SUB | NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

Werk

Titel: Commentarii Mathematici Helvetici
Verlag: Birkhäuser-Verlag
Jahr: 1981
Kollektion: Mathematica
Digitalisiert: Niedersächsische Staats- und Universitätsbibliothek Göttingen
Werk Id: PPN358147735_0056
PURL: http://resolver.sub.uni-goettingen.de/purl?PPN358147735_0056

Übergeordnetes Werk

Werk Id: PPN358147735
PURL: http://resolver.sub.uni-goettingen.de/purl?PPN358147735

Terms and Conditions

The Goettingen State and University Library provides access to digitized documents strictly for noncommercial educational, research and private purposes and makes no warranty with regard to their use for other purposes. Some of our collections are protected by copyright. Publication and/or broadcast in any form (including electronic) requires prior written permission from the Goettingen State- and University Library.
Each copy of any part of this document must contain there Terms and Conditions. With the usage of the library's online system to access or download a digitized document you accept the Terms and Conditions.
Reproductions of material on the web site may not be made for or donated to other repositories, nor may be further reproduced without written permission from the Goettingen State- and University Library.

For reproduction requests and permissions, please contact us. If citing materials, please give proper attribution of the source.

Contact

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Georg-August-Universität Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen
Germany
Email: gdz@sub.uni-goettingen.de

Commentarii Mathematici Helvetici

Volumen 56, 1981

Editi a Societate Mathematica Helvetica in Aedibus
Birkhäuser Verlag, Basel · Boston · Stuttgart

Commentarii Mathematici Helvetici

Redaktionskomitee - Editing committee - Comité de rédaction - Comitato di redazione
J.J. Burckhardt, A. Haefliger, M. Kervaire, A. Pfluger

Beirat - Board of trustees - Comité de patronage - Comitato consultativo
G. de Rham, M. Gut, Ch. Blanc, H. Hadwiger, G. Vincent, H. Jecklin, B. Eckmann, J. de Siebenthal, H. Huber, W. Nef, R. Bader, E. Specker, H. Kleisli, A. Delessert.

Die Commentarii Mathematici Helvetici sind Eigentum der Schweizerischen Mathematischen Gesellschaft. Sie erscheinen in Bänden von vier Heften. Es ist vorgesehen, jährlich einen Band herauszugeben. Manuskripte in Maschinenschrift sind zu senden an J.J. Burckhardt, Bergheimstr.4, CH-8032 Zürich. Die Autoren erhalten 50 Separatabzüge unentgeltlich.

Les Commentarii Mathematici Helvetici sont la propriété de la Société mathématique suisse. Ils paraissent en volumes de quatre fascicules; on prévoit un volume par an. Les mémoires doivent être remis dactylographiés à J.J. Burckhardt, 4, Bergheimstr., CH-8032 Zurich. Les auteurs reçoivent 50 tirages à part gratuitement.

I Commentarii Mathematici Helvetici sono proprietà della Società matematica svizzera. Escono in volumi di quattro fascicoli; si prevede un volume per anno. Le memorie (in dattilografia) vengono ricevute dal J.J. Burckhardt, 4, Bergheimstr., CH-8032 Zurigo. Gli autori ricevono 50 estratti gratuitamente.

The Commentarii Mathematici Helvetici are owned by the Swiss Mathematical Society. They appear in volumes of four fascicules. One volume per year is the rule. Manuscripts must be typewritten and should be submitted to J.J. Burckhardt, Bergheimstr.4, CH-8032 Zürich. 50 reprints will be provided free of charge.

Abonnementspreis eines Bandes, Prix d'abonnement par volume, Prezzo di abbonamento per volume, Price per volume	sFr. 184.-
Für Mitglieder der SMG, pour membres de la SMS, per i membri della SMS	sFr. 147.-
Einzelnummer, le numéro, il numero, single copy	sFr. 49.-

Abgewiesene Manuskripte werden nur retourniert, falls dies bei der Einsendung verlangt wird.

Publiziert mit Unterstützung der Schweizerischen Naturforschenden Gesellschaft
(Schweizerische Akademie der Naturwissenschaften).

Birkhäuser Verlag Basel

CH-4010 Basel
(Schweiz, Suisse, Switzerland)

NIEDERSÄCHS.
STAATS- U. UNIV.-
BIBLIOTHEK
GÖTTINGEN

© Birkhäuser Verlag, Basel 1980
ISSN 0010-2571

Commentarii Mathematici Helvetici

Mit dem Abschluss von Band 56 tritt der Unterzeichnete aus der Redaktion der CMH zurück.
Er dankt allen Kollegen, Mitarbeitern, Autoren, Tauschpartnern und dem Verlag für die 31-jährige freundschaftliche Zusammenarbeit.

Alle Manuskripte und Tauschsendungen sind ab sofort zu richten an:

Prof. Dr. Peter Gabriel
Mathematisches Institut
Freiestrasse 36

CH-8032 Zürich

With the completion of Volume 56, the undersigned is retiring from the Editing Committee. He wishes to thank all colleagues, co-workers, authors, and exchange partners for 31 years of amicable collaboration.

All manuscripts and exchange copies are now to be directed to:

Prof. Dr. Peter Gabriel
Mathematisches Institut
Freiestrasse 36

CH-8032 Zürich / Switzerland

Index Volumen 56, 1981

Akbulut, S., and King, H.: All knots are algebraic 339

Arbarello, E., e Cornalba, M.: Su una congettura di Petri 1

Bangert, V.: On the existence of escaping geodesics 59

Brandenburg, J., and Dyer, M.: On J.H.C. Whitehead's aspherical question I . 431

Brenton, L., Drucker, D., and Prins, G.C.E.: Graph theoretic techniques in algebraic geometry II: construction of singular complex surfaces of the rational cohomology type of CP^2 39

Brooks, R.: The fundamental group and the spectrum of the Laplacian 581

Chavel, I., and Feldman, E.A.: Spectra of manifolds with small handles 83

Cheng, S.-Y., and Li, P.: Heat kernel estimates and lower bound of eigenvalues 327

Cornalba, M., v. *Arbarello, E.:* 1

Dror, E., Dwyer, W.G., and Kan, D.M.: Self homotopy equivalences of virtually nilpotent spaces 599

Drucker, D., v. *Brenton, L.* 39

Duflot, J.: Depth and equivariant cohomology 627

Duren, P.L.: Arcs omitted by support points of univalent functions 352

Dwyer, W.G., v. *Dror, E.* 599

Dyer, M., v. *Brandenburg, J.* 431

Eisenbud, D., Hirsch, U., and Neumann, W.: Transverse foliations of Seifert bundles and self homeomorphism of the circle 638

Fehlmann, R.: Über extremale quasikonforme Abbildungen 558

Feldman, E.A., v. *Chavel, I.* 83

Fried, D., Goldman, W., and Hirsch, M.W.: Affine manifolds with nilpotent holonomy 487

Goldman, W., v. *Fried, D.* 487

Gromov, M.: Curvature, diameter and Betti numbers 179

Hayman, W.K., and Wu, J.-M.G.: Level sets of univalent functions 366

Heath-Brown, D.R.: An asymptotic series for the mean value of Dirichlet L-functions 148

Hillman, J.A.: Aspherical four-manifolds and the centres of two-knot groups . 465

Hirsch, M.W., v. *Fried, D.* 487

Hirsch, U., v. *Eisenbud, D.* 638

Hitt, L.R., and Sumners, D.W.: Many different disk knots with the same exterior 142

Hlawka, E.: Anwendung einer zahlengeometrischen Methode von C.L. Siegel auf Probleme der Analysis 66

Howie, J., and Schneebeli, H.R.: Permutation modules and projective resolutions 447

Hurder, S.: On the secondary classes of foliations with trivial normal bundles . 307

Kan, D. M., v. *Dror, E.* ... 599

Kantorovitz, S.: Characterization of unbounded spectral operators with spectrum in a half-line ... 163

Kearton, C., and Wilson, S. M. J.: Cyclic group actions on odd-dimensional spheres ... 615

King, H., v. *Akbulut, S.* ... 339

Krantz, S. G.: $H^p(R^n)$ is equidistributed with $L^p(R^n)$ ... 136

Kühnau, R.: Funktionalabschätzungen bei quasikonformen Abbildungen mit Fredholmschen Eigenwerten ... 297

Kuwabara, R.: Correction to 'On the characterization of flat metrics by the spectrum' ... 196

Lemaire, J.-M., et Sigrist, F.: Sur les invariants d'homotopie rationnelle liés à la L.S. catégorie ... 103

Li, P., v. *Cheng, S.-Y.* ... 327

McGehee, R.: Double collisions for a classical particle system with nongravitational interactions ... 524

Meeks III, W. H., and Yau, S.-T.: The equivariant Dehn's lemma and loop theorem ... 225

Moser, J.: An example of a Schroedinger equation with almost periodic potential and nowhere dense spectrum ... 198

Munkholm, H. J., and Pedersen, E. K.: Whitehead transfers for S^1-bundles, an algebraic description ... 404

Neumann, W., v. *Eisenbud, D.* ... 638

Pedersen, E. K., v. *Munkholm, H. J.* ... 404

Prins, G. C. E., v. *Brenton, L.* ... 39

Rohlfs, J.: The Lefschetz number of an involution on the space of classes of positive definite quadratic forms ... 272

Schneebeli, H. R., v. *Howie, J.* ... 447

Sigrist, F., v. *Lemaire, J.-M.* ... 103

Strebel, R.: On one-relator soluble groups ... 123

Sumners, D. W., v. *Hitt, L. R.* ... 142

Thurnheer, P.: Zu einem hyperbolischen Gitterpunktproblem ... 240

Wilson, S. M. J., v. *Kearton, C.* ... 615

Wolpert, S.: An elementary formula for the Fenchel-Nielsen twist ... 132

Wu, J.-M. G., v. *Hayman, W. K.* ... 366

Yau, S.-T., v. *Meeks III, W. H.* ... 225

Zimmermann, B.: Über Homöomorphismen n-dimensionaler Henkelkörper und endliche Erweiterungen von Schottky-Gruppen ... 474

Comment. Math. Helvetici **56** (1981) 1–38 0010–2571/81/001001–38$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Su una congettura di Petri

ENRICO ARBARELLO E MAURIZIO CORNALBA*

Introduzione

Nel suo lavoro "Über Spezialkurven," [19] p. 184, K. Petri, in modo quasi parentetico, afferma quanto, in linguaggio moderno, può essere espresso come segue

(0.1) *Dato un qualsiasi divisore D su di una curva C a moduli generali il prodotto "cup"*

$$\mu_0 : H^0(C, \mathcal{O}(D)) \otimes H^0(C, \Omega^1_C(-D)) \to H^0(C, \Omega^1_C)$$

è iniettivo.

Chiameremo (0.1) la congettura di Petri. Questa congettura giuoca un ruolo centrale nella teoria dei divisori speciali su di una curva e precisa la bella congettura di A. Brill e M. Noether che è stata recentemente dimostrata da P. Griffiths e J. Harris, [6], [1]. Ricordiamo il contenuto di questa congettura. Sia C una curva di genere g. Indichiamo con W^r_d l'insieme delle classi di equivalenza lineare dei divisori di grado d su C che si muovono in un sistema lineare di dimensione almeno r. Vi è su W^r_d una struttura naturale di varietà algebrica (anche non ridotta). La congettura di Brill e Noether afferma che

(0.2) *Se C è una curva a moduli generali allora:*
a) *W^r_d è ridotta.*
b) *Ogni componente irriducibile di W^r_d ha dimensione*

$$\rho = g - (r+1)(g-d+r)$$

* Lavoro eseguito nell'ambito dei Gruppi per la Matematica del C.N.R. e finanziato in parte dalla N.S.F. (Grant MCS-78-07348).

Nella prima parte di questo lavoro daremo un primo significato geometrico della congettura di Petri, mostrando che essa può essere pensata come un teorema di singolarità di Riemann per W^r_d su di una curva a moduli generali. Più precisamente mostreremo che

(0.3)

a) *La congettura di Petri implica* (0.2).

b) *Se vale la congettura di Petri e C è a moduli generali, il luogo singolare di* W^r_d *è esattamente* W^{r+1}_d.

Le tecniche da noi usate nello studiare la struttura di W^r_d sono simili a quelle usate da Kleiman e Laksov [14].

Seguendo il metodo usato da Kempf [12] nel dimostrare la sua generalizzazione del teorema di singolarità di Riemann mostreremo inoltre come la congettura di Petri permetta, su di una curva a moduli generali, di calcolare i coni tangenti a W^r_d e i loro gradi.

Il lavoro prosegue poi col dare una ulteriore interpretazione geometrica della congettura di Petri, e cioè:

(0.4) *Se C è una curva a moduli generali e*

$$\varphi : C \to \mathbf{P}^r$$

è un qualsiasi morfismo, ogni deformazione infinitesima di φ *è non ostruita. Più precisamente, denotando con* N_φ *il fascio normale a* φ, *si ha*

$$H^1(C, N_\varphi) = 0$$

Usando questa versione della congettura di Petri, dimostreremo la congettura stessa nel caso $r = 2$, cioè per le curve piane (il caso $r = 1$ è stato studiato classicamente e risolto da numerosi autori moderni).

La dimostrazione della congettura per $r = 2$ sarà basata su un curioso risultato di natura generale riguardante le deformazioni di curve con cuspidi. Più esattamente, dato un morfismo

$$\varphi : C \to \mathbf{P}^r$$

individueremo un sottofascio $\mathcal{K}_\varphi$ di N_φ avente supporto nei punti di diramazione di φ (le "cuspidi" di $\varphi(C)$) con le seguenti proprietà:

a) *Gli elementi di* $H^0(C, \mathcal{K}_\varphi)$ *corrispondono alle deformazioni infinitesime di* φ *che non mutano* $\varphi(C)$.

b) *Se* φ *non è composto con una involuzione, ogni deformazione effettiva di* φ, *definita, al primo ordine, da un elemento non nullo di* $H^0(C, \mathcal{K}_\varphi)$, *ha l'effetto di far diminuire la complessità delle cuspidi di* $\varphi(C)$.

I temi e le idee sviluppate in questo lavoro hanno avuto origine da una lunga e amichevole collaborazione con P. Griffiths e J. Harris (vedi [1]). Cogliamo qui l'occasione per ringraziarli vivamente entrambi. Vogliamo anche ringraziare Edoardo Sernesi per le numerose e interessanti conversazioni che abbiamo avuto con lui su questi problemi.

§1. Notazioni e preliminari

L'ambiente in cui opereremo è quello delle varietà algebriche su $\mathbf{C}$. Una tale varietà algebrica potrà anche essere non ridotta o riducibile. Poichè ciò non darà mai adito ad ambiguità, ci riterremo liberi, ove necessario, di trattare tali varietà come varietà analitiche complesse avvertendone, beninteso, il lettore.

Il termine curva sarà sempre usato nel senso di "varietà algebrica completa, irriducibile, non singolare e di dimensione uno." Per convenzione useremo il termine "fibrato in rette" in luogo di "fascio algebrico invertibile." Data una varietà algebrica liscia indicheremo con i simboli K_X, Θ_X (o più semplicemente K, Θ) il fibrato canonico su X e il fascio tangente su X.

Sia ora C una curva di genere g. Indicheremo con C_d il prodotto simmetrico d-esimo di C e con $u = u_d$ il morfismo

$$(1.1) \quad u_d : C_d \to \mathrm{Pic}^d(C)$$

che a ogni divisore D associa il fibrato in rette $\mathcal{O}(D)$. Indicheremo altresì con il simbolo C_d^r la sottovarietà di C_d di "equazioni":

$$\text{rango } u_* \leq d - r$$

Più esattamente, l'ideale di C_d^r è generato, localmente, dai minori $(d-r+1)\times(d-r+1)$ della matrice jacobiana di u. E'un fatto classico [2] che C_d^r, come insieme di punti, non è altro che il luogo dei $D \in C_d$ tali che

$$\dim H^0(C, \mathcal{O}(D)) \geq r+1.$$

Esporremo ora, per convenienza del lettore, alcuni rudimenti della teoria della

deformazioni di fibrati in rette su curve (cf. [15], [1] per una trattazione più ampia).

In primo luogo, se C è una curva di genere g e L è un fibrato in rette di grado d su C, vi è un isomorfismo naturale tra $H^1(C, \mathcal{O}_C)$ e lo spazio tangente a $\mathrm{Pic}^d(C)$ in L. Analogamente, se D è un punto di C_d, si ha una identificazione naturale tra lo spazio tangente a C_d nel punto D e $H^0(C, \mathcal{O}_D(D))$.

(1.2) *Con queste identificazioni, l'omomorfismo*

$$u_{*,D}: T_D(C_d) \to T_{u(D)}(\mathrm{Pic}^d(C))$$

si identifica all'omomorfismo cobordo

$$H^0(C, \mathcal{O}_D(D)) \to H^1(C, \mathcal{O}_C)$$

della successione esatta di coomologia di

$$0 \to \mathcal{O}_C \to \mathcal{O}_C(D) \to \mathcal{O}_D(D) \to 0$$

Indichiamo con il simbolo $\mathbf{C}[\varepsilon]$ l'anello dei numeri duali. Per la proprietà universale di $\mathrm{Pic}^d(C)$ [7], un elemento φ di $H^1(C, \mathcal{O})$ corrisponde a un fibrato in rette $\mathcal{L}$ su $C \times \mathrm{Spec}\, \mathbf{C}[\varepsilon]$ la cui restrizione a C è L. Nel seguito un tale fibrato in rette sarà chiamato una *deformazione infinitesima di* L e φ la sua *classe di Kodaira–Spencer.*

Si ha una successione esatta

$$(1.3) \quad H^0(C \times \mathrm{Spec}\, \mathbf{C}[\varepsilon], \mathcal{L}) \to H^0(C, L) \xrightarrow{f} H^1(C, L)$$

ove f è il prodotto "cup" con φ.

Analogamente, per *deformazione infinitesima* di $L \to C$ intenderemo il dato di una deformazione di C, $X \to \mathrm{Spec}\, \mathbf{C}[\varepsilon]$, e di un fibrato in rette $\mathcal{L}$ su X la cui restrizione a C è isomorfa a L.

Sia ora Σ_L l'$\mathcal{O}_C$-modulo localmente libero di rango 2 le cui sezioni sono gli operatori differenziali, di ordine al più eguale a uno, agenti su sezioni di L.

A ogni deformazione infinitesima di $L \to C$ è associato un elemento di $H^1(C, \Sigma_L)$ che è chiamato la sua *classe di Kodaira–Spencer.* Si mostra facilmente che ciò induce una corrispondenza biunivoca tra l'insieme delle classi di equivalenza di deformazioni infinitesime di $L \to C$ e $H^1(C, \Sigma_L)$.

Sia ora σ un elemento di $H^1(C, \Sigma_L)$ e $\mathcal{L} \to X \to \mathrm{Spec}\, \mathbf{C}[\varepsilon]$ la corrispondente deformazione infinitesima di $L \to C$. In analogia con quanto si è osservato per le

deformazioni infinitesime di un fibrato in rette su di una curva fissa, vi sono un prodotto "cup"

$$(1.4)\quad H^1(C, \Sigma_L)\otimes H^0(C, L)\to H^1(C, L)$$

e una successione esatta

$$H^0(X, \mathscr{L})\to H^0(C, L)\xrightarrow{F} H^1(C, L)$$

dove F è il prodotto "cup" con σ.

Si ha infine una successione esatta

$$(1.5)\quad 0\to \mathcal{O}_C\to \Sigma_L\to \Theta_C\to 0$$

e l'immagine di σ in $H^1(C, \Theta_C)$ è la classe di Kodaira–Spencer della deformazione $X\to \operatorname{Spec}\mathbf{C}[\varepsilon]$.

Sia ora $X\xrightarrow{p} S$ un morfismo liscio e proiettivo di varietà algebriche le cui fibre sono curve di genere g. Supporremo che p possegga una sezione. Esistono allora [7] una varietà $\operatorname{Pic}^d(p)$, un morfismo liscio e proiettivo $\operatorname{Pic}^d(p)\xrightarrow{\pi} S$ e un fibrato in rette $\mathscr{L}_d$ su $X\times_S \operatorname{Pic}^d(p)$, detto *fibrato di Poincaré*, con la seguente proprietà.

(1.6) *Per ogni morfismo* $S'\xrightarrow{\psi} S$ *e ogni fibrato in rette* $\mathscr{L}$ *su* $X\times_S S'$, *tale che la restrizione di* $\mathscr{L}$ *a ogni fibra di* $q: X\times_S S'\to S'$ *abbia grado d, esiste un unico morfismo* $S'\xrightarrow{\varphi}\operatorname{Pic}^d(p)$ *tale che* $\pi\circ\varphi=\psi$ *e*

$$(1_X\times\varphi)^*(\mathscr{L}_d)=\mathscr{L}\otimes q^*\mathscr{Q}$$

per qualche fibrato in rette $\mathscr{Q}$ *su* S'.

Naturalmente, se s è un punto di S

$$\pi^{-1}(s)=\operatorname{Pic}^d(p^{-1}(s)).$$

In particolare, se S è un punto, $\operatorname{Pic}^d(p)=\operatorname{Pic}^d(X)$.

§2. Le varietà W^r_d e G^r_d.

In tutto questo paragrafo indicheremo con $X\xrightarrow{p} S$ un morfismo liscio e proiettivo di varietà algebriche le cui fibre sono curve di genere g. Supporremo anche che p abbia una sezione.

Siano $\mathscr{L}_d$ un fibrato di Poincaré su $X \times_S \mathrm{Pic}^d(p)$ e

$$\mathrm{Pic}^d(p) \xrightarrow{\pi} S,$$
$$X \times_S \mathrm{Pic}^d(p) \xrightarrow{q} \mathrm{Pic}^d(p)$$

le proiezioni naturali. I teoremi fondamentali sul cambiamento di base in coomologia [8], [18], implicano che, per ogni punto x di $\mathrm{Pic}^d(p)$ esistono un intorno affine U di x e un omomorfismo di $\mathcal{O}_U$-moduli liberi

$$(2.1) \quad K^0 \xrightarrow{\alpha} K^1$$

con la seguente proprietà. Per ogni morfismo di varietà affini $V \to U$ vi sono isomorfismi funtoriali

$$H^i(q^{-1}(U) \times_U V, \mathscr{L}_{d|_{q^{-1}(U)}} \otimes_{\mathcal{O}_U} \mathcal{O}_V) = H^i(\Gamma(V, K^{\cdot} \otimes_{\mathcal{O}_U} \mathcal{O}_V))$$

In particolare se m e n sono i ranghi di K^0 e K^1, segue dal teorema di Riemann–Roch che

$$m - n = d - g + 1$$

Se si scelgono isomorfismi di $\Gamma(U, K^0)$ e $\Gamma(U, K^1)$ con $\Gamma(U, \mathcal{O}_U^m)$ e $\Gamma(U, \mathcal{O}_U^n)$, ad α è associata una matrice $n \times m$ A, con coefficienti in $\Gamma(U, \mathcal{O}_U)$. Definiamo ora $\mathrm{W}^r_{d,U}$ come la sottovarietà di U il cui ideale è il $(g-d+r)$-esimo ideale di Fitting della presentazione

$$\Gamma(U, \mathcal{O}_U)^m \to \Gamma(U, \mathcal{O}_U)^n \to H^1(q^{-1}(U), \mathscr{L}_{d|_{q^{-1}(U)}}) \to 0$$

cioè l'ideale generato dai minori di ordine $n-g+d-r+1 = m-r$ di A. La formazione degli ideali di Fitting è compatibile con il cambiamento di base; è inoltre noto [4] che due diverse presentazioni di uno stesso modulo di tipo finito su un anello noetheriano hanno gli stessi ideali di Fitting. Esiste perciò una sottovarietà $\mathrm{W}^r_d(p)$ di $\mathrm{Pic}^d(p)$ tale che

$$\mathrm{W}^r_d(p) \cap U_i = \mathrm{W}^r_{d,U_i}$$

dove $\{U_i\}$ è un opportuno ricoprimento affine di $\mathrm{Pic}^d(p)$.

Sia ora $\mathbf{G} = \mathrm{Gr}(r+1, m)$ la Grassmanniana degli $(r+1)$-piani in $\mathbf{C}^m$ e sia M la varietà delle matrici complesse $n \times m$. Su M vi è un omomorfismo naturale di fasci

$$(2.2) \quad \mathcal{O}_M^m \xrightarrow{\varphi} \mathcal{O}_M^n$$

la cui matrice è il morfismo identità di M. La matrice A associata a (2.1) può essere pensata come un morfismo di U in M. L'omomorfismo (2.1) è indotto da (2.2) tramite A. Inoltre (2.2) induce un omomorfismo

$$\mathcal{O}^m_{M\times\mathbf{G}} \xrightarrow{\tilde{\varphi}} \mathcal{O}^n_{M\times\mathbf{G}}$$

Indichiamo con F il sottofascio tautologico (localmente libero di rango $r+1$) di $\mathcal{O}^m_{\mathbf{G}}$ e poniamo $\tilde{F}=\mathcal{O}_M\times F$.

Sia V il luogo dei punti x di $M\times\mathbf{G}$ per cui $\tilde{F}\otimes k(x)$ è contenuto nel nucleo di

$$\tilde{\varphi}_x : k(x)^m \to k(x)^n$$

V è una sottovarietà liscia di $M\times\mathbf{G}$. Indichiamo con V' la sua immagine in M. E' noto [26] che l'ideale di V' è generato dai minori di ordine $m-r$. D'altro canto $W^r_d(p)\cap U$ è l'immagine inversa di V' tramite il morfismo A.

Poniamo

$$G^r_{d,U}=U\times_M V$$

$G^r_{d,U}$ è una sottovarietà di $U\times\mathbf{G}$. Riassumendo si ha che

(2.3) *I diagrammi*

$$\begin{array}{ccc} G^r_{d,U} & \longrightarrow & U \\ \downarrow & & \downarrow A \\ V & \longrightarrow & M \end{array} \qquad \begin{array}{ccc} W^r_d(p)\cap U & \longrightarrow & U \\ \downarrow & & \downarrow A \\ V' & \longrightarrow & M \end{array}$$

sono diagrammi cartesiani.

Sia ora

$$j : G^r_{d,U} \to M\times\mathbf{G}$$

il morfismo naturale e

$$g : q^{-1}(U)\times_U G^r_{d,U} \to G^r_{d,U}$$

la proiezione. Indichiamo con il simbolo $\mathcal{M}_U$ il fascio immagine inversa di $\mathcal{L}_d$ tramite il morfismo

$$q^{-1}(U)\times_U G^r_{d,U} = X\times_S G^r_{d,U} \to X\times_S \mathrm{Pic}^d\,(p)$$

Allora $\mathscr{F}_U = j^*(\tilde{F})$ è un sottofascio localmente libero di

$$g_*\mathscr{M}_U = \operatorname{Ker}(K^0 \otimes_{\mathscr{O}_U} \mathscr{O}_{G^r_{d,U}} \to K^1 \otimes_{\mathscr{O}_U} \mathscr{O}_{G^r_{d,U}})$$

Quanto sopra può essere riassunto dicendo che sul morfismo $q^{-1}(U) \to U$ vi è una famiglia naturale di g^r_d parametrizzata da $G^r_{d,U} \xrightarrow{\psi} U$, secondo la seguente:

(2.4) DEFINIZIONE. *Una famiglia di g^r_d su $X \xrightarrow{p} S$ parametrizzata da f è una terna $(f, \mathscr{L}, \mathscr{H})$ ove:*

a) $f : S' \to S$ *è un morfismo*

b) $\mathscr{L}$ *è un fibrato in rette su* $X \times_S S'$ *la cui restrizione a ogni fibra di* $X \times_S S' \xrightarrow{p'} S'$ *ha grado* d

c) $\mathscr{H}$ *è un sottofascio localmente libero di rango* $r+1$ *di* $p'_*(\mathscr{L})$ *tale che, per ogni* $s \in S'$, *l'omomorfismo*

$$\mathscr{H} \otimes k(s) \to H^0(p'^{-1}(s), \mathscr{L} \otimes k(s))$$

sia iniettivo.

Abbiamo il seguente

(2.5) TEOREMA. *Sia $X \xrightarrow{p} S$ un morfismo liscio e proiettivo di varietà algebriche le cui fibre sono curve di genere g. Supponiamo che p abbia una sezione. Sia* $\operatorname{Pic}^d(p) \xrightarrow{\pi} S$ *la proiezione naturale. Allora esistono una varietà* $G^r_d(p)$, *un morfismo proiettivo*

$$G^r_d(p) \xrightarrow{c} \operatorname{Pic}^d(p)$$

e una famiglia $(\pi \circ c, \mathscr{M}, \mathscr{F})$ di g^r_d su $X \xrightarrow{p} S$ con la seguente proprietà universale. Se $(f : S' \to S, \mathscr{L}, \mathscr{H})$ è una famiglia di g^r_d su $X \xrightarrow{p} S$, esiste un unico morfismo

$$\eta : S' \to G^r_d(p)$$

di varietà su S tale che l'immagine inversa di $(\pi \circ c, \mathscr{M}, \mathscr{F})$, tramite η, sia isomorfa a

$$(f, \mathscr{L} \otimes \mathscr{Q}, \mathscr{H} \otimes \mathscr{Q})$$

per qualche fibrato in rette $\mathscr{Q}$ su S'. Inoltre, nel caso in cui $f = \pi \circ \psi$, $S' = G^r_{d,U}$, $\mathscr{L} = \mathscr{M}_U$, $\mathscr{H} = \mathscr{F}_U$, il morfismo η è un isomorfismo di $G^r_{d,U}$ su $c^{-1}(U)$. Infine $\mathscr{M} = (1_X \times c)^* \mathscr{L}_d$, dove $\mathscr{L}_d$ è un fibrato in rette di Poincaré su $X \times_S \operatorname{Pic}^d(p)$, e c si fattorizza attraverso l'inclusione $W^r_d(p) \hookrightarrow \operatorname{Pic}^d(p)$.

Dimostrazione. Per la proprietà universale di $\mathrm{Pic}^d(p)$, si può scrivere, per un unico $\tilde{f}: S' \to \mathrm{Pic}^d(p)$,

$$f = \pi \circ \tilde{f}$$
$$(1_X \times \tilde{f})^* \mathscr{L}_d \cong \mathscr{L} \otimes \mathscr{Q}$$

Basta ora dimostrare che $\mathrm{G}^r_{d,U}$, $\mathscr{M}_U$, $\mathscr{F}_U$ hanno la proprietà universale descritta da (2.5) quando $\tilde{f}$ si fattorizza tramite l'inclusione di U in $\mathrm{Pic}^d(p)$. Sia

$$p': X \times_S S' \to S'$$

la proiezione. Il fascio $p'_* \mathscr{L}$ si identifica con il nucleo di

$$\tilde{f}^*(\alpha): \tilde{f}^* K^0 \to \tilde{f}^* K^1$$

Esiste un unico morfismo

$$h: S' \to \mathrm{Gr}(r+1, m) = \mathbf{G}$$

tale che

$$h^*(F) = \mathscr{H} \hookrightarrow \tilde{f}^* K^0 = h^*(\mathscr{O}^m_{\mathbf{G}})$$

Poichè $\mathscr{H} \subset \mathrm{Ker}\, \tilde{f}^*(\alpha)$, il morfismo

$$(\tilde{f}, h): S' \to U \times \mathbf{G}$$

è la composizione dell'inclusione

$$\mathrm{G}^r_{d,U} \hookrightarrow U \times \mathbf{G}$$

e di un morfismo

$$\eta: S' \to \mathrm{G}^r_{d,U}$$

di varietà su U tale che

$$\eta^* \mathscr{F}_U = \mathscr{H} \otimes \mathscr{Q}$$
$$(1_X \times \eta)^*(\mathscr{M}_U) = \mathscr{L} \otimes \mathscr{Q}$$

Che η sia unico segue dall'unicità di $\tilde{f}$ e h. Q.E.D.

Quando S è un punto, e X è, perciò, una curva di genere g, scriveremo $W^r_d(X)$, $G^r_d(X)$ in luogo di $W^r_d(p)$, $G^r_d(p)$. Come $\mathrm{Pic}^d(X)$ parametrizza le serie lineari complete di grado d su X, $W^r_d(X)$ e $G^r_d(X)$ parametrizzano, rispettivamente, le serie lineari complete su X di grado d e dimensione almeno r, e le serie lineari di grado d e dimensione r su X (cioè le g^r_d su X).

Sia C una curva di genere g. Ricordiamo (cf. (1.1)) che vi è un morfismo

$$u : C_d \to \mathrm{Pic}^d(C)$$

(2.6) PROPOSIZIONE (cf. [14]). $C^r_d = u^{-1}(W^r_d(C))$.

Dimostrazione. Indichiamo con p_1, p_2 le proiezioni di $C \times C_d$ sui due fattori e poniamo $\mathscr{L} = (1_C \times u)^*(\mathscr{L}_d)$. La sottovarietà $u^{-1}(W^r_d(C))$ di C_d ha come fascio di ideali il $(g-d+r)$-esimo ideale di Fitting di $R^1 p_{2*}(\mathscr{L})$. Su $C \times C_d$ vi è un divisore D tale che

$$D \cdot (C \times \{\Delta\}) = \Delta, \qquad \Delta \in C_d$$

Inoltre D è piatto su C_d. Poichè $\mathscr{L} \otimes \mathscr{O}(-D)$ è banale su ogni fibra di p_2, esiste, su C_d, un fibrato in rette $\mathscr{Q}$ tale che

$$\mathscr{O}(D) = \mathscr{L} \otimes p_2^* \mathscr{Q}$$

Perciò $u^{-1}(W^r_d(C))$ ha, come fascio di ideali, il $(g-d+r)$-esimo ideale di Fitting di $R^1 p_{2*}(\mathscr{O}(D))$. Dalla successione esatta di fasci su $C \times C_d$

$$0 \to \mathscr{O} \to \mathscr{O}(D) \to \mathscr{O}_D(D) \to 0$$

si deduce una successione esatta di fasci su C_d

$$p_{2*}\mathscr{O}_D(D) \xrightarrow{\alpha} R^1 p_{2*}\mathscr{O} \to R^1 p_{2*}\mathscr{O}(D) \to 0$$

che può essere usata per calcolare gli ideali di Fitting di $R^1 p_{2*}\mathscr{O}(D)$, poichè $p_{2*}\mathscr{O}_D(D)$ e $R^1 p_{2*}\mathscr{O}$ sono localmente liberi di ranghi d e g rispettivamente. D'altra parte, per (1.2), α si identifica all'omomorfismo

$$u_* : \Theta_{C_d} \to u^*(\Theta_{\mathrm{Pic}^d(C)}) \qquad \text{Q.E.D.}$$

§3. Il significato geometrico della congettura di Petri

In questo paragrafo esporremo alcune conseguenze della congettura di Petri riguardanti la struttura delle varietà W^r_d.

Sia C una curva di genere $g \geq 3$. Siano r e d due interi non negativi. Sia L un fibrato in rette su C di grado d, e sia

$$W \subset H^0(C, L)$$

un sottospazio di dimensione $r+1$ (una g^r_d). Indicheremo con I il punto corrispondente a una tale g^r_d in $G^r_d(C)$. Diremo che I *soddisfa la condizione di Petri* se il prodotto "cup"

$$(3.1) \quad \mu_0 : W \otimes H^0(C, K \otimes L^{-1}) \to H^0(C, K)$$

è iniettivo.

La congettura di Petri è perciò equivalente a

(3.2) CONGETTURA. *Sia C una curva di genere $g \geq 3$, a moduli generali. Allora, per ogni r e ogni d, ogni punto I di $G^r_d(C)$ soddisfa la condizione di Petri.*

Il risultato fondamentale che traduce la congettura di Petri in un enunciato geometrico è il seguente

(3.3) TEOREMA. *Sia C una curva di genere $g \geq 3$. Sia I un punto di $G^r_d(C)$. Allora I soddisfa la condizione di Petri se e solo se $G^r_d(C)$ è liscia e di dimensione*

$$\rho = g - (r+1)(g-d+r)$$

nel punto I.

Dimostrazione. Innanzitutto osserviamo che ogni componente irriducibile di $G^r_d(C)$ ha dimensione pari almeno a ρ; poichè questa affermazione è di carattere locale, possiamo lavorare su $G^r_{d,U}$ dove U è un aperto affine di $\mathrm{Pic}^d(C)$. Segue da (2.3) e dal fatto che V è una sottovarietà liscia di $M \times \mathrm{Gr}(r+1, m)$ di codimensione $n(r+1)$, che ogni componente irriducibile di $G^r_{d,U}$ ha dimensione pari almeno a ρ.

Siano ora L il fibrato in rette su C e W il sottospazio di $H^0(C, L)$ corrispondenti al punto $I \in G^r_d(C)$. Vogliamo calcolare lo spazio tangente di Zariski

$$T_I(G^r_d(C))$$

a $G^r_d(C)$ nel punto I. Per la proprietà universale di $G^r_d(C)$, $T_I(G^r_d(C))$ è in corrispondenza biunivoca con l'insieme delle classi di equivalenza di famiglie di g^r_d

su C parametrizzate da $\operatorname{Spec} \mathbf{C}[\varepsilon]$ che estendono $W \subset H^0(C, L)$. Per l'esattezza della successione (1.3), l'immagine di $T_I(G^r_d(C))$ in $T_L(\operatorname{Pic}^d(C)) = H^1(C, \mathcal{O})$ è

(3.4) $\quad H = \{\varphi \in H^1(C, \mathcal{O}) : \varphi \cdot W = 0\}$

Vi è perciò una successione esatta

(3.5) $\quad 0 \to H' \to T_I(G^r_d(C)) \to H \to 0$

dove H' è l'insieme delle classi di equivalenza di famiglie di g^r_d contenute in $|L|$ e cioè lo spazio tangente alla grassmanniana $\operatorname{Gr}(r+1, H^0(C, L))$ nel punto W. Quindi

$$H' = \operatorname{Hom}(W, H^0(C, L)/W)$$

e

$$\dim T_I(G^r_d(C)) = \dim H + (r+1)(\bar{r} - r)$$

dove si è posto $\bar{r} = \dim(H^0(C, L)) - 1$. Ora H è definito come il nucleo dell'aggiunta di (3.1):

$$H^1(C, \mathcal{O}) \to \operatorname{Hom}(W, H^1(C, L))$$

quindi

$$\dim T_I(G^r_d(C)) = (r+1)(\bar{r} - r) + \dim(\operatorname{coker} \mu_0).$$

Perciò dire che I soddisfa la condizione di Petri è equivalente a dire che

$$\begin{aligned} \dim T_I(G^r_d(C)) &= (r+1)(\bar{r} - r) + g - (r+1)(\bar{r} - d + g) \\ &= g - (r+1)(g - d + r) = \rho \end{aligned}$$

Poichè in precedenza si è osservato che ogni componente irriducibile di $G^r_d(C)$ ha dimensione pari almeno a ρ, ciò conclude la dimostrazione. Q.E.D.

Consideriamo un morfismo di varietà algebriche ridotte

(3.6) $\quad f : X \to Y$

Supponiamo che X e Y siano connesse. Ricordiamo che f si dice una *risoluzione*

razionale se X è liscia, Y è normale e di Cohen–Macaulay, f è propria e birazionale e le immagini dirette superiori $R^q f_*(\mathcal{O}_X)$, $q>0$, sono nulle. Se X e Y non sono connesse e $X_1, \ldots, X_n$ sono le componenti connesse di X, diremo che (3.6) è una risoluzione razionale se, per ogni i,

$$f|_{X_i}: X_i \to f(X_i)$$

è una risoluzione razionale e Y è unione disgiunta degli $f(X_i)$.

(3.7) TEOREMA. *Se C è una curva di genere $g \geq 3$ che soddisfa la condizione di Petri per ogni g^r_d, ove d è un intero minore o eguale a $g-1$, allora $G^r_d(C)$ è non singolare e di dimensione*

$$\rho = g-(r+1)(g-d+r).$$

Inoltre $W^r_d(C) = c(G^r_d(C))$ *e*

$$G^r_d(C) \to W^r_d(C)$$

è una risoluzione razionale.

La dimostrazione si basa su due lemmi che enunciamo senza dimostrazione.

(3.8) LEMMA (cf. [12], Lemma 2). *Siano X, Z, Z' varietà lisce e connesse e*

$$g: X \to Z, \qquad m: Z' \to Z$$

morfismi. Supponiamo che g sia proprio e che il morfismo indotto $f: X \to Y = g(X)$ sia una risoluzione razionale. Siano

$$\begin{array}{ccc} X' & \xrightarrow{g'} & Z' \\ \downarrow & & \downarrow m \\ X & \xrightarrow{g} & Z \end{array} \qquad \begin{array}{ccc} Y' & \xrightarrow{j} & Z' \\ \downarrow & & \downarrow m \\ Y & \longrightarrow & Z \end{array}$$

diagrammi cartesiani. Se X' è liscia e $\dim X' - \dim Z' = \dim X - \dim Z$, *allora l'immagine di g' è $j(Y')$ e*

$$f': X' \to Y'$$

è una risoluzione razionale.

Prima di enunciare l'altro lemma, richiamiamo alcune notazioni del §2. Indichiamo con M la varietà delle matrici $n \times m$ a coefficienti complessi, con V la sottovarietà (liscia) di $M \times \mathrm{Gr}(r+1, m)$ costituita dalle coppie (B, W) tali che $B \cdot W = 0$, e con V' la sottovarietà di M il cui ideale è generato dai minori di ordine $m-r$.

(3.9) LEMMA (cf. [13], [26]). *Se $n \geq m$, V' è l'immagine di V tramite la proiezione di $M \times \mathrm{Gr}(r+1, m)$ su M e*

$$V \to V'$$

è una risoluzione razionale.

La dimostrazione di (3.7) è ora immediata. La prima parte segue dal Teorema (3.3) e dal teorema di esistenza per i divisori speciali [11], [14], [25]. Per quanto riguarda la seconda parte, la questione è di natura locale su $\mathrm{Pic}^d(C)$, perciò basta mostrare che ogni punto di $\mathrm{Pic}^d(C)$ possiede un intorno U tale che $c(\mathrm{G}^r_{d,U}) = \mathrm{W}^r_d(C) \cap U$ e che $\mathrm{G}^r_{d,U}$ sia una risoluzione razionale di $\mathrm{W}^r_d(C) \cap U$. Si può anche supporre che $\mathrm{W}^r_d(C) \cap U$ sia connesso: allora anche $\mathrm{G}^r_{d,U}$ lo è, poichè c ha fibre connesse. Il risultato segue ora da (2.3) e dai Lemmi (3.8) e (3.9) dopo aver notato che, per il Teorema (3.3) $\mathrm{G}^r_{d,U}$ è liscia e

$$\dim U - \dim \mathrm{G}^r_{d,U} = (r+1)(r+g-d)$$

mentre

$$\begin{aligned} \dim M - \dim V &= nm - (r+1)(m-r-1) + n(m-r-1) \\ &= (r+1)(n-m+r+1) = (r+1)(r+g-d) \end{aligned}$$

poichè, nel nostro caso, $m - n = d - g + 1$. Ciò conclude la dimostrazione di (3.7).

(3.10) TEOREMA. *Sia C una curva di genere $g \geq 3$. Sia d un intero minore o uguale a $g-1$. Sia I un punto di $\mathrm{G}^r_d(C)$, e siano L il fibrato in rette su C e W il sottospazio $(r+1)$-dimensionale di $H^0(C, L)$ corrispondenti a I. Supponiamo che I soddisfi la condizione di Petri. Allora il cono tangente a $\mathrm{W}^r_d(C)$ nel punto L è la sottovarietà ridotta T di $T_L(\mathrm{Pic}^d(C)) = H^1(C, \mathcal{O})$ di supporto*

$$\mathrm{Supp}\, T = \{\varphi \in H^1(C, \mathcal{O}) : \varphi \cdot W = \{0\} \subset H^1(C, L), \text{ per qualche } W \in \mathrm{Gr}(r+1, H^0(C, L))\}$$

Inoltre se N è il fibrato normale alla fibra di

$$c : G^r_d(C) \to W^r_d(C)$$

sopra L, il morfismo $N \to T$ è una risoluzione razionale.

Dimostrazione. Poniamo $\bar{r} = \dim H^0(C, L) - 1$ e $i = \dim H^1(C, L)$. Come segue dalla esattezza della successione (3.5), c_* applica lo spazio normale a $c^{-1}(L)$ in I isomorficamente sul sottospazio H di $H^1(C, \mathcal{O})$ definito da (3.4). Perciò l'immagine di N in $H^1(C, \mathcal{O})$ è precisamente T. Indichiamo con $\mathcal{V}$ la sottovarietà liscia di

$$\mathrm{Hom}\,(H^0(C, L),\ H^1(C, L)) \times \mathrm{Gr}\,(r+1,\ H^0(C, L))$$

costituita dalle coppie (ψ, W) tali che $\psi(W) = 0$. Per il Lemma (3.9) $\mathcal{V}$ è una risoluzione razionale della sua immagine $\mathcal{V}'$ in $\mathrm{Hom}\,(H^0(C, L), H^1(C, L))$. Inoltre il supporto di $\mathcal{V}'$ è l'insieme degli omomorfismi di $H^0(C, L)$ in $H^1(C, L)$ di rango non superiore a $\bar{r} - r$. Vi è un diagramma cartesiano

$$\begin{array}{ccc} N & \longrightarrow & \mathcal{V} \\ \downarrow & & \downarrow \\ H^1(C, \mathcal{O}) & \xrightarrow{\mu_0^*} & \mathrm{Hom}\,(H^0(C, L),\ H^1(C, L)) \end{array}$$

Definiamo T' come l'immagine inversa di $\mathcal{V}'$ tramite μ_0^*. Poichè

$$\dim N - \dim H^1(C, \mathcal{O}) = \rho - g = (r+1)(d - g - r)$$

$$\begin{aligned} \dim \mathcal{V} - \dim \mathrm{Hom}\,(H^0(C, L),\ H^1(C, L)) &= (r+1)(\bar{r} - r) + i(\bar{r} - r) - (r+1)i \\ &= (r+1)(\bar{r} - r - i) = (r+1)(d - g - r) \end{aligned}$$

e N è liscia, il Lemma (3.8) dice the $T = T'$ e che N è una risoluzione razionale di T. In particolare T è normale e birazionale a N. Per concludere si usa il seguente lemma, che enunciamo senza dimostrazione

(3.11) LEMMA (cf. [12], Proposition 1). *Sia $f : X \to Y$ un morfismo proprio tra varietà liscie. Sia y un punto di Y tale che $f^{-1}(y)$ sia liscia. Indichiamo con $\mathcal{N}$ il fibrato normale a $f^{-1}(y)$ e con $\mathcal{T}$ la sua immagine in $T_y(Y)$. Se $\mathcal{N}$ è birazionale a $\mathcal{T}$ e $\mathcal{T}$ è normale, allora $\mathcal{T}$ è il cono tangente a $f(X)$ in y.*

(3.12) *Osservazione.* Sempre nelle ipotesi del Teorema (3.10), supponiamo che C non sia iperellittica e identifichiamo C con la sua immagine canonica in $\mathbf{P}^{g-1} = \mathbf{P}H^1(C, \mathcal{O})$. Se D è un divisore su C indichiamo con $\bar{D}$ il sottospazio lineare generato da D. Allora il Teorema (3.10) ci dice che il proiettivizzato $\tilde{T}$ del cono tangente a $W^r_d(C)$ in L è

$$\tilde{T} = \bigcup \left\{ \bigcap_{D \in W} \bar{D} : W \in \mathbf{P}\,\mathrm{Gr}\,(r, |L|) \right\} \subset \mathbf{P}^{g-1} = \mathbf{P},$$

dove si è indicata con $\mathbf{P}\,\mathrm{Gr}\,(r, |L|)$ la grassmaniana degli r-piani nello spazio proiettivo $|L|$. Parimenti indichiamo con $\tilde{N}$ il proiettivizzato del fibrato normale a $c^{-1}(L) = \mathrm{Gr}\,(r+1, H^0(C, L)) = \mathbf{G}$, e, come al solito, poniamo $\bar{r} = \dim H^0(C, L) - 1$, $i = \dim H^1(C, L)$. Vogliamo calcolare la classe di coomologia di $\tilde{N}$ in $\mathbf{P} \times \mathbf{G}$. Indichiamo con p e q le proiezioni di $\mathbf{P} \times \mathbf{G}$ sui due fattori. Immergiamo $\mathbf{P} \times \mathbf{G}$ in $\mathbf{P} \times \mathbf{P}'$ tramite l'immersione di Plücker di $\mathbf{G}$ in $\mathbf{P}' = \mathbf{P}(\Lambda^{r+1} H^0(C, L))$. Indichiamo con h la classe di coomologia di un iperpiano in $\mathbf{P}$ e con h' la classe di coomologia di un iperpiano in $\mathbf{P}'$. Se $\mathbf{P} \times \mathbf{P}' \to \mathbf{P}''$ è l'immersione di Segre, $\tilde{N}$ è tagliata da una sottovarietà lineare di $\mathbf{P}''$. Poichè la codimensione di $\tilde{N}$ in $\mathbf{P} \times \mathbf{G}$ è $i(r+1)$, la classe di $\tilde{N}$ è

$$\nu = (p^* h + q^* h')^{i(r+1)}$$

Perciò la classe di $\tilde{T}$ in $\mathbf{P}$ è

$$p_* \nu = d(r, \bar{r}) \binom{i(r+1)}{(\bar{r}-r)(r+1)} h^{(r+1)(g-d+r)}$$

dove

$$d(r, \bar{r}) = \frac{1!2! \cdots r!(r+1)(\bar{r}-r)!}{(\bar{r}-r)!(\bar{r}-r+1)! \cdots \bar{r}!}$$

è il grado della Grassmanniana $\mathbf{G}$ nella immersione di Plücker [9]. Concludendo

(3.13) PROPOSIZIONE. *Nelle ipotesi del Teorema* (3.10) *il proiettivizzato* $\tilde{T}$ *del cono tangente a* $W^r_d(C)$ *in* L *è una sottovarietà di* $\mathbf{P}^{g-1} = \mathbf{P}H^1(C, \mathcal{O})$ *di grado*

$$\frac{[(r+1)(g-d+\bar{r})]!1!2! \cdots r!}{[(r+1)(g-d+r)]!(\bar{r}-r)! \cdots \bar{r}!}$$

dove si è posto $\bar{r} = \dim H^0(C, L) - 1$.

§4. g^r_d su di una curva mobile

Sia C una curva di genere $g \geq 3$. Ricordiamo [20] che esistono varietà liscie e irriducibili X, S e un morfismo liscio e proiettivo

$$p : X \to S$$

tali che

1) Ogni fibra di p è una curva di genere g e una di queste è isomorfa a C.
2) Per ogni punto $t \in S$ l'omomorfismo di Kodaira–Spencer

$$T_t(S) \to H^1(p^{-1}(t), \Theta_{p^{-1}(t)})$$

è un isomorfismo (ciò si esprime dicendo che la famiglia $p : X \to S$ è *completa* e *effettivamente parametrizzata*, e implica che il morfismo naturale da S allo spazio dei moduli delle curve di genere g è *finito* e ha *immagine densa*).

3) p possiede una sezione.

Fissiamo ora, una volta per tutte, una famiglia con le proprietà 1), 2), 3). Per ogni $s \in S$ porremo

$$p^{-1}(s) = C_s, \qquad K_{C_s} = K_s, \qquad \Theta_{C_s} = \Theta_s$$

Riferendoci alle notazioni introdotte nel §2 porremo, in modo suggestivo

$$Pic^d = \mathrm{Pic}^d(p)$$
$$\mathcal{W}^r_d = \mathrm{W}^r_d(p)$$
$$\mathcal{G}^r_d = \mathrm{G}^r_d(p)$$

cosicchè si hanno morfismi

$$c : \mathcal{G}^r_d \to Pic^d, \qquad \pi : Pic^d \to S$$

Sia I un punto di Pic^d, corrispondente a una fibra $C_t = p^{-1}(t)$ e a un fibrato in rette $L \to C_t$. Le argomentazioni svolte nel §1 consentono di identificare lo spazio tangente a Pic^d in I con $H^1(C_t, \Sigma_L)$. Se poi $I \in \mathcal{W}^r_d$ e w è il punto di $\mathcal{G}^r_d$ corrispondente a un sottospazio $(r+1)$-dimensionale $W \subset H^0(C_t, L)$, considerazioni analoghe a quelle contenute nella dimostrazione del Teorema (3.3) mostrano che vi è una successione esatta

$$0 \to \mathrm{Hom}(W, H^0(C_t, L)/W) \to T_w(\mathcal{G}^r_d) \xrightarrow{c_*} T_I(Pic^d)$$

e inoltre che

$$c_*(T_w(\mathscr{G}^r_d)) = \{\sigma \in H^1(C_t, \Sigma_L) : \sigma \cdot W = 0\}$$

Ricordiamo la fondamentale applicazione

$$\mu_0 : W \otimes H^0(C_t, K_t \otimes L^{-1}) \to H^0(C_t, K_t)$$

Seguendo le idee sviluppate in [1], consideriamo l'applicazione lineare

$$\mu : W \otimes H^0(C_t, K_t \otimes L^{-1}) \to H^0(C_t, K_t \otimes \Sigma_L^*)$$

definita per dualità a partire da (1.4), e definiamo una applicazione lineare

$$\mu_1 : \operatorname{Ker} \mu_0 \to H^0(C_t, K_t^2)$$

a mezzo del diagramma commutativo

$$\begin{array}{ccc} & & 0 \\ & & \uparrow \\ W \otimes H^0(C_t, K_t \otimes L^{-1}) & \xrightarrow{\mu_0} & H^0(C_t, K_t) \\ \| & & \uparrow \\ W \otimes H^0(C_t, K_t \otimes L^{-1}) & \xrightarrow{\mu} & H^0(C_t, K_t \otimes \Sigma_L^*) \\ \uparrow & & \uparrow \\ \operatorname{Ker} \mu_0 & \xrightarrow{\mu_1} & H^0(C_t, K_t^2) \\ & & \uparrow \\ & & 0 \end{array} \tag{4.1}$$

L'osservazione centrale che mette in relazione l'omomorfismo μ_1 con l'applicazione

$$\pi \circ c : \mathscr{G}^r_d \to S$$

costruita all'inizio di questo paragrafo, sfrutta, in modo essenziale, la proprietà di completezza 2) soddisfatta, per ipotesi, dalla famiglia $X \to S$ e si lascia enunciare dal seguente

(4.2) LEMMA (cf. [1]). *Supponiamo che il morfismo $\pi \circ c : \mathscr{G}^r_d \to S$ sia suriettivo. Sia t un punto generale di S. Per ogni L e ogni W si ha che*

$$\mu_1 = 0$$

Dimostrazione. Essendo t generale si può assumere che per ogni $w \in (\pi \circ c)^{-1}(t)$

$$T_w(\mathscr{G}_d^r) \xrightarrow{(\pi \circ c)_*} T_t(S) \cong H^1(C_t, \Theta_t)$$

sia suriettivo. Da (4.1), per dualità, si ottiene un diagramma commutativo

$$\begin{array}{ccc} H^1(C_t, \Sigma_L) & \xrightarrow{\mu^*} & \mathrm{Hom}\,(W, H^1(C_t, L)) \\ \downarrow{\scriptstyle \pi_*} & & \downarrow \\ H^1(C_t, \Theta_t) & \xrightarrow{\mu_1^*} & (\mathrm{Ker}\ \mu_0)^* \end{array}$$

D'altra parte, l'ipotesi fatta su t significa che

$$\begin{array}{c} \mathrm{Ker}\ \mu^* = \{\sigma \in H^1(C_t, \Sigma_L) : \sigma \cdot W = 0\} = c_* T_w(\mathscr{G}_d^r) \\ \downarrow{\scriptstyle \pi_*} \\ H^1(C_t, \Theta_t) \end{array}$$

è suriettiva. Q.E.D.

§5. Il fascio normale

Incominciamo col richiamare alcuni risultati elementari della teoria di Horikawa [10].

Dato un morfismo non banale

$$\varphi : C \to M$$

dove C è una curva e M una varietà liscia, il *fascio normale* $N = N_\varphi$ *al morfismo* φ è, per definizione, il conucleo dell'omomorfismo iniettivo di fasci

$$\varphi_* : \Theta_C \to \varphi^*(\Theta_M)$$

Si ha quindi una successione esatta

$$(5.1) \quad 0 \to \Theta_C \to \varphi^* \Theta_M \to N_\varphi \to 0$$

La teoria di Horikawa mette in corrispondenza biunivoca lo spazio vettoriale

$H^0(C, N)$ con l'insieme delle classi di equivalenza di deformazioni infinitesime di φ.

Una *deformazione infinitesima* di φ, lo ricordiamo, è il dato di una deformazione infinitesima di C,

$$p : \mathscr{C} \to \operatorname{Spec} \mathbf{C}[\varepsilon]$$

e di un morfismo

$$\tilde{\varphi} : \mathscr{C} \to M$$

tale che $\tilde{\varphi}|_C = \varphi$. La nozione di equivalenza tra deformazioni infinitesime di φ è quella ovvia.

Supponiamo ora che il genere di C sia almeno 3, che $M = \mathbf{P}^r$, $r \geq 1$, e che il morfismo φ sia associato a un sottospazio $(r+1)$-dimensionale $W \subset H^0(C, L)$, per un fibrato in rette L su C di grado d, di modo che la g^r_d definita da W non ha punti base. In questo caso la successione del fibrato normale (5.1) e la successione (1.5) sono legate dalla ben nota successione di Eulero, e, insieme, queste successioni formano il seguente diagramma commutativo:

(5.2)
$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & & & \\
 & & \downarrow & & \downarrow & & (\Theta = \varphi^*(\Theta_{\mathbf{P}^r})) & & \\
 & & \mathscr{O}_C & = & \mathscr{O}_C & & & & \\
 & & \downarrow & & \downarrow & & & & \\
0 & \longrightarrow & \Sigma_L & \xrightarrow{\lambda} & L^{\oplus(r+1)} & \longrightarrow & N & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \| & & \\
0 & \longrightarrow & \Theta_C & \xrightarrow{\varphi_*} & \Theta & \longrightarrow & N & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & & & \\
 & & 0 & & 0 & & & &
\end{array}$$

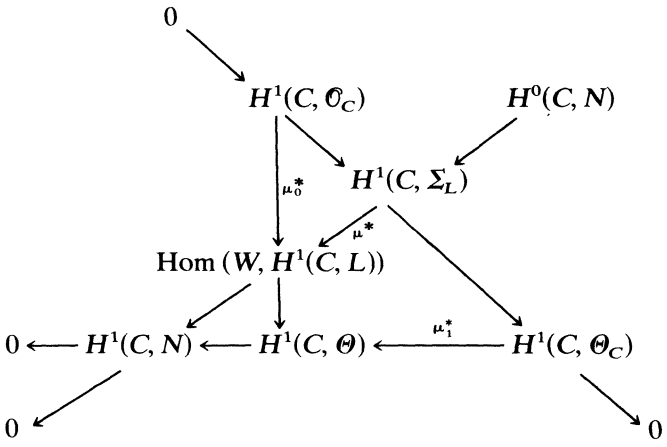
L'unico morfismo che va definito è λ. Per fare ciò notiamo che intrinsecamente, $\mathbf{P}^r = \mathbf{P}W^*$ e $L^{\oplus(r+1)} = L \otimes_{\mathbf{C}} W^*$. Sia ora $x_0, \ldots, x_r$ una base di E e $\xi_0, \ldots, \xi_r$ la base duale di W^*. Una sezione locale di Σ_L è un operatore differenziale ∇ di ordine ≤ 1 operante sulle sezioni di L. Porremo

$$\lambda(\nabla) = \Sigma \nabla x_i \otimes \xi_i$$

Il lettore non avrà alcuna difficoltà nel dimostrare la commutatività del dia-

gramma (5.2). Da (5.2) si ottiene un diagramma commutativo di coomologia

(5.3)



dove si è fatta l'identificazione

$$\text{Hom}\,(W, H^1(C, L)) = H^1(C, L \otimes_{\mathbf{C}} W^*) = H^1(C, L^{\oplus(r+1)})$$

Una prima immediata conseguenza del diagramma (5.3) è

(5.4) PROPOSIZIONE. *Il nucleo di* μ_0 *si identifica allo spazio duale di* $H^1(C, \Theta)$. *I nuclei di* μ *e* μ_1 *si identificano allo spazio duale di* $H^1(C, N)$.

Vogliamo sottolineare che l'interpretazione dei nuclei di μ_0, μ, μ_1 data da (5.4) è valida solo solo per g^r_d senza punti base. A questo punto è anche utile osservare che la congettura di Petri così come formulata in (3.2) si riduce in modo naturale alla corrispondente asserzione per le serie prive di punti base.

Tenuto conto del Lemma (4.2), ecco dunque una nuova e assai suggestiva formulazione della congettura di Petri.

(5.5) CONGETTURA. *Sia C una curva a moduli generali di genere* $g \geq 3$. *Sia*

$$\varphi : C \to \mathbf{P}^r \quad (r \geq 1)$$

un morfismo non degenere. Allora

$$H^1(C, N_\varphi) = 0$$

Come applicazione immediata delle considerazioni fin qui svolte dimostriamo la congettura di Petri per le g^1_d. E'questo un risultato da alcuni considerato noto classicamente e (ri)dimostrato da numerosi autori moderni [3], [16], [22], [24], [27].

Lax [16] lo enuncia esplicitamente nella forma equivalente data dal Teorema (3.3), per $r = 1$. Abbiamo dunque il seguente

(5.6) SCOLIO. *Sia C una curva a moduli generali di genere $g \geq 3$. Allora C soddisfa la condizione di Petri per le g^1_d; in altri termini, per ogni g^1_d su C si ha*

$$\operatorname{Ker} \mu_0 = \{0\}$$

Dimostrazione. Basta considerare il caso di una g^1_d senza punti fissi, che dà luogo a un morfismo

$$\varphi : C \to \mathbf{P}^1$$

Poichè N_φ è concentrato sul divisore di ramificazione di φ si ha $H^1(C, N_\varphi) = 0$. Q.E.D.

Un'altra versione di (5.6) anch'essa classicamente nota [24] è:

(5.7) COROLLARIO. *Sia C una curva a moduli generali di genere $g \geq 3$ e D un divisore su C tale che* $\dim |D| \geq 1$. *Allora*

$$H^0(C, K(-2D)) = 0$$

Dimostrazione. Siano s_1, s_2 sezioni linearmente indipendenti di $H^0(C, \mathcal{O}(D))$ e sia t una sezione di $H^0(C, K(-2D))$. Allora

$$s_1 \otimes ts_2 - s_2 \otimes ts_1$$

appartiene al nucleo di μ_0 ed è nulla se e solo se t lo è. Q.E.D.

Concludiamo queste considerazioni con la seguente applicazione.

(5.8) PROPOSIZIONE. *Sia C una curva a moduli generali di genere $g \geq 3$. Sia*

$$\varphi : C \to \mathbf{P}^r \quad (r \geq 2)$$

un morfismo non degenere definito da una g^r_d speciale. Allora φ stabilisce una birazionalità tra C e la sua immagine $\Gamma = \varphi(C)$.

Dimostrazione. Poichè C è a moduli generali, per un risultato classico (della cui dimostrazione il lettore troverà una moderna versione in [1]) si ha che, o Γ è birazionale a C, oppure il genere di Γ è zero. Questa eventualità non può presentarsi. Se così fosse, detto $n = 2$ il grado della curva Γ, il divisore Δ, immagine inversa su C di un punto di Γ, avrebbe la proprietà che

$$\dim H^0(C, \mathcal{O}(\Delta)) \geq 2$$
$$H^0(C, K(-n\Delta)) \neq 0$$

in contraddizione con il Corollario (5.7). Q.E.D.

La conclusione della Proposizione (5.8) è chiaramente falsa se non si assume che φ sia definita da una g^r_d speciale. Come curiosità il lettore noterà che un'altra conseguenza di (5.6) è che la conclusione della Proposizione (5.8) continua a valere ove si assuma che la g^r_d che definisce φ sia completa (non necessariamente speciale), a meno che non sia $r = 2$, $d = g + 2$ e g pari.

§6. Un fenomeno assai curioso

In questo paragrafo ci porremo nella categoria degli spazi analitici complessi. I risultati che otterremo saranno applicabili alla nostra situazione algebrica per i noti teoremi di paragone di J. P. Serre [23].

Consideriamo un morfismo analitico non banale

$$\varphi : C \to M$$

dove C è una curva di genere $g \geq 1$ e M una varietà analitica liscia. Come nella situazione algebrica descritta nel §5, vi è un fascio normale a φ, N_φ, le cui sezioni sono in corrispondenza biunivoca con le classi di equivalenza di deformazioni infinitesime di φ.

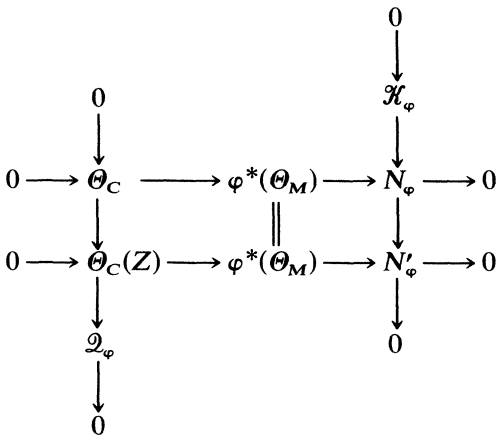
Sia Z il *divisore di ramificazione di* φ, cioè il divisore degli zeri del differenziale di φ. Si ha allora un omomorfismo iniettivo

$$\varphi_* : \Theta_C(Z) \to \varphi^*(\Theta_M)$$

il cui conucleo è un fascio localmente libero che denoteremo con $N' = N'_\varphi$. Si ha

dunque un diagramma commutativo di successioni esatte

(6.1)



In cui $\mathcal{K} = \mathcal{K}_\varphi$ e $\mathcal{Q} = \mathcal{Q}_\varphi$ sono (non-canonicamente) isomorfi al fascio strutturale $\mathcal{O}_Z$. Una prima immediata osservazione è che

$$H^1(C, N) = H^1(C, N')$$

Le considerazioni che ora svolgeremo hanno lo scopo di dare una interpretazione geometrica degli elementi di $H^0(C, \mathcal{K}_\varphi)$. Per fare ciò è innanzitutto necessario rendere esplicita l'interpretazione, in termini di deformazioni infinitesime di φ, degli elementi di $H^0(C, N)$.

Scegliamo un ricoprimento $\{U_\alpha\}$ di C costituito da dischi coordinati su cui è stato scelto un parametro locale z_α. Si può assumere che, per ogni α, $\varphi(U_\alpha) \subset V_\alpha$, dove V_α è un aperto coordinato di M, con parametri locali $w_\alpha = (w_\alpha^1, \ldots, w_\alpha^n)$. Poniamo

$$z_\alpha = f_{\alpha\beta}(z_\beta) \quad \text{in} \quad U_\alpha \cap U_\beta$$
$$w_\alpha = g_{\alpha\beta}(w_\beta) \quad \text{in} \quad V_\alpha \cap V_\beta$$

e sia

$$w_\alpha = \psi_\alpha(z_\alpha)$$

l'espressione di φ in queste coordinate. Naturalmente

$$g_{\alpha\beta}(\psi_\beta(z_\beta)) = \psi_\alpha(f_{\alpha\beta}(z_\beta)) \tag{6.2}$$

Sia ora

$$(6.3)\qquad \begin{array}{ccc} \mathscr{C} & \xrightarrow{\tilde{\varphi}} & M \\ {\scriptstyle p}\downarrow & & \\ \operatorname{Spec} \mathbf{C}[\varepsilon] & & \end{array}$$

una deformazione infinitesima di φ. Questa deformazione è definita dai seguenti dati:

a) Funzioni di transizione per $\mathscr{C}$:

$$z_\alpha = \tilde{f}_{\alpha\beta}(z_\beta, \varepsilon) = f_{\alpha\beta}(z_\beta) + \varepsilon b_{\alpha\beta}(z_\beta)$$

b) Espressioni locali di $\tilde{\varphi}$

$$w_\alpha = \tilde{\psi}_\alpha(z_\alpha, \varepsilon) = \psi_\alpha(z_\alpha) + \varepsilon a_\alpha(z_\alpha)$$

Il cociclo $\{b_{\alpha\beta}(\partial/\partial z_\alpha)\}$ rappresenta la classe di Kodaira–Spencer di $p : \mathscr{C} \to \operatorname{Spec} \mathbf{C}[\varepsilon]$. Si ha inoltre una condizione di compatibilità analoga a (6.2). Questa condizione è equivalente a (6.2) insieme con la condizione

$$(6.4)\qquad \sum_j \frac{\partial g_{\alpha\beta}}{\partial w_\beta^j} a_\beta^j = a_\alpha + \frac{\partial \psi_\alpha}{\partial z_\alpha} b_{\alpha\beta}$$

Gli a_α definiscono un elemento di $H^0(C, N)$ che chiameremo la *classe di Horikawa di* (6.3). Naturalmente l'omomorfismo cobordo

$$H^0(C, N) \xrightarrow{\delta} H^1(C, \Theta_C)$$

applica la classe di Horikawa di (6.3) sulla classe di Kodaira-Spencer di $p : \mathscr{C} \to \operatorname{Spec} \mathbf{C}[\varepsilon]$.

Scriviamo $Z = \sum_{i=1}^{s} \nu_i p_i$, dove i p_i sono punti distinti. Possiamo supporre che ciascuno dei punti p_i sia contenuto in un unico aperto U_{α_i} e che $z_{\alpha_i}(p_i) = 0$. Ogni elemento $\{a_\alpha\}$ di $H^0(C, \mathscr{K})$ è della forma

$$(6.5)\qquad \begin{cases} a_{\alpha_i} = c_i(z_{\alpha_i}) \dfrac{\partial \psi_{\alpha_i}}{\partial z_{\alpha_i}} z_{\alpha_i}^{-\nu_i}, & i = 1, \ldots, s \\ a_\beta = 0, \quad \text{se} \quad \beta \neq \alpha_i, & i = 1, \ldots, s, \end{cases}$$

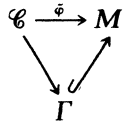
dove c_i è un polinomio di grado al più $\nu_i - 1$. Naturalmente la corrispondente

classe di Kodaira–Spencer è data da

$$(6.6)\qquad \begin{aligned} b_{\alpha_i\beta} &= -c_i(z_{\alpha_i})z_{\alpha_i}^{-\nu_i} \\ b_{\alpha\beta} &= 0 \quad \text{se} \quad \alpha \neq \alpha_i \neq \beta \quad i = 1, \ldots, s \end{aligned}$$

Nel caso in cui $c_i(z) = z^{\nu_i - 1}$ e $c_j = 0$ per ogni $j \neq i$, la classe di Kodaira–Spencer definita da (6.6) non è altro che la *variazione di Schiffer* [21] associata al punto p_i, e come tale è un elemento non nullo di $H^1(C, \Theta_C)$.

Le formule (6.5) dicono che il morfismo $\tilde{\varphi}$ associato a una classe $\{a_\alpha\} \in H^0(C, \mathcal{K})$ si fattorizza attraverso l'inclusione di $\Gamma = \varphi(C)$ in M:



Il fenomeno ora descritto permette quindi di identificare gli elementi di $H^0(C, \mathcal{K})$ con le deformazioni infinitesime del morfismo φ che lasciano fissa (al prim'ordine) la curva immagine Γ. Dunque nel caso in cui φ sia una birazionalità tra C e Γ, *la presenza di "cuspidi" su Γ, comporta l'esistenza, dal punto di vista infinitesimo, di più di un modello liscio della curva Γ*, se così ci possiamo esprimere.

Veniamo ora a un semplice lemma di natura locale che ci permetterà di usare costruttivamente questo fenomeno, a prima vista paradossale.

Diamo innanzitutto la seguente definizione. Sia Δ il disco unitario nello spazio di una coordinata complessa z e sia

$$\psi : \Delta \to \mathbf{C}^n, \qquad n > 1$$

una applicazione iniettiva tale che $\psi(0) = \psi'(0) = 0$. Con una opportuna scelta di coordinate in $\mathbf{C}^n$ si può assumere che

$$\psi(z) = (\psi_1, \psi_2, \ldots)$$

dove ψ_i si annulla di ordine $k_i > 1$ per $z = 0$ e

$$(6.7)\qquad \begin{cases} k_1 < k_2 < \cdots \\ k_1 \nmid k_2 \end{cases}$$

E'immediato verficare che gli interi k_1, k_2 così definiti sono invarianti di ψ.

L'intero k_1 è *l'indice di ramificazione di* ψ *in* 0 e k_2 verrà chiamato il *tipo* del punto di ramificazione in questione.

(6.8) LEMMA. *Sia*

$$\psi : \{(z, t) \in \mathbf{C}^2 : |z| < 1, |t| < 1\} = \Delta^2 \to \mathbf{C}^n, \qquad n > 1$$

un morfismo analitico. Supponiamo che, comunque fissato t, $\psi_t(z) = \psi(z, t)$ *sia iniettiva, abbia un unico punto di ramificazione, e che l'indice e il tipo di questo punto siano indipendenti da* t. *Supponiamo anche che esista una funzione meromorfa* $f(z)$ *tale che*

$$(6.9) \qquad \frac{\partial \psi}{\partial t}(z, 0) = f(z) \frac{\partial \psi}{\partial z}(z, 0)$$

Allora $f(z)$ *è olomorfa.*

Dimostrazione. E'semplice verificare che la validità (o meno) dell'ipotesi (6.9) e della tesi non è inficiata ove si effettuino, in $\mathbf{C}^n$, un arbitrario cambiamento di coordinate e, in Δ^2, un cambiamento di coordinate del tipo $z' = z'(z, t)$, $t' = t$. Sia $V \subset \Delta^2$ il luogo dei punti di ramificazione degli ψ_t. Poichè, per ipotesi, V si proietta bijettivamente su $\{t \in \mathbf{C} : |t| < 1\}$, V è una sottovarietà liscia di Δ^2 e si può supporre che la sua equazione sia $z = 0$. L'indice di ramificazione di ψ_t non dipende da t e lo indichiamo con h; quindi, a meno di un cambiamento di coordinate in $\mathbf{C}^n$ e della moltiplicazione di z per una funzione mai nulla si può supporre che

$$\psi = (\psi_1, \ldots, \psi_n)$$

sia della forma

$$(6.10) \qquad \begin{aligned} \psi_1(z, t) &= \alpha(t) + z^h \\ \psi_i(z, t) &= P_i(z^h, t) + \gamma_i(t) z^k + [k]; \qquad i = 2, \ldots, n \end{aligned}$$

dove:

a) $k > h$
b) h non divide k
c) $P_i(\zeta, t)$ è un polinomio in ζ
d) le funzioni γ_i non sono tutte identicamente nulle
e) il simbolo $[k]$ denota una somma di termini di ordine superiore a k in z.

L'ipotesi che il tipo del punto di ramificazione di ψ_t non dipenda da t significa che una tra le funzioni γ_i, diciamo γ_2, non si annulla per $t=0$. La (6.9) e la (6.10) implicano intanto che

$$\frac{\partial \alpha}{\partial t}(0) = hf(z)z^{h-1}$$

Perciò $f(z) = cz^{1-h}$, dove c è una costante. Usando ancora (6.9) e (6.10) si ottiene che

$$\frac{\partial P_2}{\partial t}(z^h, 0) + \frac{\partial \gamma_2}{\partial t}(0)z^k + [k] = c\left(h\frac{\partial P_2}{\partial \zeta}(z^h, 0) + k\gamma_2(0)z^{k-h} + [k-h]\right).$$

Poichè $\gamma_2(0) \neq 0$ deve essere $c=0$. Dunque, in questo particolare sistema di coordinate, f è la funzione olomorfa identicamente nulla. Q.E.D.

(6.11) COROLLARIO. *Sia* $\mathscr{C} \xrightarrow{p} \Delta = \{t \in \mathbf{C} : |t| < 1\}$ *una famiglia analitica di curve di genere* g, *e*

$$\varphi : \mathscr{C} \to M$$

un morfismo analitico di $\mathscr{C}$ *in una varietà analitica liscia* M. *Poniamo* $C_t = p^{-1}(t)$, $\varphi_t = \varphi|_{C_t}$. *Supponiamo che*

a) φ_t *stabilisca una birazionalità tra* C_t *e* $\varphi(C_t)$.

b) *Il numero, l'indice e il tipo dei punti di ramificazione di* φ_t *non dipendano da* t.

Allora la classe di Horikava di $(\mathscr{C}, p, \varphi)$, *per* $t=0$, *non appartiene a* $H^0(C_0, \mathscr{H}_{\varphi_0})$, *a meno che non sia nulla.*

Dimostrazione. Sia V il luogo dei punti di ramificazione degli φ_t. L'ipotesi b) implica che V è un rivestimento non diramato di Δ. Si può quindi scegliere un ricoprimento $\{U_\alpha\}$ di $\mathscr{C}$ e una fuzione z_α in U_α di modo che $V \cap U_\alpha = \{z_\alpha = 0\}$ e che (z_α, t) siano coordinate locali in U_α. Possiamo anche supporre che $\varphi(U_\alpha)$ sia contenuto in un aperto coordinato V_α di M. Siano $w_\alpha = (w_\alpha^1, \ldots, w_\alpha^n)$ coordinate locali in V_α. In queste coordinate φ è data da

$$w_\alpha = \psi_\alpha(z_\alpha, t).$$

La classe di Horikawa di $(\mathscr{C}, p, \varphi)$ per $t = 0$ è la sezione s di N_{φ_0} rappresentata da

$$\left\{\frac{\partial \psi_\alpha}{\partial t}(z, 0)\right\}$$

Ricordando (6.5), dire che questa classe appartiene a $H^0(C_0, \mathscr{K}_{\varphi_0})$ significa che vi sono funzioni meromorfe f_α tali che

$$\frac{\partial \psi_\alpha}{\partial t}(z_\alpha, 0) = f_\alpha(z_\alpha)\frac{\partial \psi_\alpha}{\partial z_\alpha}(z_\alpha, 0).$$

Per (6.8) le f_α sono olomorfe e quindi s è la sezione nulla di N_{φ_0}. Q.E.D.

In definitiva il significato del Lemma (6.8) e del Corollario (6.11) è il seguente. *Supponiamo di avere una sezione non nulla di $H^0(C, \mathscr{K})$ e di poter prolungare la deformazione infinitesima di $C \xrightarrow{\varphi} M$, ad essa corrispondente, in una deformazione effettiva. Allora "lungo questa deformazione la complessità della ramificazione di φ diminuisce."*

§7. La congettura di Petri per le g^2_d

Il nostro scopo principale, in questo paragrafo, è quello di dimostrare la congettura di Petri per le g^2_d. Otterremo anche risultati analoghi, ma più deboli, per le g^3_d.

Il teorema principale è il seguente.

(7.1) TEOREMA. *Sia C una curva a moduli generali di genere $g \geq 3$. Allora*

a) *C soddisfa la condizione di Petri per le g^2_d: in altri termini, per ogni intero d, per ogni fibrato in rette L di grado d su C e ogni sottospazio tri-dimensionale $W \subset H^0(C, L)$, il prodotto "cup"*

$$\mu_0 : W \otimes H^0(C, K_C \otimes L^{-1}) \to H^0(C, K_C)$$

è iniettivo.

Equivalentemente

b) *Per ogni morfismo non degenere*

$$\varphi : C \to \mathbf{P}^2$$

si ha che

$$H^1(C, N_\varphi) = 0.$$

Prima di esporre la dimostrazione di (7.1) introduciamo alcune notazioni che ci saranno utili nel seguito. Sia

$$X \xrightarrow{p} S$$

la famiglia di curve costruita nel §4. Useremo le stesse notaziono adottate in quel paragrafo. Inoltre denoteremo con il simbolo $\mathscr{G}_d^{r,k}$ il luogo dei punti di $\mathscr{G}_d^r$ corrispondenti a g_d^r per cui Ker μ_0 ha dimensione pari almeno a k. Naturalmente $\mathscr{G}_d^{r,k}$ è una sottovarietà chiusa di $\mathscr{G}_d^r$.

Dimostrazione del Teorema (7.1). Notiamo innanzitutto che a) e b) sono equivalenti per (4.2) e per (5.4). Indichiamo con il simbolo π' il morfismo $\pi \circ c$ di $\mathscr{G}_d^r$ in S. Dobbiamo mostrare che, per ogni d, $\pi'(\mathscr{G}_d^{2,1})$ è una sottovarietà propria di S. Supponiamo che ciò non sia vero, e sia d il minimo intero per cui $\pi'(\mathscr{G}_d^{2,1}) = S$. Allora un punto generale di $\mathscr{G}_d^{2,1}$ corrisponde a una g_d^2 speciale senza punti base. Vi sono perciò una varietà liscia B, un morfismo aperto

$$g : B \to S$$

e un morfismo

$$\varphi : X' = X \times_S B \to \mathbf{P}^2$$

tali che, per ogni $b \in B$,

(7.2) $\quad H^1(C_{g(b)}, N_{\varphi_b}) \neq 0$

dove si è indicata con φ_b la restrizione di φ alla fibra $C_{g(b)}$ di $X' \to B$ sopra b. Vogliamo mostrare che ciò è assurdo. Innanzitutto, se b è un punto di B, si ha un diagramma commutativo

$$\begin{array}{ccc} & H^0(C_{g(b)}, N_{\varphi_b}) & \\ {\scriptstyle h}\nearrow & & \searrow{\scriptstyle \delta} \\ g_*: T_b(B) & \to & T_{g(b)}(S) = H^1(C_{g(b)}, \Theta_{g(b)}) \end{array}$$

dove h associa ad ogni elemento di $T_b(B)$ la classe di Horikawa corrispondente.

Se poi b è un punto generale di B, g_* è suriettivo. Inoltre, sempre se b è un punto generale di B, φ_b induce, per la Proposizione (5.8), una birazionalità tra $C_{g(b)}$ e $\varphi_b(C_{g(b)})$ ed il numero, l'indice e il tipo dei punti di ramificazione di φ_t sono costanti al variare di t in un opportuno intorno di b. Perciò per il Corollario (6.11)

$$h(T_b(B)) \cap H^0(C_{g(b)}, \mathcal{H}_{\varphi_b}) = 0$$

e quindi

$$3g-3 \leq \dim H^0(C_{g(b)}, N'_{\varphi_b}).$$

Poichè, in questo caso, N'_{φ_b} è un fibrato in rette, deve essere

$$H^1(C_{g(b)}, N'_{\varphi_b}) = 0$$

e quindi

$$H^1(C_{g(b)}, N_{\varphi_b}) = 0$$

in contraddizione con (7.2). Q.E.D.

Una prima immediata conseguenza di (7.1) è che, se C è una curva a moduli generali di genere $g \geq 3$, per ogni g^2_d su C vale la diseguaglianza

(7.3) $\quad d \geq \frac{2}{3}g + 3$

La dimostrazione del Teorema (7.1) mostra anche che, se C è una curva di genere $g \geq 3$ a moduli generali e

$$\varphi : C \to \mathbf{P}^2$$

un qualsiasi morfismo che induca una birazionalità tra C e $\varphi(C)$, allora

$$\text{grado}\,(N'_\varphi) \geq 4g-4.$$

D'altra parte il grado di N'_φ può essere calcolato usando i diagrammi (5.2) e (6.1), e risulta essere pari a

$$\text{grado}\,(N'_\varphi) = 2g-2-\text{grado}\,(Z)+3d$$

dove Z è il divisore degli zeri di φ_*. Perciò

(7.4) $\text{grado}\,(Z) \leq 3d - 2g + 2$

La diseguaglianza (7.4) implica il seguente risultato, meno generale, ma certo più suggestivo.

"*Sia* $\Gamma \subset \mathbf{P}^2$ *una curva algebrica irriducible ma eventualmente singolare di genere* $g \geq 3$ *e grado d. Se* Γ *ha moduli generali allora* Γ *non ha più di* $3d - 2g + 2$ *cuspidi.*"

La limitazione (7.4) non è la migliore possibile. Ci riserviamo di ritornare sulla questione in un successivo lavoro.

Terminiamo questo paragrafo applicando le considerazioni precedenti al caso delle g^3_d. Ne trarremo la seguente conclusione

(7.5) PROPOSIZIONE. *Sia C una curva a moduli generali di genere* $g \geq 3$. *Allora*

a) *Per ogni intero d, per ogni fibrato in rette L di grado d su C e ogni sottospazio quadridimensionale* $W \subset H^0(C, L)$, *il nucleo del prodotto "cup"*

$$\mu_0 : W \otimes H^0(C, K_C \otimes L^{-1}) \to H^0(C, K_C)$$

ha dimensione al più uno.

Equivalentemente

b) *Per ogni morfismo non degenere*

$$\varphi : C \to \mathbf{P}^3$$

si ha che

$$\dim H^0(C, N_\varphi) \leq 1$$

Dimostrazione. Basta considerare il caso in cui la g^3_d corrispondente a φ è speciale e sappiamo che, in questa situazione, $\varphi(C)$ è birazionale a C. La dimostrazione si suddivide in tre passi successivi.

