseteq P(S)$. Suppose that $S = \{s_1, \dots, s_m\}$ and $A = \{\sigma_1, \dots, \sigma_n\}$. Let $r_i = (\sigma_1(s_i), \dots, \sigma_n(s_i))$ and $c_i = (s_i, \dots, s_i)$, for $i = 1, 2, \dots, m$, be regarded as elements of $F(S)^n$, the direct product of n copies of the free group on the set S. Let T be the subgroup of $F(S)^n$ generated by $r_1, \dots, r_m$ and $c_1, \dots, c_m$ and define $\theta = \prod_{i=1}^m r_i c_i^{-1} \in T$. There is a homomorphism $\lambda\colon T \to T/[T, T]$, where $[T, T]$ is the commutator subgroup of T. The coherence number, $c(A)$, of A is defined to be the order of $\lambda(\theta)$, which may be infinite. It is easily verified that $c(A)$ is independent of the ways in which S and A are ordered. $\square$

We now specialize this definition to groups by letting the group act on itself as a set of permutations.

DEFINITION 3. Let G be any finite group. We can regard G as a subset of $P(G)$ by letting G act on itself on the left by translation. Specifically, for $g \in G$, we define $\bar{g} \in P(G)$ by putting $\bar{g}(h) = gh$. The coherence number of $\{\bar{g} | g \in G\} \subseteq P(G)$ will be called the coherence number of the group G and will be denoted by $c(G)$. We will use T_G to denote the subgroup corresponding to T in Definition 2. $\square$

In [**R3**] we showed that the coherence numbers obtained by letting G act on itself on the right or the left are the same. We have chosen Definition 3 as the most convenient in this context.

3. The combinatorial part of the transfer. In what follows we shall denote the action of G on a point, function, homotopy class, etc. by a $*$ and let the meaning be determined from the context.

Definition 4. Let X and Y be sets and let $\{A_i\}$ be a set of n subsets that cover X. If $f_i\colon X \to Y$ and $g_i\colon X \to Y$ for $i = 1, \dots, n$ are two families of maps, then we will say that $(f_1, \dots, f_n) \sim (g_1, \dots, g_n)$ if the restrictions of f_i and g_i to A_i are the same for all $i = 1, \dots, n$. □

Our next lemma contains the essential combinatorial idea that makes the transfer work. Although it is only stated in terms of continuous functions, the identical proof works for arbitrary functions. In the following lemma, $(L, 1)$ is a topological group and $\zeta\colon (Y, y_0) \to (L, 1)$ is a map. $\mathrm{Hom}((X, x_0), (Y, y_0))$ will denote the set of maps from (X, x_0) to (Y, y_0) and $\bar{\zeta}$ denotes the induced map from $\mathrm{Hom}((X, x_0), (Y, y_0))$ to $\mathrm{Hom}((X, x_0), (L, 1))$. We will regard $\mathrm{Hom}((X, x_0), (L, 1))$ as a group under pointwise multiplication.

Lemma 1. *Let G be a finite group with finite group with finite coherence number. Let (X, x_0) be a Hausdorff space and let (Y, y_0) be a space on which G acts. Let σ, $\omega \in$ $\mathrm{Hom}((X, x_0), (Y, y_0))$ with $\rho \circ \sigma = \rho \circ \omega$. Let M be the subgroup of* $\mathrm{Hom}((X, x_0), (L, 1))$ *generated by all elements of the form $\zeta \circ (g * \sigma)$ and $\zeta \circ (g * \omega)$, where g is any element of G. Let $\delta\colon M \to M^{\mathrm{ab}}$ denote the projection. Then we have*

$$c(G) \sum_{g \in G} \delta \circ \bar{\zeta}(g * \sigma) = c(G) \sum_{g \in G} \delta \circ \bar{\zeta}(g * \omega)$$

in M^{ab}.

Proof. Recall, from Definitions (3) and (4), that the group T_G is generated by the elements $r_i = (g_i g_1, \dots, g_i g_n)$ and $c_i = (g_i, \dots, g_i)$ for $i = 1, \dots, n$. The first step in our proof is to show that the homomorphism from T_G into $\mathrm{Hom}((X, x_0), (L, 1))$ given on generators by $r_i \mapsto \zeta \circ (g_i * \sigma)$ and $c_i \mapsto \zeta \circ (g_i * \omega)$ is well defined.

Let $A_i = \{x \in X | \sigma(x) = g_i * \omega(x)\}$. Note that, since $\rho \circ \sigma = \rho \circ \omega$, these sets cover X. It is clear from the definition of the A_i and Definition 4 that $(\sigma, \dots, \sigma) \sim (g_1 * \omega, \dots, g_n * \omega)$. From this, we see in general that $(g * \sigma, \dots, g * \sigma) \sim ((gg_1) * \omega, \dots, (gg_n) * \omega)$.

We will construct a homomorphism $\psi_{x,k}\colon T_G \to L$ as follows. Let ρ_k be the projection of the subgroup T_G of $F(G)^{|G|}$ onto the kth factor. For each $x \in X$, we define a map of sets, $\eta_x\colon G \to L$, by $\eta_x(g) = \zeta \circ (g * \omega)(x)$, which then induces a homomorphism $\bar{\eta}_x\colon F(G) \to F(L)$. Let $\upsilon\colon F(L) \to L$ be the canonical homomorphism and define $\psi_{x,k} = \upsilon \circ \bar{\eta}_x \circ \rho_k$.

Now suppose that $x \in A_j \cap A_k$. We claim that $\psi_{x,j} = \psi_{x,k}$. To see this, the reader can easily check that equality holds on the two types of generators for the group T_G.

Since the sets A_k cover the set X, it follows that the homomorphisms $\psi_{x,k}$, obtained from each pair x and k with $x \in A_k$, can be "glued" together into a single homomorphism ψ from T_G into the set of continuous functions from (X, x_0) into $(L, 1)$. It is immediate from our construction of ψ that $\psi(r_i) = \zeta \circ (g_i * \sigma)$ and $\psi(c_i) = \zeta \circ (g_i * \omega)$ and hence we obtain a well-defined map from T_G into $\operatorname{Hom}((X, x_0), (L, 1))$.

Clearly the image of ψ is precisely the subgroup M of $\operatorname{Hom}((X, x_0), (L, 1))$ and so ψ induces a homomorphism $\psi^{\mathrm{ab}}\colon T_G^{\mathrm{ab}} \to M^{\mathrm{ab}}$. Since the coherence number of G was assumed to be finite, we know that $c(G)\lambda(\theta) = 0$, where $\lambda(\theta)$ is the element of T_G^{ab} defined in Definition 2. Applying ψ^{ab}, we see that $c(G)\psi^{\mathrm{ab}}(\lambda(\theta)) = 0$. Expanding this, we obtain that

$$\begin{aligned} &c(G)\psi^{\mathrm{ab}}\left(\lambda\left(\prod_{i=1}^{m} r_i c_i^{-1}\right)\right) \\ &\quad = \left(c(G)\sum_{g\in G} \delta \circ \bar{\zeta}(g * \sigma)\right) - \left(c(G)\sum_{g\in G} \delta \circ \bar{\zeta}(g * \omega)\right) \end{aligned}$$

which equals zero and gives our result. □

The reader should check that, if the topological group L used above happened to be an abelian monoid, then we would not need the factor $c(G)$.

4. Properties of the transfer map.

Here is the crucial property of the transfer map.

THEOREM 4. *If* τ_α *is given as in Definition* 1 *and* $z \in [(X, x_0), (Y, y_0)]$ *then*

$$c(G)\left(\tau_\alpha \circ \bar{p}(z) - \sum_{g\in G} \beta \circ \bar{\alpha}(g * z)\right) = 0$$

in $[(X, x_0), (W, w_0)]^{\mathrm{ab}}$.

Proof. Define $\zeta = \varepsilon \circ \alpha\colon (Y, y_0) \to (L, 1)$ and choose a representative $z\colon X \to Y$. Let M be the subgroup of $\operatorname{Hom}((X, x_0), (L, 1))$

generated by all the elements of the form $\zeta\circ(g*z)$ and $\zeta\circ(g*\{s\circ\rho\circ z\})$ for all $g\in G$. By Lemma 1,

$$c(G)\sum_{g\in G}\delta\circ\bar{\zeta}(g*z)=c(G)\sum_{g\in G}\delta\circ\bar{\zeta}(g*\{s\circ\rho\circ z\})$$

in M^{ab}.

From this it follows that

$$c(G)\sum_{g\in G}[[\bar{\zeta}(g*z)]]=c(G)\sum_{g\in G}[[\bar{\zeta}(g*\{s\circ\rho\circ z\})]]$$

in $[(X, x_0), (L, 1)]^{\mathrm{ab}}$, where the inner and outer brackets indicate the two equivalence relations of homotopy and abelianization respectively. The result now follows after applying the isomorphism

$$(\bar{\varepsilon}^{-1})^{\mathrm{ab}}\colon [(X, x_0), (L, 1)]^{\mathrm{ab}}\to[(X, x_0), (W, w_0)]^{\mathrm{ab}}. \qquad \square$$

Since the subgroup generated by a single element is abelian, it turns out that, under an additional hypothesis, we can say something about the order of elements in $[(X, x_0), (W, w_0)]$. The reader can easily furnish the proof of the following by modifying the proof of Theorem 4.

THEOREM 5. *Assume the context of Definition* 1 *and in addition that* $[X, Y/G]$ *is trivial. Suppose that for some* $[f]\in[X, Y]$ *and all* $g\in G$ *the elements* $\bar{\alpha}([g*f])$ *coincide in* $[X, W]$. *Then the order of* $\bar{\alpha}([f])$ *in* $[X, W]$ *divides* $c(G)|G|$. □

We next examine the simplest form of our transfer, which occurs when Y is an H space which has the homotopy type of a topological group and X is a co-H space. In this case, the transfer map τ is a homomorphism from $[X, Y/G]$ to $[X, Y]$. It is in this setting that we feel our use of the term transfer is most justified, since we obtain a transfer map with exactly the same properties as the transfer for singular homology theory, except for the appearance of the factor $c(G)$. This factor can be looked on as the "correction" term that is required because homotopy theory is "not as abelian", in a certain sense (see **[R1]** for a precise description), as we might like it to be.

THEOREM 6. *In addition to the assumptions of Definition* 1, *assume that* $Y = W$ *and that* X *is a co-H space. Then the transfer map*

$\tau\colon [X, Y/G] \to [X, Y]$ is a homomorphism with the following properties:

$$(1) \qquad c(G)\left(\tau \circ \bar{p}(z) - \sum_{g\in G} g * z\right) = 0 \quad \textit{for all } z \in [X, Y],$$

$$(2) \qquad \bar{p} \circ \tau(v) = |G|v \quad \textit{for all } v \in [X, Y/G].$$

Proof. Immediate from Theorem 4 and the definitions. □

5. The transfer for a generalized homology theory. Next, we develop a transfer for the reduced homology theory associated with any spectrum.

We will follow the definitions and notation for spectra given in [Sw]. Recall that the reduced homology theory, $\tilde{h}_*$, associated with a spectrum E is defined by

$$\tilde{h}_n(Y) = \operatorname*{dirlim}_k \pi_{k+n}(E_k \wedge Y, *),$$

where the connecting homomorphisms ϕ_k are defined by the composition

$$\pi_{k+n}(E_k \wedge Y, *) \xrightarrow{\Sigma_k} \pi_{k+n+1}(S^1 \wedge E_k \wedge Y, *) \xrightarrow{\iota_k} \pi_{k+1+n}(E_{k+1} \wedge Y, *).$$

It will be convenient in what follows to denote by φ_k the corresponding homomorphisms used to define $\tilde{h}_n(Y/G)$.

Proof of Theorem 1. The proof will be a straightforward "direct limit version" of Theorem 4. Note that we have not had to make any assumptions involving topological groups. This is because suspending maps is equivalent to mapping into spaces of the form ΩSY, where Y is a CW complex, and as, is well known, see [St] for example, these spaces are of the homotopy types of topological groups. In what follows, we will extend the group action of G on Y to an action of G on $E_k \wedge Y$ in the obvious way.

We begin by taking the spaces X, Y and W in Theorem 4 to be S^{n+k}, $E_k \wedge Y$ and $\Omega S(E_k \wedge Y)$, respectively, and the map $\alpha_k\colon E_k \wedge Y \to \Omega S(E_k \wedge Y)$ to be the adjoint of the suspension of the identity map. By Theorem 4, there is a transfer homomorphism

$$\tau_{\alpha_k}\colon \pi_{n+k}(E_k \wedge Y/G, *) \to \pi_{n+k}(\Omega S(E_k \wedge Y), *).$$

Let $\mathrm{adj}_k\colon \pi_{n+k}(\Omega S(E_k \wedge Y), *) \to \pi_{k+n+1}(S^1 \wedge E_k \wedge Y, *)$ denote the adjoint homomorphism and define

$$\tau_k = \iota_k \circ \mathrm{adj}_k \circ \tau_{\alpha_k}\colon \pi_{n+k}(E_k \wedge Y/G, *) \to \pi_{k+1+n}(E_{k+1} \wedge Y, *).$$

The theorem now follows immediately from the following easily verified properties:

(a) $\phi_{k+1} \circ \tau_k = \tau_{k+1} \circ \varphi_k$,
(b) $\rho_{k+1} \circ \tau_k(v) = |G|\varphi_k(v)$,
(c) $c(G)(\tau_k \circ \rho_k(z) - \sum_{g\in G} \phi_k(g * z)) = 0$. □

Using the fact that Eilenberg-Mac Lane spaces have the homotopy types of abelian monoids and the remark following Lemma 1, it can be shown that the coherence number factor $c(G)$ can be dropped in Theorem 1 to obtain a transfer resembling the usual transfer for ordinary homology theory.

6. Proofs of Theorems (2) and (3).

Proof of Theorem 2. From Theorem 1, the transfer gives that $c(G)\sum_{g\in G} g * z = 0$. Since $\tilde{h}_n(X) = \mathbb{Z}$, we have that $\sum_{g\in G} g * z = 0$. This can only happen if there is a nontrivial homomorphism from G into the automorphism group $\mathbb{Z}_2$ of $\mathbb{Z}$. The result follows immediately. □

Proof of Theorem 3. By way of contradiction, assume that the orbit map has a section. The units of order 4 in $\mathbb{Z}_{16} = \{0, 1, \dots, 15\}$ are 3, 5, 11 and 13. Since the action is effective, $\rho: \tilde{h}_n(W) \to \tilde{h}_n(W/\mathbb{Z}_4)$ cannot be an isomorphism. ρ is a split epimorphism since the orbit map admits a section. Therefore the image of the map in homology induced by the section is a direct summand of $\mathbb{Z}_{16}$. It follows that this image and the image of the transfer map must be trivial. Since $\mathbb{Z}_4$ is cyclic, its coherence number is 1 by **[R3]**. Theorem (1) gives that $\tau \circ \rho(z) = \sum_{g\in G} g * z$ which equals either $8z$ or $12z$ depending on which unit of order 4 we take. This contradicts the triviality of τ. □

REFERENCES

[Ba] M. G. Barratt, *Spaces of finite characteristic*, Quart. J. Math., **11** (1960), 124–136.

[Br] G. E. Bredon, *Introduction to Compact Transformation Groups*, Academic Press, London and New York, (1972).

[C] F. R. Cohen, *Orders of Certain Compositions*, Seminar on Current Trends in Algebraic Topology (1981), CMS Conference Proceedings, vol. 2 part 1, Amer. Math. Soc., (1982), 289–295.

[KL] D. Kraines and T. Lada, *A Counterexample to the Transfer Conjecture*, Lecture Notes in Mathematics #741, Springer-Verlag, New York, Heidelberg, Berlin, (1979), 588–624.

[R1] W. J. Ralph, *Category and group rings in homotopy theory*, Trans. Amer. Math. Soc., **299** (1987), 205–223.

[R2] ——, *An extension of singular homology to Banach algebras*, Pacific J. Math., **123**, no. 2 (1986), 391–405.

[R3] ——, *The coherence number of a finite group*, J. Algebra, **126** (1989), 61–79.

[St] J. Stasheff, *H-Spaces from a Homotopy Point of View*, Lecture Notes in Mathematics #161, Springer-Verlag, New York, Heidelberg, Berlin, (1970).

[Sw] R. M. Switzer, *Algebraic Topology-Homotopy and Homology*, Springer-Verlag, New York, Heidelberg, Berlin, (1975).

Received March 5, 1990 and in revised form April 25, 1992.

BROCK UNIVERSITY
ST. CATHERINE'S, ONTARIO
CANADA L2S 3A1

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

MAPS BETWEEN SEIFERT FIBERED SPACES OF INFINITE π_1

YONGWU RONG

A theorem of A. Edmonds says that any nonzero degree map between closed surfaces is homotopic to a composition of a pinch map and a branched covering. Here we consider the analogous problem in dimension three. We prove that any nonzero degree map between P^2-irreducible Seifert fibered spaces of infinite π_1 is homotopic to a composition of "vertical pinches" and a fiber preserving branched covering, except for a few cases which we describe completely. In particular, any such degree one map is homotopic to a composition of vertical pinches.

0. Introduction. In this paper we study nonzero degree maps between closed P^2-irreducible Seifert fibered spaces of infinite π_1, or equivalently, closed aspherical Seifert fibered spaces. We prove that any such map is homotopic to a composition of vertical pinches (defined in §1) and a fiber preserving branched covering, except for certain cases which can be completely understood (Theorem 3.2). As a corollary, any degree one map between such spaces is homotopic to a composition of vertical pinches.

The analogous theorem for surfaces was proved by A. Edmonds [**1**]. Later R. Skora gave a simplified proof using the notion of geometric degree [**6**]. Our proof uses similar ideas as theirs. Some extra work must be done to adjust the map so that it is nice with respect to the Seifert fibrations of the manifolds.

In §1 we establish terminology. Pinches and squeezes are defined by analogy with those definitions in dimension two given by A. Edmonds [**1**]. In §2 we show our map can be homotoped into an equivariant fiber preserving map, but possibly followed by a covering between Euclidean manifolds (those which have the geometry of E^3). In §3 we give an inductive proof of our main theorem.

I would like to thank R. Fintushel for helpful conversations, and R. Skora for his useful comment.

1. Notations and terminology. For a Seifert fibered space M, h denotes either a regular fiber or its homotopy class. Tori and annuli are often regarded as Seifert fibered without singular fibers, and h has a similar meaning in these cases. For two or more spaces which are Seifert fibered, we use the same letter h to denote the regular fibers in them if this does not cause confusion. For an element x in a group, $\langle x\rangle$ denotes the subgroup generated by x. For a submanifold S, $N(S)$ denotes the regular neighborhood of S. I always denotes a closed interval, T denotes a torus, V a solid torus, and A an annulus. The abbreviation S.F.S. stands for Seifert fibered space. The geometric degree of a proper map $f\colon M^n \to N^n$, denoted by $G(f)$, is the least number d such that for some map g properly homotopic to f, and some disk in $\mathring{N}$, $g^{-1}(D)$ consists of exactly d disks and g maps each such disk homeomorphically onto D.

We now define pinches. Let M be a closed 3-manifold, F be a 2-sided closed surface which separates M into a union of M_1 and M_2. If there is a map q from M_2 onto a handlebody H such that $q|\partial$ is a homeomorphism, then we have a degree one map f $(= \mathrm{id} \cup q)$ from M to $N = M_1 \cup_F H$. We call such a map a 1-*pinch*. When F is a 2-sphere, a 1-pinch is a pinch in the usual sense, which we may call a "0-pinch". The following lemma says that a 1-pinch can always be homotoped so that it "pinches" M_2 onto a 1-dimensional complex. The proof is simple and is omitted.

LEMMA 1.1. *If $f\colon (W, \partial W) \to (H, \partial H)$ is a map from a 3-manifold onto a handlebody such that $f|\partial$ is a homeomorphism, then f can be homotoped* $\mathrm{rel}\,\partial$ *such that f sends a collar $\partial W \times [0, 1)$ of ∂W homeomorphically onto $H - c$, and sends $W - \partial W \times [0, 1)$ onto c, where c is a core of H.*

Proof. Let $s\colon \partial W \times [0, 1] \to [0, 1]$ be the projection. Let $r_t\colon H \to H$ be a deformation retract of H so that $r_0 = \mathrm{id}$, and $r_1(H) \subset c$. Let $f_t\colon W \to H$ be defined by $f_t(x) = r_{ts(x)} \circ f(x)$. It is easy to verify f_t is the desired homotopy. □

The next lemma tells us when a 1-pinch can occur:

LEMMA 1.2. *Let W be a compact 3-manifold with $\partial W \cong F$, a connected orientable surface. Then there exists a map $f\colon W \to H$ with $f|\partial$ is a homeomorphism if and only if there are g $(= \mathrm{genus}(F))$ disjoint simple closed curves on F which cut F into a $2g$-punctured sphere and bound disjoint surfaces in W.*

Proof. Let $D = \bigcup D_i$ be a system of meridian disks in H. If there is such a map f, homotop $f\,\mathrm{rel}\,\partial$ so that f is transverse to D.

Then $f^{-1}(D)$ is a surface whose boundary is a collection of s.c.c. on F satisfying the conclusion.

If there is such a collection of s.c.c. $\{l_1, \dots, l_g\}$ on F such that $l_i = \partial F_i$, and F_i's are disjoint, let H be the handlebody obtained by adding 2-handles to $F \times I$ along each l_i, and then a 3-handle to cap off the 2-sphere. There is a map from $N(\bigcup F_i)$ onto the union of the 2-handles by pinching. The map can then be extended over the remaining part onto the 3-handle by the Tietze extension theorem. □

A manifold W satisfying the above lemma is called *pinchable*.

If M is a Seifert fibered space and T is a separating vertical torus in M such that one side of T is pinchable, then the resulting manifold N after the pinching is again a Seifert fibered space with an induced Seifert fibration from M. This is true because in M_2, the fiber h is not null-homologous. Such a 1-pinch is called a *vertical* 1-*pinch* or simply a *vertical pinch.*

Next we define squeezes. Let T be an incompressible torus in a 3-manifold M with a product neighborhood $T \times I$, l be an essential simple closed curve on T. Parameterize T by $T = S^1 \times S^1$ such that $l = S^1 \times \{p\}$. Let $X = M/\sim$, where $(x, y, t) \sim (x', y, t)$ for $(x, y, t), (x', y, t) \in S^1 \times S^1 \times I$. The quotient map $q \colon M \to X$ is called a *squeeze.* Topologically X is M cut open along T, union two solid tori along the boundary such that each meridian is identified with a copy of l, and then union an annulus connecting the cores of the two tori. If T is a vertical torus in a Seifert fibered space, a squeeze along T is called a *vertical squeeze.*

If a map $f \colon M \to N$ factors through a squeeze, then we say f admits a squeeze. If N is aspherical, then f admits a squeeze along a torus T iff f sends an essential s.c.c. on T onto a null-homotopic loop in N.

2. Equivariant fiber preserving maps. Let M be a Seifert fibered space. Fix a Seifert fibration of M and regard M as an S^1-bundle over its base orbifold O_M [**5**]. The transition group is SO(2) if the bundle is orientable and is O(2) otherwise. In any case, there is a well-defined local S^1-action on M. We denote the image of the action by tx for $t \in S^1$ and $x \in M$. Globally tx is well-defined up to changing t into t^{-1}.

Let a be a path in M connecting two fibers c_0 and c_1. An S^1-action tx on c_0 extends along a to an S^1-action $t_a x$ on c_1. If a' is another arc connecting c_0 and c_1, then $t_a x = t_{a'} x$ or $t_{a'}^{-1} x$

depending whether $w_1(a \cup a') = 0$ or 1, where w_1 is the first Stiefel-Whitney class of the bundle. In particular, the S^1-action is globally well-defined iff the bundle is orientable.

We say a map $f\colon M \to N$ is an *equivariant fiber preserving map* (EFPM) if f is a bundle homomorphism. By definition f is EFPM iff f is fiber preserving and $f(t(x)) = t^n f(x)$, where n is the integer for which $f_*(h) = h^n$. (n is globally well-defined when both M and N are oriented bundles, otherwise $|n|$ is well-defined and will be called the geometric fiber degree of f in §3.)

Similar definitions can be made for annuli, tori, and Klein bottles when they are regarded as Seifert fibered spaces.

We hope to show in this section that any nonzero degree map between aspherical S.F.S. M and N is homotopic to an EFPM for some Seifert fibrations of M and N. However, the following example shows that this is not true in general.

EXAMPLE. Let $M = F_g \times S^1$, where F_g is a closed orientable surface of genus > 1. Let $S = S^1 \times S^1 \times S^1$, N be the unique S.F.S. with orbifold $S^2(3, 3, 3)$ that is covered by S. Let $\alpha = p \times \mathrm{id}\colon M \to S$, where p is a (2-dimensional) pinch from F_g onto $S^1 \times S^1$. Hence α sends the fiber of M onto the last S^1-factor of S. Let $\beta\colon S \to N$ be a covering that sends the first S^1-factor onto the fiber of N. Now define $F = \beta \circ \alpha$. Under the unique Seifert fibrations of M and N, $f_*(h) \notin \langle h \rangle$. Hence f is not homotopic to a fiber preserving map.

The core of the above example is that S has two non-isotopic Seifert fibrations, so that α and β are both fiber preserving but under different Seifert fibrations of S. In other words, when a map factors through S, S switches its fibration. Since we restrict ourself to P^2-irreducible manifolds with infinite π_1, such S must be a Euclidean manifold [5]. There are only finitely many of such Seifert fibered spaces. We will show that this is the only way our map fails to be an EFPM (Lemma 2.1 and Proposition 2.4).

LEMMA 2.1. *Let $f\colon M \to N$ be a map of nonzero degree between aspherical S.F.S. of infinite π_1. Then, for any fixed Seifert fibration of M, either*

1. *there is a Seifert fibration of N such that $f_*(h) \in \langle h \rangle$, or*
2. *there is a covering $p\colon \widetilde{N} \to N$, a lifting $\tilde{f}\colon M \to \widetilde{N}$ such that $f = p \circ \tilde{f}$, and $\tilde{f}_*(h) \in \langle h \rangle$ for some fibration of $\widetilde{N}$. Furthermore, N and $\widetilde{N}$ are both Euclidean manifolds in this case.*

Proof. If $f_*(h) = 1$, $f_*(h) \in \langle h \rangle$. (In fact this can happen only when N is a solid torus since $\deg f \neq 0$.)

Now let us assume that $f_*(h) \neq 1$. If f_* is onto, then $\langle f_*(h) \rangle$ is an infinite cyclic normal subgroup of $\pi_1(N)$. By [**2**, VI.11(e)], $f_*(h) \in \langle h \rangle$ for some Seifert fibration of N.

If f_* is not onto, let $p\colon \widetilde{N} \to N$ be the covering corresponding to $f_*\pi_1(M)$. It must be a finite covering for $\deg f \neq 0$. Let $\tilde{f}\colon M \to \widetilde{N}$ be the lift of f. Since $\tilde{f}_*$ is onto, $\tilde{f}_*(h) \in \langle h \rangle$ for some Seifert fibration of $\widetilde{N}$. $\widetilde{N}$ also has a Seifert fibration induced from N by the map p so that p is fiber preserving covering. If the two Seifert fibrations of $\widetilde{N}$ agree, then $f_*(h) \in \langle h \rangle$, and we are in case 1. If they do not agree, then $\widetilde{N}$ has two Seifert fibrations that are not isotopic to each other. By [**5**], $\widetilde{N}$ is Euclidean, and therefore N is Euclidean since it is a quotient of $\widetilde{N}$. □

From now on, we will focus on maps of case 1 in the above lemma.

LEMMA 2.2. *Let $f\colon A \to N$ be a map from an annulus into an aspherical Seifert fibered space such that f sends ∂A into fibers and $f|\partial A$ is an EFPM. Then f can be homotoped* rel ∂ *such that f is an EFPM.*

Proof. Let $\partial A = a_0 \cup a_1$, a be a spanning arc of A. We consider 3 cases:

Case 1. Both $f(a_0)$ and $f(a_1)$ belong to regular fibers. Let $f(tx) = t^n f(x)$ on a_0. Define f_1 on A by $f_1(tx) = t^n f(x)$ for $t \in S^1$, $x \in a$. Then f_1 is an EFPM, and $f_1 = f$ on ∂A. Using $\pi_2(N) = \{0\}$, and $f_1 = f$ on a, we can easily see that $f_1 \simeq f \operatorname{rel} \partial$.

Case 2. $f(a_0)$ belongs to a regular fiber but not $f(a_1)$. Let c_1 be the singular fiber containing $f(a_1)$. The index p_1 of c_1 must divide $\deg\{f|\colon a_1 \to c\}$. Let a_1' be a parallel copy of a_1 in A, and $A_1 \subset A$ be the annulus between a_1' and a_1. We can homotop $f \operatorname{rel} \partial$ so that it is an EFPM on A_1, and $f(a_1')$ belongs to a regular fiber. By the previous case the result holds.

Case 3. $f(a_0)$ and $f(a_1)$ belong to singular fibers c_0 and c_1 respectively. Suppose that $f(a_i)$ covers c_i with degree n_i. Let O_N be the base orbifold of N and $\widetilde{O}_N$ be its universal covering. Let $\bar{c}_i \in \pi_1(O_N)$ be the coset of c_i in the quotient group $\pi_1(N)/\langle h \rangle \cong \pi_1(O_N)$. The map f gives a conjugacy relation $\bar{c}_0^{n_0} = \alpha \bar{c}_1^{n_1} \alpha^{-1}$ in $\pi_1(O_N)$. But $\bar{c}_i$ acts on $\widetilde{O}_N$ as a rotation around some point over the ith cone point. Hence the relation cannot hold unless either the two rotation angles are both multiples of 2π, or the two rotations have the same fixed point and α is also a rotation around this point. The first possibility implies that n_i is a multiple of the index of c_i, and the result then follows by a similar reason as in Case 2. For the second possibility, $c_0 = c_1$, and $f(a)$ is a loop whose homotopy class lies in $\langle c_0 \rangle$. Hence f is homotopic to the constant fiber c_0. □

Similar to Case 1 above, the following lemma is proved using $\pi_3(N) = \{0\}$:

LEMMA 2.3. *Let* $f\colon V \to N$ *be a map from a Seifert fibered torus of type* $(1, 0)$ *into an aspherical Seifert fibered space* N *such that* $f|\partial V$ *is an EFPM. Then* $f \simeq f_1 \operatorname{rel} \partial$, *where* f_1 *is an EFPM.*

PROPOSITION 2.4. *Let* $f\colon (M, \partial M) \to (N, \partial N)$ *be a nonzero degree map between aspherical S.F.S. of infinite* π_1 *such that* $1 \neq f_*(h) \in \langle h \rangle$. *Then* f *can be homotoped into an EFPM. If* $f|\partial$ *is already an EFPM, then the homotopy can be chosen to be fixed on the boundary.*

Proof. Consider first the case when M and N are closed. We can write M as a union $M = H_0 \cup H_1 \cup H_2$, where $H_0 = N(h \cup c_i)$, h is a regular fiber, $\bigcup c_i$ is the union of all the singular fibers, $H_1 = N(\bigcup A_j)$, where A_j's are disjoint vertical annuli such that $N(\partial A_j) \subset \partial H_0$, and $H_2 = N(h_1)$, where h_1 is a regular fiber and $\partial H_2 \subset \partial(H_0 \cup H_1)$. (This is called a round-handle decomposition of M [**3**].)

Since $f_*(h) \in \langle h \rangle$, $f_*(c_i)$ is a multiple of a fiber for each i [**2**, VI.11(f)]. Hence we can homotop f so that $f|H_0$ is EFPM. Then we apply Lemma 2.2 and Lemma 2.3 to finish the proof.

For manifolds with boundary, we first use the fact $1 \neq f_*(h) \in \langle h \rangle$ to homotop $f|\partial$ so that $f|\partial$ is an EFPM. The rest is similar as before. □

Let Z_n denote the cyclic subgroup of order n in S^1. Let $\sim_n$ be the equivalence relation on the S.F.S. M defined by $x \sim_n t(x)$ for all $x \in M$ and all $t \in Z_n$. Let $M_n = M/\sim_n$. Then M_n is a Seifert fibered space, and the quotient map is a fiber preserving branched covering. An EFPM f with $f_*(h) = h^n$ satisfies $f(t(x)) = t^{\pm n}(f(x))$. Thus $f(t(x)) = f(x)$ for all $t \in Z_n$. It follows that f factors through M_n:

COROLLARY 2.5. *Under the same hypothesis as in Proposition 2.4,* $f \simeq f_1 \circ p$, *where* $p\colon M \to M_1$ *is a fiber preserving branched covering, and* $f_{1*}(h) = h$.

3. Main theorem. We prove our main theorem in this section. First we give some definitions. Let $f\colon (M, \partial M) \to (N, \partial N)$ be a map of Seifert fibered spaces which is homotopic to a fiber preserving map. The *geometric orbifold degree* of f, denoted by $G_{ob}(f)$, is the minimum number of regular fibers in $f_1^{-1}(h)$, where f_1 ranges in all the maps which are properly homotopic to f, fiber preserving, and transverse to h. The *geometric fiber degree* of f, denoted by $G_h(f)$, is $|n|$ where $f_*(h) = h^n$. Notice that these definitions depend on the choice of Seifert fibrations of the spaces.

Denote by f_∂ the restriction of f on the boundary. We have

$$G(f) \leq G_h(f)G_{ob}(f) \leq G(f_\partial).$$

We say a fiber preserving map f is *allowable* if f_∂ is a covering of degree $G_{ob}(f)G_h(f)$. If $\partial M = \partial N = \varnothing$, then any fiber preserving map is allowable. From the definition, the following observation is easy to see and will be used later in our proof:

REMARK 1. If $f\colon M \to N$ is allowable, X is a vertical set (vertical annulus, vertical solid torus, etc.) in N, and $f_1 = f|\colon M - f^{-1}(N(X)) \to N - N(X)$, then f_1 is allowable if either

$f|f^{-1}(X)$ is a covering of degree $G_{ob}(f)G_h(f)$, or
$f|f^{-1}(X)$ is a covering of degree $G(f)$. In the later case, $G(f) = G_{ob}(f)G_h(f) = G(f_\partial)$.

PROPOSITION 3.1. *Let f be an allowable map between aspherical S.F.S. of infinite π_1 with $G(f) \neq 0$. Then $F \simeq g\pi$, where π is a composition of finitely many of vertical pinches, and g is a fiber preserving branched covering branched over fibers.*

This proposition together with Lemma 2.1 and Proposition 2.4 imply our main theorem:

THEOREM 3.2. *Let $f\colon M \to N$ be a nonzero degree map between closed P^2-irreducible S.F.S. of infinite π_1. Then $f \simeq p \circ g \circ \pi$, where π is a composition of finitely many vertical pinches, g is a fiber preserving branched covering, and p is a covering.*

REMARK 2. The covering p can be chosen to be the identity map unless N is an Euclidean manifold. There are only 10 such manifolds N **[5]**. For each such N, the covering space $\tilde{N}$ must have two different Seifert fibrations and is Euclidean. There are only two such manifolds $\widetilde{N}$, the three torus and the double of the twisted I-bundle of the Klein bottle. Hence the possibilities of p are very limited.

When $G(f) = 1$, we have

COROLLARY 3.3. *Any degree one map between closed P^2-irreducible Seifert fibered spaces of infinite π_1 is a composition of finitely many vertical pinches.*

Before we prove Proposition 3.1, we first prove a lemma which serves the initial step of the induction.

LEMMA 3.4. *Let $f\colon M \to V$ be an allowable map, where M is a S.F.S. of infinite π_1, and V is a Seifert fibered solid torus. Assume that f does not admit a nontrivial vertical pinch, and $G_h(f) = 1$. Then f can be homotoped* $\operatorname{rel}\partial$ *to a fiber preserving branched covering.*

Proof. Since F has no nontrivial vertical pinch, the orbifold O_M of M must be a planar surface with at most one cone point. Let $a_1, \dots, a_k$ be proper arcs which cut O_M into a disk with possibly one cone point, let A_i be the annulus over a_i. Then $\bigcup A_i$ cut M into a Seifert fibered solid torus V_1. The map $f|A_i$ can be homotoped $\operatorname{rel}\partial$ so that $f(A_i)$ is ∂-parallel. Now $f|V_1$ is a map between tori, hence can be homotoped $\operatorname{rel}\partial$ such that off ∂V_1 it is a branched covering branched over the core c of V. It follows from Remark 1 that $f_1\colon M_1 \to N_1$ is allowable and $G_{ob}(f_1) = G_{ob}(f)$, where $N_1 = (N - N(c))^-$, $M_1 = f^{-1}(N_1)$, $f_1 = f|M_1$.

The Seifert fibered space N_1 is isomorphic to a product $A \times S^1$ where A is a horizontal annulus. Homotop $f_1 \operatorname{rel}\partial$ such that $f_1^{-1}(A)$ is incompressible. It must be either vertical or horizontal by [7]. But it cannot be vertical since $f_{1*}(h) = h$ in $\pi_1(N_1)$. Hence $f_1^{-1}(A)$ must be a union of parallel horizontal surfaces $\bigcup F$, and M_1 is a fibered manifold with fiber F and a periodic gluing map. Using $G_h(f_1) = 1$, we conclude that there is only one copy of F and the gluing map must be the identity. The map $f_1|\colon f^{-1}(A) \to A$ is allowable as a map of surfaces (note that Edmonds and Skora both defined allowable maps, but the two definitions are in fact equivalent by [**6**]); also it does not admit a pinch since f does not admit a vertical pinch. By Edmonds theorem, it is homotopic $\operatorname{rel}\partial$ to a branched covering g. This implies f_1 is homotopic $\operatorname{rel}\partial$ to $g \times \mathrm{id}\colon F \times S^1 \to A \times S^1$, which is a fiber preserving branched covering. Hence f is homotopic $\operatorname{rel}\partial$ to a fiber preserving branched covering. $\square$

Proof of Proposition 3.1. After a finite number of vertical pinches we may assume that f has no vertical pinch. Next by Corollary 2.5, we can compose f as $f = f_1 \circ g$, where g is a fiber preserving branched covering branched over fibers and $G_h(f_1) = 1$. Note that f_1 does not admit a vertical pinch since a vertical pinch of f_1 would yield a vertical pinch of f. Therefore we may assume that $G_h(f) = 1$.

If $\partial M = \varnothing$, let h be a regular fiber in N such that f is transverse to h, and $f^{-1}(h)$ is a union of $G_{ob}(f)$ regular fibers. We then consider $f|M - N(f^{-1}(h))$, which is again an allowable map of the same geometric orbifold degree. Hence we may assume that $\partial M \neq \varnothing$.

From now on, we assume that $\partial M \neq \varnothing$, f has no vertical pinch, and $G_h(f) = 1$. Under these assumptions we prove that f is homotopic to a fiber preserving branched covering branched over fibers by an induction on the complexities of M and N.

The first step is the case when $N \cong V$. This has been done by Lemma 3.4.

Next is the inductive step:

Case 1. f admits a squeeze along an incompressible vertical torus T.

We take a maximum collection of disjoint non-parallel incompressible vertical tori along which f admits squeezes. Let $X = Q \cup A$ be the space obtained after these squeezes, where $Q = \bigcup Q_i$ is a union of S.F.S., $A = \bigcup A_j$ is a union of annuli such that ∂A is a disjoint union of (possibly singular) fibers in Q. After a homotopy, $f = (g \cup \alpha) \circ q$, where Q is a composition of squeezes, $g = \bigcup g_i$ is defined on Q, and α is defined on A. Each Q_i must be of infinite π_1 for otherwise $f_*(h)$ factors through an element of finite order in $\pi_1 Q_i$ which implies $f_*(h) = 1$ in $\pi_1(N)$, and thus $G(f) = 0$.

Claim. g is allowable and $G_{ob}(g) = G_{ob}(f)$, $G_h(g) = 1$.

Proof of Claim. Clearly, $G_h(g) = G_h(f) = 1$. Next, fix a regular fiber h in N, homotop g_i so that g_i is fiber preserving and $g_i^{-1}(h)$ is a union of $G_{ob}(g_i)$ regular fibers in M. By Lemma 2.2 we can homotop $\alpha \operatorname{rel} \partial$ so that α is fiber preserving; hence we can perturb it so that $\alpha(A)$ misses h. So $f^{-1}(h) = g^{-1}(h)$. Hence $G_{ob}(f) \leq G_{ob}(g)$. It follows that $G(f_\partial) = G_{ob}(f)G_h(f) \leq G_{ob}(g)G_h(g) \leq G(g_\partial)$. Since $f_\partial = g_\partial$, all inequalities must be equalities. Hence

$G_{ob}(g)G_h(g) = G(g_\partial)$, and thus g is allowable. It also follows from the above equalities that $G_{ob}(g) = G_{ob}(f)$.

By the maximality of q, g does not admit any squeeze, thus no vertical pinch.

For each i such that Q_i is closed, by Lemma 3.5 below we can homotop g_i such that $g_i(Q_i)$ is contained in a fiber of N. Let B_1 be the union of such fibers in N.

For each i such that Q_i has boundary, $G_{ob}(g_i) \neq 0$ since g is allowable. By the inductive hypothesis, each g_i is homotopic to a fiber preserving branched covering branched over fibers. Let B_2 be the union of the branched fibers in N.

Now consider $\alpha: A \to N$. $\alpha(\partial A) = g(\partial A)$ must be fibers in N. By Lemma 2.2, we may assume that α is fiber preserving after a homotopy $\operatorname{rel}\partial$.

If a component of ∂A_j is a singular fiber c in Q of order n (> 1), $\alpha(A_j)$ must be contained in one singular fiber. Otherwise, by Lemma 2.2, α sends a fiber c' of A_j in $\operatorname{Int} A_j$ into a regular fiber in N. So $\alpha(c') \simeq h^k$ and thus $g(c) \simeq \alpha(c') \simeq h^k$. This implies that $g(h) \simeq g(c^n) \simeq h^{nk}$, contradicting the fact that $G_h(g) = G_h(f) = 1$.

If both components of ∂A_j are regular fibers in Q, we can perturb them so that they are mapped into different regular fibers in N under g. By Lemma 2.2, $\alpha(A_j)$ can be homotoped into an immersed vertical annulus. Thus we can homotop $\alpha \operatorname{rel}\partial$ so that $\alpha(A_j)$ is an embedded vertical annulus by sliding the double lines off $\alpha(\partial A_j)$.

It follows that in any case, $\alpha(A_j)$ is contained in a vertical solid torus after a homotopy $\operatorname{rel}\partial$. Let B_3 be the union of these vertical solid tori in N.

Let E be a vertical essential annulus in N missing $B_1 \cup B_2 \cup B_3$. Now $f^{-1}(E)$ covers E with degree $G_h(f)G_{ob}(f)$, thus consists of a collection of annuli. The map $f|: M - N(f^{-1}(E)) \to N - N(E)$ is again allowable, so it is homotopic $\operatorname{rel}\partial$ to a fiber preserving branched covering by the inductive hypothesis. Hence f is homotopic to a fiber preserving branched covering.

Case 2. f does not admit a squeeze along any incompressible vertical torus. Since $\partial N \neq \varnothing$, and $N \not\cong V$, there is an essential vertical annulus E in N. Homotop f such that $f^{-1}(E)$ is incompressible. It must be either vertical or horizontal [7]. If F is a horizontal component of $f^{-1}(E)$, then the subgroup A generated by $\pi_1(F)$ and h has finite index in $\pi_1(M)$. Let $B = \pi_1(E)$, which is an infinite index subgroup of $\pi_1(N)$. Since $f_*(A) \subset B$, we have

$$\begin{aligned}[\pi_1 N : B] &\le [\pi_1 N : f_*(A)] = [\pi_1 N : f_*\pi_1 M][f_*\pi_1 M : f_*A] \\ &\le [\pi_1 N : f_*\pi_1 M][\pi_1 M : A] < \infty, \quad \text{a contradiction}.\end{aligned}$$

Hence $f^{-1}(E)$ must be a union of incompressible vertical tori or annuli. It must be a union of annuli because f does not admit a squeeze. Denote the annuli by $\{E_i\}$. After a homotopy of f, $f|E_i$ either covers E or misses a boundary component of E. But the later case cannot happen, for otherwise $f^{-1}(h)$ has less than $G_{ob}(f)$ components for a regular fiber h on E that is close to ∂E, a contradiction. It follows that $f|M - N(f^{-1}(E))$ is an allowable map onto $N - N(E)$. Therefore the inductive hypothesis implies it can be homotoped rel ∂ into a fiber preserving branched covering. Hence f is homotopic to a fiber preserving branched covering. □

LEMMA 3.5. *Let $g\colon M \to N$ be a map of aspherical Seifert fibered spaces of infinite π_1 such that $1 \ne f_*(h) \in \langle h\rangle$. Assume that M is closed and $\partial N \ne \varnothing$. Then either f admits a squeeze along an incompressible vertical torus or $f \simeq f_1$, where $f_1(M) \subset$ a fiber of N.*

Proof. Let A be a union of essential vertical annuli in N which cut N into a union of solid tori. Homotop f so that $f^{-1}(A)$ is incompressible in M.

Case 1. If a component $F \subset f^{-1}(A)$ is horizontal, the group $G = \langle \pi_1(F), h\rangle$ is of finite index in $\pi_1(M)$. But $f_*(G)$ is cyclic since it is contained in an annulus group. Hence $f_*(\pi_1 M)$ contains a cyclic group of finite index. Therefore if M contains an incompressible vertical torus then f admits a squeeze along this torus. If M does not contain any incompressible vertical torus then M is a Seifert fibered space whose base orbifold is S^2 with three exceptional fibers. In this case $G = \pi_1(M)$. Thus $f_*(\pi_1 M) = f_*(G)$ is a cyclic group with a generator x. Since this group contains $f_*(h)$, and $f_*(h) \in \langle h\rangle$, x is represented by a power of fiber c in N [2, VI.11(f)]. Hence $f_*(\pi_1 M) \subset \langle c\rangle$, and so $f \simeq f_1$ where $f_1(M) \subset c$.