Primo passo. Mostriamo che due qualsiasi elementi del nucleo di μ_0 sono, punto per punto, proporzionali. Ricordiamo che Ker μ_0 è isomorfo a

$$H^0(C, N'^*_\varphi \otimes K_C).$$

Il fascio $N'_\varphi = N'$ è localmente libero di rango 2, e dai diagrammi (5.2) e (6.1) si ottiene

$$\text{grado}\,(\Lambda^2(N'^* \otimes K_C)) = 2g-2-4d+\text{grado}\,(Z)$$

dove Z è il divisore degli zeri di φ_*. Sia π la proiezione da un punto esterno a $\varphi(C)$. Applicando (7.4) a $\pi \circ \varphi$ si può dare una stima sul grado di Z ottenendo

$$\text{grado}\,(\Lambda^2(N'^* \otimes K_C)) \leq -d$$

Perciò due qualsiasi sezioni di $N'^* \otimes K_C$ sono, punto per punto, proporzionali.

Secondo passo. Mostriamo ora che il nucleo di μ_0 ha dimensione al più uguale a due. Sia $s_1, \ldots, s_4$ una base di W e supponiamo che

$$\sum_{i=1}^{4} s_i \otimes r_i, \qquad r_i \in H^0(C, K_C \otimes L^{-1}).$$

sia un elemento non nullo del nucleo di μ_0. Il primo passo della dimostrazione mostra che ogni altro elemento del nucleo di μ_0 è della forma

$$\sum s_i \otimes f r_i$$

dove f è una opportuna funzione meromorfa. Ovviamente una tale f è una sezione di $\mathcal{O}(\Delta)$, dove Δ è il divisore degli zeri comuni a $r_1, \ldots, r_4$. Si ha dunque

(7.6) $\quad \text{Ker}\,\mu_0 \cong H^0(C, \mathcal{O}(\Delta))$.

D'altro canto è immediato verificare che la congettura di Petri per le g^2_d (Teorema (7.1)) sarebbe violata se $r_1, \ldots, r_4$ non fossero linearmente indipendenti. Dunque

$$\dim H^0(C, K_C \otimes L^{-1}(-\Delta)) \geq 4$$

Poniamo $\delta = \text{grado}\,\Delta$ e supponiamo che $\dim H^0(C, \mathcal{O}(\Delta)) > 2$. Applicando la diseguaglianza (7.3) alle serie $|K_C \otimes L^{-1}(-\Delta)|$, $|L|$, $|\Delta|$ si ottengono le diseguaglianze

$$2g-2-d-\delta > \tfrac{2}{3}g$$
$$d > \tfrac{2}{3}g$$
$$\delta > \tfrac{2}{3}g$$

che sono ovviamente incompatibili. Dunque la dimensione di $H^0(C, \mathcal{O}(\Delta))$ non può essere superiore a due.

Conclusione. Applichiamo i risultati ottenuti nel secondo passo ai fibrati in rette L e $K_C \otimes L^{-1}(-\Delta)$. Si ottiene

$$(7.7) \quad \begin{cases} 10-4d+3g \leq 0 \\ 10-4d'+3g \leq 0 \end{cases}$$

dove si è posto $d' = 2g-2-d-\delta$. Se $H^0(C, \mathcal{O}(\Delta))$ avesse dimensione due si potrebbe dedurre da (5.6) che

$$8-4\delta+2g \leq 0$$

Questa diseguaglianza è chiaramente incompatibile con (7.7). Q.E.D.

§8. Un calcolo di moduli

Un problema che sorge spontaneo e che è stato oggetto di attenzione da parte di numerosi autori è quello di calcolare il numero dei parametri da cui dipende una curva di genere g che possiede una g^r_d tale che

$$\rho = g-(r+1)(g-d+r) < 0$$

o, più in generale, tale che il nucleo di μ_0 non sia nullo.

Il problema si esprime in modo naturale usando la terminologia introdotta nei paragrafi 4 e 7.

Consideriamo dunque la famiglia $X \xrightarrow{p} S$ introdotta nel §4 e, con essa, le varietà $\mathcal{G}^{r,k}_d$ introdotte nel §7, e le relative proiezioni

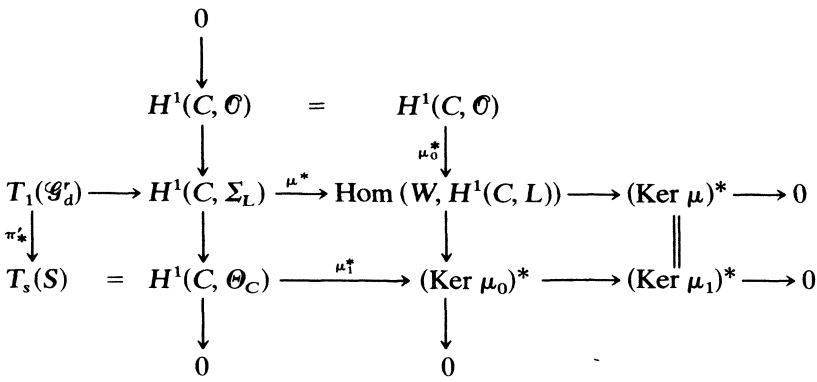
$$\pi' : \mathcal{G}^{r,k}_d \to S$$

Poichè S è naturalmente un rivestimento di un aperto di $\mathcal{M}_g$, lo spazio dei moduli delle curve di genere g (e poichè $\mathcal{M}_g$ è ricoperto da aperti di questo tipo) il calcolo di parametri cui si è accennato si traduce nel problema di calcolare la dimensione delle componenti irriducibili di $\pi'(\mathcal{G}^{r,k}_d)$.

Vogliamo offrire al lettore una discussione informale dei risultati noti in questa direzione.

Sia I un punto di $\mathcal{G}^r_d$, corrispondente a un fibrato in rette $L \to C = C_s$ e a un sottospazio $(r+1)$-dimensionale $W \subset H^0(C, L)$. Vi è un diagramma commutativo

di successioni esatte



Se ne deduce intanto che, se $\mathrm{Ker}\ \mu_1 = 0$ e $I \in \mathcal{G}_d^{r,k}$ allora

$$\dim \pi'_* T_I(\mathcal{G}_d^r) = 3g - 3 - k$$

In particolare se Y è una componente irriducibile di $\mathcal{G}_d^{r,k}$ e vale la proprietà

(8.1) "Per un punto generale di Y, $\mathrm{Ker}\ \mu_1 = 0$"

allora

(8.2) $\dim \pi'(Y) \leq 3g - 3 - k$.

Non ci è affatto chiaro in quali ipotesi si abbia l'eguaglianza in (8.2), anche supponendo che valga (8.1). Nè è chiaro per quali valori di r, k, d valga (8.1).

La condizione (8.1) è certamente verificata per $r = 1$ (vedi §5). Se poi

$$\rho < 0, \qquad k = -\rho$$

e perciò $\mathcal{G}_d^{1,k} = \mathcal{G}_d^1$, se ne deduce che $\pi'(\mathcal{G}_d^1)$ ha codimensione pari almeno a k [3], [24]. Più precisamente, B. Segre [22] ha dimostrato che la codimensione di $\pi'(\mathcal{G}_d^1)$ è esattamente $k = -\rho$.

Inoltre, ripetendo, con modifiche formali, la dimostrazione del Teorema (3.3), si può mostrare che $\mathcal{G}_d^1$ è liscia e di dimensione $3g - 3 + \rho$ [3] (si veda anche [1]). In particolare il risultato di B. Segre sopra ricordato implica che, *se una curva di genere* $g \geq 3$ *possiede una* g_d^1 *tale che* $\rho \leq 0$, *in generale ne possiede un numero finito.*

Sia ora Y una componente irriducibile di $\mathcal{G}_d^{2,k}$. Supponiamo che Y contenga un punto corrispondente a una curva C e a una g_d^2 su C (anche con punti base) tale che il morfismo

$$C \to \mathbf{P}^2$$

da essa definito induca una birazionalità tra C e l'immagine di C. Vogliamo mostrare che, in queste ipotesi, vale (8.2).

E' semplice mostrare che

$$k \leq 2 \dim H^1(C, L) \leq 2g - 4$$

Si può perciò supporre che $\pi'(Y)$ abbia dimensione pari almeno a $g+1$. Sia ora I un punto generale di Y corrispondente a un sottospazio tridimensionale W di $H^0(C, L)$. Vogliamo mostrare che Ker $\mu_1 = 0$. Se si sostituisce la g_d^2 (definita da W) con la corrispondente $g_{d'}^2$ senza punti base, la dimensione di Ker μ_1 non diminuisce. Si può perciò supporre che la g_d^2 non abbia punti base. Sia φ il morfismo di C in $\mathbf{P}^2$ associato a W. Poichè I è un punto generale di Y, $\pi'_*(T_I(\mathcal{G}_d^{2,k}))$ ha dimensione pari almeno a $g+1$. Applicando il Corollario (6.11) si ottiene

$$\dim H^0(C, N'_\varphi) \geq g + 1$$

e poichè N'_φ è un fibrato in rette si ha

$$(\text{Ker } \mu_1)^* = H^1(C_1 N_\varphi) = H^1(C, N'_\varphi) = 0$$

In conclusione si può affermare che *la sottovarietà dello spazio dei moduli $\mathcal{M}_g$ corrispondente a quelle curve di genere g che posseggono una g_d^2, non composta con una involuzione, e per cui il nucleo di μ_0 abbia dimensione eguale almeno a k, ha codimensione non inferiore a k.* E' facile vedere che l'ipotesi che la g_d^2 non sia composta con una involuzione è essenziale.

BIBLIOGRAFIA

[1] ARBARELLO, E., CORNALBA, M., GRIFFITHS, P. and HARRIS, J., *Topics in the theory of algebraic curves*, di prossima pubblicazione in Annals of Mathematics Studies.

[2] BRILL, A. and NOETHER, M., *Über die algebraischen Funktionen und ihre Anwendungen in der Geometrie*, Math. Ann. 7 (1873), 269–310.

[3] FARKAS, H. M., *Special divisors and analytic subloci of the Teichmüller space*, Amer. J. Math. *88* (1966), 881–901.

[4] FITTING, H., *Die Determinantenideale eines Moduls*, Jahresber. Deutsch. Math. Verein. *46* (1936), 195–228.

[5] FULTON, W., *Hurwitz schemes and irreducibility of moduli of algebraic curves*, Annals of Math. *90* (1969), 542–575.

[6] GRIFFITHS, P. and HARRIS, J., *The dimension of the variety of special linear systems on a general curve*, Duke Math. J. 47 (1980), 233–272.
[7] GROTHENDIECK, A., *Technique de descente et théorèmes d'existence en géometrie algébrique V. Les Schémas de Picard; théorèmes d'existence*, Séminaire Bourbaki 14^e année (1961–62) Exposé 232.
[8] ——, DIEUDONNÉ J., *Eléménts de géometrie algébrique* III 2, n. 17, I.H.E.S. Publ. Math., 1963.
[9] HODGE, W. V. and PEDOE, D., *Methods of algebraic geometry*, Cambridge 1952.
[10] HORIKAWA, E., *On deformations of holomorphic maps* I, J. Math. Soc. Japan *25* (1973), 372–396.
[11] KEMPF, G., *Schubert methods with an application to algebraic curves.* Publications of Mathematisch Centrum, Amsterdam 1971.
[12] ——, *On the geometry of a theorem of Riemann*, Ann. Math. *98* (1973), 178–185.
[13] ——, *On the collapsing of homogeneous bundles*, Inventiones Math. *37* (1976), 229–239.
[14] KLEIMAN, S. and LAKSOV, D., *On the existence of special divisors*, Amer. J. Math. *94* (1972), 431–436.
[15] KODAIRA, K. and SPENCER, D. C., *On deformations of complex analytic structures I, II*, Ann. Math. *67* (1957), 328–401, 403–465.
[16] LAX, R. F., *On the dimension of varieties of special divisors*, Trans. Amer. Math. Soc. *203* (1975), 141–159.
[17] MARTENS, H., *On the varieties of special divisors on a curve*, J. Reine Angew. Math. *227* (1967), 111–120.
[18] MUMFORD, D., *Abelian varieties*, Tata Inst. Studies in Math., Oxford University Press, London–New York 1974.
[19] PETRI, K., *Über Spezialkurven I*, Math. Ann. *93* (1924), 182–209.
[20] POPP, H., *Moduli theory and the classification theory of algebraic varieties*, Lecture Notes in Mathematics vol. 620, Springer-Verlag, Berlin–New York 1977.
[21] SCHIFFER, M. and SPENCER, D. C., *Functionals of finite Riemann surfaces*, Princeton University Press, Princeton 1954.
[22] SEGRE, B., *Sui moduli delle curve poligonali, e sopra un complemento al Teorema di esistenza di Riemann*, Math. Annalen *100* (1928), 537–551.
[23] SERRE, J. P., *Géometrie algébrique et géometrie analytique*, Ann. Inst. Fourier *6* (1955), 1–42.
[24] SEVERI, F., *Sul teorema di esistenza di Riemann*, Rendiconti del Circolo Matematico di Palermo *46* (1922), 105–116.
[25] SZPIRO, L., *Travaux de Kempf, Kleiman, Laksov sur les diviseurs exceptionnels*, Sém. Bourbaki 24^e Année (1971–72), Exposé 417.
[26] WEYL, H., *The classical groups*, Princeton, 1939.
[27] ZARISKI, O., *On the moduli of algebraic functions possessing a given monodromie group*, Amer. J. Math. *52* (1930), 150–170.

Enrico Arbarello
Department of Mathematics
Harvard University
Cambridge, Ma. 02138, U.S.A.

Maurizio Cornalba
Istituto di Matematica
Università di Pavia
I-27100 Pavia
Italia

Reçu le 19 mai 1980

Nota aggiunta all'alto della correzione delle bozze: la congettura di Petri è stata recentemente dimostrata da D. Gieseker.

Addendum 7.1.1981. Correzione a "su una congettura di Petri"

L'enunciato della Proposizione (3.13) in [1] non è corretto. L'errore nella dimostrazione sta nel calcolo della classe di coomologia di $\tilde{N}$; è ben vero che $\tilde{N}$ è tagliata da una sottovarietà lineare di $\mathbf{P}''$, però questa sottovarietà ha, in generale, codimensione superiore a $i(r+1)$.

Il testo di [1], dalla parola "Parimenti," riga 8, pagina 16, va perciò sostituito con:

"Nelle ipotesi di (3.10), indichiamo con $\bar{r}+1$ e i le dimensioni di $H^0(C, L)$ e $H^1(C, L)$, rispettivamente, e poniamo $\mathbf{G} = \mathrm{Gr}(r+1, H^0(C, L)) = c^{-1}(L)$. Segue da (3.10) che il grado del proiettivizzato del cono tangente a $W^r_d(C)$ in L è pari a

$$\xi^{g-(r+1)(g-d+r)-1}[\mathbf{P}(N)] = c_{(\bar{r}-r)(r+1)}(-N)[\mathbf{G}],$$

dove ξ è la classe di Chern del fibrato tautologico su $\mathbf{P}(N)$. D'altra parte, poiché (3.1) è iniettiva, $c(-N) = c(E^*)$, dove E è il fibrato vettoriale su $\mathbf{G}$ la cui fibra su un punto W è il sottospazio $W \otimes H^0(C, K \otimes L^{-1})$ di $H^0(C, L) \otimes H^0(C, K \otimes L^{-1})$. Se identifichiamo $\mathbf{G}$ con la grassmanniana $\mathrm{Gr}(\bar{r}-r, \bar{r}+1)$, E^* si identifica alla somma diretta di i copie del fibrato universale quoziente Q su $\mathrm{Gr}(\bar{r}-r, \bar{r}+1)$. D'altra parte è ben noto che, indicando con σ_j la classe di coomologia del circlo di Schubert dei sottospazi lineari $(\bar{r}-r)$-dimensionali di $\mathbf{C}^{\bar{r}+1}$ la cui intersezione con un sottospazio fissato di dimensione $r+2-j$ ha dimensione pari almeno a uno, si ha

$$\sigma_j = c_j(Q) \qquad j = 1, \ldots, r+1.$$

In conclusione

(3.13) PROPOSIZIONE. *Nelle ipotesi del Teorema* (3.10) *il proiettivizzato* $\tilde{T}$ *del cono tangente a* $W^r_d(C)$ *in* L *è una sottovarietà di* $\mathbf{P}^{g-1} = \mathbf{P}H^1(C, \mathcal{O})$ *di grado*

$$(1+\sigma_1+\cdots+\sigma_{r+1})^{g-d+\bar{r}[\mathrm{Gr}(\bar{r}-r,\bar{r}+1)]}$$

dove si è posto $\bar{r} = \dim(H^0(C, L)) - 1$."

Comment. Math. Helvetici **56** (1981) 39–58 0010–2571/81/001039–20$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Graph theoretic techniques in algebraic geometry II: construction of singular complex surfaces of the rational cohomology type of $\mathbf{CP}^2$

LAWRENCE BRENTON,* DANIEL DRUCKER and GEERT C. E. PRINS

Abstract. Methods of graph theory are used to obtain rational projective surfaces with only rational double points as singularities and with rational cohomology rings isomorphic to that of the complex projective plane. Uniqueness results for such cohomology $\mathbf{CP}^2$'s and for rational and integral homology $\mathbf{CP}^2$'s are given in terms of the types A_k, D_k, or E_k of singularities allowed by the construction.

This paper continues our discussion [4] of the use of graph theoretic methods in the construction of compact projective algebraic surfaces with rational singularities. In [4] we were concerned with complex spaces which compactify affine 2-space $\mathbf{C}^2$. The purpose of the present work is to apply similar techniques to provide examples of, and to classify according to singularity type, certain singular complex projective surfaces which have the same rational cohomology ring as the complex projective plane $\mathbf{CP}^2$. Our main result is:

THEOREM 1. *For each of the following twelve 8-point Dynkin diagrams Γ there exists a complete rational complex projective algebraic surface of the rational cohomology type of $\mathbf{CP}^2$ whose singularities are precisely the rational double points associated to the components of Γ: A_8, D_8, E_8, A_7+A_1, E_7+A_1, E_6+A_2, D_5+A_3, A_4+A_4, D_4+D_4, $A_5+A_2+A_1$, $D_6+A_1+A_1$, and $A_3+A_3+A_1+A_1$.*

Introduction

In an earlier paper ([2]), methods of the classical geometry of algebraic surfaces were applied to questions of existence and uniqueness of complex spaces of the homotopy or cohomology type of $\mathbf{CP}^2$. Although there is an error in the examples of [2] (corrected in [3]), the main result of [2] gives the uniqueness result below. The statement of the result requires a bit of terminology. A compact

* Supported in part by National Science Foundation grant no. MCS 77-03540.

complex surface X is a *rational homology* $\mathbf{CP}^2$ if $\forall i$,

$$H^i(X, \mathbf{Q}) \cong H^i(\mathbf{CP}^2, \mathbf{Q}) = \begin{cases} \mathbf{Q} & \text{for } i = 0, 2, 4 \\ 0 & \text{otherwise} \end{cases}$$

X is a *rational cohomology* $\mathbf{CP}^2$ if there is a natural ring isomorphism $\hat{H}^*(X, \mathbf{Z}) \to \hat{H}^*(\mathbf{CP}^2, \mathbf{Z})$, where $\hat{H}^*$ means $H^*/(\text{Torsion subgroup})$. "Natural" means that the isomorphism preserves the duals of the homology class of a point and of the entire space, regarded respectively as oriented 0- and 4-cycles.

THEOREM 2 ([2], Theorem 6). *Let X be a compact two-dimensional complex analytic space whose singularities are rational double points. Suppose further that X is a rational homology $\mathbf{CP}^2$ and that $\hat{H}^2(X, \mathbf{Z})$ is generated by an effective analytic divisor. Then X is a rational projective algebraic surface. Indeed, one of the following holds:*

(a) *X is biholomorphic to $\mathbf{CP}^2$ ($\Leftrightarrow X$ is non-singular).*

(b) *X is biholomorphic to the singular complex quadratic cone $\mathbf{Q}_0^2 = \overline{\{x^2 + y^2 + z^2 = 0\}} \subset \overline{\mathbf{C}^3} = \mathbf{CP}^3$.*

(c) *For some integer n with $3 \leq n \leq 8$, X is derived from $\mathbf{CP}^2$ by the successive application of n monoidal point transformations, followed by the blowing down of precisely n non-singular rational curves having self-intersection -2. In this case, the cohomology ring structure is determined by the fact that $g^2 = 9 - n$, where g is a generator of $H^2(X, \mathbf{Z}) \cong \mathbf{Z}$. X is a rational cohomology $\mathbf{CP}^2$ exactly when $n = 8$.*

Note: It was mistakenly supposed in [2] that for $n = 8$ these spaces are in fact homotopy equivalent to $\mathbf{CP}^2$. This error is rectified in [3]. The necessity of the condition "$\hat{H}^2(X, \mathbf{Z})$ is generated by an effective divisor" was recently shown by Mumford in [9], where an example is presented of a rational cohomology $\mathbf{CP}^2$ which is a non-singular projective surface of general type.

Following the algorithm of (c), examples of cohomology $\mathbf{CP}^2$'s were constructed in [2]. The details of the method are as follows. For some integer $m \leq 8$, start with m projective lines L_i on $\mathbf{CP}^2$ and blow up 8 points, possibly including infinitely near points, on $C = \cup_{i=1}^m L_i$. Call the resulting surface $\tilde{X}$ and let $\rho : \tilde{X} \to \mathbf{CP}^2$ be the map inverse to the monoidal transformations. Suppose that among the $m + 8$ components C_j of the curve $\rho^{-1}(C) \subset \tilde{X}$ there are precisely 8 that together comprise the exceptional set for the minimal resolution of one or more of the classical rational double points (that is, each of the 8 C_j in question satisfies $C_j^2 = -2$ and the dual intersection graph is the disjoint union of one or more of the Dynkin diagrams A_k, D_k, or E_k), while the remaining C_j are exceptional of the first kind (non-singular rational with $C_j^2 = -1$). Then the unique normal

analytic space X obtained from $\tilde{X}$ by collapsing each of the connected components of the union of these 8 C_j separately to a point will satisfy the conditions of the theorem and will have the rational cohomology type of $\mathbf{CP}^2$.

At the time that [2] was written it was not known how many different constructions of this type were possible nor what combinations of singular points the resulting surfaces could have. A priori there are 39 different graphs with 8 vertices and with components of the form A_k, D_k, or E_k, but not all 39 occur. To discover which of the 39 occur and which do not is the "thankless task" mentioned in [2], page 429. This question has assumed new interest with recent work of Ronald Fintushel on rational cohomology $\mathbf{CP}^2$'s which are singular 4-manifolds, each singularity being the cone on a rational homology 3-sphere. Considered as topological spaces, the surfaces X of type (c) constructed as above are certainly such objects – indeed, they are singular 4-manifolds with singularities of the required type which in addition support a complex analytic structure.

In this paper we will give the complete list of all rational cohomology $\mathbf{CP}^2$'s that can arise by this construction.

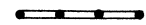
THEOREM 3. *Let X be a rational cohomology $\mathbf{CP}^2$ constructed by the above technique. Then X has at most 4 singular points x_i. Let $\Gamma = \cup_i \Gamma_i$ be the disjoint union of the Dynkin diagrams associated to the x_i and let* $\det(\Gamma)$ *be the determinant of the Cartan matrix associated to Γ. Then* $\det(\Gamma)$ *is the square of an integer less than or equal to 8. Conversely, for each 8-point graph $\Gamma \neq D_4 + 2A_2$ with 4 or fewer components, each a Dynkin diagram of type A_k, D_k, or E_k, and with* $\det(\Gamma) = j^2$ *for some integer $j \leq 8$, at least one such space X exists. Explicitly, the graphs Γ satisfying these conditions are the 12 listed in Theorem 1 above.*

The method of proof is purely graph theoretical, but at several important points the graph theory sheds light on matters of topological and geometric interest as well. We thank Paul Catlin, Daniel Frohardt, Peter Malcolmson, and the other participants of the Wayne State University Graph Theory Seminar (April, 1979) for valuable conversations about these ideas.

Preliminaries

By a *hypergraph* on a set V of m distinct vertices $v_1, \ldots, v_m$ we mean a system (V, Γ) where Γ is a set of non-empty subsets (called *edges*) of V. A *graph* is a hypergraph in which each edge has cardinality 2. A singleton edge $\{v_i\}$ is pictured as a loop at v_i. In this paper however, we will deal only with hypergraphs that have no singleton edges. Edges of cardinality $k > 2$ will be indicated by the

symbol . If a hypergraph Γ has no singleton edges and in addition satisfies the condition that for each pair of distinct vertices v_i, v_j there exists a unique edge $S \in \Gamma$ with $v_i, v_j \in S$, then Γ will be called *minimally complete.* For example, the minimally complete hypergraphs on 4 vertices are

 ,  , and 

A *weighted hypergraph* is a hypergraph in which each vertex v_i is assigned an integer "weight" n_i.

If Γ is a weighted hypergraph on m vertices v_i, with weight n_i on v_i, the *intersection matrix of* Γ is the m by m square symmetric matrix (a_{ij}) where a_{ij} equals n_i if $i=j$, -1 if $i \neq j$ and some element of Γ contains $\{v_i, v_j\}$, and 0 otherwise. By the *determinant* (respectively, *trace*) of Γ (abbreviated $\det(\Gamma)$, $\operatorname{tr}(\Gamma)$), we shall mean the determinant (trace) of the intersection matrix. Note that the trace is just the sum $\sum_{i=1}^{m} n_i$ of the weights.

Now let Y be an algebraic surface and let $C = \bigcup_{i=1}^{m} C_i$ be a curve on Y whose components C_i meet transversally with C_i meeting C_j in at most one point whenever $i \neq j$. By the *dual intersection hypergraph* associated to $C \subset Y$ we mean the weighted hypergraph on m vertices $v_1, \ldots, v_m$ defined by $\{v_{i_1}, \ldots, v_{i_k}\} \in \Gamma \Leftrightarrow C_{i_1} \cap \cdots \cap C_{i_k} \neq \emptyset$, with weight $-C_i^2$ on v_i. (We use the negatives of the self-intersection numbers to avoid having to alternate the signs of the determinants with successive monoidal transformations – see below.) A hypergraph Γ is called (*complex*) *projective planar* if it is dual to a collection of projective lines on $\mathbf{CP}^2$. A projective planar hypergraph is necessarily minimally complete, reflecting the fact that any two lines on the projective plane meet in exactly one point.

Let Γ be a weighted hypergraph on vertices $v_1, \ldots, v_m$ with weight n_i on v_i, and for $S = \{v_{i_1}, \ldots, v_{i_k}\}$ an edge of Γ define a new hypergraph $\sigma_S(\Gamma)$ on $m+1$ vertices $v_1, \ldots, v_{m+1}$ by

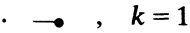
$$\sigma_S(\Gamma) = \Gamma - \{S\} \cup \{\{v_{i_1}, v_{m+1}\}, \ldots, \{v_{i_k}, v_{m+1}\}\}$$

with weights 1 on v_{m+1}, $n_i + 1$ on v_i if $v_i \in S$, and n_i on v_i if $v_i \notin S$, $i \leq m$. Similarly, if v_i is a vertex of Γ, denote by $\sigma_{v_i}(\Gamma)$ the hypergraph on $v_1, \ldots, v_{m+1}$ defined by

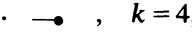
$$\sigma_{v_i}(\Gamma) = \Gamma \cup \{v_i, v_{m+1}\}$$

with weights 1 on v_{m+1}, $n_i + 1$ on v_i, and n_j on v_j for $j \neq i, m+1$. These operations are dual to the monoidal transformation for non-singular curves meeting transversally on an algebraic surface. The operation σ_S (respectively, σ_{v_i}) is called the

σ-process centered at S (resp., v_i), and applying σ_S (resp., σ_{v_i}) to Γ is called "blowing up" the edge S (resp., the vertex v_i). Any hypergraph can be transformed into a graph by blowing up all of its edges of cardinality $k>2$. If $S=\{v_{i_1},\dots,v_{i_k}\}$ is an edge of Γ of cardinality k, then $\operatorname{tr}(\sigma_S(\Gamma))=\operatorname{tr}(\Gamma)+k+1$, since each of the k weights n_{i_j}, $j=1,\dots,k$, is raised by 1 and a new vertex v_{m+1} of weight 1 is inserted, while the other weights remain the same. Similarly, $\operatorname{tr}(\sigma_{v_i}(\Gamma))=\operatorname{tr}(\Gamma)+2$. Finally, we remind the reader of the Dynkin diagrams (Coxeter graphs)

A_k:  , $k=1,2,\dots$

D_k: , $k=4,5,\dots,$ and

E_k:  , $k=6,7,$ or 8

(all weights are 2). These are the graphs associated to the fundamental root systems of the simple complex Lie algebras $\mathfrak{sl}(k+1,\mathbf{C})$, $\mathfrak{o}(2k,\mathbf{C})$, $\mathfrak{e}_k$ (cf., eg., [8], §§1, 11). They are also the weighted dual intersection graphs of the exceptional curves appearing in the minimal resolutions of the double points

$$A_k\colon\ x^2=x^{k+1}+y^2,\qquad D_k\colon\ z^2=x^{k-1}+xy^2,\qquad E_6\colon\ z^2=x^4+y^3,$$
$$E_7\colon\ z^2=x^3y+y^3,\qquad E_8\colon\ z^2=x^5+y^3.$$

These are the only two-dimensional hypersurface singularities $x\in X$ which are *rational* ($R^1\pi_*\mathcal{O}_{\tilde{X}}$ vanishes at x for $\pi:\tilde{X}\to X$ a resolution – see Artin [1]). They are precisely the singularities of the form $\mathbf{C}^2/G$ where G is a finite subgroup of $\mathrm{SL}(2,\mathbf{C})$ (namely, a cyclic group or a binary dihedral, tetrahedral, octahedral, or icosahedral group). Indeed, much attention has been directed toward understanding the relations among the various settings in which these graphs occur (see especially [10], [5], and [7]).

The determinants of these graphs are $\det(A_k)=k+1$, $\det(D_k)=4$, and $\det(E_k)=9-k$. This can be verified directly, or by the method of [6], or by computing the quotients G/G' for $G\subset\mathrm{SL}(2,\mathbf{C})$ the appropriate Kleinian group, where G' is the commutator subgroup of G. Since G/G' is also the first homology group of a spherical neighborhood of the associated rational double point, these determinants have topological significance in the construction of the complex surfaces X.

The main result

The problem of constructing rational homology and cohomology $\mathbf{CP}^2$'s by the technique described in the introduction reduces to the following question. Let n be an integer ≤ 8, and let Γ_0 be a projective planar hypergraph on m vertices, $m \leq n$, with each vertex of weight -1. Perform n successive σ-processes on Γ_0 in such a way that the final hypergraph $\tilde{\Gamma} = \sigma^n(\Gamma_0)$ has m vertices of weight 1 and n vertices – including the m original ones – of weight 2. Require further that the sub-hypergraph Γ of $\tilde{\Gamma}$ obtained by deleting all the vertices of weight 1 and all the edges adjoining them be the disjoint union of Dynkin diagrams of the form A_k, D_k, or E_k. In how many ways can this be done?

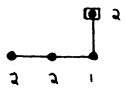
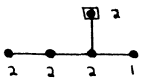
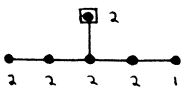
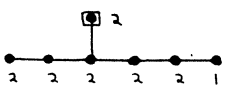
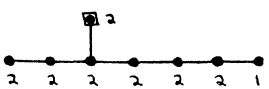
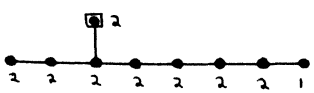
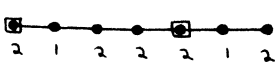
The complete solution is contained in the first 51 rows of the following table. The last 3 rows show the only 3 examples which satisfy every condition except complex projective planarity of the initial hypergraph. That is, these represent solutions to the dual graph theoretic problem for minimally complete hypergraphs, but do not translate into the geometric construction. In the last column the rational cohomology type of the resulting space X is given. By $\bar{S}_j$, $j = 3, 4, \ldots$, we mean the singular complex rational surface obtained by collapsing to a point the zero section of the $\mathbf{CP}^1$-bundle on $\mathbf{CP}^1$ with Chern class $-j$ ($\bar{S}_2$ = the singular quadric surface $\mathbf{Q}_0^2$, $\bar{S}_1 = \mathbf{CP}^2$). These are the prototypical homology $\mathbf{CP}^2$'s with cohomology ring structure given by $g^2 = j$, for g a generator of $H^*(X, \mathbf{Z})$. In the final graphs the (proper transforms of) the original vertices are denoted by the symbol ◙.

The next section will be devoted to justifying the table. Note that Theorems 1 and 3 follow from the table. That is, in the 31 rows of the table which represent cohomology $\mathbf{CP}^2$'s, the distinct Dynkin diagrams that appear are exactly the 12 listed in Theorem 1. To check the characterization in terms of determinants (Theorem 3) it is sufficient to calculate the determinants of the twelve 8-point Dynkin diagrams that appear in the table and the 27 which do not. (Recall that in all there are thirty-nine 8-point graphs with components of type A_k, D_k, or E_k.) Note that, except for $A_7 + A_1$, each of the disconnected diagrams which appear consists of two disjoint graphs Γ' and Γ'' (not necessarily connected) with $\det(\Gamma') = \det(\Gamma'')$. Thus $\det(\Gamma)$ is the square $(\det(\Gamma'))^2$. (The "missing" graph $D_4 + 2A_2$ does not have this property, although its determinant is the square of an integer ≤ 8 and it has 4 or fewer components.)

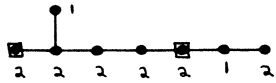
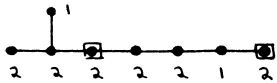
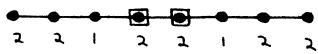
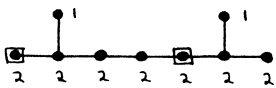
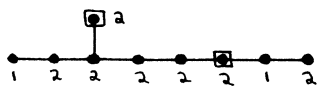
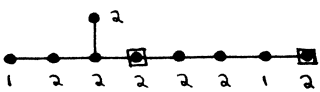
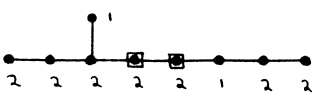
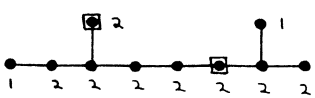
Proof of the main result. To justify the table, we must determine all initial hypergraphs Γ_0, and for each Γ_0, all sequences of $n \leq 8$ blow-ups permitted by the conditions on the final hypergraph $\tilde{\Gamma}$, as given in the first paragraph of the previous section.

First we note that since each of the original vertices of Γ_0 becomes part of the

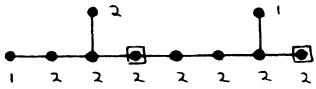
Table for the construction of rational homology $\mathbb{CP}^2$'s

Final graph	Singularities	Determinant	Cohomology type
Initial hypergraph •			
	$A_2 + A_1$	6	$\overline{S}_6$
	A_4	5	$\overline{S}_5$
	D_5	4	$\overline{S}_4$
	E_6	3	$\overline{S}_3$
	E_7	2	Q_0^2
	E_8	1	$\mathbb{P}^2$
Initial hypergraph •—•			
	$A_3 + 2A_1$	16	$\overline{S}_4$

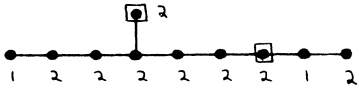
Initial hypergraph ●—● (continued)

	$A_5 + A_1$	12	$\bar{S}_3$
	$A_5 + A_1$	12	$\bar{S}_3$
	$3A_2$	27	$\bar{S}_3$
	A_7	8	Q_0^2
	$D_6 + A_1$	8	Q_0^2
	$D_6 + A_1$	8	Q_0^2
	$A_5 + A_2$	18	Q_0^2
	D_8	4	$\mathbb{P}^2$

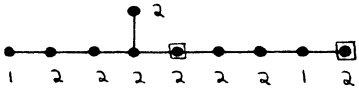
Initial hypergraph ●—● (continued)



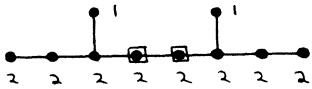
D_8 4 $\mathbb{P}^2$



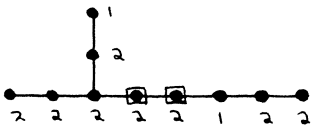
$E_7 + A_1$ 4 $\mathbb{P}^2$



$E_7 + A_1$ 4 $\mathbb{P}^2$

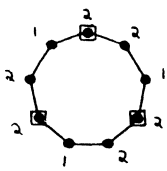


A_8 9 $\mathbb{P}^2$

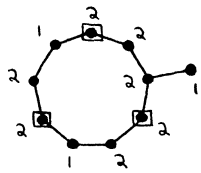


$E_6 + A_2$ 9 $\mathbb{P}^2$

Initial hypergraph 

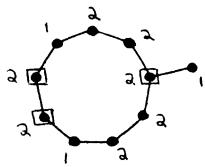
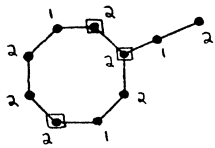
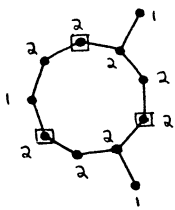
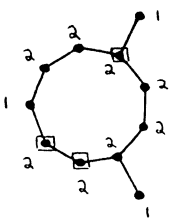
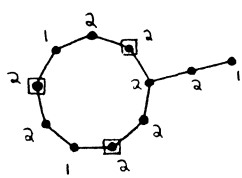
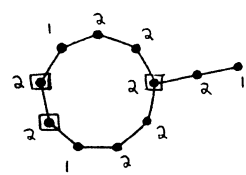


$3A_2$ 27 $\overline{S}_3$

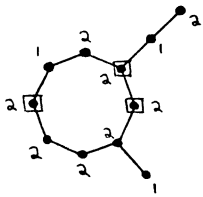


$A_5 + A_2$ 18 Q_0^2

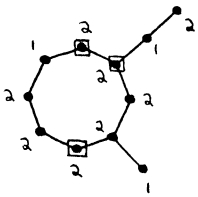
Initial hypergraph  (continued)

	$A_5 + A_2$	18	Q_0^2
	$2A_3 + A_1$	32	Q_0^2
	A_8	9	$\mathbb{P}^2$
	A_8	9	$\mathbb{P}^2$
	$E_6 + A_2$	9	$\mathbb{P}^2$
	$E_6 + A_2$	9	$\mathbb{P}^2$

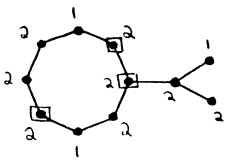
Initial hypergraph  (continued)



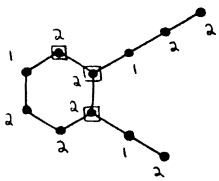
$A_7 + A_1$ 16 $\mathbb{P}^2$



$A_7 + A_1$ 16 $\mathbb{P}^2$

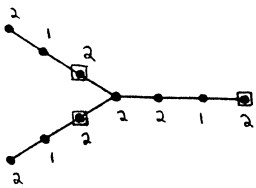


$D_5 + A_3$ 16 $\mathbb{P}^2$



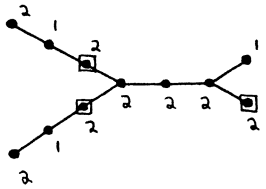
$A_5 + A_2 + A_1$ 36 $\mathbb{P}^2$

Initial hypergraph

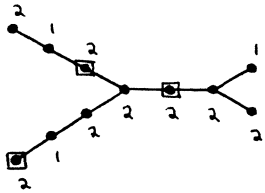


$D_4 + 3A_1$ 32 Q_0^2

Initial hypergraph (continued)

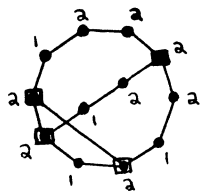


$D_6 + 2A_1$ 16 $\mathbf{P}^2$

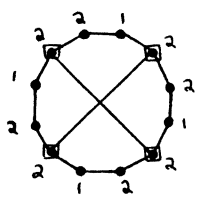


$D_6 + 2A_1$ 16 $\mathbf{P}^2$

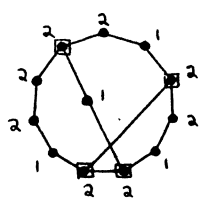
Initial hypergraph 



$D_5 + A_3$ 16 $\mathbf{P}^2$

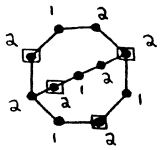


$2A_4$ 25 $\mathbf{P}^2$

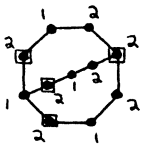


$2A_4$ 25 $\mathbf{P}^2$

Initial hypergraph 



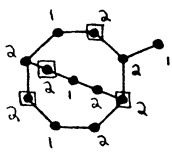
$2A_3 + A_1$ 32 Q_0^2



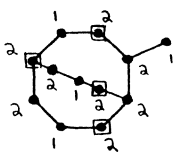
$D_4 + 3A_1$ 32 Q_0^2



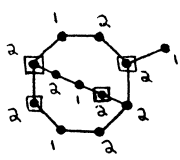
$A_7 + A_1$ 16 $\mathbb{P}^2$



$D_5 + A_3$ 16 $\mathbb{P}^2$

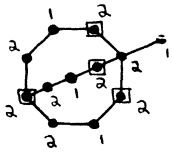
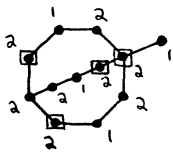
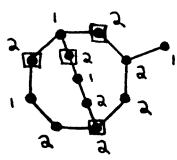


$D_5 + A_3$ 16 $\mathbb{P}^2$

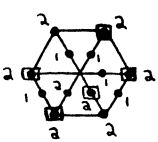
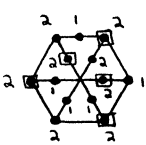


$D_5 + A_3$ 16 $\mathbb{P}^2$

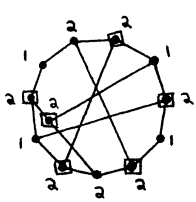
Initial hypergraph  (continued)

	$2D_4$	16	$\mathbb{P}^2$
	$2D_4$	16	$\mathbb{P}^2$
	$D_6 + 2A_1$	16	$\mathbb{P}^2$

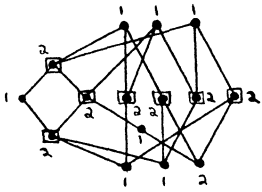
Initial hypergraph 

	$2A_4$	25	$\mathbb{P}^2$
	$A_5 + A_2 + A_1$	36	$\mathbb{P}^2$

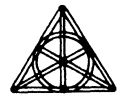
Initial hypergraph 

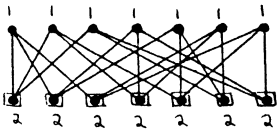
	$A_5 + A_2 + A_1$	36	$\mathbb{P}^2$

Initial hypergraph 

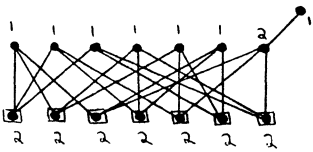


$2A_3 + 2A_1$ 64 $\mathbb{P}^2$

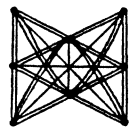
Initial hypergraph 

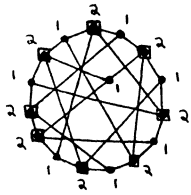


$7A_1$ 128 Does not occur over $\mathbb{C}$



$D_4 + 3A_1$ 64 Does not occur over $\mathbb{C}$

Initial hypergraph 



$4A_2$ 81 Does not occur over $\mathbb{C}$

graph $\Gamma \subset \tilde{\Gamma}$, all of the multiple edges of Γ_0 must be blown up. Furthermore, Γ_0 can have no edges of cardinality greater than 3. For if $S = \{v_{i_1}, \ldots, v_{i_k}\}$ is an edge of Γ_0 with $k \geq 4$, then $\sigma_S(\Gamma_0)$ contains k vertices v_{i_j} of weight 0 with no two in an edge of $\sigma_S(\Gamma_0)$, and these k vertices must have weight 2 in the final graph $\tilde{\Gamma}$. Thus each v_{i_j} must be at the center of 2 more blow-ups, neither of which can involve v_{i_l} for $l \neq j$. This makes a total of at least $1+2k > 8$ blow-ups, violating the restriction $n \leq 8$.

Using the fact that in a minimally complete hypergraph each pair of triple edges shares at most 1 vertex, the following lemma is quite easy to check case by case.

SUBLEMMA. *The maximum number e of triple edges in a minimally complete hypergraph on m points,* $1 \leq m \leq 8$, *is as follows*:

$m = 1, 2$	$e = 0$	$m = 6$	$e = 4$
$m = 3, 4$	$e = 1$	$m = 7$	$e = 7$
$m = 5$	$e = 2$	$m = 8$	$e = 8$

Furthermore, the minimally complete hypergraphs realizing these maxima are unique up to isomorphism.

Let x, y, and z denote the number of 3-edges, 2-edges, and vertices blown up in passing from Γ_0 to $\tilde{\Gamma}$. Then $x+y+z = n \leq 8$, and the traces satisfy $\operatorname{tr}(\tilde{\Gamma}) = \operatorname{tr}(\Gamma_0) + 4x + 3y + 2z = \operatorname{tr}(\Gamma_0) + 2n + 2x + y$. Since $\operatorname{tr}(\Gamma_0) = -m$ (m vertices, each with weight -1) and $\operatorname{tr}(\tilde{\Gamma}) = 2n + m$ (n vertices of weight 2, m vertices of weight 1) we obtain $2n + m = -m + 2n + 2x + y$, or

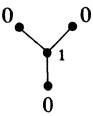
$$(*) \qquad m = (y/2) + x \leq 4 + (x/2).$$

We want to determine all possibilities for Γ_0 and the triple (x, y, z).

Case 1. $x = 0$, $m = y/2 \leq 4$. Γ_0, having no triple edges, is the complete graph K_m on m points ($1 \leq m \leq 4$), and $y = 2m$. If $m = 1$, of course z cannot be zero, for K_1 has no edge to be blown up. Likewise if $m = 2$ ($\Leftrightarrow y = 4$) then $z \neq 0$ since it is impossible to insert 4 new vertices by σ-processes into the lone edge of K_2 without increasing the weight of one of the vertices to more than 2. Thus the possible initial hypergraphs and their triples (x, y, z) are in this case:

Γ_0	(x, y, z)
K_1	$(0, 2, z)$, $1 \leq z \leq 6$
K_2	$(0, 4, z)$, $1 \leq z \leq 4$
K_3	$(0, 6, z)$, $0 \leq z \leq 2$
K_4	$(0, 8, 0)$.

Case 2. $x = 1$, $m = (y/2) + 1 \leq 4$. Since Γ_0 has a triple edge, m must be at least 3,

so the only choices for Γ_0 are and  . If $m=3$, $y=4$. After the triple edge of is blown up to obtain 

, at most one of the three 2-edges can be blown up, and no weight can be added to the other two original vertices except by blowing them up. Thus $z \geq 2$ in this case. The possibilities are

Γ_0	(x, y, z)
	$(1, 4, 2)$
	$(1, 4, 3)$
	$(1, 6, 0)$
	$(1, 6, 1)$

Case 3. $x=2$, $m=(y/2)+2 \leq 5$. By the sublemma, $m>4$. Thus $m=5$, $y=6$, and Γ_0 is the unique minimally complete hypergraph on 5 points with 2 triple edges.

Γ_0	(x, y, z)
	$(2, 6, 0)$

Case 4. $x=3$, $m=(y/2)+3 \leq 5$. By the sublemma, this cannot occur.

Case 5. $x=5$, $m=(y/2)+4 \leq 6$. The sublemma gives $m=6$, whence $y=4$.

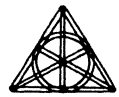
Γ_0	(x, y, z)
	$(4, 4, 0)$

Case 6. $x=5$, $m=(y/2)+4 \leq 6$. This, too, is impossible by the sublemma.

Cases 7, 8, *and* 9 similarly give the following possibilities

Γ_0	(x, y, z)
	$(6, 2, 0)$

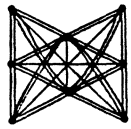
(the unique minimally complete hypergraph on 7 points with 6 triple edges)



(7, 0, 0)

(7, 0, 1)

(the Fano projective plane $\mathbf{P}^2(\mathbf{Z}_2)$)



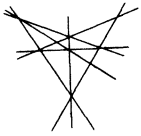
(8, 0, 0)

$m = x = 8$

It is easy to verify directly that the first 9 candidates for Γ_0 on this list are complex projective planar and that the last 2 are not. For instance, the hypergraph

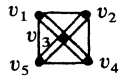


is dual to the collection of projective lines



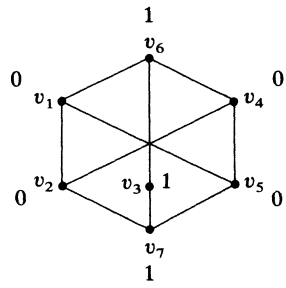
(The 7-point graph of Fano and the 8-point graph do, however, occur over fields of characteristic 2 and 3 respectively. This phenomenon will be explored in part III of this series of papers.)

From here, an easy but tedious exhaustion of cases yields the possible final graphs $\tilde{\Gamma}$ for each pair $(\Gamma_0, (x, y, z))$. We will give one example to illustrate the technique.



(2, 6, 0)

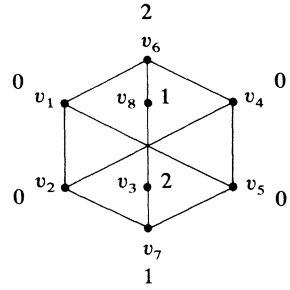
After blowing up the 2 triple edges we obtain the weighted graph

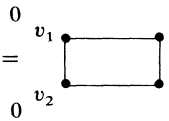


where v_6 and v_7 are the new vertices inserted by blowing up the triples $\{v_1, v_3, v_4\}$ and $\{v_2, v_3, v_5\}$ respectively. Since no vertices can be blown up ($z = 0$), the only

way to raise the weight of v_3 is by blowing up one of the two edges to which it belongs. By symmetry we may suppose that $\{v_1, v_6\}$ is blown up.

$\sigma^3(\Gamma_0)$:

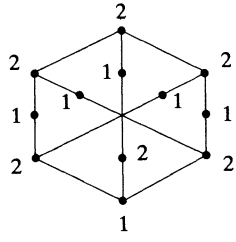


Consider the subgraph $\Gamma' =$ 

. In the 5 remaining σ-processes suppose that y' edges of Γ' are blown up, together with $z' \leq 5 - y'$ "vertices" of Γ' (i.e., edges of one of the graphs $\sigma^k(\Gamma_0)$, $3 \leq k \leq 3 + y'$ which contain only 1 of the vertices v_1, v_2, v_4, and v_5). Since the weight of each vertex of Γ' must be raised by 2, $2y' + z' = 8$. The only solutions are

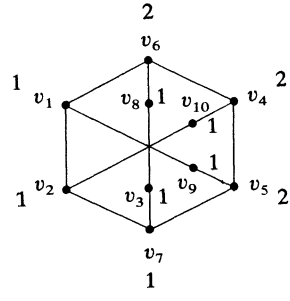
(A) $y' = 3, z' = 2,$ and (B) $y' = 4, z' = 0.$

Case (B) cannot occur, for after blowing up each edge of Γ' we obtain the dead end

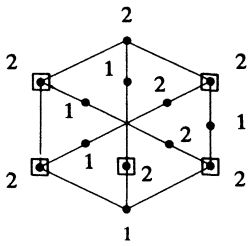


in which no further blow-ups of edges are possible. In case (A), by symmetry we may suppose that $\{v_1, v_5\}$, $\{v_2, v_4\}$, and $\{v_4, v_5\}$ are blown up.

$\sigma^6(\Gamma_0)$:

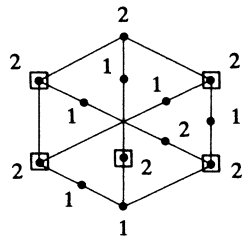


The remaining edge $\{v_1, v_2\}$ of Γ' cannot be blown up, no edge of $\sigma^6(\Gamma_0)$ can be subdivided twice, and v_2 can be involved in only one more blow-up, so it is clear that the remaining 2 σ-processes must blow up $\{v_1, v_9\}$ and either $\{v_2, v_{10}\}$ or $\{v_2, v_7\}$. The resulting final graphs are



$(A_5+A_2+A_1)$

and



(A_4+A_4)

These are lines 49 and 48 of the table, and, up to the order in which the 8 σ-processes are performed, these are the only permissible constructions with an initial hypergraph on 5 points.

The other 52 lines of the table are derived similarly. This completes the proof.

REFERENCES

[1] ARTIN, M., *On isolated rational singularities of surfaces*, Amer. J. Math. **84** (1962), 485–496.

[2] BRENTON, L., *Some examples of singular compact analytic surfaces which are homotopy equivalent to the complex projective plane*, Topology **16** (1977), 423–433.

[3] BRENTON, L., and BINDSCHADLER, D., *On singular 4-manifolds of the homology type of* $\mathbf{CP}^2$, to appear.

[4] BRENTON, L., DRUCKER, D., and PRINS, G. C. E., *Graph theoretic techniques in algebraic geometry I: the Dynkin diagram* $\tilde{E}_8$ *and minimal singular compactifications of* $\mathbf{C}^2$, to appear in the Proceedings of the Conference on Complex Analysis, Princeton, 1979.

[5] BRIESKORN, E., *Singular elements of semi-simple algebraic groups*, Actes. Congrès Inter. Math. **2** (1970), 279–284.

[6] DRUCKER, D., and GOLDSCHMIDT, D., *Graphical evaluation of sparse determinants*, Proc. A.M.S. **77** (1979), 35–39.

[7] DU VAL, P., *Homographies, quaternions, and rotations*, Oxford Univ. Press, Oxford (1957).

[8] HUMPHREYS, J., *Introduction to Lie Algebras and Representation Theory*, Graduate Text in Mathematics 9, Springer-Verlag, New York, Heidelberg, Berlin (1972).

[9] MUMFORD, D., *An algebraic surface with K ample,* $(K^2)=9$, $p_g=q=0$, Amer. J. Math. **101** No. 1 (1979), 233–244.

[10] STEINBERG, R., *Classes of elements of semisimple algebraic groups*, Proc. Int. Congr. Math., Moscow (1966), 277–289.

Department of Mathematics
Wayne State University
Detroit, Michigan 48202 U.S.A.

Received May 20, 1980

Comment. Math. Helvetici **56** (1981) 59–65

0010–2571/81/001059–07$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

On the existence of escaping geodesics

VICTOR BANGERT[1]

In [5] Cohn-Vossen remarks that to his knowledge the following question is undecided

PROBLEM. Let M be a complete Riemannian manifold homeomorphic to the plane. Does there exist an escaping geodesic $c: \mathbf{R} \to M$ without self-intersections?

Here "escaping" means that c is a proper map, cf. Section 1.

Recently the same problem, though in weaker form, has been posed by Wojtkowski [9]. In [8] and [9] Wojtkowski applies symbolical dynamics to study the geodesic flow of certain complete surfaces. The existence of escaping geodesics is a crucial prerequisite for this application.

In this paper we solve Cohn-Vossen's problem in the affirmative. We have not been able to find a general method to construct escaping geodesics without self-intersections. Thus we use different methods according to different properties of the geodesic flow. Intuitively the most difficult case is when there exists a closed geodesic on M. In this case we use Lusternik–Schnirelmann theory on the space of curves with fixed end-points. When closed geodesics do not exist we apply results from [3]. We know that in this case bounded geodesic rays do not exist either. Furthermore the possible self-intersections of geodesics are of a very restricted type. This allows us to prove that through every point of a complete plane without closed geodesics there exists a geodesic without self-intersections. Unfortunately we are not able to exclude that this geodesic is oscillating.

Hence, in a final step, we prove that, in the absence of closed geodesics, the existence of an oscillating geodesic implies the existence of an escaping geodesic without self-intersections. We note that our arguments generalize to Finsler metrics, at least if one admits self-intersections of the escaping geodesic.

I would like to thank W. Ballmann for helpful conversations on the Lusternik–Schnirelmann deformation $\mathcal{D}^{\sim}$, cf. Section 2.

---

[1] Work supported by the program Sonderforschungsbereich "Theoretische Mathematik" at the University of Bonn

1. Preliminaries

We consider a complete Riemannian manifold M homeomorphic to the plane $\mathbf{R}^2$. Such M is called a complete plane. The distance function on M is denoted by d. A curve on M is a piecewise C^1-map $\gamma: I \to M$ defined on some interval $I \subseteq \mathbf{R}$. If I is compact the energy $E(\gamma) := \int_I |\dot\gamma(t)|^2\, dt$ and the length $L(\gamma) := \int_I |\dot\gamma(t)|\, dt$ of γ are finite. For $I = [a, b]$ length and energy of γ are related by $L^2(\gamma) \leqslant (b-a)E(\gamma)$ with equality if and only if γ is parametrized proportionally to arc-length. For $p, q \in M$ we denote by $\Gamma_{p,q}$ the space of curves $\gamma: [0, 1] \to M$ with $\gamma(0) = p$, $\gamma(1) = q$. $\Gamma_{p,q}$ is endowed with the compact-open topology. By $\tilde\Gamma_{p,q}$ we denote the subspace of $\Gamma_{p,q}$ consisting of curves without self-intersections. Finally, for $\kappa \geqslant 0$, we let $\Gamma^\kappa_{p,q}$ be the set of curves in $\Gamma_{p,q}$ of energy $\leqslant \kappa$ and $\tilde\Gamma^\kappa_{p,q} := \tilde\Gamma_{p,q} \cap \Gamma^\kappa_{p,q}$.

On a complete non-compact Riemannian manifold M the geodesic rays can be divided naturally into the following three classes, cf. [8]:

DEFINITION. A geodesic ray $c: [0, \infty) \to M$ is called

(i) bounded if its image $c([0, \infty))$ is a bounded set.

(ii) escaping if c is a proper map, i.e. if $c(t_i)$ diverges for every divergent sequence t_i in $[0, \infty)$.

(iii) oscillating if neither (i) nor (ii) is true.

A complete geodesic $c: \mathbf{R} \to M$ is bounded resp. escaping if both its geodesic rays are. In [6] escaping geodesic rays are called divergent. A geodesic $c: [a, b] \to M$ is called a geodesic loop if $c(a) = c(b)$. c is called simple if $c \mid [a, b)$ is one-to-one. A geodesic $c: \mathbf{R} \to M$ is simple closed if c is non-constant, periodic and if, for its smallest period ω, $c \mid [0, \omega)$ is one-to-one.

2. Escaping geodesics in the presence of closed geodesics

In this section we prove the existence of an escaping geodesic without self-intersections on a complete plane M containing a simple closed geodesic d. Examples for such M are complete planes of finite surface area or, more generally, complete planes containing a horn, cf. [2] and [8].

According to [3], Theorem 2, the existence of such a closed geodesic on M is equivalent to the existence of a bounded geodesic ray or to the existence of a compact, locally concave set D. This last property of M is essential for our proof. Here we can take as compact, locally concave set the disc D bounded by d. The

following lemma is crucial:

LEMMA 1. *For $p, q \in M$ and some $\kappa > 0$ let there be given a continuous homotopy $g : [0, 1] \to \tilde{\Gamma}^{\kappa}_{p,q}$ such that g_0 and g_1 are curves in $M - D$ which are not homotopic in $M - D$. Then $\tilde{\Gamma}^{\kappa}_{p,q}$ contains a geodesic which intersects D.*

Proof. To illustrate the idea we first forget about possible self-intersections and only construct a geodesic $c \in \Gamma := \Gamma^{\kappa}_{p,q}$ intersecting D. Our tool is the Birkhoff curve shortening process $\mathcal{D} : \Gamma \to \Gamma$, cf. [4], Section V, 7. To define $\mathcal{D}$ choose some big $k \in \mathbf{N}$. Then $\mathcal{D} := \mathcal{D}_{k-2} \circ \cdots \circ \mathcal{D}_0$ where $\mathcal{D}_i : \Gamma \to \Gamma$, $0 \leq i \leq k-2$, is defined as follows: For $\gamma \in \Gamma$ we let

$$\mathcal{D}_i \gamma \Big| \left[\frac{i}{k}, \frac{i+2}{k}\right]$$

be the shortest geodesic from $\gamma(i/k)$ to $\gamma((i+2)/k)$ while $\mathcal{D}_i\gamma$ coincides with γ on the rest of $[0, 1]$. If k is large enough each $\mathcal{D}_i$ is well-defined and continuous. We set $\mathcal{D}^n := \mathcal{D} \circ \mathcal{D}^{n-1}$, $\mathcal{D}^1 := \mathcal{D}$. Then the following is true:

(i) For every $\gamma \in \Gamma$ there exists a convergent subsequence of $\{\mathcal{D}^n\gamma\}_{n \in \mathbf{N}}$ and every limit curve of $\{\mathcal{D}^n\gamma\}_{n \in \mathbf{N}}$ is a geodesic in Γ.

(ii) If $\gamma \in \Gamma$ is contained in $M - D$ then so is $\mathcal{D}\gamma$. Furthermore γ and $\mathcal{D}\gamma$ are homotopic in $M - D$.

Here (i) is the standard property of the Birkhoff curve shortening process while (ii) is due to the fact that D is bounded by a geodesic. Now c is obtained as follows: By (ii) the curves $\mathcal{D}^n g_0$ and $\mathcal{D}^n g_1$ are not homotopic in $M - D$. Since $\mathcal{D}^n \circ g$ is a homotopy from $\mathcal{D}^n g_0$ to $\mathcal{D}^n g_1$ the set $I_n := \{t \in [0, 1] \mid \mathcal{D}^n g_t \text{ intersects } D\}$ is non-void. I_n is compact and, by (ii), we have $I_n \supseteq I_{n+1}$. Hence there exists $t \in [0, 1]$ such that $\mathcal{D}^n g_t$ intersects D for all $n \in \mathbf{N}$. Now every limit curve c of $\{\mathcal{D}^n g_t\}_{n \in \mathbf{N}}$ has the desired properties.

We now proceed to construct a geodesic $c \in \tilde{\Gamma} := \tilde{\Gamma}^{\kappa}_{p,q}$ intersecting D. This geodesic is produced by a continuous, energy-decreasing deformation $\mathcal{D}^{\sim}_h : \tilde{\Gamma} \to \tilde{\Gamma}$ which has "almost" properties (i) and (ii). To find such deformations is known to be an intricate problem. Here we use an adaptation to the present situation of the classical Lusternik–Schnirelmann deformation $\mathcal{D}^{\sim}$, cf. [7]. $\mathcal{D}^{\sim}$ has been described with great care in [1], and is actually only defined up to the choice of a constant. The following two problems arise:

(1) $\mathcal{D}^{\sim}$ is defined so as to map closed curves without self-intersections to closed curves without self-intersections. When $\mathcal{D}^{\sim}$ is applied to curves with end-points $\mathcal{D}^{\sim}$ may not leave the end-points fixed.

(2) $\mathcal{D}^{\sim}$ fails to have property (ii).

Difficulty (2) can easily be helped since $\mathcal{D}^\sim$ almost has property (ii): For $a>0$ set $D^a:=\{p\in M \mid d(p,q)\leq a \text{ for some } q\in D\}$. Then, given $\varepsilon>0$, we can define $\mathcal{D}^\sim$ in such a way that $\mathcal{D}^\sim\gamma$ does not intersect D^ε if γ does not intersect $D^{2\varepsilon}$. This amounts to an appropriate choice of the constant involved in the definition of $\mathcal{D}^\sim\gamma$, cf. [1].

To get around difficulty (1) we slow $\mathcal{D}^\sim$ down to the identity as soon as γ comes to lie close to its end-points. This means the following: Choose two small balls B_p and B_q about p and q and a function $h: M\to[0,1]$ which vanishes in a neighborhood of p and q and is identically one outside $B_p\cup B_q$. Choose some very big $k\in\mathbf{N}$. Using the Lusternik–Schnirelmann construction we define deformations $\mathcal{D}_i^\sim$, $(3\leq i\leq k-3)$, which replace

$$\gamma\left|\left[\frac{i}{k},\frac{i}{k}+h\left(\gamma\left(\frac{i}{k}\right)\right)\frac{2}{k}\right]\right.$$

by the geodesic segment $c_{\gamma,i}$ from $\gamma(i/k)$ to

$$\gamma\left(\frac{i}{k}+h\left(\gamma\left(\frac{i}{k}\right)\right)\frac{2}{k}\right)$$

while the rest of γ is so deformed that self-intersections are avoided. This deformation takes place in some small neighborhood of $c_{\gamma,i}$ only. If k is big enough the end-points of γ will never lie in such a neighborhood. We define $\mathcal{D}_h^\sim$ by $\mathcal{D}_h^\sim=\mathcal{D}_{k-3}^\sim\circ\mathcal{D}_{k-4}^\sim\circ\cdots\circ\mathcal{D}_3^\sim$. Using similar arguments as before we can find $t\in[0,1]$ such that $(\mathcal{D}_h^\sim)^n g_t$ intersects D^{ε_n} for $\varepsilon_n=\varepsilon\cdot 2^{-n}$. When we restrict a limit curve of $\{(\mathcal{D}_h^\sim)^n g_t\}_{n\in\mathbf{N}}$ to an appropriate subinterval $[s_0,s_1]\subseteq[0,1]$ we obtain a geodesic without self-intersections which starts in B_p, ends in B_q, intersects D and has energy $\leq\kappa$. If we let the radii of B_p and B_q converge to zero we obtain a geodesic $c\in\tilde{\Gamma}$ intersecting D.

Now we can prove

THEOREM 1. *Let M be a complete plane containing a simple closed geodesic d. Then there exists an escaping geodesic $c:\mathbf{R}\to M$ without self-intersections.*

Proof. c will be obtained as a limit of geodesics constructed by means of Lemma 1. We first construct a sequence of homotopies g^n which satisfy the assumptions of Lemma 1. Take any smooth disc D' containing D in its interior where D is the set bounded by d. Then there exists $\kappa>0$ such that for all $p,q\in\partial D'$ there is a homotopy $h:[0,1]\to\tilde{\Gamma}_{p,q}^\kappa$ such that h_0 and h_1 are contained and not homotopic in $M-D$ and such that all the curves h_t are contained in D'. Obviously we can assume that the curves h_t are parametrized proportionally to

arc-length. Now choose points $p_n \neq q_n$ at distance $n \in \mathbf{N}$ from D'. There exist geodesics c_n, d_n of length n such that c_n joins p_n to D' while d_n joins D' to q_n. We define a homotopy $g^n : [0, 1] \to \tilde{\Gamma}_{p_n, q_n}$ by requiring that g_t^n be the composition of c_n, h_t and d_n, parametrized proportionally to arc-length. Then $E(g_t^n) \leq (2n + \sqrt{\kappa})^2$ for all $t \in [0, 1]$ and g^n satisfies the assumptions of Lemma 1. Thus there exists a geodesic $\gamma_n' \in \tilde{\Gamma}_{p_n, q_n}$ which intersects D and has length $L(\gamma_n') = (E(\gamma_n'))^{1/2} \leq 2n + \sqrt{\kappa}$. Let $\gamma_n : [a_n, b_n] \to M$ be a unit speed parametrization of γ_n' such that $a_n < 0 < b_n$ and $\gamma_n(0) \in D$. Then $b_n - a_n = L(\gamma_n) \leq 2n + \sqrt{\kappa}$. Since p_n and q_n are points at distance n from $D' \supseteq D$ we have $-a_n \leq n + \sqrt{\kappa}$ and $b_n \leq n + \sqrt{\kappa}$. Let v be a limit vector of the sequence $\{\dot{\gamma}_n(0)\}$ and define $c(t) = \exp(tv)$. Then $c(t) = \lim \gamma_n(t)$ and c does not have self-intersections. We have

$$d(\gamma_n(t), D) \geq d(\gamma_n(b_n), D) - d(\gamma_n(b_n), \gamma_n(t)) \geq n - (b_n - t) \geq t - \sqrt{\kappa}.$$

Now $d(c(t), D) \geq t - \sqrt{\kappa}$ implies that $c \mid [0, \infty)$ is escaping. An analogous argument shows that $c \mid (-\infty, 0]$ escapes as well.

Remark. Our geodesic c even has the following property which is stronger than being escaping: Both its rays $c \mid [0, \infty)$ and $c \mid (-\infty, 0]$ are almost minimizing in the sense of [6], Definition 7.1; i.e. there exists $A > 0$ such that $d(c(t), c(0)) \geq |t| - A$.

3. Escaping geodesics in the absence of closed geodesics

On a complete plane M the absence of simple closed geodesics has surprisingly strong implications on the geodesic flow, cf. [3], Section 2. As far as self-intersections are concerned the geodesics on M have similar properties as geodesics on complete planes of positive Gaussian curvature, cf. [5]. Hence it is not surprising that [5], Satz 11 has an analogue in our situation:

THEOREM 2. *Let M be a complete plane without simple closed geodesics. Then through every point of M there exists a geodesic $c : \mathbf{R} \to M$ without self-intersections.*

Remark. By [3], Theorem 2, the geodesic rays of c are not bounded.

Proof. According to [3], Corollary 2 a non-constant geodesic $c : \mathbf{R} \to M$ either has no self-intersections or c determines uniquely a simple geodesic loop $c \mid [t_1, t_2]$. Hence we can repeat Cohn-Vossen's proof of [5], Satz 11: Assume that all geodesics through $p \in M$ have self-intersections. For every unit vector $v \in T_pM$ let c_v denote the simple geodesic loop determined by the geodesic $c(t) = \exp(tv)$.

Obviously c_v and c_{-v} bound the same disc but with the opposite orientation. Since this orientation depends on v continuously we obtain a contradiction.

We are not able to conclude that the geodesic c from Theorem 2 is escaping. This is actually part of the following open problem: Can oscillating geodesic rays exist on complete planes without simple closed geodesics?

Regardless of the answer to this question the following proposition completes the proof for the existence of an escaping geodesic without self-intersections on every complete plane:

PROPOSITION. *Let M be a complete plane without simple closed geodesics. Suppose there exists an oscillating geodesic ray $d:[0,\infty)\to M$. Then there exists an escaping geodesic without self-intersections.*

Proof. We first recall that a totally convex set $C\subseteq M$ is a closed set such that every geodesic which joins two points of C is contained in C. The totally convex hull A^c of a set $A\subseteq M$ is the intersection of all totally convex sets containing A. Using this concept and a result from [3] we will find a sequence s_i diverging to ∞ such that $\dot{d}(s_i)$ converges to the initial vector of an escaping geodesic without self-intersections.

Choose a sequence t_i diverging to ∞ such that $\lim d(t_i)=:p$ exists. Let B be a compact ball about p and, for $n\in\mathbf{N}$, let K_n denote the closure of $B\cap d([n,\infty))$. Then the totally convex hull K_n^c of K_n contains $d([t,\infty))$ for some $t>0$. According to [3], Corollary 1, there exist points $q_n\in K_n\cap\partial K_n^c$. Let q be a limit point of the sequence q_n. Since K_n^c is a decreasing sequence of closed sets we have $q\in\bigcap K_n^c=:C$. Now $q_n\in\partial K_n^c$ implies $q\in\partial C$. Since $q_n\in K_n$ we can find a sequence s_i diverging to ∞ such that $\lim\dot{d}(s_i)=:w$ exists and $w\in T_qM$. Then the geodesic $c:\mathbf{R}\to M$ with initial vector $\dot{c}(0)=w$ is contained in C since $c(t)=\lim d(s_i+t)$ for all $t\in\mathbf{R}$. Actually c is contained in the boundary ∂C of the totally convex set C since $q=c(0)\in\partial C$. Hence c is either a simple closed geodesic or an escaping geodesic without self-intersections. The first case is excluded by assumption.

Finally we note that an escaping geodesic without self-intersections can easily be constructed on all complete non-compact Riemannian manifolds which contain a non-separating compact hypersurface or which have at least two ends. In particular the answer to our problem is yes for all complete non-compact surfaces M which are not planes. For such surfaces one can even obtain geodesics without self-intersections which "start" and "end" in the same end of M unless $M\simeq S^1\times\mathbf{R}$.

Obviously one can ask many interesting questions in this context, even in the 2-dimensional case. We just mention two of them:

1. Are there infinitely many escaping geodesics on every complete plane M, or even one through every point of M?

2. What is the minimal number of escaping geodesics without self-intersections on complete planes?

REFERENCES

[1] BALLMANN, W., *Der Satz von Lusternik und Schnirelmann.* Bonner Math. Schriften Nr. 102, Bonn 1978.
[2] BANGERT, V., *Closed geodesics on complete surfaces.* Math. Ann. *251* (1980), 83–96.
[3] ——, *Geodesics and totally convex sets on surfaces. Preprint.*
[4] BIRKHOFF, G. D., *Dynamical Systems.* Colloq. Publ. IX, Amer. Math. Soc., Providence R.I. 1927.
[5] COHN-VOSSEN, S., *Totalkrümmung und geodätische Linien auf einfachzusammenhängenden offenen vollständigen Flächenstücken.* Mat. Sbornik (Recueil Mathématique) *1* (*43*) (1936), 139–163.
[6] EBERLEIN, P. and O'NEILL, B., *Visibility manifolds.* Pacific J. Math. *46* (1973), 45–109.
[7] LUSTERNIK, L. and SCHNIRELMANN, L., *Sur le problème de trois géodésiques fermées sur les surfaces de genre* 0. C.R. Acad. Sci. Paris *189* (1929), 269–271.
[8] WOJTKOWSKI, M., *Oscillating geodesics on two-dimensional manifolds. Astérisque 51* (1978), 443–456.
[9] ——, *Geodesics on open surfaces containing horns.* Preprint, Uniwersytet Warszawski, Warszawa (1979).

Universität Freiburg
Hebelstr., 29
7800 *Freiburg*
Federal Republic of Germany

Received June 5, 1980

Comment. Math. Helvetici **56** (1981) 66–82 0010-2571/81/001066-17$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Anwendung einer zahlengeometrischen Methode von C.L. Siegel auf Probleme der Analysis

E. HLAWKA

Prof. Chandrasekharan zum 60. Geburtstag gewidmet

In dem bekannten Buch von Chandrasekharan "Introduction to Analytic Number Theory" (Springer Verlag 1968) wird im 9. Kapitel "Geometry of Numbers" der Fundamentalsatz von Minkowski behandelt, der folgendermaßen lautet: Es sei K ein konvexer Körper im s-dimensionalen Raum, d.h. mit je 2 Punkten von K liegen auch alle Punkte der Verbindungsstrecke in K. Er besitze innere Punkte und weiterhin den Punkt $O=(0,\ldots,0)$ als Mittelpunkt, d.h. wenn der Punkt x in K liegt, dann liegt auch $-x$ in K. Ist nun sein Volumen $V(K)>2^s$, dann enthält K einen Gitterpunkt g, d.h. einen Punkt mit ganzzahligen Koordinaten, der aber von O verschieden ist und zwar im Inneren von K. Für diesen Satz sind viele Beweise gegeben worden. Chandrasekharan gibt in seinem Buch einen analytischen Beweis, der von C. L. Siegel (Acta mathematica 1935 Gesammelte Abhandlungen I, Springer Verlag 1965, S. 311) zum ersten Mal gegeben wurde. Der Beweis beruht auf einer Identität (Siegelsche Identität genannt).

Man betrachtet mit K alle Translate $K+2g$, wo g alle Gitterpunkte durchläuft. Es gilt dann

$$V(K)+\sum_{g\neq 0} V(K\cap K+2g)=\frac{1}{2^s}\left(V^2(K)+\sum_g\left|\int_K e^{\pi i\langle gx\rangle}\,dx\right|^2\right)$$

Dabei bedeutet $\langle gx\rangle$ das skalare Produkt von g und x, und dx das s-dimensionale Volumselement. Wenn nun K keinen Gitterpunkt $g\neq 0$ im Inneren enthält, dann muß $V(K\cap K+2g)=0$ sein, denn sonst würde es ein x in K geben, so daß auch $x-2g$ in K wäre, daher auch g in K.

Dann folgt aber aus der Identität, daß $V(K)\leq 2^s$ ist. Das ist ein Widerspruch zur Voraussetzung im Minkowskischen Satz. Die Siegelsche Beweismethode besteht darin, daß man die Funktion $\sum_g \chi_K(x+2g)$, wo χ_K die charakteristische Funktion von K ist, in eine Fourierreihe entwickelt und dann die Parseval'sche Gleichung benützt. Diese Siegelsche Methode gestattet noch andere Anwendungen.

So werden wir in §1 der vorliegenden Arbeit eine Ungleichung von N. Wiener, (vgl. Shapiro, Quart. J. *26*, 1975) verallgemeinern (siehe Formel (3) und Formel (13)).

Im §2 verallgemeinern wir Untersuchungen von A. Renyi (Acta Mat. Szeged *13*, 77–92, 1949–50) und vertiefen frühere Untersuchungen des Verfassers (E. Hlawka, Math. Annalen *125* (1952) 183–207, insbesondere S. 200).

Im §3 bauen wir die Überlegungen zu einem Summierungsverfahren für Fouriersche Reihen aus, das im eindimensionalen Spezialfall das Riemannsche Summierungsverfahren enthält. Der Grundgedanke beruht auf einer Behandlung des Riemannschen Summierungsverfahrens von Schechter (Monatshefte für Math. *22* (1911) 224–34).

Im §4 leiten wir eine Ungleichung ab, die im eindimensionalen Spezialfall eine Ungleichung enthält, die verallgemeinert, in der Theorie des großen Siebes, erhöhte Aufmerksamkeit auf sich gezogen hat.

§1

Es sei K ein konvexer Körper im R^s mit dem Mittelpunkt O und Volumen $V(K)$. Mit αK bezeichnen wir, wie üblich, die Menge aller αx, wo $x \in K$. Sein Volumen ist $\alpha^s V(K)$. Es sei E^s der s-dimensionale Einheitswürfel $0 \leq \varepsilon_1 < 1, \ldots, 0 \leq \varepsilon_s < 1$ und G das Würfelgitter, d.h. die Menge aller Punkte g im R^s mit ganzzahligen Koordinaten. Wir betrachten eine quadratische integrierbare Funktion F auf R^s, welche G als Periodengitter besitzt, d.h. es ist für alle $x \in R^s$ und alle $x \in G$

$$F(x+g) = F(x). \tag{1}$$

Die Fourierkoeffizienten $A(F, g)$ sind gegeben durch

$$A(F, g) = \int_{E^s} F(x) e(-\langle xg \rangle)\, dx, \tag{2}$$

wo $e(\alpha) = e^{2\pi i \alpha}$ ist.

SATZ 1. *Sind alle Fourierkoeffizienten $A(F, g)$ der Funktion F nicht negativ, enthält der Körper $2K$ keinen Gitterpunkt $g \neq 0$, und betrachten wir $\tilde{K} = \bigcup_g (g+2K)$, dann ist*

$$\int_{E^s \cap \tilde{K}} |F(x)|^2 \geq V(K) \int_{E^s} |F(x)^2|\, dx. \tag{3}$$

Ist $s=1$ und K ein Intervall $|x|\leq\alpha$, wo $\alpha<1/2$ ist, dann erhalten wir eine Ungleichung von N. Wiener (vgl. Shapiro loc. cit.). Wir gehen nun zum Beweis über. Wir setzen $\psi(K,x)$ definiert durch $\psi(K,x)=\sum_g \chi_K(x+g)$, dabei sei noch χ_K die charakteristische Funktion einer *beliebigen*, quadrierbaren Menge. Wir definieren

$$c(K,g)=A(\psi,g)=\int_{E^s}\sum_g \chi_K(x+g)e(-\langle xg\rangle)\,dx.$$

Da die Mengen E^s+g, wenn g alle Gitterpunkte durchläuft, den R^s einfach und lückenlos überdecken, so ist

$$c(K,g)=\int_{R^s}\chi_K(x+g)e(-\langle xg\rangle)\,dx. \tag{4}$$

Wir betrachten die Funktion ψ_t $(t\in R^s)$, definiert durch $\psi_t(x)=\psi(x+t)$ für alle x. Es ist $A(\psi_t,g)=c(K,g)e(\langle gt\rangle)$. Nach der Parsevalschen Gleichung ist

$$\int_{E^s}\psi(x)\bar{\psi}_t(x)\,dx=\sum_g |c(K,g)|^2\,e(-\langle gt\rangle|. \tag{5}$$

Wir beachten noch, daß nach (4)

$$c(K,0)=\text{Volumen von } K=V(K) \text{ ist.}$$

Weiters betrachten wir

$$\rho(K,t)=\frac{1}{V(K)}\int_{E^s}\psi(x)\bar{\psi}_t(x)\,dx. \tag{I}$$

Es ist

$$V(K)\rho(K,t)=\int_{E^s}\sum_k \chi_K(x+k)\bar{\psi}_t(x)\,dx,$$

also ist

$$V(K)\rho(K,t)=\int_{R^s}\chi_K(x)\bar{\psi}_t(x)\,dx. \tag{6}$$

Es ist stets für alle t

$$\rho(K, t) \geq 0 \tag{6'}$$

Aus (5) entnehmen wir, daß $\rho(g+t) = \rho(t)$ für alle g und alle t ist. Es ist

$$V(K)\rho(K, t) = \int_{R^2} \chi_K(x) \sum_k \chi_K(x+k+t), \tag{6''}$$

also ist

$$V(K)\rho(K, t) = \sum_k \int \chi_K(x)\chi_K(x+k+t),$$

also

$$V(K)\rho(K, t) = \sum_k V(K, K+k+t)$$

wo $V(K, K+x)$ das Volumen des Durchschnittes von K und der Translate $K+x$ ist.

Es sei nun K *konvex* und O sein Mittelpunkt. Dann ist stets

$$\rho(K, t) \leq \sum_g \chi_{2K}(g+t). \tag{7}$$

Dies ist klar, wenn $\rho = 0$ ist. Nehmen wir nun an, daß $\rho(t) > 0$ ist. Dann muß es nach (6) ein $\bar{x}$ geben, welches in K liegt und für welches

$$\psi_t(x) = \sum_g \chi_K(\bar{x}+g+t) > 0$$

ist. Dann muß es ein $\bar{g}$ geben, so daß

$$\bar{y} = \bar{x} + \bar{g} + t \in K$$

ist. Da O Mittelpunkt von K ist, gehört auch $-\bar{x}$ zu K, und da K konvex ist, gehört $(\bar{y}-\bar{x})/2$ zu K, also $\bar{g}+t$ zu $2K$. Es ist also

$$\sum_g \chi_{2K}(g+t) \geq 1.$$

Andererseits ist stets $\rho \leq 1$. Es ist nämlich stets $\chi_K(x) \leq 1$. Es ist aber auch $\psi_t(x) \leq 1$, denn wäre $\psi_t(x) > 1$, dann gäbe es Gitterpunkte g_1, g_2, wo $g_1 \neq g_2$, und einen Punkt $\bar{x}$, sodaß $\bar{x} + g_1 + t$ und $\bar{x} + g_2 + t$ in K. Dann würde aber, wenn man wie vorher schließt, $g = g_1 - g_2$ in $2K$, wobei $g \neq 0$ ist. Dies soll aber nicht sein. Es ist also

$$\int_{R^s} \chi_K(x)\psi_t(x)\,dx \leq \int_{R^s} \chi_K(x)\,dx = V(K),$$

also $\rho \leq 1$. Damit ist (7) bewiesen.

Es sei nun S eine integrierbare Funktion, so daß für alle t stets $S(t) \geq 0$ ist. Weiter seien alle ihre Fourierkoeffizienten $A(S, g) \geq 0$. Es ist natürlich

$$A(S, 0) = \int_{E^s} S(t)\,dt.$$

Dann folgt aus (7)

$$\rho(t)S(t) \leq S(t) \sum_g \chi_{2K}(g + t)\,dt,$$

also

$$\int_{E^s} S(t)\rho(t)\,dt \leq \int_{E^s} S(t) \sum_g \chi_{2K}(g + t)\,dt. \tag{8}$$

Nun ist nach (5)

$$V(K) \int_{E^s} S(t)\rho(t)\,dt = \sum_g |c(K, g)|^2 \int_{E^s} S(t)e(-\langle gt \rangle\,dt),$$

also

$$V(K) \int_{E^s} S(t)\rho(t)\,dt = \sum_g |c(K, g)|^2 A(S, g).$$

Es ist also

$$\int_{E^s} S(t)\rho(t)\,dt \geq V(K)A(S, 0). \tag{9}$$

Aus (8) und (9) folgt

$$\int_{E^s \cap \bar{K}} S(t)\, dt = \int_{E^s} S(t) \sum_g \chi_{2K}(g+t)\, dt \geq V(K) \int_{E^s} S(t)\, dt \tag{10}$$

eine auch an sich interessante Ungleichung.

Nun können wir Satz 1 in üblicher Weise herleiten. Es sei F eine integrierbare Funktion und

$$S(t) = |F(t)|^2.$$

Ist $\sum_g A(F, g)e(\langle xg \rangle)$ die Fourierreihe von F, so ist

$$\sum_{g,k} A(F, g)\bar{A}(F, k)e(\langle g-k, x \rangle) = \sum_L e(\langle Lx \rangle) \sum_{g-k=L} A(F, g)\bar{A}(F, k)$$

die Fourierreihe von S. Es ist

$$A(S, L) = \sum_{g-k=L} A(F, g)\bar{A}(F, k) \geq 0.$$

Es ist insbesondere nach der Parsevalschen Gleichung

$$A(S, 0) = \sum_g |A(F, g)|^2 = \int_{E^s} |F(x)|^2\, dt.$$

Daraus folgt (3).