Case 2. If $f^{-1}(A)$ is a nonempty union of incompressible vertical tori, then f admits squeezes along these tori.

Case 3. If $f^{-1}(A) = \varnothing$, then $f(M)$ is contained in one of the tori of N cut along A. Hence $f(M)$ can be homotoped into the core of this torus, which is a fiber of N. □

REFERENCES

[1] A. Edmonds, *Deformation of maps to branched coverings in dimension two*, Ann. of Math., **110** (1979), 113–125.

[2] W. Jaco, *Lectures on three-manifold topology*, CBMS Lecture Series No. 43, Amer. Math. Soc., (1980).

[3] J. Morgan, *Nonsingular Morse-Smale flows on 3-dimensional manifolds*, Topology, **18**, No. 1 (1979), 41–53.

[4] Y. Rong, *Degree one maps between geometric 3-manifolds*, Trans. Amer. Math. Soc., **332** (1992), 411–436.

[5] P. Scott, *The geometries of 3-manifolds*, Bull. London Math. Soc., **15** (1983), 401–487.

[6] R. Skora, *The degree of a map between surfaces*, Math. Ann., **276**, No. 3 (1987), 415–423.

[7] F. Waldhausen, *Eine Klasse von 3-dimensionalen Mannigfaltigkeiten*, II, Invent. Math., **4** (1967), 87–117.

Received June 10, 1991 and in revised form July 14, 1992.

MICHIGAN STATE UNIVERSITY
EAST LANSING, MI 48824
E-mail address: rong@math.gwu.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

SOME NUMERIC RESULTS ON ROOT SYSTEMS

JIAN-YI SHI

Let Φ be an irreducible root system (sometimes we denote Φ by $\Phi(X)$ to indicate its type X). Choose a simple root system Π in Φ. Let Φ^+ (resp. Φ^-) be the corresponding positive (resp. negative) root system of Φ. By a subsystem Φ' of Φ (resp. of Φ^+), we mean that Φ' is a subset of Φ (resp. of Φ^+) which itself forms a root system (resp. a positive root system). We refer the readers to Bourbaki's book for the detailed information about root systems. Among all subsystems of Φ, the subsystems of Φ of rank 2 and of type $\neq A_1 \times A_1$ are of particular importance in the theory of Weyl groups and affine Weyl groups (see the papers by Jian-yi Shi). In the present paper, we shall compute the number of such subsystems of Φ for an irreducible root system Φ of any type. Some interesting properties of Φ are also obtained.

1. The number $h(\alpha)$. Let $\langle\ ,\ \rangle$ be an inner product of the euclidean space E spanned by Φ. For any $\alpha \in \Phi$, we denote by $|\alpha|$ the length of α, by $\alpha^\vee$ the dual root $2\alpha/\langle\alpha, \alpha\rangle$ of α and by s_α the reflection in E which sends any vector $v \in E$ to $s_\alpha(v) = v - \langle v, \alpha^\vee\rangle\alpha$. For $\alpha, \beta \in \Phi$, we write $\alpha < \beta$ if $\beta - \alpha$ is a sum of some positive roots.

For $\alpha \in \Phi$, we define the sets $D(\alpha) = \{\beta \in \Phi \mid \alpha + \beta \in \Phi\}$, $D^+(\alpha) = D(\alpha) \cap \Phi^+$ and $D^-(\alpha) = D(\alpha) \cap \Phi^-$. Let $d(\alpha)$ be the cardinality of the set $D^+(\alpha)$. Also, we denote by $\operatorname{ht}(\alpha)$ the height of α, i.e. $\operatorname{ht}(\alpha) = \sum_{\beta\in\Pi} a_\beta$ if $\alpha = \sum_{\beta\in\Pi} a_\beta \beta$ with $a_\beta \in \mathbb{Z}$.

For any $\alpha \in \Phi^+$, there exists a sequence ξ of roots $\alpha_1 = \alpha, \alpha_2, \ldots, \alpha_r$ in Φ^+ such that $\alpha_r \in \Pi$ and for every i, $1 < i \leq r$, we have $\alpha_{i-1} > \alpha_i = s_{\delta_i}(\alpha_{i-1})$ for some $\delta_i \in \Pi$. Such a sequence ξ is called a root path from α to Π. We denote by $h(\alpha, \xi)$ the length r of ξ. We shall deduce a formula for the number $h(\alpha, \xi)$, from which we shall see that $h(\alpha, \xi)$ is actually independent on the choice of a root path ξ from α to Π but only dependent on the root α.

Note that if the root system Φ contains roots of two different lengths and if $\alpha = \sum_{\beta\in\Pi} a_\beta \beta$ is a long root of Φ with $a_\beta \in \mathbb{Z}$ then each coefficient a_β with β short is divisible by $|\alpha|^2/|\beta|^2$.

LEMMA 1.1. *Let $\alpha = \sum_{\beta\in\Pi} a_\beta \beta$, $a_\beta \in \mathbb{Z}$, be a root of Φ^+ and let ξ be a root path from α to Π. Then*

(i) *If either all the roots of* Φ *have the same length or* α *is a short root of* Φ *with* Φ *containing roots of two different lengths, then* $h(\alpha, \xi) = \mathrm{ht}(\alpha)$;

(ii) *If* α *is a long root of* Φ *with* Φ *containing roots of two different lengths, then*

$$h(x, \xi) = \sum_{\beta\in\Pi} \frac{|\beta|^2}{|\alpha|^2} a_\beta.$$

Proof. Let $\alpha_1 = \alpha, \alpha_2, \ldots, \alpha_r$ be a root path from α to Π. Then in case (i), we have $\mathrm{ht}(\alpha_i) = \mathrm{ht}(\alpha_{i+1}) + 1$ for any i, $1 \le i < r$, by the fact that $\langle \alpha_i, \delta_i^\vee \rangle = 1$, where $\delta_i \in \Pi$ satisfies the relation $\delta_i(\alpha_{i-1}) = \alpha_i$. So assertion (i) follows immediately by applying induction on $\mathrm{ht}(\alpha) \ge 1$. Next assume that we are in case (ii). Again apply induction on $\mathrm{ht}(\alpha) \ge 1$. If $\mathrm{ht}(\alpha) = 1$, then $\alpha \in \Pi$ and the result is obviously true. Now assume $\mathrm{ht}(\alpha) > 1$. Let $\xi : \alpha_1 = \alpha, \alpha_2, \ldots, \alpha_r$ be a root path from α to Π. Then $\xi' : \alpha_2, \alpha_3, \ldots, \alpha_r$ is a root path from α_2 to Π with $\mathrm{ht}(\alpha_2) < \mathrm{ht}(\alpha)$ and $\alpha_2 = s_\delta(\alpha)$ for some $\delta \in \Pi$. Note that α_2 is a long root of Φ. Write

$$\alpha_2 = \sum_{\beta\in\Pi} a'_\beta \beta, \qquad a'_\beta \in \mathbb{Z}.$$

Then by inductive hypothesis, we have

$$h(\alpha_2, \xi') = \sum_{\beta\in\Pi} \frac{|\beta|^2}{|\alpha_2|^2} a'_\beta.$$

Since $\langle \alpha, \delta^\vee \rangle = |\alpha|^2/|\delta|^2$ by the assumption $s_\delta(\alpha) < \alpha$, we have

$$\alpha = \alpha_2 + \frac{|\alpha|^2}{|\delta|^2}\delta = \sum_{\substack{\beta\in\Pi \\ \beta\neq\delta}} a'_\beta \beta + \left(a'_\delta + \frac{|\alpha|^2}{|\delta|^2}\right)\delta.$$

This implies that

$$\begin{aligned} h(\alpha, \xi) = h(\alpha_2, \xi') + 1 &= \sum_{\beta\in\Pi} \frac{|\beta|^2}{|\alpha_2|^2} a'_\beta + 1 \\ &= \sum_{\substack{\beta\in\Pi \\ \beta\neq\delta}} \frac{|\beta|^2}{|\alpha_2|^2} a'_\beta + \frac{|\delta|^2}{|\alpha_2|^2}\left(a'_\delta + \frac{|\alpha_2|^2}{|\delta|^2}\right) = \sum_{\beta\in\Pi} \frac{|\beta|^2}{|\alpha|^2} a_\beta \end{aligned}$$

by noting $|\alpha| = |\alpha_2|$. □

We see from Lemma 1.1 that, for any $\alpha \in \Phi^+$, the length of a root path ξ from α to Π is only dependent on α but not on the choice of the path ξ. So we can denote $h(\alpha, \xi)$ simply by $h(\alpha)$.

Let $\Phi^\vee$ be the dual root system of Φ, i.e. $\Phi^\vee = \{\alpha^\vee \mid \alpha \in \Phi\}$. Then $\Pi^\vee = \{\alpha^\vee \mid \alpha \in \Pi\}$ and $(\Phi^\vee)^+ = \{\alpha^\vee \mid \alpha \in \Phi^+\}$ are a simple root system and the corresponding positive root system of $\Phi^\vee$, respectively. We can define the number $h^\vee(\alpha^\vee)$ for any $\alpha^\vee \in (\Phi^\vee)^+$ in the same way as that for a root of Φ. That is, $h^\vee(\alpha^\vee)$ is the length of a root path from $\alpha^\vee$ to $\Pi^\vee$ in $(\Phi^\vee)^+$.

LEMMA 1.2. *For any* $\alpha \in \Phi^+$, *we have* $h(\alpha) = h^\vee(\alpha^\vee)$.

Proof. For any $\delta \in \Pi$, we have the following equivalence.

$$s_\delta(\alpha) < \alpha \Leftrightarrow \langle \alpha, \delta^\vee \rangle > 0 \Leftrightarrow \langle \alpha^\vee, \delta \rangle > 0 \Leftrightarrow s_{\delta^\vee}(\alpha^\vee) < \alpha^\vee. \tag{1}$$

Apply induction on $h(\alpha) \geq 1$. When $h(\alpha) = 1$, we have $\alpha \in \Pi$ and hence $\alpha^\vee \in \Pi^\vee$. So $h^\vee(\alpha^\vee) = 1$, and the result is true in this case. Now assume $h(\alpha) > 1$. Then there exists some $\delta \in \Pi$ with $\langle \alpha, \delta^\vee \rangle > 0$. So $h(s_\delta(\alpha)) = h(\alpha) - 1$. By inductive hypothesis, we have

$$h(s_\delta(\alpha)) = h^\vee((s_\delta(\alpha))^\vee) = h^\vee(s_{\delta^\vee}(\alpha^\vee)). \tag{2}$$

But by (1), we have

$$h^\vee(s_{\delta^\vee}(\alpha^\vee)) = h^\vee(\alpha^\vee) - 1.$$

Thus we get $h(\alpha) = h^\vee(\alpha^\vee)$. □

2. The number $d(\alpha)$. We shall deduce a formula for the number $d(\alpha)$ for any $\alpha \in \Phi^+$.

For $\alpha, \beta \in \Phi$, we call all roots of the form $\alpha + i\beta$ $(i \in \mathbb{Z})$ the β-string through α. Let $\alpha \in \Phi^+$ and $\delta \in \Pi$ satisfy the inequality $\langle \alpha, \delta^\vee \rangle > 0$. Then it is easily seen that $\alpha, \alpha - \delta, \dots, \alpha - \langle \alpha, \delta^\vee \rangle \delta$ is the δ-string through α except for the case when α is the highest short root of the root system of type G_2.

LEMMA 2.1. *Given* $\alpha \in \Phi^+$ *and* $\delta \in \Pi$ *with* $\langle \alpha, \delta^\vee \rangle > 0$. *Let* $\alpha' = s_\delta(\alpha)$. *Then* (i) $D(\alpha') = s_\delta(D(\alpha))$.

(ii) $s_\delta(D^+(\alpha')) = D^+(\alpha) \cup \{-\delta\}$, *provided that* α *is not the highest short root of the root system of type* G_2;

(iii) $d(\alpha') = d(\alpha) + 1$ *under the same assumption as that in* (ii).

Proof. (i) $\beta \in D(\alpha') \Leftrightarrow \beta+\alpha' \in \Phi \Leftrightarrow s_\delta(s_\delta(\beta)+\alpha) \in \Phi \Leftrightarrow s_\delta(\beta)+\alpha \in \Phi \Leftrightarrow s_\delta(\beta) \in D(\alpha) \Leftrightarrow \beta \in s_\delta(D(\alpha))$.

(ii) First we shall show $s_\delta(D^+(\alpha)) \subset D^+(\alpha')$. Let $\beta \in s_\delta(D^+(\alpha))$. Then $\beta \in D(\alpha')$ by (i). If $\beta \in D^-(\alpha') \subseteq \Phi^-$, then by the fact $s_\delta(\beta) \in D^+(\alpha) \subseteq \Phi^+$, we have $\beta = -\delta$. Since $\alpha, \alpha-\delta, \ldots, \alpha - \langle \alpha, \delta^\vee\rangle\delta$ is the δ-string through α by the above remark, we see that $\alpha+s_\delta(\beta) = \alpha+\delta \notin \Phi$ which contradicts the condition $s_\delta(\beta) \in D^+(\alpha)$. Thus we have $\beta \in D^+(\alpha')$ and so $s_\delta(D^+(\alpha)) \subset D^+(\alpha')$, i.e. $D^+(\alpha) \subset s_\delta(D^+(\alpha'))$.

It is obvious that $\{-\delta\} \subseteq s_\delta(D^+(\alpha'))$. Thus it remains to show the reversing inclusion. Now assume $\beta \in s_\delta(D^+(\alpha'))$. Then $s_\delta(\beta) \in D^+(\alpha')$. This implies that $s_\delta(\beta) + \alpha' \in \Phi$ and $s_\delta(\beta) \in \Phi^+$. Hence $\beta + \alpha \in \Phi$ and $s_\delta(\beta) \in \Phi^+$. But then we have either $\beta \in D^+(\alpha)$ or $\beta = -\delta$, which implies $s_\delta(D^+(\alpha')) \subseteq D^+(\alpha) \cup \{-\delta\}$.

(iii) This is an immediate consequence of (ii). □

REMARK. In the case when the type of Φ is G_2, let $\Pi = \{\gamma, \delta\}$ with δ short. Then $D^+(2\delta+\gamma) = \{\delta, \delta+\gamma\}$, $D^+(\delta+\gamma) = \{\delta, 2\delta+\gamma\}$ and $\delta+\gamma = s_\delta(2\delta+\gamma)$. Thus the results (ii), (iii) of Lemma 2.1 do not hold in this case.

In Φ^+, let α^l be the highest long root and let α^s be the highest short root, where we stipulate $\alpha^s = \alpha^l$ in the case when all the roots of Φ have the same length.

THEOREM 2.2. *Given* $\alpha \in \Phi^+$.

(i) *If* α *is short and if the type of* Φ *is not* G_2, *then*

$$h(\alpha) + d(\alpha) = \mathrm{ht}(\alpha^l).$$

(ii) *If* α *is long, then*

$$h(\alpha) + d(\alpha) = \mathrm{ht}(\alpha^s).$$

Proof. First assume that the result has been shown to be true in the case when $\alpha = \alpha^s$ in (i) and $\alpha = \alpha^l$ in (ii). Apply reversing induction on $h(\alpha) \le h(\alpha^s)$ in (i) and on $h(\alpha) \le h(\alpha^l)$ in (ii). Now assume that α is either short with $h(\alpha) < h(\alpha^s)$ or long with $h(\alpha) < h(\alpha^l)$. Then there must exist some $\delta \in \Pi$ with $\langle \alpha, \delta^\vee\rangle < 0$. So $\alpha' = s_\delta(\alpha) > \alpha$ with $h(\alpha') = h(\alpha) + 1$. We see $\langle \alpha', \delta^\vee\rangle > 0$. By Lemma 2.1(iii), we have $d(\alpha') = d(\alpha) - 1$. So by inductive hypothesis, we get

$$\begin{aligned} h(\alpha) + d(\alpha) &= (h(\alpha') - 1) + (d(\alpha') + 1) \\ &= h(\alpha') + d(\alpha') \\ &= \begin{cases} \operatorname{ht}(\alpha^l) & \text{if } \alpha \text{ is short,} \\ \operatorname{ht}(\alpha^s) & \text{if } \alpha \text{ is long,} \end{cases} \end{aligned}$$

by noting $|\alpha| = |\alpha'|$.

Thus it remains to show that assertion (i) is true for $\alpha = \alpha^s$ and that assertion (ii) is true for $\alpha = \alpha^l$.

In the case when the Dynkin diagram is simply laced, we have $h(\alpha^s) = \operatorname{ht}(\alpha^s)$ by Lemma 1.1(i). Clearly, $d(\alpha^s) = 0$. So our result is true in this case. Now assume that Φ contains roots of two different lengths. If Φ has type B_n, then $h(\alpha^s) = n$, $d(\alpha^s) = n - 1$, $\operatorname{ht}(\alpha^l) = 2n - 1$, $d(\alpha^l) = 0$ and $h(\alpha^l) = h^\vee((\alpha^l)^\vee) = \operatorname{ht}((\alpha^l)^\vee) = \operatorname{ht}(\alpha^s) = 2n - 2$ by Lemmas 1.2 and 1.1(i). If Φ has type C_n, then $h(\alpha^s) = 2n - 2$, $d(\alpha^s) = 1$, $\operatorname{ht}(\alpha^l) = 2n - 1$ and $d(\alpha^l) = 0$. We also have

$$h(\alpha^l) = h^\vee((\alpha^l)^\vee) = \operatorname{ht}((\alpha^l)^\vee) = \operatorname{ht}(\alpha^s) = n$$

by Lemmas 1.2 and 1.1(i). If Φ has type F_4, then $h(\alpha^s) = 8$, $d(\alpha^s) = 3$, $\operatorname{ht}(\alpha^l) = 11$ and $d(\alpha^l) = 0$. By the same reason as above, we have

$$h(\alpha^l) = h^\vee((\alpha^l)^\vee) = \operatorname{ht}((\alpha^l)^\vee) = \operatorname{ht}(\alpha^s) = 8.$$

If Φ has type G_2, then $d(\alpha^l) = 0$ and $h(\alpha^l) = \operatorname{ht}(\alpha^s) = 3$. Thus in all the cases, our result is true. □

COROLLARY 2.3. *Assume that the type of* Φ *is not* G_2. *Then for any short root* α *of* Φ^+, *we have the equation*

$$\operatorname{ht}(\alpha) + d(\alpha) = h - 1,$$

where h *is the Coxeter number of* Φ.

Proof. We have $h(\alpha) = \operatorname{ht}(\alpha)$ by Lemma 1.1(i). Since $\operatorname{ht}(\alpha^l) = h - 1$, our result follows immediately from Theorem 2.2(i). □

3. The number of certain rank 2 subsystems in Φ. Let $g(\Phi)$ be the number of subsystems of Φ of rank 2 and of type other than $A_1 \times A_1$. Then $g(\Phi)$ is also equal to the number of positive subsystems of Φ^+ of rank 2 and of type $\neq A_1 \times A_1$. In this section, we shall compute the number $g(\Phi)$ for Φ of any type.

LEMMA 3.1. *If the Dynkin diagram of Φ is simply laced, then*

$$g(\Phi)=\frac{1}{2}\sum_{\alpha\in\Phi^+} d(\alpha). \tag{3}$$

Proof. Under our assumption, the only possible type for a subsystem of Φ^+ of rank 2 and of type $\neq A_1\times A_1$ is A_2. Each of such subsystems could be obtained by first taking a root $\alpha\in\Phi^+$ and then taking any root β in the set $D^+(\alpha)$ to form a subsystem $\{\alpha, \beta, \alpha+\beta\}$. Since such a subsystem is obtained twice in the above way, this implies the required formula (3) for the number $g(\Phi)$. □

Define

$$H(\Phi)=\sum_{\alpha\in\Phi^+}\mathrm{ht}(\alpha),\quad H^s(\Phi)=\sum_{\substack{\alpha\in\Phi^+\\ \text{short}}}\mathrm{ht}(\alpha)\quad\text{and}$$

$$H^l(\Phi)=\sum_{\substack{\alpha\in\Phi^+\\ \text{long}}}\mathrm{ht}(\alpha).$$

These numbers could be computed for any irreducible root system Φ. Define $\binom{m}{n}=\frac{m!}{n!(m-n)!}$ for any integers $m, n, 0\leq n\leq m$.

LEMMA 3.2.

Type of Φ	$H(\Phi)$	$H^s(\Phi)$	$H^l(\Phi)$
A_n $(n\geq 1)$	$\binom{n+2}{3}$		
B_n $(n\geq 2)$	$\frac{n(n+1)(4n-1)}{6}$	$\binom{n+1}{2}$	$4\binom{n+1}{3}$
C_n $(n\geq 2)$	$\frac{n(n+1)(4n-1)}{6}$	$\frac{n(n-1)(4n+1)}{6}$	n^2
D_n $(n\geq 4)$	$\frac{n(n-1)(2n-1)}{3}$		
E_6	156		
E_7	399		
E_8	1240		
F_4	110	46	64
G_2	16	6	10

□

Now we can compute the numbers $g(\Phi)$ for Φ of types A_n, $n \geq 1$, D_m, $m \geq 4$, and E_i, $i = 6, 7, 8$ as follows.

THEOREM 3.3.

Type of Φ	$g(\Phi)$
A_n $(n \geq 1)$	$\binom{n+1}{3}$
D_n $(n \geq 4)$	$4\binom{n}{3}$
E_6	120
E_7	336
E_8	1120

Proof. By Corollary 2.3 and Lemma 3.1, we have

$$g(\Phi) = \frac{1}{2}\sum_{\alpha\in\Phi^+} d(\alpha) = \frac{1}{2}\sum_{\alpha\in\Phi^+}(h-1-\mathrm{ht}(\alpha))$$
$$= \frac{1}{2}((h-1)|\Phi^+| - H(\Phi)).$$

Thus we have $g(\Phi(A_n)) = \frac{1}{2}(n\binom{n+1}{2} - \binom{n+2}{3}) = \binom{n+1}{3}$ for $n \geq 1$. For $n \geq 4$, we have

$$g(\Phi(D_n)) = \frac{1}{2}\left((2n-3)n(n-1) - \frac{n(n-1)(2n-1)}{3}\right) = 4\binom{n}{3}.$$

Also, we have $g(\Phi(E_6)) = \frac{1}{2}(11\cdot 36 - 156) = 120$,

$$g(\Phi(E_7)) = \frac{1}{2}(17\cdot 63 - 399) = 336,$$

and $g(\Phi(E_8)) = \frac{1}{2}(29\cdot 120 - 1240) = 1120$. □

Now assume that Φ contains roots of two different lengths and that the type of Φ is not G_2. Then the possible types for a subsystem Φ' of Φ of rank 2 and of type $\neq A_1\times A_1$ are A_2 and B_2. Let $u(\Phi)$ be the cardinality of the set

$$\{\{\alpha, \beta\} \mid \alpha, \beta \in \Phi^+ \text{ have different lengths with } \alpha+\beta\in\Phi^+\}.$$

Then it is easily seen that the following formula for $g(\Phi)$ holds.

$$g(\Phi) = \frac{1}{2}\sum_{\alpha\in\Phi^+} d(\alpha) - u(\Phi). \tag{4}$$

First let us consider the case when Φ has type C_n, $n \geq 2$. We see that a subsystem Φ' of Φ has type A_2 only if all the roots in Φ' are short. This implies that for each long root $\beta \in \Phi^+$, the set $D^+(\beta)$ contains no long root and hence $u(\Phi) = \sum_{\beta \in \Phi^+ \text{ long}} d(\beta)$. So by (4), we get

$$g(\Phi) = \frac{1}{2} \sum_{\alpha \in \Phi^+} d(\alpha) - \sum_{\substack{\beta \in \Phi^+ \\ \text{long}}} d(\beta) = \frac{1}{2} \left(\sum_{\substack{\alpha \in \Phi^+ \\ \text{short}}} d(\alpha) - \sum_{\substack{\beta \in \Phi^+ \\ \text{long}}} d(\beta) \right)$$

$$= \frac{1}{2} \left(\sum_{\substack{\alpha \in \Phi^+ \\ \text{short}}} (h - 1 - \mathrm{ht}(\alpha)) - \sum_{i=1}^{n} (i-1) \right)$$

by Theorem 2.2, Corollary 2.3 and Lemma 1.2. Then by Lemma 3.2, we have

$$g(\Phi) = \frac{1}{2} \left((2n-1)n(n-1) - \frac{n(n-1)(4n+1)}{6} - \frac{n(n-1)}{2} \right)$$
$$= \frac{n(n-1)(4n-5)}{6}.$$

Since the root system of type B_n is the dual of the one of type C_n, there exists a bijection from the set of subsystems of the root system of type C_n to that of type B_n by sending Φ' to $\Phi'^{\vee}$. Such a bijective map preserves the ranks of subsystems and also preserves the types of them whenever their ranks are not greater than 2. This implies that we also have $g(\Phi) = \frac{n(n-1)(4n-5)}{6}$ when Φ has type B_n.

Next assume that Φ has type F_4. By Theorem 2.2, Lemma 3.2 and Lemmas 1.1, 1.2, we get

$$\frac{1}{2} \sum_{\alpha \in \Phi^+} d(\alpha) = \frac{1}{2} \left(\sum_{\substack{\alpha \in \Phi^+ \\ \text{short}}} (\mathrm{ht}(\alpha^l) - \mathrm{ht}(\alpha)) + \sum_{\substack{\beta \in \Phi^+ \\ \text{long}}} (\mathrm{ht}(\alpha^s) - \mathrm{ht}(\beta^{\vee})) \right)$$

$$= \frac{1}{2} \left(\frac{1}{2} |\Phi^+| (\mathrm{ht}(\alpha^l) + \mathrm{ht}(\alpha^s)) - 2H^s(\Phi) \right)$$

$$= \frac{1}{2} \left(\frac{1}{2} \cdot 24 \cdot (11 + 8) - 92 \right)$$

$$= 68.$$

Also, by a direct computation, we get $u(\Phi) = 18$. So by (4), we have

$$g(\Phi) = 68 - 18 = 50.$$

Finally, it is easily seen that $g(\Phi) = 3$ when Φ has type G_2. Summing up, we get the following table.

THEOREM 3.4.

Type of Φ	$g(\Phi)$
B_n or C_n $(n \geq 2)$	$\frac{n(n-1)(4n-5)}{6}$
F_4	50
G_2	3

□

From the above discussion, we can deduce even more precise conclusion. We note that in any irreducible root system Φ, there exist at most two different types of subsystems which have rank 2 and types $\neq A_1 \times A_1$. Let $g'(\Phi)$ be the number of subsystems of Φ of type A_2 and let $g''(\Phi)$ be the number of subsystems of Φ of type B_2 or G_2. Then by Theorem 3.3, we have

$$g'(\Phi(B_n)) = g'(\Phi(C_n)) = g(\Phi(D_n)) = 4\binom{n}{3} \quad \text{for } n \geq 4$$

by noting that all the long (resp. short) roots of $\Phi(B_n)$ (resp. $\Phi(C_n)$) form a root system of type D_n. Hence we also have

$$\begin{aligned} g''(\Phi(B_n)) = g''(\Phi(C_n)) &= g(\Phi(B_n)) - g'(\Phi(B_n)) \\ &= \frac{n(n-1)(4n-5)}{6} - 4\binom{n}{3} \\ &= \binom{n}{2}. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} &g''(\Phi(F_4)) = u(\Phi(F_4)) = 18 \quad \text{and} \\ &g'(\Phi(F_4)) = g(\Phi(F_4)) - g''(\Phi(F_4)) = 50 - 18 = 32. \end{aligned}$$

Finally, it is obvious that $g'(\Phi(G_2)) = 2$ and $g''(\Phi(G_2)) = 1$. Summing up, we have the following table.

THEOREM 3.5.

Type of Φ	$g'(\Phi)$	$g''(\Phi)$
B_n, C_n $(n \geq 2)$	$4\binom{n}{3}$	$\binom{n}{2}$
F_4	32	18
G_2	2	1

□

Proof. By the above discussion, it remains to show the result for Φ being of types B_m or C_m, $m = 2, 3$. But this could be checked directly.

REFERENCES

[1] N. Bourbaki, *Groupes et Algèbres de Lie*, Ch. 4–6, Hermann, Paris, 1968.
[2] Jian-yi Shi, *Alcoves corresponding to an affine Weyl group*, J. London Math. Soc., **35** (1987), 42–55.
[3] ——, *Sign types corresponding to an affine Weyl group*, J. London Math. Soc., **35** (1987), 56–74.

Received November 15, 1991. Supported by the National Science Foundation of China and by the Science Foundation of the University Doctoral Program of CNEC.

EAST CHINA NORMAL UNIVERSITY
3663 ZHONGSHAN ROAD (NORTHERN)
SHANGHAI 200062, CHINA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 160, No. 1, 1993

SINGULAR HOMOLOGY AND COHOMOLOGY WITH LOCAL COEFFICIENTS AND DUALITY FOR MANIFOLDS

E. SPANIER

This article contains an application of the author's previous work on cohomology theories on a space to an exposition of singular theory. After a summary of the relevant concepts concerning cohomology theories in general, singular homology and singular cohomology with local coefficients are defined. Each of these is presented in two versions, one with compact supports and one with arbitrary closed supports. It is shown that each version satisfies an appropriate duality theorem for arbitrary (i.e. nonorientable) topological manifolds.

1. Introduction. This paper is a presentation of singular homology and cohomology theory with local coefficients. Included is a treatment of the usual singular homology with compact supports (which is based on finite chains) and the singular homology based on locally finite chains. The former is a weakly additive theory and the latter is an additive theory.

Similarly, there are two types of singular cohomology, one with compact supports and one with arbitrary supports. In an n-manifold X the basic duality theorem asserts the isomorphism of the two types of q-dimensional homology for an open pair (U, V) in X to the corresponding two types of $(n-q)$-dimensional cohomology of the complementary closed pair $(X-V, X-U)$ with coefficient systems suitably related.

Our approach is to study homology and cohomology on a fixed space X and to prove the duality theorem referred to above by comparing two cohomology theories on X, one being the appropriate homology of the open pair in complementary dimension and the other being the corresponding cohomology theory of the complementary closed pair. For this we present the relevant concepts concerning such theories and a review of the comparison theorem for them.

Thus, the paper is divided into two parts, §§2 through 5 are devoted to general concepts concerning covariant and contravariant functors defined on pairs in a space, and §§6 through 10 are devoted to applications of these ideas to singular homology and cohomology and to a proof of the duality theorem for manifolds.

Many of the results in the paper seem to be known in some form but not readily available in the literature. What is new is the generality of our treatment and the methods used. The presentation is reasonably self contained except for two results. The first is the main comparison theorem (stated below in Theorem 2.1) whose proof can be found in various forms in any of the references [**9**, **12**, **13**]. The second is the existence of a Thom class for arbitrary manifolds (needed in §10 for the duality theorems) a proof of which can be found in [**7**].

Section 2 contains definitions of a cohomology (homology) functor on X as a contravariant (covariant) functor from closed (open) pairs in X to the category of graded modules, together with a suitable natural transformation, such that continuity, excision, and exactness are satisfied. The main comparison theorem for cohomology functors is stated as well as a dual for homology functors.

In §3 we review some notation and terminology for chain (cochain) complexes, and in §4 we consider chain (cochain) functors on a space. These are often used in constructions to obtain homology (cohomology) functors. In §5 we introduce chain (cochain) prefunctors. These may be obtained by applying the hom functor to cochain (chain) functors. By taking suitable direct limits of a prefunctor one obtains a corresponding functor. Since the hom functor converts direct sums of modules to direct products, it takes weakly additive functors into additive prefunctors.

In §6 we define the singular chain complex with a local system as coefficients. This is a weakly additive theory. In §7 we consider locally finite singular chains and obtain an additive singular chain functor with local coefficients. We prove that in a locally compact finite polyhedron the corresponding homology is isomorphic to the cellular homology of the polyhedron based on infinite chains. In §8 we introduce singular cohomology with local coefficients. By using the hom functor on singular cochains with compact support we construct another additive homology functor. This is compared with the one based on locally finite singular chains, and the two are shown to be isomorphic on manifolds.

Sections 9 and 10 are devoted to a proof of the duality theorem for manifolds. The algebraic machinery necessary to compare singular homology of an open pair in X with cohomology of the complementary closed pair is set up in §9. This uses a suitable cohomology class U in $(X \times X, X \times X - \delta(X))$ where $\delta(X)$ is the diagonal of $X \times X$. In §10 we consider the case where X is an n manifold and U is its

Thom class and deduce the duality theorem. Some variants of duality are also discussed.

2. Cohomology and homology functors. We assume throughout that the space X is paracompact and Hausdorff. This isn't absolutely essential for some of the results but it simplifies the presentation and suffices for our applications. R will denote a fixed principal ideal domain. All modules will be over R.

A *cohomology functor* H^*, δ^* on X consists of:

(a) a contravariant functor H^* from the category of closed pairs (A, B) in X and inclusion maps between them to the category of graded R modules $(H^*(A, B) = \{H^q(A, B)\}_{q \in \mathbf{Z}})$ and homomorphisms of degree 0 between them, and

(b) for every closed triple (A, B, C) in X a natural transformation

$$\delta^* \colon H^*(B, C) \to H^*(A, B) \quad \text{of degree 1},$$

such that the following three properties are valid:

Continuity. For every closed pair (A, B) in X there is an isomorphism

$$\rho \colon \varinjlim\{H^*(M, N) \mid (M, N) \text{ a closed neighborhood of } (A, B)\} \approx H^*(A, B)$$

where $\rho\{u\} = u|(A, B)$ for $u \in H^*(M, N)$.

Excision. For closed sets A, B in X there is an isomorphism

$$\rho \colon H^*(A \cup B, B) \approx H^*(A, A \cap B)$$

where $\rho(u) = u|(A, A \cap B)$ for $u \in H^*(A \cup B, B)$.

Exactness. For every closed triple (A, B, C) in X the following sequence is exact

$$\cdots \to H^q(A, B) \xrightarrow{\rho} H^q(A, C) \xrightarrow{\rho'} H^q(B, C) \xrightarrow{\delta^*} H^{q+1}(A, B) \to \cdots.$$

What is here called continuity was called tautness in [**8**, **9**]. This definition of a cohomology functor is equivalent to that of an ES Theory [**10**, **11**, **12**] although it is formally different. The definition given here is more convenient for dualization. It is a consequence that every cohomology functor defines a cohomology theory

H', δ' on X (as defined in **[9, 12]**) in which $H'(A) = H^*(A, \varnothing)$ and $\delta' \colon H'(A \cap B) \to H'(A \cup B)$ is suitably defined. The cohomology functor is *nonnegative* if $H^q(A, B) = 0$ for $q < 0$ and all closed pairs (A, B).

A family $\{S_j\}_{j \in J}$ of subsets of X is *discrete* if each point of X has a neighborhood meeting at most one element of the family. Given a discrete family $\{(A_j, B_j)\}_{j \in J}$ of closed pairs in X, for each $j \in J$ let $C_j = \bigcup_{i \neq j} A_i$, a closed set in X. For each $j \in J$ there are restriction homomorphisms

$$H^*(A_j \cup C_j, B_j \cup C_j) \xrightarrow{\rho'_j} H^*\left(\bigcup_{j \in J} (A_j, B_j)\right) \xrightarrow{\rho_j} H^*(A_j, B_j).$$

The cohomology functor is said to be *weakly additive* if the homomorphisms $\{\rho'_j\}$ define an isomorphism

$$\imath \colon \bigoplus_{j \in J} H^*(A_j \cup C_j, B_j \cup C_j) \approx H^*\left(\bigcup_{j \in J} (A_j, B_j)\right)$$

for every discrete family $\{(A_j, B_j)\}_{j \in J}$, and it is *additive* if the homomorphisms $\{\rho_j\}$ define an isomorphism

$$\sigma \colon H^*\left(\bigcup_{j \in J} (A_j, B_j)\right) \approx \prod_{j \in J} H^*(A_j, B_j)$$

for every discrete family $\{(A_j, B_j)\}_{j \in J}$. (Note that, by excision, $H^*(A_j \cup C_j, B_j \cup C_j) = H^*(A_j \cup (B_j \cup C_j), B_j \cup C_j) \approx H^*(A_j, B_j)$ for each $j \in J$, so weak additivity as defined above is equivalent to that defined in **[13]**.)

A *homomorphism* $\varphi \colon H_1^*, \delta_1^* \to H_2^*, \delta_2^*$ between two cohomology functors on the same space X is a natural transformation from H_1^* to H_2^* (of degree 0) which commutes up to sign with δ_1^*, δ_2^*. The following is a consequence of the main comparison theorem for cohomology theories **[9, 12, 13]** and the five lemma.

THEOREM 2.1. *Let $\varphi \colon H_1^*, \delta_1^* \to H_2^*, \delta_2^*$ be a homomorphism between cohomology functors on X and suppose both are weakly additive or both are additive. Suppose there is n such that $\varphi \colon H_1^*(x, \varnothing) \to H_2^*(x, \varnothing)$ is an n-equivalence for all $x \in X$. If both H_1^*, δ_1^* and H_2^*, δ_2^* are nonnegative or if X is locally finite dimensional, then*

$\varphi\colon H_1^*(A, B) \to H_2^*(A, B)$ is an n-equivalence for all closed (A, B) in X.

Dually, a *homology functor* H_*, ∂_* on X consists of:

(a′) a convariant functor H_* from the category of open pairs (U, V) in X to the category of graded R-modules ($H_*(U, V) = \{H_q(U, V)\}_{q\in\mathbb{Z}}$), and

(b′) for every open triple (U, V, W) in X a natural transformation $\partial_*\colon H_*(U, V) \to H_*(V, W)$ of degree -1,

such that the following three properties are valid:

Continuity. For every open pair (U, V) there is an isomorphism

$$i\colon \varinjlim\{H_*(U', V') | (U', V') \text{ open and } (\overline{U}', \overline{V}') \subset (U, V)\} \approx H_*(U, V)$$

where $i\{z\} = i'(z)$ for $z \in H_*(U', V')$ (and $i' : H_*(U', V') \to H_*(U, V)$ is induced by the inclusion map $(U', V') \subset (U, V)$).

Excision. For open U, V in X there is an isomorphism

$$i\colon H_*(U, U \cap V) \approx H_*(U \cup V, V)$$

induced by the inclusion $(U, U \cap V) \subset (U \cup V, V)$.

Exactness. For every open triple (U, V, W) the following sequence is exact

$$\cdots \to H_q(V, W) \xrightarrow{i} H_q(U, W) \xrightarrow{i'} H_q(U, V) \xrightarrow{\partial_*} H_{q-1}(V, W) \to \cdots.$$

The homology functor is *nonnegative* if $H_q(U, V) = 0$ for $q < 0$ and all open pairs (U, V). Given a discrete family $\{(U_j, V_j)\}_{j\in J}$ of open pairs in X, for $j \in J$ let $W_j = \bigcup_{i\neq j} U_i$. For each $j \in J$ there are homomorphisms induced by inclusion

$$H_*(U_j, V_j) \xrightarrow{i_j} H_*\left(\bigcup_{j\in J}(U_j, V_j)\right) \xrightarrow{i'_j} H_*(U_j \cup W_j, V_j \cup W_j).$$

The homology functor is *weakly additive* if the homomorphisms $\{i_j\}$ define an isomorphism

$$i\colon \bigoplus_{j\in J} H_*(U_j, V_j) \approx H_*\left(\bigcup_{j\in J}(U_j, V_j)\right)$$

for every discrete family $\{(U_j, V_j)\}_{j\in J}$, and it is *additive* if the homomorphisms $\{i'_j\}$ define an isomorphism

$$\sigma' \colon H_*\left(\bigcup_{j\in J}(U_j, V_j)\right) \approx \prod_{j\in J} H_*(U_j \cup W_j, V_j \cup W_j)$$

for every discrete family $\{(U_j, V_j)\}_{j\in J}$.

Complementation and sign changing interchanges homology functors and cohomology functors. That is, the equation

$$H_q(X - B, X - A) = H^{-q}(A, B)$$

can be used to define a covariant functor H_* on open pairs if H^* is given on closed pairs, or conversely, defines H^* on closed pairs if H_* is given on open pairs. In each case there is a similar way to relate δ^* and ∂_* so that corresponding to a homology functor H_*, ∂_* there is a cohomology functor and conversely. Complementation in this form does not preserve nonnegativity but does preserve weak additivity and additivity and it is involutive.

A *homomorphism* $h\colon H_*, \partial_* \to H'_*, \partial'_*$ between two homology functors on the same space is a natural transformation from H_* to H'_* commuting up to sign with ∂_* and ∂'_*. The following comparison theorem is one consequence of Theorem 2.1 obtained by complementation and sign changing.

THEOREM 2.2. *Let* $h\colon H_*, \partial_* \to H'_*, \partial'_*$ *be a homomorphism between two homology functors on* X *and suppose that both are weakly additive or both are additive. If* $h\colon H_*(X, X - x) \to H'_*(X, X - x)$ *is an isomorphism for all* $x \in X$ *and* X *is locally finite dimensional, then* $h\colon H_*(U, V) \to H'_*(U, V)$ *is an isomorphism for every open pair* (U, V) *in* X.

3. Chain complexes. In the next section we will see that a homology (or cohomology) functor can be obtained from a functor from open (or closed) pairs to the category of chain (or cochain) complexes having properties analogous to the continuity, excision, and exactness properties.

In this section we summarize some definitions and properties of chain complexes over R. By changing the sign of the degree in such a complex we obtain a cochain complex and vice versa. This procedure will be referred to as "the sign changing trick" and implies that results valid for chain (or cochain) complexes have analogues valid for cochain (or chain) complexes.

A chain transformation between two chain complexes is called a *weak chain equivalence* [3] if it induces isomorphisms of the respective homology modules. Every chain equivalence is a weak chain equivalence, and every weak chain equivalence between free chain complexes is a chain equivalence (recall that R was assumed to be a principal ideal domain). In a similar fashion we define the concept of weak cochain equivalence between cochain complexes.

Let $\theta\colon C \to C'$ be a chain transformation and let G be an R-module. Then $\theta \otimes 1\colon C \otimes G \to C' \otimes G$ is a chain transformation, and if θ is a chain equivalence, so is $\theta \otimes 1$. However, if θ is a weak chain equivalence, then $\theta \otimes 1$ need not be a weak chain equivalence. We will replace $C \otimes G$ by another chain complex which is a functor of C and G such that a weak chain equivalence of C will induce a weak chain equivalence on the new chain complex. The main interest in this construction is in the case of chain complexes which are not free.

Let $0 \to P_1 \xrightarrow{\partial} P_0 \xrightarrow{\varepsilon} G \to 0$ be a free presentation of G. Then $P = \{P_0, P_1, \partial\}$ is a free chain complex with

$$H_q(P) \approx \begin{cases} 0 & \text{if } q \neq 0, \\ G & \text{if } q = 0, \end{cases}$$

and there is a weak chain equivalence $\varepsilon\colon P \to (G, 0)$ where $(G, 0)$ is the chain complex with G in degree 0 and trivial chain modules in degrees other than 0. If P' is another free chain complex with $\varepsilon'\colon P' \to (G, 0)$ a weak chain equivalence, there is a chain equivalence $\tau\colon P \to P'$ such that $\varepsilon' \circ \tau = \varepsilon$.

Consider the complex $C \otimes P$. Since P is free it is a consequence of the Künneth formula [8] that for every q there is a split short exact sequence (universal coefficient formula)

$$0 \to H_q(C) \otimes G \to H_q(C \otimes P) \to \operatorname{tor}(H_{q-1}(C), G) \to 0.$$

Furthermore, if $\theta\colon C \to C'$ is a weak chain equivalence, so is $\theta \otimes 1\colon C \otimes P \to C' \otimes P$, and if C is a free chain complex, then $1 \otimes \varepsilon\colon C \otimes P \to C \otimes G$ is a weak chain equivalence. Finally, if P' is another free chain complex with $\varepsilon'\colon P' \to (G, 0)$ a weak chain equivalence, then $C \otimes P$ and $C \otimes P'$ are chain equivalent. Therefore, $H_*(C \otimes P)$ depends canonically on C and G.

If C, C' are chain complexes there is a chain complex $\hom(C, C')$ [3] where $\hom(C, C')_q$ is the module of homomorphisms φ from C to C' of degree q (so $\varphi(C_i) \subset C'_{i+q}$ for all i) and with

$$\partial''\colon \hom(C, C')_q \to \hom(C, C')_{q-1}$$

defined by $\partial''(\varphi) = \partial' \circ \varphi + (-1)^q \varphi \circ \partial$. Then $\partial''\partial'' = 0$ so $\hom(C, C')$ is a chain complex. Note that $\hom(C, (G, 0))_q = \hom(C_{-q}, G)$ so that if C is a nonnegative chain complex, $\hom(C, (G, 0))$ is a nonpositive chain complex. In this case we consider $\hom(C, (G, 0))$ as a nonnegative cochain complex by changing the sign of the degree. Similarly if C^* is a nonnegative cochain complex we change the sign of its degree and obtain a nonnegative chain complex $\hom(C^*, (G, 0))$.