§2

Es sei wieder K ein konvexer Körper mit Mittelpunkt O, es sei aber jetzt für alle Gitterpunkte $g \neq 0$ stets K und $K+g$ disjunkt. Es sei weiter eine endliche Menge B von Punkten b gegeben.

Wir bilden uns die Menge

$$K(B) = \bigcup_b (K + b + g). \tag{I}$$

Wir verlangen, daß die konvexen Mengen $K+b+g$ nie übereinandergreifen sollen. Wir betrachten die Funktion $\rho(K(B), t)$, wie sie in §1(I) für den Fall, daß die Menge $B = \{0\}$ ist, definiert wurde. Es gilt nun, wenn $|B|$ die Kardinalzahl von

B ist,

$$|B|\, c_K(0)\rho(K(B), t) = \sum_g |c(K, g)|^2\, |S(B, g)|^2\, e(-\langle gt\rangle). \tag{II}$$

Dabei ist

$$S(B, g) = \sum_b e(-\langle bg\rangle). \tag{1}$$

Beweis. Es ist nach §1(4)

$$c(K(B), g) = \sum_b \int_{R^s} \chi_{K+b}(x+g) e(-\langle xg\rangle)\, dx = \sum_b e(-\langle bg\rangle) c(K, g). \tag{2}$$

Es ist insbesondere $c(K(B), 0) = |B|\, c(K, 0)$.

Aus §1(5) folgt die Behauptung

Wie machen nun die *Voraussetzung*, daß es nicht negative Q und ϑ gibt, so daß für alle b aus B und alle $g \neq 0$.

$$||S(B, g)|^2 - Q| < \frac{\vartheta Q V}{1-V} \tag{III}$$

gilt. Dabei ist $V = V(K) = c(0)$. Aus (II) folgt, wenn $B = \{0\}$ und $t = 0$ ist, daß

$$|c(K, 0)| = \sum_g |c(K, g)|^2, \tag{3}$$

also $c(K, 0) \leq 1$ ist. Es ist ja nach §1(6)

$$\rho(K, 0) c(K, 0) = \int_{R^s} \chi_K(x) \sum_g \chi_K(x+g)\, dx = c(K, 0), \tag{4}$$

da ja die Mengen $K+g$ nicht übereinandergreifen sollen (Es liegt wieder der Minkowskische Fundamentalsatz vor). Wir behaupten im folgenden

SATZ 2. *Es ist für alle* $t \in R^s$

$$\rho(K(B), t) \geq V(K)\left(|B| - \frac{(Q(l+\vartheta)}{|B|}\right). \tag{IV}$$

Beweis. Es ist ja nach (II)

$$|B|\, c(K,0)\rho(K(B),t) = |C(K,0)|^2\, |B|^2 + A, \tag{5}$$

wo

$$A = \sum_{g\neq 0} |c(K,g)|^2\, |S(B,g)|^2\, e(-\langle gt\rangle) \tag{6}$$

ist. Nun ist nach Voraussetzung (III)

$$|S(B,g)|^2 = Q + \vartheta_g R, \tag{7}$$

wo

$$R = \frac{\vartheta QV}{1-V} \tag{8}$$

gesetzt ist. Es ist $|\vartheta_g| \leq 1$. Es ist also

$$A = \sum_{g\neq 0} |c(K,g)|^2\, (Q + \vartheta_g R) e(-\langle gt\rangle),$$

also

$$A = Q \sum_{g\neq 0} |c(K,g)|^2\, e(-\langle gt\rangle) + R_1, \tag{9}$$

wo

$$R_1 = R \sum_{g\neq 0} |c(K,g)|^2\, \vartheta_g e(-\langle gt\rangle)$$

ist. Es ist nun

$$\sum_{g\neq 0} |c(K,g)|^2\, e(-\langle gt\rangle| = c(K,0)\rho(K,t) - c^2(K,0)$$

wie aus (II) für $B = \{0\}$ folgt. Weiter ist

$$\sum_{g\neq 0} |c(K,g)|^2\, \vartheta_g e(-gt) \geq -\sum_{g\neq 0} |c(K,g)|^2.$$

Nun ist wieder nach (II) und (4)

$$\sum_{g\neq 0} |c(K,g)|^2 = c(K,0)\rho(K,0) - c^2(K,0) = c(K,0) - c^2(K,0).$$

Wir erhalten also

$$R_1 \geq -R(c(K,0)\rho(K,0) - c^2(K,0)) \tag{9'}$$

und

$$\sum_{g \neq 0} |c(K,g)|^2 \geq c(K,0) - c^2(K,0) = V - V^2. \tag{9''}$$

Wir erhalten also aus (8), (9'), (9'')

$$A \geq -QV^2 - R(V - V^2).$$

Es ist also nach (5)

$$|B|\, V\rho(K(B),t) \geq V^2 |B|^2 - QV^2 - RV(1-V),$$

und dies ergibt nach (8) gerade (IV). Es hat Satz 2 nur dann eine Bedeutung, wenn

$$|B|^2 \geq Q(1+\vartheta)$$

ist. Ein bemerkenswerter Fall liegt vor, wenn statt (III) die folgende Voraussetzung gilt: Es gibt eine natürliche Zahl d, so daß für alle $g \neq 0$ und $b \in B$

$$|S(B,g)|^2 = |B| - d \tag{V}$$

gilt. Es gilt dann

SATZ 3. *Unter der Voraussetzung* (V) *ist für alle* $t \in R^s$

$$|B|\, \rho(K(B),t) \geq dV(K) \tag{10}$$

Es gilt sogar schärfer

$$\rho(K(B),t) \geq V\left(|B| - 1 + \frac{d}{|B|}\right). \tag{10'}$$

Es gilt, wenn wir (5) und (6) benützen

$$\begin{aligned} A &= \sum_{g \neq 0} |c(K,0)|^2 (|B| - d) e(-\langle gt \rangle) \\ &= (|B| - d)(\rho(K(t) - |c(K,0)|^2), \end{aligned}$$

also ist

$$A \geq -(|B|-d)\,|c(K,0)|^2,$$

also ist

$$|B|\, V\rho(K(B), t) \geq V^2(|B|^2-|B|+d) \geq V^2 d.$$

Damit ist alles bewiesen.

Wie wollen nun einen Fall betrachten, wo (V) erfüllt ist. Es sei C eine endliche Obermenge von B (ihre Elemente bezeichnen wir mit c), so daß $\bigcup_{g,c}(g+c)$ eine Gruppe in bezug auf die Vektoraddition ist, also eine Obergruppe des Gitters G aller ganzzahligen Punkte von R^s. Es sei Γ^* die duale Gruppe zu Γ, welche in G enthalten ist. Es ist Γ^* die Menge aller Punkte l, für die stets für alle $\gamma \in \Gamma$, $\langle \gamma, l\rangle$ in Z liegt.

Es gilt dann bekanntlich für alle $l \in \Gamma^*$

$$\sum_c e(\langle l, c\rangle) = 0 \tag{11}$$

Dabei erstreckt sich in (11) die Summe über alle $c \in C$, also ist

$$\sum_{c \neq 0} e(\langle l, c\rangle) = -1. \tag{12}$$

Weiters ist für alle $c \neq 0$

$$\sum_l e(\langle lc\rangle) = 0 \tag{12'}$$

wo sich die Summe über alle inkongruenten l aus G modulo Γ^* erstreckt.

Wir nennen nun B eine *Differenzenbasis* von C von der Ordnung d, wenn sich jedes $c \neq 0$ genau d-mal in der Form $b'-b''$ schreiben läßt, wo b', b'' aus B sind. (Man vergleiche für den eindimensionalen Fall Renyi loc. cit. und für den allgemeinen Fall E. Hlawka loc. cit.) Wir zeigen nun

SATZ 4. *Es ist B genau dann Differenzenbasis von C von der Ordnung d, wenn* (V) *gilt.*

Beweis. Wenn B eine Differenzenbasis von C von der Ordnung d ist, so ist

$$|S(B,g)|^2 = \sum_{b',b'' \in B} e(\langle g(b'-b'')\rangle) = |B| + D$$

wo

$$D = \sum_{b',b'',b' \neq b''} e(g(b'-b'')) = \sum_{c \neq 0} A(c) e(\langle gc \rangle)$$

ist. Dabei ist $A(c)$ die Anzahl der b', b'', für die $b'-b''=c$ ist. Nun soll für alle $c \neq 0$ ja $A(c)=d$ sein. Dann ist also nach (11)

$$D = d \sum e(\langle g, c \rangle) = -d.$$

Damit ist (V) schon bewiesen.

Wir zeigen nun die Umkehrung: Es sei also für alle $g \neq 0$

$$\sum_{c \neq 0} A(c) e(\langle gc \rangle) = -d.$$

Setzen wir $K(c) = A(c) - d$ für $c \neq 0$ und $K(0) = 0$, so ist also

$$\sum_{c} K(c) e(\langle gc \rangle) = 0 \tag{13}$$

für alle $g \neq 0$. Es gilt also insbesonderes (13) für alle inkongruenten Gitterpunkte l modulo Γ^*, ausgenommen $l=0$. Es ist nun für ein beliebiges c' aus C

$$\sum_{l} e(-\langle lc' \rangle) \sum_{c} K(c) e(\langle lc \rangle) = K(c') |C|,$$

($|C|$ Kardinalzahl von C). Es ist ja nach (12')

$$\sum_{l} e(\langle l(c-c') \rangle) = 0, \quad \text{falls} \quad c \neq c'.$$

Andrerseits ist

$$\sum_{l} e(-\langle lc' \rangle) \sum_{c} K(c) e(\langle lc \rangle) = \sum_{l} \sum_{c} K(c) e(\langle l(c-c') \rangle),$$

also ist diese Summe gleich

$$\sum_{c} K(c) + \sum_{l \neq 0} e(-\langle lc' \rangle) \sum_{c} K(c) (e(\langle lc \rangle),$$

also nach (13) gleich $\sum_c K(c)$. Es ist also

$$K(c') |C| = \sum_{c} K(c).$$

Es sind also alle $K(c)$ einander gleich und wir haben nach (13)

$$K(c) \sum_{c \neq 0} e(\langle gc \rangle) = 0.$$

Da nach (12) diese Summe gleich -1 ist, so ist $K(c) = 0$ für alle $c \neq 0$, also ist $A(c) = d$. Damit ist Satz 4 bewiesen.

§3

Wir können die bisherigen Überlegungen auch zu einem Summierungsverfahren von Fourierschen Reihen periodischer Funktionen ausbauen. Wir hatten nach § 1 (6′) (es ist ja $c(K,0) = V(K)$)

$$\frac{1}{V(K)} \rho(K, t) = \frac{1}{V^2(K)} \sum_{g} \int_{R^s} \chi_K(x) \chi_K(x + g + t)\, dx. \tag{1}$$

Andererseits ist ($c(K, 0) = c_0$ gesetzt)

$$\frac{1}{c_0} \rho(K, t) = 1 + \frac{1}{c_0^2} \sum_{g \neq 0} |c(K, g)|^2 e(-\langle gx \rangle). \tag{2}$$

Es ist

$$\frac{1}{c_0} \int_{E^s} \rho(K, t)\, dt = 1. \tag{3}$$

Weiter ist

$$\rho(K, t) = 0 \tag{4}$$

für alle $t \in E^s - \tilde{K}$, wo $\tilde{K} = \bigcup (2K + g)$, wie in §1 definiert. Ist nämlich $\rho(t) > 0$, dann gibt es nach (1) ein $x \in K$ und einen Gitterpunkt g, so daß $x + g + t \in K$. Dann ist aber $-x$ in K, also $g + t$ in $2K$, also ist t in $\tilde{K}$, was nicht sein soll (vgl. die Überlegung in §1). Wir betrachten nun statt K den konvexen Körper hK, dann ist also

$$\rho(hK, t) = 0 \tag{4′}$$

für alle t in $E^s - \widetilde{hK}$. Es sei nun Φ eine stetige periodische Funktion, so daß also für alle g stets $\Phi(x+g) = \Phi(x)$ ist. Es besitze also Φ die Fourierreihe

$$\sum_g A(\Phi, g) e(\langle gt \rangle). \tag{5}$$

Wir behaupten nun

SATZ 5. *Es ist*

$$\lim_{h \to 0} \sum_g |c(hK, g)|^2 A(\Phi, g) e(\langle gx \rangle) = \Phi(x). \tag{6}$$

Beweis. Nach der Parsevalschen Gleichung ist

$$\sum_g |c(hK, g)|^2 A(\Phi, g) e(\langle gx \rangle) = \int_{E^s} \Phi(x+t) \frac{\rho(hK, t)}{c_0(hK)},$$

wo $c_0(hK) = c(hK, 0)$ gesetzt ist. Es ist nach (4')

$$\int_{E^s} (\Phi(x+t) - \Phi(x)) \rho(hK, t) c_0^{-1}(hK)\, dt$$

gleich

$$\int^* (\Phi(x+t) - \Phi(x)) \rho(hK, t) c_0^{-1}(hK)\, dt, \tag{7}$$

wo sich die Integration in (7) über $E^s \cap \widetilde{hK} = H(K)$ erstreckt. Nun ist nach (3)

$$\left| \int^* (\Phi(x+t) - \Phi(x)) \rho(hK, t) c_0^{-1}(hK)\, dt \right| \leq \sup_{t \in H} |\Phi(x+t) - \Phi(x)|.$$

Es ist also

$$\left| \sum_g |c(hK, g)|^2 A(\Phi, g) e(\langle gx \rangle) - \Phi(x) \right| \leq \sup_{t \in H} |\Phi(x+t) - \Phi(x)|. \tag{7'}$$

Wir wollen nun *voraussetzen*, daß die Körper $2K + g$ nicht übereinandergreifen. Das gleiche gilt dann für $2hK + g$, wenn $h \leq 1$.

Es ist also

$$\sup_{t\in H}|\Phi(x+t)-\Phi(x)|\leq\omega(\Phi, 2hK)$$

wobei

$$\omega(\Phi, 2hK)=\sup_{t\in 2hK}|\Phi(x+t)-\Phi(x)|$$

ist. Es ist $\omega(\Phi, 2hK)$ ein Maß für den Fehler in diesem Summierungsverfahren. Es ist $\lim_{h\to 0}\omega(\Phi, 2hK)=0$.

Betrachten wir ein Beispiel: Es sei K das Parallepiped

$$\begin{aligned}|L_1(x)|&=|b_{11}x_1+\cdots+b_{1s}x_s|<\lambda_1\\ |L_s(x)|&=|b_{s1}x_1+\cdots+b_{ss}x_s|<\lambda_s,\end{aligned}\tag{8}$$

wobei die Matrix $B=(b_{ik})$ nicht singulär sein soll. Es ist dann

$$c(K, g)=\frac{1}{|d(B)|}\prod_{j=1}^{s}\frac{\sin\pi l_j\lambda_j}{\pi l_j},$$

wobei l_j die j-te Komponente zur $(B^{-1})^*g$ ist. Es ist also:

$$\lim_{h\to 0}\sum_g\frac{1}{d^2(B)}\prod_{j=1}^{s}\frac{\sin^2\pi l_j\lambda_j h}{(\pi l_j)^2}A(\Phi, g)e(\langle gx\rangle)=\Phi(x)\tag{7''}$$

Für $s=1$, $\lambda_1=1$, $b_{11}=1$ ist es das Riemannsche Summierungsverfahren. Man kann noch ein anderes Summierungsverfahren entwickeln. Es gilt

SATZ 5

$$\lim_{h\to 0}\sum_g c(hK, g)A(\Phi, g)e(\langle gx\rangle)=\Phi(x).$$

Wir betrachten statt $\rho(K, t)$ jetzt $\psi(K, t)$, wo $\psi(K, t)$ in §1 definiert ist. Es ist

$$\frac{1}{c_0}\int_{E^s}\psi(K, t)\,dt=1,\tag{4'}$$

da ja

$$\frac{\psi(K,t)}{c_0}=1+\sum_g \frac{c(K,g)}{c_0}e\langle gt\rangle.$$

Es ist wieder

$$\psi(K,t)=0,$$

wenn $t\in E^s-\hat{K}$ ist, wo $\hat{K}=\bigcup_g (K+g)$ ist. Wir haben wieder

$$\sum_g c(hK,g)A(\Phi,g)e(\langle gx\rangle)=\int_{E^s}\Phi(x+t)\frac{\psi(hK,t)}{c_0(hK)}\,dt.$$

Es ist wieder

$$\int_{E^s}\Phi(x)\cdot\psi(hK,t)\,dt=\int^{*}\Phi(x)\psi(hK,t)\,dt,$$

wo sich die Integration rechts über $\hat{hK}=\hat{H}(K)$ erstreckt. Es ist also

$$\left|\int_{E^s}(\Phi(x+t)-\Phi(x))\frac{\psi(hK,t)}{c_0(hK)}\right|\leq\sup_{t\in\hat{H}}|\Phi(x+t)-\Phi(x)|.$$

Wenn wir voraussetzen, daß die Körper $K+g$ nicht übereinandergreifen, so gilt auch für $h\leq 1$, daß $hK+g$ nicht übereinandergreifen. Es ist also

$$\sup_{t\in\hat{H}}|\Phi(x+t)-\Phi(x)|\leq\omega(\Phi,hK).$$

Nehmen wir das Beispiel (8), so erhalten wir

$$\lim_{h\to 0}\sum_g\prod_{j=1}^{s}\frac{\sin l_j\lambda_j h}{d(B)l_j}A(\Phi,g)e(\langle gx\rangle)=\Phi(x)$$

Der Fall $s=1$ wurde von Lebesgue betrachtet.

Nehmen wir die Kugel $K:|x|\leq\rho$, so erhalten wir

$$\lim_{h\to 0}\sum_l\frac{J_{s/2}(|l|\,h\rho)}{|l|^{s/2}}A(\Phi,g)e(\langle gx\rangle)=\Phi(x).$$

§4

Im Zusammenhang mit dem zweiten Summierungsverfahren in §3 steht die folgende Ungleichung:

Es sei wieder Φ eine periodische integrierbare Funktion mit einer Fourierentwicklung §3(5). Es sei wieder K ein konvexer Körper mit Mittelpunkt O, so daß für alle g die Körper $K+g$ nicht übereinandergreifen. Dann gilt

SATZ 6. *Es ist stets*

$$\sum_{k,l} A(\Phi, k)\bar{A}(\Phi, l)c(K, k-1) \leq (1-V(K)) \sum_{g} |A(\Phi, g)|^2. \tag{1}$$

Wir können (1) noch verallgemeinern: Es sei außer Φ noch eine weitere periodische Funktion Φ_1 gegeben, dann gilt

$$\sum_{k,l} A(\Phi, k)\bar{A}(\Phi_1, l)c(K, k-l) \leq (1-V(K))\sqrt{S_k S_l}, \tag{2}$$

wo

$$S = \sum_{g} |A(\Phi, g)|^2 \quad \text{und} \quad S_1 = \sum_{g} |A(\Phi_1, g)|^2$$

ist.

Beweis. Nach der Parsevalschen Gleichung ist

$$J = \int_{E^s} \Phi(t)\bar{\Phi}_1(t)(\psi(K, t) - c(K, 0)) \tag{3}$$

gleich

$$\int_{E^s} \sum_{g,k,l} c(K, g)A(\Phi, k)A(\Phi_1, l)e(\langle l-k+g\rangle t)\, dt. \tag{4}$$

Nun ist

$$\int_{E^s} e(\langle l-k+g, t\rangle)\, dt = 0, \quad \text{außer für} \quad g = k-l,$$

also ist nach (4) und (3)

$$J = \sum_{k,l} c(K, k-l)A(\Phi, k)A(\Phi_1, l) \leq \sup_{t \in E^s} |\psi(K, t) - c(K, 0)|\, J_1,$$

wo

$$J_1 = \int_{E^s} |\Phi(t)\bar{\Phi}_1(t)|\, dt$$

ist. Nun ist $0 \leq \psi(K, t) \leq 1$, da die $K + g$ nicht übereinandergreifen. Weiter ist

$$J_1^2 \leq \int_{E^s} |\Phi(t)|^2\, dt \int_{E^s} |\Phi_1(t)|^2\, dt = SS_1,$$

damit ist alles gezeigt. Nehmen wir wieder ein Beispiel §3(8), so erhalten wir

$$\frac{1}{d(B)} \sum_{k,l} A(\Phi, k) A(\Phi_1, l) \prod_{j=1}^{s} \frac{\sin \pi l_j \lambda_j (k_j - l_j)}{\pi(k_j - l_j)} \leq \left(1 - \frac{\lambda_1 \lambda_s}{d(B)}\right) \sqrt{SS_1},$$

wenn die Ungleichungen

$$|L_1(g)| < 2\lambda_1, \ldots, |L_s(g)| < 2\lambda_s$$

nur die triviale Lösung $g = (0, \ldots, 0)$ hat. Für $s = 1$, $b_{11} = 1$, $\lambda_s = \lambda$ erhalten wir

$$\sum_{k,l} A(\Phi, k) \bar{A}(\Phi, l) \frac{\sin \pi(k - l)}{\pi(k - l)} \leq (1 - 2\lambda) \sqrt{SS_1},$$

Dabei muß $\lambda < 1/2$ sein, denn dann hat $|g| < 2\lambda$ nur die Lösung $g = O$. Ein anderes Beispiel ist $K: |x| < \rho$. Dann erhalten wir

$$\sum_{k,l} A(\Phi, k) \bar{A}(\Phi, l) \frac{J_{s/2}(\pi(k - l)\rho)}{|(k - l)\rho|^{s/2}} \leq (1 - V(J)\rho^s) \sqrt{SS_1},$$

wo $V(J)$ das Volumen der s-dimensionalen Einheitskugel ist, und aus $|g| < \rho$ folgt $g = O = (0 \ldots 0)$.

Mathematisches Institut,
Strudlhofgasse 4,
A-1090, Wien

Eingegangen den 20. August 1980.

Comment. Math. Helvetici **56** (1981) 83–102 0010-2571/81/001083-20$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Spectra of manifolds with small handles

I. CHAVEL[1] and E. A. FELDMAN[1]

To H. E. RAUCH, in memoriam

In this paper we consider a compact connected C^∞ Riemannian manifold M of dimension $n \geq 2$ and study the effect, on the spectrum of the associated Laplace–Beltrami operator Δ acting on functions, of adding a "small" handle to M.

The handles we consider are defined as follows: Fix two distinct points p_1, p_2 in M and for $\varepsilon > 0$ define

$B_\varepsilon \equiv:$ union of the open geodesic disks about p_1, p_2 of radius ε,
$\Omega_\varepsilon \equiv: M - \overline{B_\varepsilon}$,
$\Gamma_\varepsilon \equiv:$ common boundary of B_ε and Ω_ε,
$S_\varepsilon \equiv: (n-1)$-sphere in R^n of radius ε,
$S \equiv: S_1$.

For positive ε which is less than $\frac{1}{4}$ the injectivity radius of M and less than $\frac{1}{4}$ the distance from p_1 to p_2, let M_ε be a compact connected C^∞ Riemannian manifold with $\overline{\Omega_\varepsilon}$ isometrically imbedded in M_ε, and with a diffeomorphism

$$\Psi_\varepsilon : M_\varepsilon - \Omega_{2\varepsilon} \to [-1, 1] \times S$$

such that

$$C_\varepsilon \equiv: M_\varepsilon - \Omega_\varepsilon = \Psi_\varepsilon^{-1}[[-\tfrac{1}{2}, \tfrac{1}{2}] \times S].$$

We refer to such an M_ε as *obtained from M by adding the handle C_ε across Γ_ε*.

Denote the respective spectra of M, M_ε by

$$\text{spec}\,(M) \equiv: \{0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots\},$$

$$\text{spec}\,(M_\varepsilon) \equiv: \{0 = \sigma_0(\varepsilon) < \sigma_1(\varepsilon) \leq \sigma_2(\varepsilon) \leq \cdots\},$$

[1] Partially supported by N.S.F. Grant MCS 77-02757

where each distinct eigenvalue is repeated according to its multiplicity; and denote the associated theta functions by

$$\Theta(t) \equiv: \sum_{j=0}^{\infty} e^{-\lambda_j t}, \qquad \Theta_\varepsilon(t) \equiv: \sum_{j=0}^{\infty} e^{-\sigma_j(\varepsilon) t}.$$

Our interest in this paper is in determining whether the family of Riemannian manifolds M_ε can be chosen so that

$$\lim \sigma_j(\varepsilon) = \lambda_j \quad \text{as} \quad \varepsilon \downarrow 0 \tag{1}$$

for all $j = 1, 2, \ldots$.

Our first comment is that even if (1) is valid for all j, we do not expect that it be valid uniformly in j. In fact, when M is 2-dimensional the Minakshisundaram–Pleijel asymptotic expansion reads as [10, p. 45; 1, pp. 204–222]

$$\Theta(t) \sim \frac{A(M)}{4\pi t} + \frac{\chi(M)}{6} + O(t), \qquad \Theta_\varepsilon(t) \sim \frac{A(M_\varepsilon)}{4\pi t} + \frac{\chi(M_\varepsilon)}{6} + O(t), \tag{2}$$

as $t \downarrow 0$ (where $A(\cdot)$, $\chi(\cdot)$ denote area and Euler-characteristic, respectively). If (1) were valid uniformly in j then (2) would imply, by an easy argument, that $\chi(M_\varepsilon) = \chi(M)$ – an impossibility.

THEOREM A. *We always have*

$$\limsup \sigma_j(\varepsilon) \leq \lambda_j \quad \text{as} \quad \varepsilon \downarrow 0 \tag{3}$$

for all $j = 1, 2, \ldots$. *A necessary condition that* (1) *be valid for all* j *is that* $\nu(\varepsilon)$, *the lowest eigenvalue of* C_ε *with Dirichlet data on* Γ_ε, *satisfy*

$$\lim \nu(\varepsilon) = +\infty \quad \text{as} \quad \varepsilon \downarrow 0. \tag{4}$$

In particular, if for a fixed $l > 0$, *the ("long-thin") cylinder* $[-l/2, l/2] \times S_\varepsilon$ *is an isometrically imbedded open submanifold of* C_ε *for every* ε, *then* $\nu(\varepsilon) \leq \pi^2/l^2$ *and* (1) *cannot be satisfied for all* j.

To give a sufficient condition we require a definition,

DEFINITION 1. For any compact Riemannian manifold X of dimension

$n \geq 2$, we define the *isoperimetric constant* $c_1(X)$ by

$$c_1(X) = \inf_Y \frac{\{\mathrm{vol}_{n-1}(Y)\}^n}{\{\min(\mathrm{vol}_n(X_1), \mathrm{vol}_n(X_2))\}^{n-1}} \tag{5}$$

where $\mathrm{vol}_k(\cdot)$ denotes k-dimensional Riemannian measure, and Y ranges over all compact $(n-1)$-dimensional submanifolds of X which divide X into 2 open submanifolds X_1, X_2 each having boundary Y.

THEOREM B. *Assume there exists a constant* $c > 0$ *such that*

$$c_1(M_\varepsilon) \geq c > 0 \tag{6}$$

for all ε. *Then* (1) *is valid for all* $j = 1, 2, \ldots$.

That (6) is an indication of the "smallness" of C_ε is given by

LEMMA 1. *The sufficient condition* "(6) *for all* ε" *implies*

$$\mathrm{vol}_n(C_\varepsilon) = 0(\varepsilon^n), \tag{7}$$

$$\nu(\varepsilon) \geq \mathrm{const}/\varepsilon^2 \tag{8}$$

as $\varepsilon \downarrow 0$.

Indeed, one proves (7) by picking $Y = \Gamma_\varepsilon$, and $X_1 = C_\varepsilon$, $X_2 = \Omega_\varepsilon$.

In order to prove (8) from (6) and (7) let us recall, a definition and Cheeger's inequality for manifolds with boundary [4; 14].

DEFINITION. Let M be a compact manifold with boundary ∂M. We define the constant $h(M)$ by

$$h(M) = \inf_Y \frac{\mathrm{vol}_{n-1}(Y)}{\mathrm{vol}_n(X)}$$

where Y ranges over all compact $(n-1)$ dimension submanifolds such that $\partial M \cap Y = \emptyset$, which divide M into X and X' where $\partial \bar{X} \cap \partial M = \emptyset$.

Cheeger's argument [4] shows that $\lambda_1(M) \geq h^2/4$ where $\lambda_1(M)$ is the first eigenvalue for the Laplacian with Dirichlet boundary data.

Let $M = C_\varepsilon$, $\partial M = \Gamma_\varepsilon$ and X and Y submanifolds of M as in the above

definition then $\mathrm{vol}_{n-1}(Y) \geq c^{1/n}\ \mathrm{vol}_n(X)^{n-1/n}$ and

$$h(C_\varepsilon) \geq \inf_Y c^{1/n}\ \mathrm{vol}_n(X)^{-1/n} \geq k/\varepsilon$$

follow from (6) and (7). Therefore (8) follows from Cheeger's inequality.

We next remark that whereas the necessary condition for the validity of (1) for all j is a consequence of the max-min characterization of eigenvalues and thus best interpreted via vibration phenomena, the sufficient condition is obtained by working with the respective fundamental solutions of the heat equation on M, M_ε.

Most important is the interpretation of these fundamental solutions via Brownian motion, viz., if

$$p : M \times M \times (0, \infty) \to R$$

is the fundamental solution of the heat equation on M, then $p(x, y, t)$ is the probability density for a Brownian path in M starting at x at time 0 to be at y at time t. Of course one has a similar statement for

$$p_\varepsilon : M_\varepsilon \times M_\varepsilon \times (0, \infty) \to R,$$

the fundamental solution of the heat equation on M_ε. Similarly, if we let

$$q_\varepsilon : \Omega_\varepsilon \times \Omega_\varepsilon \times (0, \infty) \to R$$

denote the fundamental solution of the heat equation on Ω_ε with Dirichlet data on Γ_ε then $q_\varepsilon(x, y, t)$ is the probability density that a Brownian path starting at $x \in \Omega_\varepsilon$ at time 0 will be at $y \in \Omega_\varepsilon$ at time t *without having hit* Γ_ε *between time* 0 *and time* t. In particular, for x, y in

$$M_0 \equiv : M - \{p_1, p_2\}$$

(we now think of q_ε as vanishing on the complement of $\Omega_\varepsilon \times \Omega_\varepsilon$) $q_\varepsilon(x, y, t)$ is a decreasing function in ε, and

$$q_\varepsilon \leq p, \qquad q_\varepsilon \leq p_\varepsilon \tag{9}$$

on $M \times M \times (0, \infty)$, $M_\varepsilon \times M_\varepsilon \times (0, \infty)$, cf. [7; 12] for the application and details in Euclidean space, and [9] for the construction on general Riemannian manifolds.

Our final concern is that we *can* construct manifolds M, M_ε for which (6) is satisfied for all ε.

MAIN THEOREM. *Let M be a compact 2-dimensional Riemannian manifold, $K: M \to R$ its Gaussian curvature and $\tilde{M} = \{M - K^{-1}[0]\} \cup \{\text{int } K^{-1}[0]\}$. Then $\tilde{M}$ is open and dense in M. Given any two distinct points p_1, p_2 in $\tilde{M}$ then M_ε may be constructed so that there exists $c > 0$ for which (6) is valid for all ε. Thus M_ε may be constructed so that $A(M_\varepsilon) \to A(M)$ as $\varepsilon \downarrow 0$ and so that (1) is valid for all $j = 1, 2, \ldots$.*

The theorem suggests that to the question "Can you hear the shape of a drum?" [7] one should answer "For a compact 2-manifold – not really." For to determine the Euler-characteristic, via (2), by actually listening to its tones (square roots of the eigenvalues) one would have to know *a priori* that what is heard in fact approximates *all* the tones with *uniform* accuracy. Anything less could lead the listener astray in determining the Euler-characteristic.

We wish to thank our colleagues S. Kaplan and B. Randol for many helpful discussions, and A. Heller for help with Lemma 5.

This paper is dedicated to the inspiring memory of H. E. Rauch, whom both authors knew and admired as a friend, teacher, and mathematician.

§1. Proof of Theorem A

Denote the spectrum of Ω_ε with Dirichlet boundary data (distinct eigenvalues are repeated according to multiplicity) by

$$\text{spec}\,(\Omega_\varepsilon) \equiv: \{0 < \lambda_1(\varepsilon) < \lambda_2(\varepsilon) \leq \lambda_3(\varepsilon) \leq \cdots\}.$$

Then the max-min characterizations of the eigenvalues [5, Chap. VI] imply that $\lambda_j(\varepsilon)$ is an increasing function of ε, and the validity of the inequalities

$$\lambda_j(\varepsilon) \geq \lambda_{j-1}, \qquad \lambda_j(\varepsilon) \geq \sigma_{j-1}(\varepsilon) \tag{10}$$

for $j = 1, 2, \ldots$. Moreover, in [3] it was shown (cf. [13] for the case of domains in Euclidean space) that

$$\lambda_j(\varepsilon) \to \lambda_{j-1} \quad \text{as} \quad \varepsilon \downarrow 0 \tag{11}$$

for all $j = 1, 2, \ldots$. Then (10), (11) imply (3).

With these preliminaries, establishing the necessary condition is done as follows: Let the union of the spectra of C_ε, Ω_ε with Dirichlet data on Γ_ε be

denoted by

$$\operatorname{spec}(C_\varepsilon) \cup \operatorname{spec}(\Omega_\varepsilon) \equiv: \{0 < \mu_0(\varepsilon) \leq \mu_1(\varepsilon) \leq \cdots\}$$

where the eigenvalues have been re-listed in non-decreasing order and repeated according to multiplicity. Then a max-min argument [5, p. 408] implies

$$\sigma_j(\varepsilon) \leq \mu_j(\varepsilon) \tag{12}$$

for all $j = 0, 1, 2, \ldots$. Assume

$$\alpha \equiv: \liminf \nu(\varepsilon) \quad \text{as} \quad \varepsilon \downarrow 0$$

is finite, and let λ_k be the first eigenvalue of M which is strictly greater than α (in particular, $\lambda_{k-1} < \lambda_k$). Then for any ε for which we have

$$\nu(\varepsilon) < \lambda_k$$

we also have $\nu(\varepsilon) < \lambda_k \leq \lambda_{k+1}(\varepsilon)$, i.e.,

$$\nu(\varepsilon) \in \{\mu_0(\varepsilon), \ldots, \mu_k(\varepsilon)\},$$

which implies

$$\sigma_k(\varepsilon) \leq \mu_k(\varepsilon) \leq \max\{\nu(\varepsilon), \lambda_k(\varepsilon)\}.$$

Thus $\alpha < \lambda_k$ implies by (11) that

$$\liminf \sigma_k(\varepsilon) \leq \liminf \max\{\nu(\varepsilon), \lambda_k(\varepsilon)\} = \max\{\alpha, \lambda_{k-1}\} < \lambda_k$$

as $\varepsilon \downarrow 0$. It is therefore impossible that $\delta_k(\varepsilon) \to \lambda_k$ as $\varepsilon \downarrow 0$.

COROLLARY 1. *If*

$$\liminf \nu(\varepsilon) = \alpha < +\infty \quad \textit{as} \quad \varepsilon \downarrow 0,$$

and λ_k *is the first eigenvalue of* M *greater than* α, *then*

$$\liminf \sigma_k(\varepsilon) < \lambda_k \quad \textit{as} \quad \varepsilon \downarrow 0.$$

Remark 1. We note that (4) is also a necessary condition that $\Theta_\varepsilon(t) \to \Theta(t)$,

for any given $t>0$, $\varepsilon\downarrow 0$. Indeed, (12) implies that

$$\Theta_\varepsilon(t)\geq e^{-\nu(\varepsilon)t}+\sum_{j=1}^{\infty} e^{-\lambda_j(\varepsilon)t}.$$

But in [3] it was proved that the series on the right-hand side of the above inequality tends to $\Theta(t)$, uniformly on compact subsets of $(0,\infty)$, as $\varepsilon\downarrow 0$. That (4) is a consequence of $\Theta_\varepsilon(t)\to\Theta(t)$ is immediate.

§2. Proof of Theorem B

LEMMA 2. *Let dM, dM_ε denote the respective volume elements of M, M_ε (of course they agree on Ω_ε), and let f be any bounded measurable function compactly supported on M_0. Then*

$$\lim \int_{M_\varepsilon} p_\varepsilon(x, w, t)f(w)\, dM_\varepsilon = \int_M p(x, y, t)f(y)\, dM(y),$$

uniformly in $(x, t)\in$ compact subsets of $M_0\times(0,\infty)$, as $\varepsilon\downarrow 0$. In particular we have

$$\lim p_\varepsilon(x, y, t) = p(x, y, t) \quad as \quad \varepsilon\downarrow 0 \tag{13}$$

on $M_0\times M_0\times(0,\infty)$.

Proof. In [3] it was shown (cf. [13] for the case of domain in Euclidean space) that

$$\lim q_\varepsilon(x, y, t) = p(x, y, t) \quad \text{as} \quad \varepsilon\downarrow 0$$

uniformly on compact subsets of $M_0\times M_0\times(0,\infty)$. Let K be a compact subset of M_0 and pick ε sufficiently small so that Ω_ε contains K and the support of f. Then for $x\in K$, $t\in[a, b]\subseteq(0,\infty)$ we have

$$\begin{aligned}
&\left|\int_{M_\varepsilon}\{p_\varepsilon(x, w, t)-q_\varepsilon(x, w, t)\}f(w)\, dM_\varepsilon(w)\right| \\
&\leq \max|f|\int_{M_\varepsilon}\{p_\varepsilon(x, w, t)-q_\varepsilon(x, w, t)\}\, dM_\varepsilon(w) \\
&= \max|f|\left\{1-\int_{\Omega_\varepsilon} q_\varepsilon(x, y, t)\, dM(y)\right\} = \max|f|\int_M\{p(x, y, t)-q_\varepsilon(x, y, t)\}\, dM(y),
\end{aligned}$$

since

$$\int_{M_\varepsilon} p(x, w, t)\, dM_\varepsilon(w) = 1 = \int_M p(x, y, t)\, dM(y).$$

Thus

$$\left| \int_{M_\varepsilon} p_\varepsilon(x, w, t) f(w)\, dM_\varepsilon(w) - \int_M p(x, y, t) f(y)\, dM(y) \right|$$

$$\leq 2 \max |f| \int_M \{p(x, y, t) - q_\varepsilon(x, y, t)\}\, dM(y)$$

which goes to 0, uniformly in $(x, t) \in K \times [a, b]$, as $\varepsilon \downarrow 0$. Thus the lemma is proven.

To prove Theorem B we first reduce the problem to showing that for t bounded away from 0, $p_\varepsilon(z, w, t)$ is uniformly bounded above *independent of* ε.

Assume that this has in fact been accomplished. Then one has by the Sturm–Liouville expansion (cf. below) of p_ε, p that for any fixed $t > 0$,

$$\Theta_\varepsilon(t) - \Theta(t) = \int_{M_\varepsilon} p_\varepsilon(z, z, t)\, dM_\varepsilon(z) - \int_M p(x, x, t)\, dM(x)$$

$$= \int_{C_\varepsilon} p_\varepsilon(z, z, t)\, dM_\varepsilon(z) - \int_{B_\varepsilon} p(x, x, t)\, dM(x)$$

$$+ \int_{\Omega_\varepsilon} \{p_\varepsilon(x, x, t) - p(x, x, t)\}\, dM(x) \to 0 \quad \text{as} \quad \varepsilon \downarrow 0.$$

Indeed, the first two integrands are bounded and the volumes of C_ε, B_ε tend to 0. The convergence of the third integral follows from (13) and Lebesgue's dominated convergence theorem. Thus p_ε uniformly bounded independent of ε implies for $t > 0$,

$$\lim \Theta_\varepsilon(t) = \Theta(t) \quad \text{as} \quad \varepsilon \downarrow 0. \tag{14}$$

Finally assume there exists $k \geq 1$ such that $\liminf \sigma_k(\varepsilon) < \lambda_k$ as $\varepsilon \to 0$. Let ε_l be a sequence going to 0, with $\sigma_k(\varepsilon_l) \to \sigma_k < \lambda_k$ as $l \to \infty$. Then by (14), (3) and Fatou's lemma we have

$$\Theta(t) = \lim \Theta_{\varepsilon_l}(t) \geq \sum_{j=0}^{\infty} \liminf \exp(-\sigma_j(\varepsilon_l)t) = \sum_{j=0}^{\infty} \exp(-\limsup \sigma_j(\varepsilon_l)t)$$

$$= e^{-\sigma_k t} + \sum_{j \neq k} \exp(-\limsup \sigma_j(\varepsilon_l)t) > \Theta(t)$$

which implies a contradiction.

So to prove Theorem B we must bound p_ε above independently of ε. To do so we require some estimates of P. Li [8].

DEFINITION 2. Given a compact Riemannian manifold X of dimension $n \geq 2$ we define the *Sobolev constant of* X, $c_0(X)$, by

$$c_0(X) \equiv: \inf_f \left[\left\{ \int_M |\nabla f| \right\}^n \Big/ \inf_{\beta \in \mathbf{R}} \left\{ \int_M |f-\beta|^{n/(n-1)} \right\}^{n-1} \right]$$

where f ranges over the Sobolev space of functions with L^1-derivatives.

LEMMA (P. Li) 3. *Let* $v = \mathrm{vol}_n(X)$, $c_0 = c_0(X)$. *Then there exist constants depending only on n such that*

for any eigenfunction f with eigenvalue $\tau \neq 0$ *we have*

$$\|f\|_\infty^2 \leq \text{const} \begin{cases} \|f\|_2^2 (\tau^{n/2}/c_0) \exp\{\text{const}\,(c_0/v)^{2/n}/\tau\}, & n \geq 3 \\ \|f\|_2^2 (\tau^2 v/c_0^2) \exp\{\text{const}\, c_0/\tau v\}, & n = 2 \end{cases} \tag{15}$$

for the k^{th} *eigenvalue* τ_k *of* X *we have*

$$k \leq \text{const} \begin{cases} \{\tau_k (v/c_0)^{2/n}\}^{n-1} & n \geq 3 \\ \{\tau_k v/c_0\}^2 & n = 2 \end{cases} \tag{16}$$

for all $k = 1, 2, \ldots$.

Before turning to the proof of Theorem B we remark (as in [8]) that the argument of [2, Section 3], when applied to compact X without boundary, yields

$$c_1(X) \leq c_0(X) \leq 2c_1(X). \tag{17}$$

Also, by considering arbitrarily small geodesic disks, one has under *all* circumstances

$$c_1(X) \leq n^{n-1} \mathrm{vol}_{n-1}(S). \tag{18}$$

We now prove Theorem B; recall that we must establish an upper bound on p_ε which is independent of ε. Assume (6). Then Lemma 1 implies

$$\lim \mathrm{vol}_n(M_\varepsilon) = \mathrm{vol}_n(M) \quad \text{as} \quad \varepsilon \downarrow 0. \tag{19}$$

Also, $\sigma_1(\varepsilon)$ is bounded away from zero, by either using Cheeger's inequality [4]

with (19) or by using (17), (19), and (16) for $k=1$. Thus we have that $\{\sigma_1(\varepsilon), \mathrm{vol}_n(M_\varepsilon), c_0(M_\varepsilon)\}$ are all restricted to a compact subset of $(0,\infty)$. Then there exist constants *independent of* ε for which Li's estimates now read as

$$\|f\|_\infty^2 \leq \text{const}\, \|f\|_2^2 \begin{cases} \tau^{n/2} & n \geq 3 \\ \tau^2 & n=2 \end{cases} \tag{15'}$$

$$\tau_k \geq \text{const} \begin{cases} k^{1/(n-1)} & n \geq 3 \\ k^{1/2} & n=2. \end{cases} \tag{16'}$$

Now fix $t>0$ and let $\{\Phi_j(\varepsilon)\}$ be an orthonormal basis of $L^2(M_\varepsilon)$ consisting of eigenfunctions corresponding respectively to $\{\sigma_j(\varepsilon)\}$. Then the eigenfunction expansion of p_ε is given by, and satisfies,

$$p_\varepsilon(z,w,t) = \sum_{j=0}^{\infty} e^{-\sigma_j(\varepsilon)t} \Phi_j(\varepsilon)(z)\Phi_j(\varepsilon)(w)$$

$$\leq \sum_{j=0}^{\infty} e^{-\sigma_j(\varepsilon)t} \|\Phi_j(\varepsilon)\|_\infty^2.$$

We proceed with estimate for the case $n=2$ as this is the situation in which we will construct our explicit examples (the argument for $n>2$ is similar). From (15') we have

$$p_\varepsilon(z,w,t) \leq \text{const} \left\{ 1 + \sum_{j=1}^{\infty} \sigma_j^2(\varepsilon) e^{-\sigma_j(\varepsilon)t} \right\}$$

with the constant independent of ε. Now (16') implies the existence of a positive integer J, independent of ε, such that for all $j \geq J$ we have

$$\sigma_j^5(\varepsilon) e^{-\sigma_j(\varepsilon)t} \leq 1.$$

Then (16') implies that

$$\sum_{j=1}^{\infty} \sigma_j^2(\varepsilon) e^{-\sigma_j(\varepsilon)t} \leq \sum_{j<J} \sigma_j^2(\varepsilon) + \text{const} \sum_{j \geq J} j^{-3/2}$$

which is bounded above, independently of ε, by (3).

This concludes the proof of Theorem B.

§3. The construction of M_ε for the main theorem

Let M be 2-dimensional and $p \in \tilde{M}$, i.e., either $K(p) \neq 0$ or K vanishes identically on some neighborhood of p. To p we associate a number $\alpha(p)$ with the following list of properties:

(i) α will be less than the convexity radius of M (in particular, it is less than $\frac{1}{2}$ the injectivity radius of M). If K, the Gauss curvature of M, has maximum equal to κ then α will be chosen so that it is also less than $\pi/2\sqrt{\kappa}$.

(ii) Set

$$B(p; r) \equiv: \text{metric disk about } p \text{ of radius } r.$$

Then we require that K either vanishes identically on $B(p; \alpha)$ or never vanishes on $B(p; \alpha)$.

Should K never vanish on $B(p; \alpha)$ then α will be sufficiently small so that

$$\inf |K(q)| > (\tfrac{2}{3}) \sup |K(q)| \tag{20}$$

where q ranges over $B(p; \alpha)$.

(iii) Let dA denote the Riemannian element of area and, as in the introduction, $A(\cdot)$ denote the area. Then we require that

$$A(B(p; \alpha)) < A(M)/8 \tag{21}$$

$$\iint_{B(p;\alpha)} K \, dA < \pi/2. \tag{22}$$

In particular for

$$K^+ \equiv: \max\{K, 0\}$$

we have

$$\iint_{B(p;\alpha)} K^+ \, dA < \pi/2. \tag{23}$$

Remark 2. In a moment we shall change the Riemannian metric in a compact subset of $B(p; \alpha)$, when $K(p) \neq 0$, such that the new Gauss curvature does not change sign. The Gauss–Bonnet formula implies that the left-hand side of (22) does *not* change, hence (23) remains valid in the new Riemannian metric.

If $K(p) \neq 0$, then for every $\varepsilon \in (0, \alpha/2)$ we introduce a new Riemannian metric on $B(p; \varepsilon)$. The details will be given for $K(p) > 0$, as the case $K < 0$ is similar. Let

$$k_1 \equiv: \inf K, \qquad k_2 \equiv: \sup K$$

on $B(p; \alpha)$. Then $k_1/k_2 > \frac{2}{3}$ implies that for all r satisfying $0 < r < \alpha$ we have

$$\cos \sqrt{k_1}\, r < 2\left\{\frac{\sin \sqrt{k_2}\, r}{\sqrt{k_2}\, r} - \frac{1}{2}\right\}.$$

Introduce geodesic polar coordinates (r, θ) about p and write the given Riemannian metric as

$$ds^2 = dr^2 + \eta^2(r, \theta)\, d\theta^2.$$

Then η satisfies Jacobi's equation

$$\frac{\partial^2}{\partial r^2}\eta + K\eta = 0$$

with initial data

$$\eta(0, \theta) = 0, \qquad (\partial\eta/\partial r)(0, \theta) = 1.$$

Standard Sturmian arguments imply

$$\frac{\partial \eta}{\partial r}(r, \theta) \leq \cos \sqrt{k_1}\, r < 2\left\{\frac{\sin \sqrt{k_2}\, r}{\sqrt{k_2}\, r} - \frac{1}{2}\right\}$$

$$\leq 2\left\{\frac{\eta(r, \theta)}{r} - \frac{1}{2}\right\} = \frac{\eta(r, \theta) - r/2}{r/2}$$

i.e.,

$$\frac{\partial \eta}{\partial r} \leq \frac{\eta - r/2}{r/2}. \tag{24}$$

Geometrically, (24) implies that for each fixed θ, the tangent line to the curve $y = \eta(x, \theta)$ (in the (x, y)-plane) at $x = r$ intersects the line $y = x$ for some $x(r, \theta)$ satisfying $r/2 < x(r, \theta) < r$.

Given ε satisfying $0 < \varepsilon < \alpha/2$, set $x_1 \equiv: x(3\varepsilon/4, \theta)$ and replace $y = \eta(x, \theta)$ for

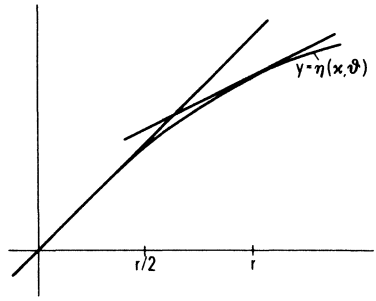


Figure 1

$0 \leq x \leq \varepsilon$ by $y = \tilde{\eta}(x, \theta)$ where

$$\tilde{\eta}(x, \theta) = x \quad \text{for} \quad 0 \leq x \leq x_1,$$

$\tilde{\eta}(x, \theta)$ is given by the tangent line to $y = \eta(x, \theta)$ at $x = 3\varepsilon/4$ for $x_1 \leq x \leq 3\varepsilon/4$, $\tilde{\eta}(x, \theta) = \eta(x, \theta)$ for $3\varepsilon/4 \leq x \leq \alpha$.

Now smooth $\tilde{\eta}$ to η_ε which is a C^∞ function in (r, θ) and which satisfies

$$\begin{aligned} &\eta_\varepsilon(r, \theta) = r && 0 \leq r \leq 5\varepsilon/16 \\ &\eta_\varepsilon(r, \theta) = \eta(r, \theta) && \varepsilon \leq x \leq \alpha \\ &\frac{\partial \eta_\varepsilon}{\partial r} > 0, \quad \frac{\partial^2 \eta_\varepsilon}{\partial r^2} \leq 0 && 0 \leq x \leq \alpha. \end{aligned} \tag{25}$$

Finally replace

$$ds^2 = dr^2 + \eta^2(r, \theta)\, d\theta^2$$

on $B(p; \alpha)$, by

$$ds_\varepsilon^2 = dr^2 + \eta_\varepsilon^2(r, \theta)\, d\theta^2.$$

Then the new metric is flat on the geodesic disk of radius $5\varepsilon/16$, has non-negative Gauss curvature on $B(p; \alpha)$, and agrees with the original metric on $B(p; \alpha) \cap \overline{\Omega_\varepsilon}$. One sees easily that the smoothing may be chosen so that (21) can be replaced by

$$A(B(p; \alpha)) < A(M)/6. \tag{21'}$$

A similar argument can be carried out when $K(p_1) < 0$.

We are now ready to define M_ε. Given $p_1, p_2 \in \tilde{M}$ fix α to be less than $\alpha(p_1)$, $\alpha(p_2)$ and less than $\frac{1}{2}$ the distance from p_1 to p_2. For each $i = 1, 2$, if $K(p_i) = 0$ leave well enough alone. If $K(p_i) \neq 0$ then for every $\varepsilon < \alpha/2$ introduce the change of metric in $B(p_i; \varepsilon)$ just described. Call the new Riemannian manifold M_ε^*.

Now attach the tube

$$T \equiv: [-\varepsilon/4, \varepsilon/4] \times S_{\varepsilon/4}$$

to M_ε^* "at right angles" to the geodesic circles about p_1, p_2 of radius $\varepsilon/4$, identifying $\{-\varepsilon/4\} \times S_{\varepsilon/4}$ with the geodesic circle about p_1, and $\{\varepsilon/4\} \times S_{\varepsilon/4}$ with that about p_2, of radius $\varepsilon/4$. Put differently, in the polar coordinate system about each p_i, replace

$$\eta_\varepsilon(r, \theta) = r \qquad 0 \leq r \leq 5\varepsilon/16$$

with

$$\zeta_\varepsilon(r, \theta) = \begin{cases} \varepsilon/4 & 0 \leq r \leq \varepsilon/4 \\ r & \varepsilon/4 \leq r \leq 5\varepsilon/16 \end{cases}$$

and identify the two circles $\{0\} \times S_{\varepsilon/4}$.

The resulting manifold with "creased" Riemannian metric will be our M_ε and we shall estimate $C_1(M_\varepsilon)$ for this "creased" metric from below. Once we, in fact, verify the existence of c for which (6) is valid for all ε, it is easy to smooth the "crease," with non-positive curvature near the crease, such that (6) is valid with c replaced by $c/2$ for all ε.

§ 4. Estimating the isoperimetric constant

LEMMA (F. Fiala [6, p. 336; 11, p. 12]) 3. *Let M be a complete Riemannian surface with Riemannian measure dA, Gauss curvature K, $K^+ = \max\{K, 0\}$, and D a simply connected domain in M of area A with smooth boundary of length L. Then*

$$\iint_D K^+ \, dA < 2\pi$$

implies

$$L^2 - 4\pi A + 2A \iint_D K^+ \, dA \geq 0.$$

COROLLARY 1. *In the above,*

$$\iint_D K^+ \, dA < \pi$$

implies

$$L^2 \geq 2\pi A.$$

LEMMA (S. T. Yau [4, p. 489]) 4. *For a compact n-dimensional Riemannian manifold M, to evaluate $c_1(M)$ it suffices to let Y range over those compact $(n-1)$-manifolds which separate M into connected open submanifolds X_1, X_2.*

Remark 3. In [14] Yau proves this fact for Cheeger's isoperimetric constant. However his induction argument and the Minkowski inequality (Triangle inequality if $\dim M = 2$) immediately yields the above lemma.

In what follows $L_\varepsilon(\cdot)$, dA_ε, $A_\varepsilon(\cdot)$, K_ε will denote length, area element, area, and Gauss curvature of M_ε. We also write $K_\varepsilon^+ = \max\{K_\varepsilon, 0\}$.

To start with our estimate of $c_1(M_\varepsilon)$ from below, let γ be a compact 1-manifold imbedded in Ω_ε, separating M_ε into M_1, M_2, and assume $C_\varepsilon \subseteq M_1$. Set $M_1^* = (M_1 - C_\varepsilon) \cup B_\varepsilon$, where B_ε is, as originally defined, the union of the open geodesic disks of radius ε in M about p_1, p_2. One easily checks that $A_\varepsilon(C_\varepsilon)/A(B_\varepsilon)$ is bounded away from $0, +\infty$ independently of ε. Thus

$$\begin{aligned} L_\varepsilon^2(\gamma)/\min(A_\varepsilon(M_1), A_\varepsilon(M_2)) &\geq \text{const}\, L^2(\gamma)/\min(A(M_1^*), A(M_2)) \\ &\geq \text{const}\, c_1(M) > 0 \end{aligned}$$

which is independent of ε.

Also if γ is any compact imbedded 1-manifold of length $\geq \alpha$ separating M_ε into M_1, M_2, then

$$L_\varepsilon^2(\gamma)/\min(A_\varepsilon(M_1), A_\varepsilon(M_2)) \geq \alpha^2/A_\varepsilon(M_\varepsilon) \geq \text{const} > 0$$

by (19).

Thus our task is to estimate

$$L_\varepsilon^2(\gamma)/\min(A_\varepsilon(M_1), A_\varepsilon(M_2))$$

where γ ranges over compact imbedded 1-manifolds separating M_ε into connected M_1, M_2 and such that

$$L_\varepsilon(\gamma) < \alpha, \qquad \gamma \cap \operatorname{int}(C_\varepsilon) \neq \emptyset.$$

Since $\varepsilon < \alpha/2$ we immediately have

$$\gamma \subseteq M_\varepsilon - \overline{\Omega_\alpha}.$$

We will always have one of the domains, say M_1, contained in $M_\varepsilon - \overline{\Omega_\alpha}$ and (19), (21′) allow us to assume $A_\varepsilon(M_1) < A_\varepsilon(M_\varepsilon)/2$; so we are always estimating

$$L_\varepsilon^2(\gamma)/A_\varepsilon(M_1)$$

from below.

Since $M_\varepsilon - \overline{\Omega_\alpha}$ is topologically a cylinder, we have that γ is an imbedded circle bounding a disk $M_1 \subseteq M_\varepsilon - \overline{\Omega_\alpha}$ or γ is a pair of imbedded circles bounding a cylinder $M_1 \subseteq M_\varepsilon - \overline{\Omega_\alpha}$. In the first case, then by smoothing the "crease" with non-positive curvature we would have (from Remark 2)

$$\int_{M_\varepsilon - \bar{\Omega}_\alpha} K_\varepsilon^+ \, dA_\varepsilon < \pi$$

and therefore

$$L_\varepsilon^2(\gamma) \geq 2\pi A_\varepsilon(M_1)$$

in the approximating metric, which implies the inequality for our "creased" metric. So we need only consider the second case.

First let $d(,)$ denote distance in M_ε^* and define for $i = 1, 2$

$$\Gamma_i(t) \equiv: \{q \in M_\varepsilon^* : d(q, p_i) = t\}$$

$$W_i \equiv: \{q \in M_\varepsilon^* : \varepsilon/4 \leq d(q, p_i) < \alpha\}.$$

We are now considering $\gamma = \gamma_1 \cup \gamma_2$ where γ_1, γ_2 are imbedded circles bounding a cylinder M_1 in $M_\varepsilon - \overline{\Omega_\alpha}$. We think of γ_i as "closest" to $\Gamma_i(\alpha)$. First we shall assume that γ_i does not cross $\Gamma_i(\varepsilon/4)$ for either $i = 1, 2$, and then reduce the general case (viz., where at least one γ_i crosses $\Gamma_i(\varepsilon/4)$) to this one. The assumption that neither γ_i crosses $\Gamma_i(\varepsilon/4)$ involves three essential possibilities: (i) $\gamma_1, \gamma_2 \subseteq T$, (ii) $\gamma_1 \subseteq W_1$, $\gamma_2 \subseteq W_1 \cup T$, (iii) $\gamma_1 \subseteq W_1$, $\gamma_2 \subseteq W_2$.

(i) If γ_1, $\gamma_2 \subseteq T$ then $M_1 \subseteq T$, and $L_\varepsilon(\gamma_i) \geq \pi\varepsilon/2$, $A_\varepsilon(M_1) \leq \pi\varepsilon^2/4$, which implies

$$L_\varepsilon^2(\gamma) \geq 4\pi A_\varepsilon(M_1).$$

(ii) If $\gamma_1 \subseteq W_1$, $\gamma_2 \subseteq W_1 \cup T$ then $M_1 \subseteq T \cup (M_1 \cap W_1)$, which implies

$$A_\varepsilon(M_1) \leq \pi\varepsilon^2/4 + A_\varepsilon(M_1 \cap W_1).$$

Now think of γ_1 as bounding a disk $D \subseteq B(p_1; \alpha) \subseteq M_\varepsilon^*$. Then

$$A_\varepsilon(D) = \pi(\varepsilon/4)^2 + A_\varepsilon(M_1 \cap W_1) \geq A_\varepsilon(M_1)/4,$$

and by Remark 2 and Fiala's inequality we therefore have

$$L_\varepsilon^2(\gamma) \geq L_\varepsilon^2(\gamma_1) \geq 2\pi A_\varepsilon(D) \geq (\pi/2) A_\varepsilon(M_1).$$

This argument, of course, covers the possibility: $\gamma_2 \subseteq W_2$, $\gamma_1 \subseteq W_2 \cup T$.

(iii) If $\gamma_1 \subseteq W_1$, $\gamma_2 \subseteq W_2$ then our argument is similar to the one just given. Let D_i correspond to γ_i as D corresponded to γ_1 in (ii). Then we have

$$\begin{aligned} L_\varepsilon^2(\gamma) \sum_i L_\varepsilon^2(\gamma_i) &\geq \sum_i 2\pi A_\varepsilon(D_i) \\ &= 2\pi \sum_i \{\pi\varepsilon^2/16 + A_\varepsilon(M_1 \cap W_i)\} \\ &> \sum_i \{\pi\varepsilon^2/8 + A_\varepsilon(M_1 \cap W_i)\} = \pi A_\varepsilon(M_1). \end{aligned}$$

In summary, when neither γ_i crosses $\Gamma_i(\varepsilon/4)$ we have

$$L_\varepsilon^2(\gamma) \geq (\pi/2) A_\varepsilon(M_1). \tag{26}$$

For the general case, we shall assume γ_1 crosses $\Gamma_1(\varepsilon/4)$ transversally with an even number of intersections, and show that we can replace γ_1 with $\tilde{\gamma}_1$ having fewer intersections (therefore, ultimately no intersections), shorter length, and enclosing with γ_2 larger area. Thus (26) will remain valid in the general case.

So we now have γ_1 crossing $\Gamma_1(\varepsilon/4)$ transversally. We shall think of $M_\varepsilon - \overline{\Omega_\alpha}$ as the cylinder $R^2 - \{0\}$, with γ_1 and $\Gamma_1(\varepsilon/4)$ winding about 0 once, let

$$D \equiv: \text{component of } R^2 - \{\gamma_1\} \text{ not containing } M_1,$$

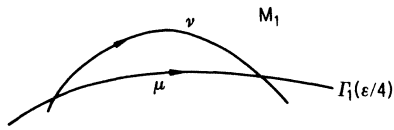


Figure 2

assume for convenience that D contains 0, and let

$$\{D_1, \ldots, D_k\} \equiv: \text{components of } D - \Gamma_1(\varepsilon/4).$$

We say that a component D_l is *simple* if its boundary consists of the union of 2 smooth arcs ν and μ, with ν part of γ_1 and μ part of $\Gamma_1(\varepsilon/4)$.

Given a simple D_l we have that either $D_l \subseteq W_1$ or $D_l \subseteq W_2 \cup T$, with $\nu \subseteq W_1$ or $\nu \subseteq W_2 \cup T$ respectively. However, in either case

$$L(\nu) > L(\mu)$$

– note that we are using here the hypothesis that α was picked to be less than the convexity radius of M. Then replace γ_1 by $\tilde{\gamma}_1$ by first replacing ν by μ, then sliding μ a drop to the side of $\Gamma_1(\varepsilon/4)$ opposite to D_l, and, finally, smoothing the corners. Define $\tilde{\gamma} = \tilde{\gamma}_1 \cup \gamma_2$. Then $L_\varepsilon(\tilde{\gamma}) \leq L_\varepsilon(\gamma)$, and $\tilde{M}_1$, the domain in $M_\varepsilon - \overline{\Omega_\alpha}$ bounded by $\tilde{\gamma}$, contains $M_1 \cup D_l$ which implies $A_\varepsilon(\tilde{M}_1) \geq A_\varepsilon(M_1)$. Thus

$$L_\varepsilon^2(\gamma)/A_\varepsilon(M_1) \geq L_\varepsilon^2(\tilde{\gamma})/A_\varepsilon(\tilde{M}_1).$$

The curve $\tilde{\gamma}$ has the same properties as γ, so we may repeat the argument just given until we are left with a curve not intersecting $\Gamma_1(\varepsilon/4)$. In this last case we already have the estimate (26). Thus the last thing for us to verify is that given γ, a simple D_l exists.

LEMMA 5. *Let Γ, Λ be two simply closed smooth curves in $R^2 - \{0\}$ which wind about the origin once and which meet each other transversally. Let G be the component of $R^2 - \Lambda$ containing 0, and let $\{G_1, \ldots, G_r\}$ be the components of $G - \Gamma$. Then there exists a component G_s whose boundary consists of two smooth arcs; one a part of Γ and one a part of Λ.*

Proof. We can assume Γ is the unit circle which Ω divides into a finite number of arcs alternately belonging to G and $R^2 - (G \cup \Lambda)$.

Let G_0 be the component of $G - \Gamma$ containing 0, and consider the segments

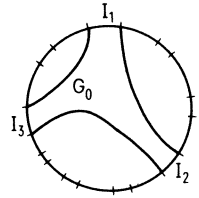


Figure 3

$I_1, \ldots, I_{l_1}$ of $\overline{G_0} \cap \Gamma$. (Of course if $l_1 = 1$ then G_0 is simple. So assume otherwise.) *To each* I_{j_1}, $j_1 = 1, \ldots, l_1$ associate the component G_{j_1} of $G - \Gamma$ distinct from G_0 having I_{j_1} as part of its boundary. $G_1, \ldots, G_{l_1}$ are all distinct for otherwise Λ would not be simple by the Jordan curve theorem.

If there exists G_{j_1} having only I_{j_1} for the intersection of ∂G_{j_1} with Γ then we are done, for then G_{j_1} is then simple. So assume the opposite. Then to each $j_1 = 1, \ldots, l_1$ associate the segments $I_{j_1 1}, \ldots, J_{j_1 l_2}$ of $\overline{G_{j_1}} \cap \Gamma - I_{j_1}$ and to each segment $I_{j_1 j_2}$ associate the component of $G - \Gamma$ distinct from G_{j_1}, $G_{j_1 j_2}$, having $I_{j_1 j_2}$ as part of its boundary. Again, the collection $\{G_{j_1 j_2}\}$ are all distinct.

By continuing this process, if necessary, we exhaust all the arcs of $G - \Gamma$ and the process stops. But then every component of the last step is simple.

This concludes the proof of Lemma 5 and, with it, the proof of the main theorem.

REFERENCES

[1] BERGER, M., GAUDUCHON, P., and MAZET, E., *Le Spectre d'une Variété Riemannienë*, Lecture Notes in Mathematics, #194 (1971), Springer-Verlag, New York.

[2] BOMBIERI, E., *Theory of Minimal Surfaces and a Counterexample to the Bernstein Conjecture in High Dimensions*, Lecture Notes (1970), Courant Institute, N.Y.U., New York.

[3] CHAVEL, I. and FELDMAN, E. A., *Spectra of domains on compact manifolds*, J. of Functional Anal., *30* (1978), 192–222.

[4] CHEEGER, J., *A lower bound for the smallest eigenvalue of the Laplacian, Problems in Analysis, A symposium in honor of S. Bochner*, Princeton University Press, Princeton, N.J., (1970), 195–199.

[5] COURANT, R. and HILBERT, D., *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.

[6] FIALA, F., *Le probleme des isoperimètres sur les surfaces ouvertes a courbure positive*, Comment. Math. Helv., *13* (1941), 293–346.

[7] KAC, M., *Can you hear the shape of a drum?*, Amer. Math. Monthly, *73* (1966), 1–23.

[8] LI, P., *On the Sobolev constant and the p-spectrum of a compact Riemannian manifold*, preprint.

[9] MCKEAN, H. P., *Stochastic Integrals*, Academic Press, New York.

[10] MCKEAN, H. P. and SINGER, I. M., *Curvature and the eigenvalues of the Laplacian*, J. of Diff. Geometry, *1* (1967), 43–70.

[11] OSSERMAN, R., *Bonnerson-style isoperimetric inequalities*, Amer. Math. Monthly, (1979), 1–29.

[12] PORT, S. C. and STONE, C. J., *Brownian Motion and Classical Potential Theory*, Academic Press, New York, 1978.
[13] RAUCH, J. and TAYLOR, M., *Potential and scattering theory on wildly perturbed domains*, J. of Functional Anal., *18* (1975), 27–59.
[14] YAU, S. T., *Isoperimetric constants and the first eigenvalue of a compact manifold*, Ann. Sci. Ecole Norm. Sup. (4° Sci), *8* (1975), 487–507.

City College of C.U.N.Y.
New York, N.Y. 10031
and
Graduate Center of C.U.N.Y.
New York, N.Y. 10036

Received May 28, 1980

Comment. Math. Helvetici **56** (1981) 103–122 0010–2571/81/001103–20$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Sur les invariants d'homotopie rationnelle liés à la L.S. catégorie

JEAN-MICHEL LEMAIRE ET FRANÇOIS SIGRIST

§0. Introduction

Le but de cet article est de montrer l'utilité du modèle minimal homologique ("modèle de Quillen") dans la détermination de la catégorie de Lusternik–Schnirelmann rationnelle. Nous avons choisi un espace X pour lequel la détermination de cat (X) nécessite le calcul explicite d'un invariant de Hopf délicat (au sens de Hilton–Berstein). La simplicité du calcul à l'aide du modèle minimal homologique nous a paru convaincante, d'autant plus que les estimations classiques de cat (X) sont inopérantes pour l'espace en question.

Un résultat prélable, que le modèle minimal homologique fournit rapidement et facilement, est la détermination de la catégorie des espaces formels. Nous en prenons prétexte pour déployer simultanément les propriétés principales des espaces formels et une panoplie d'estimations de cat (X), tant par défaut que par excès. Un contre-exemple récent [5] ne laisse qu'une conjecture en suspens: un espace de catégorie n est-il toujours un cône itéré n fois? L'équivalence de cette conjecture avec une propriété géométrique simple nous incline vers l'affirmative.

Ce travail est le fruit d'une longue collaboration, encouragée par le Troisième cycle romand de mathématiques, que nous tenons à remercier ici.

Pour faciliter la lecture, nous donnons ici un résumé télégraphique de ce travail.

§1. Topologie algébrique.	Catégorie de Lusternik–Schnirelmann, définitions équivalentes. Estimations par défaut, contrôle de qualité.
§2. Modèle de Quillen.	Description et premières propriétés. Nouveaux invariants estimant cat (X) par excès. Contrôle de qualité.
§3. Espaces formels.	Description axiomatique. Etude des applications formalisables. Coïncidence des invariants.
§4. Exemple explicite.	

§1. Invariants

Nous supposerons dans cet article, avec les abus de langage traditionnels, que les espaces topologiques sont rationnels, CW et nilpotents sinon simplement connexes.

La catégorie de Lusternik–Schnirelmann d'un espace X est un invariant homotopique entier dont la définition est:

$\operatorname{cat}(X) \leqslant n \Leftrightarrow X$ est réunion de $n+1$ ouverts contractiles dans X.

En approche directe, la détermination de $\operatorname{cat}(X)$ est malaisée; quelques échantillons utiles cependant:

LEMME (1.1). $\operatorname{cat}(X) \leqslant 1 \Leftrightarrow X$ *est un co-H-espace*

LEMME (1.2). *Soit* Cf *le cône de* $f: X \to Y$. *Alors* $\operatorname{cat}(Cf) \leqslant \operatorname{cat}(Y)+1$

LEMME (1.3). *Si* $\operatorname{cat}(X) \leqslant n$, *tout cup-produit de longueur* $n+1$ *dans* $\tilde{H}^*(X)$ *est nul.*

Les progrès décisifs dans l'estimation de la catégorie de Lusternik–Schnirelmann s'appuient sur deux interprétations dues respectivement à G. Whitehead [17] et T. Ganea [6].

Première interprétation. Notons $\Delta: X \to X^{n+1}$ la diagonale, $i: T^n X \to X^{n+1}$ l'inclusion du bouquet garni, et (pour la suite) $q: X^{n+1} \to X^{(n+1)}$ le quotient. Alors $\operatorname{cat}(X) \leqslant n \Leftrightarrow$ il existe $\phi: X \to T^n X$ factorisant homotopiquement $\Delta: i \circ \phi \simeq \Delta$.

Deuxième interprétation. Considérons la construction de Milnor de $B\Omega X$, classifiant du H-espace ΩX. Il en résulte une filtration de $B\Omega X$ par ses espaces projectifs successifs:

$$\Sigma\Omega X = \Omega X P(1) \to \Omega X P(2) \to \cdots \to \Omega X P(n) \to \cdots \to B\Omega X \simeq X.$$

Notons abusivement $\gamma: \Omega X P(n) \to X$ la conversion de l'application $\Omega X P(n) \to B\Omega X$ en fibration de base X. Pour la suite également désignons par $\kappa: X \to C\gamma$ la cofibre de γ. Alors $\operatorname{cat}(X) \leqslant n \Leftrightarrow$ la fibration γ a une section.

Outre les références originales, le lecteur consultera avec profit un article de Gilbert [7], bien documenté, qui contient la démonstration de l'équivalence des deux interprétations ci-dessus; nous nous contenterons de signaler que cette

équivalence résulte de l'existence d'un carré

$$\begin{array}{ccc} \Omega XP(n) & \xrightarrow{\gamma} & X \\ \downarrow & & \downarrow \Delta \\ T^n X & \xrightarrow{i} & X^{n+1} \end{array}$$

cartésien à homotopie près (homotopy cartesian).

Avec les notations introduites, définissons maintenant:

$c(X)$ = "cup-length" de X:

$$c(X) \leqslant n \Leftrightarrow (q \circ \Delta)_* = 0.$$

$e(X)$ = longueur d'Eilenberg–Moore de X:

$$e(X) \leqslant n \Leftrightarrow \kappa_* = 0.$$

w cat (X) = catégorie faible de X:

$$w \text{ cat } (X) \leqslant n \Leftrightarrow q \circ \Delta \simeq *$$

Gw cat (X) = catégorie faible de Ganea de X:

$$Gw \text{ cat } (X) \leqslant n \Leftrightarrow \kappa \simeq *.$$

L'examen des définitions nous donne immédiatement la

PROPOSITION (1.4).

$$c(X) \leqslant \begin{matrix} w \text{ cat } (X) \\ e(X) \end{matrix} \leqslant Gw \text{ cat } (X) \leqslant \text{cat } (X).$$

Ces différents invariants peuvent coïncider dans des cas de stabilité bien précis. Par exemple:

LEMME (1.5). *Soit* X $(r-1)$*-connexe de dimension* $\leqslant 2(k+1)r-2$. *Alors*

$$c(X) \leqslant k \Rightarrow w \text{ cat } (X) \leqslant k$$

$$e(X) \leqslant k \Rightarrow Gw \text{ cat } (X) \leqslant k.$$

Démonstration. $(q \circ \Delta)_* = 0$ équivaut à $\Sigma(q \circ \Delta) \simeq *$. C'est un résultat général d'homotopie rationnelle que nous démontrerons ci-après (Proposition 3.12). Dans ce cas particulier, il s'agit du résultat de Toomer [16]: $c(X) = \text{conil}(\Sigma X)$. Les restrictions dimensionnelles sont celles du théorème de suspension de Freudenthal, nous pouvons donc conclure que $(q \circ \Delta) \simeq *$. Pour la deuxième assertion, le raisonnement est identique.

LEMME (1.6). *Soit* X $(r-1)$*-connexe de dimension* $\leqslant (k+2)r-2$. *Alors*

$w\,\text{cat}(X) \leqslant k \Rightarrow \text{cat}(X) \leqslant k$

$Gw\,\text{cat}(X) \leqslant k \Rightarrow \text{cat}(X) \leqslant k.$

Démonstration. La fibre homotopique commune à $i: T^k X \to X^{k+1}$ et à $\gamma: \Omega X P(k) \to X$ est le joint k fois itéré $\Omega X * \Omega X * \cdots * \Omega X$. Les hypothèses dimensionnelles sont celles du théorème de Blakers–Massey qui donne directement les deux assertions.

Les méthodes de détermination inductive de $\text{cat}(X)$ font appel aux diverses notions d'invariant de Hopf. De l'article fondamental de Berstein–Hilton [2], nous extrayons le résultat qui nous sera utile par la suite.

THÉORÈME (1.7). *Soit* Y $(r-1)$*-connexe de dimension* $\leqslant (k+1)r-2$ *et* $f: S^N \to Y$ *avec* $N \geqslant r+2$. *Notions* $i: Y \to X = Cf$. *Alors*

(a) $\text{cat}\, Y \leqslant k$. *En outre* $\phi: Y \to T^k Y$ *factorisant* Δ *est unique à homotopie près.*

(b) $\text{cat}\, X \leqslant k \Leftrightarrow T^k(i) \circ \phi \circ f \simeq * : S^N \to T^k X.$

§2. Le modèle de Quillen

(*2.1*) Soit X un espace simplement connexe. Le modèle de Quillen L_X de X est une $\mathbf{Q}$-algèbre de Lie différentielle graduée minimale [1] qui détermine le type d'homotopie rationnelle de X au sens suivant: L_X est défini à isomorphisme près, et tout choix de foncteur $X \rightsquigarrow L_X$ est une équivalence de catégories d'homotopie entre les espaces 1-connexes rationnels et les $\mathbf{Q}$-algèbres graduées connexes minimales.

La description de L_X peut se résumer ainsi: (cf. [11]) L_X est l'algèbre de Lie libre $\mathbf{L}(s^{-1}\tilde{H}(X))$ sur la désuspension de l'homologie rationnelle réduite, munie d'une différentielle dont la partie quadratique (restreinte aux générateurs) est à désuspension près la diagonale réduite de $\tilde{H}$. En d'autres termes, L_X est isomorphe comme algèbre de Lie aux primitifs de la cobar-construction $P\Omega H(X)$,

et la différentielle de L_X est de la forme $d+\delta$, où d est la différentielle de $P\Omega H(X)$ et δ une dérivation qui augmente la filtration de Lie de +2.

Rappelons enfin la propriété fondamentale de L_X: il existe un isomorphisme naturel d'algèbres de Lie: $H(L_X)\cong\pi(X)=\pi_*(\Omega X)\otimes\mathbf{Q}$ où $\pi(X)$ est munie du crochet de Samelson.

L'existence de cet isomorphisme résulte de l'équivalence de catégories mentionnée ci-dessus.

Une autre conséquence de cette équivalence de catégories d'homotopie, qui résulte du travail fondamental de Quillen ([13]; cf. [8] (4.4.6)) est la suivante:

LEMME (2.2). *Soit X un espace 1-connexe rationnel, $f: S^{n+1}\to X$ continue, et soit $Y=X\bigcup_f e^{n+2}$. Soit $\alpha\in L_X$ un cycle représentant $[f]\in H_n(L_X)=\pi_{n+1}(X)$. Alors L_Y est un modèle minimal de l'algèbre de Lie $L_X\amalg\mathbf{L}(a)$, où $|a|=n+1$ et $da=\alpha$.*

Ceci suggère la définition suivante: soit L une DG algèbre de Lie libre, connexe.

DÉFINITION (2.3). Une filtration admissible sur L est la donnée

(a) d'un espace vectoriel générateur minimal $V\subset L$ (de sorte que $L\cong\mathbf{L}(V)$)

(b) d'une filtration croissante (F_pV) de V vérifiant $F_0V=0$ et $d(F_pV)\subset\mathbf{L}(F_{p-1}V)\subset L$.

On dira que l'algèbre L est de longueur $\leqslant n$ si elle admet une filtration admissible de longueur n, i.e. vérifiant $F_nV=V$.

Remarque (2.4). Si L est de longueur n, il est en général faux que tout espace générateur minimal porte une filtration admissible de longueur n: ainsi, soit

$$L=\mathbf{L}(a,b,c,e,f)$$

$$|a|=|b|=1,\qquad |c|=3,\qquad |e|=4,\qquad |f|=6$$

munie de la différentielle définie par $da=0=db$, $dc=[a,b]$, $de=0$, $df=[a,e]+[a,[b,c]]$.

La filtration admissible la plus courte sur ces générateurs est de longueur 3, mais si l'on pose $e'=e+[b,c]$ le système de générateurs a,b,c,e',f admet une filtration admissible de longueur 2.

A partir de la définition (2.3), nous introduisons les invariants d'homotopie rationnelle f et F comme suit:

DÉFINITION (2.5). Soit X un espace 1-connexe; $f(X)$ est le plus petit entier n tel qu'il existe une algèbre de Lie libre de longueur n quasi-isomorphe à L_X.

DÉFINITION (2.6). $F(X)$ est la longueur de L_X.

Le lemme (2.2) fournit immédiatement une interprétation géométrique des invariants f et F:

PROPOSITION (2.7). *Soit X un espace rationnel 1-connexe.*

(a) *On a $f(X) \leqslant n$ si et seulement si X a le type d'homotopie d'un cône itéré X_n, défini par*

$$X_0 = *, \qquad X_i = C(f_i : Y_i \to X_{i-1})$$

où les Y_i sont des bouquets de sphères (rationalisés) pour $i = 1, 2, \ldots, n$.

(b) *On a $F(X) \leqslant n$ si de plus les applications d'attachement f_i sont instables ($\Sigma f_i \sim *$) pour $i = 2, \ldots, n$.*

Le lemme (1.2) fournit immédiatement les inégalités:

PROPOSITION (2.8). $\operatorname{cat}(X) \leqslant f(X) \leqslant F(X)$.

Parvenus à ce point, il nous a longtemps semblé que la conjecture raisonnable était $\operatorname{cat}(X) = F(X)$. Nous verrons d'ailleurs que les espaces formels et les espaces coformels vérifient cette propriété, ainsi que l'exemple qui termine cet article. C'est à Yves Félix que nous devons le premier contre-exemple, à savoir un espace vérifiant $\operatorname{cat} = f = 3$, $F = 4$. L'invariant F est donc un majorant strict de la catégorie rationnelle: il permet néanmoins dans certains cas d'obtenir des majorations intéressantes:

On a tout d'abord:

PROPOSITION (2.9) *Si $F(X) \leqslant 3$, alors $F(X) = f(X)$.*

Démonstration. Il est clair que $F(X) = 1$ entraîne $f(X) = 1$. Compte tenu des inégalités (2.8) il suffit d'établir que le modèle minimal d'une algèbre de longueur 2 est de longueur $\leqslant 2$. Une algèbre de longueur 2 peut s'écrire $\mathbf{L}(V \oplus W)$ avec $d \mid V = 0$ et $dW \subset \mathbf{L}(V)$; soit $\pi : \mathbf{L}(V) \to Q\mathbf{L}(V) = V$ la projection canonique sur

les indécomposables, soit $W' = \mathrm{Ker}(\pi d)$ et soit W'' un supplémentaire de W' dans W. Soit $V' = d(W'') \subset \mathbf{L}(V)$ et soit V'' un sous-espace de $\mathbf{L}(V)$ tel que $V' \oplus V''$ soit un espace générateur minimal de $\mathbf{L}(V)$. Le quotient de $\mathbf{L}(V \oplus W)$ par l'idéal engendré par $V' \oplus W''$ est quasi-isomorphe à $\mathbf{L}(V \oplus W)$ d'après [1] (Proposition 1.5) or par construction il est isomorphe à $\mathbf{L}(V'' \oplus W')$ avec une différentielle nulle sur V'' et $dW' \subset [\mathbf{L}(V''), \mathbf{L}(V'')]$: cette dernière algèbre est donc le modèle minimal de $\mathbf{L}(V \oplus W)$ et elle est de longueur 2.

Ce résultat est le meilleur possible:

PROPOSITION (2.10) [5]. *Il existe des espaces X_n, pour tout $n \geqslant 0$, vérifiant $f(X_n) = 3$ et $F(X_n) = 3 + n$.*

La seule question qui reste ouverte est donc:

CONJECTURE (2.11). Pour tout espace 1-connexe X, on a $\mathrm{cat}(X) = f(X)$.

Nous concluons ce paragraphe en montrant comment l'invariant F permet d'établir des inégalités de contrôle dimensionnel de $\mathrm{cat}(X)$.

Le corollaire (2.14) ci-dessous est une reformulation partielle de (1.5) et (1.6), le corollaire (2.15) est nouveau.

DÉFINITION (2.12). Soit (x_i) un système minimal de générateurs de l'algèbre de Lie libre L. Une chaîne descendante de longueur k est une suite $x_{i_k}, \ldots, x_{i_1}$ d'éléments de ce système telle que $dx_{i_1} = 0$, et, pour tout $j = 1, \ldots, k-1$, $dx_{i_{j+1}}$ admet au moins une coordonnée non-nulle suivant un crochet basique où figure x_{i_j}.

Il est clair que toute algèbre de Lie de longueur $\geqslant n$ admet une chaîne descendante de longueur $\geqslant n$. On remarque alors:

LEMME (2.13). *Soit X un espace $(r-1)$-connexe, et soit $(x_k, \ldots, x_1)$ une chaîne descendante de L_X. On a*

$$|x_k| \geqslant kr - 1$$

En particulier, $\dim X \geqslant kr$.