If C, C' are chain complexes and $\theta\colon C \to C'$ is a weak chain equivalence, then

$$\hom(\theta, 1)\colon \hom(C', (G, 0)) \to \hom(C, (G, 0))$$

need not be a weak cochain equivalence. Because of this we consider an injective resolution of G

$$0 \to G \xrightarrow{\eta} Q^0 \xrightarrow{\delta} Q^1 \to 0.$$

Here, $Q = \{Q^0, Q^1, \delta\}$ is an injective cochain complex with

$$H^q(Q) \approx \begin{cases} 0, & q \neq 0, \\ G, & q = 0, \end{cases}$$

and $\eta\colon (G, 0) \to Q$ is a weak cochain equivalence.

Consider the cochain complex $\hom(C, Q)$ ($\hom(C, Q)^q$ consists of pairs (φ_0, φ_1) where $\varphi_0\colon C_q \to Q^0$ and $\varphi_1\colon C_{q-1} \to Q^1$ and $\delta(\varphi_0, \varphi_1) = ((-1)^q \varphi_0 \circ \partial, (-1)^q \varphi_1 \circ \partial + \delta \circ \varphi_0))$. For every q there is a split short exact sequence (universal coefficient formula)

$$0 \to \operatorname{ext}(H_{q-1}(C), G) \to H^q(\hom(C, Q)) \to \hom(H_q(C), G) \to 0.$$

Furthermore, if $\theta\colon C \to C'$ is a weak chain equivalence so is $\hom(\theta, 1)\colon \hom(C', Q) \to \hom(C, Q)$, and if C is a free cochain complex, then

$$\hom(1, \eta)\colon \hom(C, (G, 0)) \to \hom(C, Q)$$

is a weak cochain equivalence. Finally if Q' is another injective cochain complex with a weak cochain equivalence $\eta'\colon (G, 0) \to Q'$, then there is a cochain equivalence $\hom(C, Q) \to \hom(C, Q')$. Therefore, $H^*(\hom(C, Q))$ depends canonically on C and G. Similarly if C^* is a cochain complex then $\hom(C^*, Q)$ is a chain complex (with $\hom(C^*, Q)_q$ consisting of pairs (ψ_0, ψ_1) where $\psi_0\colon C^q \to Q^0$ and $\psi_1\colon C^{q+1} \to Q^1$).

4. Chain and cochain functors. It is frequently the case that a homology (cohomology) functor is obtained from a suitable chain (cochain) functor on the space. This section contains the relevant definitions.

A *chain functor* C_* on a topological space X is a covariant functor from the category of open pairs in X to the category of chain complexes of R modules ($C_*(U, V) = \{C_q(U, V)\}_{q\in\mathbb{Z}}$) such that the following three properties are valid:

Continuity. For every open pair (U, V) there is a weak chain equivalence

$$i\colon \varinjlim\{C_*(U', V') \mid (U', V') \text{ open and } (\overline{U}', \overline{V}') \subset (U, V)\} \\ \to C_*(U, V).$$

Excision. For open sets U, V in X there is a weak chain equivalence

$$i\colon C_*(U, U \cap V) \to C_*(U \cup V, V).$$

Exactness. For every open triple (U, V, W) in X there is a short exact sequence

$$0 \to C_*(V, W) \xrightarrow{i} C_*(U, W) \xrightarrow{i'} C_*(U, V) \to 0.$$

The chain functor is *nonnegative* if $C_q(U, V) = 0$ for $q < 0$ and all open pairs (U, V) in X. It is *weakly additive* if for every discrete family $\{(U_j, V_j)\}_{j\in J}$ of open pairs in X there is a weak chain equivalence

$$i\colon \bigoplus_{j\in J} C_*(U_j, V_j) \to C_*\left(\bigcup_{j\in J}(U_j, V_j)\right)$$

induced by the maps $i_j\colon C_*(U_j, V_j) \to C_*(\bigcup_{j\in J}(U_j, V_j))$. It is *additive* if for every discrete family $\{(U_j, V_j)\}_{j\in J}$ there is a weak chain equivalence

$$\sigma'\colon C_*\left(\bigcup_{j\in J}(U_j, V_j)\right) \to \prod_{j\in J} C_*(U_j \cup W_j, V_j \cup W_j)$$

where $W_j = \bigcup_{i\neq j} U_i$ and σ' is induced by the maps

$$i'_j\colon C_*\left(\bigcup_{j\in J}(U_j, V_j)\right) \to C_*(U_j \cup W_j, U_j \cup V_j).$$

THEOREM 4.1. *If C_* is a chain functor on X and G is an R module, there is a homology functor $H_*(\cdot, \cdot; G)$ on X with $H_q(U, V; G) = H_q(C_*(U, V) \otimes P)$ (where P is a free resolution of G) and ∂_* the connecting homomorphism corresponding to the exact sequence*

$$0 \to C_*(V, W) \otimes P \to C_*(U, W) \otimes P \to C_*(U, V) \otimes P \to 0.$$

If C_* is nonnegative or weakly additive, the same is true of $H_*(\cdot, \cdot; G)$, ∂_*.

Proof. The operation on chain complexes of forming their tensor product with P commutes with direct sums and direct limits, takes weak chain equivalences into weak chain equivalences, and takes short exact sequences into short exact sequences. Therefore, the continuity, excision, and exactness properties of C_* yield the corresponding properties of $H_*(\cdot, \cdot; G)$, ∂_*. Nonnegativity of C_* clearly implies nonnegativity of $H_*(\cdot, \cdot; G)$ and weak additivity of C_* implies weak additivity of $H_*(\cdot, \cdot; G)$. □

Additivity of C_* does not imply additivity of $H_*(\cdot, \cdot; G)$ because tensor product of chain complexes with P does not commute with infinite products.

Thus, weakly additive homology functors can be obtained from weakly additive chain functors. To get additive homology functors, in the next section we shall use the hom$(\cdot, Q)$ construction applied to a weakly additive cochain functor. Therefore, we now introduce the concept of a cochain functor, dual to that of a chain functor.

A *cochain functor* C^* on X is a contravariant functor from closed pairs in X to the category of cochain complexes of R modules ($C^*(A, B) = \{C^q(A, B)\}_{q \in \mathbb{Z}}$) such that the following properties are valid:

Continuity. For every closed pair (A, B) there is a weak cochain equivalence

$$\rho\colon \varinjlim\{C^*(M, N) \mid (M, N) \text{ a closed neighborhood of } (A, B)\} \to C^*(A, B).$$

Excision. For closed sets A, B in X there is a weak cochain equivalence

$$\rho\colon C^*(A \cup B, B) \to C^*(A, A \cup B).$$

Exactness. For every closed triple (A, B, C) in X there is a short exact sequence

$$0 \to C^*(A, B) \xrightarrow{\rho} C^*(A, C) \xrightarrow{\rho'} C^*(B, C) \to 0.$$

The cochain functor is *nonnegative* if $C^q(A, B) = 0$ for $q < 0$ and all closed (A, B) in X. It is *weakly additive* if for every discrete family $\{(A_j, B_j)\}_{j \in J}$ of closed pairs in X the homomorphisms

$$\rho'_j \colon C^*(A_j \cup C_j, B_j \cup C_j) \to C^*\left(\bigcup_{j \in J} (A_j, B_j)\right)$$

induce a weak cochain equivalence

$$\iota \colon \bigoplus_{j \in J} C^*(A_j \cup C_j, B_j \cup C_j) \to C_*\left(\bigcup_{j \in J} (A_j, B_j)\right)$$

(where $C_j = \bigcup_{i \neq j} A_i$ for each j), and it is *additive* if there is a weak cochain equivalence

$$\sigma \colon C^*\left(\bigcup_{j \in J} (A_j, B_j)\right) \to \prod_{j \in J} C^*(A_j, B_j)$$

induced by the maps $\rho_j \colon C^*(\bigcup_{j \in J}(A_j, B_j)) \to C^*(A_j, B_j)$.

Complementations and sign changing

$$(C^q(A, B) = C_{-q}(X - B, X - A))$$

interchanges chain and cochain functors preserving weak additivity and additivity (but not nonnegativity). The following analogue of Theorem 4.1 is obtained by complementation and sign changing.

THEOREM 4.2. *If C^* is a cochain functor on X and G is an R module, there is a cohomology functor $H^*(\cdot, \cdot; G)$, δ^* on X with $H^q(A, B; G) = H^q(C^*(A, B) \otimes P)$ (where P is a free resolution of G) and δ^* is the connecting homomorphism corresponding to the exact sequence*

$$0 \to C^*(A, B) \otimes P \to C^*(A, C) \otimes P \to C^*(B, C) \otimes P \to 0.$$

If C^* is weakly additive the same is true of $H^*(\cdot, \cdot; G)$.

5. Prefunctors. We have viewed homology on X as a covariant functor on open pairs of X and so our chain functors are also covariant functors on open pairs of X. Dually cohomology on X is a contravariant functor from closed pairs of X and the corresponding cochain functors are also contravariant functors on closed pairs. It may happen that we encounter or construct a covariant functor from *closed* pairs of X or a contravariant functor from *open* pairs. If such functors satisfy suitable hypotheses they are called chain (cochain) prefunctors on X. By passing to appropriate direct limits they give rise to chain (cochain) functors.

One way of constructing a chain (cochain) prefunctor is to apply $\hom(\cdot, Q)$ to a cochain (chain) functor (where Q is an injective resolution of a module G). This procedure applied to a weakly additive cochain (chain) functor yields an additive chain (cochain) prefunctor.

A *cochain prefunctor* C^* on X is a contravariant functor from the category of open pairs (U, V) of X to the category of cochain complexes of R modules $(C^*(U, V) = \{C^q(U, V)\}_{q\in\mathbb{Z}})$ such that the following are valid:

Excision. For open sets U, V in X there is a weak cochain equivalence

$$\rho\colon C^*(U \cup V, V) \to C^*(U, U \cap V).$$

Exactness. For every open triple (U, V, W) in X there is a short exact sequence

$$0 \to C^*(U, V) \xrightarrow{\rho} C^*(U, W) \xrightarrow{\rho'} C^*(V, W) \to 0.$$

Nonnegativity, weak additivity, and *additivity* are defined for cochain prefunctors in fashion analogous to their definition for cochain functors. There is no continuity property involved in the definition of cochain prefunctor.

Dually a *chain prefunctor* C_* on X is a covariant functor C_* from the category of closed pairs (A, B) in X to the category of chain complexes of R modules satisfying excision and exactness.

THEOREM 5.1. *Let C_* be a chain functor on X and G an R module. For an open (U, V) in X define*

$$C^*(U, V) = \hom(C_*(U, V), Q)$$

where Q is an injective resolution of G. Then C^* is a cochain prefunctor on X. If C_* is weakly additive, then C^* is additive. Dually, if C^* is a (weakly additive) chain functor define, for a closed pair (A, B) in X,

$$C_*(A, B) = \hom(C^*(A, B), Q).$$

Then $C_*(A, B)$ is an (additive) chain prefunctor on X.

Proof. Because Q is an injective complex, the excision property of C_* implies that of C^* and the exactness property of C_* implies that of C^*. Thus, C^* is a cochain prefunctor.

Assume C_* is weakly additive. We prove C^* is additive. Let $\{(U_j, V_j)\}_{j\in J}$ be a discrete family of open pairs. Then there is a weak chain equivalence

$$i\colon \bigoplus_{j\in J} C_*(U_j, V_j) \to C_*\left(\bigcup_{j\in J}(U_j, V_j)\right).$$

Therefore, there is a weak cochain equivalence

$$\begin{aligned}\hom(i, 1)\colon \hom&\left(C_*\left(\bigcup_{j\in J}(U_j, V_j)\right), Q\right)\\ &\to \hom\left(\bigoplus_{j\in J} C_*(U_j, V_j), Q\right).\end{aligned}$$

Since $\hom(\bigoplus_{j\in J} C_*(U_j, V_j), Q) \approx \prod_{j\in J} \hom(C_*(U_j, V_j), Q)$ we obtain a weak chain equivalence

$$\sigma\colon C^*\left(\bigcup_{j\in J}(U_j, V_j)\right) \to \prod_{j\in J} C^*(U_j, V_j)$$

proving that C^* is additive.

This completes the proof of the statement about the cochain prefunctor. The dual statement about chain prefunctors is proved similarly. □

THEOREM 5.2. *Let C^* be a cochain prefunctor. Define, for (A, B) a closed pair in X,*

$$\overline{C}^*(A, B) = \varinjlim\{C^*(U, V) | (U, V) \text{ an open neighborhood of } (A, B)\}.$$

Then $\overline{C}^*$ is a cochain functor and if C^* is additive, so is $\overline{C}^*$. Similarly if C_* is a chain prefunctor define, for open (U, V),

$$\underline{C}_*(U, V) = \varinjlim\{C_*(A, B) | (A, B) \text{ closed } \subset (U, V)\}.$$

Then $\underline{C}_*$ is a chain functor and if C_* is additive, so is $\underline{C}_*$.

Proof. Since X is normal, if a closed pair (A, B) is contained in an open pair (U, V) there is a closed pair of neighborhoods (M, N)

of (A, B) with $(M, N) \subset (U, V)$. Using this and the definition of $\overline{C}^*$ it follows that $\overline{C}^*$ is continuous.

Since the cohomology of a direct limit of cochain complexes is the direct limit of the cohomology of the cochain complexes, the excision property for $\overline{C}^*$ follows from that for C^*.

Since the direct limit of exact sequences is exact, the exactness property for $\overline{C}^*$ follows from that for C^*.

Therefore, $\overline{C}^*$ is a cochain functor. To prove it is additive if C^* is, note that, since X is paracompact, it is collectionwise normal. Hence, if $\{(A_j, B_j)\}_{j\in J}$ is a discrete family of closed pairs in X there exist discrete families of open pairs $\{(U_j, V_j)\}_{j\in J}$ with $(A_j, B_j) \subset (U_j, V_j)$ for each j, and as these discrete families vary over such neighborhoods their unions $\bigcup_{j\in J}(U_j, V_j)$ form a cofinal family of open neighborhoods of $\bigcup_{j\in J}(A_j, B_j)$. Since C^* is additive, there is a weak cochain equivalence

$$\sigma\colon C^*\left(\bigcup_{j\in J}(U_j, V_j)\right) \to \prod_{j\in J} C^*(U_j, V_j).$$

Taking the direct limit of both sides as (U_j, V_j) vary over discrete neighborhoods of (A_j, B_j) yields the additivity of $\overline{C}^*$. This proves the result for $\overline{C}^*$.

The result for $\underline{C}_*$ follows similarly. □

Combining Theorems 5.1 and 5.2 we obtain the following.

COROLLARY 5.3. *Let* C_* *be a chain functor on* X *and* G *an* R *module. For a closed pair* (A, B) *in* X *define*

$$\overline{C}^*(A, B; G) = \varinjlim\{\hom(C_*(U, V), Q) \mid (U, V) \text{ an open neighborhood of } (A, B)\}$$

where Q *is an injective resolution of* G. *Then* $\overline{C}^*(\cdot, \cdot; G)$ *is a cochain functor on* X. *If* C_* *is weakly additive, then* $\overline{C}^*(\cdot, \cdot; G)$ *is additive. Dually, if* C^* *is a* (*weakly additive*) *cochain functor on* X *define, for* (U, V) *open in* X,

$$\underline{C}_*(U, V; G) = \varinjlim\{\hom(C^*(A, B), Q) \mid (A, B) \text{ closed } \subset (U, V)\}.$$

Then $\underline{C}_*$ *is an* (*additive*) *homology functor on* X. □

REMARK 5.4. In Corollary 5.3 if C_* (or C^*) is nonnegative it is not true that $\overline{C}^*$ (or $\underline{C}_*$) is nonnegative; however, the corresponding cohomology functor $\overline{H}{}^*$ (or homology functor $\underline{H}_*$) determined by $\overline{C}{}^*$ ($\underline{C}_*$) is nonnegative because of the universal coefficient formula.

Even for prefunctors which are neither weakly additive nor additive we can obtain weakly additive functors by using limits over compact (or cobounded) sets as we now describe. (Recall that a set $A \subset X$ is *cobounded* if $X - A$ has compact closure.)

THEOREM 5.5. *Let* C^* *be a cochain prefunctor on* X. *For* (A, B) *a closed pair in* X *define*

$$\overline{C}{}^*_c(A, B) = \varinjlim\{C^*(U, V) | (U, V) \textit{ open cobounded neighborhood of } (A, B)\}.$$

Then $\overline{C}{}^*_c$ *is a weakly additive cochain functor. Dually, if* C_* *is a chain prefunctor on* X *define, for open* (U, V) *in* X,

$$\underline{C}{}^c_*(U, V) = \varinjlim\{C_*(A, B) | (A, B) \textit{ compact } \subset (U, V)\}.$$

Then $\underline{C}{}^c_*$ *is a weakly additive chain functor on* X.

Proof. The proof that $\overline{C}{}^*_c$ is a cochain functor (and $\underline{C}{}^c_*$ is a chain functor) is analogous to the proof in Theorem 5.2 that $\overline{C}{}^*$ ($\underline{C}_*$) is a cochain (chain) functor.

We show $\overline{C}{}^*_c$ is weakly additive. Let $\{(A_j, B_j)\}_{j \in J}$ be a discrete family of closed pairs in X and let (U, V) be an open cobounded neighborhood of $\bigcup_{j \in J}(A_j, B_j)$. Then there is a finite set $F \subset J$ such that $j \notin F$ implies $A_j \subset V$ (since $X - V$ is compact it can meet only finitely many A_j's). If $u \in \overline{C}{}^*_c(\bigcup_{j \in J}(A_j, B_j))$ has the form $u = \{v\}$ where $v \in C^*(U, V)$ then $u | \bigcup_{j \notin F}(A_j, B_j) = 0$ so (by the analogue of the Lemma in **[13]**) it follows that $\overline{C}{}^*_c$ is weakly additive. □

6. Singular homology. In this section we introduce the usual singular homology of X with coefficients in a local system. This is a weakly additive homology functor on X. A corresponding additive homology functor will be introduced in the next section. Our treatment of singular theory dates back to Eilenberg **[5]** and is the one most commonly used since the appearance of **[5]**.

We begin by recalling some properties of local systems. Local systems were defined by Steenrod **[14]**. The definitions below are equivalent to his. A *local system* **[8]** Γ of R modules on a space X is a function associating to every $x \in X$ an R module Γ_x and to every path $\omega: I \to X$ a homomorphism $\Gamma_\omega: \Gamma_{\omega(0)} \to \Gamma_{\omega(1)}$ (this is the reverse of the definition in [**8**]) such that:

(1) If ε_x is the constant path at x, then $\Gamma_{\varepsilon_x} = 1_{\Gamma(x)}$.

(2) If $\omega(1) = \omega'(0)$ the product path $\omega * \omega'$ is defined and $\Gamma_{\omega*\omega'} = \Gamma_{\omega'} \circ \Gamma_\omega \colon \Gamma_{\omega(0)} \to \Gamma_{\omega'(1)}$.

(3) If ω and ω' are homotopic paths (i.e. homotopic relative to $\{0, 1\}$), then $\Gamma_\omega = \Gamma_{\omega'} \colon \Gamma_{\omega(0)} \to \Gamma_{\omega(1)}$.

Let $f \colon Y \to X$ be a continuous map and Γ a local system on X. A Γ *section* of f is a function s assigning to every $y \in Y$ an element $s(y) \in \Gamma_{f(y)}$ such that, for every path ω in Y, $\Gamma_{f\circ\omega}(s(\omega(0))) = s(\omega(1))$. The set of all such Γ sections of f is an R module $\Gamma(f)$ under pointwise operations of functions. In case Y is path connected and every closed path in Y is mapped by f into a null homotopic path in X (e.g. if Y is simply connected), then for every $y_0 \in Y$ the map $\varphi_0 \colon \Gamma(f) \to \Gamma_{f(y_0)}$ defined by $\varphi_0(s) = s(y_0)$ is an isomorphism.

If Γ is a local system and G is an R module, $\hom(\Gamma, G)$ is the local system $(\hom(\Gamma, G))_x = \hom(\Gamma_x, G)$ and, for ω a path in X, $(\hom(\Gamma, G))_\omega \colon (\hom(\Gamma, G))_{\omega(0)} \to (\hom(\Gamma, G))_{\omega(1)}$ is equal to $(\hom(\Gamma_\omega, 1))^{-1} \colon \hom(\Gamma_{\omega(0)}, G) \to \hom(\Gamma_{\omega(1)}, G)$ (so for $\varphi \in \hom(\Gamma, G)_{\omega(0)}$, $(\hom(\Gamma, G))_\omega(\varphi) = \varphi \circ \Gamma_\omega^{-1} \in \hom(\Gamma, G)_{\omega(1)})$.

Two local systems Γ and Γ' are *paired* to an R module G if there is bilinear map $\langle \cdot, \cdot \rangle \colon \Gamma_x \otimes \Gamma'_x \to G$ for each x such that if ω is a path in X then for $\gamma \in \Gamma_{\omega(0)}$, $\gamma \in \Gamma'_{\omega(0)}$ we have $\langle \gamma, \gamma' \rangle = \langle \Gamma_\omega(\gamma), \Gamma'_\omega(\gamma') \rangle$.

EXAMPLE 6.1. There is a pairing of $\hom(\Gamma, G)$ and Γ to G defined by $\langle \varphi, \gamma \rangle = \varphi(\gamma)$ for $\gamma \in \Gamma_x$ and $\varphi \in (\hom(\Gamma, G))_x = \hom(\Gamma_x, G)$.

If Γ and Γ are paired to G then for every path connected Y and map $f \colon Y \to X$ there is a pairing of $\Gamma(f)$ and $\Gamma'(f)$ to G defined by $\langle s, s' \rangle = \langle s(y), s'(y) \rangle$ for $y \in Y$ (the value of $\langle s(y), s'(y) \rangle$ is independent of $y \in Y$ because Y is path connected).

For $q < 0$ define $\Delta_q(X; \Gamma) = 0$ and for $q \geq 0$ define $\Delta_q(X; \Gamma)$ to be set of all finitely non-zero functions c which assign to every singular q-simplex $\sigma \colon \Delta^q \to X$ an element $c(\sigma) \in \Gamma(\sigma)$. Then $\Delta_q(X; \Gamma)$ is an R module under pointwise operations of functions. If $g_\sigma \in \Gamma(\sigma)$ let $g_\sigma \sigma$ denote the element of $\Delta_q(X; \Gamma)$ such that

$$g_\sigma(\sigma') = \begin{cases} 0 & \text{if } \sigma \neq \sigma', \\ g_\sigma & \text{if } \sigma = \sigma'. \end{cases}$$

Then every element $c \in \Delta_q(X; \Gamma)$ has the form $c = \sum_\sigma g_\sigma \sigma$ where $g_\sigma = 0$ except for a finite set of σ's.

If $g_\sigma \in \Gamma(\sigma)$ and $\sigma^{(i)}$ is the ith face of σ, then $g_\sigma | \sigma^{(i)} \in \Gamma(\sigma^{(i)})$.