Démonstration. Comme $L_X \cong \mathbf{L}(s^{-1}\tilde{H}(X))$, on a, $\forall x \in L_X$, $|x| \geqslant r-1$ ou $x = 0$ et comme L_X est minimale, tout bord est de filtration de Lie au moins 2, d'où

$$\forall i \in \{1, 2, \ldots, k-1\}, \qquad |x_{i+1}| = |dx_{i+1}| + 1 \geqslant |x_i| + r - 1 + 1.$$

d'où le résultat par sommation. ■

Remarquons qu'on retrouve, au moins pour les espaces rationnels, le résultat classique:

$$\operatorname{cat}(X) \leqslant \frac{\dim X}{r}.$$

COROLLAIRE (2.14) (Toomer [16]). *Si X est $(r-1)$-connexe, on a*:

$$(c(X)=k \quad et \quad F(X) \geqslant k+1) \Rightarrow \dim X \geqslant (k+2)r-1.$$

Démonstration. Soit $x_{k+1}, \ldots, x_1$ une chaîne descendante de L_X. Comme la partie quadratique de la différentielle d'un générateur est la transposée du cup-produit, il existe au moins un indice i_0 tel que x_{i_0} figure dans un crochet basique de longueur $\geqslant 3$ apparaissant dans dx_{i_0+1}. On a alors $|x_{i_0+1}| \geqslant |x_{i_0}|+2(r-1)+1$ d'où

$$\dim X - 1 \geqslant |x_k| \geqslant (k+1)r-1+r-1. \qquad \blacksquare$$

On obtient de même le résultat suivant:

COROLLAIRE (2.15). *Soit X un espace $(r-1)$-connexe à cup-produit rationnel trivial. Alors*

$$\operatorname{cat} X \leqslant \frac{\dim X + r - 1}{2r-1} < \frac{\dim X}{2r-1} + \frac{1}{2}.$$

Démonstration. Cette inégalité résulte de l'implication:

$$(c(X)=1 \quad \text{et} \quad F(X) \geqslant k) \Rightarrow \dim X \geqslant (2k-1)r-k+1$$

qu'on obtient en remarquant que tout bord de L_X est de filtration de Lie $\geqslant 3$.

§3. Espaces et applications formelles

Nous rassemblons dans ce paragraphe quelques résultats sur les espaces et les applications formelles qui sont nécessaires pour la justification des propriétés de notre exemple du paragraphe suivant, ainsi que quelques autres résultats intéressants. Les notations adoptées sont celles de [1].

THÉORÈME ET DÉFINITION (3.1). *Les propositions suivantes sont équivalentes: tout espace 1-connexe X vérifiant l'une d'elles sera dit formel.*

(HF 1) $L_X \cong P\Omega H_*(X)$ (*comme DG algèbres de Lie*)

(HF 2) *Il existe un espace vectoriel générateur minimal* $V \subset L_X$ *tel que* $d(V) \subset [V, V] \subset L_X$

(HF 3) *Soit* E_X *le modèle de Sullivan. La projection canonique* $ZE_X \to H(E_X) = H^*(X)$ *admet une extension multiplicative à* E_X*:*

$$\begin{array}{ccc} ZE_X & \longrightarrow & H^*(X) \\ \downarrow & \nearrow & \\ E_X & & \end{array}$$

(HF 4) *Il existe un quasi-isomorphisme d'algèbres de cochaînes*

$$E_X \xrightarrow{\sim} H^*(X).$$

(*3.2*) *Exemples d'espaces formels*

(a) Les sphères et les $K(\mathbf{Q}, n)$ sont formels.

(b) Un produit, un wedge d'espaces formels est formel.

(c) (R. Douglas) Un rétracte d'espace formel est formel (d'après HF 3).

(d) Une variété kaehlérienne compacte, un espace riemannien homogène symétrique compact sont des espaces formels ([3], [15]) lorsqu'ils sont 1-connexes.

Pour une autre caractérisation des espaces formels en termes de "modèles filtrés", voir [14].

La catégorie rationnelle d'un espace formel est des plus faciles à calculer:

PROPOSITION (3.3). *Si X est formel, on a*

$$c(X) = F(X) = \operatorname{cat}(X).$$

Démonstration. D'après HF 1, on a $L_X \cong P\Omega H_*(X)$. Or la filtration de co-algèbre de $H_*(X)$ est admissible, et de longueur $c(X)$. ■

Ainsi, tous les invariants que nous avons cités ou introduits coïncident sur les espaces formels.

Remarque (3.4). Les espaces coformels ([18], [4]) ou π-formels ([9], [11]) vérifient par définition les conditions suivantes, duales des (HF_i).

(πF_1) $E_X \cong \mathrm{Hom}(\mathscr{C}\pi(X), \mathbf{Q})$ où $\mathscr{C}$ est la construction de Koszul, ([1], (0.18)).

(πF_2) Il existe un espace vectoriel générateur minimal $U \subset E_X$ tel que $d(U) \subset S^2U \subset E_X$.

(πF_3) Il existe une extension multiplicative

$$\begin{array}{ccc} ZL_X & \longrightarrow & \pi(X) \\ \downarrow & \nearrow & \\ L_X & & \end{array}$$

(πF_4) Il existe un quasi-isomorphisme d'algèbres de Lie

$$L_X \xrightarrow{\sim} \pi(X).$$

Plus précisément, on peut "dualiser" les idées de Stasheff et Halperin ([14]; cf. [11], [4]) et exprimer plus précisément (πF_4) en disant que L_X est le modèle bigradué de $\pi(X)$. La filtration associée au premier degré est admissible, et coïncide sur les générateurs de L_X – à savoir $s^{-1}\tilde{H}_*(X)$ – avec la filtration d'Eilenberg–Moore. On établit ainsi:

PROPOSITION (3.5). *Si X est coformel, on a*

$$e(X) = F(X)$$

et par conséquent $e(X) = Gw\ \mathrm{cat}\,(X) = \mathrm{cat}\,(X) = F(X)$.

Ainsi $(S^3 \vee S^3) \bigcup_\gamma e^{2N+2}$, où γ est un élément quelconque de filtration de Lie N de l'algèbre de Lie libre $\pi_*(S^3 \vee S^3) \otimes \mathbf{Q}$, est un espace coformel, vérifiant $e = \mathrm{cat} = F = 2$, (cf. [10], Ex. 1) et $c = 1$. Notons que cet espace vérifie $w\ \mathrm{cat} = 1$ pour tout N, d'après ([2] Th. 4.3.). (Pour $N \leqslant 4$ cela résulte évidemment de (1.5).)

Un espace formel est un espace dont le type d'homotopie rationnel est déterminé par l'homologie. La notion analogue pour les applications peut être introduite, moyennant quelques précautions dues au fait que les modèles ne sont définis qu'à isomorphisme près.

Rappelons d'abord que l'on a des isomorphismes *naturels:*

$$\alpha_X : QL_X \xrightarrow{\cong} s^{-1}\tilde{H}_*(X; \mathbf{Q})$$

$$\beta_X : H(E_X) \xrightarrow{\cong} H^*(X, \mathbf{Q}).$$

La naturalité tient au fait que les modèles minimaux sont uniques à isomorphisme près, lui-même unique à homotopie près. Un espace X est formel s'il existe $\varphi_X : L_X \to P\Omega H_*(X)$ tel que $Q\varphi_X = \alpha_X$, ou encore s'il existe $\psi_X : E_X \to H^*(X)$ tel que $H(\psi_X) = \beta_X$.

Ceci conduit à la définition suivante:

THÉORÈME ET DÉFINITION (3.6). *Les conditions suivantes sont équivalentes pour une application $f : X \to Y$ entre espaces formels X et Y:*

(AF 1) *Il existe des isomorphismes φ_X et φ_Y, tels que $Q\varphi_X = \alpha_X$, $Q\varphi_Y = \alpha_Y$ et que le diagramme suivant*

$$\begin{array}{ccc} L_X & \xrightarrow{L_f} & L_Y \\ {\scriptstyle \varphi_X}\downarrow{\scriptstyle \cong} & & {\scriptstyle \varphi_X}\downarrow{\scriptstyle \cong} \\ P\Omega H_*(X) & \xrightarrow{P\Omega f_*} & P\Omega H_*(Y) \end{array} \tag{3.6.1}$$

commute à homotopie près.

(AF 2) *Il existe des quasi-isomorphismes ψ_X et ψ_Y, tels que $(\psi_X)^* = \beta_X$ et $(\psi_Y)^* = \beta_Y$, et que le diagramme suivant*

$$\begin{array}{ccc} E_X & \xleftarrow{E_f} & E_Y \\ {\scriptstyle \sim}\downarrow{\scriptstyle \psi_X} & & {\scriptstyle \sim}\downarrow{\scriptstyle \psi_Y} \\ H^*(X) & \xleftarrow[f^*]{} & H^*(Y) \end{array} \tag{3.6.2}$$

commute à homotopie près.

On dira que f est formalisable si elle vérifie (AF 1) *ou* (AF 2).

L'équivalence de (AF 1) et (AF 2) provient du fait que ([1], §3) les foncteurs $\min P\Omega \operatorname{Hom}(?, \mathbf{Q})$ et $\min \operatorname{Hom}(\mathscr{C}?, \mathbf{Q})$ sont des équivalences de catégories d'homotopie entre $[\min \mathrm{Alg}^*]$ et $[\min \mathrm{Lie}_*]$, compatibles avec l'homologie (ou la cohomologie).

Remarque (3.7). Nous évitons l'usage de l'adjectif "formelle" dans ce contexte, dans la mesure où il n'existe pas de choix naturel pour φ_X ou ψ_X. Un tel choix est cependant possible sur certaines sous-catégories d'espaces topologiques, par exemple la catégorie des variétés kaehlériennes compactes et des applications holomorphes: la théorie de Hodge fournit alors un quasi-isomorphisme naturel

entre les formes différentielles complexes et la cohomologie: ceci justifie l'assertion «les applications holomorphes entre variétés kaehlériennes compactes sont formelles» qui se trouve dans [3].

On remarquera en particulier qu'il n'y a aucune raison pour que la composée de deux applications formalisables le soit: nous donnons un exemple de cette situation en (3.10).

EXEMPLES (3.8). Soient X et Y deux espaces formels. Les projections $X\times Y\to X, Y$, les diagonales $\Delta^n: X\to X^n$, sont des applications formalisables: ceci résulte de l'isomorphisme naturel $E_{X\times Y}=E_X\otimes E_Y$.

On a évidemment:

LEMME (3.9). *Une application formalisable et homologiquement triviale est (rationnellement) homotope à zéro.*

EXEMPLE (3.10). Soit $X=S^7$, $Y=\Sigma(S^3\times S^3)=S^4\vee S^4\vee S^7$, $Z=S^4$. Soit $j: X\to Y$ l'inclusion de S^7, et $f: Y\to Z$ la suspension de la multiplication de S^3. La composée $f\circ j$ est l'application de Hopf $S^7\to S^4$ qui est rationnellement non-triviale, et par conséquent non-formalisable, d'après le lemme précédent. Or j est évidemment formalisable, et f l'est aussi. En effet

$$L_Y=\mathbf{L}(a_3, b_3, c_6), \qquad d=0$$

$$L_Z=\mathbf{L}(e_3)$$

$L_f: a\mapsto e,\ b\mapsto e,\ c\mapsto\lambda[e,e]$, (pour un certain $\lambda\in\mathbf{Q}^*$) et il suffit de choisir $\varphi_X: L_X\to P\Omega H_*(X)=L_X$ défini par

$$\varphi_X(a)=a, \qquad \varphi_X(b)=b, \qquad \varphi_X(c)=c-\lambda[a,b]$$

et $\varphi_Y=id$, pour obtenir le carré commutatif (3.6.1) ■

EXEMPLE (3.11). On peut montrer qu'une suspension est toujours formalisable (cf. l'exemple précédent).

Nous nous contenterons d'établir l'énoncé suivant, qui résulterait de (3.9) et (3.11), et que nous avons utilisé dans (1.5):

PROPOSITION (3.12). *Soit $f: X\to Y$ une application continue entre espaces rationnels. Si $0=f_*: \bar{H}_*(X;\mathbf{Q})\to\bar{H}_*(Y;\mathbf{Q})$, alors Σf est homotope à zéro.*

Preuve. Nous aurons besoin d'introduire la notion suivante: soit $u : L \to M$ un morphisme d'algèbres de Lie libres. Filtrons L et M par la filtration de Lie. Si $u \neq 0$, il existe un entier $k \geqslant 0$ et un seul tel que u augmente la filtration de k et pas de $k+1$. Le morphisme u induit alors une application polynomiale homogène de degré d'homogénéité $k+1$

$$p(u) : QL = F^1L/F^2L \to E_0^{k+1}M = F^{k+1}M/F^{k+2}M$$

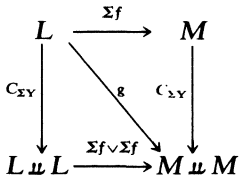
que nous appellerons la partie principale de u. Comme une suspension a le type d'homotopie rationnel d'un bouquet de sphères, on a

$$L_{\Sigma X} = (\mathbf{L}(\tilde{H}_*(X)), d = 0) = L$$

$$L_{\Sigma Y} = (\mathbf{L}(\tilde{H}_*(Y)), d = 0) = M$$

et $L_{\Sigma f} : L_{\Sigma X} \to L_{\Sigma Y}$ est homotope à zéro si et seulement si elle est nulle. Posons abusivement pour alléger l'écriture, $\Sigma f = L_{\Sigma f}$.

Exprimons à présent que Σf est un morphisme de co-H-espaces: on obtient le diagramme *commutatif* (car les différentielles sont nulles!):



où $C_{\Sigma X}$ et $C_{\Sigma Y}$ sont les comultiplications des suspensions ΣX et ΣY. Il est clair que la partie principale de ces comultiplications est la diagonale $x \to (x, x)$ sur les indécomposables. Posant $g = (\Sigma f \vee \Sigma f) \circ C_{\Sigma X} = C_{\Sigma Y} \circ \Sigma f$, la partie principale de g vérifie:

$$\forall x \in QL,\ p(g)(x) = (p(g)(x))' + (p(g)(x))''$$

où $'$ et $''$ désignent les inclusions de M dans $M \sqcup M$. En composant avec les codiagonales $L \sqcup L \to L$ et $M \sqcup M \to M$, on obtient:

$$\forall x \in QL, \qquad p(\Sigma f)(2x) = 2p(\Sigma f)(x)$$

ce qui montre que $p(\Sigma f)$ doit être homogène de degré 1, et alors

$$p(\Sigma f) = Q\Sigma f = f_* \neq 0, \quad \text{ou bien} \quad \Sigma f = 0.$$

Ceci achève la démonstration. ■

Nous terminons ce paragraphe par l'étude de la diagonale $\Delta^{n+1} : X \to X^{n+1}$ $(n \geq 1)$ d'un espace formel de catégorie k.

D'après (3.3), l'espace X est de catégorie k si et seulement si $c(X) = k$. Par ailleurs, la diagonale Δ^{n+1} est formalisable. Le choix approprié du foncteur L permet donc de supposer que $L_{\Delta^{n+1}} = P\Omega\Delta_*^{n+1}$, où

$$\Delta_*^{n+1} : H_*(X) \to H_*(X^{n+1}) = H_*(X)^{\otimes n+1}$$

est la comultiplication itérée de $H_*(X)$, qui est un morphisme de coalgèbres. Dire que $c(X) \leq k$ revient à dire que l'image de Δ_*^{n+1} est contenue dans $F_kH_*(X)$, où $F_*H_*(X)$ est la filtration canonique de coalgèbre; si $k \leq n$, on a la factorisation:

$$L_{\Delta^{n+1}} = P\Omega\Delta_*^{n+1} : L_X = P\Omega H_*(X) \to P\Omega(F_nH_*(X)) \subset L_{X^{n+1}}.$$

Cette factorisation fournit la factorisation "catégorique" de Δ^{n+1}! En effet:

LEMME (3.13). *Si X est formel, le bouquet garni $T^nX \subset X^{n+1}$ est formel.*

COROLLAIRE (3.14). *On a $L_{T^nX} \cong P\Omega(F_nH_*(X))$.*

Preuve du corollaire. On a bien $H_*(T^nX) = F_nH_*(X)$. ■

Preuve de (3.13). On considère T^nX comme somme amalgamée d'exemplaires de X^n le long d'exemplaires de X^{n-1}. Le lemme résulte alors du:

LEMME (3.14). *Soit*

$$\begin{array}{ccc} X & \xrightarrow{i_1} & X_1 \\ {\scriptstyle i_2}\downarrow & & \downarrow{\scriptstyle i_2} \\ X_2 & \xrightarrow{j_1} & Y \end{array}$$

un carré cocartésien dans lequel les flèches i_1, i_2 sont des cofibrations fermées, les espaces X, X_1, X_2 sont formels et les inclusions i_1, i_2 sont formalisables et injectives

en homologie. Alors Y est formel et j_1, j_2 sont formalisables et injectives en homologie.

Preuve. En choisissant convenablement les modèles minimaux, on peut supposer que

$$L_{i_j} = P\Omega(i_j)_* : P\Omega H_*(X) \to P\Omega H_*(X_j), \qquad j = 1, 2.$$

Comme le foncteur L est une équivalence de catégories d'homotopie qui respecte les fibrations et les cofibrations, on a:

$$L_Y \cong P\Omega H_*(X_1) \amalg_{P\Omega H_*(X)} P\Omega H_*(X_2)$$

L'algèbre L_Y est engendrée par $H_*(X_1) + H_*(X_2) = H_*(X_1) \oplus_{H_*(X)} H_*(X_2) = H_*(Y)$, et la restriction de la différentielle est quadratique, donc $L_Y \cong P\Omega H_*(Y)$ d'après (HF 2).

Remarque (3.15). On peut démontrer (3.13) au moyen du modèle de Sullivan, à partir du fait que E_{T^nX} est le modèle minimal de l'algèbre $(E_X)^{\otimes n+1}/I$, où $I = \bar{E}_X^{\otimes n+1}$ est le produit tensoriel des idéaux d'augmentation de chaque facteur, et ceci pour X connexe quelconque, formel ou non (Y. Félix, non publié).

Remarquons enfin que l'étude qui précède fournit une autre démonstration de l'inégalité $\operatorname{cat}(X) \leqslant c(X)$ pour un espace formel X.

§4. Un exemple

A l'appui de notre conjecture (2.9), nous avons choisi un espace à quatre cellules se prêtant à la détermination de tous les invariants introduits aux paragraphes 1 et 2. Considérons

$$X = (S^2 \vee \mathbf{C}P^2) \bigcup_{\alpha} e^7$$

le complexe cellulaire dans lequel l'application attachante α est le produit de Whitehead de l'identité de S^2 et de l'application de Hopf $\gamma : S^5 \to \mathbf{C}P^2$. Notre but est le

THÉORÈME (4.1).

$$c(X) = w\operatorname{cat}(X) = e(X) = Gw\operatorname{cat}(X) = 2 \qquad \operatorname{cat}(X) = f(X) = F(X) = 3.$$

La détermination de cat (X) nécessite un calcul détaillé de modèles minimaux, et constitue la partie principale du théorème; au préalable, nous commentons les autres propriétés.

Il est clair que $c(X) = 2$. Quant à $e(X) = 2$, il faut soit faire appel à l'autopsie des meurtres [10], soit raisonner avec la nilpotence homologique du modèle de Sullivan. Cette dernière technique fournit une démonstration assez aisée, mais sort totalement du cadre de cet article.

Comme X est simplement connexe de dimension 7, on obtient w cat $(X) = 2$ et Gw cat $(X) = 2$ par le lemme (1.5). Bien entendu, le lemme (1.6) est inapplicable!

Finalement, il est immédiat que $F(X) \leqslant 3$ par la proposition (2.7): l'écriture même de la décomposition cellulaire de X le montre comme triple cône avec des attachements homologiquement triviaux, donc instables par la proposition (3.12). Vu le résultat final, il est inutile de démontrer que $F(X) = 3$. Il reste cat (X) à déterminer.

En notant $Y = S^2 \vee \mathbf{CP}^2$, nous nous trouvons exactement dans les notations et les hypothèses du théorème de Berstein–Hilton (1.7). Pour établir que cat $(X) = 3$, il suffit donc de montrer que $T^2(i) \cdot \phi \cdot \alpha : S^6 \to T^2X$ est homotopiquement non-triviale. Par approximation cellulaire, nous sommes ramenés à étudier l'application

$$S^6 \xrightarrow{\alpha} Y \xrightarrow{\phi} T^2Y \subset\longrightarrow T^2Y \underset{\alpha_1}{\cup} e^7 \underset{\alpha_2}{\cup} e^7 \underset{\alpha_3}{\cup} e^7 = Z$$

ce que nous ferons à l'aide des modèles minimaux.

PROPOSITION (4.2). *Le modèle minimal de* $\mathbf{C}P^2$ *est* $\mathbf{L}(y, z)$ *avec*

$$|y| = 1 \qquad dy = 0$$

$$|z| = 3 \qquad dz = \tfrac{1}{2}[y,y].$$

Dans ce modèle, l'application de Hopf $\gamma : S^5 \to \mathbf{C}P^2$ *a* $[y, z]$ *comme cycle représentatif.*

Démonstration. Le modèle minimal de la sphère S^2 est $\mathbf{L}(y)$, avec $|y| = 1$ et $dy = 0$. Il est donc additivement engendré, comme son homologie, par y en dimension 1 et $[y, y]$ en dimension 2. (Rappelons que $[y, [y, y]] = 0$.) L'homotopie rationnelle de S^2 a donc deux générateurs correspondant à y et $[y, y]$. En notation habituelle, ceuxci sont $\iota \in \pi_2(S^2)$ et $[\iota, \iota] \in \pi_3(S^2)$. Un cycle représentant l'application de Hopf $S^3 \to S^2$ est donc $\frac{1}{2}[y, y]$. Par le Lemme (2.2), le modèle minimal de $\mathbf{C}P^2$ a bien la structure annoncée. Il en résulte que $\mathbf{C}P^2$ est formel, puisque les

différentielles du modèle sont purement quadratiques (HF 2). A l'aide de (HF 1), nous pouvons confirmer le résultat en explicitant la coalgèbre d'homologie de $\mathbf{C}P^2$. En notant abusivement y et z les générateurs homologiques, il vient en effet

$$\Delta_*(y) = 1\otimes y + y\otimes 1 \qquad \tilde{\Delta}_*(y) = 0$$

$$\Delta_*(z) = 1\otimes z + y\otimes y + z\otimes 1 \qquad \tilde{\Delta}_*(z) = y\otimes y = \tfrac{1}{2}[y, y].$$

Le calcul (facile) de l'homologie du modèle minimal de $\mathbf{C}P^2$ fournit un seul générateur en dimension 4, représenté par le cycle $[y, z]$.

L'application de Hopf $\gamma : S^5 \to \mathbf{C}P^2$ a donc un cycle représentatif qui est un multiple de $[y, z]$, purement quadratique. Il s'ensuit que l'espace $\mathbf{C}P^3$ est formel. Le calcul de la coalgèbre d'homologie de $\mathbf{C}P^2$ analogue au précédent, fournit alors facilement $[\gamma] = [y, z]$ achevant la démonstration. Bien entendu, la formalité de $\mathbf{C}P^3$ résulte aussi de [3]!

COROLLAIRE (4.3). *Le modèle minimal de* $Y = S^2 \vee \mathbf{C}P^2$ *est* $\mathbf{L}(x, y, z)$ *avec*

$$|x| = |y| = 1 \qquad dx = dy = 0$$

$$|z| = 3 \qquad dz = \tfrac{1}{2}[y, y].$$

PROPOSITION (4.4). *Le modèle minimal de* T^2Y *est* $\mathbf{L}(x_i, y_i, z_i, x_ix_j, x_iy_j, y_ix_j, y_iy_j, x_iz_j, z_ix_j, y_iz_j, z_iy_j, z_iz_j)$ $i = 1, 2, 3$, $(i, j) = (1, 2), (1, 3), (2, 3)$, *avec*

$$|x_i| = |y_i| = 1 \qquad dx_i = dy_i = 0$$

$$|z_i| = 3 \qquad dz_i = [y_i, y_i]$$

$$|x_ix_j| = 3 \qquad dx_ix_j = [x_i, x_j]$$

$$|x_iy_j| = 3 \qquad dx_iy_j = [x_i, y_j]$$

$$|y_ix_j| = 3 \qquad dy_ix_j = [y_i, x_j]$$

$$|y_iy_j| = 3 \qquad dy_iy_j = [y_i, y_j]$$

$$|x_iz_j| = 5 \qquad dx_iz_j = [x_i, z_j] + [x_iy_j, y_j]$$

$$|z_ix_j| = 5 \qquad dz_ix_j = [z_i, x_j] + [y_ix_j, y_i]$$

$$|y_iz_j| = 5 \qquad dy_iz_j = [y_i, z_j] + [y_iy_j, y_j]$$

$$|z_iy_j| = 5 \qquad dz_iy_j = [z_i, y_j] + [y_iy_j, y_i]$$

$$|z_iz_j| = 7 \qquad dz_iz_j = [z_i, z_j] + [y_iz_j, y_i] + [z_iy_j, y_j] + \tfrac{1}{2}[y_iy_j, y_iy_j].$$

Démonstration. La liste des générateurs est bien sûr $s^{-1}\tilde{H}_*(T^2Y)$. Le corollaire (3.14) ramène à nouveau le calcul des différentielles à celui des coproduits en homologie. Nous donnons seulement deux échantillons de calcul explicite, qui suffisent à expliquer les notations et les quelques abus de langage:

$$\Delta_*(x_1z_2) = (1\otimes x_1 + x_1\otimes 1)(1\otimes z_2 + y_2\otimes y_2 + z_2\otimes 1) =$$
$$1\otimes x_1z_2 + x_1\otimes z_2 + y_2\otimes x_1y_2 + x_1y_2\otimes y_2 + z_2\otimes x_1 + x_1z_2\otimes 1$$

$$\tilde{\Delta}_*(x_1z_2) = [x_1, z_2] + [x_1y_2, y_2]$$

$$\Delta_*(z_1z_2) = (1\otimes z_1 + y_1\otimes y_1 + z_1\otimes 1)(1\otimes z_2 + y_2\otimes y_2 + z_2\otimes 1)$$
$$= 1\otimes z_1z_2 + y_1\otimes y_1z_2 + z_1\otimes z_2 + y_2\otimes z_1y_2 + y_1y_2\otimes y_1y_2 + z_1y_2\otimes y_2$$
$$+ z_2\otimes z_1 + y_1z_2\otimes y_1 + z_1z_2\otimes 1$$

$$\tilde{\Delta}_*(z_1z_2) = [z_1, z_2] + [y_1z_2, y_1] + [z_1y_2, y_2] + \tfrac{1}{2}[y_1y_2, y_1y_2].$$

PROPOSITION (4.5). *L'application* $\phi: Y \to T^2Y$ *factorisant* Δ^3 *est donnée sur les modèles minimaux par:*

$$\phi_*(x) = x_1 + x_2 + x_3$$

$$\phi_*(y) = y_1 + y_2 + y_3$$

$$\phi_*(z) = z_1 + z_2 + z_3 + y_1y_2 + y_1y_3 + y_2y_3.$$

Démonstration. Y est un espace formel de catégorie 2. Il suffit donc de remplacer ϕ_* par Δ^3_*, en vertu de Lemme (3.13). Le calcul est alors immédiat.

PROPOSITION (4.6). *Le modèle minimal de* $Z = T^2Y \underset{\alpha_1}{\cup} e^7 \underset{\alpha_2}{\cup} e^7 \underset{\alpha_3}{\cup} e^7$ *est* $\mathbf{L}_{T^2Y} \amalg \mathbf{L}(u_1, u_2, u_3)$ *avec* $|u_i| = 6$ *et* $du_i = [x_i, [y_i, z_i]]$.

Démonstration. L'application $\alpha: S^6 \to S^2 \vee \mathbf{C}P^2$ a $[x, [y, z]]$ comme cycle représentatif. Comme les différentielles du_i sont décomposables, le modèle obtenu par l'application du Lemme (2.2) est minimal.

Démonstration du théorème (4.1). Nous devons démontrer que l'application $S^6 \xrightarrow{\alpha} Y \xrightarrow{\phi} T^2Y \hookrightarrow Z$ est non triviale. Par la Proposition (4.5), cette composition a

dans L_Z

$$[x_1+x_2+x_3, [y_1+y_2+y_3, z_1+z_2+z_3+y_1y_2+y_1y_3+y_2y_3]]$$

comme cycle représentatif. Tout revient donc à montrer que ce cycle n'est pas un bord. Une inspection très simple des différentielles de L_Z mène rapidement au résultat: les seules différentielles de L_Z qui contiennent le générateur y_1y_2 sont

$$dy_1z_2=[y_1, z_2]+[y_1y_2, y_2]$$

$$dz_1y_2=[z_1, y_2]+[y_1y_2, y_1]$$

$$dz_1z_2=[z_1, z_2]+[y_1z_2, y]+[z_1y_2, y_2]+\tfrac{1}{2}[y_1y_2, y_1y_2].$$

Aucun bord ne peut donc atteindre $[x_1, [y_3, y_1y_2]]$ qui est manifestement touché par le cycle donné. Donc $\operatorname{cat}(X)=3$, achevant la démonstration du théorème (4.1).

BIBLIOGRAPHIE

[1] Baues, H. and Lemaire, J. M., *Minimal models in homotopy theory.* Math. Ann. *225*, (1977) 219–242.

[2] Berstein, I. and Hilton, P., *Category and generalized Hopf invariants.* Illinois J. Math. *4*, (1960), 437–451.

[3] Deligne, P., Griffiths, P. A., Morgan, J., Sullivan, D., *Real homotopy type of Kähler manifolds.* Invent. Math. *29*, (1975) 245–274.

[4] Felix, Y., *Modèles et comodèles.* Rapport interne N°88, Université de Louvain, (Nov. 1979).

[5] Felix, Y., Lemaire, J. M. and Sigrist, F., *Sur les cônes itérés et la catégorie de Lusternik–Schnirelman rationnelle.* C.R. Acad. Sc. Paris. *290A* (1980) 905–907.

[6] Ganea, T., *Lusternik–Schnirelmann category and strong category.* Illinois J. Math. *11*, (1967) 417–427.

[7] Gilbert, W., *Some examples for weak category and conilpotency.* Illinois J. Math. *12*, (1968) 421–431.

[8] Lemaire, J. M., *Algèbres connexes et homologie des espaces de lacets.* Lect. Notes in Math. *422*, Springer, (1974).

[9] ——, *Modèles minimaux pour les algèbres de chaînes.* Pub. Math., Lyon (1976), *13*, N°3.

[10] ——, *Autopsie d'un meurtre....* Ann. Sc. E.N.S. *11*, (1978) 93–100.

[11] Oukili, A., *Thèse 3e cycle*, Nice 1978.

[12] Lemaire, J. M., Sigrist, F., *Dénombrement de types d'homotopie rationnelle.* C.R. Acad. Sc., Paris, *287A*, (1978) 109–112.

[13] Quillen, D., *Rational homotopy theory.* Ann. Math. *90*, (1969) 205–295.

[14] Stasheff, J., Halperin, S., *Obstructions to homotopy equivalences.* Advances Math. *32*, (1979) 233–279.

[15] Sullivan, D., *Infinitesimal computations in topology.* Pub. Math. I.H.E.S., *47* (1978) 269–331.

[16] Toomer, G., *Lusternik–Schnirelmann category and the Moore spectral sequence.* Math. Z. *138*, (1974) 123–143.

[17] WHITEHEAD, G., *The homology suspension.* Coll. top. alg., Louvain (1956), 89–65. Liège: Georges Thone; Paris: Masson & Cie. (1957).
[18] MILLER, T., NEISENDORFER, J., *Formal and Coformal spaces.* Illinois J. Math. (à paraître).

Jean-Michel Lemaire
Département de Mathématiques
L.A. au C.N.R.S. N°168
Université de Nice-Parc Valrose
F-06034 NICE CEDEX

François Sigrist
Institut de Mathématiques
Université de Neuchâtel
Chantemerle 20
CH-2000 NEUCHATEL

Reçu le 6 juin 1980

Comment. Math. Helvetici **56** (1981) 123–131 0010-2571/81/001123-09$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

On one-relator soluble groups

RALPH STREBEL

1. Introduction

1.1. Let G be a one-relator group, say

$$G = \langle x_1, x_2, \ldots, x_m ; w \rangle,$$

and $\mathfrak{V}$ a soluble variety. Following G. Baumslag we call $G/\mathfrak{V}(G)$, where $\mathfrak{V}(G)$ denotes the $\mathfrak{V}$-verbal subgroup of G, a *one-relator $\mathfrak{V}$-group*. G. Baumslag initiated the search for conditions on m, w and $\mathfrak{V}$ that force $G/\mathfrak{V}(G)$ to be infinitely related. He showed this to be the case if $\mathfrak{V} = \mathfrak{A}^2$ (i.e. the variety of all metabelian groups) and $m \geqq 3$, w arbitrary, or $m = 2$, w a proper power ([2], p. 67, Theorem F). These results were extended in [4] (p. 259, Theorem B): If $m \geqq 3$ and $\mathfrak{V}$ is a soluble variety containing a subvariety of the form $\mathfrak{A}_p \cdot \mathfrak{A}$, p a prime, then $G/\mathfrak{V}(G)$ is infinitely related. This result is actually best possible; for a soluble variety $\mathfrak{V}$ having no subvariety of the form $\mathfrak{A}_p \cdot \mathfrak{A}$ is nilpotent-by-locally finite (J. Groves [6], p. 392, Theorem A; I'd like to thank P. M. Neumann for drawing my attention to this reference). For $m \geqq 3$ the final answer to G. Baumslag's question therefore reads:

THEOREM A. *Let $G = \langle x_1, \ldots, x_m ; w \rangle$ be a one-relator group and $\mathfrak{V}$ a soluble variety. If $m \geqq 3$ then $G/\mathfrak{V}(G)$ is finitely related if and only if $\mathfrak{V}$ is nilpotent-by-locally finite. In particular, a finitely generated $\mathfrak{V}$-free group is finitely related if and only if it is nilpotent-by-finite.*

(For groups $H = \langle x_1, \ldots, x_m ; w_1, \ldots, w_n \rangle$ with $m \geqq n+2$ the answer is the same.)

1.2. As the case $m = 1$ is clear we are left with $m = 2$. This case is surprisingly complex, as was already brought to light by the examples in [3] and by the Baumslag-Boler 2-generator 2-relator group

$$G = \langle x_1, x_2; [x_1, x_2]^{x_1} = [x_1, x_2] \cdot [x_1, x_2]^{x_2}, [[x_1, x_2], [x_1, x_2]^{x_2}] = 1 \rangle$$
$$= \langle x_1, x_2; x_1^{x_2 x_1} = x_1^{x_2 x_2}, [[x_1, x_2], [x_1, x_2]^{x_2}] = 1 \rangle,$$

which is a one-relator metabelian group whose derived group is free Abelian of rank $\aleph_0$ ([2], p. 68, Ex. G(ii)).

In [4] the subcase "$m=2$ and w not a commutator word" is dealt with. The authors prove that for $\mathfrak{V}$ a variety with $\mathfrak{A}^2 \leqq \mathfrak{V} \leqq \mathfrak{N}_c\mathfrak{A}$, $c \geqq 1$, $G/\mathfrak{V}(G)$ is finitely related if and only if $G/\mathfrak{A}^2(G) = G/G''$ is, and they give an algorithm which, for a preassigned w, allows to decide whether G/G'' is finitely related or not.

In this article the remaining subcase will be treated. To state our answer concisely we introduce some notation. Let F be the free group on x_1 and x_2, let $D_1 : F \to \mathbb{Z}F$ denote the partial derivative with respect to x_1 and $^- : \mathbb{Z}F \to \mathbb{Z}F_{ab}$ the canonical projection. Choose an isomorphism $\theta : F_{ab} \xrightarrow{\sim} \mathbb{Z}^2 \subset \mathbb{R}^2$ onto the standard lattice in the Euclidean plane $\mathbb{R}^2$. If $\lambda = \sum \lambda_q \cdot q$ is an element of the group ring $\mathbb{Z}F_{ab}$ the set

$$\theta_*(\lambda) = \{\theta q \mid \lambda_q \neq 0\}$$

is a finite subset of $\mathbb{Z}^2$ and λ_q will also be called the coefficient of the point θq.

THEOREM B. *Let $G = \langle x_1, x_2; w\rangle$ be a one-relator group with w a commutator word. If $G' \neq G''$ and $l \geqq 3$ then $G/G^{(l)}$ is never finitely related. G/G'', however, may be finitely related: this happens precisely if $\theta_*(\overline{D_1 w/(1-\bar{x}_2)})$ does not lie on a straight line, if the convex polygon $\mathcal{P}$ bounding the convex hull of $\theta_*(\overline{D_1 w/(1-\bar{x}_2)}) \subseteq \mathbb{R}^2$ has no parallel edges and if for every edge s of $\mathcal{P}$ the vertex v_s of $\mathcal{P}$ having greatest distance from the straight line supporting s has coefficient ± 1.*

If, to give an illustration, G is the Baumslag-Boler group

$$G = \langle x_1, x_2;\ x_1^{x_2^2 x_1} = x_1^{x_2^2 x_2}\rangle$$

its bounding polygon $\mathcal{P}$ is a triangle all whose vertices have coefficients ± 1 and so G/G'' is finitely related, in accordance with our previous remark.

1.3. Theorem A and Theorem B together reveal that G. Baumslag's question whether $G/G^{(l+1)}$, G a one-relator group, can only be finitely related if $G/G^{(l)}$ is so ([2], p. 71, Problem 6), has a positive answer in all cases, save possibly in the case where $m=2$ and w is not a commutator word. The positive answer, moreover, suggests that the variety $\mathfrak{A}^2$ may play an exceptional role among the varieties $\mathfrak{A}^l$, $l \geqq 2$, in that a finitely generated group of $\mathfrak{A}^l$ which is "close" to being relatively free, containing e.g. an $\mathfrak{A}^l$-free subgroup of rank $\geqq 2$, can at best be finitely related if it is metabelian.

2. Proof of Theorem B: the metabelian case

2.1. Let $G=\langle x_1, x_2; w\rangle$ be a one-relator group with w a commutator word. Then $Q=G_{ab}=G/G'$ is free-Abelian of rank 2 and $A=G'_{ab}$ is a $\mathbb{Z}Q$-module (via conjugation). By [5] (Theorem A and formula (2.3)) the answer to the question whether the quotient group G/G'' is finitely related depends only on the annihilator ideal $I=\mathrm{Ann}_{\mathbb{Z}Q}(A)$ of A. We claim that I is a principal ideal generated by $\overline{D_1 w}/(1-\bar{x}_2)$.

To see this let F be free on x_1 and x_2, let $D_i: F\to\mathbb{Z}F$ denote the partial derivative with respect to x_i, $i=1,2$, and $^-:\mathbb{Z}F\twoheadrightarrow\mathbb{Z}F_{ab}$ the canonical projection. If $G=F/gp_F(w)=\langle x_1, x_2; w\rangle$ is a one-relator group with $w\in F'$, the canonical projection $\pi: F\twoheadrightarrow G$ induces an isomorphism $\pi_*: F_{ab}\xrightarrow{\sim} G_{ab}$ which will be thought of as being the identity.

LEMMA 1. *If $G=\langle x_1, x_2; w\rangle$ is a one-relator group and w a commutator word then G'_{ab} is a cyclic $\mathbb{Z}G_{ab}$-module with $\overline{D_1 w}/(1-\bar{x}_2)$ as its defining annihilator.*

Proof. There is an exact sequence of left $\mathbb{Z}G$-modules

$$\mathbb{Z}G\xrightarrow{\partial_2}\mathbb{Z}G\oplus\mathbb{Z}G\xrightarrow{\partial_1}\mathbb{Z}G\xrightarrow{\varepsilon}\mathbb{Z}\to 0 \tag{1}$$

where ε is the augmentation, ∂_1 sends $(1,0)$ to $1-\pi(x_1)$ and $(0,1)$ to $1-\pi(x_2)$, while $\partial_2(1)=((\pi\circ D_1)w, (\pi\circ D_2)w)$. If this beginning of a $\mathbb{Z}G$-free resolution is used to compute $H_1(G',\mathbb{Z})=G'_{ab}$ one gets

$$G'_{ab}\cong\ker(\mathbb{Z}\otimes_{G'}\partial_1)/\mathrm{im}\,(\mathbb{Z}\otimes_{G'}\partial_2).$$

In the special case $w_0=x_1x_2x_1^{-1}x_2^{-1}$ we have $G'=G''$ and so $\ker(\mathbb{Z}\otimes_{G'}\partial_1)$ is generated by

$$(\overline{D_1 w_0}, \overline{D_2 w_0})=(1-\bar{x}_2, -1+\bar{x}_1).$$

In the general case, $\mathrm{im}\,(\mathbb{Z}\otimes_{G'}\partial_2)$ is the cyclic submodule generated by $(\overline{D_1 w}, \overline{D_2 w})$. Since we know a priori that $\mathrm{im}\,(\mathbb{Z}\otimes_{G'}\partial_2)\subseteq\ker(\mathbb{Z}\otimes_{G'}\partial_1)$ there must exist $\mu\in\mathbb{Z}G_{ab}$ such that $\overline{D_1 w}=\mu(1-\bar{x}_2)$ and $\overline{D_2 w}=\mu(-1+\bar{x}_1)$. It follows that $G'_{ab}\cong\mathbb{Z}G_{ab}/(\mathbb{Z}G_{ab}\cdot\mu)$. Note that μ is unique, $\mathbb{Z}G_{ab}$ being a unique factorization domain. □

Remark. If $w=([x_1,x_2]^{c_1})^{f_1}\cdot\cdots\cdot([x_1,x_2]^{c_k})^{f_k}$ where $c_1,\ldots,c_k$ are integers and $f_1,\ldots,f_k$ elements of F then $\overline{D_1 w}/(1-\bar{x}_2)=c_1\bar{f}_1+\cdots+c_k\bar{f}_k$. This shows that every element of $\mathbb{Z}F_{ab}$ is the defining annihilator of G'_{ab} for a suitable w.

2.2. We describe next the necessary and sufficient condition stated in [5] for G/G'' to be finitely related. Set $A = G'_{ab}$ and $Q = G_{ab}$ for short. Let $C(A) = 1 + \mathrm{Ann}_{\mathbb{Z}Q}(A)$ denote the set of centralizers of A. By Lemma 1 we have

$$C(A) = \{1 + \mu(\overline{D_1 w}/(1-\bar{x}_2)) \mid \mu \in \mathbb{Z}Q\}.$$

Let $S^1 \subseteq \mathbb{R}^2$ denote the unit circle in the Euclidean plane $\mathbb{R}^2$, and choose an isomorphism $\theta : Q = G_{ab} \xrightarrow{\sim} \mathbb{Z}^2 \subseteq \mathbb{R}^2$ onto the standard lattice of $\mathbb{R}^2$. Define the subset $\Sigma_A \subseteq S^1$ by

$$\Sigma_A = \bigcup_{\lambda \in C(A)} \{u \mid \text{for all } q \in \mathrm{supp}\,(\lambda) \text{ one has } \langle u, \theta q\rangle > 0\}. \tag{2}$$

Then the metabelian group G/G'' is finitely related if, and only if, $\Sigma_A \cup -\Sigma_A = S^1$, where $-\Sigma_A = \{-u \in S^1 \mid u \in \Sigma_A\}$ is the set of antipodes of Σ_A ([5], Theorem A, formula (2.3) and Remark *2.3*). So it remains only to determine when $\Sigma_A \cup -\Sigma_A = S^1$.

2.3. Every $u \in S^1$ leads to an $\mathbb{R}$-grading of $\mathbb{Z}Q$ by setting $\mathbb{Z}Q = \bigoplus \{R\rho \mid \rho \in \mathbb{R}\}$, where R_ρ is the $\mathbb{Z}$-submodule of $\mathbb{Z}Q$ generated by all elements $q \in Q$ with $\langle u, \theta q\rangle = \rho$. Let $\lambda = 1 + \mu \cdot \mu_0$, $\mu_0 = \overline{D_1 w}/(1-\bar{x}_2)$, be a centralizer of A. It contributes to Σ_A as given in (2) precisely if the homogeneous components of negative degree of $\mu \cdot \mu_0$ are all zero, while the homogeneous component of degree zero equals -1. This is only possible when the homogeneous component of μ_0 having lowest degree is a unit of $\mathbb{Z}Q$, i.e., of the form $\pm q$, $q \in Q$. If, conversely, the homogeneous component of μ_0 having lowest degree is a unit of $\mathbb{Z}Q$, say e, then $1 + (-e^{-1})\mu_0$ will be a centralizer contributing to Σ_A. If, for $u \in S^1$ and $\lambda \in \mathbb{Z}Q$, we denote by λ_u the homogeneous component of λ with lowest degree in the $\mathbb{R}$-grading associated with u, we have therefore

$$\Sigma_A = \{u \in S^1 \mid (\overline{D_1 w}/(1-\bar{x}_2))_u \text{ is a unit of } \mathbb{Z}Q\}. \tag{3}$$

Assume now that $\Sigma_A \cup -\Sigma_A = S^1$ and set $\mu_0 = \overline{D_1 w}/(1-\bar{x}_2)$ for short. Then for every $u \in S^1$ either $(\mu_0)_u$ or $(\mu_0)_{-u}$ is a unit of $\mathbb{Z}Q$. Let us see what this implies for the convex polygon $\mathcal{P}$ bounding the convex hull of the finite set

$$\theta_*(\mu_0) = \{\theta q \mid q \in \mathrm{supp}\,(\mu_0)\}.$$

Let u be a unit vector orthogonal to the straight line l supporting an edge s of $\mathcal{P}$ and such that the remaining edges of $\mathcal{P}$ lie in the positive halfplane determined by l and u. Then the support of $(\mu_0)_u$ has at least two elements and so $(\mu_0)_u$ cannot be a unit, whence $(\mu_0)_{-u}$ must be a unit. It follows that no other edge can be

parallel to s and that the vertex v_s of greatest distance from l must have coefficient ± 1 (and cannot lie on l). Thus the conditions on $\mathcal{P}$ stated in Theorem B are necessary. Suppose, conversely, they are satisfied. Then $\Sigma_A \cup -\Sigma_A$ is a *closed, non-void* set. On the other hand it is immediate from (2) that Σ_A, and hence also $\Sigma_A \cup -\Sigma_A$, are *open*. Since S^1 is connected this gives $\Sigma_A \cup -\Sigma_A = S^1$. □

3. Proof of Theorem B: the meta metabelian case

3.1. We begin with some auxiliary results. Let $K \lhd L \lhd G$ be normal subgroups of a group G and suppose $K \leqq [L, L]$. For every trivial $\mathbb{Z}G$-module A the pair of extensions

$$\begin{array}{ccccc} K & \lhd & L & \twoheadrightarrow & L/K \\ \downarrow{=} & & \downarrow & & \downarrow \\ K & \lhd & G & \twoheadrightarrow & G/K \end{array} \tag{4}$$

where all maps are the canonical ones, gives rise to a commutative diagram

$$\begin{array}{ccccccc} H_2(L, A) & \to & H_2(L/K, A) & \to & K_{ab} \otimes_L A & \to & 0 \\ \downarrow & & \downarrow & & \downarrow{1_*} & & \\ H_2(G, A) & \to & H_2(G/K, A) & \to & K_{ab} \otimes_G A & \to & 0. \end{array} \tag{5}$$

Its rows are exact (here the assumption that $K \leqq [L, L]$ and A be a trivial G-module are used). Each $g \in G$ defines by conjugation an action on each of the groups occurring in (4) and, as all the canonical maps in (4) are compatible with this action, it follows that (5) is a diagram of (right) $\mathbb{Z}G$-modules and $\mathbb{Z}G$-homomorphisms. Note that $H_2(G, A)$ and $H_2(G/K, A)$ are trivial G-modules, inner automorphisms inducing the identity in homology, and that the G-action on $K_{ab} \otimes_L A$ resp. $K_{ab} \otimes_G A$ is the obvious one induced by conjugation (and so trivial in the second case). By applying the right exact functor $(?)_G = ? \otimes_G \mathbb{Z}$ to the diagram (5) one therefore obtains the commutative diagram with exact rows

$$\begin{array}{ccccccc} H_2(L, A)_G & \to & H_2(L/K, A)_G & \to & K_{ab} \otimes_G A & \to & 0 \\ \downarrow & & \downarrow & & \downarrow{=} & & \\ H_2(G, A) & \to & H_2(G/K, A) & \to & K_{ab} \otimes_G A & \to & 0. \end{array} \tag{6}$$

In particular, one gets

COROLLARY 2. *Let $L \lhd G$ be a normal subgroup of G and $A = \mathbb{Q}$. If $H_2(L, \mathbb{Q})_G$ and $H_2(G, \mathbb{Q})$ are finite dimensional then $H_2(G/[L, L], \mathbb{Q})$ is infinite dimensional if and only if $H_2(L_{ab}, \mathbb{Q})_G$ is so.*

LEMMA 3. *Let $L \lhd G$ be a normal subgroup. If $H_2(G, \mathbb{Q})$ and $H_2(L, \mathbb{Q})_G$ are finite and $H_2(G/L, \mathbb{Q})$ is infinite dimensional, then $H_2(G/[L, L], \mathbb{Q})$ is also infinite dimensional.*

Proof. The 5-term sequence associated with the extension $L \lhd G \twoheadrightarrow G/L$ shows $L_{ab} \otimes_G \mathbb{Q} \cong L/[L, G] \otimes_{\mathbb{Z}} \mathbb{Q}$ to be infinite dimensional. Hence $L/[L, G]$ has infinite torsion-free rank and the same is true of its exterior square

$$L/[L, G] \wedge L/[L, G] \cong H_2(L/[L, G], \mathbb{Z}).$$

The canonical projection $L/[L, L] \twoheadrightarrow L/[L, G]$ induces a *surjective* $\mathbb{Z}G$-module homomorphism $H_2(L_{ab}, \mathbb{Q}) \twoheadrightarrow H_2(L/[L, G], \mathbb{Q})$ and the target module is a G-trivial, infinite dimensional vector space over $\mathbb{Q}$. Consequently $H_2(L_{ab}, \mathbb{Q})_G$ is infinite dimensional and Corollary 2 permits to conclude. □

3.2. **E**-*groups.* Following [8] a group X will be called an **E**-*group* if its Abelianization X_{ab} is *torsion-free* and if there exists a $\mathbb{Z}X$-projective resolution $\cdots \to P_2 \xrightarrow{\partial_2} P_1 \to P_0 \twoheadrightarrow \mathbb{Z}$ such that $\partial_2 \otimes_X \mathbb{Z} : P_2 \otimes_X \mathbb{Z} \to P_1 \otimes_X \mathbb{Z}$ is *injective.* A basic result about **E**-groups states that the commutator subgroup of an **E**-group is again an **E**-group; in particular, $H_2(X^{(l)}, \mathbb{Z}) = 0$ for every $l \geqq 0$ (see [8], p. 303, Theorem A).

As an illustration, let $G = F/gp_F(w) = \langle x_1\, x_2; w \rangle$ be a one-relator group with $w \in F'$. Then $H_2(G, \mathbb{Z}) \neq 0$ and so G itself is not an **E**-group. But the situation improves if we pass to G'. To see this view the beginning (1) of a $\mathbb{Z}G$-free resolution as a sequence of right $\mathbb{Z}G$-modules. Then $\partial_2 \otimes_{G'} \mathbb{Z}$, being isomorphic to $\partial_2 \otimes_G \mathbb{Z}G_{ab}$, is injective if, and only if, $\overline{D_1 w} \neq 0$, and this, in turn, happens precisely when the canonical projection $F/F'' \twoheadrightarrow G/G''$ is *not* injective, i.e. if $w \notin F''$. If therefore $w \in F' \setminus F''$ and G'_{ab} is torsion-free, G' will be an **E**-group. Since $H_2(G, \mathbb{Q})$ is at most 1-dimensional iterated application of Lemma 3 will then establish

PROPOSITION 4. *Let $G = F/gp_F(w) = \langle x_1, x_2; w \rangle$ be a one-relator group with $w \in F' \setminus F''$. If G'_{ab} is torsion-free and if there is a natural number $l_0 > 1$ for which $H_2(G/G^{(l_0)}, \mathbb{Q})$ is infinite dimensional then $H_2(G/G^{(l)}, \mathbb{Q})$ will remain infinite dimensional for all $l \geqq l_0$.*

Remark. In [1] H. Abels exhibits a finitely related, soluble group L having an infinitely related quotient group $\bar{L}$, thereby answering a question P. Hall raised in

1954. By varying Abels' example one can concoct a finitely related soluble group L_1 having a homomorphic image $\bar{L}_1$ whose $H_2(\bar{L}_1, \mathbb{Q})$ is infinite dimensional. This shows that in Proposition 4 the (implicit) assumption that G' be an **E**-group is not redundant.

3.3. We are now ready to prove the remainder of Theorem B. Let $G = F/gp_F(w) = \langle x_1, x_2; w\rangle$ be a one-relator group and w a commutator word. If G/G'' is infinitely related then so is every soluble quotient of G which maps onto G/G'' ([5], Theorem B). Assume therefore G/G'' to be finitely related. Then $\Sigma_A \cup -\Sigma_A = S^1$, where $A = G'_{ab}$. (This follows from Theorem A of [5] and has been discussed in Subsection *2.2.*) Since S^1 is connected and Σ_A as well as $-\Sigma_A$ is open there exists $u \in S^1$ with $\{u, -u\} \subseteq \Sigma_A$, and because the unit vectors with rational slope are dense in S^1, u can be chosen to have rational slope. Let t be a primitive element of the free Abelian group G_{ab} which maps under the identification $\theta : G_{ab} \xrightarrow{\sim} \mathbb{Z}^2 \subseteq \mathbb{R}^2$ onto a vector orthogonal to u, and enlarge $\{t\}$ to a basis $\{s, t\}$. By (3) we know that $\{u, -u\} \subseteq \Sigma_A$ holds if and only if the homogeneous components of both lowest and highest degree of $\overline{D_1 w}/(1-\bar{x}_2)$ (in the $\mathbb{R}$-grading associated with u) are units. With respect to the basis $\{s, t\}$ the defining annihilator $\overline{D_1 w}/(1-\bar{x}_2)$ must therefore have the form

$$\overline{D_1 w}/(1-\bar{x}_2) = c_n s^n + c_{n+1} s^{n+1} + \cdots + c_{n+k} s^{n+k},$$

where each c_j, $j = n, \ldots, n+k$, is an element of the group ring $\mathbb{Z}gp(t)$, and c_n and c_{n+k} are units. Consequently $A = G'_{ab}$, viewed as a $\mathbb{Z}gp(t)$-module, is a free $\mathbb{Z}gp(t)$-module of rank k and so, in particular, free Abelian of rank $\aleph_0$, unless $G'_{ab} = 0$. As there is nothing to be proved in the latter case let us assume $G'_{ab} \neq 0$. Then G'_{ab} is torsion-free and $\overline{D_1 w} \neq 0$ and thus G' is an **E**-group. Comparing Theorem B with Proposition 4 and taking into account that $G/G^{(1)}$ cannot be finitely related if $H_2(G/G^{(1)}, \mathbb{Q})$ is infinite dimensional, we see that Theorem B will be established if we can show that $H_2(G/G''', \mathbb{Q})$ is infinite dimensional. We shall achieve this goal by a variation on the proof of Lemma 3.

Since G'_{ab} is a free $\mathbb{Z}gp(t)$-module of positive rank its exterior square $G'_{ab} \wedge G'_{ab}$ is a free $\mathbb{Z}gp(t)$-module of rank $\aleph_0$. As $H_2(G', \mathbb{Q}) = 0$, the 5-term sequence associated with $G'' \lhd G' \twoheadrightarrow G'_{ab}$ produces an isomorphism

$$H_2(G'_{ab}, \mathbb{Q}) \xrightarrow{\sim} G''/[G'', G'].$$

Hence $B = G''/[G'', G']$ is a free $\mathbb{Z}gp(t)$-module of infinite rank, and

$$\bar{B} = G''/[G'', G'] \otimes_{gp(t)} \mathbb{Z}$$

is free Abelian of rank $\aleph_0$. On the other hand, G/G'' is finitely related whence B is a finitely generated $\mathbb{Z}G_{ab}$-module and $\bar{B}$ is a finitely generated $\mathbb{Z}gp(s)$-module. As $\mathbb{Q}gp(s)$ is a principal ideal domain, $\bar{B} \otimes_{\mathbb{Z}} \mathbb{Q}$ maps onto $\mathbb{Q}gp(s)$. The exterior square of $\mathbb{Q}gp(s)$ is a free $\mathbb{Q}gp(s)$-module of infinite rank, whence $V = H_2(\mathbb{Q}gp(s), \mathbb{Q}) \otimes_{gp(s)} \mathbb{Q}$ is a G-trivial, infinite dimensional vector space over $\mathbb{Q}$. Collecting the various maps one gets a chain of $\mathbb{Q}G$-module epimorphisms

$$H_2(G''_{ab}, \mathbb{Q}) \twoheadrightarrow H_2(G''/[G'', G'], \mathbb{Q}) \xrightarrow{\sim} H_2(G''/[G'', G'] \otimes_{\mathbb{Z}} \mathbb{Q}, \mathbb{Q}) \twoheadrightarrow$$
$$\twoheadrightarrow H_2(\mathbb{Q}gp(s), \mathbb{Q}) \twoheadrightarrow V,$$

which shows that $H_2(G''_{ab}, \mathbb{Q})_G$ is infinite dimensional. The proof is completed by an appeal to Corollary 2. □

3.4. Remarks. (a) The topological argument used in *3.3.* to prove that G''_{ab} is a finitely generated $\mathbb{Z}gp(t)$-module for a suitable $t \in G_{ab}$ generalizes to arbitrary finitely related metabelian groups, proving that *every finitely related metabelian group whose Abelianization G_{ab} has torsion-free rank >1 contains a finitely generated normal subgroup $N \lhd G$ with infinite cyclic quotient group.* Moreover, if $M \lhd G$ is a finitely generated normal subgroup with infinite cyclic quotient group then all normal subgroups M_1 (with infinite cyclic quotient group) which are sufficiently close to M, will be finitely generated. (Proofs of these assertions will appear in a joint paper with Robert Bieri.) The second property holds in fact for every finitely generated group G with G_{ab} of torsion-free rank >1 (W. D. Neumann [7]).

(b) Although the subcase where $m = 2$ and w is not a commutator word looks, on the face of it, more tractable than the other subcase, I have not succeeded in proving that $G/G^{(l)}$ is infinitely related if $l \geqq 3$ and $G'' \neq G'''$. One might be tempted to proceed similarly: If G/G'' is infinitely related every soluble quotient of G which can be mapped onto G/G'' is infinitely related; in the contrary case G' can be shown to be an **E**-group and by the analogue of Proposition 4 it will then do to establish that $H_2(G''_{ab}, \mathbb{Q})_G$ is infinite dimensional. Note, however, that in the present situation $G'_{ab} \otimes_{\mathbb{Z}} \mathbb{Q}$ is finite dimensional so that the strategy used in *3.3.* fails.

BIBLIOGRAPHY

[1] ABELS, H., *An example of a finitely presented solvable group;* Homological Group Theory – Proceedings of a symposium, held at Durham in September 1977; ed. by C. T. C. Wall (pp. 205–211); London Math. Soc. Lecture Note Series. 36; Cambridge University Press 1979.

[2] BAUMSLAG, G., *Finitely presented metabelian groups*; Proc. Seconf Intern. Conf. Theory of Groups; Springer Lecture Notes in Math. *372* (1973), 65–74.

[3] BAUMSLAG, G. and STREBEL, R., *Some finitely generated, infinitely related metabelian groups with trivial multiplicator*; J. Algebra *40* (1976), 46–62.
[4] BIERI, R. and STREBEL, R., *Almost finitely presented soluble groups*; Comment. Math. Helv. *53* (1978), 258–278.
[5] ——, *Valuations and finitely presented metabelian groups*; Proc. London Math. Soc. (3) *41* (1980), 439–464.
[6] GROVES, J. R. J., *Varieties of soluble groups and a dichotomy of P. Hall*; Bull. Austral. Math. Soc. *5* (1971), 391–410.
[7] NEUMANN, W. D., *Normal subgroups with infinite cyclic quotient*; Math. Scientist *4* (1979), 143–148.
[8] STREBEL, R., *Homological methods applied to the derived series of groups*; Comment. Math. Helv. *49* (1974), 302–332.

Mathematisches Institut der
Universität Heidelberg
Im Neuenheimer Feld 288
D-6900 Heidelberg

Received November 29, 1979

Comment. Math. Helvetici **56** (1981) 132–135 0010-2571/81/001132-04$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

An elementary formula for the Fenchel–Nielsen twist

SCOTT WOLPERT[1]

A deformation of hyperbolic Riemann surfaces was investigated in the Fenchel–Nielsen manuscript. The surface is cut along a simple closed hyperbolic geodesic. The two sides are separated, rotated relative to each other and glued. The purpose of this note is to present a formula for the first variation under the twist deformation of the length of a closed geodesic. The formula is local in nature and entails the angles at the intersections of the geodesic and the cut.

The author is indebted to Troels Jørgensen for his encouragement and advice.

1. A variational formula

Let $f_t : R_0 \to R_t$ be a smooth deformation. Denote by g the hyperbolic line element on R_t. Choose a closed curve γ on R_0 and denote by $l(t)$ the length of the unique g geodesic freely homotopic to $f_t(\gamma)$ on R_t. Then

$$\frac{d}{dt} l(t) = \int_{\gamma_0} \frac{d}{dt} g_t \tag{1.1}$$

where $g_t = f_t^* g$ and γ_0 is the g geodesic on R_0.

Proof. Consider the lift $\tilde{f}_t$ of f_t to the universal covers, each represented as the upper half plane. Now $\tilde{f}_t$ is smooth in t and z, thus for A a deck transformation corresponding to γ_0 the Möbius transformation $A_t = \tilde{f}_t A \tilde{f}_t^{-1}$ varies smoothly in t. The endpoints on $\mathbb{R}$ of the axis of A_t (the fixed points of A_t) are algebraic in the matrix entries of A_t. Consequently the axes of the A_t vary smoothly in t. Now $\tilde{\gamma}_t = \tilde{f}_t^{-1}$ (axis of A_t) form a smooth 1-parameter family of curves. Denote by γ_t the projection to R_0 of $\tilde{\gamma}_t$. Then

$$\frac{d}{dt} l(t) = \frac{d}{dt} \int_{\gamma_t} g_t = \lim_{t \to 0} \frac{1}{t} \left(\int_{\gamma_t} g_t - \int_{\gamma_0} g_0 \right)$$

$$= \lim_{t \to 0} \frac{1}{t} \left(\int_{\gamma_t} g_t - \int_{\gamma_0} g_t + \int_{\gamma_0} g_t - \int_{\gamma_0} g_0 \right).$$

---

[1] Partially supported by NSF Grant #78-03838

Expanding in t, $g_t = g_o + t\dot{g} + O(t^2)$ where $\dot{g}^2$ is a symmetric 2-tensor. We then have

$$\frac{d}{dt} l(t) = \lim_{t\to 0} \frac{1}{t} \left(\int_{\gamma_t} g_0 - \int_{\gamma_0} g_0 \right) + \lim_{t\to 0} \left(\int_{\gamma_t} \dot{g} - \int_{\gamma_0} \dot{g} \right) + \int_{\gamma_0} \frac{d}{dt} g_t.$$

The curve γ_0 is a periodic geodesic and consequently the first variation of its length is zero; the second limit vanishes by continuity of the integral.

2. The Fenchel–Nielsen twist

Let $z \in H$, the upper half plane and $\theta = \arg z$. Choose $\phi(\theta)$ smooth with compact support in $(0, \pi)$, $\varphi \geq 0$ and $\int_0^\pi \varphi \, d\theta = \frac{1}{2}$. Define $\Phi(\theta) = \int_0^\theta \varphi \, d\theta$. The formula

$$w = z \exp(2t\Phi(\theta)) \tag{2.1}$$

defines a quasiconformal self map of H realizing the Fenchel–Nielsen twist deformation for the cyclic group stabilizing the imaginary axis. The boundary values of w are independent of the particular choice of φ. We recall that the boundary values determine the deformation as a point of the Teichmüller space. Now

$$w_z = \exp(2t\Phi)(1 - it\varphi), \qquad w_{\bar{z}} = \exp(2t\Phi)(it\varphi)$$

and the Beltrami differential of w is

$$\mu = w_{\bar{z}}/w_z = \frac{it\varphi}{1 - it\varphi} \frac{z}{\bar{z}}$$

The first variation of μ with respect to t is the tangent vector to the deformation.

Let Γ be a finitely generated Fuchsian group containing a hyperbolic element A stabilizing the imaginary axis. We assume A is primitive and that the imaginary axis projects to a simple closed geodesic. The infinitesimal Fenchel–Nielsen twist deformation of Γ about A is

$$\mu_\Gamma = \sum_{B \in \langle A \rangle \backslash \Gamma} B^* \mu,$$

for $\langle A\rangle$ the cyclic group generated by A, [2]. We emphasize that μ_Γ as a tangent to the deformation space is independent of the particular choice of φ.

3. The cosine formula

Let $C \in \Gamma$ correspond to a closed geodesic γ on $R = H/\Gamma$. The first variation of the length of γ under a Fenchel–Nielsen twist about the geodesic α corresponding to A is

$$\sum_{p \in \{\alpha\} \cap \{\gamma\}} \cos \beta_p.$$

The sum is over the intersection of the geodesics α and γ; β_p is the angle between the tangents at p of α and γ. The angle β_p is defined by the segment between 0 and p on the imaginary axis and that segment of the axis of C in the right half plane. The angle is well defined; note the expression $\cos \beta_p$ is anti-symmetric in the roles of α and γ. In particular for a twist about γ the contribution at p to the derivative of the length of α is $\cos(\pi - \beta_p)$.

We begin the proof by considering the solution w_t of the Beltrami equation $w_{\bar z} = \mu_\Gamma w_z$. The solution was constructed in [2]. Indeed, in a neighborhood of the imaginary axis w_t coincides with the map defined by (2.1). Now in the notation of section 1

$$g = |dw|/\operatorname{Im} w \quad \text{and}$$

$$g_t = |w_z|\,|dz + \mu_\Gamma\, d\bar z|/\operatorname{Im} w$$

$$= |1 - it\varphi|\,|dz + \mu_\Gamma\, d\bar z|/\operatorname{Im} z.$$

Our calculations only require an expression for g_t in a neighborhood of the imaginary axis, where $\mu_\Gamma = it\varphi e^{2i\theta} + O(t^2)$. Furthermore, φ will approximate $\frac{1}{2}$ the Dirac delta at $\pi/2$; consequently

$$g_t = |dz - it\varphi\, d\bar z|/\operatorname{Im} z + O(t^2, \text{ support } \varphi)$$

$$= ((dx - t\varphi\, dy)^2 + (dy - t\varphi\, dx)^2)^{1/2}/\operatorname{Im} z + O(t^2, \text{ support } \varphi).$$

Now differentiating

$$\frac{d}{dt}((dx - t\varphi\, dy)^2 + (dy - t\varphi\, dx)^2)^{1/2}/\operatorname{Im} z\big|_{t=0} = -2\varphi\, dx\, dy/(dx^2 + dy^2)^{1/2} \operatorname{Im} z.$$

We are now ready to apply formula (1.1). Let s be the arc length parameter along γ. Assume increasing s corresponds to moving to the left. The tangent of γ

at p is

$$\left(\frac{dx}{ds}, \frac{dy}{ds}\right)_p = \operatorname{Im} z(-\sin \beta_p, \cos \beta_p)_p.$$

The contribution to the integral is then $\int 2\varphi \sin \beta_p \cos \beta_p \, ds$. At p we have $\operatorname{Im} z \, d\theta = -dx = \operatorname{Im} z \sin \beta_p \, ds$ and finally the integral

$$\int 2\varphi \cos \beta_p \, d\theta.$$

Now the contribution of p to the integral (1.1) tends to $\cos \beta_p$ as φ approximates $\frac{1}{2}$ the Dirac delta. A check of the calculations shows the integral is indeed unchanged if s is replaced by $-s$. Finally each intersection of α and γ contributes such a term.

4. Remarks

Our calculations suggest that the appropriate Beltrami differential representing a Fenchel–Nielsen twist is a distribution. This is consistent with the point of view of the Fenchel–Nielsen twist found in the theory of amalgamation of Fuchsian groups.

Finally, we give an alternative formula for the cosine. Let the axis of two hyperbolic transformations A and B intersect at a point p. Let δ be the angle at p formed by the segments along the axes of A and B to their respective attractive fixed points. Then

$$\cos \delta = \frac{-\operatorname{sgn}(\operatorname{tr} A \operatorname{tr} B)(\operatorname{tr} A \operatorname{tr} B - 2 \operatorname{tr} AB)}{(\operatorname{tr}^2 A - 4)^{1/2}(\operatorname{tr}^2 B - 4)^{1/2}}$$

where tr denotes the trace of a matrix and sgn denotes the sign of a real number.

REFERENCES

[1] FENCHEL W. and NIELSEN, J., *Discontinuous groups of non-Euclidean motions*, unpublished manuscript.

[2] WOLPERT, S., *The Fenchel–Nielsen twist deformation*, preprint.

Department of Mathematics
University of Maryland
College Park, Maryland 20742

Received March 17, 1980

Comment. Math. Helvetici **56** (1981) 136–141 0010-2571/81/001136-06$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

$H^p(\mathbf{R}^n)$ is equidistributed with $L^p(\mathbf{R}^n)$[1]

STEVEN G. KRANTZ

Abstract. Let $0<p<\infty$. Let $H^p(\mathbf{R}^n)$ be the real variable Hardy spaces defined by Stein and Weiss. Let $L^p(\mathbf{R}^n)$ be the usual Lebesgue space. It is shown that for $f\in L^p$ there is an $\tilde{f}\in H^p$ with the distribution functions of $|f|$ and $|\tilde{f}|$ identical and $\|\tilde{f}\|_{H^p}\approx\|f\|_{L^p}$. The converse is trivially true.

§0

For $0<p<\infty$, let

$$L^p(\mathbf{R}^n)=\left\{f:\int_{\mathbf{R}^n}|f(x)|^p\,dx\equiv\|f\|_{L^p}^p<\infty\right\}.$$

Fix $\varphi\in C_c^\infty(\mathbf{R}^n)$, $\int\varphi(x)\,dx=1$. Let $\mathcal{S}(\mathbf{R}^n)$ be the Schwartz space, $\mathcal{S}'(\mathbf{R}^n)$ the Schwartz distributions, let $\varphi_\varepsilon(x)=\varepsilon^{-n}\varphi(x/\varepsilon)$, and define

$$f^*(x)=\sup_{\varepsilon>0}\varphi_\varepsilon * f(x),\qquad f\in\mathcal{S}'.$$

Let $\|f\|_{H^p}\equiv\|f^*\|_{L^p}$ for $f\in\mathcal{S}'$ and $H^p(\mathbf{R}^n)=\{f\in\mathcal{S}':\|f\|_{H^p}<\infty\}$. The space $H^p(\mathbf{R}^n)$ has a number of important equivalent characterizations, for which see [3].

Elements of the space of *distributions* $H^p(\mathbf{R}^n)$ may be represented by integration against $L^p(\mathbf{R}^n)$ functions satisfying certain *moment conditions* (see Section 1). It is with this in mind that all ensuing statements about H^p should be read.

Note that $\chi_{[0,1]}$ on $\mathbf{R}$ cannot represent an H^p function, $0<p\le 1$. So not all L^p functions represent H^p functions. Calderón–Zygmund operators are bounded on H^p, but not on L^p, $0<p\le 1$.

If $f:\mathbf{R}^n\to\mathbf{C}$ is measurable, let $m_f(\lambda)=|\{x:|f(x)|>\lambda\}|$, $\lambda>0$, where $|\ \ |$ denotes Lebesgue measure. Abusing terminology slightly, let us say that two functions f_1, f_2 are *equidistributed* if $m_{f_1}(\lambda)=m_{f_2}(\lambda)$ for all λ. Two function spaces X_1, X_2 on $\mathbf{R}^n$ are said to be equidistributed if to every $f_1\in X_1$ there corresponds an $f_2\in X_2$ so that f_1 and f_2 are equidistributed and *vice versa*. The main result of this paper is that

[1] Research partially supported by NSF Grant #MCS77-02213.

THEOREM A. *The spaces $H^p(\mathbf{R}^n)$ and $L^p(\mathbf{R}^n)$ are equidistributed, $0<p<\infty$. Indeed there are universal constants $C=C(p,n)$ so that each $f\in L^p$ is equidistributed with an $f'\in H^p$ satisfying $1/C\le\|f\|_{L^p}/\|f'\|_{H^p}\le C$.*

The proof of this result is an application of the atomic theory of the H^p spaces. The result emphasizes that the distinction between H^p and L^p for a given p is strictly a moment condition and does not involve size. The second inequality in the theorem is trivial with $C=1$. So it is the first inequality that we prove.

This work was motivated by a question of Colin Bennett. John Garnett independently discovered Theorem A for $q=1$, $n=1$.

§1. Proof of Theorem A

The atomic characterization of H^p proceeds as follows. Let $0<p\le 1$. Let

$$\mathcal{P}_k=\{\text{polynomials on } \mathbf{R}^n \text{ of degree not exceeding } k\},\qquad k=0,1,2,\ldots.$$

A measurable function $a:\mathbf{R}^n\to\mathbf{C}$ is said to be a *p-atom* if

$$a \text{ is supported on a ball } \overline{B(x,r)}=\{|y-x|\le r\} \tag{1.1}$$

$$|a|\le|B(x,r)|^{-1/p} \tag{1.2}$$

$$\int a(x)p(x)\,dx=0\qquad \forall p\in\mathcal{P}_{[(n/p)-n]}. \tag{1.3}$$

THEOREM 1.1 ([1], [2], [4]). *Let $0<p\le 1$. Let $f\in\mathcal{S}'(\mathbf{R}^n)$. Then $f\in H^p(\mathbf{R}^n)$ if and only if there is a sequence $\{a_i\}_{i=1}^\infty$ of p atoms and a sequence $\{\lambda_i\}_{i=1}^\infty\subseteq\mathbf{C}$ with $f=\sum\lambda_i a_i$ in the sense of distributions and*

$$(1/C)\cdot\sum|\lambda_i|^p\le\|f\|_{H^p}^p\le C\sum|\lambda_i|^p.$$

Here $C=C(n,p)$ is a universal constant.

Remark. Since, for $0<p\le 1$,

$$\int\left|\sum_i\lambda_i a_i(x)\right|^p dx\le\sum_i|\lambda_i|^p\int|a_1(x)|^p\,dx\le\sum|\lambda_i|^p\le C\|f\|_{H^p}^p,$$

it is possible to represent elements of H^p by L^p functions.

In order to prove Theorem A, it is enough to consider $0<p\leq 1$ and to prove that every $f\in L^p$ can be rearranged so that it is manifestly a linear combination of p atoms $\{a_i\}$ with coefficients $\{\lambda_i\}\in l^p$ satisfying $\|\{\lambda_i\}\|_{l^p}\approx\|f\|_{L^p}$.

Now the main technical result which is required for the proof of Theorem A is

PROPOSITION 1.2. *Let* $[a, b]\subseteq\mathbf{R}^1$ and let $f:[a, b]\to[0, \infty)$, $f=0$ *off* $[a, b]$, $f\in L^1(\mathbf{R})$. *Let* $0\leq k\in\mathbf{Z}$. *There is a measurable function* $\tilde{f}$ *on* $[0, (b-a)]$, *equidistributed with* f, *so that* $\int\tilde{f}(x)p(x)\,dx=0$ for all $p\in\mathscr{P}_k$.

Proposition 1.2 will be proved in Section 2. Taking it for granted, let us complete the proof of Theorem A. In order to simplify notation, the details will be given in $\mathbf{R}^2$ only. Fix $0<p\leq 1$. Let $f:\mathbf{R}^2\to\mathbf{C}$, $f\in L^p$. Assume for now that $|\text{supp } f|=1$. Let Nf be the non-increasing rearrangement of f (see [7]). So supp $Nf=[0, 1]$.

For $j\in\mathbf{Z}$, let $I_j=\{x: 2^j\leq Nf<2^{j+1}\}$. Then $[0, 1]=\cup I_j$, each I_j is an interval with endpoints $a_j\leq b_j$, and

$$\cdots a_1\leq b_1=a_0\leq b_0=a_{-1}\leq b_{-1}=\cdots.$$

Let $k=[(2/p)-2]$ and for each j apply Proposition 1.2 to $(Nf)|_{I_j}$ and $\mathscr{P}_k$. This yields, for each j, a function f^j on $[0, |I_j|]$ which is L^2 orthogonal to $\mathscr{P}_k$.

For each j, let $l_j=|I_j|\equiv b_j-a_j$. Write

$$0=\alpha_1^j<\beta_1^j=\alpha_2^j<\beta_2^j=\cdots=\alpha_{M_j}^j<\beta_{M_j}^j=1$$

where for each $i=1, \ldots, M_j-1$, $\beta_i^j-\alpha_i^j=l_j$, and $\beta_{M_j}^j-\alpha_{M_j}^j\leq l_j$. For each j and $1\leq i\leq M_j$, apply Proposition 1.2 to the function 1 on $[\alpha_i^j, \beta_i^j]$ and $\mathscr{P}_k$. Call the resulting function h_i^j on $[0, (\beta_i^j-\alpha_i^j)]$. Now define

$$\begin{aligned}\tilde{f}(x_1, x_2)\equiv&\sum_{j=-\infty}^{\infty}\sum_{i=1}^{M_j}\chi_{[0, l_j]}\left(x_1-\sum_{m=-\infty}^{j-1}l_m\right)\cdot f^j\left(x_1-\sum_{m=-\infty}^{j-1}l_m\right)\\&\times\chi_{[0,(\beta_i^j-\alpha_i^j)]}\left(x_2-\sum_{p=1}^{i-1}(\beta_i^j-\alpha_i^j)\right)h_i^j\left(x_2-\sum_{p=1}^{i-1}(\beta_i^j-\alpha_i^j)\right)\\\equiv&\sum_{j=-\infty}^{\infty}\sum_{i=1}^{M_j}\lambda_{ij}a_{ij}(x_1, x_2)\end{aligned}$$

where $\lambda_{ij}=2^j(l_j)^{2/p}$. Then $\tilde{f}$ is equidistributed with f since Nf is. Also, each a_{ij} is a p-atom. For each a_{ij} satisfies (i) a_{ij} is supported in a box of size $\leq l_j\times l_j$, (ii) each a_{ij} satisfies $|a_{ij}|\sim(l_j)^{-2/p}\sim|B(0, l_j)|^{-1/p}$, and (iii) each a_{ij} is orthogonal to $\mathscr{P}_k$ by Fubini's Theorem. In the verification of (iii) we have used the fact that if $p\in\mathscr{P}_k$

and $h \in R$ is fixed then the function $x \mapsto p(x+h)$ is in $\mathcal{P}_k$. Finally,

$$\sum_{i,j} |\lambda_{ij}|^p \leq \sum_j j_M 2^{jp} l_j^2 = \sum_j (2^{jp} l_j)(j_M l_j) \leq 2 \|f\|_{L^p}^p.$$

Likewise, $\sum |\lambda_{ij}|^p \geq \|f\|_{L^p}^p/4$. So $\tilde{f} \in H^p(\mathbf{R}^2)$ by Theorem 1.1, and $\|\tilde{f}\|_{H^p} \approx \|f\|_{L^p}$.

This completes the proof of Theorem A in case $|\operatorname{supp} f| = 1$. For the general case, write $f = \sum f_j$ where each f_j, except possibly one, has support of measure 1 and the odd one has measure not exceeding 1. Then each f_j gives rise to an H^p function f_j on $[0, 1] \times [0, 1]$. Let

$$\tilde{f}(x) = \sum_j \tilde{f}_j(x_1 - 4j, x_2 - 4j). \qquad \square$$

§2. Proof of Proposition 1.2

The proposition proceeds from some rather more technical lemmas about polynomials and about L^1 functions.

If $f: \mathbf{R} \to \mathbf{C}$, $h \in \mathbf{R}$, let

$$\Delta_h f(x) \equiv f(x+h) - f(x).$$

LEMMA 2.1. *If* $0 < k \in \mathbf{Z}$, $h \in \mathbf{R}$, *and* $p \in \mathcal{P}_k$, *then* $\Delta_h p \in \mathcal{P}_{k-1}$ (*where* $\mathcal{P}_{-1} = \{0\}$).

Proof. Apply the binomial theorem. $\square$

LEMMA 2.2. *Let* $h_1, \ldots, h_{k+1}$ *be non-zero real numbers with* $|h_j| > 2|h_{j+1}|$, $j = 2, 3, \ldots, k+1$. *Then for* $p \in \mathcal{P}_k$,

$$\Delta_{h_{k+1}}(\Delta_{h_k}(\ldots(\Delta_{h_1} p)\ldots) \equiv 0, \tag{2.2.1}$$

and the expression on the left side of (2.2.1) *may be written as*

$$\sum_{j=1}^{2^{k+1}} \varepsilon_j p(x + a_j) \tag{2.2.2}$$

with $\varepsilon_j = \pm 1$ *for each* j *and the* a_j *distinct.*

Proof. This follows from 2.1 by induction. □

Remark. By choosing the h_j in Lemma 2.2 adroitly, we may arrange that the a_j are equally spaced, with any preselected distance d_0 between successive a_j's. By renaming the a_j's if necessary, and by the translation invariance of Lemma 2.2, we may suppose that

$$0 > a_1 > \cdots > a_{2^{k+1}}$$

where $a_{j-1} - a_j = d_0$, $j = 2, \ldots, 2^{k+1}$.

DEFINITION 2.3. If $d_0 > 0$, $0 \leq k \in \mathbf{Z}$, and $a_1, \ldots, a_{2^{k+1}}$, $\varepsilon_1, \ldots, \varepsilon_{2^{k+1}}$ are selected according to Proposition 2.2 and the subsequent remark, let $T_k^{d_0} f(x) \equiv \sum_{j=1}^{2^{k+1}} \varepsilon_j f(x + a_j)$, any $f : \mathbf{R} \to C$.

LEMMA 2.4. *Let* $f : [0, 1] \to [0, \infty)$, $f \in L^1(\mathbf{R})$. *Let* $0 < M \in \mathbf{Z}$. *There exists a function* $f_M : [0, 1/M] \to [0, \infty)$ *such that* f *is equidistributed with* $g(x) = \sum_{j=1}^{M} f_M(x - j/M)$. *In particular,* $m_f(\lambda) = M \cdot m_{f_M}(\lambda)$ *for every* $\lambda > 0$.

Proof. Let $f_M(x) = f(Mx)$. □

Proof of Proposition 1.2. Assume without loss of generality that $a = 0$, $b = 1$. Let other notation be as in the statement of the proposition. Apply Lemma 2.4 with $M = 2^{k+1}$. The resulting function f_M is supported on $[0, 2^{-k-1}]$. Let $d_0 = 2^{-k-1}$. Using Definition 2.3 let $\tilde{f} = T_k^{d_0} f_M$. Then $\tilde{f}$ is equidistributed with f and $\operatorname{supp} \tilde{f} \subseteq [0, 1]$. Then if $p \in \mathcal{P}_k$ we have, letting $R(x) = -x$, that

$$\int p(x) \tilde{f}(x)\, dx = \int p(x)(T_k^{d_0} f_M)(x)\, dx$$

$$(\text{ch. of variable}) = \int R(T_k^{d_0} Rp)(x) f_M(x)\, dx.$$

But $Rp \in \mathcal{P}_k$ so $T_k^{d_0} Rp = 0$ whence the last line is 0 as desired. □

§3. Concluding remarks

The proof we have given of Theorem A uses the structure of $\mathbf{R}^n$ rather decisively. With considerable additional technical difficulty, a proof of the same kind appears to work for the case when H^p is the non-isotopic Kähler H^p on the boundary of the unit ball $B \subseteq C^n$ (see [4]) and L^p is the usual L^p space with

respect to rotationally invariant measure on ∂B. It would be interesting to know to what extent Theorem A, or a modification thereof, holds for the H^p spaces defined on certain spaces of homogeneous type ([2], [5]).

REFERENCES

[1] COIFMAN, R., *A real variable characterization of* H^p, Studia Math. *51* (1974), 269–274.
[2] COIFMAN, R. and WEISS, G., *Extensions of Hardy spaces and their use in analysis*, BAMS *38* (1977), 569–645.
[3] FEFFERMAN, C. and STEIN, E. M., H^p *Spaces of several variables*, Acta Math. *129* (1972), 137–193.
[4] GARNETT, J. and LATTER, R., *The atomic decomposition for Hardy spaces in several complex variables*, Duke Math. Jour. *45* (1978), 815–846.
[5] LATTER, R., Thesis, UCLA, 1977.
[6] STEIN, E. M., *Singular integrals and differentiability properties of functions*, Princeton Univ. Press, 1970.
[7] STEIN, E. M. and WEISS, G., *Introduction to Fourier analysis on Euclidean spaces*, Princeton Univ. Press, 1971.

Department of Mathematics
University of California
Los Angeles, CA 90024/*USA*

Received July 18, 1980

Added in Proof: Results related to, but distinct from, Theorem A, have recently appeared in "Hardy Spaces and Rearrangements," *Trans. A.M.S. 261* (1980), 211–233.