Thus, there is a homomorphism

$$\partial : \Delta_q(X;\Gamma) \to \Delta_{q-1}(X;\Gamma) \quad \text{for } q > 0$$

defined by $\partial(\sum_\sigma g_\sigma \sigma) = \sum_\sigma \sum_{0 \le i \le q} (-1)^i (g_\sigma | \sigma^{(i)}) \sigma^{(i)}$. It is easily verified that $\partial\partial = 0$ so that $\Delta_*(X;\Gamma) = \{\Delta_q(X;\Gamma), \partial\}$ is a nonnegative chain complex. If $c = \sum_\sigma g_\sigma \sigma$ its *support* $|c| = \bigcup\{\sigma(\Delta^q) | g_\sigma \neq 0\}$, a compact subset of X. If $A \subset X$, then $\Delta_*(A;\Gamma) = \{c \in \Delta_*(X;\Gamma) | |c| \subset A\}$ is a subcomplex of $\Delta_*(X;\Gamma)$, and we define

$$\Delta_*(X, A;\Gamma) = \Delta_*(X;\Gamma)/\Delta_*(A;\Gamma)$$

so there is a short exact sequence

$$0 \to \Delta_*(A;\Gamma) \to \Delta_*(X;\Gamma) \to \Delta_*(X, A;\Gamma) \to 0.$$

THEOREM 6.2. *Let Γ be a fixed local system on X. Then $\Delta_*(\cdot, \cdot)$ defined for (U, V) open by $\Delta_*(U, V) = \Delta_*(U, V;\Gamma)$ is a nonnegative weakly additive chain functor on X.*

Proof. Continuity follows from the fact that as (U', V') vary over open sets with $(\overline{U}', \overline{V}') \subset (U, V)$, then $\bigcup \Delta_*(U';\Gamma) = \Delta_*(U;\Gamma)$ and $\bigcup \Delta_*(V';\Gamma) = \Delta_*(V;\Gamma)$ because every element of $\Delta_*(U;\Gamma)$ (or $\Delta_*(V;\Gamma)$) has support a compact subset of U (or V) so is contained in U' (or V') for some open U' (or V') whose closure is contained in U (or V).

For excision let $\mathscr{V}$ be a collection of subsets of X and define $\Delta_*(\mathscr{V};\Gamma)$ to be the subcomplex of $\Delta_*(X;\Gamma)$ consisting of finite sums $\sum c_j$ such that for each j, $c_j \in \Delta_*(X;\Gamma)$ and there is some $V_j \in \mathscr{V}$ with $|c_j| \subset V_j$. If $\mathscr{V}$ is a collection of subsets of A such that $A = \bigcup_{V \in \mathscr{V}} \operatorname{int} V$, then the inclusion map $\Delta_*(\mathscr{V};\Gamma) \subset \Delta_*(A;\Gamma)$ is a chain equivalence (proof analogous to that of Theorem 14 on p. 178 of **[8]**). In particular if U and V are open sets in X, then $\Delta_*(U;\Gamma) + \Delta_*(V;\Gamma) \subset \Delta_*(U \cup V;\Gamma)$ is a chain equivalence. It follows that

$$\begin{aligned}[\Delta_*(U;\Gamma) + \Delta_*(V;\Gamma)]/\Delta_*(V;\Gamma) &\to \Delta_*(U \cup V;\Gamma)/\Delta_*(V;\Gamma) \\ &= \Delta_*(U \cup V, V;\Gamma)\end{aligned}$$

is a chain equivalence. Excision follows from this and the Noether isomorphism

$$\begin{aligned}\Delta_*(U, U \cap V;\Gamma) &= \Delta_*(U;\Gamma)/\Delta_*(U \cap V;\Gamma) \\ &= \Delta_*(U;\Gamma)/[\Delta_*(U;\Gamma) \cap \Delta_*(V;\Gamma)] \\ &\approx [\Delta_*(U;\Gamma) + \Delta_*(V;\Gamma)]/\Delta_*(V;\Gamma).\end{aligned}$$

If (U, V, W) is an open triple in X there is a short exact sequence

$$0 \to \Delta_*(V, W; \Gamma) \to \Delta_*(U, W; \Gamma) \to \Delta_*(U, V; \Gamma) \to 0$$

so exactness is satisfied.

Thus, $\Delta_*(\cdot, \cdot)$ is a nonnegative (by definition) chain functor. It is weakly additive because every chain has compact supports [**13**]. □

It follows from Theorem 6.2 that for every R module G the singular homology on X with coefficients $\Gamma \otimes G$, denoted by $H_*(\cdot, \cdot; \Gamma \otimes G)$ and defined by $H_q(U, V; \Gamma \otimes G) = H_q(\Delta_*(U, V; \Gamma) \otimes P)$ (where P is a projective resolution of G) is a nonnegative weakly additive homology functor on X.

7. Locally finite singular homology. In this section we introduce the locally finite singular chains to obtain an additive homology functor. In a special situation we relate this new homology to limits of the usual singular homology. This is applied to show that for a locally finite simplicial complex the locally finite singular homology is isomorphic to the simplicial homology based on infinite chains. The locally finite singular homology was considered in Sèminaire Cartan [2] where it was called "singular homology of the second kind".

For $q \geq 0$ a function c which assigns to every singular q-simplex $\sigma\colon \Delta^q \to X$ an element $c(\sigma) \in \Gamma(\sigma)$ is called a *locally finite q-chain* if $\{\sigma(\Delta^q) | c(\sigma) \neq 0\}$ is a locally finite family in X. The locally finite q-chains form an R module $\Delta_q^\infty(X; \Gamma)$ under the usual operations on functions. For $q < 0$ we define $\Delta_q^\infty(X; \Gamma) = 0$. If $c \in \Delta_q^\infty(X; \Gamma)$ with $0 \leq q$ its *support* $|c| = \bigcup\{\sigma(\Delta^q) | c(\sigma) \neq 0\}$, a closed subset of X. If $c \in \Delta_q^\infty(X; \Gamma)$ with $q < 0$ we define $|c| = \varnothing$.

If $\{c_j\}_{j\in J}$ is a family of elements of $\Delta_q^\infty(X; \Gamma)$ it is said to be *locally finite* if the family of supports $\{|c_j|\}_{j\in J}$ is locally finite. In this case the sum $\sum_{j\in J} c_j$ is defined as an element of $\Delta_q^\infty(X; \Gamma)$ (because for every singular q-simplex σ, $\sigma(\Delta^q)$ is compact so meets only finitely many $|c_j|$ and so $c_j(\sigma) = 0$ except for a finite set of j's). In particular, if $q > 0$ and $c = \sum_\sigma g_\sigma \sigma \in \Delta_q^\infty(X; \Gamma)$, then $\{\sum_{0\leq i\leq q} (-1)^i (g_\sigma | \sigma^{(i)}) \sigma^{(i)}\}_{g_\sigma \neq 0}$ is a locally finite family so

$$\sum_\sigma \sum_{0\leq i\leq q} (-1)^i (g_\sigma | \sigma^{(i)}) \sigma^{(i)} \in \Delta_{q-1}^\infty(X; \Gamma)$$

and there is a homomorphism

$$\partial\colon \Delta_q^\infty(X; \Gamma) \to \Delta_{q-1}^\infty(X; \Gamma)$$

such that $\partial(\sum_\sigma g_\sigma \sigma) = \sum_\sigma \sum_{0\leq i \leq \sigma}(-1)^i (g_\sigma | \sigma^{(i)})\sigma^{(i)}$. Then $|\partial c| \subset |c|$ and $\partial\partial c = 0$ so there is a nonnegative chain complex $\Delta_*^\infty(X; \Gamma) = \{\Delta_q^\infty(X; \Gamma), \partial\}$ of *locally finite singular chains with coefficients* Γ. We define $H_q^\infty(X; \Gamma) = H_q(\Delta_*^\infty(X; \Gamma))$.

Clearly $\Delta_*(X; \Gamma) \subset \Delta_*^\infty(X; \Gamma)$ (in fact, $\Delta_*(X; \Gamma)$ is the subcomplex of $\Delta_*^\infty(X; \Gamma)$ of chains having compact support) so there is a homomorphism

$$H_q(X; \Gamma) \to H_q^\infty(X; \Gamma).$$

If $A \subset X$ we define ${}^X\Delta_*^\infty(A; \Gamma) = \{c \in \Delta_*^\infty(X; \Gamma) \mid |c| \subset A\}$. This is a subcomplex of $\Delta_*^\infty(X; \Gamma)$ and consists of chains in A which are locally finite in X (a stronger condition than being locally finite in A). We define ${}^XH_*^\infty(A; \Gamma) = H_*({}^X\Delta_*^\infty(A; \Gamma))$. In case A is closed in X, ${}^X\Delta_*^\infty(A; \Gamma) = \Delta_*^\infty(A; \Gamma)$ so that ${}^XH_*^\infty(A; \Gamma) = H_*^\infty(A; \Gamma)$ in this case.

If $c \in {}^X\Delta_*^\infty(A; \Gamma)$ there is a set F closed in X with $F \subset A$ such that $c \in \Delta_*^\infty(F; \Gamma)$ (for example, $F = |c|$). Therefore, if $\{A_j\}_{j\in J}$ is a family of subsets of A directed upward by inclusion such that every subset of A which is closed in X is contained in A_j for some $j \in J$ then ${}^X\Delta_*^\infty(A; \Gamma) = \bigcup_{j\in J} {}^X\Delta_*^\infty(A_j; \Gamma)$ and so $\varinjlim\{{}^XH_*^\infty(A_j; \Gamma)\} \approx {}^XH_*^\infty(A; \Gamma)$.

If $B \subset A \subset X$ then ${}^X\Delta_*^\infty(B; \Gamma) \subset {}^X\Delta_*^\infty(A; \Gamma)$ and we define ${}^X\Delta_*^\infty(A, B; \Gamma) = {}^X\Delta_*^\infty(A; \Gamma)/{}^X\Delta_*^\infty(B; \Gamma)$. There is then a short exact sequence of chain complexes

$$0 \to {}^X\Delta_*^\infty(B; \Gamma) \to {}^X\Delta_*^\infty(A; \Gamma) \to {}^X\Delta_*^\infty(A, B; \Gamma) \to 0.$$

THEOREM 7.1. *For* (U, V) *an open pair in* X *define* $\Delta_*^\infty(U, V) = {}^X\Delta_*^\infty(U, V; \Gamma)$. *Then* Δ_*^∞ *is a nonnegative additive chain functor on* X.

Proof. Continuity follows from the fact that as (U', V') vary over open pairs with $(\overline{U'}, \overline{V'}) \subset (U, V)$ then $\bigcup \Delta_*^\infty(U') = \Delta_*^\infty(U)$ and $\bigcup \Delta_*^\infty(V') = \Delta_*^\infty(V)$ so that $\varinjlim\{\Delta_*^\infty(U', V')\} \approx \Delta_*^\infty(U, V)$.

For excision let $\mathscr{V}$ be a family of subsets of X and define ${}^X\Delta_*^\infty(\mathscr{V}; \Gamma)$ to be the subcomplex of $\Delta_*^\infty(X; \Gamma)$ consisting of locally finite sums $\sum_{j\in J} c_j$ such that for each $j \in J$, $c_j \in \Delta_*^\infty(X; \Gamma)$ and there is some $V_j \in \mathscr{V}$ with $|c_j| \subset V_j$. If $\mathscr{V}$ is a collection of subsets of A such that $A = \bigcup_{V\in\mathscr{V}} \operatorname{int} V$ then the inclusion map ${}^X\Delta_*^\infty(\mathscr{V}; \Gamma) \subset {}^X\Delta_*^\infty(A; \Gamma)$ is a chain equivalence (proof analogous to that of Theorem 14 on p. 178 of [**8**]). Excision follows from this (as in the proof of excision in Theorem 6.1).

If (U, V, W) is an open triple in X there is a short exact sequence

$$0 \to \Delta_*^\infty(V, W) \to \Delta_*^\infty(U, W) \to \Delta_*^\infty(U, V) \to 0$$

so exactness is satisfied.

Thus, Δ_*^∞ is a nonnegative chain functor on X. To show it is additive suppose $\{U_j\}_{j\in J}$ is a discrete family of open sets in X and $V \subset \bigcup_{j\in J} U_j$ is open. Then there are isomorphisms

$$\begin{aligned} {}^X\Delta_*^\infty\left(\bigcup_{j\in J} U_j ; \Gamma\right) &\approx \prod_{j\in J} {}^X\Delta_*^\infty(U_j ; \Gamma), \\ {}^X\Delta_*^\infty(V ; \Gamma) &\approx \prod_{j\in J} {}^X\Delta_*^\infty(V \cap U_j ; \Gamma) \end{aligned}$$

and additivity of Δ_*^∞ follows. □

Thus, usual singular homology is obtained from a weakly additive chain functor and locally finite singular homology is obtained from an additive chain functor. In case X is compact the two theories agree. The following is a generalization of this.

Lemma 7.2. *If $B \subset A$ and $A - B$ has compact closure in X then* $\Delta_*(A, B ; \Gamma) \approx {}^X\Delta_*^\infty(A, B ; \Gamma)$.

Proof. If $A - B$ has compact closure in X, then for $c \in {}^X\Delta_q^\infty(A, B ; \Gamma)$ the set $\{\sigma | c(\sigma) \neq 0 \text{ and } \sigma(\Delta^q) \cap (A - B) \neq \varnothing\}$ is finite. Therefore, ${}^X\Delta_q^\infty(A ; \Gamma) = {}^X\Delta_q^\infty(B ; \Gamma) + \Delta_q(A ; \Gamma)$ and so

$$\begin{aligned} {}^X\Delta_q^\infty(A, B ; \Gamma) &= {}^X\Delta_q^\infty(A ; \Gamma)/{}^X\Delta_q^\infty(B ; \Gamma) \\ &= [{}^X\Delta_q^\infty(B ; \Gamma) + \Delta_q(A ; \Gamma)]/{}^X\Delta_q^\infty(B ; \Gamma) \\ &\approx \Delta_q(A ; \Gamma)/[{}^X\Delta_q^\infty(B ; \Gamma) \cap \Delta_q(A ; \Gamma)] \\ &= \Delta_q(A ; \Gamma)/\Delta_q(B ; \Gamma) = \Delta_q(A, B ; \Gamma). \end{aligned}$$ □

Our next result relates H_*^∞ to limits of H_* in a special situation.

Theorem 7.3. *Assume there is a sequence $\{C_i\}_{i\geq 0}$ of subsets of X such that $X = \bigcup_i \operatorname{int} C_i$, $C_i \subset C_{i+1}$ for each i, and $\overline{C}_i$ is compact for each i. Then there is a short exact sequence*

$$\begin{aligned} 0 &\to \lim{}^1\{H_{q+1}(X, X - C_i ; \Gamma)\} \\ &\to H_q^\infty(X ; \Gamma) \to \varprojlim\{H_q(X, X - C_i ; \Gamma)\} \to 0. \end{aligned}$$

Proof. (Note that a sequence $\{C_i\}$ of subsets of X satisfying the hypotheses of the theorem exists if and only if X is a σ-compact space as defined in Dugundji [**4**].) For each i there is a quotient chain map

$$\tau_i\colon \Delta_*^\infty(X;\Gamma) \to {}^X\Delta_*^\infty(X, X - C_i;\Gamma)$$

and these define a chain map

$$\tau\colon \Delta_*^\infty(X;\Gamma) \to \varprojlim \{{}^X\Delta_*^\infty(X, X - C_i;\Gamma)\}.$$

We prove τ is an isomorphism. If $c \in \Delta_*^\infty(X;\Gamma)$ is in $\ker\tau$ then $\tau_i(c) = 0$ for each i so $c \in {}^X\Delta_*^\infty(X - C_i;\Gamma)$ for each i. Since $\bigcup_i C_i = X$, it follows that $\bigcap_i {}^X\Delta_*^\infty(X - C_i;\Gamma) = 0$ so $c = 0$ and τ is a monomorphism.

To show τ is an epimorphism let $\{c_i \in {}^X\Delta_*^\infty(X, X - C_i;\Gamma)\}$ be an element in $\varprojlim\{{}^X\Delta_*^\infty(X, X - C_i;\Gamma)\}$. The element c_i can be regarded as a locally finite sum of singular simplexes each having support which meets C_i. Similarly c_{i+1} is a locally finite sum of singular simplexes each having support which meets C_{i+1}. The condition that c_{i+1} maps to c_i implies that on singular simplexes whose support meets C_i both c_i and c_{i+1} have the same value. Therefore, there is a chain $C = \sum g_\sigma \sigma$ in X such that if $|\sigma| \cap C_i \neq \varnothing$ then $g_\sigma =$ the value of c_i on σ. We show c is a locally finite chain in X. Since $X = \bigcup \operatorname{int} C_i$, if $x \in X$ there is i such that $x \in \operatorname{int} C_i$. If $|\sigma| \cap \operatorname{int} C_i \neq \varnothing$ then $|\sigma| \cap C_i \neq \varnothing$ so g_σ is the value of c_i on σ. Since c_i is a locally finite chain, there is a neighborhood N of x such that there are only a finite number of σ's with $|\sigma| \cap N \neq \varnothing$ and $c_i(\sigma) \neq 0$. Then $N \cap \operatorname{int} C_i$ is a neighborhood of x such that there are only finitely many σ's with $|\sigma| \cap (N \cap \operatorname{int} C_i) \neq \varnothing$ and $g_\sigma \neq 0$. Therefore, $c \in \Delta_*^\infty(X;\Gamma)$ and clearly $\tau(c) = \{\tau_i(c)\} = \{c_i\}$ so τ is an epimorphism.

By Lemma 7.2 since $\overline{C_i}$ is compact, there is an isomorphism $\Delta_*(X, X - C_i;\Gamma) \approx {}^X\Delta_*^\infty(X, X - C_i;\Gamma)$ so that $\Delta_*^\infty(X;\Gamma)$ is isomorphic to $\varprojlim\{\Delta_*(X, X - C_i;\Gamma)\}$. Each of the chain maps $\Delta_*(X, X - C_{i+1};\Gamma) \to \Delta_*(X, X - C_i;\Gamma)$ is an epimorphism so by A.15 on p. 402 of [**6**], $\lim^1\{\Delta_*(X, X - C_i;\Gamma)\} = 0$, and then by A.19 on p. 407 of [**6**] there is a short exact sequence

$$\begin{aligned} 0 \to {} & \lim{}^1\{H_{q+1}(X, X - C_i;\Gamma)\} \\ & \to H_q^\infty(X;\Gamma) \to \varprojlim\{H_q(X, X - C_i;\Gamma)\} \to 0. \qquad \square \end{aligned}$$

We use this last result to show that for the space of a locally finite simplicial complex K, then $H_q^\infty(|K|;\Gamma)$ is isomorphic to the homology group of the complex of infinite simplicial chains with coefficients Γ.

Let K be a locally finite simplicial complex and let $C_*^\infty(K; \Gamma)$ be the infinite chain complex of oriented simplexes with coefficients Γ (thus an element $c \in C_q^\infty(K; \Gamma)$ is a function assigning to every oriented q-simplex σ of K an element $c(\sigma) \in \Gamma(|\sigma|)$ such that if σ' is the oppositely oriented simplex then $c(\sigma) + c(\sigma') = 0$). Similarly let $\Delta_*^\infty(K; \Gamma)$ be the infinite chain complex of ordered simplexes with coefficients Γ (so $c \in \Delta_q^\infty(K; \Gamma)$ is a function assigning to every ordered q-simplex σ of K an element $c(\sigma) \in \Gamma(|\sigma|)$). There are natural chain maps [**8**] $\mu^\infty \colon \Delta_*^\infty(K; \Gamma) \to C_*^\infty(K; \Gamma)$ and $\nu^\infty \colon \Delta_*^\infty(K; \Gamma) \to \Delta_*^\infty(|K|; \Gamma)$ which are chain equivalences for every finite complex K.

THEOREM 7.4. *If K is a locally finite simplicial complex there are isomorphisms*

$$\mu_*^\infty \colon H_q(\Delta_*^\infty(K; \Gamma)) \approx H_q(C_*^\infty(K; \Gamma))$$

and

$$\nu_*^\infty \colon H_q(\Delta_*^\infty(K; \Gamma)) \approx H_q^\infty(|K|; \Gamma).$$

Proof. The local finiteness of K implies that $K = \bigcup_{i=0}^\infty K_i$ where K_i is finite for each i and $|K_i| \subset \operatorname{int}|K_{i+1}|$. By analogues of Theorem 7.3 for $C_*^\infty(K; \Gamma)$ and $\Delta_*^\infty(K; \Gamma)$ there is a commutative diagram with exact rows (in which $K - \operatorname{int}|K_i|$ is denoted by L_i and all coefficients are in Γ)

$$\begin{tikzcd}
0 \arrow[r] & \lim^1\{H_{q+1}(C_*(K, L_i))\} \arrow[r] & H_q(C_*^\infty(K)) \arrow[r] & \varprojlim\{H_q(C_*(K, L_\iota))\} \arrow[r] & 0 \\
0 \arrow[r] & \lim^1\{H_{q+1}(\Delta_*(K, L_i))\} \arrow[r] \arrow[u, "\mu_*"'] \arrow[d, "\nu_*"'] & H_q(\Delta_*^\infty(K)) \arrow[r] \arrow[u, "\mu_*^\infty"'] \arrow[d, "\nu_*^\infty"'] & \varprojlim\{H_q(\Delta_*(K, L_\iota))\} \arrow[r] \arrow[u, "\mu_*"'] \arrow[d, "\nu_*"'] & 0 \\
0 \arrow[r] & \lim^1\{H_{q+1}(|K|, |L_\iota|)\} \arrow[r] & H_q^\infty(|K|) \arrow[r] & \varprojlim\{H_q(|K|, |L_\iota|)\} \arrow[r] & 0.
\end{tikzcd}$$

Since the vertical maps on the sides are known to be isomorphisms, the result follows from the 5-lemma. □

8. Singular cohomology. In this section we define the singular cohomology of X which coefficients in a local system, and, using this, we define another additive homology function X which we compare with the one defined in the last section.

Let Γ be a local system on X and for a pair (M, N) in X let $\Delta^q(M, N; \Gamma) = 0$ if $q < 0$ and for $q \geq 0$ let $\Delta^q(M, N; \Gamma)$ be the module of q-cochains of M which vanish on N (i.e. $u \in \Delta^q(M, N; \Gamma)$ is a function assigning to every singular q-simplex

$\sigma\colon \Delta^q \to M$ an element $u(\sigma) \in \Gamma(\sigma)$ such that if $\sigma(\Delta^q) \subset N$ then $u(\sigma) = 0$). For $u \in \Delta^q(M, N; \Gamma)$, $q \geq 0$ define $\delta u \in \Delta^{q+1}(M, N; \Gamma)$ by

$$(\delta u)(\sigma) = \sum_{0 \leq i \leq q+1} (-1)^i \overline{u(\sigma^{(i)})}$$

where $\sigma\colon \Delta^{q+1} \to M$ and if $g \in \Gamma(\sigma^{(i)})$ then $\bar{g} \in \Gamma(\sigma)$ is the unique element such that $\bar{g}|\sigma^{(i)} = g$. Then $\delta\delta = 0$ so there is a cochain complex $\Delta^*(M, N; \Gamma) = \{\Delta^q(M, N; \Gamma), \delta\}$.