Comment. Math. Helvetici **56** (1981) 142–147 0010-2571/81/001000-06$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Many different disk knots with the same exterior

L. R. Hitt[1] and D. W. Sumners

§1. Introduction

Much of codimension-two knot theory is concerned with finding and computing topological invariants of knot exteriors in order to distinguish between the knots themselves. It is well-known ([G], [L-S], [B]) that there are at most two inequivalent smooth n-sphere knots with the same exterior ($n \geq 2$), and examples of two inequivalent n-knots with the same exterior have recently been discovered ([C-S], [Go]). We show that the corresponding theory for $(n+1)$-disk knots is more complicated. Let Y denote the bounded exterior of a smooth $(n+1)$-disk knot. The *indeterminacy index* $\zeta(Y)$ is the number of inequivalent $(n+1)$-disk pairs having exteriors diffeomorphic to Y. We show that there exist disk knots with large indeterminacy indices (bigger than two, in particular). We then show that $\zeta(Y) \leq 2|\pi'|$, *where* $|\pi'|$ denotes the cardinality of π', the commutator subgroup of $\pi = \pi_1(\partial Y)$. This yields as a corollary a new and easy proof of the well-known fact that $\zeta(X) \leq 2$, where X is the exterior of an n-sphere knot, and $\zeta(X)$ its indeterminacy index.

§2. The indeterminacy index

For convenience, we work in the smooth category (the same results hold in the locally flat PL situation). We let S^n and D^{n+1} denote the standard n-sphere and $(n+1)$-disk, respectively. An *n-sphere knot* (or just *n-knot*) is the pair (S^{n+2}, kS^n) where $k : S^n \to S^{n+2}$ is an embedding. The *exterior* X of an n-knot is the complement in S^{n+2} of an open trivial 2-disk bundle neighborhood of the submanifold kS^n. An *$(n+1)$-disk knot* is the pair (D^{n+3}, gD^{n+1}) where $g : D^{n+1} \to D^{n+3}$ denotes a proper embedding, one in which the submanifold gD^{n+1} intersects ∂D^{n+3} transversely in $g(\partial D^{n+1})$. We let Y denote the $(n+1)$-disk knot exterior. Two knots are *equivalent* if there is a diffeomorphism of the ambient space throwing one submanifold onto the other (we disregard orienta-

[1] Research partially supported by the University of South Alabama Research Committee.

tions), and the *indeterminacy index* ζ is the number of inequivalent knots determined by a given knot exterior.

We will now produce examples to show that $\zeta(Y)$ can be large. The reason for this is that ∂Y contains the exterior X of the boundary sphere pair, and X can be very complicated. Recall the example of Kato [Ka 2, Theorem 4.9]:

Let $n \geq 3$, and M^{n+2} be a contractible manifold such that $\pi_1(\partial M)$ is the binary icosohedral group $G = \langle a, b \mid a^5 = b^3 = (ab)^2 \rangle$ [Ke]. Let $Y^{n+3} = S^1 \times M^{n+2}$; we will show that Y is the exterior of at least three inequivalent $(n+1)$-disk knots. Then by modifying the construction, we will show that the indeterminacy index of a disk knot exterior can be at least as large as six.

Let H be a group. A *weight element* of H is an element whose normal closure is all of H. The *automorphism class* of an element of H is the orbit of the element under the automorphism group of H. Two elements of H are *algebraically distinct* if they are in different automorphism classes.

We are interested in finding different automorphism classes of weight elements in the group $\pi_1(\partial Y) \cong Z \times G \cong \langle t, a, b \mid a^5 = b^3 = (ab)^2, ta = at, tb = bt \rangle$ where Z denotes the infinite cyclic group generated by t. An element of the form $t^n g$, for $g \in G$, is a weight element of $Z \times G$ if and only if t^n is a weight element of Z and g is a weight element of G, which forces $n = \pm 1$. To determine the weight elements of G, note that $\{1\} \lhd \{1, (ab)^2\} \lhd G$ is a composition series for G, since $\langle a, b \mid a^5 = b^3 = (ab)^2 = 1 \rangle$ is a presentation of the simple group A_5. The center of G is $C(G) = \{1, (ab)^2\}$, the cyclic group of order 2. Any element of G which is not in $C(G)$ is a weight element of G. The set of algebraically distinct weight elements of G is $\{a, a^2, b, b^2, ab\}$. That they are algebraically distinct follows from their different orders: 10, 5, 6, 3 and 4, respectively.

Therefore we have ta, ta^2, tb, tb^2, and tab as weight elements of $Z \times G$. However ta and ta^2 are in the same automorphism class in $Z \times G$, as are tb and tb^2 (e.g., the automorphism θ, induced by $\theta(t) = t(ab)^2$, $\theta(a) = a^7$, and $\theta(b) = ba^8b$ sends ta to ta^2). So our list of possibly algebraically distinct weight elements is shortened to ta, tb, and tab. That these three elements are algebraically distinct follows from the fact that the center of $Z \times G$ is $Z \times \{1, (ab)^2\}$, so $Z \times G$ modulo its center is $A_5 \cong \langle a, b \mid a^5 = b^3 = (ab)^2 = 1 \rangle$. But the center is a characteristic subgroup, so any automorphism of $Z \times G$ induces one on A_5. Since a, b, and ab have different orders in A_5, their counterparts in $Z \times G$ must be algebraically distinct.

Let $\{\sigma_i \mid 1 \leq i \leq 3\}$ denote smooth embeddings of S^1 in ∂Y representing the homotopy class in ∂Y of each of the above weight elements of $Z \times G$. Choose a trivialization of the normal bundle of each σ_i, and attach 2-handles to form the manifolds $Y \cup_{\sigma_i} h^2$. The cocore or transverse disk of each 2-handle is an $(n+1)$-disk, and $(Y \cup_{\sigma_i} h^2, \text{ cocore } (h^2)) \approx (D^{n+3}, g_i d^{n+1})$, where $g_i : D^{n+1} \to D^{n+3}$ is a

proper smooth embedding. This is because $Y \cup_{\sigma_i} h^2$ is contractible, with simply-connected boundary, and $n+3 \geq 6$. However, no two of the three disk pairs $(Y \cup_{\sigma_i} h^2, g_i D^{n+1})$ are equivalent, because any diffeomorphism of pairs between them would restrict to a diffeomorphism on Y, inducing an isomorphism on $\pi_1(\partial Y)$ taking one of the weight elements of $Z \times G$ to another, or its inverse.

In [S], it is shown that $(n+1)$-disk pairs $(n \geq 2)$ can be constructed with an arbitrarily prescribed Alexander polynomial in a single dimension p $(2 \leq p \leq n)$, and trivial Alexander polynomials elsewhere. Moreover, these disk pairs have the property that $\pi_1(Y) \cong \pi_i(\partial Y) \cong \pi_i(S^1)$ for $i < p$. Thus, by taking the boundary connected sum of the above examples with these disk pairs, one obtains infinitely many distinct $(n+1)$-disk exteriors, each with indeterminacy index $\zeta \geq 3$. This proves

THEOREM 2.1. *For each $n \geq 3$, there exist infinitely many homeomorphically distinct $(n+1)$-disk knot exteriors Y_i, each with indeterminacy index $\zeta(Y_i) \geq 3$.*

Remark. The analogue of Theorem 2.1 for $n = 2$ can be done in the topological category (non-PL embeddings). One takes $Y = S^1 \times (c * \Sigma^3)$, where $c * \Sigma^3$ is the cone on Σ^3, the Poincare' 3-sphere. Then Y is a topological manifold [Ca], and arguments of Scharlemann [Sc] can be used to prove that the various handle attachments give rise to different non-PL disk pairs (D^5, gD^3).

We can modify the above construction to increase the lower bound for the indeterminacy index. Consider the group

$$\begin{aligned} G \times G \times G = \langle a, b, c, d, e, f \mid\ & a^5 = b^3 = (ab)^2,\ c^5 = d^3 = (cd)^2,\ e^5 = f^3 = (ef)^2, \\ & ac = ca,\ ad = da,\ bc = cb,\ bd = db,\ ae = ea,\ af = fa,\ be = eb,\ bf = fb, \\ & ce = ec,\ cf = fc,\ de = ed,\ df = fd \rangle. \end{aligned}$$

Now $G \times G \times G$ is finitely presented, and $H_1(G \times G \times G) = H_2(G \times G \times G) = 0$, so by Kervaire [Ke], for $n \geq 4$ there exists a contractible manifold M^{n+2} with $\pi_1(\partial M) \cong G \times G \times G$. As before, $Y = S^1 \times M$, and $\pi_1(\partial Y) \cong Z \times G \times G \times G$. Since the center of $Z \times G \times G \times G$ is the product of the center of each factor, we see that $Z \times G \times G \times G$ modulo its center is $A_5 \times A_5 \times A_5$. Then, as before, *tacf*, *tacef*, *tbdef*, *tace*, *tbdf*, and *tabcdef* are all algebraically distinct since their projections modulo the center have orders 15, 10, 6, 5, 3, and 2, respectively. We have the following

COROLLARY 2.2. *For each $n \geq 4$, there exist infinitely many homeomorphically distinct $(n+1)$-disk knot exteriors Y_i, each with indeterminacy index $\zeta(Y_i) \geq 6$.*

§3. An upper bound for the indeterminacy index

Now that we have seen that in some cases the lower bound of ζ can be large, we are interested in finding upper bounds. Along these lines, we have the following

THEOREM 3.1. *Let* Y^{n+3} *be an* $(n+1)$*-disk knot exterior* $(n \geq 2)$. *Then* $\zeta(Y) \leq 2\,|\pi'|$, *where* $|\pi'|$ *denotes the cardinality of the commutator subgroup* π' *of* $\pi = \pi_1(\partial Y)$.

Proof. Consider the disk pair (D^{n+3}, gD^{n+1}). Choose a trivialization $G : D^2 \times D^{n+1} \to N(gD^{n+1})$ of the tubular neighborhood of the submanifold; thus $G(\{0\} \times y) = g(y)$ for $y \in D^{n+1}$. We have that the exterior $Y = D^{n+3} - G(\mathring{D}^2 \times D^{n+1})$. Regarding $N(gD^{n+1})$ as a 2-handle attached to Y via the meridian attaching curve $G(\partial D^2 \times \{0\})$, we have $(D^{n+3}, gD^{n+1}) \approx (Y \cup_G h^2, \text{cocore}\,(h^2))$. We now wish to study the number of different ways it is possible to attach a 2-handle to Y to produce D^{n+3}. We first count the maximum number of possible isotopy classes in ∂Y of attaching curves for a 2-handle which produce a contractible manifold after handle attachment is performed. If $\pi = \pi_1(\partial Y)$, and π' is the commutator subgroup of π, we have the short exact sequence

$$1 \to \pi' \to \pi \to Z \to 1. \tag{3.2}$$

Denoting the generator of the infinite cyclic multiplicative group by t, we have a semi-direct product structure for π, and once a splitting for (3.2) is chosen, we can write each element $x \in \pi$ uniquely as $x = t^a g$ where a is an integer and $g \in \pi'$. By abuse of notation, let $t^a g$ represent an embedding of S^1 in the same homotopy class, and choose a trivialization of its normal bundle. In order for $Y \cup_{t^a g} h^2$ to be acyclic, we must have $a = \pm 1$, because $H_1(Y; Z)$ is infinite cyclic on the generator t. In order for $Y \cup_{t^a g} h^2$ to be contractible, $i_*(t^a g)$ must be a weight element of $\pi_1(Y)$, where $i_* : \pi_1(\partial Y) \to \pi_1(Y)$ is the inclusion homomorphism. In order for $\partial(Y \cup_{t^a g} h^2)$ to be simply-connected, $t^a g$ must be a weight element of $\pi_1(\partial Y)$. The upper bound we are aiming at is very crude, coming just from the homology condition $(a = \pm 1)$, so we are in fact counting the ways it is possible to complete Y to obtain an integral homology disk. The set of elements of $\pi_1(\partial Y)$ producing acyclic manifolds upon handle attachment is $\{t^{\pm 1} g \mid g \in \pi'\}$. But since the sign of the exponent of t in an element of $\pi_1(\partial Y)$ is reversed by changing the orientation of the attaching curve of h^2 (or equivalently, reversing the orientation on the cocore D^{n+1}), the set of elements of π corresponding to possibly different manifold pairs is $\{tg \mid g \in \pi'\}$, a set of the cardinality of π'. Now since we are in the

dimension range $(n+2) \geq 4$ for ∂Y, homotopy of embedded one-spheres gives rise to isotopy, so the number of possible isotopy classes of attaching curves in ∂Y giving rise to acyclic manifolds is bounded above by $|\pi'|$. Now, given a represenatative of an isotopy class of attaching curves in ∂Y, there are precisely two ways to attach the 2-handle h^2, corresponding to the $\pi_1(SO) = Z_2$ ways of choosing a trivialization of the normal bundle of the curve. Hence the number of possible handle attachments yielding acyclic manifolds is bounded above by $2|\pi'|$.

COROLLARY 3.3. *Suppose that* Y^{n+3} $(n \geq 2)$ *is an* $(n+1)$*-disk knot exterior, and that* $\pi_1(\partial Y) = Z$. *Then* $\zeta(Y) \leq 2$, *and the two possibly different disk pairs are obtained, each from the other, by re-attaching the* 2*-handle corresponding to the normal bundle over the submanifold via the non-trivial element of* $\pi_1(SO)$.

Corollary 3.3 yields an easy proof of the well-known result that there are at most two inequivalent n-knots with the same exterior:

COROLLARY 3.4. ([B], [L–S], [Ka 1], [Sw]). *Let* X^{n+2} $(n \geq 3)$ *be an* n*-sphere exterior. Then* $\zeta(X) \leq 2$. *Moreover, if* $(X \cup_\gamma (D^2 \times S^n), \{0\} \times S^n)$ *denotes a sphere pair obtained by sewing* $D^2 \times S^n$ *onto* X *via some trivialization of the* S^n*-bundle over the meridian curve* $\gamma = S^1 \times \{*\} \subset \partial X$, *then the possibly different sphere pair is* $(X \cup_{\tilde{\gamma}} (D^2 \times S^n), \{0\} \times S^n)$, *where* $\tilde{\gamma}$ *denotes the same meridian curve with different trivialization of the* S^n*-bundle* (i.e., $D^2 \times S^n$ *is sewn in with a* $\pi_1(SO)$*-twist*).

Proof. There is a one-to-one correspondence between n-sphere knots and n-disk knots with unknotted boundary $(n-1)$-sphere pair, obtained by removing an unknotted disk pair (the neighborhood of a point on the submanifold) from the sphere pair to obtain the required disk pair. An n-sphere knot and its corresponding n-disk knot have the same exterior X. But $\partial X \approx S^1 \times S^n$, and $\pi_1(\partial X) = Z$, so by Corollary 2.4, $\zeta(X) \leq 2$. That is, X (thought of as a disk exterior) determines at most two inequivalent disk pairs. Therefore, thinking of it as a sphere pair exterior, then $\zeta(X) \leq 2$ as well.

§4. Some questions

1. Given a positive integer N, does there exist an $(n+1)$-disk exterior Y with $\zeta(Y) \geq N$?
2. Is there an $(n+1)$-disk exterior Y with $\zeta(Y) = +\infty$?
3. If X is an n-sphere exterior and $\pi_1(X) = Z$, must if follow that $\zeta(X) = 1$?

REFERENCES

[B] W. BROWDER, *Diffeomorphisms of 1-connected manifolds*, Trans. Amer. Math. Soc. 128 (1967), 155–163.

[Ca] J. W. CANNON, *Shrinking cell-like decompositions of manifolds: codimension three*, Ann. of Math. *110* (1979), 83–112.

[C–S] S. CAPELL and J. L. SHANESON, *There exist inequivalent knots with the same complement*, Ann. of Math. *103* (1976), 349–353.

[G] H. GLUCK, *The embedding of two-spheres in the four-sphere*, Trans. Amer. Math. Soc. *104* (1962), 308–333.

[Go] C. M. GORDON, *Knots in the 4-sphere*, Comment. Math. Helv. *51* (1976), 585–596.

[Ka 1] M. KATO, *A concordance classification of* PL *homeomorphisms of* $S^p \times S^q$, Topology *8* (1969), 371–383.

[Ka 2] M. KATO, *Higher dimensional* PL *knots and knot manifolds*, Jour. Math. Soc. Jap. *21* (1969), 458–480.

[Ke] M. KERVAIRE, *Smooth homology spheres and their fundamental groups*, Trans. Amer. Math. Soc. *144* (1969), 67–72.

[L–S] R. K. LASHOF and J. S. SHANESON, *Classification of knots in codimension two*, Bull. Amer. Math. Soc. *75* (1969), 171–175.

[Sc] M. SCHARLEMANN, *Stably trivial non-*PL *knotted ball pairs*, Notices Amer. Math. Soc. *23* (1976), pg. A–22, Abstract #76T-G5.

[S] D. W. SUMNERS, *Homotopy torsion in codimension two knots*, Proc. Amer. Math. Soc. *24* (1970), 229–240.

[Sw] G. A. SWARUP, *A note on higher-dimensional knots*, Math. Zeit. *121* (1971), 141–144.

University of South Alabama
Mobile, Al 36688

Florida State University
Tallahassee, FL 32306

Received November 22, 1979/November 30, 1980

Comment. Math. Helvetici **56** (1981) 148–161 0010-2571/81/001000-14$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

An asymptotic series for the mean value of Dirichlet L-functions

D. R. HEATH-BROWN

1. Introduction

Let χ denote a typical Dirichlet character (mod q), where $q>1$, and let $L(s, \chi)$ be the corresponding L-function. These functions have many applications to the distribution of primes and other arithmetic objects in the various arithmetic progressions to modulus q. For such applications one frequently needs information on the mean values

$$\sum_{\chi(\mathrm{mod}\, q)} |L(\tfrac{1}{2}+it, \chi)|^{2k} \tag{1}$$

and

$$\sum_{\chi(\mathrm{mod}\, q)} \int_0^T |L(\tfrac{1}{2}+it, \chi)|^{2k}\, dt \tag{2}$$

for a positive integer k. Although one requires a certain degree of uniformity in t or T, an arbitrary power of $2+|t|$ or $2+T$ in any error terms will usually suffice. Moreover it is normally sufficient to have a good upper bound rather than an asymptotic formula. In this respect the estimate of Montgomery [4; Theorem 10.1], namely

$$\sum_{\chi(\mathrm{mod}\, q)}^{*} \int_0^T |L(\tfrac{1}{2}+it, \chi)|^4\, dt \ll \phi(q) T(\log qT)^4, \qquad (T \geqslant 2),$$

where $\sum^*$ denotes summation over primitive characters only, is the best currently available. None the less, it is of theoretical interest to investigate the mean values (1) and (2) more closely, and it is this line of enquiry we shall pursue here, by examining the behaviour of the sum (1), with $k=1$, for large values of q. There are many respects in which the behaviour of $\zeta(\frac{1}{2}+it)$, for varying t, resembles that of $L(\frac{1}{2}+it, \chi)$ with respect to q. In particular one might expect the behaviour of

(1), when $k=1$, to be similar to that of

$$I=\int_0^T |\zeta(\tfrac{1}{2}+it)|^2\,dt.$$

This is in fact not so, for it is known that

$$I=T\left(\log\frac{T}{2\pi}+2\gamma-1\right)+E(T),$$

in which, by Balasubramanian [1] we have $E(T)\ll T^{\frac{1}{3}+\varepsilon}$, and by Good [2] we have $E(T)=\Omega(T^{1/4})$. In contrast we shall prove:—

THEOREM. *We have*

$$\sum_{\chi(\mathrm{mod}\,q)} |L(\tfrac{1}{2},\chi)|^2=\frac{\phi(q)}{q}\sum_{k\mid q}\mu(q/k)T(k), \tag{3}$$

where $T(k)$ *has the asymptotic expansion*

$$T(k)=k\left(\log\frac{k}{8\pi}+\gamma\right)+2\zeta(\tfrac{1}{2})^2k^{1/2}+\sum_{n=0}^{2N-1}c_nk^{-n/2}+O(k^{-N}), \tag{4}$$

for any $N\geqslant 1$. *Here the* c_n *are numerical constants and* γ *is Euler's constant.*

COROLLARY. *If* $q=p$ *is prime then we have an asymptotic expansion*

$$\sum_{\chi(\mathrm{mod}\,p)} |L(\tfrac{1}{2},\chi)|^2=(p-1)\left(\log\frac{p}{8\pi}+\gamma\right)+2\zeta(\tfrac{1}{2})^2p^{1/2}+\sum_{n=0}^{2N-1}d_np^{-n/2}+O(p^{-N}), \tag{5}$$

for any $N\geqslant 1$.

The behaviour is therefore much closer to that of the mean value

$$\int_0^\infty |\zeta(\tfrac{1}{2}+it)|^2\,e^{-2t/T}\,dt,$$

which has the asymptotic series

$$\frac{\log\dfrac{T}{4\pi}+\gamma}{2\sin\dfrac{1}{T}}+\sum_{n=0}^{N-1}c_n'T^{-n}+O(T^{-N}),$$

for which see Kober [3]. However one should note the occurrence of the half-integer powers in the expansion (4). Note also that for general q the error terms in (4) will decrease at different rates for the different factors k of q. Thus we cannot give an asymptotic series for the sum (3) itself, except when, as in the corollary, q has no small prime factors.

Instead of examining sums of the form (3) one could treat the expressions

$$\sum_{\chi (\mathrm{mod}\, q)}^{*} |L(\tfrac{1}{2}+it, \chi)|^2,$$

or

$$\sum_{\chi (\mathrm{mod}\, q)} |L(\tfrac{1}{2}+it, \chi^*)|^2,$$

where χ^* is the primitive character which induces χ. However these present additional difficulties without simplifying the form of the results. Thus if $q = p$ is prime and $t = 0$, for example, the above sums differ from (5) by $O(1)$, so that the $p^{1/2}$ term will still be present.

I should like to thank the Eidgenössiche Technische Hochschule, Zürich, for their hospitality during the period in which this paper was written.

2. Functional equations

Let

$$T_1(q, s) = \sum_{\chi (\mathrm{mod}\, q)} L(s, \chi) L(1-s, \bar{\chi}).$$

Here $L(s, \chi)$ is holomorphic for non-principal χ. If χ is principal and $q \neq 1$ then $L(s, \chi)$ will have a simple pole at $s = 1$ and $L(1-s, \bar{\chi})$ will have a zero; moreover $L(s, \chi)$ will be regular elsewhere. Thus $T_1(q, s)$ is holomorphic for $q > 1$. We shall also need to know that $T_1(q, s)$ satisfies the growth condition

$$T_1(q, \sigma+it) \ll (1+|t|)^2, \qquad (-1 \leqslant \sigma \leqslant 2) \tag{6}$$

for fixed q; but this follows from a trivial bound for $L(s, \chi)$. Finally, since $\bar{\chi}$ runs over all characters (mod q) as χ does, we find

$$T_1(q, s) = T_1(q, 1-s). \tag{7}$$

We now express $L(s, \chi)$ in terms of the Hurwitz Zeta-function

$$\zeta(s, \alpha) = \sum_{n=0}^{\infty} (n+\alpha)^{-s}, \qquad (0 < \alpha \leqslant 1,\ \mathrm{Re}\,(s) > 1)$$

and apply the functional equation (for which see Titchmarsh [5; §2.17]) for $\zeta(1-s, \alpha)$. This yields for $\mathrm{Re}\,(s) > 1$,

$$L(s, \chi)L(1-s, \bar{\chi}) = \left(q^{-s} \sum_{u=1}^{q} \chi(u)\zeta(s, u/q)\right)\left(q^{s-1} \sum_{v=1}^{q} \bar{\chi}(v)\zeta(1-s, v/q)\right)$$
$$= q^{-1} \sum_{u,v} \chi(u)\bar{\chi}(v)\zeta(s, u/q)\left\{\frac{2\Gamma(s)}{(2\pi)^s} \sum_{l=1}^{\infty} \frac{\sin(2\pi lv/q + \pi(1-s)/2)}{l^s}\right\}.$$

We now sum over χ, using the orthogonality relation

$$\sum_{\chi (\mathrm{mod}\, q)} \chi(u)\bar{\chi}(v) = \begin{cases} \phi(q), & u \equiv v (\mathrm{mod}\, q), \quad (uv, q) = 1, \\ 0, & \text{otherwise.} \end{cases}$$

We then obtain

$$T_1(q, s) = \frac{\phi(q)}{q} \sum_{\substack{v=1 \\ (v,q)=1}}^{q} \zeta(s, v/q)\left\{\frac{2\Gamma(s)}{(2\pi)^s} \sum_{l=1}^{\infty} \frac{\sin(2\pi lv/q + \pi(1-s)/2)}{l^s}\right\}$$
$$= \frac{\phi(q)}{q} 2\Gamma(s)\left(\frac{q}{2\pi}\right)^s \sum_{\substack{v=1 \\ (v,q)=1}}^{q} \sum_{\substack{h=1 \\ h \equiv v (\mathrm{mod}\, q)}}^{\infty} h^{-s} \sum_{l=1}^{\infty} \frac{\sin(2\pi lv/q + \pi(1-s)/2)}{l^s}$$
$$= \frac{\phi(q)}{q} 2\Gamma(s)\left(\frac{q}{2\pi}\right)^s \sum_{\substack{h=1 \\ (h,q)=1}}^{\infty} h^{-s} \sum_{l=1}^{\infty} \frac{\sin(2\pi lh/q + \pi(1-s)/2)}{l^s}$$
$$= \frac{\phi(q)}{q} 2\Gamma(s)\left(\frac{q}{2\pi}\right)^s \sum_{h=1}^{\infty} \left(\sum_{\substack{d|q \\ d|h}} \mu(d)\right) h^{-s} \sum_{l=1}^{\infty} \frac{\sin(2\pi lh/q + \pi(1-s)/2)}{l^s}$$
$$= \frac{\phi(q)}{q} 2\Gamma(s)\left(\frac{q}{2\pi}\right)^s \sum_{d|q} \mu(d) d^{-s} \sum_{j,l=1}^{\infty} (jl)^{-s} \sin(2\pi djl/q + \pi(1-s)/2).$$

Now writing $d = q/k$ we see that

$$T_1(q, s) = \frac{\phi(q)}{q} \sum_{k|q} \mu(q/k) T(k, s), \tag{8}$$

where

$$T(k, s) = 2\Gamma(s)\left(\frac{k}{2\pi}\right)^s \sum_{n=1}^{\infty} d(n)n^{-s} \sin(2\pi n/k + \pi(1-s)/2). \tag{9}$$

Clearly the sum on the left of (3) is $T_1(q, \frac{1}{2})$.

3. A weighted sum

The sum (9) for $T(k, s)$ does not converge absolutely if $s = \frac{1}{2}$, so we use the functional equation (7) to produce a weighted sum. Let

$$F(s) = s^{-1} \cos(\pi s) \exp(s^2).$$

Then $F(s)$ is an odd function of s and is holomorphic, except for a simple pole at the origin, of residue 1. Moreover it satisfies the growth condition

$$F(\sigma + it) \ll e^{-2|t|}, \qquad |t| \geqslant 1 \tag{10}$$

in any fixed vertical strip. Now consider the integral

$$I = \frac{1}{2\pi i} \int_{(1)} T_1(q, s+\tfrac{1}{2}) F(s)\, ds,$$

where the symbol (σ) denotes integration along a line from $\sigma - i\infty$ to $\sigma + i\infty$. After moving the line of integration to (-1) and allowing for the pole at $s = 0$, we obtain

$$I = T_1(q, \tfrac{1}{2}) + \frac{1}{2\pi i} \int_{(-1)} T_1(q, s+\tfrac{1}{2}) F(s)\, ds. \tag{11}$$

Then, using (7) together with the relation $F(-s) = -F(s)$, we deduce that $T_1(q, \frac{1}{2}) = 2I$.

We evaluate I by termwise integration, using (8) and (9). This yields

$$T_1(q, \tfrac{1}{2}) = 2I = \frac{\phi(q)}{q} \sum_{k \mid q} \mu(q/k) T(k),$$

where

$$T(k)=4\sum_{1}^{\infty}d(n)K_1\left(\frac{2\pi n}{k}\right) \tag{12}$$

and

$$\begin{aligned}K_1(x)&=\frac{1}{2\pi i}\int_{(1)}\Gamma(s+\tfrac{1}{2})x^{-s-\frac{1}{2}}\sin\left(x+\frac{\pi}{4}-\frac{\pi s}{2}\right)F(s)\,ds\\&=\mathrm{Re}\left\{\frac{1}{2\pi i}\int_{(1)}\Gamma(s+\tfrac{1}{2})x^{-s-\frac{1}{2}}e^{ix-i\pi/4-i\pi s/2}F(s)\,ds\right\}\\&=\mathrm{Re}\,(K(x)),\end{aligned}$$

say. Here

$$K(x)=e^{-i\pi/4}e^{ix}x^{-1/2}J(x) \tag{13}$$

and

$$J(x)=\frac{1}{2\pi i}\int_{(1)}\Gamma(s+\tfrac{1}{2})x^{-s}e^{-i\pi s/2}F(s)\,ds. \tag{14}$$

Our choice of $F(s)$ ensures that the integrand of $J(x)$ is holomorphic except for a simple pole at $s=0$, with residue $\Gamma(\frac{1}{2})=\pi^{1/2}$. Thus, on moving the line of integration to $(\pm M)$, where $M>0$, we find

$$J(x)\ll x^{-M},\qquad (x\geqslant 1), \tag{15}$$

$$J(x)-\pi^{1/2}\ll x^{M},\qquad (0<x\leqslant 1), \tag{16}$$

for any $M>0$. Similarly, after differentiating under the sign l (>0) times, we obtain

$$J^{(l)}(x)\ll\begin{cases}x^{-M}, & (x\geqslant 1),\\ x^{M}, & (0<x\leqslant 1),\end{cases} \tag{17}$$

there now being no pole at $s=0$. From these bounds it follows in particular that

$$K(x)\ll\begin{cases}x^{-M}, & (x\geqslant 1),\\ x^{-1/2}, & (0<x\leqslant 1),\end{cases} \tag{18}$$

$$K'(x)\ll\begin{cases}x^{-M}, & (x\geqslant 1),\\ x^{-3/2}, & (0<x\leqslant 1).\end{cases} \tag{19}$$

We now sum (12) by parts, writing

$$\sum_{n\leqslant x} d(n) = D(x) = x(\log x + 2\gamma - 1) + \Delta(x)$$

whether $x \geqslant 1$ or not. This yields

$$T(k) = 4 \operatorname{Re}(S(k)),$$

with

$$\begin{aligned} S(k) &= -\frac{2\pi}{k}\int_0^\infty D(x)K'\left(\frac{2\pi x}{k}\right)dx \\ &= -\frac{2\pi}{k}\int_0^\infty x(\log x + 2\gamma - 1)K'\left(\frac{2\pi x}{k}\right)dx - \frac{2\pi}{k}\int_0^\infty \Delta(x)K'\left(\frac{2\pi x}{k}\right)dx, \end{aligned} \tag{20}$$

the integrals converging absolutely, by (19).

4. The leading terms

An application of (19), together with Voronoï's bound $\Delta(x) \ll x^{\frac{1}{3}+\varepsilon}$, shows that the second term in (20) is $O(k^{1/2})$. On the other hand, the first term in (20) becomes, on substituting $x = ky/(2\pi)$,

$$-\frac{k}{2\pi}\int_0^\infty y\left(\log\frac{k}{2\pi} + \log y + 2\gamma - 1\right)K'(y)\,dy = A\frac{k}{2\pi}\log\frac{k}{2\pi} + B\frac{k}{2\pi},$$

where the constants A and B are given by

$$A = -\int_0^\infty yK'(y)\,dy, \qquad B = -\int_0^\infty y(\log y + 2\gamma - 1)K'(y)\,dy.$$

Integration by parts yields

$$A = \int_0^\infty K(y)\,dy, \qquad B = 2\gamma A + \int_0^\infty K(y)\log y\,dy.$$

Put $L(y) = 1$ or $\log y$. Then by (13) and (14), with the line of integration moved to

$(\frac{1}{4})$, we have

$$\int_0^1 K(y)L(y)\,dy = \frac{1}{2\pi i}\int_0^1 \int_{(1/4)} L(y)\Gamma(s+\tfrac{1}{2})y^{-s-\frac{1}{2}}e^{iy-i\pi/4-i\pi s/2}F(s)\,ds\,dy$$

$$= \frac{1}{2\pi i}\int_{(1/4)} \Gamma(s+\tfrac{1}{2})e^{-i\pi/4-i\pi s/2}F(s)\int_0^1 L(y)y^{-s-\frac{1}{2}}e^{iy}\,dy\,ds, \qquad (21)$$

the interchange of integrations being justified by absolute convergence (it is for precisely this reason that we move the line of integration to $(\frac{1}{4})$). Similarly we find

$$\int_1^\infty K(y)L(y)\,dy = \frac{1}{2\pi i}\int_{(1)} \Gamma(s+\tfrac{1}{2})e^{-i\pi/4-i\pi s/2}F(s)\alpha(s)\,ds, \qquad (22)$$

where

$$\alpha(s) = \int_1^\infty L(y)y^{-s-\frac{1}{2}}e^{iy}\,dy.$$

We next modify the path of integration in the last integral so as to run from 1 to $i\infty$. We then see that $\alpha(s)$ is regular for all s and satisfies

$$\alpha(s) \ll \exp\left(\frac{\pi}{2}|t|\right), \qquad (\tfrac{1}{4} \leqslant \sigma \leqslant 1),$$

for $s = \sigma + it$. We may therefore move the line of integration in (22) back to $(\frac{1}{4})$ and add the result to (21) to obtain

$$\int_0^\infty K(y)L(y)\,dy = \frac{1}{2\pi i}\int_{(1/4)} \Gamma(s+\tfrac{1}{2})e^{-i\pi/4-i\pi s/2}F(s)\beta(s)\,ds,$$

where

$$\beta(s) = \int_0^1 L(y)y^{-s-\frac{1}{2}}e^{iy}\,dy + \int_1^{i\infty} L(y)y^{-s-\frac{1}{2}}e^{iy}\,dy$$

$$= \int_0^{i\infty} L(y)y^{-s-\frac{1}{2}}e^{iy}\,dy.$$

We therefore see that $\beta(s) = i^{\frac{1}{2}-s}\Gamma(\frac{1}{2}-s)$ for $L(y) = 1$ and

$$\beta(s) = i^{\frac{1}{2}-s}\left(\frac{i\pi}{2}\Gamma(\tfrac{1}{2}-s) + \Gamma'(\tfrac{1}{2}-s)\right)$$

for $L(y) = \log y$. Now

$$\Gamma(\tfrac{1}{2}+s)\Gamma(\tfrac{1}{2}-s) = \pi \sec(\pi s) \tag{23}$$

so that

$$A = \frac{1}{2\pi i}\int_{(1/4)} e^{-i\pi s}\pi \sec(\pi s)F(s)\,ds$$

and

$$\int_0^\infty K(y)\log y\,dy = \frac{i\pi}{2}A + \frac{1}{2\pi i}\int_{(1/4)} e^{-i\pi s}\pi \sec(\pi s)\frac{\Gamma'(\frac{1}{2}-s)}{\Gamma(\frac{1}{2}-s)}F(s)\,ds.$$

Thus

$$\operatorname{Re}(A) = \frac{1}{2\pi i}\int_{(1/4)} \pi F(s)\,ds$$

and

$$\operatorname{Re}\left(\int_0^\infty K(y)\log y\,dy\right) = \frac{1}{2\pi i}\int_{(1/4)} \frac{\pi^2}{2}\tan(\pi s)F(s)\,ds + \frac{1}{2\pi i}\int_{(1/4)} \pi\frac{\Gamma'(\frac{1}{2}-s)}{\Gamma(\frac{1}{2}-s)}F(s)\,ds.$$

By logarithmic differentiation of the relation (23) we find

$$\frac{\Gamma'(\frac{1}{2}+s)}{\Gamma(\frac{1}{2}+s)} - \frac{\Gamma'(\frac{1}{2}-s)}{\Gamma(\frac{1}{2}-s)} = \pi\tan(\pi s),$$

so that it follows that

$$G(s) = \frac{\pi^2}{2}\tan(\pi s) + \pi\frac{\Gamma(\frac{1}{2}-s)}{\Gamma(\frac{1}{2}-s)}$$

is an even function of s. Then

$$I = \frac{1}{2\pi i}\int_{(1/4)} G(s)F(s)\,ds = G(0) + \frac{1}{2\pi i}\int_{(-1/4)} G(s)F(s)\,ds$$

$$= G(0) + \frac{1}{2\pi i}\int_{(1/4)} G(-w)F(-w)\,dw = G(0) - \frac{1}{2\pi i}\int_{(1/4)} G(w)F(w)\,dw,$$

whence

$$I = \tfrac{1}{2}G(0) = -\frac{\pi}{2}(\gamma + \log 4).$$

Similarly

$$\frac{1}{2\pi i}\int_{(1/4)} \pi F(s)\, ds = \frac{\pi}{2}.$$

Thus, finally, we have

$$4\,\mathrm{Re}\left(A\frac{k}{2\pi}\log\frac{k}{2\pi} + B\frac{k}{2\pi}\right) = k\left(\log\frac{k}{8\pi} + \gamma\right)$$

as required.

5. Lower order terms

On writing

$$H(x) = e^{-i\pi/4}e^{ix}x^{-1/2}$$

and

$$\delta(x) = \begin{cases} 1, & (0 < x \leqslant 1), \\ 0, & (x \geqslant 1), \end{cases}$$

the estimates (15), (16) and (17) yield

$$K^{(1)}\left(\frac{2\pi x}{k}\right) = H^{(1)}\left(\frac{2\pi x}{k}\right)\Gamma(\tfrac{1}{2})\delta\left(\frac{2\pi x}{k}\right) + M_1\left(\frac{2\pi x}{k}\right),$$

where for any $M > 0$ we have

$$M_l(x) \ll \begin{cases} x^{-M}, & (x \geqslant 1), \\ x^{M}, & (0 < x \leqslant 1). \end{cases} \tag{24}$$

In particular it follows that the second term of (20) is

$$-\frac{2\pi}{k}\int_0^\infty \Delta(x)K'\left(\frac{2\pi x}{k}\right)dx = -\left(\frac{k}{2\pi}\right)^{1/2}\Gamma(\tfrac{1}{2})e^{-i\pi/4}\int_0^{k/(2\pi)}\Delta(x)\frac{d}{dx}(e^{2\pi ix/k}x^{-1/2})\,dx$$

$$+O\left(\int_0^{k/2\pi}k^{-1}|\Delta(x)|\,dx\right)+O\left(\int_{k/2\pi}^\infty k\,|\Delta(x)|\,x^{-2}\,dx\right)$$

$$=\left(\frac{k}{8}\right)^{1/2}e^{-i\pi/4}\int_0^{k/2\pi}\Delta(x)(x^{-3/2}+O(x^{-1/2}k^{-1}))\,dx$$

$$+O(k^{\frac{1}{3}+\varepsilon})$$

$$=\frac{k^{1/2}(1-i)}{4}\int_0^\infty \Delta(x)x^{-3/2}\,dx+O(k^{\frac{1}{3}+\varepsilon}).$$

On noting that (see Titchmarsh [5; (12.5.5)])

$$\zeta^2(s)=s\int_0^\infty \Delta(x)x^{-s-1}\,dx, \qquad (\tfrac{1}{3}<\mathrm{Re}\,(s)<1),$$

we obtain $2\zeta^2(\frac{1}{2})$ for the coefficient of $k^{1/2}$ in (4), as required.

We now define, inductively, $\Delta_1(x)=\Delta(x)$ and

$$\Delta_{l+1}(x)=\int_0^x \Delta_l(t)\,dt.$$

Then, for $l\geqslant 3$, we have

$$\Delta_l(x)=\frac{1}{2\pi i}\int_{(3/4)}\zeta^2(s)\frac{x^{s+l-1}}{s(s+1)\cdots(s+l-1)}\,ds. \tag{25}$$

Here we have

$$\zeta^2(\sigma+it)\ll(1+|t|)^{l-1-\varepsilon}, \qquad (1-\tfrac{1}{2}l+\tfrac{1}{2}\varepsilon\leqslant\sigma\leqslant\tfrac{3}{4}),$$

so that, on moving the line of integration to $(1-\frac{1}{2}l+\frac{1}{2}\varepsilon)$, we obtain

$$\Delta_l(x)=\sum_{1\leqslant n<l/2}K_{l,n}x^{l-n}+O(x^{\frac{1}{2}(l+\varepsilon)}) \tag{26}$$

for certain constants $K_{l,n}$. Next, on integrating by parts $l-1$ times, we find

$$-\frac{2\pi}{k}\int_0^\infty \Delta(x)K'\left(\frac{2\pi x}{k}\right)dx = (-1)^l\left(\frac{2\pi}{k}\right)^l\int_0^\infty \Delta_l(x)K^{(l)}\left(\frac{2\pi x}{k}\right)dx$$

$$= (-1)^l\left(\frac{2\pi}{k}\right)^l\Gamma(\tfrac{1}{2})\int_0^{k/(2\pi)} \Delta_l(x)H^{(l)}\left(\frac{2\pi x}{k}\right)dx$$

$$+(-1)^l\left(\frac{2\pi}{k}\right)^l\int_0^\infty \Delta_l(x)M_l\left(\frac{2\pi x}{k}\right)dx.$$

We now write

$$H^{(l)}(y) = y^{-\frac{1}{2}-l}e^{iy}p_1(y),$$

where $p_1(y)$ is a polynomial of degree l. Moreover

$$e^{iy} = p_2(y) + y^{l+1}p_3(y),$$

where $p_2(y)$ is also a polynomial of degree l, and where $p_3(y) \ll 1$ for $0 \leqslant y \leqslant 1$. Combining these, we may write

$$H^{(l)}(y) = y^{-\frac{1}{2}-l}p(y) + q(y),$$

where $p(y)$ is a polynomial of degree at most l, and $q(y) \ll 1$ for $0 \leqslant y \leqslant 1$. Finally let

$$N_l(y) = M_l(y) + \Gamma(\tfrac{1}{2})q(y)\delta(y),$$

so that, by (24)

$$N_l(y) \ll \begin{cases} y^{-M}, & (y \geqslant 1), \\ 1, & (0 < y \leqslant 1). \end{cases} \tag{27}$$

We have thus to consider

$$I_{l,m} = k^{-l}\int_0^{k/(2\pi)} \Delta_l(x)\left(\frac{x}{k}\right)^{-\frac{1}{2}-l+m}dx$$

for $0 \leqslant m \leqslant l$, and

$$I_l = k^{-l} \int_0^\infty \Delta_l(x) N_l\left(\frac{2\pi x}{k}\right) dx.$$

We aim to show that each of these expressions can be written in the form

$$\sum_{-1 \leqslant n \leqslant l-3} C_{l,n} k^{-n/2} + O(k^{1-(l-\varepsilon)/2}),$$

with constants $C_{l,n}$ depending on l and n alone. On choosing $l = 2N+3$, this will show that

$$T(k) = k\left(\log \frac{k}{8\pi} + \gamma\right) + 2\zeta(\tfrac{1}{2})^2 k^{1/2} + \sum_{n=0}^{2N-1} c_{n,N} k^{-n/2} + O(k^{-N}).$$

However it is clear that the constants $c_{n,N}$ must be independent of N, so that the theorem will follow.

In the case of $I_{l,m}$ we first examine

$$k^{-m+\frac{1}{2}} \int_0^1 \Delta_l(x) x^{-\frac{1}{2}-l+m}\, dx.$$

Since $\Delta_l(x) \ll x^{l-\frac{1}{4}}$, by (25), the integral converges. Thus this contribution to $I_{l,m}$ is of the required form. For the remaining part, namely

$$k^{-m+\frac{1}{2}} \int_1^{k/(2\pi)} \Delta_l(x) x^{-\frac{1}{2}-l+m}\, dx, \tag{28}$$

the main terms of (26) yield

$$\sum_n k^{-m+\frac{1}{2}} \int_1^{k/(2\pi)} K_{l,n} x^{m-n-\frac{1}{2}}\, dx = \sum_{1 \leqslant n < l/2} K_{l,n} (m-n+\tfrac{1}{2})^{-1} \left(\left(\frac{k}{2\pi}\right)^{m-n+\frac{1}{2}} - 1\right) k^{-m+\frac{1}{2}},$$

which is of the required form. As to the error term in (26), $E_l(x)$ say, this contributes to (28) a quantity $\ll k^{1+(\varepsilon-l)/2}$ if $m \geqslant (l-1)/2$, and otherwise

$$k^{-m+\frac{1}{2}} \int_1^\infty E_l(x) x^{-\frac{1}{2}-l+m}\, dx + O\left(k^{-m+\frac{1}{2}} \int_{k/(2\pi)}^\infty x^{(\varepsilon-1-l)/2+m}\, dx\right)$$
$$= C'_{l,m} k^{-m+\frac{1}{2}} + O(k^{1+(\varepsilon-l)/2}),$$

which again is of the required form.

For I_l the leading terms of (26) yield

$$\sum_n K_{l,n} k^{-l} \int_0^\infty x^{l-n} N_l\left(\frac{2\pi x}{k}\right) dx = \sum_{1 \leqslant n < l/2} K_{l,n} (2\pi)^{n-l-1} k^{1-n} \int_0^\infty y^{l-n} N_l(y)\, dy.$$

The bound (27) shows that the integrals here converge, so that the above expression has the form we seek. Finally the error term of (26) contributes

$$\ll k^{-l} \int_0^\infty x^{(l+\varepsilon)/2} \left| N_l\left(\frac{2\pi x}{k}\right)\right| dx$$

$$\ll k^{-l} \cdot k^{1+(l+\varepsilon)/2} \int_0^\infty y^{(l+\varepsilon)/2} |N_l(y)|\, dy \ll k^{1+(\varepsilon-l)/2},$$

and this too has the required shape. The proof of the theorem is now complete.

6. The corollary

The theorem shows that the sum on the left of (5) is

$$\frac{p-1}{p}(T(p)-T(1)).$$

Here, since $T(1)$ is a numerical constant, given by (12), the term $-(p-1)p^{-1}T(1)$ may be incorporated in the asymptotic expansion. Moreover the factor $1-p^{-1}$ in front of $T(p)$ merely changes the constants in the asymptotic expansion.

REFERENCES

[1] Balasubramanian, R., *An improvement on a theorem of Titchmarsh on the mean square of* $|\zeta(\frac{1}{2}+it)|$, Proc. London Math. Soc. (3), *36* (1978), 540–576.
[2] Good, A., *Ein Ω-Resultat für das quadratische Mittel der Riemannschen Zetafunktion auf der kritische Linie*, Invent. Math., *41* (1977), 233–251.
[3] Kober, H., *Eine Mittelwertformel der Riemannschen Zetafunktion*, Compositio Math., *3* (1936), 174–189.
[4] Montgomery, H. L., Topics in multiplicative number theory (Springer, Berlin, 1971).
[5] Titchmarsh, E. C., The theory of the Riemann Zeta-function (Clarendon Press, Oxford, 1951).

Magdalen College
Oxford OX1 4AU

Received August 8, 1980

Comment. Math. Helvetici **56** (1981) 163–178 0010–2571/81/002163–16$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Characterization of unbounded spectral operators with spectrum in a half-line

SHMUEL KANTOROVITZ[1]

Abstract. Let T be a possibly unbounded linear operator in the Banach space X such that $R(t)=(t+T)^{-1}$ is defined on $\mathbf{R}^+$. Let $S=TR(I-TR)$ and let $B(.,.)$ denote the Beta function.

THEOREM 1.1. *T is a scalar-type spectral operator with spectrum in $[0,\infty)$ if and only if*

$$\sup\left\{B(k,k)^{-1}\int_0^\infty |x^*S^k(t)x|\,dt/t;\ \|x\|\leq 1,\ \|x^*\|\leq 1,\ k\geq 1\right\}<\infty.$$

A "local" version of this result is formulated in Theorem 2.2.

Introduction

A problem of fundamental importance in Mathematical Physics consists of the discovery of "practical" criteria for the self-adjointness of symmetric operators. Extensive work done in this general direction is efficiently summarized in Reed and Simon [6]. Selfadjointness has found a proper generalization to Banach space in the concept of scalar-type spectrality invented by N. Dunford in the early 1950's and developed since then by his students and followers. The monograph of N. Dunford and J. T. Schwartz [1; Part III] contains a detailed exposition of the subject and an updated bibliography up to 1968. We shall freely use the terminology of [1].

Let T be a possibly unbounded linear operator with domain $D(T)$ in a Banach space X. Suppose there exists a (strongly) countably additive spectral measure E on the Borel σ-field $\mathfrak{B}(\mathbf{R})$ such that

$$D(T)=\left\{x\in X|\ \text{strong}\lim_n\int_{-n}^{n} sE(ds)x\ \text{exists}\right\}$$

---

[1] Research partially done while the author was a visiting member at the Forschungsinstitut für Mathematik, ETH, Zürich.

Research supported by the Israel Academy of Science and Bar-Ilan University Research Authority.

and

$$Tx = \lim_n \int_{-n}^{n} sE(ds)x, \qquad x \in D(T).$$

One then says that T is a (real) scalar-type spectral operator, with resolution of the identity E.

In [1, Chapters XIX and XX], the (scalar-type) spectrality of various perturbations $T = S + P$ is studied, and many deep results providing sufficient conditions are obtained, when S is either a relatively simple differential operator, or a selfadjoint operator in Hilbert space, or a multiplication operator. Accordingly, the methods are very specialized and have to overcome many technical difficulties. Even the statements of the theorems involve heavily technical assumptions (see for example Theorems XIX.4.16; XX.1.12, 2.21, 2.22, 4.9, etc. . . .). Methods run from direct efforts toward the construction of the resolution of the identity, to the Friedrichs method of similar operators, to the Kato-Kuroda method of wave operators and similarity. The authors remark that the material in their exposition is still somewhat "fragmentary", "and is presented in the hope that it may stimulate research in the topics treated".

Motivated by the fact that the selfadjointness analysis for *semibounded* symmetric operators in Hilbert space is enormously simpler than the general case, we consider in this paper the construction of a (scalar-type) spectrality criterion for operators whose resolvent set contains a half-line. Our main result (Theorem 1.1) provides a criterion in terms of the asymptotic behavior of adequate functions of the resolvent operator. We also obtain a "localized version" of this result (Theorem 2.2), in the spirit of [4]. Also in the spirit of previous work of ours [2, 3, 4, 5], we shall use transform methods – in this case, naturally, the Stieltjes transform. The following classical result of Widder's is the key to our analysis. Let M and D denote respectively the formal operators of multiplication

$$M: f(t) \to tf(t)$$

and differentiation $D: f \to f'$.

The Widder (formal) differential operators L_k are given by

$$L_k = c_k M^{k-1} D^{2k-1} M^k \qquad k = 1, 2, \ldots$$

where $c_1 = 1$ and $c_k = (-1)^{k-1}[k!(k-2)!]^{-1}$ for $k \geq 2$.

The STR (Stieltjes Transform Representation) *Theorem.*
Let f be a C^∞ complex function on $\mathbf{R}^+ = (0, \infty)$ such that

$$K = \sup_{k \geq 1} \int_0^\infty |L_k f(t)|\, dt < \infty. \tag{*}$$

Then there exists a unique complex regular Borel measure μ on $\mathbf{R}^+$ such that

$$f(t) = \int_0^\infty (t+s)^{-1} \mu(ds)$$

$$\text{var } \mu \leq K'$$

where $K' = 2K + |A|$ and $A = \lim_{t \to 0+} tf(t)$. A (*).

In [7; Theorem 16, p. 361], the STR theorem is stated (in more classical terminology) for *real* f. The complex ease is a trivial consequence, and the estimate for var μ follows easily from the proof in the cited reference. Note that the *existence* of the limit A follows from Condition (*).

We conclude this introduction with a list of notation. The (normed) dual space of the Banach space X is denoted by X^*. The norm-closed unit balls of X and X^* are X_1 and X_1^* respectively. $B(X)$ denotes the Banach algebra of all bounded linear operators in X. For $s, t > 0$, $B(s, t)$ stands for the Beta function

$$B(s, t) = \Gamma(s)\Gamma(t)\Gamma(s+t)^{-1}$$

$\mathfrak{B}(\mathbf{R}^+)$ is the Borel σ-field of $\mathbf{R}^+$, and $\| \cdot \|_1$ denotes the $L^1(\mathbf{R}^+, dt/t)$-norm (*attention*! the measure is dt/t!). Strong limits mean limits in the given norm of X.

1. Operators with spectrum in a half-line

Since spectrality is preserved by the map $T \to \alpha I + \beta T$ ($\alpha, \beta \in \mathbf{C}$), we may restrict our discussion, without loss of generality, to the case where the half-line of the title is $[0, \infty)$.

Let R denote the resolvent of $-T$, that is $R(t) = (t+T)^{-1}$. Thus, R is defined on $\mathbf{R}^+$, and so is the $B(X)$-valued function

$$S = TR(I - TR). \tag{1.1}$$

For any weakly measurable function $F: \mathbf{R}^+ \to B(X)$, we let

$$|||F||| = \sup \{ \| x^* F(\cdot) x \|_1 ;\ x \in X_1,\ x^* \in X_1^* \}. \tag{1.2}$$

Clearly, $|||\cdot|||$ is a semi-norm on the vector space $\mathfrak{F}$ (under pointwise operations) of all such functions F with $|||F|||<\infty$, and

$$\|x^*F(\cdot)x\|_1 \leq |||F||| \, \|x\| \, \|x^*\| \tag{1.3}$$

for all $F\in\mathfrak{F}$, $x\in X$ and $x^*\in X^*$.

Since $S(t)=tR(t)[I-tR(t)]$, the functions S^k $(k=1,2,\ldots)$ are of class C^∞ on $\mathbf{R}^+$; in particular, $|||S^k|||$ is well-defined (finite or infinite). We are now prepared to state our spectrality criterion for T, which bears some analogy with the Hille-Yosida criterion for semi-group generation.

THEOREM 1.1 *Let T be a possibly unbounded operator in the reflexive Banach space X, whose resolvent set contains the half-line $(-\infty, 0)$. Then T is spectral of scalar type with spectrum in $[0,\infty)$ if and only if*

$$|||S^k|||\leq MB(k,k) \qquad k=1,2,\ldots \tag{1.4}$$

for some constant M.

Proof. Necessity. Since $TR(t)=I-tR(t)$ for $t\notin\sigma(T)$, therefore

$$S(t)=t[I-tR(t)]R(t)=tTR(t)^2. \tag{1.5}$$

Let T be spectral of scalar type with spectrum in $[0,\infty)$, and let E denote its resolution of the identity. By (1.5), we have for $t>0$ and $k=1,2,\ldots$

$$S^k(t)=t^kT^k(t+T)^{-2k}=t^k\int_0^\infty s^k(t+s)^{-2k}E(ds)=\int_0^\infty (s/t)^k(1+s/t)^{-2k}E(ds).$$

Therefore, for all $x\in X_1$ and $x^*\in X_1^*$,

$$\|x^*S^kx\|_1\leq\int_0^\infty\int_0^\infty (s/t)^k(1+s/t)^{-2k}\,|x^*E(ds)x|\,dt/t.$$

If we interchange the order of integration and substitute $u=s/t$ in the inner integral, the right hand side becomes

$$\int_0^\infty\int_0^\infty u^k(1+u)^{-2k}\,du/u\,|x^*E(ds)x|=B(k,k)\ \text{var}\,(x^*Ex)\leq B(k,k)\ \text{var}\,E<\infty.$$

Applying Tonelli's theorem, we conclude that

$$|||S^k|||\leq \operatorname{var} E \cdot B(k,k) \qquad k=1,2,\ldots. \tag{1.6}$$

Observe that the reflexivity assumption on X is not needed for the necessity part of the theorem.

Sufficiency. Let L_k $(k=1,2,\ldots)$ be the Widder formal differential operators mentioned in the introduction. A straightforward calculation using Leibnitz' rule shows that

$$L_k = c'_k \sum_{j=0}^{k} \Gamma(k+j)^{-1}\binom{k}{j} M^{k+j-1}D^{k+j-1} \tag{1.7}$$

where $c'_1=1$ and $c'_k=(-1)^{k-1}B(k-1,k+1)$ for $k\geq 2$. Fix $x\in X$ and $x^*\in X^*$, and consider the C^∞ function on $\mathbf{R}^+$

$$f(t)=x^*R(t)x \qquad t\in\mathbf{R}^+.$$

Since

$$D^{k+j-1}f=(-1)^{k+j-1}\Gamma(k+j)x^*R^{k+j}x,$$

we have

$$L_kf(t)=c''_kt^{-1}x^*(tR)^k\sum_{j=0}^{k}\binom{k}{j}(-tR)^jx=c''_kt^{-1}x^*S^k(t)x \tag{1.8}$$

where $c''_1=1$ and $c''_k=B(k-1,k+1)^{-1}$ for $k\geq 2$.

Therefore, by (1.3) and (1.4),

$$\begin{aligned}\int_0^\infty |L_kf(t)|\,dt &= c''_k\,\|x^*S^kx\|_1 \leq c''_k\,|||S^k|||\,\|x\|\,\|x^*\| \\ &\leq M\,\|x\|\,\|x^*\| \qquad k=1,2,\ldots,\end{aligned} \tag{1.9}$$

since $c''_1\,|||S|||\leq M$ and for $k\geq 2$

$$c''_k\,|||S^k|||\leq MB(k,k)/B(k-1,k+1)=M(k-1)/k\leq M.$$

By the STR theorem, it follows that there exists a unique complex regular Borel

measure $\mu = \mu(\cdot \mid x, x^*)$ on $\mathbf{R}^+$ such that

$$x^*R(t)x = \int_0^\infty (t+s)^{-1}\mu(ds \mid x, x^*), \ (t>0, x \in X, x^* \in X^*). \tag{1.10}$$

In particular,

$$\sup_{t>0} |x^*tR(t)x| = \sup_{t>0} \left| \int_0^\infty t(t+s)^{-1}\mu(ds \mid x, x^*)\right| \leq \operatorname{var} \mu < \infty.$$

By the Uniform Boundedness Theorem, it follows that

$$H = \sup_{t>0} \|tR(t)\| < \infty. \tag{1.11}$$

Hence, for our function f, $|tf(t)| \leq H\|x\|\,\|x^*\|$ for all $t>0$, and therefore $|A| \leq H\|x\|\,\|x^*\|$ (with notation as in the STR theorem). We conclude that

$$\operatorname{var} \mu(\cdot \mid x, x^*) \leq M'\|x\|\,\|x^*\| \tag{1.12}$$

where $M' = 2M + H$.

The uniqueness of the representation (1.10) implies that for each fixed $\delta \in \mathfrak{B}(\mathbf{R}^+)$, $\mu(\delta \mid ., .)$ is a bilinear form on $X \times X^*$. By (1.12), the form is bounded (with bound M'), and since X is reflexive, there exists a unique function $E: \mathfrak{B}(\mathbf{R}^+) \to B(X)$ such that

$$\mu(\cdot \mid x, x^*) = x^*E(\cdot)x \tag{1.13}$$

for all $x \in X$ and $x^* \in X^*$.

We shall now verify that E is a spectral measure on $\mathfrak{B}(\mathbf{R}^+)$. By (1.13) and (1.12), E is weakly, hence strongly countably additive, and $\|E(\delta)\| \leq M'$ for all $\delta \in \mathfrak{B}(\mathbf{R}^+)$.

Let $U \in B(X)$ be such that $UD(T) \subset D(T)$ and $TUx = UTx$ for all $x \in D(T)$ (i.e., $TU \supset UT$, which is the usual definition of commutativity with an unbounded operator T). Equivalently, $UR(t) = R(t)U$ for all $t>0$. By (1.10), for all $t>0$, $x \in X$ and $x^* \in X^*$,

$$\int_0^\infty (t+s)^{-1}\mu(ds \mid x, U^*x^*) = [U^*x^*]R(t)x = x^*R(t)Ux = \int_0^\infty (t+s)^{-1}\mu(ds \mid Ux, x^*).$$

By the uniqueness statement of the STR theorem, it follows that

$$\mu(\delta \mid x, U^*x^*) = \mu(\delta \mid Ux, x^*)$$

for all $\delta \in \mathfrak{B}(\mathbf{R}^+)$, $x \in X$ and $x^* \in X^*$.

Hence, by (1.13),

$$UE(\delta) = E(\delta)U \qquad \delta \in \mathfrak{B}(\mathbf{R}^+). \tag{1.14}$$

Taking in particular $U = R(u)$ $(u > 0)$, we obtain that $E(\delta)D(T) \subset D(T)$ for all $\delta \in \mathfrak{B}(\mathbf{R}^+)$. For $t > u > 0$, we have by the First Resolvent Equation

$$tR(t)R(u) = t(t-u)^{-1}R(u) - [tR(t)](t-u)^{-1}.$$

Fixing u and letting $t \to \infty$, it follows from (1.11) that $tR(t)R(u) \to R(u)$ in the uniform operator topology. By (1.10) and the Lebesgue dominated convergence theorem, we have for all $u > 0$, $x \in X$, and $x^* \in X^*$

$$\begin{aligned} x^*R(u)x &= \lim_{t\to\infty} x^*tR(t)R(u)x \\ &= \lim_{t\to\infty} \int_0^\infty t(t+s)^{-1}\mu(ds \mid R(u)x, x^*) \\ &= \mu(\mathbf{R}^+ \mid R(u)x, x^*) = x^*E(\mathbf{R}^+)R(u)x. \end{aligned}$$

Hence, by (1.14) with $U = R(u)$, $R(u)E(\mathbf{R}^+) = R(u)$. Applying $u + T$ to both sides (on the left), we conclude that $E(\mathbf{R}^+) = I$. For $t, u > 0$, $t \neq u$, the First Resolvent Equation implies that

$$\begin{aligned} \int_0^\infty (t+s)^{-1}\mu(ds \mid R(u)x, x^*) &= x^*R(t)R(u)x \\ &= (t-u)^{-1}\int_0^\infty [(u+s)^{-1} - (t+s)^{-1}]\mu(ds \mid x, x^*) \\ &= \int_0^\infty (t+s)^{-1}(u+s)^{-1}\mu(ds \mid x, x^*). \end{aligned}$$

The uniqueness claim in the STR theorem gives

$$\mu(\delta \mid R(u)x, x^*) = \int_0^\infty (u+s)^{-1}\chi_\delta(s)\mu(ds \mid x, x^*) \tag{1.15}$$

for all $u > 0$, $\delta \in \mathfrak{B}(\mathbf{R}^+)$, etc. ...

On the other hand, by (1.13), (1.14) with $S = R(u)$, and (1.10),

$$\mu(\delta \mid R(u)x, x^*) = x^*R(u)E(\delta)x$$

$$= \int_0^\infty (u+s)^{-1}\mu(ds \mid E(\delta)x, x^*). \tag{1.16}$$

Comparing the two relations, we obtain (again by uniqueness!)

$$\mu(\sigma \mid E(\delta)x, x^*) = \int_0^\infty \chi_\sigma(s)\chi_\delta(s)\mu(ds \mid x, x^*) = \mu(\sigma \cap \delta \mid x, x^*)$$

for all $\sigma, \delta \in \mathfrak{B}(\mathbf{R}^+)$, etc. . . . , that is, by (1.13),

$$E(\sigma)E(\delta) = E(\sigma \cap \delta)$$

for all $\sigma, \delta \in \mathfrak{B}(\mathbf{R}^+)$.

In conclusion, E is a (strongly) countably additive spectral measure on $\mathfrak{B}(\mathbf{R}^+)$, which sends $D(T)$ into itself and commutes with every $U \in B(X)$ which commutes with t. By (1.13), we may rewrite (1.10) in the form

$$R(t) = \int_0^\infty (t+s)^{-1}E(ds) \qquad (t>0) \tag{1.17}$$

where the integral makes sense in $B(X)$ as described for example in [1, Part II; pp. 891–892].

From (1.17) and [1, Part III; Theorems XVIII.2.17 and XVIII.2.11(h)] we may easily conclude that T is spectral of scalar type with resolution of the identity E and spectrum in $[0, \infty)$.

However the presentation can be made direct and selfcontained with only a small additional effort.

By (1.17), for all $t>0$,

$$\int_0^\infty s(t+s)^{-1}E(ds) = \int_0^\infty [1 - t(t+s)^{-1}]E(ds)$$

$$= I - tR(t) = TR(t). \tag{1.18}$$

For all $n>0$, $y \in X$, and $x^* \in X^*$, we have by (1.13) and (1.16)

$$x^* \int_0^n sE(ds)R(t)y = \int_0^n s\mu(ds \mid R(t)y, x^*)$$

$$= \int_0^n s(t+s)^{-1}\mu(ds \mid y, x^*) = x^* \int_0^n s(t+s)^{-1}E(ds)y$$

Therefore, by (1.15),

$$\int_0^n sE(ds)R(t)y = \int_0^n sR(t)E(ds)y = \int_0^n s(t+s)^{-1}E(ds)y \tag{1.19}$$

for all $n, t>0$ and $y \in X$.

Suppose now that $x \in D(T)$. Then $x = R(t)y$ for $t>0$ fixed and suitable $y \in X$. By (1.19),

$$\int_0^n sE(ds)x = \int_0^n s(t+s)^{-1}E(ds)y \xrightarrow[n\to\infty]{} \int_0^\infty s(t+s)^{-1}E(ds)y$$

(strong convergence).

By (1.18), the above limit equals $TR(t)y = Tx$, that is

$$Tx = \text{strong} \lim_{n\to\infty} \int_0^n sE(ds)x, \qquad x \in D(T). \tag{1.20}$$

On the other hand, if $x \in X$ is such that the limit on the right of (1.20) exists, then denoting this limit by z, we have for any $t>0$

$$\begin{aligned} R(t)z &= \text{strong} \lim_{n\to\infty} \int_0^n sR(t)E(ds)x \\ &= \text{strong} \lim_{n\to\infty} \int_0^n s(t+s)^{-1}E(ds)x = \int_0^\infty s(t+s)^{-1}E(ds)x \\ &= x - tR(t)x, \end{aligned}$$

where we used (1.19) and (1.18). Hence $x = R(t)[z+tx] \in D(T)$. This shows that

$$D(T) = \left\{ x \in X \mid \text{strong} \lim_{n\to\infty} \int_0^n sE(ds)x \text{ exists} \right\},$$

and together with (1.20), this completes the proof of the scalar-type spectrality of T. Since the (uniquely determined) resolution of the identity E is supported by $[0,\infty)$, the spectrum of T is necessarily contained in $[0,\infty)$. Q.E.D.

2. Local spectral analysis

As in the preceding section, T is a possibly unbounded operator in the reflexive Banach space X, whose resolvent set contains a half-line, which we

assume to be $(-\infty, 0)$ without loss of generality. We shall use the function $S:\mathbf{R}^+ \to B(X)$ in order to obtain a so-called "local" spectral analysis for any such operator T (cf. [4] for an analogous treatment of *bounded* operators).

For any $x \in X$, we let

$$|||x|||_0 = \sup\{B(k,k)^{-1}\|x^*S^kx\|_1;\ x^* \in X_1^*,\ k = 1,2,\ldots\},$$
$$|||x||| = \max\{\|x\|, |||x|||_0\}, \tag{2.1}$$

and

$$Z = \{x \in X;\ |||x||| < \infty\}. \tag{2.2}$$

The functional $|||\cdot|||$ is a norm on the linear manifold Z, and in fact $|||x||| \geq \|x\|$ on Z. An easy calculation shows for example that if $x \in D(T)$ is an eigenvector of T corresponding to an eigenvalue $\lambda \in \mathbf{R}$, then $|||x||| = \|x\|$, so that Z contains all such eigenvectors.

If $U \in B(X)$ commutes with T (that is $UT \subset TU$), then Z is U-invariant, and for all $x \in Z$

$$|||Ux||| \leq \|U\|\,|||x|||. \tag{2.3}$$

It is also clear that

$$\|x^*S^kx\|_1 \leq \|x^*\|\,|||x|||\,B(k,k) \tag{2.4}$$

for all $x \in Z$, $x^* \in X^*$, and $k = 1, 2, \ldots$.

PROPOSITION 2.1. *$(Z, |||\cdot|||)$ is a Banach space.*

Proof. Let $\{x_n\} \subset Z$ be $|||\cdot|||$-Cauchy. Since $|||\cdot||| \geq \|\cdot\|$, $\{x_n\}$ converges to some $x \in X$ in the given norm. For each $k \geq 0$ and $x^* \in X_1^*$, $x^*S^k(\cdot)x_n \to x^*S^k(\cdot)x$ pointwise as $n \to \infty$, and by (2.4),

$$B(k,k)^{-1}\|x^*S^kx_n\|_1 \leq |||x_n||| \leq K. \qquad \text{(for all } n\text{)}.$$

By Fatou's lemma, it follows that $|||x||| \leq K$, i.e., $x \in Z$. Now, given $\varepsilon > 0$, let n_0 be such that $|||x_n - x_m||| < \varepsilon$ for $n > m > n_0$. Since $x^*S^k(x - x_m) = \lim_{n\to\infty} x^*S^k(x_n - x_m)$ pointwise, we have again by Fatou's lemma and (2.4) $B(k,k)^{-1}\|x^*S^k(x - x_m)\|_1 \leq \varepsilon$ for all $k \geq 1$, $x^* \in X_1^*$, and $m > n_0$, that is, $|||x - x_m||| \leq \varepsilon$ for $m > n_0$. ∎

For $t \in \mathbf{R}^+$, the linear manifold $R(t)Z = Z_0$ is contained in $Z \cap D(T)$ (since

$R(t)$ commutes with T), and is independent of t. Indeed, for $s, t>0$ and $z \in Z$, we have by the first resolvent equation

$$R(t)z = R(s)[(s-t)R(t)z + z] \in R(s)Z,$$

hence $R(t)Z \subset R(s)Z$, and the equality follows by symmetry.

We denote by $\mathbf{T}(Z)$ the set of all (linear) operators with domain Z and range contained $\mathbf{Z}$. For the usual operations, $\mathbf{T}(Z)$ is an algebra with the identity $I \mid Z$. We are now ready to state our local version of Theorem 1.1.

THEOREM 2.2. *There exists an algebra homomorphism E of $\mathfrak{B}(\mathbf{R}^+)$ (as a Boolean algebra) into $\mathbf{T}(Z)$ with the following properties* (1)–(4):

(1) *$E(\mathbf{R}^+) = I \mid Z$, and for each $x \in Z$, $E(\cdot)x$ is a countably additive measure on $\mathfrak{B}(\mathbf{R}^+)$ (necessarily bounded and strongly countably additive)*:

(2) *$E(\delta)$ commutes with every $U \in B(X)$ which commutes with T (for each $\delta \in \mathfrak{B}(\mathbf{R}^+)$)*;

(3) *$Z_0 = \{x \in Z \mid \text{strong } \lim_{n\to\infty} \int_0^n sE(ds)x$ exists and belongs to $Z\}$*;

(4) *$Tx = \text{strong } \lim_{n\to\infty} \int_0^n sE(ds)x$ for all $x \in Z_0$.*

Proof. Fix $x \in Z$ and $x^* \in X^*$, and let f be as in Section 1. By (1.8) and (2.4),

$$\int_0^\infty |L_k f(t)|\, dt \leq \|x^*\|\, |||x|||\, c_k'' B(k, k) \leq \|x^*\|\, |||x|||$$

for all $k \geq 1$. By the STR theorem, there exists a unique complex regular Borel measure $\mu(\cdot \mid x, x^*)$ such that (1.10) is valid for all $x \in Z$ and $x^* \in X^*$. As before, we obtain that for each $x \in Z$,

$$H_x = \sup_{t>0} \|tR(t)x\| < \infty, \tag{2.5}$$

and

$$\text{var } \mu(\cdot \mid x, x^*) \leq M_x \|x^*\| \qquad (x \in Z, x^* \in X^*) \tag{2.6}$$

where $M_x = 2|||x||| + H_x$.

As before, the uniqueness of the representation (1.10) and the reflexivity of X imply the existence of a unique function $E(\cdot)x : \mathfrak{B}(\mathbf{R}^+) \to X$ (for each fixed $x \in Z$) such that $\mu(\delta \mid x, x^*) = x^*E(\delta)x$ for all $\delta \in \mathfrak{B}(\mathbf{R}^+)$ and $x^* \in X^*$. Clearly, $E(\delta)$ is a linear operator with domain Z, and $\|E(\delta)x\| \leq M_x$ for all $\delta \in \mathfrak{B}(\mathbf{R}^+)$ and $x \in Z$. For fixed $x \in Z$, $E(\cdot)x$ is weakly, hence strongly countably additive.

If $U \in B(X)$ commutes with T, we already observed that $UZ \subset Z$, and the commutativity $UE(\delta) \subset E(\delta)U$ is verified as (1.14) ($\delta \in \mathfrak{B}(\mathbf{R}^+)$). Fix $u > 0$. For each $x \in Z$ and $x^* \in X^*$,

$$x^*R(u)E(\mathbf{R}^+)x = x^*E(\mathbf{R}^+)R(u)x = \mu(\mathbf{R}^+ \mid R(u)x, x^*)$$

$$= \lim_{t\to\infty} \int_0^\infty t(t+s)^{-1}\mu(ds \mid R(u)x, x^*)$$

$$= \lim_{t\to\infty} x^*tR(t)R(u)x = x^*R(u)x,$$

where we used the first resolvent equation and (2.5). Since $R(u)$ is one-to-one, we conclude that $E(\mathbf{R}^+) = I \mid Z$.

We prove now that for all $\delta \in \mathfrak{B}(\mathbf{R}^+)$ and $x \in Z$,

$$|||E(\delta)x||| \leq M_x. \tag{2.7}$$

In particular, $E(\delta) \in \mathbf{T}(Z)$ for each $\delta \in \mathfrak{B}(\mathbf{R}^+)$.

Clearly, (1.15) is valid for all variables as before and x varying in Z. Briefly

$$\mu(ds \mid R(u)x, x^*) = (u+s)^{-1}\mu(ds \mid x, x^*) \tag{2.8}$$

($u > 0$, $x \in Z$, $x^* \in X^*$), and inductively

$$\mu(ds \mid R(u)^k x, x^*) = (u+s)^{-k}\mu(ds \mid x, x^*),$$

for $k = 0, 1, 2, \ldots$. Hence

$$\mu(ds \mid p(R(u))x, x^*) = p((u+s)^{-1})\mu(ds \mid x, x^*)$$

for any polynomial p. In particular

$$\mu(ds \mid S^k(u)x, x^*) = u^k(u+s)^{-k}[1-u(u+s)^{-1}]^k\mu(ds \mid x, x^*)$$

$$= (us)^k(u+s)^{-2k}\mu(ds \mid x, x^*) \tag{2.9}$$

($u > 0$, $x \in Z$, $x^* \in X^*$: $k = 1, 2, \ldots$).

Since $S^k(u)$ commutes with $R(t)$ (that is, with T), it commutes with $E(\delta)$, and therefore, by (2.9)

$$x^*S^k(u)E(\delta)x = x^*E(\delta)S^k(u)x = \mu(\delta \mid S^k(u)x, x^*)$$

$$= \int_\delta (us)^k(u+s)^{-2k}\mu(ds \mid x, x^*).$$

Hence, by Tonelli's theorem and (2.6)

$$\|x^*S^k(u)E(\delta)x\|_1 \leq \int_{\delta}\int_0^{\infty}(us)^k(u+s)^{-2k}du/u\,|\mu|\,(ds\mid x, x^*)$$

$$= \int_{\delta}\int_0^{\infty}t^k(1+t)^{-2k}dt/t\,|\mu|\,(ds\mid x, x^*)$$

$$= B(k,k)\,|\mu|\,(\delta\mid x, x^*) \leq B(k,k)M_x\,\|x^*\|.$$

Therefore $|||E(\delta)x||| \leq M_x$.

Using (1.15) and (1.16) (with $x \in Z$), we obtain that for all $\sigma, \delta \in \mathfrak{B}(\mathbf{R}^+)$, $E(\sigma)E(\delta) = E(\sigma \cap \delta)$ in $\mathbf{T}(Z)$.

We verify finally Properties (3) and (4). Denote the set on the right of (3) by Z_1. Let $x \in Z_0$, that is $x = R(u)y$ for $u > 0$ and suitable $y \in Z$. By (2.8),

$$\int_0^n sE(ds)x = \int_0^n sE(ds)R(u)y = \int_0^n s(u+s)^{-1}E(ds)y$$

$$\xrightarrow[n\to\infty]{} \int_0^{\infty}s(u+s)^{-1}E(ds)y \qquad \text{(strongly)}$$

$$= \int_0^{\infty}[1-u(u+s)^{-1}]E(ds)y = [I-uR(u)]y \quad (\in Z)$$

$$= TR(u)y = Tx.$$

Thus $Z_0 \subset Z_1$ and (4) is valid for $x \in Z_0$. On the other hand, if $x \in Z_1$, denote the limit in (3) by $z \in Z$. As in the proof of Theorem 1.1., we obtain that for any $t > 0$,

$$x = R(t)[z + tx] \in R(t)Z = Z_0.$$

Therefore $Z_0 = Z_1$ and the proof is complete.

Remarks. 1. One has

$$Z_0 = \{x \in D(T) \cap Z \mid Tx \in Z\}. \tag{2.10}$$

Indeed, if $x \in Z_0$, then $x = R(t)z$ for $t > 0$ and suitable $z \in Z$; therefore $x \in D(T) \cap Z$ and $Tx = TR(t)z = [I - tR(t)]z \in Z$. On the other hand, if $x \in D(T) \cap Z$ and $Tx \in Z$, then $x = R(t)y$ with $y = tx + Tx \in Z$, i.e., $x \in Z_0$. By (2.10), Z_0 is the natural domain of the restriction $T \mid Z$.

2. Let $\mathbf{B}(\mathbf{R}^+)$ denote the Banach algebra of all bounded Borel functions on

$\mathbf{R}^+$, with pointwise operations and supremum norm. For $x \in Z$ and $h \in \mathbf{B}(\mathbf{R}^+)$, $h(T)x = \int_0^\infty h(s)E(ds)x$ belongs to Z by (2.9). Hence $h(T) \in \mathbf{T}(Z)$ for each $h \in \mathbf{B}(\mathbf{R}^+)$. Using the density of simple Borel functions in $\mathbf{B}(\mathbf{R}^+)$, (2.7), (2.6), and the relation $E(\sigma)E(\delta) = E(\sigma \cap \delta)$ (in $\mathbf{T}(Z)$), one deduces easily that the map $h \to h(T)$ is an algebra homomorphism of $\mathbf{B}(\mathbf{R}^+)$ into $\mathbf{T}(Z)$, which is continuous in the following sense: if $h_n \to h$ in $\mathbf{B}(\mathbf{R}^+)$, then for each $x \in Z$, $h_n(T)x \to h(T)x$ with respect to the $|||\cdot|||$-norm. This "operational calculus" may then be extended to unbounded Borel functions h in the "usual" way.

Let $h_n = h$ if $|h| \leq n$ and $h_n = 0$ otherwise. Define

$$D(h(T)) = \{x \in Z \mid \text{strong } \lim_{n\to\infty} h_n(T)x \text{ exists and belongs to } Z\} \tag{2.11}$$

and

$$h(T)x = \text{strong } \lim_{n\to\infty} h_n(T)x, \qquad x \in D(h(T)). \tag{2.12}$$

By Theorem 2.2 and Remark 1 above, $h(T) = T \mid Z$ for $h(s) = s$. The details of the analysis of this operational calculus will be omitted. One point however should be stressed. If p is any polynomial of degree k, and $p(T)$ has the usual meaning, the "natural" domain of $p(T) \mid Z$ is the set

$$D(p(T) \mid Z) = \{x \in D(T^k) \mid T^j x \in Z, 0 \leq j \leq k\}.$$

As in Remark 1, one verifies easily that this domain coincides with $R(t)^k Z$ (for any $t > 0$).

Let $x \in D(p(T) \mid Z)$, that is $x = R(t)^k z$ for some $z \in Z$. Then, since $p(s)(t+s)^{-k} \in \mathbf{B}(\mathbf{R}^+)$,

$$p_n(T)x = \int_0^\infty p_n(s)(t+s)^{-k}E(ds)z \xrightarrow[n\to\infty]{} \text{(strongly)}$$

$$\int_0^\infty p(s)(t+s)^{-k}E(ds)z \in Z,$$

i.e. $x \in D(p(T))$ in the sense of (2.11). Also, writing $p(s) = \sum_{j=0}^k \alpha_j s^j$, the last integral is equal to

$$\sum_{j=0}^k \alpha_j \int_0^\infty [s(t+s)^{-1}]^j (t+s)^{-(k-j)}E(ds)z$$

$$= \sum_{j=0}^k \alpha_j \int_0^\infty [1 - t(t+s)^{-1}]^j (t+s)^{-(k-j)}E(ds)z$$

$$= \sum_{j=0}^k \alpha_j [I - tR(t)]^j R(t)^{k-j} z = \sum_{j=0}^k \alpha_j T^j R(t)^k z$$

$$= p(T)x \text{ (the "classical" meaning of } p(T)),$$

where we used the multiplicativity of the operational calculus with *bounded* functions. Thus the definition of $p(T)$ according to (2.11) and (2.12) coincides with the usual definition on $D(p(T)\,|\,Z)$.

3. Some corollaries

If T is an operator with real spectrum, we may apply Theorem 1.1 to T^2, since $\sigma(T^2)\subset[0,\infty)$. Let

$$S(t)=tR(t)[I-tR(t)], \qquad t>0$$

where

$$R(t)=(t+T^2)^{-1}.$$

COROLLARY 3.1. *Let T be a possibly unbounded linear operator with real spectrum. Then T has a scalar-type spectral square if and only if*

$$\sup_{k\geq 1} B(k,k)^{-1}\,|||S^k|||<\infty. \tag{3.1}$$

An equivalent formulation is given below.

COROLLARY 3.2. *Let T be a possibly unbounded linear operator with real spectrum. Then there exists a scalar-type spectral operator $|T|$ with $\sigma(|T|)\subset[0,\infty)$ such that $|T|^2=T^2$ if and only if Condition* (3.1) *is valid.*

Proof. The necessity of Condition (3.1) is a trivial consequence of Corollary 3.1, since the square of a scalar-type spectral (s.t.s) operator is s.t.s. Conversely, Condition (3.1) implies that T^2 is s.t.s. with spectrum in $[0,\infty)$.

If E denotes its resolution of the identity, define $|T|$ as the operator corresponding to the function $f(s)=s^{1/2}$ on $\mathbf{R}^+$ in the operational calculus (o.c.) associated with E (cf. [1; Definition 10, p. 2238]).

By Theorem 17 in [1; p. 2244], $|T|$ is s.t.s. with spectrum in $[0,\infty)$, and $|T|^2\subset T^2$ by [1; Theorem 11(f), p. 2238]. However, if $x\in D(T^2)$, then $x=R(t)y$ for $t>0$ and suitable y. Hence

$$\int_0^{n^2} s^{1/2}E(ds)x=\int_0^{n^2} s^{1/2}(t+s)^{-1}E(ds)y\xrightarrow[n\to\infty]{}\int_0^{\infty} s^{1/2}(t+s)^{-1}E(ds)y,$$

since the integrand is bounded on $\mathbf{R}^+$. Therefore $x\in D(|T|)$ and $|T|x$ is given by

the above limit. Next, by the multiplicativity of the o.c. with bounded functions,

$$\int_0^{n^2} s^{1/2}E(ds)\,|T|x = \int_0^{n^2} s^{1/2}E(ds)\int_0^{\infty} s^{1/2}(t+s)^{-1}E(ds)y$$

$$= \int_0^{n^2} s(t+s)^{-1}E(ds)y \xrightarrow[n\to\infty]{} \int_0^{\infty} s(t+s)^{-1}E(ds)y,$$

that is $|T|x \in D(|T|)$, showing that $x \in D(|T|^2)$. Hence $|T|^2 = T^2$. Q.E.D.

REFERENCES

[1] DUNFORD, N. and SCHWARTZ, J. T., *Linear operators*, Parts II and III, Wiley-Interscience, New York (1963 and 1971).

[2] KANTOROVITZ, S., *On the characterization of spectral operators*, Trans. Amer. Math. Soc. *111* (1964), 152–181.

[3] ——, *Classification of operators by means of their operational calculus*, Trans. Amer. Math. Soc. *115* (1965), 194–224.

[4] ——, *The semi-simplicity manifold of arbitrary operators*, Trans. Amer. Math. Soc. *123* (1966), 241–252.

[5] ——, *Characterization of C^n-operators*, Indiana University Math. J. *25* (1976), 119–133.

[6] REED, M. and SIMON, B., *Methods of modern mathematical physics*, II. Fourier Analysis, Self-adjointness. IV. Analysis of operators. Academic Press, New York (1975 and 1978).

[7] WIDDER, D. V., *The Laplace transform*, Princeton Univ. Press, Princeton, N.J. (1941).

Temple University
Philadelphia, Pennsylvania 19122 USA.

Bar-Ilan University
Ramat-Gan
Israel

Received June 16, 1980

Comment. Math. Helvetici **56** (1981) 179–195 0010-2571/81/002179-17$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Curvature, diameter and Betti numbers

MICHAEL GROMOV

We give an upper bound for the Betti numbers of a compact Riemannian manifold in terms of its diameter and the lower bound of the sectional curvatures. This estimate in particular shows that most manifolds admit no metrics of non-negative sectional curvature.

§0. Introduction

0.1. *Sectional curvature*

Let V denote a compact (without boundary) connected Riemannian manifold of dimension n. We denote by K the sectional curvature of V and we set $\inf K = \inf_\tau K(\tau)$ where τ runs over all tangent 2-planes in V. One calls V *a manifold of non-negative curvature* if $\inf K \geqslant 0$. This condition has the following geometric meaning.

An n-dimensional Riemannian manifold has non-negative curvature iff for each point $v \in V$ there is a positive number ε and a map f of the n-dimensional Euclidean ε-ball B into V with the following two properties:

(a) *f sends B diffeomorphicly onto the ε-ball in V with the center v.*

(b) *The map f is distance non-increasing, that is for any two points x and y in B one has.*

$$\operatorname{dist}(f(x), f(y)) \leqslant \operatorname{dist}(x, y),$$

where the first "dist" denotes the Riemannian distance in V and the second one is the Euclidean distance in $B \subset \mathbf{R}^n$.

Such a map f when it exists, is unique and it coincides with the so called exponential map (see [2], [4], [17]). In particular, f sends the center of B to v.

Observe, that the more general condition $\inf K \geqslant k$, $k \in (-\infty, +\infty)$, can be also interpreted geometrically. One should only use an ε-ball in the space of constant curvature k instead of the Euclidean ball B. For $k > 0$ one takes the sphere of radius $k^{-1/2}$ and for $k < 0$ one uses the hyperbolic space of curvature k.

Examples. Most known manifolds of non-negative curvature have the group theoretic origin. For instance, if V admits a smooth transitive action of a compact Lie group, then there is a Riemannian metric on V of non-negative curvature (see [4]). For each dimension ≥ 3 there are infinitely many homotopy types of such manifolds. Among other examples we mention only an exotic 7-sphere with a metric of non-negative curvature (see [8]) and the connected sum of two copies of the complex projective space (see [3]).

Counterexamples. The first topological obstruction for the existence of a metric of non-negative curvature on a compact manifold V was found by Bochner (see [1]).

Let V be a compact n-dimensional Riemannian manifold of non-negative curvature. Then $\operatorname{Dim} H_1(V, \mathbf{R}) \leq n$ *and the equality takes place only if V is flat.*

In fact, this theorem of Bochner remains true for a manifold V *with non-negative Ricci curvature. Furthermore, the universal covering of every manifold of non-negative curvature metrically splits into the product of* $\mathbf{R}^m$ *and a compact simply connected manifold* V^{n-m} (see [4], [6]).

This theorem reduces the problem to the case when the fundamental group $\pi_1(V)$ is finite.

There is another general obstruction for the existence of metrics of non-negative curvature (see [18]) and, in fact, this obstruction already appears for the manifolds with positive *scalar* curvature. Without going into details we mention only a few facts.

There are exotic 9*-spheres that carry no metrics of positive scalar curvature* (see [16]). *In particular they admit no metrics of non-negative sectional curvature.*

The product of an arbitrary manifold by the sphere S^m, $m \geq 2$, *admits a metric of positive scalar curvature. Furthermore, connected sums of manifolds of positive scalar curvature admit metrics with positive scalar curvature* (see [13], [19]).

We shall see below that most of these manifolds admit no metrics with non-negative *sectional* curvature.

Non compact manifolds. Every open manifold admits a *noncomplete* metric with positive sectional curvature (see [9]). On the other hand, when such a V is complete it must be homeomorphic to $\mathbf{R}^n$ (see [7]). When the curvature of a complete manifold V is non-negative, then V is homeomorphic to a vector bundle over a compact manifold of non-negative curvature (see [5]). This theorem brings us back to the compact case.

0.2. *Estimates for Betti numbers*

Fix a field F and denote by $b_i = b_i(V; F)$ the dimension over F of the homology group $H_i(V; F)$.

0.2.A. *There exists a constant $\mathscr{C} = \mathscr{C}(n)$, such that every compact connected n-dimensional Riemannian manifold V of non-negative sectional curvature satisfies*

$$\sum_0^n b_i \leqslant \mathscr{C}.$$

COROLLARY. *The connected sums of sufficiently many copies of the products of spheres $S^p \times S^{n-p}$, $0 < p < n$, or of the complex projective spaces, admit no metrics of non-negative curvature.*

Remarks. The n-dimensional torus is, probably, topologically the largest manifold of non-negative curvature, but our estimate for $\mathscr{C}(n)$ is very far from $2^n = \sum_0^n b_i(T^n)$. Even for $b_1(V, \mathbf{Z}_p)$ we can not get the expected estimate $b_1(V, \mathbf{Z}_p) \leqslant n$.

Let us replace now the condition $\inf K \geqslant 0$ by $\inf K \geqslant -\kappa^2$, $\kappa \geqslant 0$, and denote by D the diameter of V.

0.2B. *There exists a constant $\mathscr{C} = \mathscr{C}(n)$ such that every compact connected manifold V satisfies*

$$\sum_0^n b_i \leqslant \mathscr{C}^{1+\kappa D}.$$

Remarks

(a) When $\kappa = 0$ this theorem reduces to 0.2.A.

(b) The minimal number of generators of the fundamental group $\pi_1(V)$ is also bounded from above by $\mathscr{C}^{1+\kappa D}$ (see [10]).

(c) The connected sum of k copies of the product $S^p \times S^{n-p}$ can be equipped with a metric such that $2k+2 = \sum_0^n b_i \geqslant (1.01)^{1+\kappa D}$.

(d) The theorem 0.2B can be, probably, generalized to the manifolds with the *Ricci* curvature bounded from below, that is with $\inf_t (\mathrm{Ric}\,(t, t) \geqslant -\delta^2$, where t runs over all unite tangent vectors in V. But all known results on estimating topology of V by δD are tied up with the non torsion part of the fundamental group. For example, one can show that $b_1(V; \mathbf{R}) \leqslant n - 1 + \mathscr{C}^{\delta D}$ (this generalizes Bochner's theorem) but it is unknown whether this estimate holds for $b_1(V, \mathbf{Z}_2)$, even when V has positive Ricci curvature. We shall discuss the π_1-related estimates elsewhere (see also [11], [12]).

The proof of the theorem 0.2.A and 0.2.B is given in §1–§3. The curvature assumption essentially appears in this proof only once, in §1 for an analysis of the critical points of the Riemannian distance function as in [15]. This analysis is

based on Toponogov's comparison theorem (see §1). Although the curvature assumption is also present for estimating the number of small balls needed for a covering of a larger ball (compare to [20]), we could equally use for this purpose the Ricci curvature instead of the sectional curvature.

Acknowledgements. I owe my gratitude to Karsten Grove for several important critical remarks.

§1. Distance function

1.1. *Critical points*

Take a complete Riemannian manifold and a point x in V. Denote by $\mathrm{dist}_x : V \to \mathbf{R}_+$ the distance function $\mathrm{dist}_x(y) = \mathrm{dist}(x, y)$, $y \in V$. This function is not smooth but one can develop a complete Morse theory for this kind of a function. We shall need here only a few simple facts.

A point $y \in V$, $y \neq x$, is called *critical* for the function dist_x, or simply for x, if for every non-zero tangent vector $t \in T_y(V)$ there is a minimizing geodesic segment γ between x and y, such that the angle between t and γ at y is at most $\pi/2$. Recall, that a segment γ between x and y is called minimizing if

$$\mathrm{length}(\gamma) = \mathrm{dist}(x, y).$$

If a point $y_0 \in V$ is *not* critical for x, then there is a neighbourhood U of y_0 and a smooth vector field in U, that is $t(y)$, $y \in U$, such that for every point $y \in U$ the angle between the vector $t(y) \in T_y(V)$ and an arbitrary minimizing segment between x and y is an *acute* angle. It follows that the function dist_x is strictly decreasing along each integral curve of the field $t(y)$. This leads to the following fact that is a slight modification of a result of Grove–Shiohama [15].

ISOTOPY LEMMA. *Take two concentric balls B_1 and $B_2 \subset B_1$ in V centered at $x \in V$ and suppose that the closed annulus A between these balls, that is $A = Cl(B_1 \setminus B_2)$, contains no critical points of the function dist_x. Then there exists an isotopy of V which sends B_1 into B_2 and which is fixed outside any given neighbourhood of B_1.*

Proof. With the local fields $t(y)$ above one constructs a field $\tilde{t}$ on V which has its support in a small neighborhood of A and such that the function dist_x strictly decreases along the integral curves of $\tilde{t}$. This field performs the required isotopy.

1.2. *Comparison theorems*

Take three points x, y_1 and y_2 in V and take some minimizing segments γ_1 and γ_2 joining x with y_1 and with y_2 correspondingly. Denote by α the angle between γ_1 and γ_2 at x. Let l_1 denote length $(\gamma_1) = \text{dist}(x, y_1)$ and let l_2 denote length $(\gamma_2) = \text{dist}(x, y_2)$.

Toponogov's theorem (see [4], [17]). *If V is a complete manifold of non-negative curvature then*

$$\text{dist}(y_1, y_2) \leqslant \sqrt{l_1^2 + l_2^2 - 2l_1 l_2 \cos \alpha}.$$

Notice that for the Euclidean space $\mathbf{R}^n$ this inequality becomes an equality. We shall later use Topagonov's inequality only in the following two cases.

1.2.A. *Let $l_1 \geqslant l_2$ and let $\alpha \leqslant \pi/2$. Then*

$$\text{dist}(y_1, y_2) \leqslant l_1 + \tfrac{1}{2} l_2.$$

1.2.B. *Let again $l_1 \geqslant l_2$ and suppose that $\alpha \leqslant \frac{1}{6} \simeq \pi/18$. Then*

$$\text{dist}(y_1, y_2) \leqslant l_1 - \tfrac{3}{4} l_2.$$

Toponogov's inequality generalizes to all complete manifolds (see [4], [17]). In particular one has.

1.2.C. *If $\inf K \geqslant -\kappa^2$, $\kappa \geqslant 0$, and if the product $l_1 \kappa$ is sufficiently small, for example, if $l_1 \kappa \leqslant 10^{-10}$, then the inequalities* 1.2.A *and* 1.2.B *hold true.*

1.3. *An inequality for a critical point*

Take three points x, y and z in V and suppose that y is a critical point for x. Suppose further that $\text{dist}(z, x) \geqslant 2 \text{dist}(x, y)$.

If V has non-negative curvature, then

$$\text{dist}(z, x) \leqslant \text{dist}(z, y) + \tfrac{1}{2} \text{dist}(x, y). \qquad (*)$$

Proof. Take a minimizing segment γ_1 between z and y. According to the definition of the critical point there is a minimizing segment γ_2 between x and y such that the angle between γ_1 and γ_2 at y is at most $\pi/2$. The inequality $\text{dist}(z, x) \geqslant 2 \text{dist}(x, y)$ implies that length $(\gamma_1) \geqslant$ length (γ_2) and so we can apply 1.2.A.

Notice, that by the remark 1.2.C *the inequality* (*) *holds for a manifold* V *with* $\inf K \geqslant -\kappa^2$ *if* $\kappa(\operatorname{dist}(z, x)) \leqslant 10^{-10}$.

1.4. *An inequality for two critical points*

Take a point $x \in V$ and two critical points y_1 and y_2 for the distance function dist_x. Take some minimizing segments γ_1 and γ_2 joining x with y_1 and y_2 correspondingly and denote by α the angle between γ_1 and γ_2 at x.

If $\inf \kappa \geqslant 0$ *and if* $l_1 = \operatorname{dist}(x, y_1) \geqslant 2l_2 = 2 \operatorname{dist}(x, y_2)$, *then* $\alpha > \frac{1}{6}$

Proof. If $\alpha \leqslant \frac{1}{6}$, then, by 1.2.B, we have

$$\operatorname{dist}(y_1, y_2) \leqslant l_1 - \tfrac{3}{4}l_2.$$

Now we use the inequality (*) above with y_1 in place of z and with y_2 in place of y. We get

$$l_1 = \operatorname{dist}(y_1, x) \leqslant \operatorname{dist}(y_1, y_2) + \tfrac{1}{2}l_2, \qquad l_2 = \operatorname{dist}(x, y_2).$$

It follows that $l_2 = 0$, that is $x = y_2$, but this is not allowed by the definition of a critical point.

1.5. *Non compact manifolds*

Let us start with an obvious fact.

1.5.A. *Let* $t_1, \ldots, t_k$ *be non zero vectors in* $\mathbf{R}^n$, *such that the angle between any two of these vectors is at least* $\frac{1}{6}$. *Then the number* k *of these vectors does not exceed a universal constant,* $\operatorname{const}_n < (100)^n$.

Consider now a complete n-dimensional manifold V of non-negative sectional curvature and the distance function at a point x in V.

All critical points of the function dist_x *are contained in a compact ball around* x.

Indeed, we could find otherwise some critical points $y_1, \ldots, y_k$ such that $k > (100)^n$ and $\operatorname{dist}(x, y_i) \geqslant 2 \operatorname{dist}(x, y_j)$ for all $1 \leqslant i < j \leqslant k$. Take some minimizing segments $\gamma_1, \ldots, \gamma_k$ between x and $y_1, \ldots, y_k$ and denote by $t_1, \ldots, t_k$ their tangent vectors at x. According to 1.5.A some of these angles must be less than $\frac{1}{6}$, but this contradicts to 1.4.

As a corollary we get a weak version of a theorem of Cheeger–Gromoll (see [4], [5]).

The manifold V has "finite topological type" that is V is homeomorphic to the interior of a compact manifold with boundary.

Proof. Use the isotopy lemma in 1.1.

Our argument generalizes to a class of manifolds whose curvatures are "not very negative at infinity." Since this is a digression we leave the proof of the following theorem to the reader.

Take a point x in a complete manifold V and denote by $K_-(R)$ the infinimum of the sectional curvature of V outside the R-ball centered at x.

If $R^2K_-(R)\to 0$ as $R\to\infty$, then the function $\mathrm{dist}_x: V\to \mathbf{R}_+$ has its all critical points contained in a compact ball. In particular V is homeomorphic to the interior of a compact manifold V with boundary.

It will become clear later that the boundary V_0 of V is rather special. It must satisfy the inequality.

$$\sum_0^{n-1} b_i(V_0)\leqslant \mathscr{C}=\mathscr{C}(n).$$

§2. Coverings by balls

2.1. *Volumes of balls*

Let V be a complete n-dimensional manifold, such that $\inf K\geqslant -\kappa^2$. Denote by $b(R)$ the volume of a radius R ball in the hyperbolic space with curvature $-\kappa^2$. Take two concentric balls B_1 and $B_2\subset B_1$ in V of radii R_1 and R_2. The volumes of these balls are related as follows.

$$\frac{\mathrm{Vol}\,(B_1)}{\mathrm{Vol}\,(B_2)}\leqslant\frac{b(R_1)}{b(R_2)}. \tag{*}$$

See [2] for the proof. Notice that (*) also holds for $\inf_t \mathrm{Ric}\,(t,t)\geqslant -((n-1)\kappa)^2$ (see [2]). When V has non-negative curvature the inequality (*) says that

$$\frac{\mathrm{Vol}\,(B_1)}{\mathrm{Vol}\,(B_2)}\leqslant\frac{R_1^n}{R_2^n}.$$

If the balls are not supposed to be concentric the inequality (*) takes the following form

$$\frac{\mathrm{Vol}\,(B_1)}{\mathrm{Vol}\,(B_2)}\leqslant\frac{b(R_1+2d)}{b(R_2)}, \tag{**}$$

where d denotes the distance between the centers of B_1 and B_2. We shall use only the following two crude corrollaries of (**).

2.1.A. *If the balls B_1 and B_2 of the radii R_1 and $R_2 \leqslant R_1$ have a non-empty intersection then the inequality*

$$\frac{\mathrm{Vol}\,(B_1)}{\mathrm{Vol}\,(B_2)} \leqslant \frac{(10R_1)^n}{R_2^n} \qquad (***)$$

holds in the following two cases

(a) $\mathrm{Inf}\, K \geqslant 0$,

(b) $\mathrm{Inf}\,(K) \geqslant -\kappa^2$ *and the product* κR_1 *is sufficiently small, for example,* $\kappa R_1 \leqslant \exp(-n^n)$.

2.1.B. *Let V be a compact manifold of diameter D and let* $\mathrm{Inf}\, K = -\kappa^2$. *Then each ε ball B in V satisfies the following inequality.*

$$\frac{\mathrm{Vol}\,(V)}{\mathrm{Vol}\,(B)} \leqslant 10^n D^n \varepsilon^{-n} \exp(n\kappa D).$$

2.2. *Minimal coverings*

Take some sets $\{B_i\}_{i=1,\ldots,N}$ in V and denote first by I the set of all multiindices $(i_1 < i_2 < \cdots < i_l)$, $l = 1, \ldots, N$. Denote by I_+ the subset of I consisting of all multiindices $(i_1, \ldots, i_l)$, such that the intersection $\bigcap_1^l B_{i_j}$, $j = 1, \ldots, l$, is not empty. The number of the elements in I_+ is called the *index* of the system $\{B_i\}$. Clearly, the index takes the values between N and 2^N.

Take a ball B in V of radius R and cover it by some ε-balls, $0 < \varepsilon \leqslant R$, as follows. Take the *maximal* system of points x_i in B such that the distance between any two of them is greater than $\varepsilon/2$. In this case the ε-balls R_i around x_i cover B. We call such a system $\{B_i\}$ *a minimal ε-covering of B*. We want to estimate from above the index of such a covering in terms of the ratio $\varepsilon^{-1}R$ and $\mathrm{Inf}\, K$.

Here and in future for a ball B of radius r we denote by λB, $\lambda > 0$ the concentric ball of radius λr.

Let us return to our minimal ε-covering $\{B_i\}$, $i = 1, \ldots, N$. Observe that the balls $\frac{1}{4}B_i$ are disjoint and they are all contained in the ball $2B$. Now we invoke 2.1.A and conclude.

2.2.A. *If* $\operatorname{Inf} K \geqslant 0$, *or more generally, if* $\inf K \geqslant -\kappa^2$ *and* $2\kappa R \leqslant \exp(-n^n)$, *then the number* N *of the balls* B_i *does not exceed* $\operatorname{const}(n, \varepsilon^{-1}R) \leqslant (80\varepsilon^{-1}R)^n$. *Therefore the index of the covering does not exceed* 2^M, $M = (80\varepsilon^{-1}R)^n$.

This Lemma gives, in particular, a reasonable upper bound for the indices of minimal covering of V for $K \geqslant 0$, but in the general case of $\inf K \geqslant -k^2$, $K > 0$, the following sharper estimate is needed.

2.2.B. *Let* V *be as in* 2.1.B *and let* $\{B_i\}$, $i = 1, \ldots, N$, *be a minimal* ε*-covering of* V. *Take a number* $\lambda > 1$ *and let the product* $\lambda\varepsilon\kappa$ *be sufficiently small, for example,* $4\lambda\varepsilon\kappa < \exp(-n^n)$. *Then the index of the concentric covering* $\{\lambda B_i\}$ *does not exceed*

$$2^M 80^n D^n \varepsilon^{-n} \exp(n\kappa D), \qquad M = (160\,\lambda)^n.$$

Proof. First, we conclude as above that $N \leqslant 80^n D^n \varepsilon^{-n} \exp(n\kappa D)$. Now, if some balls λB_i intersect a fixed ball λB_{i_0}, then the centers of these balls must be contained in the ball $2\lambda B_{i_0}$ and so, by 2.2.A, each ball is involved in no more than 2^M intersections. Q.E.D.

2.3. *Topological Lemma*

Let V be an arbitrary complete Riemannian manifold of dimension n. Fix a coefficient field F and define *the content* of a ball B in V as the rank of the inclusion homomorphism

$$H_*(\tfrac{1}{5}B; F) \to H_*(B; F).$$

The number $\frac{1}{5}$ plays no essential role here, but it is convenient for our further constructions. Observe that the homology $H_*(\frac{1}{5}B; F)$ may be not finitely generated but the content of B is finite just the same. Notice also that balls of radii $>$ diam V are equal to V and so

$$\operatorname{cont}(V) = \sum_0^n b_i(V; F).$$

Take a ball B and cover the concentric ball $\frac{1}{5}B$ by some open balls B_i,

$i = 1, \ldots, N$, all of the same radius. Consider also the concentric coverings $\{\lambda_j B_i\}$, $j = 0, 1, \ldots, n+1$, $\lambda_j = 10^j$. Suppose that all balls $5\lambda_j R_i$, $j = 1, \ldots, n+1$, $i = 1, \ldots, N$, are contained in B and let the contents of all these balls be bounded by a constant p, that is

$$\text{Cont}\,(5\lambda_j B_i) \leqslant p, \qquad j = 0, \ldots, n+1, \qquad i = 1, \ldots, N.$$

Denote by J the index of the system $\{5\lambda_{n+1} B_i\}$, $i = 1, \ldots, N$.

The content of B satisfies the following inequality

$$\text{Cont}\,(B) \leqslant (n+1)pJ.$$

Proof. The ranks of the inclusion homomorphisms between all non-empty intersections of our balls,

$$H_*(\lambda_j B_{i_1} \cap \cdots \cap \lambda_j B_{i_k}) \to H_*(\lambda_{j+1} B_{i_1} \cap \cdots \cap \lambda_{j+1} B_{i_k}),$$

are estimated in terms of contents by interpolating pairs of balls,

$$\lambda_j B_{i_1} \cap \cdots \cap \lambda_j B_{i_k} \subset \lambda_j B_{i_1} \subset 5\lambda_j B_{i_1} \subset \lambda_{j+1} B_{i_1} \cap \cdots \cap \lambda_{j+1} B_{i_k}.$$

Then Leray's spectral sequence applies. See Appendix for the details.

2.4. *Main covering lemmas*

We return to the ball B of radius R as in 2.2 and we assume that $\inf K \geqslant 0$, or more generally, that $2\kappa R \leqslant \exp(-n^n)$, for $-\kappa^2 \leqslant \inf K$.

2.4.A. *Let for some number $p > 0$, the content of each ball of radius $r \leqslant 0.01R$ which intersects the ball $\frac{1}{5}B$ is bounded from above by p. Then*

$$\text{Cont}\,(B) \leqslant (n+1)pJ, \quad \text{for} \quad J = 2^M \quad \text{and} \quad M = 8^n\, 10^{n^2+4n}.$$

Remark. The numbers 0.01 and $\frac{1}{4}$ play no role here, but we shall need them later on.

Proof. According to 2.2.A the ball $\frac{1}{5}B$ can be covered by M balls B_i of radius $2 \cdot 10^{-n-4} R$ and the topological lemma applies.

2.4.B. *Let V be a compact manifold of diameter D and let $\inf K \geqslant -\kappa^2$. Let p and ε_0 be positive numbers such that each ε-ball, $\varepsilon \leqslant \varepsilon_0$, has content at most p, and*

such that $\varepsilon_0\kappa \leqslant \exp(n^{-n})$. *Then*

$$\operatorname{Cont}(V) \doteqdot \sum_0^n b_i(V; F) \leqslant (n+1)pJ, \quad \textit{for} \quad J = \text{const}\, D^n \varepsilon_0^{-n} \exp(n\kappa D),$$

where for const = const (n) *one can take* 2^M, $M = 10^{n^2+5n}$.

Proof. Use a minimal ε-covering of V for $\varepsilon = 5 \cdot 10^{-n-3}\varepsilon_0$ and apply 2.2.B and 2.3.

§3. Proof of the theorems 0.2.A and 0.2B

3.1 *Rank and Corank*

A ball B of radius R in a complete manifold V is called $\frac{1}{10}$-*critical* if there is a point $y \in V$ such that it is critical for the center x of B and such that $\operatorname{dist}(x, y) = 10R$.

Now let V have non-negative curvature. Take an arbitrary set A and define its *corank*, corank (A), as the maximal integer k, such that there exist some $\frac{1}{10}$-critical balls $B_1, \ldots, B_k$ with the following two properties.

(1) The radii R_i of B_i satisfy the inequalities $R_i \geqslant 3R_{i+1}$, $i = 1, \ldots, k-1$.

(2) The intersection $\bigcap_1^k B_i$ contains A.

There exists a positive integer $k_0 \leqslant (100)^n$, $n = \dim(V)$, *such that for every set* A *we have*

$$\operatorname{corank}(A) \leqslant k_0.$$

Remark. This proposition bounds the number of "essential directions" in V and the condition $K \geqslant 0$ is crucial.

Proof. Let $x_i \in V$ denote the centers of the balls B_i, $i = 1, \ldots, k$, and let y_i be the corresponding critical points for x_i with $\operatorname{dist}(x_i, y_i) = 10R_i$. Take a point z in $A \subset \bigcap_1^k B_i$ and join it by shortest segments γ_i with each of the points y_i.

If $k > (100)^n$, then there are two segments, γ_{i_1} and γ_{i_2}, $i_2 > i_1$, such that the angle between them at z is at most $\frac{1}{6}$ (see 1.5.A.). Now we argue as in Section 1.4. Set

$$l_1 = \operatorname{dist}(z, y_{i_1}) = \operatorname{length}(\gamma_{i_1}), \qquad l_2 = \operatorname{dist}(z, y_{i_2}) = \operatorname{length}(\gamma_{i_2}),$$

$$r_1 = \operatorname{dist}(z, x_{i_1}) \leqslant R_{i_1}, \qquad r_2 = \operatorname{dist}(z, x_{i_2}) \leqslant R_{i_2}, \qquad l = \operatorname{dist}(y_{i_1}, y_{i_2}).$$

The triangle inequality implies that

$$l_1 \geqslant 10R_{i_1} - r_1 \geqslant 9R_{i_1}, \qquad l_2 \leqslant 10R_{i_2} + r_2 \leqslant 11R_{i_2}.$$

Since $R_{i_1} \geqslant 3R_{i_2}$, we conclude that $l_1 > l_2$. Using 1.2.B, we get

$$l \leqslant l_1 - \tfrac{3}{4}l_2. \qquad (*)$$

Let d denote the distance between x_{i_2} and y_{i_1}, $d = \operatorname{dist}(x_{i_2}, y_{i_1})$. By the triangle inequality we have

$$d = l_1 - r_2 \geqslant 10R_{i_1} - r_1 - r_2 \geqslant 8R_{i_1} \geqslant 24R_{i_2} \geqslant 20R_{i_2} = 2\operatorname{dist}(x_{i_2}, y_{i_2}),$$

and so we can apply the inequality (*) in 1.3 with y_{i_1} in place of z and with x_{i_2} and y_{i_2} in place of x and y. We get $d \leqslant l + 5R_{i_2}$, and by the triangle inequality we have

$$l_1 \leqslant d + r_2 \leqslant d + R_{i_2} \leqslant l + 6R_{i_2}. \qquad (**)$$

The triangle inequality also implies that $l_2 \geqslant 10R_{i_2} - r_{i_2} \geqslant 9R_{i_2}$, and together with (*) this yields $l \leqslant l_1 - \frac{27}{4}R_{i_2}$, but this contradicts to (**). Q.E.D.

Now, if we have a manifold V with $\inf K \geqslant -\kappa^2$, we change the notion of the corank by adding the condition $2R_1\kappa \leqslant 10^{-10}$ and the inequality $\operatorname{corank}(A) \leqslant k_0 \leqslant (100)^n$ holds true. Now we set:

$$k_0 = \sup_{A \subset V} \operatorname{corank}(A), \quad \text{and} \quad \operatorname{rank}(A) = k_0 - \operatorname{corank}(A).$$

3.2. *Inductive lemmas*

Let B be a ball in V of rank zero. Then the content of this ball (see 3.2) is equal to one. In fact, if we look at the distance function dist_x, where x is the center of B, we shall see that it has no critical points in B; otherwise, for a sufficiently small concentric ball εB, we would get a contradiction, $\operatorname{corank}(\varepsilon B) \geqslant \operatorname{corank}(B) + 1$. The isotopy lemma (see 1.1) now shows that B is contractible and so $\operatorname{cont}(B) = 1$.

Denote by $\mathfrak{B}(k)$ the set of all balls in V of rank $\leqslant k$ and let p_k denote the upper bound

$$\sup_{B \in \mathfrak{B}(k)} \operatorname{cont}(B).$$

3.2.A. *Let V be a complete n-dimensional manifold of non-negative curvature. Then for each $k=0,1,2,\ldots$, the number p_{k+1} satisfies the inequality $p_{k+1}\leq (n+1)Jp_k$, where the constant $J=J(n)$ is the same as in* 2.4.A.

Since $k\leq k_0\leq(100)^n$, this lemma shows that

$$\sum_0^n b_i = \operatorname{cont}(V)\leq((n+1)J)^{100^n},$$

and this implies theorem 0.2.A. The proof of 3.2.A is given in the next section.

Notice that the lemma 3.2.A and its proof immediately extend to the general case of $\inf K\leq -\kappa^2<0$ if one modifies the definition of the numbers p_k by replacing the set $\mathfrak{B}(k)$ by the subset consisting of the balls of radius $<\varepsilon_0$ for $\varepsilon_0=\frac{1}{2}\kappa^{-1}\exp(-n^n)$. In view of 2.4.B, this general form of 3.2.A yields theorem 0.2.B.

3.3. *Incompressible balls*

Let V be a complete Riemannian manifold. A ball B in V of radius $R>0$ is called *compressible* if there exists a ball B' in V of radius $R'\leq\frac{1}{2}R$, such that B' is contained in B and such that there is an isotopy of V which is fixed outside B and which sends the ball $\frac{1}{5}B$ into $\frac{1}{5}B'$. It is clear that $\operatorname{cont}(B')\geq\operatorname{cont}(B)$, and so we conclude.

Each ball B contains an incompressible ball B_0 such that $\operatorname{cont}(B_0)\geq\operatorname{cont}(B)$. Now the inclusion $B_0\subset B$ implies $\operatorname{rank}(B_0)\leq\operatorname{rank}(B)$, and lemma 3.2.A becomes equivalent to the following more special lemma.

3.3.A *Let V be as in* 3.2.A, *and let B be an incompressible radius R ball in V of rank $k+1$, $k=0,1,\ldots$. Then*

$$\operatorname{cont}(B)\leq(n+1)Jp_k.$$

Proof. According to 2.4.A, we only have to show that each ball $\tilde{B}$, of radius $r<0.01R$ with the center at a point $\tilde{x}$ in the ball $\frac{1}{4}B$, has rank at most k. Look at the distance function $\operatorname{dist}_{\tilde{x}}$ and let us find an appropriate critical point of this function. Take the concentric ball $\tilde{B}'=(R/2r)\tilde{B}$ of radius R/2. This ball is contained in B but it contains the ball $\frac{1}{5}B$. Since by our hypothesis the ball B can not be compressed to $\tilde{B}'$, we conclude, in view of the isotopy (see 1.1), that the function $\operatorname{dist}_{\tilde{x}}$ must have a critical point $\tilde{y}$ such that

$$\tfrac{1}{10}R\leq\operatorname{dist}(\tilde{x},\tilde{y})\leq\tfrac{1}{2}R.$$

Now, by the definition of rank, there are some $\frac{1}{10}$-critical balls $B_1, \ldots, B_l$, $l = k_0 - k$, containing B and we take for B_{l+1} the ball concentric to B of radius $\frac{1}{10}$ dist (x, y). This ball contains $\tilde{B}$ and its radius is at least ten times less than the radius of the minimal of the balls $B_1, \ldots, B_l$. So the conditions (1) and (2) in 3.1 are met and rank $(\tilde{B}) <$ rank (B). Q.E.D.

Appendix: Leray spectral sequence

(1) *Filtered and graded spaces.* Recall, that a filtered vector space $\{F^i X\}$, $i = 0, 1, \ldots, n+1$, is defined as a decreasing sequence of subspaces

$$X = F^0 X \supset F^1 X \supset \cdots \supset F^n X \supset F^{n+1} X = \{0\}$$

The associated graded space to a filtered space $\{F^i X\}$ is the space

$$\operatorname{Gr} X = \bigoplus_{i=0}^{n} \operatorname{Gr}^i X,$$

where

$$\operatorname{Gr}^i X = F^i X / F^{i+1} X$$

A homomorphism f between two filtered spaces $\{F^i X\}$ and $\{F^i Y\}$ is, by definition, a linear map $f: X \to Y$ such that it sends each subspace $F^i X$ to $F^i Y$, $i = 0, \ldots, n+1$. Every such f gives rise to a *graded homomorphism* Gr f, that is a linear map Gr $X \to$ Gr Y which sends each Gr$^i X$ to Gr$^i Y$. It is clear that

$$\operatorname{rank} (\operatorname{Gr} f) \leqslant \operatorname{rank} (f),$$

but the equality does not in general hold.

Now, consider a sequence of filtered spaces $\{F^i X_j\}$, $i, j = 0, 1, \ldots, n+1$, and a sequence of homorphisms

$$f_j : \{F^i X_j\} \to \{F^i X_{j+1}\}.$$

Denote by $f: \{F^i X_0\} \to \{F^i X_{n+1}\}$ the composition, $f = f_n \circ f_{n-1} \circ \cdots \circ f_0$.

LEMMA. *The rank of the homomorphism f satisfies the following inequality*

$$\operatorname{rank} (f) \leqslant \sum_{j=0}^{n} \operatorname{rank} (\operatorname{Gr} f_j)$$

Proof. A standard induction reduces the lemma to the case of $n = 1$. Now, we have the following commutative diagram where the horizontal lines are exact

$$\begin{array}{ccccc} & & F^1X_0 & \to & X_0 & \to & \mathrm{Gr}^0 X_0 \\ & & \downarrow & & \downarrow f_0 & & \downarrow \mathrm{Gr}^0 f_0 \\ & & F^1X_1 & \to & X_1 & \to & \mathrm{Gr}^0 X_1 \\ F^1f_1 & \downarrow & & & \downarrow f_1 & & \downarrow \\ & & F^1X_2 & \to & X_2 & \to & \mathrm{Gr}^0 X_2. \end{array}$$

It is clear that

$$\mathrm{rank}\,(f_1 \circ f_0) \leq \mathrm{rank}\,(\mathrm{Gr}^0 f_0) + \mathrm{rank}\,(F^1 f_1).$$

(2) *Coverings and spectral sequences.* Let us recall some relevant fact on Leray's sequence (see [14]). For a set A in an n-dimensional manifold V we denote by $H_*(A)$ the total homology of A over a fixed coefficient field F,

$$H_*(A) = \bigoplus_0^n H_i(A; F).$$

Let $B_1, \ldots, B_N$ be some open sets in V. Then the homology of the union $A = \bigcup_1^N B_k$ carries a natural filtration $\{F^iH_*(A)\}$, $i = 0, 1, \ldots, n+1$. This is *not* the filtration associated to the grading $H_* = \oplus_0^n H_i$. If we take some larger open sets $B_k' \supset B_k$, then for their union A' the inclusion map $H_*(A) \to H_*(A')$ is a (filtered) homomorphism.

With the sets B_k above one associates Leray's spectral sequence that is a sequence $E_1, E_2, \ldots$, of vector spaces with the following properties.

(i) For each multiindex $\mu = \{i_1, \ldots, i_l\} \in I_+$ (see 2.2) we denote by H_*^μ the homology of the intersection $B_{i_1} \cap B_{i_2} \cap \cdots \cap B_{i_l}$. Then the space E_1 is isomorphic to $\oplus_{\mu \in I_+} H_*^\mu$.

(ii) Each space $E_1, E_2, \ldots$, has an additional structure of a complex, that is there are differentials $d_1 : E_1 \to E_1$, $d_1^2 = 0$, $d_2 : E_2 \to E_2$, $d_2^2 = 0, \ldots$. Furthermore each space E_{i+1} is obtained as the homology group of (E_i, d_i).

These structures are functional, that is the inclusions $B_i \to B_i'$ induce some homomorphisms $E_i \to E_i'$ which commute with d_i. The first homomorphism, $E_1 \to E_1'$, corresponds to the inclusion homomorphisms

$$f_\mu : H_*(B_{i_1} \cap \cdots \cap B_{i_l}) \to H_*(B_{i_1}' \cap \cdots \cap B_{i_l}'), \qquad \mu = (i_1, \ldots, i_l).$$

In particular, *the rank of the homomorphism* $E_1 \to E_1'$ *is equal to the sum* $\sum_{\mu \in I_+} \operatorname{rank}(f_\mu)$. Since each space E_{i+1} is the homology of (E_i, d_i) the rank of each homomorphism $E_i \to E_i'$ is bounded by the sum $\sum_\mu \operatorname{rank}(f_\mu)$.

(iii) For a sufficiently large i_0 the differentials d_i, $i > i_0$, vanish and so the sequence E_i stabilizes. The stable terms are denoted by E_∞. This space E_∞ is *functorially* isomorphic to the graded space associated to the filtered homology $\{F^i H_*(A)\}$. The word "functorially" means that the homomorphism $E_\infty \to E_\infty'$ corresponding to the inclusions $B_k \to B_k'$ is equal to the graded homomorphism associated to the (filtered) inclusion homomorphism $H_*(A) \to H_*(A')$.

(3) The following proposition generalizes the topological lemma of 2.3.

Let $B_k^i \subset V$, $k = 1, \ldots, N$, $i = 0, 1, \ldots, n+1$, $n = \dim(V)$, *be some open sets such that*

$$B_k^0 \subset B_k^1 \subset \cdots \subset B_k^{n+1}, \qquad k = 1, \ldots, N.$$

Let A^i *denote the unions* $\bigcup_{k=1}^n B_k^i$ *and let* f_μ^i *denotes the inclusion homomorphisms*

$$H_*(B_{i_1}^i \cap \cdots \cap B_{i_l}^i) \to H_*(B_{i_1}^{i+1} \cap \cdots \cap B_{i_l}^{i+1}), \qquad (i_1, \ldots, i_l) = \mu \in I_+.$$

Then the rank of the inclusion homomorphism $H_*(A^0) \to H_*(A^{n+1})$ *is bounded from above by the sum*

$$\sum_{\substack{i=0,\ldots,n \\ \mu \in I_+}} \operatorname{rank}(f_\mu^i).$$

Proof. The properties of the spectral sequence imply that the rank of each homomorphism $E_\infty^i \to E_\infty^{i+1}$ is bounded by

$$\sum_{\mu \in I_+} \operatorname{rank}(f_\mu^i),$$

and the lemma in (1) applies.

REFERENCES

[1] BOCHNER S. and YANO, K. *Curvature and Betti number*, Ann. Math. Stud. Princeton [1953].
[2] BISHOP, R. and CRITTENDEN, R. *Geometry of Manifolds*, A.e. Press. N.Y. [1964].
[3] CHEEGER, J. *Some examples of manifolds of non-negative curvature*, J. Diff. Geom. *8* (4) [1973], 623–629.

[4] —— and EBIN, D. *Comparison theorems in Riemannian geometry*, North Holland, N.Y. [1975].
[5] —— and GROMOLL, D. *The structure of complete manifolds of non-negative curvature*, Ann. Math. *96* [1972], 413–443.
[6] —— and GROMOLL, D. *The splitting theorem for manifolds of non-negative Ricci curvature*, J. Diff. Geom. *6* (1) [1971], 19–129.
[7] GROMOLL, D. and MEYER, W. *On complete open manifolds of positive curvature*, Ann. Math. *90* (1) [1969], 75–90.
[8] —— and ——, *An exotic sphere with non-negative sectional curvature*, Ann. Math. *100* (1) [1974], 401–406.
[9] GROMOV, M. *Stable maps of foliations into manifolds*, Isv. Ak. Nauk SSR. (192) [1969], 707–734.
[10] ——, *Almost flat manifolds*. J. Diff. Geom. *13* (2) [1978], 231–243.
[11] ——, *Synthetic geometry in Riemannian manifolds*, Proc, ICM-1978. Helsinki, [1980], 415–421.
[12] ——, *Volume and bounded cohomology*, to appear in Publ. IHES.
[13] —— and LAWSON, B. *The classification of simply connected manifolds of positive scalar curvature*, Ann. Math. *111* (1980) 423–435.
[14] GROTHENDIECK, A. *Sur quelques points d'algèbre homologique*, Toh. Math. J. (9) [1957], 119–221.
[15] GROVE, K. and SHIOHAMA, K. *A generalized sphere theorem*, Ann. Math. *106* [1977], 201–211.
[16] HITCHEN, N. *Harmonic spinors*, Adv. Math. *14* [1974], 1–55.
[17] KLINGENBERG, W., GROMOLL, D. and MEYER, W. *Riemannische Geometric im Grossen*, Springer Lecture Notes 55, [1978].
[18] LICHNEROWICZ, A. *Spineurs harmoniques*, C.R. Ac. Sci. Ser. A–B 257 [1963], 7–9.
[19] SCHOEN, R. and YAU, S. *On the structure of manifolds with positive scalar curvature* [1979], Manuscripta Math. *28* (1979) 159–183.
[20] WEINSTEIN, A. *On the homotopy type of positively pinched manifolds*, Arch. Math. (Basel) *18* [1967], 523–524.

Institut des Hautes Etudes Scientifiques
35, route de Chartres
91440 Bures-sur-Yvette

Received June 25, 1980

Comment. Math. Helvetici **56** (1981) 196–197 0010–2571/81/002196–02$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Correction to "On the characterization of flat metrics by the spectrum"

Comment. Math. Helvetici 55 (1980), 427–444

RUISHI KUWABARA

In order to derive Proposition 5.3 on page 437, we must give a more precise definition of the weak C^∞ Riemannian structure on the normed space X.

Consider a map $x(\in X)\mapsto \beta(x)$ (the weak inner product of X) such that the topology of $X_{\beta(x)}$ does not depend on x. Set $\beta=\beta(0)$. If $x\mapsto\beta(x)$ is a C^∞ map of X into $L(X_\beta, X_\beta; \boldsymbol{R})$, we call this map a *weak C^∞ Riemannian structure* on X.

Under this definition, subsequent discussions are valid by changing Lemma 6.2 on page 438 as follows.

LEMMA 6.2. *Let* $L_g=(1+\bar{\Delta}_g)^2$, *and* $s>(n/2)+3$, $s\geqq k$. *Then the maps*

$$\mathcal{R}^s\times S_2^k\to S_2^{k-4};\qquad (g,h)\mapsto L_g h$$

and

$$\mathcal{R}^s\times S_2^{k-4}\to S_2^k;\qquad (g,h)\mapsto L_g^{-1}h$$

are of C^∞ class.

Proof. Since the differential operator $(1+\bar{\Delta}_g)^2$ is an injective self-adjoint elliptic operator, L_g is an injective Fredholm operator from S_2^k to S_2^{k-4} for each $g\in\mathcal{R}^s$. First, we see that L_g has a continuous linear inverse $L_g^{-1}\colon S_2^{k-4}\to S_2^k$. For $k-4\geqq 0$, L_g is surjective by the decomposition theorem (e.g. [1, Ch. XI]). For a general k, $S_2^{k'}$ is a dense subspace of S_2^k if $k'\geqq k$, hence $L_g(S_2^k)$ is dense in S_2^{k-4} from the above. On the other hand, $L_g(S_2^k)$ is closed because L_g is a Fredholm operator. Thus, L_g is surjective and accordingly has a continuous inverse by the open mapping theorem. Next, it is easily shown that the maps are of C^∞ class on the same lines as the proofs by Omori [2, Lemmas 1.3 and 2.11]. □

REFERENCES

[1] R. PALAIS, Seminar on the Atiyah-Singer index theorem, Ann. of Math. Studies, No. 57, Princeton University Press, Princeton, 1965.
[2] H. OMORI, *On the group of diffeomorphisms on a compact manifold*, Proc. Sympos. Pure Math. 15, Amer. Math. Soc., 1970, 167–183.

Department of Applied Mathematics and Physics
Faculty of Engineering
Kyoto University
Kyoto 606, Japan

Received January 16, 1981

Comment. Math. Helvetici **56** (1981) 198–224 0010–2571/81/002198–27$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

An example of a Schroedinger equation with almost periodic potential and nowhere dense spectrum*

JÜRGEN MOSER

§1. Introduction

It is well known that the spectrum of a second order differential operator

$$L = -\left(\frac{d}{dx}\right)^2 + q(x)$$

with a continuous periodic potential $q(x) = q(x+l)$ with period $l > 0$ is given by an infinite or finite sequence of intervals extending to $+\infty$. In the latter case one of the intervals extends to infinity. Here we consider the unique selfadjoint extension of the above operator considered in $C''_{\text{comp}}(R)$, the space of twice continuously differential functions on the real line R with compact support. Such a selfadjoint extension, with dense domain in $L^2(R)$, is unique since this problem is in the "limit point case."

Little is known about the spectrum of such operators if $q(x)$ is a continuous almost periodic function, since one has too little information about the solutions of the differential equation $L\phi = \lambda\phi$ in this case. One knows examples of some hyperelliptic functions $q(x)$, which are almost periodic, even quasi-periodic for which the spectrum again consists of a finite number of intervals, one of which extends to $+\infty$. These hyperelliptic functions are, in fact, meromorphic on the Riemann surface constructed from two copies of the complex plane slit along the intervals which constitute the spectrum. For these considerations see Dubrovin, Novikov and Matveev [3, 9].

These examples are, however, in no way typical but rather exceptional, and generally one has to expect a much more complicated spectrum for almost periodic potentials. We want to mention a paper by Dynaburg and Sinai [4] in which it is shown that for quasi-periodic potentials with certain number theoretical restrictions on the frequency basis there exists a Cantor set extending to

* This work was partially supported by the National Science Foundation Grant, MCS 77-01986.

infinity which is contained in the spectrum. Recently Rüssmann [8] gave a simpler proof of this theorem and sharpened the result. However, these theorems only assert the containment of the Cantor set in the spectrum and it is not excluded that the spectrum is actually a halfline.

Here we want to construct a simple example of an almost periodic potential for which the spectrum is in fact a Cantor set. Actually, the potential is constructed as a limit periodic function, which is a function which can be uniformly approximated by periodic functions (see Besicovitch [1]). These form a special class of almost periodic functions for which the frequencies are rational multiples of one fixed positive number, which we could normalize to 1. Even for this simple kind of potential one finds such a complicated spectrum. We will show that *in the neighborhood of any periodic potential in $C(R)$-topology there are limit periodic potentials with nowhere dense spectrum.*

The idea of the construction is simply the following: We observe that for a periodic potential the gaps in the complement of the spectrum are given by the instability intervals, i.e. those λ-intervals for which the solutions are unbounded. If $M = M(\lambda)$ is the 2 by 2 matrix which takes

$$\begin{pmatrix} y(0) \\ y'(0) \end{pmatrix} \quad \text{into} \quad \begin{pmatrix} y(l) \\ y'(l) \end{pmatrix}$$

then the eigenvalues μ, μ^{-1} of M, the so-called Floquet multipliers, are either real or on the unit circle. They are the solutions of the quadratic equation

$$\mu^2 - \Delta\mu + 1 = 0$$

where $\Delta = \Delta(\lambda) = \operatorname{tr} M$ is the so-called discriminant. The instability intervals are characterized as those real λ for which $|\mu| \neq 1$ or $\Delta^2 > 4$, and their boundaries by $\mu^2 = 1$ or $\Delta = \pm 2$. It is possible that such an instability interval collapses, which amounts to a double root of $\Delta^2 - 4$. This situation occurs precisely if all solutions of the eigenvalue equation $L\phi = \lambda\phi$ are periodic of period $2l$; therefore one speaks of coexistence of periodic solutions, a case which has been frequently discussed in the literature [6, 7].

A potential $q(x)$ of period l has, of course, also $2l, 3l, \ldots$ etc. as periods, and considered as a function of period ml the matrix M has to be replaced by M^m and μ by μ^m, and Δ by $\Delta_m = \mu^m + \mu^{-m}$.

For $m = 2, 3, \ldots$ the roots of the equation $\Delta_m^2 = 4$ correspond to roots of unity $\mu : \mu^{2m} = 1$ and for these λ one has a periodic solution of period $2ml$. However, these roots, which are not also roots of $\Delta^2 = 4$, are all double roots of $\Delta_m^2 = 4$, i.e. we always have coexistence for all roots of $(\Delta_m^2 - 4)/(\Delta^2 - 4)$. This is simply due to

the fact that for eigenvalues μ, $\mu^{-1} = \bar{\mu}$ which are roots of unity different from ± 1 the eigenvectors of M are linearly independent and are the initial values of the desired periodic solutions. This coexistence is due to the fact that the potential has the period l while the period of the solutions is $2ml$. However, if we perturb the given potential with a function of period ml the coexistence will in general be destroyed and an interval of instability will be formed. If we apply such a sequence of perturbations with longer and longer periods the instability intervals will become dense and the resulting potential a limit periodic function.

The fact that generically coexistence of periodic solutions is destroyed is not surprising. But it is important to observe that a double root of $\Delta_m^2 - 4$ will not disappear under perturbation but generically split into two real roots. Assume we have $\Delta_m = +2$ at such a double root, then it turns out that after a small perturbation the local maximum of the corresponding discriminant is always $\geqq 2$ and generically >2. This fact can be derived from a closer investigation of the second variation of Δ_m (see Section 4).

Aside from proving the above statement we will discuss the Floquet theory from a geometrical point of view. Our discussion is based on the rotation number

$$\alpha = \lim_{x \to \infty} \frac{\arg (y(x) - iy'(x))}{x}$$

which exists for any periodic potential, and is the same for any nontrivial solution y of $Ly = \lambda y$. This defines a continuous monotone increasing function $\alpha = \alpha(\lambda) \geqq 0$ which is constant on the intervals of instability, but strictly increasing on the spectrum. In fact, in the interior of the spectrum one has

$$\frac{d(\alpha^2)}{d\lambda} \geqq 1$$

where equality can occur at any point only if q is a constant. The constant values of α on the instability interval are of the form $(\pi/l)j$, $j = 1, 2, 3, \ldots$ In [3] the quantity $\alpha(\lambda)$ is referred to as "quasimomentum."

It is conceivably a generic phenomenon, in the sense of Baire category, that for a limit periodic, or even an almost periodic potential the spectrum is such a Cantor set. This will not be proven here.* In particular, it would require the existence and continuity of the above rotation number α. Actually this limit exists and depends continuously on λ for an arbitrary continuous almost periodic potential, as was proven recently by Russell Johnson, USC. Moreover, a spectral theoretical interpretation of $\alpha = \alpha(\lambda)$ can be given, showing that the set of constancy intervals of $\alpha(\lambda)$ agree with the complement of the spectrum.

* Added in proof: We learned that J. Avron and B. Simon just have proved such a result.

These results will be presented elsewhere and here we restrict ourselves to proving the underlined statement.

Our aim will be to construct a limit periodic potential $q(x)$ for which the function $\alpha = \alpha(\lambda)$ also exists and is a Cantor function: It is continuous and takes any rational value of the form $j/2^n$, $j \geqq 1$, $n \geqq 0$ both integers, on an interval of positive length. The complement of these intervals forms the spectrum of the operator $L = -(d/dx)^2 + q(x)$.

It is a known phenomenon that the rotation number $\alpha = \alpha(\lambda)$ of a circle mapping $M(\lambda)$ depending on a parameter generally takes all rational values on intervals of positive length. This is connected with the "lock in" phenomenon of nonlinear oscillation. A rational mapping of this type can be found in [5]. For the circle mapping induced by

$$\begin{pmatrix} y \\ y' \end{pmatrix} \to M(\lambda) \begin{pmatrix} y \\ y' \end{pmatrix}$$

on the rays through the origin this does not happen however, because of the linearity of this mapping. The point of the present note can be seen in the observation that the phenomenon of dense resonance can occur even in the linear case, if the coefficients are almost periodic.

The paper [10] by G. Scharf is devoted to the study of the spectrum for almost periodic potentials. His result can be also used to effectively investigate the rotation number $\alpha(\lambda)$, and describe $\alpha(\lambda)$ as the boundary values of a positive harmonic function in the upper halfplane. However, the phenomenon of dense spectrum was not noted there. In some recent preliminary draft Joseph Avron and Barry Simon ("Cantor Sets, Almost Periodic Hill's Equations, and the Rings of Saturn," Cal. Inst. Technology) announced various statements about the spectrum of almost periodic potentials, as well as the appearance of Cantor sets as spectrum.

Finally, I should like to thank Russell Johnson for supplying me with a proof of the existence proof for the rotation number for almost periodic potential as well as for substantial advice. Also I am grateful to P. Deift and E. Zehnder for reading the manuscript and suggesting improvements.

§2. The rotation number

We consider the differential equation

$$y'' = Q(x)y \tag{2.1}$$

where $Q(x)$ is a real continuous function of period l. Later on we will take $Q(x)=q(x)-\lambda$ so that Q depends on the parameter λ. With this equation one can associate a flow on, a torus by identifying points $c\begin{pmatrix} y \\ y' \end{pmatrix}$ for $c\neq 0$ in the plane. Or introducing polar coordinates (Prüfer transformation, see [7])

$$y=r\sin\theta, \qquad y'=r\cos\theta \quad \text{for} \quad r>0$$

one finds the differential equation

$$\frac{d\theta}{dx}=\cos^2\theta-Q(x)\sin^2\theta \tag{2.2}$$

which can be viewed as a vectorfield on the torus by identifying $\theta \bmod 2\pi$ and $x \bmod l$. For such a vectorfield the rotation number

$$\alpha=\lim_{x\to\infty}\frac{\theta(x)}{x}$$

exists and is independent of the choice of $\theta(0)$. Since $\arg(y-iy')=\theta-\pi/2 \pmod{2\pi}$ it follows that for any non vanishing solution $y(x)$ of (2.1)

$$\alpha=\lim_{x\to\infty}\frac{\arg(y(x)-iy'(x))}{x}$$

exists and is independent of the particular solution. It also does not matter that the argument is defined only mod 2π since this constant drops out in the limit.

It is useful to express α in terms of the number $N(x)$ of zeroes of a nontrivial solution in the interval $(0, x)$. Since at zero ξ of $y(x)$ the angle θ is an integral multiple of π one sees from (2.2) that $d\theta/dx=1>0$ at $x=\xi$, i.e. θ increases at such a value. From this fact it follows readily that

$$\theta(x)-\pi N(x)$$

is bounded for all x and hence

$$\alpha=\pi\lim_{x\to\infty}\frac{N(x)}{x}. \tag{2.3}$$

By Sturm's comparison theorem [6] one sees at once: If $Q(x)\leqq\tilde{Q}(x)$ are two

periodic function and $\alpha, \tilde{\alpha}$ the corresponding rotation numbers then

$$\alpha \geqq \tilde{\alpha}, \tag{2.4}$$

for any choice of the corresponding nontrivial solutions $y, \tilde{y}$. This shows again the independence of α from the particular solution, if we take $\tilde{Q} = Q$. Moreover, by (2.3),

$$\alpha \geqq 0. \tag{2.5}$$

Finally we relate α to Floquet theory, in the case that the Floquet multipliers μ, μ^{-1} are not real, the stable case. In that case (2.1) admits a complex solution $w(x)$ satisfying

$$w(x+l) = \mu w(x), \tag{2.6}$$

or with $\mu = e^{i\beta}$, β being defined only mod 2π,

$$w(x) = e^{i\beta x/l} p(x)$$

where $p(x)$ has period l and does not vanish. Therefore $p(x)$ describes a closed curve with an integer winding number

$$\frac{1}{2\pi i} \int_0^l \frac{p'}{p}\, dx = j.$$

Replacing β by $\beta + 2\pi j$ we can achieve that this winding number is zero and this way define β completely and not only mod 2π. We can take $\beta \geqq 0$, otherwise we replace w by $\bar{w}$ and hence β by $-\beta$.

This solution can be written in the form

$$w(x) = e^{i(\beta x/l)} \rho(x) e^{is(x)} \tag{2.7}$$

where both $\rho(x) > 0$, $s(x)$ have period l. We show that

$$\beta = l\alpha. \tag{2.8}$$

For this purpose we consider the real solution

$$y = \operatorname{Re} w(x) = \rho(x) \cos\left(\frac{\beta x}{l} + s(x)\right)$$

and note that the number $N(x)$ of zeroes in $(0, x)$ satisfies

$$N(x)-\frac{\beta}{\pi}\frac{x}{l}=0(1)$$

so that (2.3) implies (2.8).

The Floquet solution (2.6) is determined only up to a complex factor $c\neq 0$. One usually normalized $|c|$ by requiring that the Wronskian

$$w\bar{w}'-\bar{w}w'=-2i.$$

One computes from (2.7) that this Wronskian has the value

$$-2i\rho^2\left(\frac{\beta}{l}+s'\right)$$

so that

$$\frac{\beta}{l}+s'=\rho^{-2}.$$

Inserting this into (2.7) we find the differential equation

$$\rho''=Q\rho+\rho^{-3}. \tag{2.9}$$

From the above relation we find

$$\alpha=\frac{\beta}{l}=\lim_{x\to\infty}\frac{1}{x}\int_0^x \rho^{-2}(t)\,dt.$$

On the other hand ρ being a positive function of period l we find

$$\alpha=\frac{1}{l}\int_0^l \rho^{-2}\,dx=\frac{1}{l}\int_0^l |w(x)|^{-2}\,dx \tag{2.10}$$

showing again $\alpha\geqq 0$, and even $\alpha>0$ in case of stability.

For an actual calculation of $\alpha \pmod{(2\pi/l)}$ we return to the Floquet matrix: Let $Y(x)$ be the two by two matrix satisfying

$$Y'=\begin{pmatrix}0 & 1\\ Q(x) & 0\end{pmatrix}Y, \qquad Y(0)=1. \tag{2.11}$$

Then the Floquet matrix M is given by $M = Y(l)$ and its eigenvalues μ by

$$\mu^2 - \Delta\mu + 1 = 0 \quad \text{where} \quad \Delta = \operatorname{tr} M.$$

Hence in the stable case, $\mu = e^{i\beta}$ we have

$$\Delta = \mu + \mu^{-1} = 2\cos\beta$$

or

$$\Delta = 2\cos l\alpha. \tag{2.12}$$

Now we take $Q(x) = q(x) - \lambda$, so that, for fixed $q(x)$, α becomes a function of λ. By (2.4) it is a monotone increasing which is, moreover continuous. This is a well known property of the rotation number, due to P. Bohl (see, e.g. [5]). For $\lambda < \min q(x)$ one has $Q(x) > 0$, and so any solution has at most one zero, i.e. $\alpha(\lambda) = 0$. On the other hand, again by Sturm comparison theorems $\alpha(\lambda) \to +\infty$ as $\lambda \to \infty$, so that α takes on all nonnegative values. In the stability intervals one has

$$\mu = e^{il\alpha}$$

for the Floquet multiplier. Because of the continuity and monotonicity one concludes that in the instability intervals α is a constant, an integer multiple of π/l.

If we define the clipped discriminant

$$\Delta_{cl}(\lambda) = \begin{cases} +2 & \text{if} \quad \Delta(\lambda) > 2 \\ \Delta(\lambda) & \text{if} \quad -2 \leqq \Delta(\lambda) \leqq +2 \\ -2 & \text{if} \quad \Delta(\lambda) < -2 \end{cases}$$

then we have from (2.12)

$$\Delta_{cl}(\lambda) = 2\cos l\alpha(\lambda).$$

Incidentally, although α is differentiable in the interior of the stability interval it does not even have a one-sided derivative at the boundary of an uncollapsed instability interval, since α behaves like $\sqrt{|\lambda - \lambda_0|}$ there.

§3. $\alpha(Q)$ as a functional

We saw that $\alpha(Q)$ is a monotone decreasing functional in the sense of (2.4). It is generally not differentiable but for a potential Q for which one has nonreal

multipliers μ it has a functional derivative $\delta\alpha/\delta Q = -k(x)$, defined by

$$\frac{d}{d\varepsilon}\,\alpha(Q+\varepsilon\hat{q})\,|_{\varepsilon=0} = -\int_0^l k(x)\hat{q}(x)\,dx$$

for any continuous $\hat{q}(x)$ of period l. We insert the minus sign, since $k(x)$ will turn out to be positive. In particular, if $Q = q(x) - \lambda$ we set $\alpha(Q) = \alpha(\lambda, q)$ and get, with $\hat{q} = -1$,

$$\frac{\partial\alpha}{\partial\lambda} = \int_0^l k(x)\,dx. \tag{3.1}$$

Dubrovin [2] showed that

$$k(x) = \frac{\rho^2}{2l} = \frac{|w|^2}{2l} \tag{3.2}$$

where w is the Floquet solution (2.6) normalized by $w\bar{w}' - \bar{w}w' = -2i$. We will derive the relation (3.2) at the end of this section. Thus $k(x)$ is positive (if the Floquet multiplier is not real) which shows that $\alpha = \alpha(\lambda, q)$ satisfies the estimate

$$|\delta\alpha| < \sup_x |\delta q| \cdot \int_0^l k(x)\,dx = \sup_x |\delta q| \cdot \frac{\partial\alpha}{\partial\lambda}. \tag{3.3}$$

We also derive from (3.1) and (2.10)

$$\frac{d(\alpha^2)}{d\lambda} = 2\alpha\frac{\partial\alpha}{\partial\lambda} = \frac{1}{l^2}\int_0^l \rho^{-2}\,dx \int_0^l \rho^2\,dx \geqq 1$$

by Schwarz' inequality. Equality on the right occurs only if ρ, and hence q, is a constant. This proves a statement of the introduction.

Although the functional $\alpha = \alpha(Q)$ is not everywhere differentiable, the discriminant $\Delta = \operatorname{tr} Y(l)$ is.

By solving the differential equation

$$Z' = \begin{pmatrix} 0 & 1 \\ Q & 0 \end{pmatrix} Z + (\delta Q) E Y; \qquad Z(0) = 0$$

for the first variation $Z = \delta Y$ of Y where $Y(x)$ is the fundamental solution of (2.11), we verify readily

$$\frac{\delta\Delta}{\delta Q} = \operatorname{tr}\left(Y(l)Y^{-1}(x)EY(x)\right), \qquad E = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \tag{3.4}$$

In similar forms this formula can be found in [7], [9]. We rewrite it in several different forms.

If we write the fundamental matrix $Y(x)$ in the form

$$Y(x) = \begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix}$$

then y_1, y_2 are linearly independent solutions of (2.1) normalized at $x = 0$. In terms of these solutions (3.4) takes the form

$$\begin{aligned} &\frac{\delta \Delta}{\delta Q} = A y_1^2(x) + B y_1(x) y_2(x) + C y_2^2(x) \\ &A = y_2(l), \qquad B = y_2'(l) - y_1(l), \qquad C = y_1'(l). \end{aligned} \tag{3.5}$$

The three products $y_1^2, y_1 y_2, y_2^2$ are linearly independent functions since their Wronskian is

$$2(y_1 y_2' - y_1' y_2)^3 = 2 \neq 0.$$

In particular, we see from (3.5) that Δ has a critical point, i.e. $\delta \Delta / \delta Q = 0$, if and only if $A = B = C = 0$, i.e. if $Y(l)$ is a multiple of the identity. Since its determinant is 1 this means that $Y(l) = \pm 1$ which in turn means that all solutions of (2.1) satisfy

$$y(x + l) = \pm y(x), \tag{3.6}$$

or $y^2(x)$ has period l. We conclude: *The discriminant $\Delta(Q)$ has a critical point at Q if and only if $y^2(x)$ has period l for all solutions of* (2.1). In terms of the function $\Delta = \Delta(\lambda, q)$ the condition is equivalent for $\Delta^2 - 4$ to have a double zero, as function of λ (see [7]).

The gradient $\delta \Delta / \delta q$ can be represented in another way as

$$\frac{\delta \Delta}{\delta Q}(\xi) = \frac{\phi(\xi + l)}{\phi'(\xi)} \tag{3.7}$$

where $\phi(x)$ is a nontrivial solution of (2.1) with $y(\xi) = 0$, $y'(\xi) \neq 0$; of course, $\phi(x)$ depends also on ξ.

It suffices to prove this result for $\xi = 0$, since the general result follows by translation. Setting $x = 0$ in (3.5) we find

$$\frac{\delta \Delta}{\delta Q}(0) = y_2(l)$$

which proves (3.7) since $y_2(0)=0$, $y_2'(0)=1$. We conclude from (3.7) that $\delta\Delta/\delta Q$ has period l.

Finally, we express $\delta\Delta/\delta Q = K(x)$ in terms of the Floquet solution (2.7), assuming now that the Floquet multiplier of (2.1) is not real. Then on account of (3.5) $\delta\Delta/\delta Q$ can be expressed as a linear combination of the 3 products w^2, $w\bar{w}$, $\bar{w}^2$ and since it is of period l it must be a multiple of $w\bar{w}=\rho^2$. If we normalize w again by $w\bar{w}'-w'\bar{w}=-2i$ the constant is determined by

$$\frac{\delta\Delta}{\delta Q}=\rho^2 \sin l\alpha,$$

as we will show now. This will lead to a proof of (3.2) if we differentiate $\Delta = 2\cos l\alpha$ and use $\delta\alpha/\delta Q = -k$.

The determination of this constant could be done as follows: We use that the products y_1y_2 of solutions of (2.1) as well as K satisfy the third order differential equation

$$K'''-4QK'-2Q'K=0.$$

In the stable case K is a solution of period l of this equation, by which it is determined up to a constant. This equation possesses the integral

$$K(K''-2QK)-\tfrac{1}{2}K'^2$$

as one verifies immediately. Its value can be determined at $x=0$, since from (3.5) one finds $K(0)=y_2(l)$; $K'(0)=y_2'(l)-y_1(l)$, $K''(0)=2(y_2(l)Q(l)-y_1'(l))$ that the above integral has the value

$$\begin{aligned}
&=-2y_1'(l)y_2(l)-\tfrac{1}{2}(y_2'(l)-y_1(l))^2\\
&=-\tfrac{1}{2}(y_2'(l)+y_1(l))^2+2(y_1(l)y_2'(l)-y_1'(l)y_1(l))\\
&=-\tfrac{1}{2}(\operatorname{tr} Y(l))^2+2\\
&=\tfrac{1}{2}(4-\Delta^2)=-\tfrac{1}{2}(\mu-\mu^{-1})^2=2\sin^2 l\alpha.
\end{aligned}$$

On the other hand from the differential equation (2.9) we obtain for $R=\rho^2$

$$R(R''-2QR)-\tfrac{1}{2}R'^2=2$$

so that indeed

$$K=\pm\rho^2 \sin l\alpha.$$

To show that the sign is + we can use the monotonicity of α.

§4. The second variation of Δ

The collapse of an instability interval at a point λ^* corresponds to a double zero of Δ^2-4, which by the previous section is tantamount to the vanishing of the first variation $\delta\Delta/\delta Q$ at $Q=q(x)-\lambda^*$. At such a point we will need the second variation of Δ, which we now compute.

Therefore we assume that for the considered function $Q(x)$ all solutions of (2.1) satisfy (3.6). Replacing $Q(x)$ by $Q(x)+\varepsilon\hat{q}(x)$ we determine the corresponding fundamental matrix

$$Y(x)+\varepsilon Y_1+\varepsilon^2 Y_2+\cdots.$$

It satisfies the equations

$$Y'=FY$$

$$Y_1'=FY_1+\hat{q}EY$$

$$Y_2'=FY_2+\hat{q}EY_1$$

where

$$F=\begin{pmatrix}0 & 1\\ Q & 0\end{pmatrix};\qquad E=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix};\qquad Y_1(0)=Y_2(0)=0.$$

With

$$S(x)=Y^{-1}(x)EY(x)$$

one solves the equations in the form

$$Y_1(x)=Y(x)\int_0^x S(t)\hat{q}(t)\,dt$$

$$Y_2(x)=Y(x)\int_0^x\int_0^t S(t)S(s)\hat{q}(t)\hat{q}(s)\,dx\,dt.$$

From these relations one computes readily

$$\frac{1}{2}\frac{d^2\Delta}{d\varepsilon^2}\bigg|_{\varepsilon=0}=\operatorname{tr} Y_2(l)=\pm\frac{1}{2}\int_0^l\int_0^l \operatorname{tr}\,(S(t)S(s))\hat{q}(t)\hat{q}(s)\,ds\,dt$$

where the sign corresponds to that of $Y(l) = \pm 1$. Further we find

$$\operatorname{tr}(S(t)S(s)) = -(y_1(t)y_2(s) - y_1(s)y_2(t))^2.$$

This expression is the same for any basis of solutions $\phi_1(x)$, $\phi_2(x)$ with Wronskian $+1$, so we have

$$\pm\frac{1}{2}\frac{d^2\Delta}{d\varepsilon^2}\Big|_{\varepsilon=0} = -\frac{1}{2}\int_0^l\int_0^l (\phi_1(t)\phi_2(s) - \phi_1(s)\phi_2(t))^2\hat{q}(t)\hat{q}(s)\,ds\,dt$$
$$= -\int_0^l \phi_1^2\hat{q}\,dx\int_0^l \phi_2^2\hat{q}\,dx + \left(\int_0^l \phi_1\phi_2\hat{q}\,dx\right)^2. \tag{4.1}$$

Finally, if we introduce the 3 functions

$$\Phi_1 = \tfrac{1}{2}(\phi_1^2 + \phi_2^2), \qquad \Phi_2 = \tfrac{1}{2}(\phi_1^2 - \phi_2^2), \qquad \Phi_3 = \phi_1\phi_2 \tag{4.2}$$

of period l we get

$$\pm\frac{1}{2}\frac{d^2\Delta}{d\varepsilon^2}\Big|_{\varepsilon=0} = -\left(\int_0^l \Phi_1\hat{q}\,dx\right)^2 + \left(\int_0^l \Phi_2\hat{q}\,dx\right)^2 + \left(\int_0^l \Phi_3\hat{q}\,ds\right)^2. \tag{4.3}$$

We see from this formula that the second variation is a quadratic form of rank 3 and its null space N is the orthogonal complement of the space of Φ_1, Φ_2, Φ_3. Moreover, the quadratic form (4.3) has a 1-dimensional subspace on which it is negative definite and a two-dimensional subspace on which it is positive; it is of type (2, 1).

We obtain for $Q = q(x) - \lambda$, $\hat{q} = -1$ the second λ-derivative, and get from (4.1)

$$\pm\frac{1}{2}\frac{d^2\Delta}{d\lambda^2} = -\int_0^l \phi_1^2\,dx\int_0^l \phi_2^2\,dx + \left(\int_0^l \phi_1\phi_2\,dx\right)^2 < 0.$$

The negativity follows from Schwarz' inequality. Hence, at a point λ^* at which $\Delta(\lambda^*) = +2$, $\Delta'(\lambda^*) = 0$ one has always a nondegenerate maximum, and similarly for $\Delta(\lambda^*) = -2$, $\Delta'(\lambda^*) = 0$ a nondegenerate minimum.

PROPOSITION 1. *Consider the potential* $q(x) + \varepsilon q_1(x)$ *of period* l *where for* λ^*

$$\Delta(\lambda^*, q) = \pm 2, \qquad \Delta'(\lambda^*, q) = 0.$$

Unless q_1 *lies in the space span* $\{1, N\}$ (*where* N *is the orthogonal complement of*

Φ_1, Φ_2, Φ_3) then for $\varepsilon \neq 0$ sufficiently small at some point $\lambda^*(\varepsilon)$ near λ^* we have

$$|\Delta(\lambda^*(\varepsilon), q + \varepsilon q_1)| > 2.$$

This proposition expresses that generically a collapsed instability interval will "open up."

Proof. We give the proof in the case $\Delta(\lambda^*, q) = 2$. With some constant κ we set

$$Q(x, \varepsilon) = q(x) - \lambda^* + \varepsilon \hat{q}(x), \qquad \hat{q} = q_1(x) - \kappa.$$

If we determine κ so that

$$0 = \int_0^l \Phi_1 \hat{q}\, dx = \int_0^l \Phi_1 q_1\, dx - \kappa \int_0^l \Phi_1\, dx$$

– this is possible since $\Phi_1 > 0$ – then the right hand side of (4.3) is positive, unless $\hat{q}$ is orthogonal to Φ_2, Φ_3. By construction it is orthogonal to Φ_1, i.e. $\hat{q} = q_1 - \kappa \in N$, which would contradict our assumption. Hence for $\lambda^*(\varepsilon) = \lambda^* + \kappa\varepsilon$ we have

$$Q(x, \varepsilon) = q(x) - \lambda^*(\varepsilon) + \varepsilon q_1(x)$$

and from the Taylor expansion of $\Delta(Q)$

$$\Delta(\lambda^*(\varepsilon), q + \varepsilon q_1) > 2$$

if $\varepsilon \neq 0$ is sufficiently close to zero, as was to be proven. The same argument holds clearly for the case $\Delta(\lambda^*, q) = -2$.

In particular, we can always choose q_1 so, that the local maximum of $\Delta(\lambda, q + \varepsilon q_1)$ near λ is larger than 2.

This can be made more explicit if one makes use of the fact that the basis ϕ_1, ϕ_2 can be replaced by $(a\phi_1 + b\phi_2, c\phi_1 + d\phi_2)$ where $ad - bc = 1$ by which Φ_1, Φ_2, Φ_3 are replaced by

$$\sum_{k=1}^{3} c_{jk} \Phi_k$$

where the matrix $C=(c_{jk})$ is given by

$$C=\begin{pmatrix} \frac{1}{2}(a^2+b^2+c^2+d^2) & \frac{1}{2}(a^2-b^2+c^2-d^2) & ab+cd \\ \frac{1}{2}(a^2+b^2-c^2-d^2) & \frac{1}{2}(a^2-b^2-c^2+d^2) & ab-cd \\ ac+bd & ac-bd & ad+bc \end{pmatrix}.$$

PROPOSITION 2. *It is always possible to find a basis* ϕ_1, ϕ_2 *of* (1.2) *with Wronskian* $=1$ *such that the three functions* Φ_1, Φ_2, Φ_3 *of* (4.2) *satisfy*

$$\int_0^l \Phi_j \Phi_k \, dx = 0 \quad \text{for} \quad j \neq k.$$

In other words, for this basis we have

$$\int_0^l \phi_1^4 \, dx = \int_0^l \phi_2^4 \, dx; \qquad \int_0^l \phi_1^3 \phi_2 \, dx = \int_0^l \phi_1 \phi_2^3 \, dx = 0.$$

We indicate the proof. The main observation is that the mapping $x \to Cx$; $x \in R^3$ preserves the quadratic form $x_1^2 - x_2^2 - x_3^2$, so that $C \in SO(1,2)$. In fact the mapping

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \to C$$

defines a homomorphism of $Sl(2)$ into $SO(1,2)$. This mapping is not onto, but only onto the component of $SO(1,2)$ containing the identity, which is characterized by $c_{11} > 0$.

Now we apply the theorem of linear algebra that the two quadratic forms

$$\int_0^l \left(\sum_{j=1}^3 x_j \Phi_j \right)^2 dx \quad \text{and} \quad x_1^2 - x_2^2 - x_3^2$$

can be simultaneously diagonalized, i.e. brought into the form

$$\kappa_1 x_1^2 + \kappa_2 x_2^2 + \kappa_3 x_3^2, \qquad x_1^2 - x_2^2 - x_3^2$$

by a linear transformation. Such a transformation belongs to $SO(1,2)$ since the second form is preserved. Since we can still replace x_j by $\pm x_j$ we can assume that the linear transformation belongs to the component of $SO(1,2)$ containing the

identity. Therefore this transformation can be realized by a change of basis $(\phi_1, \phi_2) \to (a\phi_1 + b\phi_2, c\phi_1 + d\phi_2)$.

If we write $\hat{q}$ in terms of such a basis

$$\hat{q} = \sum_{j=1}^{3} \xi_j \Phi_j + n, \qquad n \in N$$

then (4.3) takes the form

$$\pm \frac{1}{2} \frac{d^2 \Delta}{d\varepsilon^2} \bigg|_{\varepsilon=0} = -A_1 \xi_1^2 + A_2 \xi_2^2 + A_3 \xi_3^2 \tag{4.4}$$

where

$$A_j = \int_0^l \Phi_j^2 \, dx.$$

We mention that the above basis ϕ_1, ϕ_2 is unique up to the obvious changes

$$(\phi_1, \phi_2) \to \pm(\phi_1, \phi_2), \qquad \pm(\phi_2, -\phi_1)$$

if $A_2 \neq A_3$, i.e. if

$$\int_0^l (\phi_1^4 + \phi_2^4) \, dx \neq 6 \int_0^l \phi_1^2 \phi_2^2 \, dx.$$

If this condition is violated the basis is fixed only up to a rotation.

The representation (4.4) makes it clear that for $\hat{q} = \xi_2 \Phi_2 + \xi_3 \Phi_3 \neq 0$ the maximum of $\Delta(\lambda, q + \varepsilon \hat{q})$ near λ^* is increased.

§5. Construction of the example

We normalize the basic period to $l_0 = \pi$ and consider an arbitrary continuous function $q_0(x)$ of period l_0. In any neighborhood of q_0 we will construct a limit periodic function $q(x)$ with a Fourier series

$$q(x) = a_0 + \sum_{j,s=1}^{\infty} (a_{js} \cos s2^{-j}x + b_{js} \sin s2^{-j}x)$$

for which the rotation number $\alpha(\lambda, q) = \alpha(\lambda)$ takes on every rational number of the form $s \cdot 2^{-t}$, $s \geqq 0$, $t \geqq 0$ on an interval of positive length. More precisely, we will prove the

THEOREM. *Given $\eta > 0$; $q_0 \in C(R)$ of period π there exists a limit periodic continuous function q with basic frequencies 2^{-j} $(j = 0, 1 \cdots)$ in $\|q - q_0\| < \eta$ for which* (i) *the rotation number $\alpha(\lambda) = \alpha(\lambda, q)$ exists and is continuous,* (ii) *$\alpha(\lambda)$ takes on each rational numbers of the form $s \cdot 2^{-t}$, $s, t \geqq 0$ on an interval of positive length and* (iii) *the union*

$$E = \bigcup_{s,t \geqq 0} \operatorname{int} \alpha^{-1}(s \cdot 2^{-t})$$

of these intervals is dense on R. E is contained in the complement of the spectrum $\sigma(L)$ of $L = -D^2 + q$; hence $\sigma(L)$ is nowhere dense.

Remark. According to some recent result by R. Johnson the rotation number exists and is continuous for every almost periodic continuous function $q(x)$, so that (i) follows. For completeness we give a proof of (i) on our simpler situation.

Proof. (a) We write the rational numbers of the form $s \cdot 2^{-t} s \geqq 1$, $t \geqq 0$ as a sequence r_j. We choose the labelling in the following way: $r_1 = \frac{1}{2}$, $r_2 = 1$, then $r_3, r_4, \ldots, r_8$ denote all rationals $\leqq 2$ with denominator $2^2 = 4$ not listed before, and more general, $r_{\nu(N-1)+1} < \cdots < r_{\nu(N)} = N$ denote all rationals of denominator 2^N in $0 < r \leqq N$ not listed before. Thus $r_1, r_2, \ldots, r_{\nu(N)}$ is a list of all positive rationals with denominators $1, 2, 2^2, \ldots, 2^N$ in the interval $0 < r \leqq N$. One verifies that $\nu(N) = N \cdot 2^N$.

We set also $l_j = \pi 2^N$ for $\nu(N-1) + 1 \leqq j \leqq \nu(N)$, $N \geqq 0$ so that

$$r_j = \frac{s_j}{l_j} \pi = \frac{s_j}{2^N} \tag{5.1}$$

with some integer s_j.

We will construct $q(x)$ as the limit of continuous functions $q_\nu(x)$, $\nu \geqq 0$ of period l_ν for which the rotation number $\alpha_\nu(\lambda) = \alpha(\lambda, q_\nu)$ takes on the value r_ν on an interval of positive length.

In the following we will constantly use the fact mentioned already above, that for any periodic function p the rotation number $\alpha(\lambda, p)$ depends continuously on λ, and also on p in the class of continuous functions of fixed period. Moreover, the spectrum of $-D^2 + p$ agrees with the set of λ where $\alpha(\lambda, p)$ is strictly increasing. These facts, not obvious, for a limit periodic function will be forced for the function q by the construction of q_ν for which $q = \lim_{\nu \to \infty} q_\nu$.

(b) Construction of q_ν: We assume that $q_j(x)$ of period l_j $j=1,2,\ldots,\nu$ are constructed already so that $\alpha_j(\lambda)=\alpha(\lambda, q_j)$ takes on the values $r_1, r_2, \ldots, r_j$ on intervals of positive length. Moreover, $\varepsilon_\nu>0$ is any positive number which may depend on $q_0, q_1, \ldots, q_\nu$, and $\eta>0$; then we construct a continuous function $q_{\nu+1}(x)$ of period $l_{\nu+1}$ such that

$$\|q_{\nu+1}-q_\nu\|<\varepsilon_\nu \tag{5.2}$$

and such that $\alpha_{\nu+1}(\lambda)=\alpha(\lambda, q_{\nu+1})$ takes on $r_1, r_2, \ldots, r_{\nu+1}$ in intervals of positive length.

Since $\alpha_\nu(\lambda)$ is a continuous monotone increasing function taking on all positive values there exists a $\lambda^*=\lambda_\nu^*$ such that

$$\alpha_\nu(\lambda^*)=r_{\nu+1}. \tag{5.3}$$

If α_ν takes this value on an interval of positive length, then we set $q_{\nu+1}=q_\nu$. If, however, λ^* is uniquely determined by (5.3) then the discriminant $\Delta(Q)$ (with respect to the period $l_{\nu+1}$) has a critical point at $Q=q_\nu-\lambda^*$. In particular, this will be the case if $l_{\nu+1}>l_\nu$. According to §3 all solutions of $y''=(q_\nu(x)-\lambda^*)y$ satisfy (3.6) with $l=l_{\nu+1}$. Therefore, if ϕ_1, ϕ_2 is a basis with Wronskian 1, then the functions Φ_1, Φ_2, Φ_3 defined by (4.2) have period $l_{\nu+1}$. If Φ_1, Φ_2 are normalized so that the Φ_j are orthogonal then we set

$$q_{\nu+1}=q_\nu+\varepsilon\Phi_3. \tag{5.4}$$

Then $\Delta(\lambda, q_{\nu+1})$ has the value $r_{\nu+1}$ are a noncritical value, if $\varepsilon>0$ is near zero. It also has the period $l_{\nu+1}$ since l_ν divides $l_{\nu+1}$. Moreover, if ε is chosen small enough the noncritical values $r_1, r_2, \ldots, r_\nu$ are also noncritical for $q_{\nu+1}$ and (5.2) will be satisfied. This completes the construction of the sequence of periodic potentials q_ν of period l_ν.

(c) Smallness condition: Now we will select the numbers ε_ν recursively so that

(i) q_ν converges uniformly to a limit periodic function q in $\|q-q_0\|<\eta$.

(ii) $\alpha_\nu(\lambda)=\alpha(\lambda, q_\nu)$ converges uniformly on compact intervals to a continuous function $\alpha(\lambda)$, having the properties given in the theorem.

(iii) The rotation number $\alpha(\lambda, q)$ exists and agrees with $\alpha(\lambda)$.

(iv) The interior of the intervals of constancy of $\alpha(\lambda)$ form a dense set E on the real axis.

(v) The spectrum of $-D^2+q$ is contained in the complement of E, hence is nowhere dense.

In the following we will prove these five claims which imply the above theorem.

In order to define the ε_ν we introduce the resolvents

$$R_\nu(\lambda)=(-D^2+q_\nu-\lambda)^{-1}; \qquad R(\lambda)=(-D^2+q-\lambda)^{-1}. \tag{5.5}$$

By construction the interval $I_{j,\nu}=\alpha_\nu^{-1}(r_j)$ is for $j=0,1,2,\ldots,\nu$ an interval of positive length whose interior belongs to the resolvent set of $-D^2+q_\nu$. For $j\geqq 1$ these intervals are bounded and for $j=0$ it extends to $-\infty$. Let $\tau_\nu>0$ be a monotone decreasing sequence tending to 0 and pick for $j\geqq 1$ a closed subinterval $I'_{j,\nu}\subset \operatorname{int} I_{j,\nu}$, with the same center as $I_{j,\nu}$ and so that the measure

$$m(I_{j,\nu}-I'_{j,\nu})<2\tau_\nu. \tag{5.6}$$

For $j=0$ we take $I'_{0,\nu}$ as half infinite interval with

$$m(I_{0,\nu}-I'_{0,\nu})<\tau_\nu.$$

Then $R_\nu(\lambda)$ is bounded in $I'_{j,\nu}$ and we set

$$B_\nu=\max_{0\leqq j\leqq \nu}\sup_{\lambda\in I'_{j,\nu}}|R_\nu(\lambda)|.$$

Let $\eta>0$ be the number given in the theorem and δ_ν a positive sequence satisfying

$$\sum_{\nu=1}^{\infty}\delta_\nu=1.$$

Then we set $\varepsilon_0=\eta\delta_1$ and for $\nu\geqq 1$

$$\varepsilon_\nu=\min\left\{\eta\delta_{\nu+1},\frac{\delta_\nu}{2B_1},\frac{\delta_{\nu-1}}{2B_2},\ldots,\frac{\delta_1}{2B_\nu}\right\}. \tag{5.7}$$

Then from (5.2), (5.7) we have, in particular,

$$\|q_{\nu+1}-q_\nu\|<\eta\delta_{\nu+1}.$$

Therefore the sequence q_ν converges uniformly to a limit periodic function $q(x)$ with

$$\|q-q_0\|<\eta\sum_{\nu=1}^{\infty}\delta_\nu=\eta$$

verifying (i) of our claim.