In case Γ, Γ' are local systems on X paired to an R module G as in §6 there is a pairing $\Delta^q(M, N; \Gamma) \otimes \Delta_q(M, N; \Gamma') \to G$ defined by $\langle u, c\rangle = \sum \langle u(\sigma), g_\sigma\rangle$ where $c = \sum_\sigma g_\sigma \sigma$. In case $u \in \Delta^q(M, N; \Gamma)$ and $c = \sum g_\sigma \sigma \in \Delta_{q+1}(M, N; \Gamma')$ we have

$$\langle \delta u, c\rangle = \sum_{g_\sigma} \langle \delta u(\sigma), g_\sigma\rangle = \sum_\sigma \sum_{0 \leq i \leq q+1} (-1)^i \langle \overline{u(\sigma^i)}, g_\sigma\rangle.$$

Now for any $g \in \Gamma(\sigma^{(i)})$ we have $\langle \bar{g}, g_\sigma\rangle = \langle g, g_\sigma|\sigma^{(i)}\rangle$. Therefore,

$$\begin{aligned}\langle \delta u, c\rangle &= \sum_\sigma \sum_{0 \leq i \leq q+1} (-1)^i \langle u(\sigma^{(i)}), g_\sigma|\sigma^{(i)}\rangle \\ &= \sum_\sigma \langle u, \partial(g_\sigma \sigma)\rangle = \langle u, \partial c\rangle.\end{aligned}$$

In case $\Gamma = \hom(\Gamma', G)$ we have

$$\Delta^q(M, N; \hom(\Gamma', G)) \approx \hom(\Delta_q(M, N; \Gamma'), G)$$

an δ corresponds to $\hom(\partial, 1)$ under the above pairing.

The *singular cohomology* $H^q(M, N; \Gamma)$ is defined to equal $H^q(\Delta^*(M, N; \Gamma))$.

Consider the cochain complex $\Delta^*(U, V; \Gamma)$ as a functor of open pairs (U, V) in X. For an open triple (U, V, W) there is a short exact sequence

$$0 \to \Delta^*(U, V; \Gamma) \to \Delta^*(U, W; \Gamma) \to \Delta^*(V, W; \Gamma) \to 0.$$

If U, V are open sets in X then (by an argument similar to the proof of excision in Theorem 6.2) there is a weak cochain equivalence

$$\Delta^*(U \cup V, V; \Gamma) \to \Delta^*(U, U \cap V; \Gamma).$$

Furthermore, if $\{(U_j, V_j)\}_{j \in J}$ is a discrete family of open pairs in X there is an isomorphism

$$\Delta^*\left(\bigcup_{j \in J}(U_j, V_j); \Gamma\right) \approx \prod_{j \in J} \Delta^*(U_j, V_j; \Gamma).$$

Thus, $\Delta^*(\cdot, \cdot; \Gamma)$ is a nonnegative additive cochain prefunctor on X.

By Theorem 5.2 there is a corresponding cochain functor $\overline{\Delta}^*$ which associates to a closed pair (A, B) in X the cochain complex

$$\overline{\Delta}^*(A, G; \Gamma) = \varinjlim\{\Delta^*(U, V; \Gamma) \mid (U, V) \text{ an open neighborhood of } (A, B)\}$$

whose cohomology functor is denoted by $\overline{H}^*(A, B; \Gamma)$. It is a nonnegative additive cohomology functor on X. Note that $\overline{H}^*(A, B; \Gamma)$ is not the singular cohomology of (A, B) with coefficients Γ but is the limit of the singular cohomology of open neighborhoods of (A, B). (However, because $(X, \varnothing)$ is an open pair, $\overline{H}^*(X, \varnothing; \Gamma) = H^*(X, \varnothing; \Gamma)$ is the singular cohomology of $(X, \varnothing)$.) In case X is an HLC space, $\overline{H}^*(A, B; \Gamma)$ is isomorphic to the Čech-Alexander cohomology of (A, B) with coefficients Γ **[9]**.

Similarly by Theorem 5.5 there is a weakly additive cochain functor $\overline{\Delta}_c^*$ which assigns to a closed pair (A, B) in X the cochain complex

$$\overline{\Delta}_c^*(A, B; \Gamma) = \varinjlim\{\Delta^*(U, V; \Gamma) \mid (U, V) \text{ open cobounded neighborhood of } (A, B)\}$$

whose cohomology functor is denoted by $\overline{H}_c^*(A, B; \Gamma)$. It is a nonnegative weakly additive cohomology functor on X. In case X is HLC, $\overline{H}_c^*(A, B; \Gamma)$ is isomorphic to the Čech-Alexander cohomology of (A, B) with coefficients Γ and with compact supports.

For a pair (M, N) in X we define

$$\begin{aligned}&\underline{\Delta}_*(M, N; \hom(\Gamma, G))\\ &\quad = \varinjlim\{\hom(\overline{\Delta}_c^*(A, B; \Gamma), Q) \mid (A, B) \text{ closed } \subset (M, N)\}\end{aligned}$$

where Q is, as usual, an injective resolution of G. We define $\underline{H}_q(M, N; \hom(\Gamma, G)) = H_q(\underline{\Delta}_*(M, N; \hom(\Gamma, G)))$. Then for a closed pair (A, B), we have

$$\underline{\Delta}_*(A, B; \hom(\Gamma, G)) = \hom(\overline{\Delta}_c^*(A, B; \Gamma), Q).$$

By Corollary 5.3 there is an additive chain functor $\underline{\Delta}_*$ which associates to an open pair (U, V) in X the chain complex $\underline{\Delta}_*(U, V; \hom(\Gamma, G))$. We want to compare the additive homology functor

$$\underline{H}_*(\cdot, \cdot; \hom(\Gamma, G))$$

just defined with the additive homology functor defined in §7 by using locally finite singular chains. Note that $\underline{\Delta}_*$ (and $\underline{H}_*$) are defined for local coefficients of the form $\hom(\Gamma, G)$ for some local system Γ and some module G. It is not clear how much of a restriction this imposes on the resulting local system $\hom(\Gamma, G)$.

We begin with a pairing

$$\Delta^q(U', V'; \Gamma) \times {}^X\Delta_q^\infty(A, B; \hom(\Gamma, G)) \to G$$

for a closed pair (A, B) contained in an open cobounded pair (U', V'). Let $c = \sum_\sigma g_\sigma \sigma \in {}^X\Delta_q^\infty(A, B; \hom(\Gamma, G))$ and let $u \in \Delta^q(U', V'; \Gamma)$. Then u vanishes on every singular simplex σ in V'. Since $X - V'$ is compact there are only a finite number of σ's such that $g_\sigma \neq 0$ and $|\sigma| \cap (X - V') \neq \varnothing$. Therefore

$$\langle u, c\rangle = \sum_\sigma g_\sigma(u(\sigma))$$

is a finite sum of elements of G. In case

$$c = \sum_\sigma g_\sigma \sigma \in {}^X\Delta_{q+1}^\infty(A, B; \hom(\Gamma, G))$$

then $\partial c = \sum_\sigma \sum_{0\le i\le q+1}(-1)^i(g_\sigma|\sigma^{(i)})\sigma^{(i)}$ and

$$\begin{aligned}\langle u, \partial c\rangle &= \sum_\sigma \sum_{0\le i\le q+1} (-1)^i(g_\sigma|\sigma^{(i)})(u(\sigma^{(i)}))\\ &= \sum_\sigma \sum_{0\le i\le q+1} (-1)^i g_\sigma(\overline{u(\sigma^{(i)})})\\ &= \sum_\sigma g_\sigma(\delta u(\sigma)) = \langle \delta u, c\rangle.\end{aligned}$$

Passing to the direct limit as (U', V') varies over open cobounded neighborhoods of (A, B) we obtain a pairing

$$\overline{\Delta}_c^q(A, B; \Gamma) \times {}^X\Delta_q^\infty(A, B; \hom(\Gamma, G)) \to G$$

which corresponds to a homomorphism

$$\varphi\colon {}^X\Delta_q^\infty(A, B; \hom(\Gamma, G)) \to \hom(\overline{\Delta}_c^q(A, B; \Gamma), G)$$

such that $(\varphi(c))(\{u\}) = \langle u, c\rangle$ for $c \in {}^X\Delta_q^\infty(A, B; \hom(\Gamma, G))$ and $u \in \Delta^q(U', V'; \Gamma)$ where (U', V') is an open cobounded neighborhood of (A, B). For the injective resolution

$$0 \to G \xrightarrow{\eta} Q^0 \to Q^1 \to 0$$

we see that $\eta\varphi\colon {}^X\Delta_q^\infty(A, B; \hom(\Gamma, G)) \to \hom(\overline{\Delta}_c^q(A, B; \Gamma), Q)$ and

$$\begin{aligned}(\eta\varphi)(\partial c)(\{u\}) &= \eta\langle u, \partial c\rangle = \eta\langle \delta u, c\rangle = (\eta\varphi)(c)(\{\delta u\})\\ &= \partial(\eta\varphi(c))(\{u\})\end{aligned}$$

so that $(\eta\varphi)\partial = \partial(\eta\varphi)$.

Passing to the direct limit of both sides as (A, B) varies over closed pairs in (U, V) yields the homomorphism

$$\bar{\varphi}\colon {}^X\Delta_q^\infty(U, V; \hom(\Gamma, G)) \to \underline{\Delta}_q(U, V; \hom(\Gamma, G))$$

and $\bar{\varphi}\partial = \partial\bar{\varphi}$. So $\bar{\varphi}$ is a natural chain map and determines a homomorphism $\bar{\varphi}_*$ from ${}^XH_*^\infty(\cdot, \cdot; \hom(\Gamma, G))$, ∂_* to $\underline{H}_*(\cdot, \cdot; \hom(\Gamma, G))$, ∂_* both nonnegative additive homology functors on X.

If X is finite dimensional it would follow from Theorem 2.2 that if $\bar{\varphi}_*$ were an isomorphism for pairs of the form $(X, X - x)$ for $x \in X$ then $\bar{\varphi}_*$ would be an isomorphism for all open (U, V) in X. To obtain the local result we consider the case of an n manifold (i.e. a paracompact Hausdorff space in which each point has an open neighborhood homomorphic to $\mathbb{R}^n$).

Lemma 8.1. *If X is an n manifold, then*

$$\bar{\varphi}_*\colon {}^XH_*^\infty(X, X - x; \hom(\Gamma, G)) \approx \underline{H}_*(X, X - x; \hom(\Gamma, G))$$

is an isomorphism for all $x \in X$.

Proof. Since $X - (X - x) = x$ is compact, it follows from Lemma 7.2 that

$${}^XH_q^\infty(X, X - x; \hom(\Gamma, G)) \approx H_q(X, X - x; \hom(\Gamma, G)).$$

Let U be an open neighborhood of x with $\overline{U}$ homomorphic to a closed n ball. Then

$$H_q(\overline{U}, \overline{U} - x; \hom(\Gamma, G)) \approx H_q(X, X - x; \hom(\Gamma, G)).$$

Since $\overline{U}$ is simply connected, $\hom(\Gamma, G)$ is equivalent to a constant system on $\overline{U}$ so that

$$H_q(\overline{U}, \overline{U} - x; \hom(\Gamma, G)) \approx \begin{cases} 0, & q \neq n, \\ \hom(\Gamma_x, G), & q \neq n. \end{cases}$$

For the other group note that cofinal in the family of all closed pairs in $(X, X-x)$ are pairs of the form $(X, X-U)$ where U is as above. Then $\overline{\Delta}_c^*(X, X - U; \Gamma)$ and $\overline{\Delta}_c^*(\overline{U}, \overline{U} - U; \Gamma)$ are chain equivalent. Since $\overline{U}$ is simply connected,

$$H^q(\overline{\Delta}_c^*(\overline{U}, \overline{U} - U; \Gamma)) \approx \begin{cases} 0, & q \neq n, \\ \Gamma_x, & q = n. \end{cases}$$

Therefore,

$$\begin{aligned}
&\underline{H}_q(X, X-x; \hom(\Gamma, G)) \\
&\quad = H_q(\underline{\Delta}_*(X, X-x; \hom(\Gamma, G))) \\
&\quad = H_q(\varinjlim\{\hom(\overline{\Delta}^*_c(X, X-U; \Gamma), Q)\}) \\
&\quad \approx \varinjlim\{H_q(\hom\overline{\Delta}^*_c(X, X-U; \Gamma), Q)\} \\
&\quad \approx \varinjlim\{H_q(\hom\overline{\Delta}^*_c(\overline{U}, \overline{U}-U; \Gamma), Q)\} \\
&\quad \approx \begin{cases} 0, & q \neq n, \\ \hom(\Gamma_x, G), & q = n, \end{cases}
\end{aligned}$$

the last isomorphism by the universal coefficient formula.

Finally, we observe that the pairing we defined induces a pairing

$$^X H_n(\overline{U}, \overline{U}-x; \hom(\Gamma, G)) \times H^n_c(\overline{U}, \overline{U}-U; \Gamma) \to G$$

which is isomorphic to the evaluation pairing

$$\hom(\Gamma_x, G) \times \Gamma_x \to G.$$

Therefore

$$\bar{\varphi}_*\colon {}^X H^\infty_n(X, X-x; \hom(\Gamma, G)) \to \underline{H}_n(X - X - x; \hom(\Gamma, G))$$

is an isomorphism (corresponding to the identity map of $\hom(\Gamma_x, G) \to \hom(\Gamma_x, G)$ resulting from the evaluation pairing above). □

THEOREM 8.2. *For an n manifold X*

$$\bar{\varphi}_*\colon {}^X H^\infty_*(U, V; \hom(\Gamma, G)) \approx \underline{H}_*(U, V; \hom(\Gamma, G))$$

for every open pair (U, V) in X.

Proof. Using Lemma 8.1 the result follows immediately from Theorem 2.2 □

From the definitions we see that if (A, B) is closed in X then $\underline{H}_*(A, B; \hom(\Gamma, G)) = H_*(\hom(\overline{\Delta}^*_c(A, B; \Gamma), Q))$. We already remarked that if X is HLC then $\overline{\Delta}^*_c(A, B; \Gamma)$ has cohomology isomorphic to the Čech-Alexander cohomology of (A, B) with compact supports. It follows that in a locally compact HLC space X the groups $\underline{H}_*(A, B; \hom(\Gamma, G))$ are isomorphic to the Borel-Moore homology groups [1] of (A, B) with coefficients $\hom(\Gamma, G)$.

9. Homology-cohomology comparison. In this section we construct a homomorphism from singular homology ${}^X H_*^\infty(U, V)$ to singular cohomology $H^*(X - V, X - U)$ with suitable coefficient systems. This homomorphism will depend on a cohomology class of $H^*(X \times X, X \times X - \delta(X))$ where $\delta(X) = \{(x, y) \in X \times X | x = y\}$ is the diagonal of $X \times X$.

We shall assume $\mathscr{V}$ is an open covering of X such that if V, $V' \in \mathscr{V}$ and $V \cap V' \neq \varnothing$ then every closed path in $V \cup V'$ is null homotopic in X. If X is a paracompact space in which each point has an open neighborhood W with the property that every closed path in W is null homotopic in X, such a covering can be obtained as follows. Let $\mathscr{W}$ be an open covering of X by sets W such that every closed path in W is null homotopic in X ($\mathscr{W}$ is assumed to exist) and let $\mathscr{V}$ be an open star refinement of $\mathscr{W}$ covering X (such star refinements exist because X is paracompact). Then $\mathscr{V}$ has the desired property.

In §§6 and 7 it was noted that for any $A \subset X$ the inclusion maps

$$ {}^X\Delta_*^\infty(\mathscr{V} \cap A; \Gamma) \subset {}^X\Delta_*^\infty(A; \Gamma), \quad \Delta_*(\mathscr{V} \cap A; \Gamma) \subset \Delta_*(A; \Gamma) $$

are chain equivalences (where $\mathscr{V} \cap A = \{V \cap A | V \in \mathscr{V}\}$). If σ and σ' are singular simplexes such that $|\sigma| \subset V$, $|\sigma'| \subset V'$ for V, $V' \in \mathscr{V}$ and $|\sigma| \cap |\sigma'| \neq \varnothing$ then if $x \in |\sigma| \cap |\sigma'|$ there are isomorphisms

$$ \Gamma(\sigma) \underset{\approx}{\rightarrow} \Gamma_x \underset{\approx}{\leftarrow} \Gamma(\sigma'). $$

If x' is another point in $|\sigma| \cap |\sigma'|$ let ω be a path in $|\sigma|$ from x to x' and ω' a path in $|\sigma'|$ from x to x'. There are commutative triangles

$$\begin{tikzcd}
\Gamma(\sigma) \arrow[rr, "\approx"'] \arrow[dr, "\approx"'] & & \Gamma_x \arrow[dl, "\approx"', "\Gamma_\omega"] \\
& \Gamma_{x'} &
\end{tikzcd}
\qquad
\begin{tikzcd}
\Gamma_x \arrow[dr, "\Gamma_{\omega'}"', "\approx"] & & \Gamma(\sigma') \arrow[ll, "\approx"] \arrow[dl, "\approx"] \\
& \Gamma_{x'} &
\end{tikzcd}$$

Since ω, ω' have the same endpoints and lie in $V \cup V'$ they are homotopic in X so $\Gamma_\omega = \Gamma_{\omega'}: \Gamma_x \to \Gamma_{x'}$. It follows that the composite of the isomorphisms

$$ \Gamma(\sigma) \underset{\approx}{\rightarrow} \Gamma_x \underset{\approx}{\leftarrow} \Gamma(\sigma') $$

is independent of the choice of $x \in |\sigma| \cap |\sigma'|$ so there is a well-defined isomorphism $\lambda_{\sigma\sigma'}: \Gamma(\sigma) \approx \Gamma(\sigma')$.

If σ, σ', σ'' are singular simplexes such that $|\sigma| \subset V$, $|\sigma'| \subset V'$, $|\sigma''| \subset V''$ for V, V', $V'' \in \mathscr{V}$ and $|\sigma| \cap |\sigma'|$, $|\sigma| \cap |\sigma''|$, $|\sigma'| \cap |\sigma''|$ are all non-empty, then $\lambda_{\sigma\sigma''} = \lambda_{\sigma'\sigma''} \lambda_{\sigma\sigma'}: \Gamma(\sigma) \approx \Gamma(\sigma'')$.

Using this notation if $g \in \Gamma(\sigma)$, then $g|\sigma^{(j)} = \lambda_{\sigma\sigma(j)}(g)$. It follows that if $c = \sum_\sigma g_\sigma \sigma \in \Delta_*^\infty(\mathscr{V}; \Gamma)$, then

$$\partial c = \sum_\sigma \sum_j (-1)^j \lambda_{\sigma\sigma(j)}(g_\sigma)\sigma^{(j)}$$

and if $u \in \Delta^q(\mathscr{V}; \Gamma)$, then

$$(\delta u)(\sigma) = \sum_i (-1)^i \lambda_{\sigma^{(i)}\sigma}(u(\sigma^{(i)})).$$

If Γ and Γ' are local systems on X define $\Gamma \times \Gamma'$ to be the local system on $X \times X$ such that $(\Gamma \times \Gamma')_{(x,y)} = \Gamma_x \otimes \Gamma'_y$ and $(\Gamma \times \Gamma')_\omega = \Gamma_{\mathrm{pr}_1 \omega} \otimes \Gamma_{\mathrm{pr}_2 \omega}$ where $\mathrm{pr}_1 \colon X \times X \to X$ and $\mathrm{pr}_2 \colon X \times X \to X$ are projections to the first and second coordinates, respectively.

Let $U \in H^n(X \times X, X \times X - \delta(X); \hom(R \times \Gamma, R))$ be a given cohomology class (here we regard R as a constant local system on X in forming $R \times \Gamma$ on $X \times X$) and let

$$u \in \hom(\Delta_n(X \times X, X \times X - \delta(X); R \times \Gamma), R)$$

be a cocycle representing U.

Let $\tau \colon \Delta_*(X; R) \otimes \Delta_*(X; \Gamma) \to \Delta_*(X \times X, R \times \Gamma)$ be an Eilenberg-Zilber map [8]. Then τ is a natural chain map (so $\tau(\Delta_*(Y; R) \otimes \Delta_*(Z; \Gamma)) \subset \Delta_*(Y \times Z, R \times \Gamma)$ for all Y, $Z \subset X$), and in dimension 0, if $[x]$ denotes the 0-singular simplex at the point x, then

$$\tau(\alpha[x] \otimes g[y]) = \alpha g[(x, y)]$$

for $\alpha \in R$, $g \in \Gamma_y$ and x, $y \in X$. For an arbitrary local system Γ' on X define

$$\theta \colon {}^X\Delta_q^\infty(\mathscr{V}; \Gamma \otimes \Gamma') \to \Delta^{n-q}(\mathscr{V}; \Gamma')$$

by requiring it to be R linear and such that if σ' is an $(n-q)$ singular simplex in V' where $V' \in \mathscr{V}$ and $c \in {}^X\Delta_q^\infty(\mathscr{V} \cap (X - |\sigma|); \Gamma \otimes \Gamma')$ then

$$\theta(c)(\sigma') = 0$$

while if $c = (g \otimes g')\sigma \in \Delta_q(\mathscr{V}, \Gamma \otimes \Gamma')$ then

$$\theta(c)(\sigma') = \langle u, \tau(\sigma' \otimes g\sigma)\rangle \lambda_{\sigma\sigma'}(g') = \lambda_{\sigma\sigma'}(\langle u, \tau(\sigma' \otimes g\sigma)\rangle g')$$

(the right-hand side is 0 if $|\sigma| \cap |\sigma'| = \varnothing$ because, in that case

$$\tau(\sigma' \otimes g\sigma) \subset X \times X - \delta(X)$$

and u vanishes on $\Delta_n(X \times X, X \times X - \delta(X); R \times \Gamma)$). This uniquely defines θ because, for every singular simplex σ', given $c \in {}^X\Delta_q^\infty(\mathscr{V}; \Gamma \otimes \Gamma')$ there exist $c_1 \in \Delta_q(\mathscr{V}; \Gamma \otimes \Gamma')$ and $c_2 \in {}^X\Delta_q^\infty(\mathscr{V} \cap (X - |\sigma'|); \Gamma \otimes \Gamma)$ such that $c = c_1 + c_2$ (so that $\theta(c_2)(\sigma') = 0$ and $\theta(c_1)(\sigma')$ is a finite sum of terms of the form described above).