From the other restrictions of (5.7) follows that for any fixed ν and $p = 0, 1, \ldots$

$$\|q_{\nu+p+1} - q_{\nu+p}\| \leqq \frac{\delta_{p+1}}{2B_\nu}$$

and hence

$$\|q_{\nu+p} - q_\nu\| \leqq \frac{1}{2B_\nu}(\delta_1 + \delta_2 + \cdots) = \frac{1}{2B_\nu}. \tag{5.8}$$

From this we conclude that $I'_{j,\nu}$ belongs to the resolvent set of $-D^2 + q_{\nu+p}$ for any $p \geqq 0$. Indeed, from the resolvent identity

$$R_{\nu+p}(\lambda) - R_\nu(\lambda) = R_{\nu+p}(q_\nu - q_{\nu+p})R_\nu$$

we obtain the estimate

$$|R_{\nu+p}(\lambda)| \leqq |R_\nu(\lambda)|\,(1 - \|q_{\nu+p} - q_\nu\|\,|R_\nu(\lambda)|)^{-1}.$$

For $\lambda \in I'_{j,\nu}$ we have by (5.8)

$$\|q_{\nu+p} - q_\nu\|\,|R_\nu(\lambda)| \leqq \frac{1}{2B_\nu}|R_\nu(\lambda)| \leqq \tfrac{1}{2}$$

and hence

$$|R_{\nu+p}(\lambda)| \leqq 2B_\nu < \infty$$

which proves that $I'_{j,\nu}$ belongs to the resolvent set of $-D^2 + q_{\nu+p}$.

The same holds true for any potential

$$q^t = tq_\nu + (1-t)q_{\nu+p}, \qquad 0 \leqq t \leqq 1,$$

and hence the value of its rotation number is rational for $\lambda \in I'_{j,\nu}$. Because of the continuous dependence the value of $\alpha(\lambda, q^t)$ is independent of t, and hence

$$\alpha_{\nu+p}(\lambda) = \alpha_\nu(\lambda) = r_j \quad \text{for} \quad \lambda \in I'_{j\nu} \text{ and } j \leqq \nu \leqq \nu + p.$$

Therefore

$$I'_{j,\nu} \subset I_{j,\nu+p} \quad \text{for} \quad j \leqq \nu \leqq \nu + p. \tag{5.9}$$

In particular, if λ_j is the midpoint of $I_{j,j}$ we have $\lambda_j \in I'_{j,j}$ and

$$\alpha_\nu(\lambda_j) = \alpha_j(\lambda_j) = r_j \quad \text{for} \quad \nu \geqq j. \tag{5.10}$$

This proves that for $\lambda = \lambda_j$ the sequence of monotone functions $\alpha_\nu(\lambda)$ converges for $\nu \to \infty$. We claim that it converges for all real λ. Indeed, given any λ, $\varepsilon > 0$ we can find λ_j, λ_k so that

$$\lambda_j < \lambda < \lambda_k, \qquad r_j = \alpha_j(\lambda_j) \leqq \alpha_\nu(\lambda) \leqq \alpha_\nu(\lambda_k) = r_k < r_j + \varepsilon$$

for all $\nu \geqq j, k$. This is possible since all numbers r_j – which form a dense set – are taken on by $\alpha_\nu(\lambda)$. Hence for any $p > 0$

$$|\alpha_\nu(\lambda) - \alpha_{\nu+p}(\lambda')| < 2\varepsilon \quad \text{if} \quad \lambda_j < \lambda,\ \lambda' < \lambda_k.$$

In particular, for $\lambda' = \lambda$ we see that $\alpha_\nu(\lambda) \to \alpha(\lambda)$ and

$$|\alpha(\lambda) - \alpha(\lambda')| < 2\varepsilon \quad \text{for} \quad \lambda_j < \lambda,\ \lambda' < \lambda_k.$$

This shows that $\alpha_\nu(\lambda)$ converges to a continuous monotone increasing function $\alpha(\lambda)$, taking on all values $\geqq 0$. Moreover, by (5.10),

$$\alpha(\lambda_j) = r_j.$$

Similarly, we see that the convergence $\alpha_\nu \to \alpha$ is uniform on compact sets. Indeed, otherwise there would exist a sequence $\lambda^{(n)} \to \lambda$ for which $|\alpha_{\nu_n}(\lambda^{(n)}) - \alpha(\lambda)| > \delta > 0$. But if $\varepsilon = \frac{1}{2}\delta$ and if we construct λ_j, λ_k as before such that $\lambda_j < \lambda < \lambda_k$, $0 < r_k < r_j + \varepsilon$ then for any sufficiently large n we have $\lambda_j < \lambda^{(n)} < \lambda_k$ and hence

$$|\alpha_{\nu_n}(\lambda^n) - \alpha(\lambda)| < 2\varepsilon = \delta,$$

which gives the desired contradiction.

Next we show that $\alpha(\lambda)$ is indeed the rotation number for $y'' = (q(x) - \lambda)y$: Let

$$\alpha^* = \pi \overline{\lim_{x \to \infty}} \frac{N(x, \lambda)}{x}, \qquad \alpha_* = \pi \underline{\lim_{x \to \infty}} \frac{N(x, \lambda)}{x}$$

for any nontrivial solution $y(x)$ of our differential equation, say one with $y(0) = 0$, $y'(0) = 1$. Set

$$\eta_\nu = \|q - q_\nu\|$$

and

$$\underline{q_\nu} = q_\nu - \eta_\nu, \qquad \overline{q_\nu} = q_\nu + \eta_\nu$$

so that

$$\underline{q_\nu}(x) \leqq q(x) \leqq \overline{q_\nu}(x)$$

and $\underline{q_\nu}, \overline{q_\nu}$ are periodic functions of period l_ν. By the comparison argument we have for the number of zeros $N(x, \lambda, p)$ of a nontrivial solution of $y'' = (p - \lambda)y$ with $y(0) = 0$

$$N(x, \lambda, \underline{q_\nu}) \geqq N(x, \lambda, q) \geqq N(x, \lambda, \overline{q_\nu})$$

and hence

$$\alpha_\nu(\lambda + \eta_\nu) = \alpha(\lambda + \eta_\nu, q_\nu) = \alpha(\lambda, \underline{q_\nu}) \geqq \alpha^*(\lambda) \geqq \alpha_*(\lambda) \geqq \alpha(\lambda, \overline{q_\nu}) = \alpha(\lambda - \eta_\nu, q_\nu) = \alpha_\nu(\lambda - \eta_\nu).$$

Because of the uniform convergence on compact sets of $\alpha_\nu \to \alpha$ we conclude that

$$\alpha(\lambda) \geqq \alpha^*(\lambda) \geqq \alpha_*(\lambda) \geqq \alpha(\lambda),$$

i.e. we have equality in all places. This proves (iii) of the claim.

Next we show that the set

$$E = \bigcup_{j=0}^{\infty} \operatorname{int} \alpha^{-1}(r_j) \tag{5.11}$$

is dense on the real axis. For this purpose we make use of the special labelling of the rational numbers r_j, which was chosen so that $r_1, r_2, \ldots, r_{\nu(N)}$ is a complete list of all rationals with denominators 2^N in the half open interval $(0, N]$. We set

$$Q_N(x) = q_{\nu(N)}(x), \qquad A_N(\lambda) = \alpha(\lambda, Q_N)$$

and analogue to $I_{j,\nu}$ we define the intervals

$$J_{j,N} = A_N^{-1}(r_j).$$

We pick a closed subinterval $J'_{j,N} \subset \operatorname{int} J_{j,N}$, $j \geqq 1$ with the same center as $J_{j,N}$ and

so that the measure

$$m(J_{j,N} - J'_{j,N}) < 2t_N \tag{5.12}$$

where

$$t_N = \tau_{\nu(N)}.$$

The half interval $J'_{0,N}$ is defined similarly as $I'_{0,\nu}$. We show more than (5.11) if we establish that the set

$$E' = \bigcup_{N=1}^{\infty} \bigcup_{j=0}^{\nu(N)} J'_{j,N} \subset E$$

is dense.

For this purpose it is obviously sufficient to show that for any open interval $\Delta \subset R^1$

$$\Delta \cap \bigcup_{j=0}^{\nu(N)} J'_{j,N} \neq \varnothing \tag{5.13}$$

if N is large enough.

We chose a compact interval $[\Lambda_1, \Lambda_2] \supset \Delta$ with $\Lambda_2 > 0$, $\Lambda_1 < \min_x (q_0(x) - \eta)$ so that $A_N(\lambda) = 0$ for $\lambda \leqq \Lambda_1$. Moreover, we pick N so large that

$$A_N(\lambda) \geqq r_{\nu(N)} = N \quad \text{for} \quad \lambda \geqq \Lambda_2. \tag{5.14}$$

This is possible, since by (2.2), for $\lambda > 0$

$$\frac{d\theta}{dx} \leqq 1 + \|Q_N\| + \lambda < 1 + \|q_0\| + \eta + \lambda$$

hence

$$A_N(\lambda) \leqq 1 + \|q_0\| + \eta + \lambda$$

and (5.14) follows if we choose

$$N \geqq 1 + \|q_0\| + \eta + \Lambda_2.$$

By construction A_N takes on all rationals $r_1, r_2, \ldots, r_{\nu(N)}$ on intervals of positive length, i.e. none of those intervals in $[\Lambda_1, \Lambda_2]$ collapse to a point.

We estimate the distance between two neighboring instability intervals

$$[\xi_N, \eta_N] = J_N = A_N^{-1}(r), \qquad [\xi_N^+, \eta_N^+] = J_N^+ = A_N^{-1}(r^+) \tag{5.15}$$

where $r, r^+ = r + 2^{-N}$ are two neighboring numbers of the set $r_0, r_1, \ldots, r_{\nu(N)}$, so that

$$\xi_N < \eta_N < \xi_N^+ < \eta_N^+.$$

Then the distance between J_N and J_N^+ can be estimated by

$$0 < \xi_N^+ - \eta_N < 3N2^{-N}. \tag{5.16}$$

For the proof we recall that $A_N(\lambda)$ is strictly monotone increasing in $\eta_N < \lambda < \xi_N^+$ and by the inequality below (3.3) satisfies there the inequality

$$\frac{d}{d\lambda}(A_N^2) \geqq 1.$$

Hence, with $r = s2^{-N}$,

$$\xi_N^+ - \eta_N \leqq \int \frac{d}{d\lambda}(A_N^2)\, d\lambda = (r^+)^2 - r^2 = \frac{2s+1}{4^N} \leqq 3N2^{-N}$$

since $1 \leqq s \leqq N \cdot 2^N$, which proves our estimate (5.16). This shows that the gaps between adjacent instability intervals tends to zero. Because of the construction of the $J'_{j,N}$ the distance of these intervals intersecting $[\Lambda_1, \Lambda_2]$ will be at most $2t_N = 2\tau_{\nu(N)}$ bigger, i.e.

$$\leqq 3N \cdot 2^{-N} + 2t_N.$$

Since this sequence tends to zero this number will be smaller than the length of Δ if N is large enough. This proves (5.13), and hence the claim (iv).

Finally we observe that by construction the intervals $J'_{j,N}$, $J = 0, 1, \ldots, \nu(N)$ belong to the resolvent set of $L = -D^2 + q$. Therefore, since E' is dense, the spectrum of L is nowhere dense, and all our claims of (c) are proven.

The theorem at the beginning of this section is clearly a consequence of these claims.

Remark 1. We showed that the spectrum of $L = -D^2 + q$ is contained in the

complement of the set E, defined by (5.11). Actually this spectrum agrees with the complement of E, as we want to verify now.

For this purpose we first show that the intervals

$$J_N = A_N^{-1}(r) = [\xi_N, \eta_N]$$

tend for $N \to \infty$ to the interval

$$\alpha^{-1}(r) = [\xi, \eta]$$

for any $r = r_j$.

By our construction the interval J'_N belongs to the resolvent set of $-D^2 + q_M$ for every $M \geqq N$. This implies

$$A_M^{-1}(r) = J_M \supset J'_N = [\xi'_N, \eta'_N] \quad \text{for} \quad M \geqq N. \tag{5.17}$$

On the other hand, the right endpoint η_M of J_M stays below the interval $(J_N^+)'$ contained in

$$J_N^+ = A_N^{-1}(r^+), \qquad r^+ = r + 2^{-N}.$$

Using our estimate (5.16) for the gap between J_N and J_N^+ we find

$$\eta_M \leqq \xi_N^+ + t_N \leqq \eta_N + 3N2^{-N} + t_N$$

and because of (5.17)

$$\eta_M \geqq \eta'_N > \eta_N - t_N.$$

for all $M \geqq N$. These relations show that η_N is a Cauchy sequence, and let $\eta^* = \lim_{N\to\infty} \eta_N$. Similarly $\xi_N \to \xi^*$.

From the convergence of $A_N(\lambda)$ to $\alpha(\lambda)$ it follows that

$$[\xi^*, \eta^*] \subset [\xi, \eta] = \alpha^{-1}(r),$$

hence $\eta^* \leqq \eta$. If, however, $\eta^* < \eta$ then also

$$\xi_N^+ \leqq \eta_N + 3N2^{-N} < \eta$$

for large N which implies

$$r < r^+ = A_N(\xi_N^+) \leqq A_N(\eta).$$

This is a contradiction since the right hand side tends to $\alpha(\eta) = r$. Hence $\eta^* = \eta$ and similarly one proves $\xi^* = \xi$, i.e.

$$[\xi^*, \eta^*] = \alpha^{-1}(r).$$

Finally we show that any point $\zeta \notin E$ belongs to the spectrum of $L = -D^2 + q$. Since E is dense there exists a sequence $\zeta_\nu \in E$ converging to E. Moreover, we may take ζ_ν as the left endpoints ξ_{N_ν} of J_{N_ν} which belong to the continuous spectrum of $L_{N_\nu} = -D^2 + q_{N_\nu}$.

It is a standard result that the cluster spectrum of L_N converges to that of L since

$$|L - L_N| = \|q - q_N\| \to 0.$$

Hence ζ belongs to the spectrum $\sigma(L)$ of L, hence $E^c \subset \sigma(L)$. Since we know already that $\sigma(L) \subset E^c$ these two sets agree.

Remark 2. The spectrum $\sigma(L)$ is a Cantor set, in particular a nowhere dense, perfect set. It remains to show that every point $\lambda^* \in \sigma(L)$ is the cluster point of $\sigma(L)$. For this purpose we set $\alpha(\lambda^*) = a$ and choose a sequence $a_\nu \to a$, where a_ν are different irrational numbers. Then the $\alpha^{-1}(a_\nu) = \lambda_\nu$ belong to $\sigma(L)$ and, because of the continuity of $\alpha(\lambda)$, λ^* is the only cluster-point of $\{\lambda_\nu\}$. This proves our claim, since $\lambda_\nu \in E^c = \sigma(\lambda)$.

Remark 3. By the same argument one can construct such limit periodic potentials which are real analytic and not only continuous. One would pick q_0 real analytic and replace the norm $\| \ \|$ by

$$\|q\| = \sup_{|\mathrm{Im}\, x| < \rho} |q(x)|$$

for some fixed $\rho > 0$. The above argument goes through and yields a real analytic potential $q(x)$ with the above properties.

Remark 4. We mention without proof that the functional $\alpha = \alpha(q(x) - \lambda)$ for fixed period l is Hölder continuous in the sense that

$$|\alpha(q_1 - \lambda) - \alpha(q_2 - \lambda)| \leqq c \, \|q_1 - q_2\|^{1/2} \quad \text{if} \quad \|q_1 - q_2\| < 1$$

where $c = c(l, M)$ depends on l and a bound for

$$\|q_1\|, \|q_2\| \leqq M.$$

However, c grows exponentially with l, like $\exp 3lM$, and one can not hope to have Hölder continuity for limit periodic potentials.

Finally we remark that the existence and continuity of $\alpha(\lambda, q)$ can actually be established for any almost periodic function $q(x)$, as was shown by R. Johnson. Moreover, one can relate this rotation number to the spectral resolution. In particular, it is a constant in any interval contained in the resolvent set of $-D^2+q$. In our case $2\alpha(\lambda, q)$ takes the values

$$\frac{s}{2^j}$$

in the resolvent set which are precisely the nonnegative number of the frequency mod 1. It is interesting that this phenomenon generalizes to quasi-periodic potentials, i.e. $2\alpha(\lambda, q)$ takes on values of the frequency module when λ is in the resolvent set. These and other properties of the rotation number will be published elsewhere.

REFERENCES

[1] BESICOVITCH, *Almost periodic functions*, Dover 1954, chap. I, §6.
[2] DUBROVIN, B. V., *The periodic problem for the Korteweg–de Vries equation in the class of finite-zone potentials*. Funkt. Anal. i. Prilozin, 9:3, 1975.
[3] DUBROVIN, B. A., MATVEEV, V. B. and NOVIKOV, S. P., *Non-linear Equations of Korteweg–de Vries Type, finite zone linear operators and Abelian Varieties*, Russian Math. Surveys 31: *1* (1976) 59–146.
[4] DYNABURG, E. I. and SINAI, Ja. G., *On the one dimensional Schrödinger Equation with quasi-periodic potential*, Funkt. Anal. i. Prilozin, vol. *9(4)* (1975) 8–21.
[5] HERMAN, M. R., *Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations*. Inst. des Hautes Etudes Scientifiques, Publ. Math. n. *49* (1979) Chap. III, 30 ff.
[6] INCE, E. L., *Ordinary Differential Equations*. Dover 1956, esp. sect. 7.41 and 15.62.
[7] MAGNUS, W. and WINKLER, S., *Hill's Equation*. Dover 1979, esp. sect. 7.
[8] RUESSMANN, H., *On the one dimensional Schrödinger equation with a quasi-periodic potential*. To appear in the Annals of the New York Academy of Sciences, 1980.
[9] MCKEAN, H. P. and VAN MOERBEKE, P., *The spectrum of Hill's equation*, Inventiones Math., *30* (1975) 217–274.
[10] SCHARF, G., *Fastperiodische Potentiale*, Helvetica Acta 38, Heft *6* (1965) 573–605.

Mathematik, ETH Zentrum 8092 *Zürich.*

Received Febr. 10, 1981

Comment. Math. Helvetici **56** (1981) 225–239 0010–2571/81/002225–15$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

The equivariant Dehn's lemma and loop theorem

WILLIAM H. MEEKS III and SHING-TUNG YAU

Introduction

In [4] the authors observed that the topological methods in the theory of three-dimensional manifolds can be modified to settle some old problems in the classical theory of minimal surfaces in euclidean space (see also [1], [12]). In [4] and [5] we found that we could use the theory of minimal surfaces to extend the theorems of Papakriakopoulous, Whitehead and Shapiro, Stalling and Epstein on the Dehn's lemma, loop theorem and sphere theorem. The key point to our approach to these topological theorems is the following: Given a certain family of maps of the disk or sphere into our three-dimensional manifold M, we minimize the area of the maps (with respect to the pulled back metric) in this family and prove the existence of the minimal map. Then by using the area minimizing property of the map and the tower construction in topology, we prove that any area minimizing map in the family is an embedding. In this way, we realize the solutions to the above topological theorems by minimal surfaces. In [4] and [5] we used the above area minimizing solutions to prove equivariant versions of the loop and the sphere theorem, and we applied these new theorems to the classification of compact group actions on $\mathbf{R}^3$ in [11].

In this paper we generalize some of the theorems in [4] and [5] to compact planar domains by proving the existence of embedded planar domains of least area of a given genus and by proving a certain disjointness property for planar domains of least area. We then use this disjointness property to prove the equivariant Dehn's lemma for planar domains.

On the other hand, we use a different variation approach to get a geodesic version of the loop theorem. More precisely, we prove the following: suppose that the induced map $i_*: \pi_1(\partial M) \to \pi_1(M)$ of the inclusion of the boundary has nontrivial kernel K. Then for any metric on ∂M, any nontrivial geodesic of least length in K is embedded and any two such geodesics are equal or disjoint. This geodesic loop theorem coupled with the above equivariant Dehn's lemma yields a new version of the equivariant loop theorem in [5]. As the placement of curves on a surface is easier to understand this new equivariant loop theorem is easier to

apply to study group actions. Applications of this theorem to classification of group actions on $\mathbf{R}^3$ will appear in [11].

Throughout this paper we will be working with compact three-dimensional Riemannian manifolds M with convex boundary. For simplicity we sometimes refer to such an M as a *convex* manifold.

1. Dehn's lemma for planar domains

THEOREM 1 (Dehn's lemma for planar domains of a given genus). *Let $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ be a collection of disjoint unoriented Jordan curves on the boundary of a three-dimensional orientable convex manifold M. Suppose that these Jordan curves bound a continuous mapping g from a smooth compact planar domain (possibly disconnected). Let F_k be the family of all piecewise smooth maps mapping from a compact planar domain with k components into M whose boundary consists of curves in Γ. Let A_k be the infimum of the areas of the maps in F_k. If A_k is strictly less than A_{k+1}, then there exists a branched minimal immersion which has least area among maps in F_k. Furthermore, any branched minimal immersion of least area in F_k is an embedding.*

Proof. The existence of a map $f: \Omega \to M$ of planar domain with k components and least area follows from the inequality $A_k < A_{k+1}$, from Morrey [7] and from Theorem 1 in [4]. From the approximation technique in the proof of Theorem 5 in [4], we may assume that the map f is a simplicial immersion with respect to some triangulations of Ω and M.

Since $f: \Omega \to M$ is a map of least area for a given genus, f restricted to each component Ω' of Ω is a map of least area from a planar domain with boundary curves $f(\partial\Omega')$. By Theorem 5 in [4], $f \mid \Omega'$ is an embedding. Suppose that there are two distinct components Ω_1, Ω_2 of Ω such that $f(\Omega_1)$ and $f(\Omega_2)$ intersect. In this case it is shown in [4] that there are Jordan curves $\gamma_1: S^1 \to \mathring{\Omega}_1$ and $\gamma_2: S^1 \to \mathring{\Omega}_2$ such that $f(\gamma_1(t)) = f(\gamma_2(t))$. The standard cutting and gluing argument (see the end of the Proof of Theorem 5 in [4]) along the image curve $f(\gamma_1) = f(\gamma_2)$ produces a map of a planar surface with the same Euler characteristic as Ω and with the same area as f. However, the area of the new map can be decreased along the folding curve $f(\gamma_1)$. Since the Euler characteristic of a planar domain with n boundary curves determines the genus and the number of components, the existence of the new map contradicts the least area property for f. This contradiction proves Theorem 1.

In [4] the authors also proved a disjointness property for least area disks when Γ in the above theorem consists of one curve γ. In that paper we prove that any two geometrically distinct least area disks intersect only along their boundary. This

disjointness property for least area disks is useful in proving equivariant group action theorems. For this reason we would like to generalize the disjointness property to the case of planar domains given in the above theorem. However, in the following example two Jordan curves in parallel planes in $\mathbf{R}^2$ are given which bound two distinct embedded annuli of least area that intersect their interiors.

EXAMPLE. Let δ_{-1000} be a circle of radius 10 in the xy plane centered at the point $(0, -1000, 0)$ and let δ_{1000} be a circle of radius 10 in the xy plane centered at the point $(0, 1000, 0)$. Let γ_1 be the connected sum of δ_{-1000} and δ_{1000} along part of the interval I joining $(0, -1000, 0)$ to $(0, 1000, 0)$ in such a way that γ_1 is the union of parts of δ_{-1000}, δ_{1000}, and the intervals $I+(-1, 0, 0)$ and $I+(1, 0, 0)$. Let $\gamma_2 = \gamma_1 + (0, 0, 1)$ be the curve on the plane of distance one from the xy plane. A least area annulus $f: \Omega \to \mathbf{R}^3$ connecting γ_1 and γ_2 appears as in Figure 1. Let $R: \mathbf{R}^3 \to \mathbf{R}^3$ be rotation by 180 degrees around the z-axis. Then the least area annuli $f(\Omega)$ and $R \circ f(\Omega)$ intersect in their interiors. (A rigorous proof of the existence of Ω produced in this example can be found by using the bridge theorem in [6].)

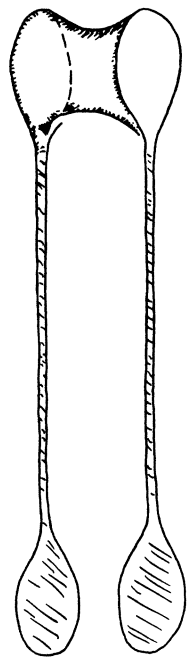


Figure 1.

In spite of this example, the disjointness property holds when the following assumptions on Γ hold.

THEOREM 2. *Let $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ be a collection of disjoint unoriented Jordan curves on the boundary of a compact three-dimensional orientable convex manifold M. Suppose that γ_1 is homotopically nontrivial when n equals two or that Γ generates a rank $(n-1)$ subgroup of the first homology group of M. If there exists a continuous map g of a compact planar domain into M with boundary Γ, then*

(1) *there exists a branched minimal immersion of a compact planar domain which bounds Γ and has (finite) least area among all such maps.*
(2) *Every such map is an embedding of a connected planar domain.*
(3) *Any two such least area maps intersect only along their boundary Γ or else they differ by a conformal reparametrization.*

Proof. Part (1) is just the statement of Theorem 5 in [4]. Part (2) follows because the condition that the curves in Γ represent $n-1$ independent homology classes implies the connectedness of the surface. The proof of part (3) is based on the proof of Theorem 6 in [4]. The nontrivial approximation procedure in Theorem 6 in [4] reduces part (3) to the special case that the two least area maps $f : \Omega_1 \to M$ and $g : \Omega_2 \to M$ are simplicial with respect to some fixed triangulations of Ω_1, Ω_2 and M.

Suppose now that $X = f_1(\Omega_1) \cap f_2(\Omega_2)$ is not equal to the union of Γ. In [4] it is shown that X is a finite one-dimensional subcomplex of M with every vertex in X meeting at least two edges in X and the intersection of $f(\Omega_1)$ and $f(\Omega_2)$ is traverse except possibly at the vertices. A simple induction argument (see Lemma 10 in [4]) proves that X contains a closed Jordan curve α which is not contained in the union of Γ or for some i and $k > 0$ there is a unique Jordan arc $\sigma : [0, 1] \to X$ with $\sigma([0, 1]) \cap \Gamma = \{\sigma(0), \sigma(1)\}$ and $\sigma(0) \in \gamma_i$ and $\sigma(1) \in \gamma_{i+k}$.

Suppose that σ exists. By the classification of compact planar surfaces, there would be a smooth Jordan curve τ in the interior of Ω_1 such that $\tau \cap X = \tau \cap \sigma$ is one point which is not vertex and the intersection of τ and σ on Ω_1 is transverse at this point. As τ intersects Ω_2 transversely in one point, $[\tau] \cap [\Omega_2]$ is nonzero where $\cap$ denotes the intersection pairing on homology in M with Z_2-coefficients. However, as Ω_1 is a compact planar domain, τ is homologous with Z_2-coefficients to some sum of boundary curves of Ω_1. As the boundary curves of Ω_1 and Ω_2 are the same, some boundary curve γ_i of Ω_2 must intersect Ω_2 nontrivially in homology. However, M is orientable and therefore we can push γ_i off Ω_2 to create a curve γ_i' which is disjoint from Ω_2. This curve is homologous to γ_i but

does not intersect Ω_2. This contradicts the intersection equation on homology and therefore σ can not exist. Hence there must be a Jordan curve α in X which is not contained in Γ.

Let $\alpha_1 : S^1 \to \Omega_1$ and $\alpha_2 : S^1 \to \Omega_2$ be the Jordan curves with $f(\alpha_1(t)) = g(\alpha_2(t))$ and $f(\alpha_1) = \alpha$. Suppose for the moment that α_1 and α_2 are contained in the interior of Ω_1 and Ω_2. The curve α_i disconnects Ω_i into two planar domains Ω_i', Ω_i'' where Ω_i' is the planar domain containing the Jordan curve γ_1.

Now consider the surface Σ obtained by gluing $f(\Omega_1'')$ and $f(\Omega_2'')$ along α. If Σ has a nonempty boundary, then for some i different from 1, an oriented boundary curve γ_j of Σ is homologous in Σ to a collection of curves in $\{\pm\gamma_2, \pm\gamma_3, \ldots, \pm\hat{\gamma}_j, \ldots, \pm\gamma_n\}$ where for the moment the curves Γ are oriented in an arbitrary manner. Therefore the curves $\Gamma - \{\gamma_1, \gamma_j\}$ generate a subgroup of $H_1(M, Z)$ with the same rank as Γ which is $n-1$. For $n \geq 3$, this contradicts our assumptions. If $n = 2$, then X is a disk and so γ_2 is homotopically trivial. This also contradicts our assumptions and so Σ must have no boundary.

As Σ has no boundary, the surfaces $f(\Omega_1')$ and $f(\Omega_2')$ have the same boundary curves. The usual cutting and gluing argument shows that f or g does not have least area and hence part (3) is valid if the Jordan curve α_1 lies in the interior of Ω_1. Actually the only reason that we chose the case "α_1 lies in the interior of Ω_1" was to make visualization of the intersection easier. The same argument still produces a contradiction when part of α intersects the union of the curves in Γ. This proves part (3) and completes the proof of the theorem.

Remark. Theorem 2 can be proved by assuming appropriate conditions about areas of planar domains which bound some subcollection of curves in Γ rather than topological conditions. For example, suppose that $\Gamma = \{\gamma_1, \gamma_2\}$ and that either γ_1 or γ_2 does not bound a disk with area less than twice the area of some annular region joining them. Then the planar domain of least area joining γ_1 and γ_2 will be an embedded annulus and any two such annular surfaces intersect only along their boundary curves. Note that this area condition fails for the example described before Theorem 2.

2. Embedding of the partially free boundary value problem

Another type of embedding theorem that can be proved using the topological tower construction is the partially free boundary value problem considered in Courant's book [2]. In its simplest topological form the partially free boundary value problem can be stated as follows. Let M be a compact three-dimensional Riemannian manifold and γ_1 be a Jordan curve on a boundary component ∂_1 of M which is freely homotopic to a closed curve γ_2 on another component ∂_2 of the

boundary of M. Let F be the family of all maps from the annulus Ω to M which maps one boundary curve of Ω homeomorphically onto γ_1 and the other boundary curve of Ω into ∂_2. Then we say that a minimal immersion $f:\Omega \to M$ is a solution to the partially free boundary value problem for γ_1 and ∂_2 if $f \in F$ and f has least area in F.

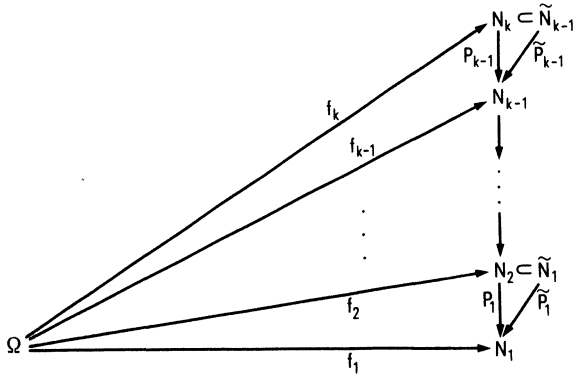
THEOREM 3. *Suppose M is a compact orientable Riemannian three-dimensional manifold with convex boundary, γ_1 is a Jordan curve on a component ∂_1 of the boundary of M and γ_1 is freely homotopic to a curve on a different component ∂_2 of the boundary. Then*

(1) *There exists a solution $f:\Omega \to M$ to the partially free boundary value problem for γ_1 and ∂_2 if the infimum of areas of maps in F is strictly less than the area of any map of a disk with boundary γ_1. Furthermore f is continuous in Ω and smooth in the interior of Ω.*

(2) *Any such solution f is one-to-one and everywhere orthogonal to ∂_2.*

Proof. The existence of a solution to the partially free boundary value problem can be proved using the methods in the proof of the free boundary value problem in [5].

After conformal reparametrization we may assume that Ω is a circular domain where the inner circle is the unit circle S^1 and $f(S^1) = \gamma_1$. From the approximation arguments in [4] and [5] we may assume that the map f is simplicial with respect to some triangulations of Ω and M. Therefore the image surface $f(\Omega)$ has a regular neighborhood N_1 in M. After restricting the range space of f to N_1, there is a new map $f_1:\Omega \to N_1$. Let H be the subgroup of $H_1(N_1, Z_2)$ generated by $f_1(S^1)$. If H is not all of $H_1(N_1, Z_2)$, then there exists a surjective homomorphism $\rho: H_1(N_1, Z_2) \to Z_2$ with $\rho([f_1(S^1)]) = 0$. This homomorphism induces a surjective



homomorphism $\bar{\rho}: \pi_1(N_1) \to Z_2$. Since the kernel of $\bar{\rho}$ has index two in $\pi_1(N_1)$, there is a 2-sheeted covering space $\tilde{P}_1: \tilde{N}_1 \to N_1$ associated to this subgroup. Since the map $f: \Omega \to N_1$ satisfies $f_{1*}(\pi_1(\Omega)) \subset \tilde{P}_{1*}(\pi_1(\tilde{N}_1)) = \ker(\bar{\rho})$, the lifting theorem for covering spaces implies that f_1 lifts to a map $\tilde{f}_1: \Omega \to \tilde{N}_1$. After restricting the range of $\tilde{f}_1$ to a regular neighborhood N_2 of $\tilde{f}_2(\Omega)$, we get a new map $f_2: \Omega \to N_2$.

Repeating this construction k-times yields the tower below. As was discussed in [4] or [5], this construction terminates with a map $f_k: \Omega \to N_k$ with $f_k(S^1)$ generating $H_1(N_k, Z_2)$. Here P_i is the restriction of $\tilde{P}_i$ to N_{i+1} and each N_i is a Riemannian manifold with the pulled back Riemannian metric.

ASSERTION 1. *$f_k: \Omega \to N_k$ is one-to-one.*

Proof. As $H_1(N_k, Z_2)$ is generated by $f_k(S^1)$, $H_1(N_k, Z_2)$ is equal to the trivial group or the group Z_2. If $H_1(N_k, Z_2)$ is the trivial group, it is straightforward to check that the boundary of N_k consists entirely of spheres (see [4] for a proof). In this case $\gamma = f_k(S^1)$ lies on some sphere S in the boundary of N_k.

In [4] and [5] it is shown that there exists, after subdivision, a simplicial retraction $R: N_k \to f_k(\Omega)$ such that (1) $R \mid (\partial N_k - f_k(\partial\Omega))$ is locally one-to-one, and (2) $R \mid \partial N_k$ covers each 2-simplex of $f_x(\Omega)$ exactly two times.

The Jordan curve γ disconnects the sphere S into two disks D_1 and D_2. Computing areas, we have

$$\text{Area}(R \mid D_1) + \text{Area}(R \mid D_2) = \text{Area}(R \mid S) \leq \text{Area}(R \mid \partial N_k) \leq 2 \text{ Area}(f_k).$$

Hence either the area of $R \mid D_1$ or $R \mid D_2$ is not greater than Area (f_k) = Area (f_1). Therefore we may assume that the area of, say, $g = P_1 \circ P_2 \circ \cdots \circ P_{k-1} \circ R \mid D_1$ is not greater than the area of f. Furthermore, the area of g can be decreased along a folding curve which is a self-intersection curve of $f_k(\Omega)$ in the case $f_k(D)$ is not embedded (see Theorem 4 in [4] for a rigorous proof of this fact). This contradicts the original assumption that f is a solution to the partially free boundary value problem.

Thus we may assume that $H_1(N_k, Z_2)$ is Z_2. In this case it is easy to show that $f_k(\partial\Omega)$ is contained in a torus component T of the boundary of N_k (see the proof of Theorem 5 in [4]). Furthermore, as $H_1(N_k, Z_2)$ is generated by $f(S_1)$, the boundary curves of $f_k(\Omega)$ are disjoint and are nontrivial homology classes on ∂N_k. From the simple topology of curves on a torus we may conclude that $f_k(\partial\Omega)$ disconnect T into a collection of closed planar domains, two of which are annular regions A_1 and A_2 where the boundary of the annular region A_i consists of $f_k(S^1)$ and part of the other boundary curve of $f_k(\Omega)$.

Let $R: N_k \to f_k(\Omega)$ be the retraction discussed above. Then, as before, Area $(R \mid A_1)$ + Area $(R \mid A_2) \leq$ Area (f_k), and so we may assume that

Area $(R \mid A_1)$ is strictly less than Area (f_k). However, the boundary curves of $g = (P_1 \circ P_2 \circ \cdots \circ P_{k-1} \circ R) \mid A_1$ consists of γ_1 and a curve on the boundary component ∂_2. As Area $(g) <$ Area (f), we arrive at a contradiction which shows the map f_k is one-to-one and proves Assertion 1.

ASSERTION 2. *f_{k-1} is one-to-one.*

Proof. If f_{k-1} is not one-to-one, then the map f_{k-1} has singular points which are double points. As f_{k-1} is everywhere orthogonal to the boundary of N_{k-1}, the maximum principle or Lemma 5 in [4] implies that the image of the boundary component of Ω different from S^1 is not completely contained in the singular set $S(f_{k-1})$. The arguments in [4] show that there exists a Jordan curve $\alpha_1 : S^1 \to \Omega$ or a Jordan arc $\alpha_1 : [0, 1] \to \Omega$ with $\alpha(0)$, $\alpha(1) \in \partial\Omega$ which bounds with some part of $\partial\Omega$ a closed connected domain Ω_1 in Ω with $\Omega_1 \cap S(f_{k-1}) = \alpha_1$. Let α_2 be the double curve corresponding to α_1. By our choice of α_1, the Jordan curve α_2 will bound, with some parts of $\partial\Omega$, a closed subdomain Ω_2 of Ω whose interior is disjoint from Ω_1.

A cutting and gluing argument shows that we can interchange the region Ω_1 and Ω_2 to get a new continuous piecewise smooth map $g : \Omega \to f_{k-1}(\Omega)$ with the same area as $f_{k-1}(\Omega)$ and such that $G = P_1 \circ P_2 \circ \cdots \circ P_{k-2} \circ g$ is a candidate for a solution to the partially free boundary value problem. However, the area of G can be decreased along the folding curve α_1 which contradicts the least area property for f. This contradiction proves the assertion which in turn implies part (2) of the theorem.

Remarks. The previous theorem can be generalized in a number of interesting ways. For example, one can replace γ_1 by a collection $\Gamma_1 = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ of pairwise disjoint Jordan curves and γ_2 by $\Gamma_2 = \{\alpha_1, \ldots, \alpha_n\}$ a collection of curves which lies on distinct boundary components of ∂M different from the boundary components containing the Jordan curves in Γ_1. In this case we assume that there is a map of a planar domain into M whose boundary curves are $\Gamma_1 \cup \Gamma_2$. One can then pose a partially free boundary value problem and if there is a least area solution to this problem, one can prove that the solution is embedded. The proof of this fact can be shown using the techniques of proof given in Theorem 3 and in Theorem 5 of [4].

It is important to note that the existence of embedded solutions to other free boundary value problems can also be shown. For example, suppose we replace the condition that γ_1 and γ_2 lie on distinct components of the boundary of M, by the condition that γ_2 lies in the complement of some compact piece P of the boundary surface containing γ_1. Then if a solution to this free boundary value problem exists and the boundary of the map is disjoint from ∂P, then the solution

is an embedding. Such free boundary value problems occur naturally for, say, certain convex subsets of euclidean three space.

The solution to the free boundary value problem in [5] can be generalized to annular or even planar domains. For example, suppose that γ_1 is a loop on a boundary component ∂_1 of a convex M, which is homotopically nontrivial in M. Suppose γ_1 is homotopic to a loop γ_2 on a different boundary component ∂_2 of M. Then there exists an immersion $f:\Omega \to M$ of an annulus of least area with one boundary curve on ∂_1 and the other boundary curve on ∂_2 and so that the induced map on fundamental groups is nontrivial. Furthermore, f is as regular as the metric of M and any such f is one-to-one.

3. The equivariant Dehn's lemma

In [5] we proved the equivariant loop theorem by using the disjointness property of least area disks. The disjointness property in Theorem 2 for least area planar domains can also be used to prove the following equivariant theorem.

THEOREM 4 (Equivariant Dehn's lemma for planar domains). *Suppose* $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ *is a collection of smooth disjoint unoriented Jordan curves on the boundary of an orientable three-dimensional manifold* M. *Suppose either the curves in* Γ *generate a rank* $(n-1)$ *subgroup of the first homology group of* M *or* n *equals to two and the curve* γ_1 *is homotopically nontrivial in* M. *Suppose also that the collection* Γ *is the image of the boundary of a map of a compact planar domain into* M. *If* G *is a compact subgroup of* $\mathrm{Diff}^+(M)$ *which acts freely on the union of* Γ, *then* Γ *is the boundary of an embedded compact planar domain in* M *which is invariant under those elements of* G *that leave some* γ_i *in* Γ *invariant.*

Proof. As G is compact, we may assume that G acts on M as a group of isometries. Furthermore, it is elementary to construct an invariant metric on M with convex boundary by averaging the metric on ∂M and taking the product metric in a neighborhood of ∂M. We may also assume that M is compact by restricting the manifold to a regular neighborhood of the G orbits of the image of the map of the compact planar domain given in the hypothesis.

By Theorem 1, there exists a smooth embedded connected compact planar domain Ω of least area in M with boundary curves in Γ. By Theorem 2, any two such least area planar domains are either disjoint in the interior of M or equal.

Suppose now that $g:M \to M$ is an element of G which leaves invariant the Jordan curve γ_i and suppose $g(\Omega) \neq \Omega$. As g leaves γ_i invariant and has no fixed points on γ_i, g acts on a regular neighborhood of γ_i as a rotation. As $g(\Omega) \neq \Omega$ and $g(\Omega)$ is another planar domain of least area, $g(\Omega)$ is disjoint from Ω in the

interior of M. This implies that $g(\Omega)$ lies locally on one side of Ω. By convexity (see for example [4] or [6]), the surfaces Ω and $g(\Omega)$ are immersed and transverse to the boundary of M.

Let $J(\gamma_i'(t))$ denote the vector obtained by rotating the tangent vector $\gamma_i'(t)$ clockwise by 90 degrees in the tangent space of $T_{\gamma_i(t)}\partial M$ with respect to the induced orientation. Define $\alpha_\Omega(\gamma_i(t))$ and $\alpha_{g(\Omega)}(\gamma_i(t))$ as the oriented angle between the vector $J(\gamma_i'(t))$ and the tangent planes of the corresponding surfaces. After integrating along γ_i, we have

$$\theta_\Omega = \int_0^1 \alpha_\Omega(\gamma(t))\,dt \quad \text{and} \quad \theta_{g(\Omega)} = \int_0^1 \alpha_{g(\Omega)}(\gamma(t))\,dt.$$

As g acts as rotation on the regular neighborhood of γ_i, $\alpha_\Omega(g(\gamma(t))) = \alpha_{g(\Omega)}(\gamma(t))$ and hence $\theta_\Omega = \theta_{g(\Omega)}$. On the other hand, as Ω lies locally on one side of $g(\Omega)$, either $\alpha_\Omega(\gamma_i(t)) \le \alpha_{g(\Omega)}(\gamma_i(t))$ for all t or else $\alpha_\Omega(\gamma_i(t)) \ge \alpha_{g(\Omega)}(\gamma_i(t))$ for all t. As the integrals are the same, $\alpha_\Omega(\gamma_i(t)) = \gamma_{g(\Omega)}(\gamma_i(t))$. This shows that $g(\Omega)$ and Ω are everywhere tangential to each other along γ_1. Therefore, the maximum principle (or Lemma 5 in [4]) implies that Ω and $g(\Omega)$ intersect in an open set. Hence the disjointness property of Ω implies that $\Omega = g(\Omega)$. This completes the proof of the theorem.

THEOREM 5 (Equivariant Dehn's lemma for disks). *Suppose* $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$ *is a collection of disjoint Jordan curves on the boundary of an orientable three-dimensional manifold* M. *Suppose each* γ_i *is homotopically trivial in* M. *If* G *is a compact group acting on* M *as a group of orientation preserving diffeomorphism which acts freely on the union of* Γ, *then there exists a collection of embedded invariant disks* $\{D_1, D_2, \ldots, D_n\}$ *which are pairwise disjoint with* $\partial D_i = \gamma_i$ *and whose union is invariant under* G.

Proof. After picking an invariant metric, G acts as a group of isometries. As in the previous lemma, we can assume that this metric is convex and M is compact. Let D_1 be a disk of least area with boundary curve γ_1 and let $G \cdot D_1$ denote the union of the least area disks which are images of D_1 under G. By the argument given in the previous theorem, D_1 is the only disk in $G \cdot D_1$ whose boundary curve is γ_1. This implies that each of the curves in $G \cdot \gamma_1$ bound a unique disk in $G \cdot D_1$.

If $G \cdot \gamma_1$ is not all of Γ, then let D_2 be a disk of least area with a boundary curve in $\Gamma \setminus (G \cdot \gamma_1)$ and $G \cdot D_2$ be the union of the orbits of D_2 under the action of G. As before, these are embedded and disjoint. As the disks in $G_1 \cdot D_1$ and $G \cdot D_2$ can only intersect in their interiors and as they have least area, they do not intersect. This last fact is proved in [4] where we show that if two embedded

minimal disks intersect only in their interiors, then there is a closed Jordan curve in their intersection which bounds two least area disks. Then by a cutting and gluing argument we can decrease the area of one of the $G(D_i)$ which is impossible.

If $G \cdot \gamma_1 \cup G \cdot \gamma_2$ does not exhaust the curves in Γ, then we can find a new least area disk D_3 with boundary curve in $\Gamma - (G \cdot \gamma_1 \cup G \cdot \gamma_2)$. Let $G \cdot D_3$ be the orbits of D_3. Continuing this process eventually, we can produce the required disks $G \cdot D_1, G \cdot D_2, \ldots, G \cdot D_k$.

COROLLARY. *Suppose $\tau : M \to M$ is an orientation preserving diffeomorphism of a three-dimensional manifold M which is an isometry with respect to some metric on M. If τ leaves invariant a Jordan curve γ on the boundary of M which is homotopically trivial in M, then τ has a fixed point on M.*

Proof. Let G be the closure in $\mathrm{Diff}^+(M)$ of the cyclic subgroup generated by τ. As γ lies on the boundary of an orientable three-dimensional manifold, G restricts to an effective action on γ. Here G is either a finite cyclic group or S^1. By the previous theorem there is either a fixed point of τ on γ or else there is a disk in M which is invariant under τ. If τ has no fixed points on γ, then the Brower fixed point theorem implies that τ has a fixed point on the invariant disk. This proves the corollary.

4. The equivariant loop theorem

In this section we are going to prove the equivariant loop theorem by first proving a disjointness property of a certain generating set of closed geodesics on the boundary of the three-dimensional manifold and then applying the equivariant Dehn's lemma of Section 3. We begin with the following

DEFINITION. Let M be an n-dimensional compact Riemannian manifold and let H be a normal subgroup of $\pi_1(M)$. Then a collection $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n, \ldots\}$ of closed geodesics is said to be a *short generating set* for H if for each n, γ_n represents a closed curve in H of least length in the complement of the normal subgroup of H generated by the free homotopy classes $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_{n-1}\}$.

LEMMA 1. *Suppose γ_1 and γ_2 are embedded distinct closed geodesics on a boundary surface Σ of a three-dimensional Riemannian manifold M. If γ_1 and γ_2 intersect nontrivially and are homotopically trivial in M, then one of these geodesics, say γ_2, can be expressed as the product of two closed nongeodesic curves in $\gamma_1 \cup \gamma_2$,*

each with length less than or equal to length of γ_2, and these nongeodesic curves are homotopically trivial in M.

Proof. Since an embedded geodesic is determined by its tangent vector at a single point and the exponential map is a local diffeomorphism, it is easily seen that γ_1 and γ_2 intersect transversally in a finite number of points. Hence we may consider γ_1 and γ_2 as simplicial curves on Σ with respect to some triangulation of M. By Dehn's lemma there exist embedded piecewise linear disks D_1 and D_2 with boundary curves γ_1 and γ_2 respectively, which are in general position.

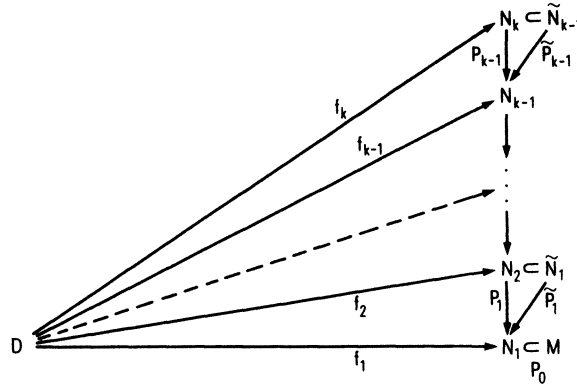
Since D_1 and D_2 are in general position, they intersect in a compact one-dimensional manifold with boundary. Let I be an interval component in $D_1 \cap D_2$. The interval I disconnects D_1 into two closed subdisks D_{11} and D_{12} and disconnects D_2 into two closed subdisks D_{21} and D_{22}. Let $\alpha_{ij} = D_{ij} \cap \Sigma$ and suppose that α_{11} is the shortest such arc. Then the length of the boundary of each of the disks $\tilde{D}_1 = D_{11} \cup_I D_{21}$ and $\tilde{D}_2 = D_{11} \cup_I D_{22}$ is less than or equal to the length of γ_2. Here $\cup_I$ means that we paste the disks along their common boundary arc I. On the other hand, γ_2 can be expressed as a product of $\partial\tilde{D}_1 \cdot \partial\tilde{D}_2 = (\alpha_{21}\alpha_{11}) \cdot (\alpha_{11}^{-1}\alpha_{22})$. Since $\partial\tilde{D}_1$ and $\partial\tilde{D}_2$ are not geodesics, $\partial\tilde{D}_1$ and $\partial\tilde{D}_2$ are the required closed curves. This completes the proof of the lemma.

THEOREM 6. *Suppose M is a compact orientable three-dimensional Riemannian manifold with a boundary component Σ. Let $K = \operatorname{Ker}(i_*)$ be the kernel of the map $i_*: \pi_1(\Sigma) \to \pi_1(M)$ induced by inclusion. Then with respect to any fixed metric on Σ, there exists a finite short generating set $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ for K. For any such generating set the geodesics in Γ are embedded. Furthermore, any two geodesics in the union of any two short generating sets are either equal or disjoint.*

Proof. We first show that there is a minimal generating set $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ for K consisting of embedded geodesics. Since there are only a finite number of free homotopy classes on a compact surface having length less than a given constant, we can choose a short generating set for K by sequentially picking the next free homotopy class of least length. To be precise, suppose by induction $\Gamma_{n-1} = \{\gamma_1, \gamma_2, \ldots, \gamma_{n-1}\}$ have been chosen. If Γ_{n-1} is not a short generating set, then we let γ_n be a closed geodesic of least length in the complement of the normal subgroup of K generated by Γ_{n-1}. We will now show that γ_n is embedded.

Since $\gamma_n: S^1 \to M$ is a closed geodesic, it is determined by its tangent vector at a point and its multiplicity which is the number of times it transverses the same path. As a geodesic of multiplicity one is always in general position with respect to itself, we may assume that $\gamma_n(S^1)$ is a simplicial curve with respect to some triangulation of M. Hence $\gamma_n: S^1 \to M$ is also simplicial with respect to the pulled back triangulation on S^1.

Let $f: D \to M$ be a branched simplicial immersion such that $f \mid \partial D = \gamma_n$. Then after restricting to a regular neighborhood N_1 of $f(D)$ in M, there is a new map $f_1: D \to N_1$. As in the proof of Dehn's lemma (see [4] or [1]), we can construct a tower where we may assume that the boundary of N_k consists entirely of spheres and each of the manifolds N_i are Riemannian with respect to the pulled back metric. Here $\tilde{P}_{i-1}: \tilde{N}_{i-1} \to N_{i-1}$ is the universal covering space of N_{i-1} and N_i is a regular neighborhood of the image of some lift $\tilde{f}_{i-1}$ to this universal covering space.



ASSERTION 1. *The lift f_k has an embedded boundary curve.*

Proof. Since $C = f_k(\partial D)$ lies on a sphere, every Jordan curve in the 1-complex C is homotopically trivial in N_{k-1}. As the fundamental group of C is generated as a $\pi_1(C, p)$ module by Jordan curves, there is a Jordan curve γ' in C such that $\gamma' = P_0 \circ P_1 \circ \cdots \circ P_{k-1}(\gamma)$ does not lie in the normal subgroup of K generated by Γ_{n-1}. If C is not a Jordan curve, then the length of γ_n is not minimal. This shows that C is a Jordan curve. Since C has less length than any nontrivial multiple of C and C is homotopically trivial, the lift $f_k \mid \partial D$ must be an embedding.

ASSERTION 2. *γ_n is embedded.*

Proof. If γ_n is not embedded, then there exists a smallest $m > 0$ such that $f_m \mid \partial D$ is not embedded. By the previous assertion $f_{m+1} \mid \partial D$ exists and is one-to-one. Let $\tilde{f}_m = i \circ f_{m+1}$ be the composition of f_{m+1} with the inclusion map into the total space of the universal covering space $\tilde{P}_m: \tilde{N}_m \to N_m$. By definition of $\tilde{f}_m$, $\tilde{f}_m$ is a lift of the map f_m to its universal covering space. Since f_m is not one-to-one, two points on $\tilde{f}_m(\partial D)$ must be identified under a nontrivial covering transformation τ: $\tilde{N}_m \to \tilde{N}_m$.

First, suppose $\tau(\tilde{f}_m(\partial D)) \neq \tilde{f}_m(\partial D)$. Then with respect to the pulled back metric on N_m, Lemma 1 implies that one of these geodesics, say $\tilde{f}_m(\partial D)$, can be expressed as a product of two closed nongeodesic curves α_1, α_2 with length $(\alpha_i) \leq$ length $(f_m(\partial D)) =$ length (γ_n). Hence either $P_0 \circ \cdots \circ P_{m-1} \circ \tilde{P}_m(\alpha_1)$ or $P_0 \circ \cdots \circ P_{m-1} \circ \tilde{P}_m(\alpha_2)$ does not lie in the normal subgroup of K generated by $\Gamma_{n-1} = \{\gamma_1, \gamma_2, \ldots, \gamma_{n-1}\}$ and has length less than the length of γ_n. This contradicts the least length property of γ_n and shows that γ_n is embedded in the case $\tau(f_m(\partial D)) \neq f_m(\partial D)$.

If $\tau(\tilde{f}_m(\partial D)) = \tilde{f}_m(\partial D)$, then by the Corollary to Theorem 5, τ has a fixed point in $\tilde{N}_m$ which implies that τ is the identity map contrary to our hypothesis about τ. This shows that this case can not occur and that γ_n is embedded. This ends the proof of Assertion 2.

By induction we can continue this process to find a short generating set Γ for K consisting of embedded geodesics. The argument given above also implies that any short generating set consists of embedded geodesics.

Let $\Gamma = \{\gamma_1, \ldots, \gamma_n, \ldots\}$ be a possibly infinite short generating set for K. We will now show that the embedded geodesics in Γ are disjoint and the number of elements in Γ are bounded by $3g$ where g is the genus of Σ. Suppose γ_i and γ_{i+k} are geodesics in Γ which intersect each other and where $k > 0$. Lemma 1 shows that the free homotopy class of one of these geodesics can be expressed as the sum of two homotopy classes of less length. This immediately contradicts the least length property for these geodesics and thereby proves the geodesics in Γ are disjoint. This argument also proves the last statement in the theorem.

If M^2 is a compact orientable surface of genus g and $\Gamma = \{\gamma_1, \ldots, \gamma_{3g+1}\}$ is a collection of $3g+1$ disjoint Jordan curves on the surface, then the classification theorem for compact surfaces can be used to show that two of these Jordan curves are isotopic. Hence there are at most $3g+1$ elements in a short generating set for K where g is the genus of Σ. This last observation completes the proof of the theorem.

THEOREM 7 (Equivariant loop theorem). *Suppose G is a finite group which acts on a compact orientable three-dimensional manifold M with boundary as a group of orientation preserving diffeomorphisms. Then there exists a collection* $\Delta = \{D_1, D_2, \ldots, D_n\}$ *of embedded pairwise disjoint disks in M which satisfy*

(1) $D_i \cap \partial M = \partial D_i$.

(2) *The normal subgroup generated by* $\Gamma = \{\partial D_1, \partial D_2, \ldots, \partial D_n\}$ *is the kernel K of the inclusion map of the fundamental group of each component* Σ *of the boundary of M into M.*

(3) *The union of* Δ *is G invariant.*

Proof. If we produce a collection of disjoint Jordan curves $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ on ∂M such that the normal subgroup of $\pi_1(\partial M)$ generated by Γ is K and G acts freely on the union of Γ, then the theorem will follow from Theorem 5. To prove the existence of such a Γ, we first consider a short generating set $\Gamma' = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ given by Theorem 6. If G acts freely on the union of Γ', then Γ' is the required collection of Jordan curves. If G has a fixed point on Γ', then we carry out the following procedure.

Let N_i be a regular neighborhood of the curve γ_i on ∂M that is small enough so that the collection of these neighborhoods is invariant under G and these neighborhoods are pairwise disjoint. Clearly, N_i is diffeomorphic to $S^1 \times [0, 1]$. Let Γ be the collection of all the boundary circles of these regular neighborhoods. As G acts as a group of orientation preserving transformations of the boundary of M, and N_i is an annulus, any element $g \in G$ which has a fixed point on ∂N_i must be equal to the identity on N_i and hence the identity on M. Therefore G acts freely on the union of Γ. By the previous discussion this completes the proof of the theorem.

REFERENCES

[1] ALMGREN, F. J. Jr. and SIMON, L., *Existence of embedded solutions to Plateau's problem*, Ann. Scuola Norm. Sup. Pisa 6 (1979), 447–495.

[2] COURANT, R., *Dirichlet's principle, conformal mappings and minimal surfaces*, Interscience, New York, 1950.

[3] MEEKS, W. H. III, Lectures on Plateau's Problem, IMPA, Rio de Janeiro, Brazil, 1978.

[4] MEEKS, W. H. III, and YAU, S. T., *The classical Plateau problem and the topology of three dimensional manifolds*, to appear in Arch. Rational Mech. Anal.

[5] ——, *Topology of three dimensional manifolds and the embedding problems in minimal surface theory*, Ann. of Math. *112* (1980), 441–485.

[6] ——, *The existence of embedded minimal sufaces and the problem of uniqueness*, to appear in Math. Zeit.

[7] MORREY, C. B., *Multiple Integrals in the calculus of variations*, Springer-Verlag, Berlin, 1966.

[8] PAPAKYRIAKOPOULOUS, C. D., *On Dehn's lemma and the asphericity of knots*, Ann. of Math. *66* (1957), 1–26.

[9] SHAPIRO, A. and WHITEHEAD, J. H. C., *A proof and extension of Dehn's lemma*, Bull. Amer. Math. Soc. *64* (1958), 174–178.

[10] STALLINGS, J. R., *Group theory and three-dimensional manifolds*, Yale University Press, New Haven, 1971.

[11] The proof of the Smith Conjecture, Proc. of Conference at Columbia University, New York, 1979, in preparation.

[12] TOMI, F. and TROMBA, A. J., *Extreme curves bound embedded minimal surfaces of the type of the disk*, Math. Z. *158* (1978), 137–145.

Institute for Advanced Study
Princeton, N.Y. 08540 U.S.A.

Received July 18, 1979/March 12, 1981

Comment. Math. Helvetici **56** (1981) 240–271 0010–2571/81/002240–32$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Zu einem hyperbolischen Gitterpunktproblem

PETER THURNHEER

I. Einleitung

Sei H der dreidimensionale hyperbolische Raum, Λ eine Bewegungsgruppe von H mit kompaktem messbarem Fundamentalbereich $\mathcal{F}$. Sei Λ diskontinuierlich in dem Sinne, dass das hyperbolische Gitter $\{Tp \mid T \in \Lambda\}$ keine endlichen Häufungspunkte besitzt für alle $p \in H$.

Wie in einem klassischen euklidischen Problem, betrachtet man die Anzahl $N^0(\Lambda, p, q, x)$ der Punkte eines solchen von $p \in H$ erzeugten Gitters – jeder so oft gezählt, wie er als Bild von p unter Λ auftritt – im Innern und auf der hyperbolischen Kugel vom Radius x um $q \in H$, sowie die entsprechenden Rieszschen Summen $N^k(\Lambda, p, q, x)$ positiver Ordnung k. Schreibt man $\rho(v, w)$ für den hyperbolischen Abstand der Punkte $v \in H$, $w \in H$, so ist

$$N^k(\Lambda, p, q, x) := \sum_{\substack{T \in \Lambda \\ \rho(Tp,q) \leq x}} \{x - \rho(Tp, q)\}^k, \quad k \geq 0, \quad x \geq 0, \quad p \in H, \quad q \in H.$$

Man interessiert sich dabei vor allem für das Wachstum dieser Funktionen, wenn der Radius x gegen ∞ geht, insbesondere im Falle $k=0$ und für $p=q$.

In dieser Arbeit werden mehrere Resultate hergeleitet, die Aufschluss geben über die Art dieses Wachstums – eines mit Hilfe einer Verallgemeinerung des Satzes von Wiener–Ikehara, welche im letzten Kapitel formuliert und bewiesen wird. Sie entsprechen alle den klassischen Ergebnissen, die man im analogen euklidischen Problem kennt.

Eine wichtige Rolle bei verschiedenen Beweisen spielt die Laplacetransformierte der Funktion N^k, von der man unter anderem zeigen kann, dass sie einer einfachen Funktionalgleichung genügt.

Die Funktion $N^0(\Lambda, p, p, x)$ wurde – in zwei Dimensionen – erstmals eingeführt und untersucht von H. Huber ([12], [13]), der dabei gezeigt hat, dass ihr Verhalten eng zusammenhängt mit dem Laplace-Beltrami-Eigenwertproblem auf H/Λ.

Diese Arbeit entstand aus der Dissertation, die ich bei Professor K. Chandrasekharan geschrieben und im Herbst 1979 abgegeben habe (man siehe auch [17] und [18]). Ich möchte ihm, sowie seinen früheren Assistenten Dr. H. Joris

und Dr. A. Good danken für Ermutigungen, Anregungen und wertvolle Hinweise.

II. Zum Laplace-Beltrami-Eigenwertproblem

Seien H, Λ und $\mathfrak{F}$ definiert wie in der Einleitung. Das Laplace-Beltrami-Eigenwertproblem auf H/Λ wird von F. Fricker beschrieben und studiert in [10]. Es ergibt sich dabei, dass auch im Falle wo Λ elliptische Elemente enthalten darf, die folgenden Tatsachen gelten: Das Spektrum ist diskret, die Eigenwertfolge $\{\lambda_n \mid n = 0, 1, \ldots\}$ reell und bei geeigneter Numerierung von der Form

$$0 = \lambda_0 < \lambda_1 < \lambda_2 < \ldots, \lim_{n\to\infty} \lambda_n = \infty.$$

Die Vielfachheit jedes Eigenwertes λ, das heisst die Dimension des zu λ gehörigen Eigenraumes ε_λ ist endlich. Es existiert ein vollständiges orthonormiertes System $\{\phi_j \mid j = 0, 1, \ldots\}$ von auf H definierten und Λ-automorphen Eigenfunktionen. Somit ist

$$\int_{\mathfrak{F}} \phi_j(p)\bar{\phi}_m(p)\, d\omega_p = \begin{cases} 1, & j = m. \\ 0, & j \neq m. \end{cases} \tag{1}$$

Dabei steht $d\omega$ für das durch die Metrik in H induzierte Mass. Es bezeichne τ denjenigen Index, für den $\lambda_{\tau-1} \leq 1 < \lambda_\tau$ ist. Man definiert

$$\alpha_n := (1-\lambda_n)^{1/2} \begin{cases} \geq 0 & \text{für} \quad n < \tau. \\ := i\beta_n, \quad \beta_n \geq \beta_\tau > 0 & \text{für} \quad n \geq \tau. \end{cases}$$

Die Menge $\{\alpha_n \mid n = 0, 1, \ldots\}$ ist dann eine diskrete Teilmenge der komplexen Ebene, ohne Häufungspunkte im Endlichen, und für die endlich vielen n mit $0 \leq \lambda_n \leq 1$ ist α_n reell und $1 \geq \alpha_n \geq 0$. Für die übrigen n ist α_n reinimaginär. Es ist $\alpha_0 = 1$ und $\alpha_m = 0$, falls $\lambda_m = 1$. Für $p \in H$, $q \in H$, setzt man weiter

$$K_m(p, q) := \sum_{\phi_j \in \varepsilon_{\lambda_m}} \phi_j(p)\overline{\phi_j}(q), \qquad K^*(p, q) := \sum_{\phi_j \in \varepsilon_1} \phi_j(p)\overline{\phi_j}(q).$$

Dann sind $K_m(p, q)$, $m = 0, 1, \ldots$ und $K^*(p, q)$ unabhängig von der speziellen Wahl des Systems $\{\phi_j \mid j = 0, 1, \ldots\}$, also reell. Schreibt man $|\mathfrak{F}|$ für den hyper-

bolischen Inhalt des Fundamentalbereichs $\mathcal{F}$, so gilt zudem nach [10]

$$K_0(p, q) = 1/|\mathcal{F}| \quad \text{für alle} \quad p \in H,\ q \in H.$$

Es bezeichne Λ_p, $p \in H$, die Untergruppe derjenigen $T \in \Lambda$, für die $Tp = p$ ist. Die Ordnung Ord Λ_p dieser Gruppe ist wegen der Diskontinuität von Λ endlich. Nach A. Selberg ([15]), enthält Λ einen fixpunktfreien Normalteiler N von endlichem Index n. In [10] wird die Existenz eines Repräsentantensystems $\{U_j \mid j = 1, 2, \ldots, n\}$ der Restklassen von Λ modulo N nachgewiesen, für welches $\{U_j \mid j = 1, 2, \ldots, \text{Ord } \Lambda_p\}$ eine Teilmenge von Λ_p ist. Also gilt

$$U_j p = \begin{cases} p, & j = 1, 2, \ldots, \text{Ord } \Lambda_p. \\ r \not\equiv p \bmod \Lambda, & j = \text{Ord } \Lambda_p + 1, \text{Ord } \Lambda_p + 2, \ldots, n. \end{cases} \tag{2}$$

Da N fixpunktfrei ist, wird H/N eine geschlossene dreidimensionale Riemannsche Mannigfaltigkeit und im Zusammenhang mit dem Laplace-Beltrami-Eigenwertproblem auf H/N gelten die klassischen Resultate [1], [14]. Es bezeichne $0 = \lambda_0^N < \lambda_1^N < \cdots$ die entsprechende Eigenwertfolge, $\varepsilon_{\lambda_j^N}$ den zu λ_j^N gehörigen Eigenraum und $\{\phi_j^N \mid j = 0, 1, \ldots\}$ ein vollständiges, orthonormiertes und N-automorphes System von Eigenfunktionen. Setzt man

$$K_m^N(p, q) := \sum_{\phi_i^N \in \varepsilon_{\lambda_m^N}} \phi_j^N(p) \phi_j^N(q), \qquad p \in H, \qquad q \in H,$$

so ist insbesondere nach [1, Satz 1] und [11, Satz 5.1]:

$$\sum_{\lambda_m^N \leq x} K_m^N(p, q) = \frac{\delta_{pq}}{6\pi^2} x^{3/2} + O(x), \qquad \delta_{pq} := \begin{cases} 1, & p \equiv q \bmod \Lambda. \\ 0, & \text{sonst.} \end{cases} \tag{3}$$

Für $p \equiv q \bmod \Lambda$ gilt die Abschätzung gleichmässig in $p \in \mathcal{F}$.

Bemerkungen. (i) Leere Summen sind in der ganzen Arbeit gleich 0 zu setzen. Insbesondere ist also $K^*(p, q) = 0$, falls 1 kein Eigenwert ist.

(ii) Ist die Funktion f rechtsseitig differenzierbar an der Stelle x, so bezeichnet man ihre rechtsseitige Ableitung bei x mit $d^+ f(x)/dx$.

III. Ergebnisse

Sei H der dreidimensionale hyperbolische Raum, Λ eine diskontinuierliche Bewegungsgruppe von H mit kompaktem messbarem Fundamentalbereich $\mathcal{F}$.

Einer Idee in [10] folgend, überträgt man das Resultat [1] von G. V. Avacumović auf den hier betrachteten Fall, wo Λ elliptische Elemente enthalten darf:

$$\sum_{\lambda_n \le x} K_n(p,p) = \frac{\operatorname{Ord} \Lambda_p}{6\pi^2} x^{3/2} + O(x), \tag{4}$$

und

$$\sum_{\lambda_n \le x} |K_n(p,q)| = O(x^{3/2}), \text{ gleichmässig in } (p,q) \in \mathfrak{F} \times \mathfrak{F}. \tag{5}$$

Für $k \ge 0$, $p \in H$, $q \in H$ und reelles x definiert man

$$N^k(\Lambda, p, q, x) := \sum_{\substack{T \in \Lambda \\ \rho(Tp,q) \le x}} \{x - \rho(Tp,q)\}^k,$$

$$Q^k(\Lambda, p, q, x) := \frac{\pi \Gamma(k+1)}{|\mathfrak{F}|\, 2^{k+1}} e^{2x} + \pi\Gamma(k+1) \sum_{0 < \lambda_n < 1} \frac{K_n(p,q)}{\alpha_n(\alpha_n+1)^{k+1}} e^{(\alpha_n+1)x}$$
$$+ 2\pi\Gamma(k+1) K^*(p,q) e^x (x-k-1), \tag{6}$$

$$N^k(\Lambda, p, q, x) := Q^k(\Lambda, p, q, x) + R^k(\Lambda, p, q, x). \tag{7}$$

Es wird sich zeigen, dass Q^k der Hauptterm, R^k der Restterm ist von N^k. Ausgangspunkt aller weiteren Ueberlegungen ist eine Reihendarstellung für $N^k(\Lambda, p, q, x)$, $k > 0$, insbesondere der Fall $k > 1$, in dem diese absolut konvergiert.

SATZ 1. *Für $k > 0$ ist*

$$N^k(\Lambda, p, q, x) = \sum_{n=0}^{\infty} a_n^k(x) K_n(p,q), \tag{8}$$

mit

$$a_n^k(x) = \begin{cases} 4\pi \displaystyle\int_0^x (x-\rho)^k \rho \operatorname{Sinh} \rho \, d\rho, & \lambda_n = 1. \\ \dfrac{4\pi}{\alpha_n} \displaystyle\int_0^x (x-\rho)^k \operatorname{Sinh} \rho \operatorname{Sinh} (\alpha_n \rho) \, d\rho, & \lambda_n \ne 1. \end{cases} \tag{9}$$

Für $k > 1$ konvergiert die Reihe absolut und gleichmässig in $(p,q) \in \mathfrak{F} \times \mathfrak{F}$, sowie gleichmässig in x aus einer beliebigen kompakten Teilmenge von $[0, \infty)$. Für $0 < k \le 1$ ist die Konvergenz gleichmässig in x aus einer beliebigen kompakten Teilmenge von $(0, \infty)$.

Für $k>1$ lässt sich die Identität (8) herleiten durch Anwendung einer speziellen, in [10] gegebenen Form eines klassischen Entwicklungssatzes ([14]). Ihr Gültigkeitsbereich kann auf Grund von (4) ausgedehnt werden nach $k>0$ durch ein Verfahren von K. Chandrasekharan und R. Narasimhan, welches auf einem Aequikonvergenzsatz von A. Zygmund beruht ([4]).

Unter (vii) überlegt man sich direkt, dass die Funktion $N^k(\Lambda, p, q, x)=O(e^{2x})$ ist für $k\geq 0$. Also ist ihre Laplacetransformierte

$$G^k(\Lambda, p, q, s):=\int_0^\infty N^k(\Lambda, p, q, x)e^{-xs}\,dx, \qquad s:=\sigma+it, \quad \sigma>2, \quad k\geq 0$$

analytisch in der Halbebene $\sigma>2$. Mit Hilfe von Satz 1 lässt sie sich meromorph fortsetzen, und zwar auf Grund der Reihendarstellung von

SATZ 2. *Für alle komplexen $s\notin P$, $P:=\{\pm 1\pm\alpha_n \mid n=0, 1, \ldots\}$, und alle $k\geq 0$, gelten die Gleichungen*

$$s^kG^k(\Lambda, p, q, s)=8\pi\Gamma(k+1)\sum_{n=0}^{\infty}\frac{K_n(p,q)}{(s^2-(\alpha_n+1)^2)(s^2-(\alpha_n-1)^2)}, \tag{10}$$

$$G^k(\Lambda, p, q, s)=(-1)^kG^k(\Lambda, p, q, -s).$$

Die Reihe unter (10) *konvergiert absolut und gleichmässig in s aus einer beliebigen kompakten Teilmenge von $\mathbf{C}-P$, sowie gleichmässig in $(p, q)\in\mathcal{F}\times\mathcal{F}$.*

Somit ist $s^kG^k(\Lambda, p, q, s)$, $k\geq 0$, eine in der ganzen komplexen s-Ebene meromorphe Funktion mit der Polmenge P. Die in Satz 2 gegebene einfache Funktionalgleichung für $G^k(\Lambda, p, q, s)$ folgt unmittelbar aus (10). Beim Beweis der Reihendarstellung (10) benützt man die Tatsache, dass sich die Koeffizienten $a_n^k(x)$ für ganzes k explizit berechnen lassen – für $k=2$ siehe man die Formel (58) – sowie die Beziehung

$$s^kG^k(\Lambda, p, q, s)=\Gamma(k+1)G^0(\Lambda, p, q, s), \quad k\geq 0, \quad s\notin P. \tag{11}$$

Speziell für $k=0$ ergibt sich aus (10):

$$G^0(\Lambda, p, q, s)=\frac{1}{s}\sum_{T\in\Lambda}e^{-\rho(Tp,q)s}=8\pi\sum_{n=0}^{\infty}\frac{K_n(p,q)}{(s^2-(\alpha_n+1)^2)(s^2-(\alpha_n-1)^2)}, \quad \sigma>2. \tag{12}$$

Die Sätze 1 und 2 ermöglichen den Beweis verschiedener Resultate, die Aufschluss geben über das Wachstum der Funktion $N^k(\Lambda, p, q, x)$, $k\geq 0$, wenn x gegen ∞ geht.

SATZ 3. *Gleichmässig in* $(p, q) \in \mathcal{F} \times \mathcal{F}$ *ist*

$$R^k(\Lambda, p, q, x) = \begin{cases} O(e^x), & k > 1. \quad (13) \\ O(x^k e^{(3-k)x/2}), & 0 \leq k \leq 1. \quad (14) \end{cases}$$

Um abzuklären, wie präzies die Aussagen (13) und (14) sind, beweist man

SATZ 4. *Für* $k \geq 0$ *ist*

$$R^k(\Lambda, p, q, x) = \Omega \pm (e^x). \quad (15)$$

Ist $R^k(\Lambda, p, p, x) = O(e^{(1+\varepsilon)x})$ *für alle positiven* ε, *so gilt*

$$\liminf_{x \to \infty} R^k(\Lambda, p, p, x)/e^x = -\infty, \quad 0 \leq k < 1. \quad (16)$$

Formel (15) bedeutet $\limsup_{x \to \infty} \inf R^k(\Lambda, p, q, x)/e^x \geqq 0$. Bezeichnet $W^k(\Lambda, p, q, x)$ die Anzahl Vorzeichenwechsel der Funktion $R^k(\Lambda, p, q, y)$ auf dem Intervall $0 \leq y \leq x$, dann geht also $W^k(\Lambda, p, q, x)$ gegen ∞ für $x \to \infty$. Diese Aussage wird präzisiert durch

SATZ 5. *Für* $k \geq 0$ *gilt*

$$W^k(\Lambda, p, q, x) \geq \frac{\beta_\tau}{\pi} x - d_k, \text{ wobei } d_k \text{ unabhängig ist von } x. \quad (17)$$

Aus (16) folgt speziell $R^k(\Lambda, p, p, x) \neq O(e^x)$, für $0 \leq k < 1$, aus (15) $R^k(\Lambda, p, q, x) \neq o(e^x)$, für $x \to \infty$ und $k \geq 0$. Das zeigt, dass das Resultat (13) in dieser Form bestmöglich ist. Anders verhält es sich für $0 \leq k < 1$, denn es gilt

SATZ 6. *Bezüglich des durch die Metrik in* H *induzierten Masses ist für fast alle* $q \in \mathcal{F}$, *mit beliebigem* $\delta > 0$:

$$R^k(\Lambda, p, q, x) = O(e^{(1+\delta)x}), \qquad k \geq 0. \quad (18)$$

Nach Satz 6 ist zu vermuten, dass die Abschätzung (14) noch wesentlich verbessert werden könnte.

Die Sätze 3, 4 und 5 lassen sich beweisen mit Hilfe der Methoden, die von K. Chandrasekharan und R. Narasimhan entwickelt wurden im Rahmen ihrer Theorie über arithmetische Funktionen und Funktionalgleichungen, ([5] bis [8]),

sowie einer Idee, die von J. Steinig im selben Zusammenhang eingeführt wurde ([16]), obwohl die Funktionalgleichung für $G^k(\Lambda, p, q, s)$ nach Satz 2 nicht von derselben Art ist. So erhält man insbesondere einen neuen Beweis für das Resultat (14) im Falle $k=0$ und $p=q$, in dem dieses schon von F. Fricker hergeleitet wurde ([10]).

Die Funktion $G^0(\Lambda, p, q, s)$ lässt sich nach (12) darstellen durch eine Dirichletreihe:

$$G^0(\Lambda, p, q, s) = \frac{1}{S} \sum_{T \in \Lambda} \lambda_T^{-s}, \qquad \lambda_T := e^{\rho(Tp,q)}, \quad \sigma > 2.$$

Anders als zum Beispiel in den Teilerproblemen, kennt man die Verteilung der Grössen λ_T nicht genau. Aus diesem Grunde dürfte $G^0(\Lambda, p, q, s)$ schwerer zu kontrollieren sein, auch in der Halbebene der absoluten Konvergenz und scheint es sinnvoll, beim Beweis von Satz 6 den Hilfssatz aus Kapitel V zu verwenden, anstelle der Laplaceumkehr- respektive Perronschen Formel. Man benützt dabei, dass gilt

$$(d/ds)^m G^0(\Lambda, p, q, s) = O(|t|^{-1+\omega}), \quad \text{für alle ganzen } m \geq 0, \text{ beliebiges } \omega > 1/(4m+10) \text{ und fast alle } q \in \mathcal{F}, \text{ sowie gleichmässig in } \sigma \text{ mit } |\sigma \pm 1| \geq \kappa, \ \kappa \text{ beliebig positiv.} \tag{19}$$

Dies lässt sich zeigen auf Grund von Satz 2.

III. Beweise

(i) *Hilfsbeziehungen.* Sei $\delta > 0$, $x > 0$ und $s := \sigma + i\beta$. Definiert man

$$f(x, s, \delta) := -(\delta - 1) \int_0^\infty (x + w/s)^{\delta-2} e^{-w}\, dw \text{ für } \sigma > 0, \text{ oder für } \sigma < 0 \text{ und } \beta \neq 0, \tag{20}$$

so ist

$$\int_0^x (x-\rho)^{\delta-1} e^{s\rho}\, d\rho = \frac{e^{sx}}{s^\delta}\, \Gamma(\delta) - \frac{x^{\delta-1}}{s} + \frac{1}{s^2} f(x, s, \delta)$$

für $\sigma > 0$, oder für $\sigma < 0$ und $\beta \neq 0$. (21)

Dies folgt durch eine partielle Integration aus der Formel

$$\int_0^x (x-\rho)^{\delta-1} e^{s\rho}\, d\rho = \frac{e^{sx}}{s^\delta}\Gamma(\delta) - \frac{1}{s}\int_0^\infty (x+w/s)^{\delta-1} e^{-w}\, dw,$$

$$x>0, \quad \sigma\neq 0, \quad \delta>0. \tag{22}$$

Für beliebige $\varepsilon>0$, $\beta^*>0$ und

für $0<\delta\leq 2$ gilt gleichmässig in $x\geq\varepsilon$, sowie $\beta\geq\beta^*$:

$$\begin{cases} f(x, \pm 1\pm i\beta, \delta) = O(1). & (23)\\ df(x, \pm 1\pm i\beta, \delta)/dx = O(1). & (24)\end{cases}$$

Beachtet man, dass $|x+w/(\pm 1\pm i\beta)|\geq \varepsilon\beta^*/(1+\beta^*)$ ist, für $\beta\geq\beta^*$, $x\geq\varepsilon$ und $w\in\mathbf{R}$, so ergeben sich die Abschätzungen (23) und (24) aus der Voraussetzung $\delta\leq 2$. Rechnungen, analog denjenigen, welche auf (21) bis (24) führten, zeigen, dass für $s:=\sigma+i\beta$, $\chi:=\min(1,\delta)$ und beliebige $\varepsilon>0$, $\beta^*>0$ auch gilt

$$\int_0^x (x-\rho)^{\delta-1} e^{s\rho}\, d\rho = \begin{cases} O(e^x/|s|^\chi),\ \delta>0,\ \sigma=\pm 1, \\ \qquad\qquad \text{gleichmässig in } x\geq 0 \text{ und } |\beta|\geq\beta^*. & (25)\\ \dfrac{e^{sx}}{s^\delta}\Gamma(\delta) + O(x^\delta/|s|^\chi),\ \delta>0, \\ \qquad\qquad \text{gleichmässig in } \varepsilon\leq\sigma\leq 3,\ x\geq\varepsilon \text{ und } \beta\in\mathbf{R}. & (26)\\ O(x^\delta/|s|^\chi),\ \delta>0, \\ \qquad \text{gleichmässig in } -3\leq\sigma\leq-\varepsilon,\ x\geq\varepsilon \text{ und } \beta\in\mathbf{R}. & (27)\end{cases}$$

Um Beziehungen zwischen den Resttermen R^k verschiedener Ordnungen k zu erhalten, überlegt man sich, dass für die Rieszschen Summen N^k gilt ([2]):

$$kN^{k-1}(\Lambda, p, q, x) = \begin{cases} dN^k(\Lambda, p, q, x)/dx,\ k>1, \\ \qquad\qquad \text{oder } k=1 \text{ und } x\neq\rho(Tp, q), \quad T\in\Lambda, & (28)\\ d^+N^k(\Lambda, p, q, x)/dx,\ k=1 \text{ und } x=\rho(Tp, q), \quad T\in\Lambda,\end{cases}$$

$$N^{k+1}(\Lambda, p, q, x) = (k+1)\int_0^x N^k(\Lambda, p, q, v)\, dv, \quad k\geq 0, \tag{29}$$

und

$$N^k(\Lambda, p, q, x) = k\int_0^x (x-v)^{k-1} N^0(\Lambda, p, q, v)\, dv, \quad k>0. \tag{30}$$

Aus der Definition (6) folgt direkt

$$kQ^{k-1}(\Lambda, p, q, x) = dQ^k(\Lambda, p, q, x)/dx$$

und mit Hilfe von (22):

$$Q^k(\Lambda, p, q, x) = k \int_0^x (x-v)^{k-1} Q^0(\Lambda, p, q, v)\, dv + O(x^{k+1}),$$
$$k > 0, \text{ gleichmässig in } (p, q) \in \mathcal{F} \times \mathcal{F}. \quad (31)$$

Die erste dieser beiden Formeln impliziert wegen (28):

$$kR^{k-1}(\Lambda, p, q, x) = \begin{cases} dR^k(\Lambda, p, q, x)/dx, \ k > 1, \\ \qquad \text{oder } k = 1 \text{ und } x \neq \rho\ (Tp, q),\ T \in \Lambda, \\ d^+R^k(\Lambda, p, q, x)/dx, \ k = 1 \text{ und } x = \rho(Tp, q),\ T \in \Lambda, \end{cases} \quad (32)$$

sowie, wegen (29):

$$R^{k+1}(\Lambda, p, q, x) = (k+1) \int_0^x R^k(\Lambda, p, q, v)\, dv - Q^{k+1}(\Lambda, p, q, 0), \qquad k \geq 0. \quad (33)$$

Zusammen mit (30) ergibt sich aus (31)

$$R^k(\Lambda, p, q, x) = k \int_0^x (x-v)^{k-1} R^0(\Lambda, p, q, v)\, dv + O(x^{k+1}),$$
$$\text{gleichmässig in } (p, q) \in \mathcal{F} \times \mathcal{F}. \quad (34)$$

(ii) *Beweis von* (4) *und* (5), *Folgerungen.* Summiert man bei gegebenem $x \geq 0$ beide Seiten der Gleichung (4) in [10, Seite 406] über alle $\lambda \leq x$ und setzt für $\{S_j \mid j = 1, 2, \ldots, n\}$ das spezielle Repräsentantensystem $\{U_j \mid j = 1, 2, \ldots, n\}$ ein, so ergibt sich

$$\sum_{\lambda_j \leq x} K_j(p, q) = \sum_{m=1}^{n} \sum_{\lambda_j^N \leq x} K_j^N(U_m p, q). \quad (35)$$

Zusammen mit (2) und (3) erhält man daraus (4), wenn man $q = p$ wählt. Für $p = q$ gilt die Abschätzung (3) gleichmässig in $p \in \mathcal{F}$, woraus mit der Cauchyschen Ungleichung folgt

$$\sum_{\lambda_j^N \leq x} |K_j^N(p, q)| = O(x^{3/2}), \text{ gleichmässig in } (p, q) \in \mathcal{F} \times \mathcal{F},$$

was unter Beachtung von (35) die Behauptung (5) beweist. Unter Verwendung von (5) zeigt man mit Hilfe Abelscher partieller Summation:

$$\text{Gleichmässig in } (p,q)\in\mathscr{F}\times\mathscr{F} \text{ gilt:} \begin{cases} \sum\limits_{\lambda_n \geq x>1} |K_n(p,q)|/|\alpha_n|^{\kappa} = O(x^{(3-\kappa)/2}), & \kappa>3. \quad (36) \\ \sum\limits_{1<\lambda_n\leq x} |K_n(p,q)|/|\alpha_n|^{\kappa} = \begin{cases} O(x^{(3-\kappa)/2}) & \kappa<3. \\ O(\log x), & \kappa=3. \end{cases} & (37) \end{cases}$$

Da nur endlich viele $\lambda_n \leq 1$ sind, folgt aus (36):

$$\sum_{\lambda_n\neq 1} |K_n(p,q)|/|\alpha_n|^{\kappa} \text{ konvergiert für } \kappa>3 \text{ gleichmässig in } (p,q)\in\mathscr{F}\times\mathscr{F}. \quad (38)$$

(iii) *Reihendarstellung von* $N^k(\Lambda,p,q,x)$, $k>2$. Definiert man bei gegebenem $x\geq 0$

$$F_x^k(v) := \begin{cases} (x-v)^k, & 0\leq v\leq x, \\ 0, & v>x, \end{cases}$$

so hat $F_x^k(v)$ kompakten Träger und ist für $k>2$ zweimal stetig differenzierbar nach v auf $[0,\infty)$. Zudem ist $N^k(\Lambda,p,q,x)=\sum_{T\in\Lambda} F_x^k(\rho(Tp,q))$. Dies erlaubt auf Grund der Entwicklung in [10, Seite 411] die in Satz 1 gegebene Reihendarstellung der Funktion $N^k(\Lambda,p,q,x)$, $k>2$, zu gewinnen.