Given σ' in $\Delta_{n-q+1}(V')$ for some $V' \in \mathscr{V}$ let $c = c_1 + c_2$ where $c_1 \in \Delta_q(\mathscr{V}; \Gamma \otimes \Gamma')$ and $c_2 \in {}^X\Delta_q^\infty(\mathscr{V} \cap (X - |\sigma'|); \Gamma \otimes \Gamma')$. Then $\delta(\theta(c)) = \delta(\theta(c_1)) + \delta(\theta(c_2))$ and $(\delta(\theta(c_2)))(\sigma') = 0$ so $(\delta(\theta(c)))(\sigma') = (\delta(\theta(c_1)))(\sigma')$. To calculate $(\delta(\theta(c_1)))(\sigma')$ we need only calculate it for c_1 of the form $(g \otimes g')\sigma$ where $|\sigma| \cap |\sigma'| \neq \varnothing$. In this case

$$\begin{aligned}(\delta\theta((g \otimes g')\sigma))(\sigma') &= \sum_i (-1)^i \lambda_{\sigma'^{(i)}\sigma'}[\theta((g \otimes g')\sigma)(\sigma'^{(i)})] \\ &= \sum_i (-1)^i \lambda_{\sigma'^{(i)}\sigma'}[\langle u, \tau(\sigma'^{(i)} \otimes g\sigma)\rangle \lambda_{\sigma\sigma'^{(i)}}(g')]\end{aligned}$$

where the corresponding term on the right is 0 if $|\sigma| \cap |\sigma'^{(i)}| = \varnothing$. If $|\sigma| \cap |\sigma'^{(i)}| \neq \varnothing$, then $\lambda_{\sigma'^{(i)}\sigma'}\lambda_{\sigma\sigma'^{(i)}} = \lambda_{\sigma\sigma'}$ so we obtain

$$\sum_i (-1)^i \langle u, \tau(\sigma'^{(i)} \otimes g\sigma)\rangle \lambda_{\sigma\sigma'}(g') = \langle u, \tau(\partial\sigma' \otimes g\sigma)\rangle \lambda_{\sigma\sigma'}(g').$$

Now $\partial\tau(\sigma' \otimes g\sigma) = \tau(\partial\sigma' \otimes g\sigma) + (-1)^{n-q+1}\tau(\sigma' \otimes \partial(g\sigma))$ so that $\tau(\partial\sigma' \otimes g\sigma) = \partial\tau(\sigma' \otimes g\sigma) + (-1)^{n-q}\tau(\sigma' \otimes \partial(g\sigma))$, and we obtain

$$\begin{aligned}&\langle u, \partial\tau(\sigma' \otimes g\sigma)\rangle \lambda_{\sigma\sigma'}(g') + (-1)^{n-q}\langle u, \tau(\sigma' \otimes \partial(g\sigma))\rangle \lambda_{\sigma\sigma'}(g') \\ &\quad = \langle \delta u, \tau(\sigma' \otimes g\sigma)\rangle \lambda_{\sigma\sigma'}(g') \\ &\qquad + (-1)^{n-q}\sum_j (-1)^j \langle u, \tau(\sigma' \otimes \lambda_{\sigma\sigma^{(j)}}(g)\sigma^{(j)})\rangle \lambda_{\sigma\sigma'}(g').\end{aligned}$$

Because $\delta u = 0$ since u is a cocycle, this equals

$$(-1)^{n-q}\sum_j (-1)^j \langle u, \tau(\sigma' \otimes \lambda_{\sigma\sigma^{(j)}}(g)\sigma^{(j)})\rangle \lambda_{\sigma\sigma'}(g').$$

The corresponding term on the right equals 0 if $|\sigma^{(j)}| \cap |\sigma'| = \varnothing$ and if $|\sigma^{(j)}| \cap |\sigma'| \neq \varnothing$, then $\lambda_{\sigma\sigma'} = \lambda_{\sigma^{(j)}\sigma'}\lambda_{\sigma\sigma^{(j)}}$ so we obtain

$$\begin{aligned}&(-1)^{n-q}\sum_j (-1)^j \langle u, \tau(\sigma' \otimes \lambda_{\sigma\sigma^{(j)}}(g)\sigma^{(j)})\rangle \lambda_{\sigma^{(j)}\sigma'}\lambda_{\sigma\sigma^{(j)}}(g') \\ &\quad = (-1)^{n-q}\sum_j (-1)^j \lambda_{\sigma^j\sigma'}[\langle u, \tau(\sigma' \otimes \lambda_{\sigma\sigma^{(j)}}(g)\sigma^{(j)})\rangle \lambda_{\sigma\sigma^{(j)}}(g')]\end{aligned}$$

and, by definition, this equals $(-1)^{n-q}\theta(\partial(g \otimes g')\sigma)(\sigma')$. Therefore, θ maps the chain complex ${}^X\Delta_*^\infty(\mathscr{V}; \Gamma \otimes \Gamma')$ into the cochain complex $\Delta^*(\mathscr{V}; \Gamma')$ so that it commutes up to sign with ∂ and δ.

If $B \subset X$ is arbitrary, then

$$\tau(\Delta_*(\mathscr{V} \cap (X - B)\,;\, R) \otimes \Delta_*(\mathscr{V} \cap B\,;\, \Gamma)) \subset \Delta_*((X - B) \times B\,;\, R \times \Gamma)$$

and u vanishes on the latter. Therefore, θ maps

$${}^X\Delta_*^\infty(\mathscr{V} \cap (X - B)\,;\, \Gamma \otimes \Gamma') \quad \text{into} \quad \Delta^*(\mathscr{V}\,,\, \mathscr{V} \cap B\,;\, \Gamma').$$

If $B \subset A \subset X$, the following diagram is commutative (the top row is a short exact sequence with coefficients $\Gamma \otimes \Gamma'$, the bottom row is a short exact sequence with coefficients Γ', and the right-hand vertical map is defined to make the diagram commutative)

$$\begin{tikzcd}
0 \arrow[r] & {}^X\Delta_*^\infty(\mathscr{V} \cap (X - A)) \arrow[r] \arrow[d, "\theta"'] & {}^X\Delta_*^\infty(\mathscr{V} \cap (X - B)) \arrow[r] \arrow[d, "\theta"'] & {}^X\Delta_*^\infty(\mathscr{V} \cap (X - B)\,,\, \mathscr{V} \cap (X - A)) \arrow[r] \arrow[d, "\theta"'] & 0 \\
0 \arrow[r] & \Delta^*(\mathscr{V}\,,\, \mathscr{V} \cap A) \arrow[r] & \Delta^*(\mathscr{V}\,,\, \mathscr{V} \cap B) \arrow[r] & \Delta^*(\mathscr{V} \cap A\,,\, \mathscr{V} \cap B) \arrow[r] & 0.
\end{tikzcd}$$

If $(A', B') \subset (A, B)$ there is a commutative square

$$\begin{tikzcd}
{}^X\Delta_*^\infty(\mathscr{V} \cap (X - B)\,,\, \mathscr{V} \cap (X - A)\,;\, \Gamma \otimes \Gamma') \arrow[r, "\imath"] \arrow[d, "\theta"'] & {}^X\Delta_*^\infty(\mathscr{V} \cap (X - B')\,,\, \mathscr{V} \cap (X - A')\,;\, \Gamma \otimes \Gamma') \arrow[d, "\theta"'] \\
\Delta^*(\mathscr{V} \cap A\,,\, \mathscr{V} \cap B\,;\, \Gamma') \arrow[r, "\rho"] & \Delta^*(\mathscr{V} \cap A'\,,\, \mathscr{V} \cap B'\,;\, \Gamma').
\end{tikzcd}$$

Let (A, B) be a closed pair. Taking the direct limit of the homomorphisms (as (V, W) varies over open neighborhoods of (A, B))

$$\theta\colon {}^X\Delta_*^\infty(\mathscr{V} \cap (X - W)\,,\, \mathscr{V} \cap (X - V)\,;\, \Gamma \otimes \Gamma') \to \Delta^*(\mathscr{V} \cap V\,,\, \mathscr{V} \cap W\,;\, \Gamma')$$

determines the homomorphism

$$\bar{\theta}\colon {}^X\Delta_*^\infty(\mathscr{V} \cap (X - B)\,,\, \mathscr{V} \cap (X - A)\,;\, \Gamma \otimes \Gamma') \to \overline{\Delta}^*(\mathscr{V} \cap A\,,\, \mathscr{V} \cap B\,;\, \Gamma').$$

Similarly taking the direct limit as (V, W) varies over open cobounded neighborhoods of (A, B) in X (and observing that, in case (V, W) is cobounded,

$$\begin{aligned} &{}^X\Delta_*^\infty(\mathscr{V} \cap (X - W)\,,\, \mathscr{V} \cap (X - V)\,;\, \Gamma \otimes \Gamma') \\ &\quad = \Delta_*(\mathscr{V} \cap (X - W)\,,\, \mathscr{V} \cap (X - V)\,;\, \Gamma \otimes \Gamma')) \end{aligned}$$

we obtain a homomorphism

$$\bar{\theta}^c\colon \Delta_*(\mathscr{V} \cap (X - B)\,,\, \mathscr{V} \cap (X - A)\,;\, \Gamma \otimes \Gamma') \to \overline{\Delta}_c^*(\mathscr{V} \cap A\,,\, \mathscr{V} \cap B\,;\, \Gamma')).$$

The homomorphisms $\bar{\theta}$ and $\bar{\theta}^c$ depend on the choice of the cocycle u and the Eilenberg-Zilber map τ. Altering u in the cohomology class

U or altering τ will alter $\bar\theta$ and $\bar\theta^c$ by a chain homotopy. Therefore, $\bar\theta$ induces uniquely defined homomorphisms

$$\bar\theta_U\colon {}^X H_q^\infty(\mathscr{V}\cap(X-B),\mathscr{V}\cap(X-A);\Gamma\otimes\Gamma') \\ \to \overline{H}^{n-q}(\mathscr{V}\cap A,\mathscr{V}\cap B;\Gamma')$$

and

$$\bar\theta_U^c\colon H_q(\mathscr{V}\cap(X-B),\mathscr{V}\cap(X-A);\Gamma\otimes\Gamma') \\ \to \overline{H}_c^{n-q}(\mathscr{V}\cap A,\mathscr{V}\cap B;\Gamma').$$

The homomorphism

$$\bar\theta_U\colon {}^X H_q^\infty(X-B,X-A;\Gamma\otimes\Gamma')\to\overline{H}^{n-q}(A,B;\Gamma')$$

is defined so that commutativity holds in the square

$$\begin{tikzcd}
{}^X H_q^\infty(\mathscr{V}\cap(X-B),\mathscr{V}\cap(X-A);\Gamma\otimes\Gamma') \arrow[r, "\approx"'] \arrow[d, "\bar\theta_U"'] & {}^X H_q^\infty(X-B,X-A;\Gamma\otimes\Gamma') \arrow[d, "\bar\theta_U"] \\
\overline{H}^{n-q}(\mathscr{V}\cap A,\mathscr{V}\cap B;\Gamma') & \overline{H}^{n-q}(A,B;\Gamma'). \arrow[l, "\approx"]
\end{tikzcd}$$

Similarly the homomorphism

$$\bar\theta_U^c\colon H_q(X-B,X-A;\Gamma\otimes\Gamma')\to\overline{H}_c^{n-q}(A,B;\Gamma')$$

is defined so that commutativity holds in the square

$$\begin{tikzcd}
H_q(\mathscr{V}\cap(X-B),\mathscr{V}\cap(X-A);\Gamma\otimes\Gamma') \arrow[r, "\approx"'] \arrow[d, "\bar\theta_U^c"'] & H_q(X-B,X-A;\Gamma\otimes\Gamma') \arrow[d, "\bar\theta_U^c"] \\
\overline{H}_c^{n-q}(\mathscr{V}\cap A,\mathscr{V}\cap B;\Gamma') & \overline{H}_c^{n-q}(A,B;\Gamma'). \arrow[l, "\approx"]
\end{tikzcd}$$

Both $\bar\theta_U$ and $\bar\theta_U^c$ are natural and commute up to sign with connecting homomorphisms. The image $\bar\theta_U(z)$ is the slant product U/z [7, **8**] for local coefficients.

10. Duality in manifolds. Throughout this section X will be assumed to be an n manifold. We prove various duality theorems relating homology of a pair in X to cohomology of the complementary pair.

THEOREM 10.1. *Suppose $X=\bigcup_{i=0}^\infty \operatorname{int} C_i$ where $\overline{C}_i$ is compact and $C_i\subset C_{i+1}$ for each i. Then for any local system on X,*

$$H_n^\infty(X;\Gamma)\approx\varprojlim\{H_n(X,X-C_i,\Gamma)\}.$$

Proof. By Lemma 1 on p. 299 of **[8]** $H_{n+1}(X, X - C_i, \Gamma) = 0$ for each i (the lemma referred to asserts the result for a constant local system G, but the same argument establishes the result for an arbitrary local system Γ). The theorem follows from this using Theorem 7.3. □

It is straightforward to verify that $\varprojlim \{H_n(X, X - C_i; G)\}$ is the same as $H_n^c(X; G)$ as defined on p. 299 of **[8]**. From Theorem 6 on p. 303 of **[8]** it follows that for a connected n manifold X, $H_n^\infty(X; R) \neq 0$ if and only if X is orientable over R, and from Theorem 5 on p. 302 of **[8]** there is a bijection between orientations of X over R and generators of $H_n^\infty(X; R)$.

In general there is a local system Γ^X on X with

$$\Gamma_x^X = H^n(X, X - x; R)$$

and, if ω is a path in X, $\Gamma_\omega^X : \Gamma_{\omega(0)}^X \to \Gamma_{\omega(1)}^X$ is suitably defined as in [7] by "moving along ω". The dual local system $\Gamma^{X*} = \hom(\Gamma^X, R)$ has the property that $\Gamma_x^{X*} \approx H_n(X, X - x; R)$ for all $x \in X$. In fact, $\Gamma_x^X \approx R$ for each $x \in X$ and Γ_ω^X corresponds to multiplication of R by ±1 depending on ω. Therefore, $\Gamma^{X*} \approx \Gamma^X$. On $X \times X$ the local system $R \times \Gamma^{X*}$ is isomorphic to $\hom(R \times \Gamma^X, R)$.

For $x \in X$ there are isomorphisms

$$\begin{aligned} H_n(X, X - x; \Gamma^X) &\approx \hom(H^n(X, X - x; R); \Gamma_x^X) \\ &\approx \hom(H^n(X, X - x; R), H^n(X, X - x; R)) \end{aligned}$$

and $z_x \in H_n(X, X - x; \Gamma^X)$ will denote the element corresponding under the above to the identity map of $H^n(X, X - x; R)$.

A *Thom class* on X is an element

$$U \in H^n(X \times X, X \times X - \delta(X); \hom(R \times \Gamma^X, R))$$

such that, for each $x \in X$, $\bar{\theta}_U^c(z_x) = 1 \in \overline{H}_c^0(x; R)$. It is known (Theorem 4.7 in [7]) that every manifold has a unique Thom class. In the sequel we use the Thom class U in defining $\bar{\theta}$ and $\bar{\theta}^c$ and omit specific reference to U in the notation.

THEOREM 10.2. *For every closed pair (A, B) in X and every local system Γ of R modules on X there are isomorphisms*

$$\bar{\theta} : {}^X H_q^\infty(X - B, X - A; \Gamma^X \otimes \Gamma) \approx \overline{H}^{n-q}(A, B; \Gamma)$$

and

$$\bar{\theta}^c\colon H_q(X-B\,,X-A\,;\Gamma^X\otimes\Gamma)\approx\overline{H}_c^{n-q}(A\,,B\,;\Gamma).$$

Proof. Define cohomology theories H' and $\overline{H}$ on X by $H'^j(A\,,B) = {}^XH^\infty_{n-j}(X-B\,,X-A\,;\Gamma^X\otimes\Gamma)$ and $\overline{H}^j(A\,,B)=\overline{H}^j(A\,,B\,;\Gamma)$ with δ', $\bar{\delta}$ suitably defined connecting homomorphisms. Then $\bar{\theta}$ is a homomorphism of H', δ' into $\overline{H}$, $\bar{\delta}$ which is an isomorphism for every $x\in X$ (because $H^j(x) = {}^XH^\infty_{n-j}(X\,,X-x\,;\Gamma^X\otimes\Gamma) = H_{n-j}(X\,,X-x\,;\Gamma^X\otimes\Gamma)\approx H_{n-j}(X\,,X-x\,;R)\otimes\Gamma_x\overline{H}^j(x) = \overline{H}^j(x\,;\Gamma)\approx\overline{H}^j(x\,;R)\otimes\Gamma_x$, and both sides are 0 except for $j=0$ when $\bar{\theta}$ is an isomorphism by the choice of U). Since H' and $\overline{H}$ are additive and X is finite dimensional, it follows from Theorem 2.1 that $\bar{\theta}$ is an isomorphism for all closed $(A\,,B)$ in X.

The result for $\bar{\theta}^c$ is obtained similarly because the two sides being compared are weakly additive cohomology functors and $\bar{\theta}^c$ is an isomorphism for every $x\in X$. □

REMARK 10.3. Replacing Γ by $\Gamma^X\otimes\Gamma$ and noting that $\Gamma^X\otimes\Gamma^X\approx R$ so that $\Gamma^X\otimes(\Gamma^X\otimes\Gamma)\approx R\otimes\Gamma\approx\Gamma$ we see that for $(A\,,B)$ closed in X there are also isomorphisms

$$\bar{\theta}\colon {}^XH^\infty_q(X-B\,,X-A\,;\Gamma)\approx\overline{H}^{n-q}(A\,,B\,;\Gamma^X\otimes\Gamma)$$

and

$$\bar{\theta}^c\colon H_q(X-B\,,X-A\,;\Gamma)\approx\overline{H}_c^{n-q}(A\,,B\,;\Gamma^X\otimes\Gamma)$$

for an arbitrary local system Γ.

Theorem 10.2 and Remark 10.3 express duality between the two types of singular homology groups of an open pair in X (i.e. weakly additive or additive homology) with two types of Čech-Alexander cohomology groups of the complementary closed pair (either weakly additive or additive cohomology) with arbitrary coefficient systems. This duality is not just an isomorphism of the homology groups with cohomology groups but is an isomorphism of cohomology theories with all its implications. There is also the following result which expresses duality between the Borel-Moore homology of a closed pair in X and the singular cohomology of the complementary open pair (which equals the Čech-Alexander cohomology) of the open pair.

THEOREM 10.4. *If $(A\,,B)$ is a closed pair in X and Γ is a local system on X there is an isomorphism*

$$\underline{H}_{n-q}(A\,,B\,;\mathrm{hom}(\Gamma\,,G))\approx H^q(X-B\,,X-A\,;\Gamma^X\otimes\mathrm{hom}(\Gamma\,,G)).$$

Proof. In §9 we defined a map

$$\bar{\theta}^c\colon \Delta_*(\mathscr{V}\cap(X-B),\, \mathscr{V}\cap(X-A);\, \Gamma^X\otimes\Gamma)\to\overline{\Delta}^*_c(\mathscr{V}\cap A,\, \mathscr{V}\cap B;\, \Gamma)$$

for (A, B) a closed pair in X. Define a chain complex (C, ∂) so that $C_q=\overline{\Delta}_c^{n-q}(\mathscr{V}\cap A, \mathscr{V}\cap B; \Gamma)$ and $\partial\colon C_q\to C_{q-1}$ equals $\delta\colon \overline{\Delta}_c^{n-q}(\mathscr{V}\cap A, \mathscr{V}\cap B;\Gamma)\to\overline{\Delta}_c^{n-q+1}(\mathscr{V}\cap A, \mathscr{V}\cap B;\Gamma)$. Then $\bar{\theta}^c$ is a map of degree 0 from $\Delta_*(\mathscr{V}\cap(X-B), \mathscr{V}\cap(X-A); \Gamma^X\otimes\Gamma)$ to C_* which commutes up to sign with ∂. By Theorem 10.2, $\bar{\theta}^c$ induces an isomorphism on homology. It follows that $\bar{\theta}^c$ also induces an isomorphism on cohomology

$$H^*(C_*;G)\approx H^*(\Delta_*(\mathscr{V}\cap(X-B),\, \mathscr{V}\cap(X-A);\, \Gamma^X\otimes\Gamma);\, G)$$

for any R module G. Because of the way C_* is defined, this yields an isomorphism

$$\begin{aligned}&\underline{H}_{n-q}(\mathscr{V}\cap A,\, \mathscr{V}\cap B;\, \hom(\Gamma, G))\\ &\qquad\approx H^q(\mathscr{V}\cap(X-B),\, \mathscr{V}\cap(X-A);\, \hom(\Gamma^X\otimes\Gamma, G))\end{aligned}$$

and this corresponds to an isomorphism

$$\underline{H}_{n-q}(A, B;\hom(\Gamma, G))\approx H^q(X-B, X-A;\hom(\Gamma^X\otimes\Gamma, G)).$$

Because of the special nature of the local system Γ^X it is easy to see that $\hom(\Gamma^X\otimes\Gamma, G)\approx\Gamma^X\otimes\hom(\Gamma, G)$ so that

$$\underline{H}_{n-q}(A, B;\hom(\Gamma, G))\approx H^q(X-B, X-A;\Gamma^X\otimes\hom(\Gamma, G)).$$

□

In case X is orientable, $\Gamma^X\approx R$ and the Theorems 10.2 and 10.4 assert isomorphisms of homology with coefficients in a constant system with cohomology in the same constant system. In the non-orientable case, however, if the homology is in a constant local system the corresponding cohomology has coefficients in a non-constant local system and vice versa.

REFERENCES

[1] A. Borel and J. Moore, *Homology theory for locally compact spaces*, Michigan Math. J., 7 (1960), 137–159.

[2] H. Cartan, *Séminaire de topologie algébrique*, ENS, (1948–49).

[3] A. Dold, *Zur Homotopietheorie der Kettenkomplexe*, Math. Annalen, **140** (1960), 278–298.

[4] J. Dugundji, *Topology*, Allyn and Bacon, Boston, Mass., 1968.

[5] S. Eilenberg, *Singular homology theory*, Ann. of Math., **45** (1944), 407–447.

[6] W. S. Massey, *Homology and Cohomology Theory*, Marcel Dekker, New York, 1978.

[7] E Spanier, *Duality in topological manifolds*, in *Colloque de Topologie Tenu a Bruxelles* (Centre Belge de Recherche Mathèmatiques) (1966), 91–111.

[8] ——, *Algebraic Topology*, Springer-Verlag, New York, New York, 1982.

[9] ——, *Cohomology isomorphisms*, Contemp. Math., **12** (1982), 315–329.

[10] ——, *Cohomology with supports*, Pacific J. Math., **123** (1986), 447–464.

[11] ——, *Cohomology theories on compact and locally compact spaces*, Revisita Matemática Iberoamericana, **2** (1986), 29–53.

[12] ——, *Cohomology theories on spaces*, Trans. Amer. Math. Soc., **301** (1987), 149–161.

[13] ——, *Weakly additive cohomology*, Publicaciones Matemátiques, **34** (1990), 145–150.

[14] N. E. Steenrod, *Homology with local coefficients*, Ann. of Math., **44** (1943), 610–627.

Received August 27, 1990 and in revised form November 28, 1990.

UNIVERSITY OF CALIFORNIA
BERKELEY, CA 94720

PACIFIC JOURNAL OF MATHEMATICS

Volume 160 No. 1 September 1993

0030-8730(1993)160:1;1-N