(iv) *Reihendarstellung von* $N^k(\Lambda,p,q,x)$, $k>1$. Für alle reellen $k>-1$ seien die Koeffizienten $a_n^k(x)$ definiert durch die Formel (9). Dann ist

$$a_n^{k+1}(x) = (k+1)\int_0^x a_n^k(v)\,dv, \qquad n=0,1,\ldots,k>-1, \quad (39)$$

und

$$k a_n^{k-1}(x) = da_n^k(x)/dx, \qquad n=0,1,\ldots,k>0. \quad (40)$$

Sei $\delta>0$. Für λ_n ungleich 0 oder 1 erhält man aus (9) durch zweimalige partielle Integration:

$$a_n^{1+\delta}(x) = \frac{8\pi(1+\delta)}{(\alpha_n^2-1)^2}x^{\delta} + \frac{\pi(1+\delta)\delta}{\alpha_n(\alpha_n+1)^2}\int_0^x (x-\rho)^{\delta-1}[e^{(\alpha_n+1)\rho}+e^{(-\alpha_n-1)\rho}]\,d\rho$$
$$-\frac{\pi(1+\delta)\delta}{\alpha_n(\alpha_n-1)^2}\int_0^x (x-\rho)^{\delta-1}[e^{(\alpha_n-1)\rho}+e^{(-\alpha_n+1)\rho}]\,d\rho, \quad \delta>0, \quad \lambda_n\neq {}^0_1. \quad (41)$$

Beachtet man (25), so folgt aus (41):

$$a_n^{1+\delta}(x) = O(e^x/|\alpha_n|^\chi),\ \delta > 0,\ \chi := \min(3+\delta, 4), \text{ gleichmässig in } x \geq 0 \text{ und } n \geq \tau. \tag{42}$$

Nun zeigt (42) zusammen mit (38) erstens: *Die Reihe* (8) *für* $N^k(\Lambda, p, q, x)$, $k > 2$, *konvergiert absolut und gleichmässig in* $(p, q) \in \mathcal{F} \times \mathcal{F}$, *sowie gleichmässig in* x *aus einer beliebigen kompakten Teilmenge von* $[0, \infty)$, und zweitens, wenn man (40) beachtet: *Genau dasselbe gilt für die – vorerst formal – gliedweise differenzierte Reihe.* Wegen (28) beweist dies die Gültigkeit von (8) für $k > 1$.

(v) *Reihendarstellung von* $N^k(\Lambda, p, q, x)$, $k > 0$. Im ganzen Abschnitt (v) bezeichne δ eine feste Zahl, $0 < \delta \leq 2$. Durch einmalige partielle Integration erhält man aus (9):

$$a_n^\delta(x) = \frac{\pi\delta}{\alpha_n} \int_0^x (x-\rho)^{\delta-1} \left[\frac{e^{(\alpha_n+1)\rho}}{\alpha_n+1} + \frac{e^{(-\alpha_n-1)\rho}}{-\alpha_n-1} - \frac{e^{(\alpha_n-1)\rho}}{\alpha_n-1} - \frac{e^{(-\alpha_n+1)\rho}}{-\alpha_n+1} \right] d\rho, \quad \lambda_n \neq {}^0_1. \tag{43}$$

Für $n \geq \tau$ und $x > 0$ ergibt sich daraus mit (21):

$$\begin{aligned} a_n^\delta(x) - \frac{\pi}{\alpha_n}\Gamma(\delta+1) &\left[\frac{e^{(\alpha_n+1)x}}{(\alpha_n+1)^{1+\delta}} + \frac{e^{(-\alpha_n-1)x}}{(-\alpha_n-1)^{1+\delta}} - \frac{e^{(\alpha_n-1)x}}{(\alpha_n-1)^{1+\delta}} - \frac{e^{(-\alpha_n+1)x}}{(-\alpha_n+1)^{1+\delta}} \right] \\ &= -\frac{2\pi\delta}{\alpha_n} x^{\delta-1}\{1/(\alpha_n+1)^2 - 1/(\alpha_n-1)^2\} \\ &\quad + \frac{\pi\delta}{\alpha_n}\left[\frac{f(x, \alpha_n+1, \delta)}{(\alpha_n+1)^3} + \frac{f(x, -\alpha_n-1, \delta)}{(-\alpha_n-1)^3} - \frac{f(x, \alpha_n-1, \delta)}{(\alpha_n-1)^3} \right. \\ &\qquad \left. - \frac{f(x, -\alpha_n+1, \delta)}{(-\alpha_n+1)^3} \right]. \end{aligned} \tag{44}$$

Beachtet man (23), so sieht man, dass alle Summanden auf der rechten Seite dieser Gleichung $O(|\alpha_n|^{-4})$ sind, gleichmässig in $n \geq \tau$ und x aus einem beliebigen kompakten Teilintervall I von $(0, \infty)$. Nun multipliziert man beide Seiten von (44) mit $K_n(p, q)$ und summiert über alle $n \geq \tau$. Nach (38) konvergieren dann die beidseits entstehenden Reihen absolut und man darf auf der rechten Seite beliebig umordnen. Mit

$$\xi = \xi(\Lambda, p, q, \delta) := 8\pi\delta \sum_{n \geq \tau} \frac{K_n(p, q)}{(\alpha_n^2 - 1)^2}$$

und

$$P^{\delta}_{(x)} = P^{\delta}_{(\Lambda,p,q,x)} := \sum_{n \geq \tau} \left[a_n^{\delta}(x) - \frac{\pi}{\alpha_n} \Gamma(\delta+1) \left[\frac{e^{(\alpha_n+1)x}}{(\alpha_n+1)^{1+\delta}} + \frac{e^{(-\alpha_n-1)x}}{(-\alpha_n-1)^{1+\delta}} - \frac{e^{(\alpha_n-1)x}}{(\alpha_n-1)^{1+\delta}} - \frac{e^{(-\alpha_n+1)x}}{(-\alpha_n+1)^{1+\delta}} \right] \right] K_n(p,q) \tag{45}$$

wird

$$P^{\delta}(x) = \xi x^{\delta-1} + \pi\delta \sum_{n \geq \tau} \left[\frac{f(x, \alpha_n+1, \delta) - f(x, -\alpha_n-1, \delta)}{(1+\alpha_n)^3} - \frac{f(x, -\alpha_n+1, \delta) - f(x, \alpha_n-1, \delta)}{(1-\alpha_n)^3} \right] \frac{K_n(p,q)}{\alpha_n}. \tag{46}$$

Dabei gilt:

Die Reihe (45) für $P^{\delta}(x)$ konvergiert absolut und gleichmässig in $x \in I$, wobei I ein beliebiges kompaktes Teilintervall von $(0,\infty)$ ist. (47)

Nach (23), (24) und (38) konvergiert die Summe auf der rechten Seite von (46), sowie ihre – vorerst formale – gliedweise Ableitung nach x gleichmässig in $x \in I$. Da die Funktionen $f(x, \pm 1 \pm \alpha_n, \delta)$, $\delta > 0$, $n \geq \tau$, und ihre Ableitungen nach x stetig sind in x, $x > 0$, folgt:

$P^{\delta}(x)$ ist stetig differenzierbar nach x, $0 < x < \infty$. (48)

Weiter ist

$$\sum_{m \leq \beta_n \leq m+1} \left| \frac{K_n(p,q)}{\alpha_n (\pm 1 \pm \alpha_n)^{1+\delta}} \right| = o(1) \quad \text{für} \quad m \to \infty. \tag{49}$$

Man zeigt dies mit Hilfe von (4) zuerst für $p = q$ und erhält daraus (49) wegen $|K_n(p,q)| \leq K_n(p,p) + K_n(q,q)$. Nach (38) ist

$$\sum_{n \geq \tau} \left| \frac{K_n(p,q)}{\alpha_n (\pm 1 \pm \alpha_n)^{1+\delta}} \right| \beta_n^{-1} < \infty. \tag{50}$$

Es bezeichne nun I wie oben ein kompaktes Teilintervall von $(0,\infty)$, J ein abgeschlossenes Intervall, enthalten im Innern von I. Sei μ die Länge von I. Aus

(49) und (50), folgt mit [4, Corollary 3, Seite 221]:

$$\left.\begin{array}{l} \pi\Gamma(\delta+1)\sum\limits_{n\geq\tau}\dfrac{K_n(p,q)}{\alpha_n}\dfrac{e^{(\pm1\pm\alpha_n)x}}{(\pm1\pm\alpha_n)^{1+\delta}} \quad \text{ist gleichmässig äquikonvergent in } x\in I \text{ zur abgeleiteten Fourier Reihe einer } \mu\text{-periodischen Funktion. Diese Funktion ist auf } I \text{ gleich} \\ \pi\Gamma(\delta+1)\gamma(x)\sum\limits_{n\geq\tau}\dfrac{K_n(p,q)}{\alpha_n(\pm1\pm\alpha_n)^{1+\delta}}\int_{x_0}^{x}e^{(\pm1\pm\alpha_n)v}\,dv, \quad x_0\in J, \end{array}\right\} \tag{51}$$

$$\text{mit } \gamma\in C_I^{\infty},\ \gamma(x)=1 \quad \text{für} \quad x\in J. \tag{52}$$

Sei $L^*(x)=L^*(\delta,\Lambda,p,q,x)$ die μ-periodische Funktion, für die gilt

$$L^*(x)=\gamma(x)\int_{x_0}^{x}P^{\delta}(v)\,dv, \quad \text{für} \quad x\in I.$$

Nach (52) ist dann

$$dL^*(x)/dx=P^{\delta}(x) \quad \text{für} \quad x\in J. \tag{53}$$

Ist f eine auf I definierte Funktion, so bezeichne f^* die μ-periodische Funktion mit $f^*(x)=f(x)$ für alle $x\in I$.

Wegen (48) ist die Fourier Reihe von $\{P^{\delta}(x)\}^*$ gleichmässig konvergent in $x\in J$. Auf J ist deshalb nach (53) die Ableitung der Fourier Reihe von $L^*(x)$ gleich der – gleichmässig in x aus J gegen $P^{\delta}(x)$ konvergenten – Fourier Reihe von $\{P^{\delta}(x)\}^*$. Zusammen mit (47) folgt:

Die Reihe (45) für $P^{\delta}(x)$ ist gleichmässig äquikonvergent in $x\in J$ zur Ableitung der Fourier Reihe von $L^*(x)$.

Aus (47), (51) und der Definition von $L^*(x)$ erhält man damit

$$\left.\begin{array}{l} \sum\limits_{n\geq\tau}a_n^{\delta}(x)K_n(p,q) \quad \text{ist gleichmässig äquikonvergent in } x\in J \text{ zur Ableitung der Fourier Reihe der Funktion} \\ \left[\gamma(x)\sum\limits_{n\geq\tau}K_n(p,q)\int_{x_0}^{x}a_n^{\delta}(v)\,dv\right]^*=\left[\dfrac{\gamma(x)}{\delta+1}\sum\limits_{n\geq\tau}K_n(p,q)a_n^{1+\delta}(x)-\gamma(x)c_1\right]^* \\ =\left[\dfrac{\gamma(x)}{\delta+1}\{N^{1+\delta}(\Lambda,p,q,x)-T^{1+\delta}(x)\}-\gamma(x)c_1\right]^*,\ c_1 \text{ unabhängig von } x. \end{array}\right\} \tag{54}$$

Dabei werde gesetzt

$$c_1 = c_1(\delta, \Lambda, p, q, x_0) := (1+\delta)^{-1} \sum_{n \geq \tau} a_n^{1+\delta}(x_0) K_n(p, q)$$

und für $k > -1$:

$$T^k(x) = T^k(\Lambda, p, q, x) := \sum_{n < \tau} a_n^k(x) K_n(p, q).$$

Man wird sich überlegen:

Die Ableitung der Fourier Reihe der Funktion
$\{\gamma(x)(\delta+1)^{-1}\{N^{1+\delta}(\Lambda, p, q, x) - T^{1+\delta}(x)\} - \gamma(x)c_1\}^*$ konvergiert
gleichmässig in $x \in J$. (55)

Das bedeutet wegen (54), dass $\sum_{n \geq \tau} a_n^\delta(x) K_n(p, q)$ gleichmässig äquikonvergent ist auf J zu einer gleichmässig auf J konvergenten Reihe und somit dort ebenfalls gleichmässig konvergiert. Es folgt, dass auch $\sum_{n=0}^{\infty} a_n^\delta(x) K_n(p, q)$ gleichmässig konvergiert auf J. Zusammen mit (28) und (40) vervollständigt diese Tatsache den Beweis von Satz 1. Es bleibt noch (55) zu verifizieren. Offensichtlich gilt:

Die Ableitung der Fourier Reihe der Funktion
$\{\gamma(x)(\delta+1)^{-1}\{N^{1+\delta}(\Lambda, p, q, x) - T^{1+\delta}(x)\} - \gamma(x)c_1\}^*$ konvergiert sicher dort
gleichmässig, wo die Fourier Reihe der Ableitung dieser Funktion
gleichmässig konvergiert. (56)

Nach (40) und (9) ist $dT^{1+\delta}(x)/dx = (1+\delta)T^\delta(x)$ und $T^\delta(x)$ stetig differenzierbar nach x auf I. Also konvergiert die Fourier Reihe von $d/dx\{T^{1+\delta}(x)\}^*$ gleichmässig auf J. Da ausserdem $\gamma \in C_I^\infty$ ist, folgt

Die Fourier Reihe von $d/dx\{\gamma(x)(\delta+1)^{-1}\{N^{1+\delta}(\Lambda, p, q, x) - T^{1+\delta}(x)\} -$
$\gamma(x)c_1\}^*$ konvergiert zumindest dort gleichmässig, wo die Fourier Reihe
von $d/dx\{(1+\delta)^{-1}N^{1+\delta}(\Lambda, p, q, x)\}^* = \{N^\delta(\Lambda, p, q, x)\}^*$ gleichmässig
konvergiert. (57)

Als Summenfunktion positiver stetiger Funktionen ist $N^\delta(\Lambda, p, q, x)$ stetig und nicht fallend, also von beschränkter Schwankung auf jedem kompakten Intervall. Somit ([19, Seite 410]), konvergiert die Fourier Reihe von $\{N^\delta(\Lambda, p, q, x)\}^*$ gleichmässig in $x \in J$. Das beweist (55) wegen (56) und (57).

(vi) *Die explizite Reihe für* $N^2(\Lambda, p, q, x)$. Für ganzes k lassen sich die Koeffizienten $a_n^k(x)$ durch partielle Integration explizit berechnen. Speziell für $k=2$ *erhält man:*

$$N^2_{(\Lambda,p,q,x)} = \frac{\pi}{|\mathscr{F}|}\left\{\frac{1}{2}\operatorname{Sinh} 2x - \frac{2x^3}{3} - x\right\} + 8\pi K^*_{(p,q)}\{x \operatorname{Cosh} x - 3 \operatorname{Sinh} x + 2x\} + 4\pi \sum_{\lambda_n \neq 0,1} \frac{K_n(p,q)}{\alpha_n}\left[\frac{\operatorname{Sinh}((\alpha_n+1)x)}{(\alpha_n+1)^3} - \frac{\operatorname{Sinh}((\alpha_n-1)x)}{(\alpha_n-1)^3} + \frac{4\alpha_n x}{(\alpha_n^2-1)^2}\right]. \tag{58}$$

(vii) *Beweis von Satz* 2, (11) *und* (12). Bezeichnet $I(x)$ den hyperbolischen Inhalt einer hyperbolischen Kugel vom Radius x, so ist ([10, Seite 409]) $I(x) = \pi\{\operatorname{Sinh} 2x - 2x\}$. Da $\mathscr{F}$ kompakt ist, existiert ein $\xi > 0$, sodass $\mathscr{F}$ enthalten ist in einer Kugel vom hyperbolischen Radius ξ und es gilt $N^0(\Lambda, p, q, x) = O(I(x+\xi)/|\mathscr{F}|)$. Also ist die Funktion $N^k(\Lambda, p, q, x) = O(e^{2x})$ für $k = 0$ – und damit nach (30) für alle $k \geq 0$ – und ihre Laplacetransformierte $G^k(\Lambda, p, q, s)$, $k \geq 0$, $s := \sigma + it$, ist analytisch in der Halbebene $\sigma > 2$. Dabei gilt

$$s^k G^k(\Lambda, p, q, s) = \Gamma(k+1) G^0(\Lambda, p, q, s), \quad \sigma > 2, \quad k \geq 0. \tag{59}$$

Nach dem Identitätsprinzip genügt es, die Aussage (59) für reelles $s > 2$ zu beweisen. Setzt man in die Definition von $G^k(\Lambda, p, q, s)$, $k > 0$, die Beziehung (30) ein, so erhält man (59) für $s > 2$ und $k > 0$, indem man sich überlegt, dass man die Integrationen vertauschen darf und anschliessend $s(x-v)$ substituiert. Zudem ist (59) offensichtlich richtig für $k = 0$.

Die Formeln (58) und (38) zeigen, dass die Reihe $e^{-xs} \sum_{n=0}^{\infty} a_n^2(x) K_n(p, q)$ absolut konvergiert und zwar gleichmässig in $x \in [0, \infty)$ für jedes s mit $\sigma > 2$. Setzt man also in die Definition von $G^2(\Lambda, p, q, s)$ die explizite Reihe (58) für $N^2(\Lambda, p, q, x)$ ein, so erhält man durch Vertauschung von Summation und Integration

$$G^2(\Lambda, p, q, s) = \frac{16\pi}{s^2} \sum_{n=0}^{\infty} \frac{K_n(p,q)}{(s^2-(\alpha_n+1)^2)(s^2-(\alpha_n-1)^2)}, \quad \sigma > 2. \tag{60}$$

Mit Hilfe von (60) soll $G^2(\Lambda, p, q, s)$ meromorph fortgesetzt werden nach ganz $\mathbf{C}$. Dazu definiert man $P := \{\pm 1 \pm \alpha_n \mid n = 0, 1, \ldots\}$. Dann ist P eine diskrete Punktmenge ohne Häufungspunkt im Endlichen. Wegen $\alpha_0 = 1$ gehört 0 zu P.

Sei M eine beliebige kompakte Teilmenge von $\mathbf{C}-P$. Da gilt

$1/(s^2-(\alpha_n+1)^2)(s^2-(\alpha_n-1)^2)=O(1/|\alpha_n|^4)$, gleichmässig in $n=0,1,\ldots$ und $s\in M$,

konvergiert die Reihe (60) nach (38) gleichmässig in $s\in M$ sowie in $(p,q)\in\mathfrak{F}\times\mathfrak{F}$ und

$G^2(\Lambda,p,q,s)$ ist meromorph in ganz $\mathbf{C}$ mit der Polmenge P.

Mit (59) und (60) erhält man daraus durch wiederholte Anwendung des Prinzips der analytischen Forteszung und des Identitätsprinzips Satz 2, sowie (11). Wegen der für $\sigma>2$ gültigen Beziehung

$$G^0(\Lambda,p,q,s)=\int_0^\infty e^{-xs}\sum_{\substack{T\in\Lambda\\ \rho(Tp,q)\le x}}1\,dx=\sum_{T\in\Lambda}\int_{\rho(Tp,q)}^\infty e^{-xs}\,dx$$
$$=\frac{1}{s}\sum_{T\in\Lambda}e^{-\rho(Tp,q)s},\qquad \sigma>2,$$

ist damit auch (12) bewiesen.

(viii) *Beweis von* (13). Man setzt $\theta:=\min_{n<\tau}\alpha_n$. Dann ist $0<\theta\le 1$ und unter Beachtung von (41) lassen sich mit (26) und (27) die Koeffizienten $a_n^{1+\delta}(x)$ in folgender Weise darstellen:

$$a_n^{1+\delta}(x)=\pi\frac{\Gamma(2+\delta)}{\alpha_n}\left[\frac{e^{(1+\alpha_n)x}}{(1+\alpha_n)^{2+\delta}}-\frac{e^{(1-\alpha_n)x}}{(1-\alpha_n)^{2+\delta}}\right]$$
$$+\begin{cases}O(x^\delta)+O(e^{(1-\theta)x}), & 0<\lambda_n<1.\\ O(x^\delta/|\alpha_n|^x), \text{ gleichmässig in } n\ge\tau.\end{cases}\qquad(61)$$

Dabei ist $\delta>0$ vorausgesetzt und $\chi:=\min(3+\delta,4)$. Unter Verwendung von (22) folgert man direkt aus (9):

$$a_n^{1+\delta}(x)=2\pi e^x\Gamma(2+\delta)\{x-2-\delta\}+O(x^{2+\delta}),\qquad \lambda_n=1,\qquad \delta>0,$$

$$a_0^{1+\delta}(x)=\frac{\pi}{2^{2+\delta}}\Gamma(2+\delta)e^{2x}+O(x^{2+\delta}),\qquad \delta>0.$$

Also gilt speziell für die unter (6) definierte Funktion Q^k:

$$\sum_{n<\tau} a_n^{1+\delta}(x)K_n(p,q) = Q^{1+\delta}(\Lambda, p, q, x) + O(e^{(1-\theta)x}) + O(x^{2+\delta}), \qquad \delta > 0.$$

Setzt man diese, sowie die Beziehung (61) in die Reihendarstellung (8) für $N^{1+\delta}(\Lambda, p, q, x)$ ein, so erhält man

$$\begin{aligned} N^{1+\delta}(\Lambda, p, q, x) &= Q^{1+\delta}(\Lambda, p, q, x) \\ &\quad + \pi\Gamma(2+\delta)e^x \sum_{n\geq\tau} \frac{K_n(p,q)}{\alpha_n}\left[\frac{e^{\alpha_n x}}{(1+\alpha_n)^{2+\delta}} - \frac{e^{-\alpha_n x}}{(1-\alpha_n)^{2+\delta}}\right] \\ &\quad + O(e^{(1-\theta)x}) + O(x^{2+\delta}) + O(x^{\delta}) \sum_{n\geq\tau} K_n(p,q)/|\alpha_n|^{\chi}, \\ &\qquad\qquad \delta > 0, \qquad \chi := \min(3+\delta, 4). \end{aligned} \tag{62}$$

Für $n \geq \tau$ und $x \in \mathbf{R}$ ist $|e^{\alpha_n x}| = 1$. Beachtet man noch (38), so folgt aus (62) die Behauptung (13).

(ix) *Beweis von* (14) *für* $k = 0$, $k = 1$. Die Abschätzung (14) im Falle $k = 0$ und für $p = q$, wurde bewiesen von F. Fricker ([10]), mit Hilfe einer Approximation der Funktion $N^0(\Lambda, p, p, x)$ von oben und unten durch bezüglich der Koordinaten in H zweimal stetig differenzierbare und Λ-automorphe Funktionen. Die Anwendung seines unter (iii) zitierten Entwicklungssatzes auf diese Funktionen lieferte das entsprechende Resultat.

Für $k = 0$, sowie für $k = 1$, kann man (14) auch herleiten unter Verwendung der Methode der endlichen Differenzen, ausgehend von der expliziten Reihendarstellung (58) für $N^2(\Lambda, p, q, x)$. Sei die Funktion $y = y(x)$ reellwertig und definiert auf $[0,\infty)$, $y(x) \neq 0$ für $x \geq 0$ und $y(x) = o(1)$, $x \to \infty$. Für eine beliebige, auf $[0,\infty)$ definierte Funktion F setzt man

$$\begin{aligned} \Delta_y^1(F(x)) &:= \frac{1}{y}\{F(x+y) - F(x)\}, \\ &\qquad\qquad x \geq 0, \qquad x + 2y \geq 0. \\ \Delta_y^2(F(x)) &:= \frac{1}{y^2}\{F(x+2y) - 2F(x+y) + F(x)\}, \end{aligned}$$

Durch geeignete Entwicklungen in Taylorreihen mit Restglied überlegt man sich,

dass für $j = 1, 2$ gilt

$$\Delta_y^j(x^\kappa) = O(x^{(\kappa-j)}), \qquad \kappa \in \mathbf{R}, \qquad \Delta_y^j(e^{\alpha x}) = O(e^x), \text{ gleichmässig in } -3 \leq \alpha \leq 1.$$

$$\Delta_y^j(e^{(\pm 1 \pm i\beta)x}) = \begin{cases} O(e^x/|y|^j), \\ O(e^x(1+|\beta|)^j), \end{cases} \text{ gleichmässig in } \beta \in \mathbf{R}.$$

$$\Delta_y^j(e^{(\pm 1 \pm \alpha)x}) = (\pm 1 \pm \alpha)^j e^{(\pm 1 \pm \alpha)x} + O(e^{2x}|y|), \text{ gleichmässig in } 0 \leq \alpha \leq 1.$$

$$\Delta_y^j(xe^x) = xe^x + je^x + O(xe^x|y|).$$

Mit diesen Formeln und der Definition (6) für $Q^k(\Lambda, p, q, x)$ erhält man aus den unter (58) explizit gegebenen Ausdrücken für $a_n^2(x)$:

$$\Delta_y^j(a_n^2(x)) = \begin{cases} O(e^x/|\alpha_n|^4|y|^j), \\ O(e^x/|\alpha_n|^{4-j}), \end{cases} \quad j = 1, 2, \text{ gleichmässig in } n \geq \tau, \tag{63}$$

$$\tag{64}$$

$$\Delta_y^j \left\{ \sum_{n<\tau} a_n^2(x) K_n(p, q) \right\} = 2Q^{2-j}(\Lambda, p, q, x) + O(e^{2x}|y|) + O(e^x), \quad j = 1, 2. \tag{65}$$

Nun wendet man die Operatoren Δ_y^j auf die Gleichung (8) an. Für $k = 2$ konvergiert die Reihe auf der rechten Seite von (8) absolut. Aus diesem Grunde darf man rechts Summation und Anwendung von Δ_y^j vertauschen. Auf die Terme mit $n < \tau$ wendet man (65) an, auf diejenigen mit $1 < \lambda_n \leq |y|^{-2}$ die Beziehung (64) und die restlichen Summanden schätzt man ab mit Hilfe von (63). So ergibt sich für $j = 1, 2$:

$$\Delta_y^j(N^2(\Lambda, p, q, x)) = 2Q^{2-j}(\Lambda, p, q, x) + O(e^x) \sum_{1<\lambda_n \leq |y|^{-2}} |K_n(p, q)|/|\alpha_n|^{4-j}$$

$$+ O(e^x/|y|^j) \sum_{\lambda_n > |y|^{-2}} |K_n(p, q)|/|\alpha_n|^4 + O(e^{2x}|y|) + O(e^x).$$

Die O-Abschätzungen in dieser Beziehung sind gleichmässig in $(p, q) \in \mathfrak{F} \times \mathfrak{F}$, denn nach (5) sind es auch diejenigen unter (65) und wegen (36) und (37) folgt, dass gleichmässig in $(p, q) \in \mathfrak{F} \times \mathfrak{F}$ gilt

$$\Delta_y^1(N^2(\Lambda, p, q, x)) = 2Q^1(\Lambda, p, q, x) + O(e^x|\log|y||) + O(e^x) + O(e^{2x}|y|), \tag{66}$$

$$\Delta_y^2(N^2(\Lambda, p, q, x)) = 2Q^0(\Lambda, p, q, x) + O(e^x|y|^{-1}) + O(e^{2x}|y|), \tag{67}$$

Andererseits ist nach (29):

$$\Delta_y^1(N^2(\Lambda, p, q, x)) = \frac{2}{y}\int_x^{x+y} N^1(\Lambda, p, q, v)\, dv,$$

$$\Delta_y^2(N^2(\Lambda, p, q, x)) = \frac{2}{y^2}\int_x^{x+y}\left\{\int_u^{u+y} N^0(\Lambda, p, q, v)\, dv\right\} du,$$

sodass, da die Funktionen $N^m(\Lambda, p, q, x)$, $m = 0, 1$ monoton wachsen, auch gilt

$$\begin{aligned} 2N^{2-j}(\Lambda, p, q, x) &\leq \Delta_y^j(N^2(\Lambda, p, q, x)) \quad \text{für} \quad y > 0, \\ 2N^{2-j}(\Lambda, p, q, x) &\geq \Delta_y^j(N^2(\Lambda, p, q, x)) \quad \text{für} \quad y < 0, \end{aligned} \qquad j = 1, 2.$$

Das Resultat (14) im Falle $k = 1$ ergibt sich aus diesen Ungleichungen, zusammen mit (66), wenn man $j = 1$ und $y = \pm e^{-x}$ wählt, und für $k = 0$ erhält man (14), indem man $j = 2$, $y = \pm e^{-x/2}$ setzt und (67) beachtet.

(x) *Beweis von* (14) *für* $0 < k < 1$. Mit Hilfe der unter (ix) hergeleiteten Resultate soll der Beweis von (14) vervollständigt werden. Das Vorgehen ist analog zu demjenigen beim Beweis von Satz 1.71 in [2]. Nach (34) gilt für beliebiges $\xi \geq 0$:

$$R^k(\Lambda, p, q, x) = k\left\{\int_0^{\xi} + \int_{\xi}^{x}\right\}(x-v)^{k-1} R^0(\Lambda, p, q, v)\, dv + O(x^{k+1}),$$
$$k > 0, \text{ gleichmässig in } (p, q) \in \mathfrak{F} \times \mathfrak{F}. \quad (68)$$

Man wählt nun $\xi = \xi(x) := x - xe^{-x/2}$, das heisst $x - \xi = xe^{-x/2}$. Aus dem schon bewiesenen Resultat (14) im Falle $k = 0$ ersieht man, dass dann das Integral von ξ bis x in (68) gleich $O(x^k e^{(3-k)x/2})$ ist für $0 < k < 1$ und gleichmässig in $(p, q) \in \mathfrak{F} \times \mathfrak{F}$. Zudem existiert nach dem zweiten Mittelwertsatz der Integralrechnung ([9, Seite 256]) ein u, $0 \leq u \leq \xi$, sodass gilt

$$\begin{aligned} &\int_0^{\xi}(x-v)^{k-1} R^0(\Lambda, p, q, v)\, dv \\ &= (x-\xi)^{k-1}\int_u^{\xi} R^0(\Lambda, p, q, v)\, dv + x^{k-1}\int_0^{u} R^0(\Lambda, p, q, v)\, dv \\ &= (x-\xi)^{k-1} R^1(\Lambda, p, q, v)\Big|_{v=u}^{v=\xi} + x^{k-1} R^1(\Lambda, p, q, v)\Big|_{v=0}^{v=u} + O(x^{k-1}) \\ &= O(x^k e^{(3-k)x/2}), \quad 0 < k < 1, \text{gleichmässig in } (p, q) \in \mathfrak{F} \times \mathfrak{F}. \end{aligned}$$

gleichmässig in $(p, q) \in \mathfrak{F} \times \mathfrak{F}$.

Diese letzte Abschätzung, welche man aus dem Resultat (14) für $k=1$ erhält, beweist, eingesetzt in (68) die Aussage (14) vollständig.

(xi) *Beweis von* (15) *und* (17). Wird definiert

$$S^k(\Lambda, p, q, x) := R^k(\Lambda, p, q, x) - \pi\Gamma(k+1)e^x \sum_{n\geq\tau} \frac{K_n(p,q)}{\alpha_n}\left[\frac{e^{\alpha_n x}}{(1+\alpha_n)^{k+1}} - \frac{e^{-\alpha_n x}}{(1-\alpha_n)^{k+1}}\right], \qquad k>1,$$

so ist nach (62)

$$S^k(\Lambda, p, q, x) = O(e^{(1-\theta)x}) + O(x^{k+1}), \qquad k>1. \tag{69}$$

Setzt man für $n\geq\tau$ $\arg(1-\alpha_n) := \psi_n$, so wird $1-\alpha_n = \lambda_n^{1/2} e^{i\psi_n}$, $n\geq\tau$.

Mit dieser Schreibweise ergibt sich aus der Definition von $S^{k-1}(\Lambda, p, q, x)$:

$$R^{k-1}(\Lambda, p, q, x) = 2\pi\Gamma(k)e^x \sum_{n\geq\tau} \frac{K_n(p,q)}{\beta_n \lambda_n^{k/2}} \sin(\beta_n x + \psi_n k) + S^{k-1}(\Lambda, p, q, x), \qquad k>2. \tag{70}$$

Aus (38) und der Tatsache, dass die Folge λ_j strikt monoton wachsend ist, folgert man die Existenz einer reellen Zahl $k_0>4$, sodass für alle $k\geq k_0$ gilt

$$\Gamma(k)\left[\frac{|K_\tau(p,q)|}{\beta_\tau \lambda_\tau^{k/2}} - \sum_{m\geq\tau+1} \frac{|K_m(p,q)|}{\beta_m \lambda_m^{k/2}}\right] := \gamma_k > 0. \tag{71}$$

Nach (69) gibt es zu jedem k ein $x_0 = x_0(k)$, sodass die Abschätzung

$$|S^k(\Lambda, p, q, x)| \leq \pi\gamma_k e^x \quad \text{für alle} \quad x\geq x_0 \tag{72}$$

erfüllt ist. Für $x_\nu = x_\nu(k) := (\pi/2 + \nu\pi - \psi_\tau k)/\beta_\tau$ ist $\sin(\beta_\tau x_\nu + \psi_\tau k) = (-1)^\nu$, $\nu = 0, 1, \ldots$ und aus (70), (71) und (72) folgt für $k\geq h_0$

$$R^{k-1}(\Lambda, p, q, x_\nu)\begin{cases} \leq -\pi\gamma_k e^{x_\nu}, \text{ für alle } \nu \text{ mit } (-1)^\nu K_\tau(p,q)<0, & x_\nu \geq x_0. \\ \geq \pi\gamma_k e^{x_\nu}, \text{ für alle } \nu \text{ mit } (-1)^\nu K_\tau(p,q)>0, & x_\nu \geq x_0. \end{cases} \tag{73}$$

Das bedeutet, dass $R^k(\Lambda, p, q, x)$, $k\geq k_0-1$, das Vorzeichen wechselt auf jedem x-Intervall $x_0\leq x_\nu \leq x \leq x_{\nu+1}$ der Länge π/β_τ. Somit beweist (73) sowohl (15) als auch (17) für $k\geq k_0-1$.

Genau wie in [16] folgert man mit der dort gegebenen Verallgemeinerung des Satzes von Rolle aus (73) und (28) die Gültigkeit von (17) für alle $k \geq 0$, und auch (15) ergibt sich für alle $k \geq 0$ aus (73) unter Beachtung von (29).

(xii) *Vorbereitungen zum Beweis von* (16). Aus Satz 2 lassen sich die Hauptteile von $G^k(\Lambda, p, q, s)$ in den Polen $s \in P$ bestimmen. Insbesondere ersieht man:

$$\left.\begin{array}{l} G^k(\Lambda, p, q, s) \text{ ist analytisch für } \sigma > 0,\ k \geq 0, \text{ bis auf einfache Pole bei} \\ s_n^{\pm} := 1 \pm \alpha_n, \text{ mit Residuen } v_{nk}^{\pm} = \pi\Gamma(k+1)K_n(p, q)/(\pm\alpha_n)(1 \pm \alpha_n)^{k+1}, \\ \text{für } n = 0, 1, \ldots, s_n^{\pm} \neq 1. \\ \text{Falls 1 Eigenwert ist, hat } G^k(\Lambda, p, q, s) \text{ zudem einen doppelten} \\ \text{und einen einfachen Pol bei } s = 1 \text{ mit den Hauptteilen} \\ 2\pi\Gamma(k+1)K^*(p, q)/(s-1)^2, \text{ respektive } -2\pi\Gamma(k+2)K^*(p, q)/(s-1). \end{array}\right\} \quad (74)$$

Man definiert

$$H^k(\Lambda, p, s) := \int_0^{\infty} R^k(\Lambda, p, p, x)e^{-xs}\,dx, \qquad \sigma > 3/2, \qquad k \geq 0. \quad (75)$$

Vorerst für $\sigma > 2$ ist dann

$$H^k(\Lambda, p, s) = G^k(\Lambda, p, p, s) - \int_0^{\infty} Q^k(\Lambda, p, p, x)e^{-xs}\,dx, \qquad k \geq 0.$$

Setzt man in dieser Gleichung den expliziten Ausdruck (6) für $Q^k(\Lambda, p, p, x)$ ein und führt die Integration durch, so ergibt sich zusammen mit (74):

$$\begin{array}{l} H^k(\Lambda, p, s),\ k \geq 0, \text{ ist analytisch für } \sigma > 1 - \theta,\ (0 < \theta := \min_{n<\tau} \alpha_n \leq 1), \\ \text{bis auf einfache Pole bei } s_n^{\pm} := 1 \pm \alpha_n \text{ mit Residuen } v_{nk}^{\pm} = \pi\Gamma(k+1) \times \\ K_n(p, p)/(\pm\alpha_n)(1 \pm \alpha_n)^{k+1} \text{ für alle } n \geq \tau. \end{array} \quad (76)$$

Als weitere Vorbereitung zum Beweis von (16) wählt man eine beliebige ganze Zahl $M > 0$ und betrachtet die Funktionen

$$V(x) := 2\cos^2(x/2) = 1 + \cos x = 1 + \tfrac{1}{2}(e^{ix} + e^{-ix}), \qquad W(x) := \prod_{j=\tau}^{\tau+M} V(\beta_j x).$$

Die Funktion $W(x)$ ist von der Form $W(x) := \sum_{j=0}^{Q} v_j(e^{iw_j} + e^{-iw_j})$. Dabei ist $v_j = (1/2)^{\nu}$, $\nu \geq 0$, ν ganz, also $v_j > 0$, $j = 0, 1, \ldots, Q$ und w_j durchläuft die $Q := (3^{M+1} + 1)/2$ Zahlen $r_0\beta_\tau + r_1\beta_{\tau+1} + \cdots + r_M\beta_{\tau+M}$, $r_j = 0$ oder ± 1, wobei der erste von 0 verschiedene Term der Folge $r_0, r_1, \ldots, r_M$ jeweils gleich $+1$ ist.

Die w_j seien so numeriert, dass $w_j = \beta_{\tau+j}$ ist, $j = 0, 1, \ldots, M$ und $w_Q = 0$. Dann ist $v_j = 1/2$ für $j = 0, 1, \ldots, M$ und

$$W(x) = 1 + \frac{1}{2}\sum_{j=0}^{M}\{e^{i\beta_{\tau+j}x} + e^{-i\beta_{\tau+j}x}\} + \sum_{j=M+1}^{Q-1} v_j\{e^{iw_jx} + e^{-iw_jx}\},$$

$$\text{mit} \quad w_j \neq 0, \quad j = M+1, M+2, \ldots, Q-1. \tag{77}$$

(xiii) *Beweis von* (16). Man will zeigen, dass unter der Annahme

$$R^k(\Lambda, p, p, x) = O(e^{(1+\varepsilon)x}), \qquad k \geq 0, \quad \text{für jedes} \quad \varepsilon > 0, \tag{78}$$

gilt

$$\liminf_{x\to\infty} R^k(\Lambda, p, p, x)/e^x = -\infty, \qquad 0 \leq k < 1. \tag{79}$$

Sei im ganzen Abschnitt (xiii) $0 \leq k < 1$. Mit der unter (xii) betrachteten Funktion W definiert man

$$L = L(\Lambda, p, k) := \lim_{\sigma \downarrow 1} (\sigma - 1)\int_0^\infty e^{-\sigma x} R^k(\Lambda, p, p, x) W(x)\, dx.$$

Man wird folgende Ungleichungen herleiten:

$$\liminf_{x\to\infty} R^k(\Lambda, p, p, x)/e^x \leq L \leq -\pi\Gamma(k+1)c_2 \sum_{j=\tau}^{\tau+M} K_j(p, p)/\beta_j\lambda_j^{(k+1)/2},$$

$$c_2 \text{ positiv und unabhängig von } M. \tag{80}$$

Für $n \geq \tau$ ist $\beta_n = (\lambda_n - 1)^{1/2} \sim \lambda_n^{1/2}$, $n \to \infty$, und nach [10, Seite 408] ist die Konvergenzabszisse der Dirichletreihe $\sum_{n\geq 1} K_n(p, p)/\lambda_n^s$ gleich 3/2. Somit geht die Summe auf der rechten Seite von (80) gegen ∞ für $k < 1$ und $M \to \infty$. Da M eine beliebige positive ganze Zahl war, ist mit (80) auch (79) bewiesen. Das bedeutet (78) impliziert (79), was zu zeigen war. Die erste Ungleichung in (80) ist eine Folge der offensichtlichen Tatsache, dass $W(x) \geq 0$ ist für reelles x. Beachtet man nämlich noch, dass für beliebiges festes ξ, $\xi > 0$, gilt

$$\left|(\sigma - 1)\int_0^\xi e^{-\sigma x} R^k(\Lambda, p, p, x) W(x)\, dx\right|$$

$$\leq (\sigma - 1)\xi \max_{0\leq x\leq\xi} |e^{-\sigma x} R^k(\Lambda, p, p, x) W(x)| \to 0, \qquad \sigma \downarrow 1,$$

so erhält man

$$L \geq \inf_{x \geq \xi} \frac{R^k(\Lambda, p, p, x)}{e^x} \lim_{\sigma \downarrow 1} (\sigma - 1) \int_{\xi}^{\infty} e^{-(\sigma-1)x} W(x)\, dx. \tag{81}$$

Aus der Darstellung (77) von $W(x)$ folgt aber, dass für jedes feste ξ der in (81) auftretende Limes gleich 1 ist. Da ξ beliebig positiv war, ist damit die erste Ungleichung unter (80) bewiesen.

Die andere Behauptung in (80) verifiziert man, indem man sich überlegt, dass mit (77) und der Definition (75) von $H^k(\Lambda, p, s)$ gilt

$$L = \sum_{j=0}^{M} \tfrac{1}{2} \lim_{\sigma \downarrow 1} (\sigma - 1)\{H^k(\Lambda, p, \sigma - \alpha_{j+\tau}) + H^k(\Lambda, p, \sigma + \alpha_{j+\tau})\}$$
$$+ \sum_{j=M+1}^{Q-1} v_j \lim_{\sigma \downarrow 1} (\sigma - 1)\{H^k(\Lambda, p, \sigma - iw_j) + H^k(\Lambda, p, \sigma + iw_j)\}.$$

Beachtet man (76), so ergibt sich daraus

$$L = \frac{\pi}{2} \Gamma(k+1) \sum_{j=\tau}^{\tau+M} \frac{K_j(p, p)}{\alpha_j} \{-1/(1-\alpha_j)^{k+1} + 1/(1+\alpha_j)^{k+1}\}$$
$$+ \frac{\pi}{2} \Gamma(k+1) \sum_{w_j = \beta_m} v_j \frac{K_m(p, p)}{\alpha_m} \{-1/(1-\alpha_m)^{k+1} + 1/(1+\alpha_m)^{k+1}\}. \tag{82}$$

Dabei bezeichnet $\sum_{w_j=\beta_m}$ die Summe über alle j aus der Menge $\{M+1, M+2, \ldots, Q\}$, für die $w_j = \beta_m$ ist, für irgend ein $m \geq \tau$. Mit der unter (xi) eingeführten Bezeichnung $\arg(1-\alpha_n) := \psi_n$, $n \geq \tau$, wird (82) zu

$$L = \pi \Gamma(k+1) \sum_{j=\tau}^{\tau+M} \frac{K_j(p, p)}{\beta_j \lambda_j^{(k+1)/2}} \sin((k+1)\psi_j)$$
$$+ \pi \Gamma(k+1) \sum_{w_j = \beta_m} v_j \frac{K_m(p, p)}{\beta_m \lambda_m^{(k+1)/2}} \sin((k+1)\psi_m). \tag{83}$$

Da $\beta_j \geq \beta_\tau > 0$ ist für $j \geq \tau$ gilt $0 > \psi_\tau := \arg(1-\alpha_\tau) \geq \arg(1-\alpha_j) := \psi_j \geq -\pi/2$, $j \geq \tau$, also wegen $0 \leq k < 1$:

$$0 > \psi_\tau(k+1) \geq \psi_j(k+1) \geq -\frac{\pi}{2}(k+1) > -\pi, \qquad j \geq \tau.$$

Somit existiert ein nur von k abhängiges $c_2>0$, sodass gilt

$$0>-c_2>\sin((k+1)\psi_j), \text{ gleichmässig in } j\geq\tau.$$

Unter (xii) hat man sich überlegt, dass alle $v_j>0$ sind, $j=0,1,\ldots,Q$. Also, da auch die $K_j(p,p)=|\phi_j(p)|^2>0$ sind, folgt aus (83):

$$L\leq-\pi\Gamma(k+1)c_2\sum_{j=\tau}^{\tau+M}\frac{K_j(p,p)}{\beta_j\lambda_j^{(k+1)/2}}.$$

Das beweist (80) und damit (16).

(xiv) *Beweis von* (19) *und Satz* 6. Ist $f(s)$ eine analytische Funktion von s, so setzt man $(d/ds)^m f(s):=f^{(m)}(s)$, $m=0,1,\ldots$. Für $s\notin P$, $n=0,1,\ldots$ und $m\geq 0$, m ganz, definiert man $\gamma(s,n,m):=\{1/(s^2-(\alpha_n+1)^2)(s^2-(\alpha_n-1)^2)\}^{(m)}$. Dann ist für $s\notin P$:

$$\gamma(s,n,m)=\sum_{\eta=0}^{m}\sum_{\mu=0}^{\eta}\sum_{\nu=0}^{m-\eta}\binom{m}{\eta}\binom{\eta}{\mu}\binom{m-\eta}{\nu}\left\{\frac{1}{s-\alpha_n-1}\right\}^{(\mu)}\left\{\frac{1}{s+\alpha_n+1}\right\}^{(\eta-\mu)}$$
$$\left\{\frac{1}{s+\alpha_n-1}\right\}^{(\nu)}\left\{\frac{1}{s-\alpha_n+1}\right\}^{(m-\eta-\nu)}. \quad (84)$$

Unter Beachtung von Satz 2 erhält man

$$|(d/ds)^m G^0(\Lambda,p,q,s)|^2=(8\pi)^2\sum_{n,j=0}^{\infty}K_n(p,q)K_j(p,q)\gamma(s,n,m)\bar{\gamma}(s,j,m),$$
$$s\notin P. \quad (85)$$

Für festes $s\notin P$ ist $\gamma(s,n,m)=O(1/|\alpha_n|^4)$, $m\geq 0$, gleichmässig in $n=0,1,\ldots$, sodass nach (38) die Doppelreihe unter (85) gleichmässig konvergiert in $(p,q)\in\mathcal{F}\times\mathcal{F}$. Mit der Orthogonalitätsrelation (1) folgt deshalb

$$\int_{\mathcal{F}}|(d/ds)^m G^0(\Lambda,p,q,s)|^2\,d\omega_q=(8\pi)^2\sum_{n=0}^{\infty}K_n(p,p)\,|\gamma(s,n,m)|^2, \qquad s\notin P. \quad (86)$$

Sei $0<\xi<1$. Aus Formel (84) ersieht man, dass für beliebige $\delta>0$, $\kappa>0$ und

gleichmässig in σ mit $|\sigma \pm 1| \geq \kappa$ gilt

$$|\gamma(s, n, m)|^2 = \begin{cases} O(\beta_n^{-3-\delta} |t|^{-1+\delta} |t|^{-2\xi(m+2)}), \text{ gleichmässig in } n \text{ mit} \\ \qquad\qquad \beta_n \leq |t| - |t|^{\xi},\ \beta_n \geq |t| + |t|^{\xi}. \\ O(|t|^{-4}), \text{ gleichmässig in allen übrigen } n. \end{cases}$$

Man setzt diese Abschätzungen in (86) ein und wendet (38), sowie (4) an. Setzt man dann $\xi = 1/(2m+5)$, so ist $2\xi(m+2) = 1-\xi$. Wählt man schliesslich bei gegebenem ω, $\omega > 1/(4m+10) = \xi/2$, den Parameter δ derart, dass $0 < \delta < 2\omega - \xi$ ist, so ergibt sich

$$\begin{aligned} &\int_{\mathcal{F}} |(d/ds)^m G^0(\Lambda, p, q, s)|^2 \, d\omega_q \\ &= O(|t|^{-1+\delta-2\xi(m+2)}) + O(|t|^{-4}) \sum_{|t|-|t|^{\xi} \leq \beta_n \leq |t|+|t|^{\xi}} K_n(p, p) \\ &= O(|t|^{-1+\delta-2\xi(m+2)}) + O(|t|^{-2+\xi}) = O(|t|^{-2+\delta+\xi}) \\ &= O(|t|^{-2+2\omega}), \text{ gleichmässig in } \sigma \text{ mit } |\sigma \pm 1| \geq \kappa. \end{aligned}$$

Damit ist auch (19) bewiesen. Durch wiederholte partielle Integration lässt sich mit Hilfe von (19) für fast alle $q \in \mathcal{F}$ zeigen:

$$\int_0^U G^0(\Lambda, p, q, \sigma + it) e^{ity} \, dt = O(e^{\varepsilon y}), \quad \begin{array}{l} \text{für beliebig vorgegebenes } \varepsilon > 0 \\ \text{und für feste } \xi > 1,\ \sigma > 1,\ \sigma \notin P, \\ \text{gleichmässig in } 0 \leq U \leq 2e^{2\xi y}. \end{array}$$

Mit dieser Formel und (19) sind die Voraussetzungen des nachfolgenden Hilfssatzes verifiziert für die Wahl $A(x) := N^0(\Lambda, p, q, x)$, $f(s) := G^0(\Lambda, p, q, s)$, $\sigma_0 = \beta_1 := 2$, $\gamma > 1$, $\gamma \notin P$, $s_j := 1 + \alpha_j$, $j = 0, 1, \ldots, \tau$, $w(y) := e^{\varepsilon y}$, ε beliebig positiv. Beachtet man (74) und die Definition (6) von $Q^0(\Lambda, p, q, x)$, so ergibt sich Satz 6 für $k = 0$ aus dem Hilfssatz und damit für $k > 0$ aus (34).

V. Ein Hilfssatz

Der Satz von Wiener–Ikehara gibt das asymptotische Wachstum einer reellwertigen, nicht fallenden Funktion an, unter der Voraussetzung, dass deren Laplacetransformierte in der Konvergenzabszisse σ_0 einen einfachen Pol mit

bekanntem Residuum besitzt und sonst analytisch ist in der s-Halbebene $\operatorname{Re} s \geq \sigma_0$.

Oft ist aber von der Laplacetransformierten mehr bekannt in dem Sinne, dass sie über σ_0 hinaus meromorph fortgesetzt werden kann, eventuell mit mehreren, auch mehrfachen Polen. Im folgenden soll eine Verallgemeinerung des Satzes von Wiener-Ikehara bewiesen werden, die sich auf diese Situation bezieht.

Hilfssatz. Sei die Funktion $A(x)$ definiert für $x \geq 0$, nicht negativ, monoton wachsend und das Integral $\int_0^\infty A(x)e^{-xs}\,dx$, $s := \sigma + it$, konvergiere für $\sigma > \sigma_0$ gegen die Funktion $f(s)$.

Sei $f(s)$ analytisch für $\sigma \geq \gamma$, bis auf (a_j+1)-fache Pole bei $s = s_j$, $s_j := \beta_j + i\gamma_j$, mit den Hauptteilen $v_j/(s-s_j)^{a_j+1}$, $a_j \geq 0$, a_j ganz, $j = 1, 2, \ldots, q$ und $\gamma < \beta_q \leq \beta_{q-1} \leq \cdots \leq \beta_1 = \sigma_0$. Sei

$$R(x) := A(x) - \sum_{j=1}^{q} \frac{v_j}{a_j!} x^{a_j} e^{s_j x}.$$

Ist mit einer positiven, nicht fallenden Funktion w und für beliebiges festes $\xi > \beta_1 - \gamma$

(a) $f(\sigma + it) = O(1)$, gleichmässig in $\sigma \in [\gamma, \beta_1 + 1]$,

(b) $\int_0^U f(\gamma + it)e^{ity}\,dt = O(w(y))$ für $y \to \infty$, gleichmässig in $0 \leq U \leq 2e^{2\xi y}$,

so lässt sich folgern

$$R(x) = O(e^{\gamma x} w(x)). \tag{87}$$

Bemerkungen. (i) Weil A monoton wachsend vorausgesetzt wurde, ist $\sigma_0 = \beta_1$.

(ii) Da $f(s)$ reell ist für reelles $s > \sigma_0$, folgt $f(s) = \bar{f}(\bar{s})$ für alle $s \neq s_j$ mit $\operatorname{Re} s \geq \gamma$. Dies impliziert, dass die Summe in der Definition von $R(x)$ – und damit $R(x)$ selbst – reell ist.

(iii) Die in der Voraussetzung (b) auftretende Funktion w wird im allgemeinen von der Grössenordnung der oberen Grenze des zu betrachtenden Integrals abhängen und die Aussage des Satzes kann unter Umständen besser werden, je kleiner diese Grössenordnung gewählt werden darf. Ersetzt man die beim Beweis auftretende Hilfsfunktion

$$h_1(x) := \begin{cases} 1 - |x|, & |x| \leq 1 \\ 0, & \text{sonst} \end{cases}$$

durch die n-fache Faltung

$$h_n(x) := (h_1 * h_1 * \cdots * h_1)(x) \ (n \text{ Faktoren}),$$

$$(h_1 * h_1)(x) := \int_{-\infty}^{\infty} h_1(v) h_1(x-v)\, dv,$$

so ist $h_n(t/2n\lambda) = O(1)$, gleichmässig in $t \in \mathbf{R}$ für alle $\lambda > 0$, $n = 1, 2, \ldots$ und es lässt sich zeigen, dass gilt:

Zu beliebig festem $\xi > \beta_1 - \gamma$ existiert ein ganzes $j \geq 1$, sodass die Bedingungen

(a) $f(s) = O(1)$, gleichmässig in $\sigma \in [\gamma, \beta_1 + 1]$,

(b) $\int_0^{2e^{\xi y}} f(\gamma + it) h_j(t/2je^{\xi y}) e^{ity}\, dt = O(w(y))$ für $y \to \infty$,

implizieren

$$R(x) = O(e^{\gamma x} w(x)).$$

Das bedeutet, die Grössenordnung der oberen Grenze des unter (b) zu betrachtenden Integrals kann in dieser Formulierung des Satzes halbiert werden. Die mit dem Hilfssatz hergeleitete Aussage (18) kann allerdings auf Grund dieser Feststellung nicht verbessert werden.

(iv) Indem man das Integral der Funktion A betrachtet und durch Anwendung der Methode der endlichen Differenzen, lässt sich aus dem Hilfssatz beweisen, dass allein die Voraussetzung

(a) $f(\sigma + it) = O(1)$, gleichmässig in $\sigma \in [\gamma, \beta_1 + 1]$,

impliziert

$$R(x) = O(e^{(\beta_1 - \gamma)x/2} x^{(\nu+1)/2}), \qquad \nu := \max_{\beta_j = \beta_1} a_j.$$

Beweis des Hilfssatzes. Abschätzung (87) soll hergeleitet werden durch Uebertragung eines Beweises von Landau–Bochner für den Satz von Wiener–Ikehara (man siehe dazu zum Beispiel [3]).

Man setzt $R^*(x) := R(x) e^{-\gamma x}$. In einem ersten Schritt will man zeigen, dass gilt

$$R^*(x) \leq c_3 w(x), \qquad x \geq 1, \ c_3 \text{ unabhängig von } x. \tag{88}$$

Setzt man

$$g(s) := f(s) - \sum_{j=1}^{q} v_j/(s-s_j)^{a_j+1},$$

so ist $g(s)$ analytisch für $\sigma \geq \gamma$ und

$$g(s) = \int_0^\infty R^*(x) e^{-x(s-\gamma)}\, dx, \qquad \sigma > \beta_1. \tag{89}$$

Sei $\eta > \beta_1$ und $\mu := \eta - \gamma$. Mit der Formel

$$\frac{1}{2}\int_{-2\lambda}^{2\lambda} \{1-|t|/2\lambda\} e^{itv}\, dt = \frac{\sin^2(\lambda v)}{\lambda v^2}, \qquad \lambda > 0,\ v \text{ reell},$$

erhält man für $\lambda \geq 1$, $y \geq 1$ die Gleichung

$$\frac{1}{2}\int_{-2\lambda}^{2\lambda} g(\eta+it)\{1-|t|/2\lambda\} e^{(it+\mu)y}\, dt = \int_0^\infty R^*(x) e^{\mu(y-x)} \frac{\sin^2(\lambda(y-x))}{\lambda(y-x)^2}\, dx, \tag{90}$$

indem man auf der linken Seite die Beziehung (89) einsetzt und sich überlegt, dass die Integrationsreihenfolge vertauscht werden darf. Man setzt $\lambda = \lambda(y) := e^{2\xi y}$, mit dem in der Voraussetzung b) auftretenden, festen $\xi > \beta_1 - \gamma > 0$. Für $y \geq 1$ ist dann $\lambda > 1$. Sei

$$T(y) := \operatorname{Im} \int_\eta^{\eta+2\lambda i} g(s)\left\{1 - \frac{s-\eta}{2\lambda i}\right\} e^{(s-\gamma)y}\, ds. \tag{91}$$

Wegen $g(s) = \bar{g}(\bar{s})$ für $\sigma \geq \gamma$, ist die linke Seite von (90) gleich $T(y)$ und indem man auf der rechten Seite $u := \lambda(y-x)$ substituiert, ergibt sich

$$T(y) = \int_{-\infty}^{\lambda y} R^*(y - u/\lambda) e^{\mu u/\lambda} \frac{\sin^2 u}{u^2}\, du. \tag{92}$$

Definiert man

$$P(x) := A(x)e^{-\gamma x} - R^*(x) = \sum_{j=1}^{q} \frac{v_j}{a_j!} x^{a_j} e^{(s_j-\gamma)x}, \tag{93}$$

so erhält man aus (92)

$$T(y)+\int_{-\infty}^{\lambda y} P(y-u/\lambda)e^{\mu u/\lambda}\frac{\sin^2 u}{u^2}\,du = \int_{-\infty}^{\lambda y} A(y-u/\lambda)e^{-\gamma y+\eta u/\lambda}\frac{\sin^2 u}{u^2}\,du. \quad (94)$$

Nun lassen sich die Voraussetzungen über A ausnützen. Wegen $A(x)\geq 0$, $x\geq 0$, ist der ganze Integrand auf der rechten Seite von (94) nicht negativ. Da zudem A monoton wachsend ist, folgt aus (94) für beliebige a, $1\leq a\leq \lambda y$:

$$T(y)+\int_{-\infty}^{\lambda y} P(y-u/\lambda)e^{\mu u/\lambda}\frac{\sin^2 u}{u^2}\,du \geq \int_{-a}^{a} A(y-a/\lambda)e^{-\gamma y+\eta u/\lambda}\frac{\sin^2 u}{u^2}\,du. \quad (95)$$

Man wählt $a=a(y):=\lambda(y)^{1/2}=e^{\xi y}$. Für $y\geq 1$ ist dann $1\leq a\leq \lambda y$ und durch beidseitige Subtraktion des Terms

$$\int_{-a}^{a} P(y-1/a)e^{\eta u/a^2-\gamma/a}\frac{\sin^2 u}{u^2}\,du$$

erhält man aus (95)

$$R^*(y-1/a)\leq \frac{1}{S_1(y)}\{T+S_2+S_3\}(y), \quad (96)$$

mit

$$S_1(y):=e^{-\gamma/a}\int_{-a}^{a} e^{\eta u/a^2}\frac{\sin^2 u}{u^2}\,du>0, \qquad a(y)>0,$$

$$S_2(y):=\int_{-a}^{a}\{P(y-u/a^2)-e^{\gamma(u/a^2-1/a)}P(y-1/a)\}e^{\mu u/a^2}\frac{\sin^2 u}{u^2}\,du,$$

$$S_3(y):=\left\{\int_{-\infty}^{-a}+\int_{a}^{a^2 y}\right\}P(y-u/a^2)e^{\mu u/a^2}\frac{\sin^2 u}{u^2}\,du.$$

Man wählt $\eta>\beta_1$ so, dass $\beta_1-\gamma<\eta-\gamma=\mu<\xi$ ist. Die rechte Seite von (96) soll nun in Funktion von y nach oben abgeschätzt werden. Die Funktionen $S_j(y)$, $j=1,2,3$ sind unabhängig von $f(s)$ definiert. Zu ihrer Abschätzung macht man somit keinen Gebrauch von den Voraussetzungen über $f(s)$. Offenbar ist

$$\lim_{y\to\infty} S_1(y)=\pi. \quad (97)$$

Mit der aus der Definition (93) ersichtlichen Abschätzung

$$P(y-u/a^2)-e^{\nu(u/a^2\pm 1/a)}P(y\pm 1/a)=O(e^{(\beta_1-\gamma)y}y^\nu/a)$$
$$=O(y^\nu e^{(\beta_1-\gamma-\xi)y}) \quad \text{für} \quad -a\le u\le a, \quad \nu:=\max_{\beta_i=\beta_1} a_j, \tag{98}$$

folgt wegen $\beta_1-\gamma<\xi$:

$$S_2(y)=o(1) \quad \text{für} \quad y\to\infty. \tag{99}$$

Es ist auch $\beta_1-\gamma<\mu<\xi$ und nach (93) $P(x)=O(x^\nu e^{(\beta_1-\gamma)x})=O(e^{\mu x})$, gleichmässig in $0\le x<\infty$. Also gilt $S_3(y)=O(e^{\mu y}/a)=O(e^{(\mu-\xi)y})$ und somit

$$S_3(y)=o(1) \quad \text{für} \quad y\to\infty. \tag{100}$$

Da $g(s)$ analytisch ist für $\sigma\ge\gamma$, lässt sich die unter (91) definierte Funktion $T(y)$ folgendermassen darstellen:

$$T(y)=\operatorname{Im}\int_{W_1+W_2+W_3} g(s)\left\{1-\frac{s-\eta}{2\lambda i}\right\}e^{(s-\gamma)y}\,ds,$$

mit $W_1:=\{s\mid \eta\ge s\ge\gamma\}$, $W_2:=\{s=\gamma+it\mid 0\le t\le 2\lambda\}$, $W_3:=\{s=\sigma+2\lambda i\mid \gamma\le\sigma\le\eta\}$, $\lambda=\lambda(y):=e^{2\xi y}$. Auf W_1 ist s reell, also $g(s)e^{(s-\gamma)y}$ reell und $\operatorname{Im}\{1-(s-\eta)/2\lambda i\}=O(1/\lambda)$. Auf W_3 ist $\{1-(s-\eta)/2\lambda i\}=O(1/\lambda)$. Zudem ist mit $f(s)$ – wie vorausgesetzt unter (a) – auch $g(s)=O(1)$, gleichmässig in $\sigma\in[\gamma,\beta_1+1]$. Somit sind die Beiträge der Integrale über W_1 und W_3 je gleich $O(e^{\mu y}/\lambda)=O(e^{(\mu-\xi)y})=o(1)$ für $y\to\infty$. Weiter ist

$$\operatorname{Im}\int_{W_2} g(s)\left\{1-\frac{s-\eta}{2\lambda i}\right\}e^{(s-\gamma)y}\,ds=\operatorname{Re}\int_0^{2\lambda} g(\gamma+it)\left\{1-\frac{-\mu+it}{2\lambda i}\right\}e^{ity}\,dt$$
$$=O(w(y))+O(1).$$

Die O-Abschätzung ergibt sich durch partielle Integration unter Beachtung der Voraussetzung (b) und der Definition von $g(s)$. Damit wird, da w nicht fallend vorausgesetzt ist

$$T(y)=O(w(y)), \tag{101}$$

und (97), sowie (99) bis (101), eingesetzt in (96) ergeben die zu verifizierende

Beziehung (88). Der Hilfssatz ist bewiesen, wenn noch gezeigt wird

$$R^*(x) \geq O(w(x)). \tag{102}$$

Sei wie oben $\eta > \beta_1$ so, dass $\beta_1 - \gamma < \mu := \eta - \gamma < \xi$ ist. Es kann $w(y) = O(e^{\mu y})$ angenommen werden, da andernfalls nichts zu beweisen ist. Aus (92) folgt, wenn man für $y \geq 1$ wieder setzt $a = a(y) := \lambda(y)^{1/2} = e^{\xi y}$ und (88) beachtet

$$\begin{aligned} T(y) &\leq c_3 \left\{ \int_{-\infty}^{-a} + \int_a^{a^2 y} \right\} w(y - u/a^2) e^{\mu u/a^2} \frac{\sin^2 u}{u^2}\, du \\ &\quad + \int_{-a}^{a} R^*(y - u/a^2) e^{\mu u/a^2} \frac{\sin^2 u}{u^2}\, du \\ &= O(e^{\mu y}/a) + \int_{-a}^{a} R^*(y - u/a^2) e^{\mu u/a^2} \frac{\sin^2 u}{u^2}\, du. \end{aligned}$$

Wegen $R^*(x) = A(x) e^{-\gamma x} - P(x)$ ist also mit (101) und $O(e^{\mu y}/a) = O(e^{(\mu - \xi) y}) = o(1)$ für $y \to \infty$:

$$\begin{aligned} O(w(y)) + \int_{-a}^{a} P(y - u/a^2) e^{\mu u/a^2} \frac{\sin^2 u}{u^2}\, du &\leq \int_{-a}^{a} A(y - u/a^2) e^{-\gamma y + \eta u/a^2} \frac{\sin^2 u}{u^2}\, du \\ &\leq \int_{-a}^{a} A(y + 1/a) e^{-\gamma y + \eta u/a^2} \frac{\sin^2 u}{u^2}\, du. \end{aligned} \tag{103}$$

Die letzte Ungleichung ist eine Folge davon, dass A nicht negativ und monoton wachsend vorausgesetzt wurde. Auf beiden Seiten der Ungleichung (103) subtrahiert man den Term

$$\int_{-a}^{a} P(y + 1/a) e^{\eta u/a^2 + \gamma/a} \frac{\sin^2 u}{u^2}\, du.$$

Beachtet man (93), sowie (97) und die Tatsache, dass wegen (98) gilt

$$\int_{-a}^{a} \{P(y - u/a^2) - e^{\gamma(u/a^2 + 1/a)} P(y + 1/a)\} e^{\mu u/a^2} \frac{\sin^2 u}{u^2}\, du = o(1) \quad \text{für} \quad y \to \infty,$$

so erhält man die Abschätzung (102). Damit ist der Hilfssatz bewiesen.

LITERATURANGABEN

[1] AVACUMOVIĆ, G. V., *Ueber die Eigenfunktionen auf geschlossenen Riemann'schen Manigfaltigkeiten.* Math. Zeit. *65* (1956), 327–344.
[2] CHANDRASEKHARAN, K. und MINAKSHISUNDARAM, S., *Typical Means.* (Oxford 1952).
[3] CHANDRASEKHARAN, K., Introduction to Analytic Number Theory. Springer 1968.
[4] ——, Arithmetical functions. Springer 1970.
[5] CHANDRASEKHARAN, K. und NARASIMHAN, R., *Hecke's functional equation and arithmetical identities.* Ann. of Math. *74* (1961), 1–23.
[6] ——, *Hecke's functional equation and the average order of arithmetical functions.* Acta Arithm. *6* (1961), 487–503.
[7] ——, *Functional equations with multiple gamma factors and the average order of arithmetical functions.* Ann. of Math., *76* (1962), 93–136.
[8] ——, *The approximate functional equation for a class of zeta-functions.* Math. Ann. *152* (1963), 30–64.
[9] COURANT, R., *Differential and Integral Calculus.* Vol. I, Blackie and Son, second edition.
[10] FRICKER, F., *Ein Gitterpunktproblem im dreidimensionalen hyperbolischen Raum.* Comment. Math. Helv. *43* (1968), 402–416.
[11] HÖRMANDER, L., *The spectral function of an elliptic operator.* Acta Math. *121* (1968), 193–218.
[12] HUBER, H., *Ueber eine Klasse automorpher Funktionen und ein Gitterpunktproblem in der hyperbolischen Ebene.* Comment. Math. Helv. *30* (1956), 20–62.
[13] ——, *Zur analytischen Theorie hyperbolischer Raumformen und Bewegungsgruppen.* Math. Ann. *138* (1959), 1–26. *142* (1961), 385–398, *143* (1961), 463–464.
[14] MINAKSHISUNDARAM, S. und PLEIJEL, A., *Some properties of the Eigenfunctions of the Laplace-operator on Riemannian manifolds.* Canad. Journ. of Math. *1* (1949), 242–256.
[15] SELBERG, A., *On discontinous groups in higher dimensional symmetric spaces.* Contr. to function theory (Tata Institute, Bombay 1960), 147–164.
[16] STEINIG, J., *The changes of sign of certain arithmetical error-terms.* Comment. Math. Helv. *44* (1968), 385–401.
[17] THURNHEER, P., Une généralisation du théorème de Wiener–Ikehara. C.R. Acad. Sc. Paris, t. 290 (1980), 499.
[18] ——, *Le terme de reste dans un problème de réseau hyperbolique.* C.R. Acad. Sc. Paris, t. 290 (1980), 581–583.
[19] TITCHMARSH, E. C., *The theory of functions.* Oxford University Press, 1932.

Eidgenössische Technische Hochschule,
Hg G 66.2
Rämistrasse 101
CH-8092 Zürich

Eingegangen den 21. August 1980

Comment. Math. Helvetici **56** (1981) 272–296 0010–2571/81/002272–25$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

The Lefschetz number of an involution on the space of classes of positive definite quadratic forms

J. ROHLFS[1]

Introduction

It is well known [2: §2] that the set of positive definite quadratic forms on $\mathbf{R}^n$, $n \in \mathbf{N}$, with determinant 1 can be identified with the symmetric space $X = \mathrm{SO}_n(\mathbf{R})\backslash \mathrm{Sl}_n(\mathbf{R})$. Here we use standard notation. Suppose that Γ is a subgroup of finite index in $\mathrm{Sl}_n(\mathbf{Z})$. Then two positive definite quadratic forms $q_1, q_2 : \mathbf{R}^n \to \mathbf{R}$ are called Γ-equivalent if there exists $\gamma \in \Gamma$ such that $q_1 = q_2 \circ \gamma$. The space of Γ-equivalence classes of positive definite quadratic forms with determinant 1 thus corresponds bijectively to X/Γ.

The map $A \mapsto {}^{\tau}A := {}^{t}A^{-1}$ sending a matrix A to the transpose of the inverse matrix induces an involution τ on X. If Γ is τ-stable we have an induced involution on X/Γ. Denote by $H^*(X/\Gamma, \mathbf{Q})$ the singular cohomology with coefficients $\mathbf{Q}$. It is well known that $H^*(X/\Gamma, \mathbf{Q}) = H^*(\Gamma, \mathbf{Q})$ where on the right side we have abstract group cohomology. The map τ induces maps $\tau^i : H^i(\Gamma, \mathbf{Q}) \to H^i(\Gamma, \mathbf{Q})$ with trace $\mathrm{tr}(\tau^i)$. We define the Lefschetz number of τ and Γ by

$$L(\tau, \Gamma) = \sum_{i=0}^{\infty} (-1)^i \,\mathrm{tr}\,(\tau^i).$$

For $\Gamma = \Gamma(m)$, $m \geq 3$, the full congruence subgroup mod m of $\mathrm{Sl}_n(\mathbf{Z})$ we will give an explicit formula for $L(\tau, \Gamma(m))$. If, for example, m and n are odd we get

$$L(\tau, \Gamma(m)) = m^{n(n-1)/2} \prod_{j=1}^{(n-1)/2} \left(\prod_{p \mid m} (1 - p^{-2j}) \right) \zeta(1-2j).$$

Here $\zeta(\ )$ denotes the Riemann zeta function. Hence the Lefschetz number is computable in terms of Bernoulli numbers. The second product has to be taken over all prime divisors of m. Similar formulas (see 5.6) hold in the other cases.

[1] Supported by the Sonderforschungsbereich "Theoretische Mathematik" at the University of Bonn.

Not too much is known about $H^*(\Gamma, \mathbf{Q})$ itself. Some information has been obtained by Harder [7], Borel and Serre [3], Borel [4], Lee and Szczarba [14, 15], Schwermer [20], Lee [16] and Soulé [25].

To prove the formula for the Lefschetz number we proceed as follows. In §1 we observe that the Lefschetz fixpoint formula

$$L(\tau, \Gamma) = \chi((X/\Gamma)^\tau)$$

holds. The right side denotes the Euler–Poincaré characteristic of the fixpoint set $(X/\Gamma)^\tau$ of the operation of τ *on* X/Γ. It is shown in 1.3 that the connected components of $(X/\Gamma)^\tau$ correspond bijectively to the non abelian cohomology $H^1(\mathfrak{g}, \Gamma)$, $\mathfrak{g} = \{1, \tau\}$. The cohomology set is nothing else but a set of certain equivalence classes of unimodular bilinear forms represented by symmetric matrices $\gamma \in \Gamma$. The fixpoint set $F(\gamma)$ determined by such a symmetric matrix γ is the image of the special orthogonal group $\mathrm{SO}(\gamma)(\mathbf{R})$ belonging to γ under some map $\mathrm{SO}(\gamma)(\mathbf{R}) \to \mathrm{Sl}_n(\mathbf{R}) \to X/\Gamma$. We determine $H^1(\mathfrak{g}, \Gamma)$ in §2 and §3. In §4 we compute volumes of some orthogonal groups in a local situation and determine Euler–Poincaré measures in the sense of Serre [23]. In §5 we use the formula of Minkowski–Siegel, the determination of $H^1(\mathfrak{g}, \Gamma)$ and the computation of the local volumes to obtain the final formula.

Generalisations of this procedure to get invariants of $H^*(\Gamma, \mathbf{Q})$ for other arithmetic groups Γ are possible. For galois-operations on Γ see [19]. Our result supports the idea that there should be a lifting from some automorphic forms of orthogonal groups to automorphic forms of the special linear group.

§0. Notation

0.1. Suppose that $f: M \to M$ is a map and M a set. Then we denote the fixpoint set of f by $M^f = \{m \in M / f(m) = m\}$. Let $|M|$ be the cardinality of M. The symbols $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{C}$ are used in their usual meaning.

0.2. If $a \in \mathbf{Q}$ then $[a]$ denotes the biggest integer which is smaller than or equal to a. For integers m and n we write $m \mid n$ if m divides n and $m \nmid n$ if m does not divide n. We write $(m, n) = 1$ if m and n are relatively prime, and $m \equiv j \bmod n$ if m is of the form $m = j + k \cdot n$; $j, k \in \mathbf{Z}$.

0.3. If p is a prime number we denote by $\mathbf{Q}_p$ the field of p-adic numbers and by $\mathbf{Z}_p$ the ring of p-adic integers in $\mathbf{Q}_p$. The normalized absolut value $\| \ \|_p$ on $\mathbf{Q}_p$ has the property that $\|p\|_p = p^{-1}$. A typical place of $\mathbf{Q}$ is denoted by v. If v is not finite then $v = \infty$ and $\mathbf{Q}_v = \mathbf{R}$.

0.4. Let $M_n(R)$ be the R-algebra of $n \times n$ matrices with coefficients in the ring R. Denote by $Gl_n(R)$ (resp. $Sl_n(R)$) the group of elements of $M_n(R)$ with invertible determinant in R (resp. determinant 1). For $A \in M_n(R)$ we denote by ${}^tA$ the transpose of A. If $b \in Gl_n(R)$ and $b = {}^tb$ then we define the group of isometries with determinant 1 of the bilinear form b by $SO(b)(R) = \{A \in Sl_n(R) / {}^tAb^{-1}A = b^{-1}\}$. In order to simplify the notation later on we use in this definition b^{-1} instead of b. If $b = E$ is the identity matrix we write $SO_n(R)$ instead of $SO(E)(R)$.

0.5. We suppose that a finite group $\mathfrak{g}$ operates from the left on a group G by group automorphisms. For $\sigma \in \mathfrak{g}$ the operation is written in the form $g \mapsto {}^\sigma g$, $g \in G$. In [21: Chap. I §5] the first non commutative cohomology set $H^1(\mathfrak{g}, G)$ is explained. If $\gamma = \{\gamma_\sigma\}$, $\sigma \in \mathfrak{g}$, is a cocycle for $H^1(\mathfrak{g}, G)$ the group G together with a new twisted operation of $\mathfrak{g}$ given by $g \mapsto \gamma_\sigma {}^\sigma g \gamma_\sigma^{-1}$, $g \in G$ is denoted by ${}^\gamma G$. How to determine fibres of certain maps between cohomology sets is described in [21: Chap. I §5].

§1. Fixpoints of τ and non commutative cohomology

We set ${}^\tau A = {}^tA^{-1}$ for $A \in Sl_n(\mathbf{R})$. The map τ is just the standard Cartan involution with respect to $SO_n(\mathbf{R}) \subset Sl_n(\mathbf{R})$. Let X be the symmetric space $SO_n(\mathbf{R}) \backslash Sl_n(\mathbf{R})$. Then τ operates as an isometry on X. Here X is endowed with the right invariant Riemannian metric determined by the Cartan involution and the Killing form belonging to $Sl_n(\mathbf{R})$ [8: p. 159].

Suppose that $\Gamma \subset Sl_n(\mathbf{Z})$ is a torsion free τ-stable arithmetic subgroup. For example we may take $\Gamma = \Gamma(m)$, $m \geq 3$, a full congruence subgroup mod m of $Sl_n(\mathbf{Z})$ [17: p. 203, p. 211]. Then X/Γ is a Riemannian manifold on which τ operates isometrically.

In this chapter we give a parametrisation of the connected components of the fixpoint set $(X/\Gamma)^\tau$ by certain classes of integral bilinear forms and show that the fixpoint components are locally symmetric spaces belonging to some orthogonal groups.

Some of the results and proofs of this paragraph extend to finite groups acting on a semisimple algebraic group and stabilizing an arithmetic subgroup. For galois actions see [19].

1.1. We define $\mathfrak{g} = \{1, \tau\}$ and consider a cocycle $b = (1, b)$ for the non commutative cohomology set $H^1(\mathfrak{g}, \Gamma)$. The cohomology class determined by b will be again denoted by b. By definition $b = (1, b)$, $b \in \Gamma$, is a cocycle if $b \cdot {}^tb^{-1} = b \cdot {}^\tau b = 1$, i.e., $b = {}^tb$. Two cocycles $(1, b)$ and $(1, d)$ are equivalent if there exists a $c \in \Gamma$

with $d = c^{-1} \cdot b \cdot {}^{\tau}c$, i.e., $c \cdot d \cdot {}^{t}c = b$. If for example $\Gamma = \mathrm{Sl}_n(\mathbf{Z})$ then we get a bijection of $H^1(\mathfrak{g}, \Gamma)$ with the set of $\mathrm{Sl}_n(\mathbf{Z})$-equivalence classes of integral unimodular bilinear forms as follows. To every cocycle $(1, b)$ we associate the bilinear form $b(x, y) = {}^{t}x \cdot b \cdot y$. Here $x, y \in \mathbf{Z}^n$ are considered as row vectors.

1.2. We can construct points in $(X/\Gamma)^{\tau}$ as follows. If $b = (1, b)$ is a cocycle for $H^1(\mathfrak{g}, \Gamma)$ we define by ${}^{\bar{\tau}}A = b \cdot {}^{\tau}A \cdot b^{-1}$ a new operation of $\bar{\tau}$ on $\mathrm{Sl}_n(\mathbf{R})$ and on Γ. For $x \in X$ we define ${}^{\bar{\tau}}x = {}^{\tau}x \cdot b^{-1}$. For $x \in X$ and $g \in \mathrm{Sl}_n(\mathbf{R})$ we then have ${}^{\bar{\tau}}(x \cdot g) = {}^{\bar{\tau}}x \cdot {}^{\bar{\tau}}g$. Therefore we have an operation of $\bar{\tau}$ on X/Γ and this operation coincides with the old operation of τ. If $X(b)$ denotes the set of fixpoints of $\bar{\tau}$ on X and $\Gamma(b)$ the set of fixpoints of $\bar{\tau}$ in Γ then the image $F(b)$ of the natural map

$$X(b)/\Gamma(b) \to X/\Gamma$$

clearly lies in $(X/\Gamma)^{\tau}$ and depends only on the cohomology class represented by b. Moreover as Γ is torsion free it is easy to check that we have a natural isomorphism $X(b)/\Gamma(b) \xrightarrow{\sim} F(b)$. We will see that $F(b) \neq \emptyset$ and that all fixpoints arise from this construction.

PROPOSITION 1.3

$$(X/\Gamma)^{\tau} = \bigcup_{b \in H^1(\mathfrak{g}, \Gamma)} F(b).$$

Here we have a disjoint union of non empty and connected sets $F(b)$. The fixpoint component $F(b)$ is locally convex, its fundamental group is isomorphic to $\Gamma(b)$ and its universal cover is contractible.

Proof. We observed that $F(b) \in (X/\Gamma)^{\tau}$. If $\bar{x} \in (X/\Gamma)^{\tau}$ is represented by $x \in X$, then ${}^{\tau}x = x \cdot b$ with some $b \in \Gamma$. Because Γ is torsion free b is uniquely determined. We have ${}^{\tau}({}^{\tau}x) = x = {}^{\tau}(x \cdot b) = x \cdot b \cdot {}^{\tau}b$. Therefore $b \cdot {}^{\tau}b = 1$ and $(1, b)$ is a cocycle for $H^1(\mathfrak{g}, \Gamma)$. If we take another element $y = x \cdot c$, $c \in \Gamma$, representing $\bar{x} \in (X/\Gamma)^{\tau}$, then $c^{-1} \cdot b \cdot {}^{\tau}c$ is the cocycle associated to y. Therefore every $\bar{x} \in (X/\Gamma)^{\tau}$ determines uniquely a class in $H^1(\mathfrak{g}, \Gamma)$ and we have $(X/\Gamma)^{\tau} = \bigcup F(b)$, $b \in H^1(\mathfrak{g}, \Gamma)$ as a disjoint union. By $x \mapsto {}^{\tau}x \cdot b$, b as above, we have an isometric operation on X. From [8: I, 13.5] we then see that $F(b)$ is not empty. Any two points of $X(b)$ are joined by a unique geodesic of X [8: I, 13.3]. Therefore $X(b)$ is geodesially convex and all statements are proven. q.e.d.

Remark. (i) The proof above shows that more generally the following hold. Suppose that Γ is some group acting freely from the right on a set X. Let $\mathfrak{g}$ be a finite group acting from the left on Γ and on X in a compatible way, i.e.

${}^{\sigma}(x\gamma) = {}^{\sigma}x\,{}^{\sigma}\gamma$ for $x \in X$, $\gamma \in \Gamma$, $\sigma \in \mathfrak{g}$. Suppose that for every cocycle c for $H^1(\mathfrak{g}, \Gamma)$ the c-twisted $\mathfrak{g}$-action has a fixpoint in X. If $x_0 \in X$ is a fixpoint of the $\mathfrak{g}$-action on X, then we have by $\gamma \mapsto x_0 \cdot \gamma$ a map $\Gamma \to X$ and an exact sequence $1 \to \Gamma \to X \to X/\Gamma \to 1$. We get an exact sequence of pointed sets

$$1 \to \Gamma^{\mathfrak{g}} \to X^{\mathfrak{g}} \to (X/\Gamma)^{\mathfrak{g}} \xrightarrow{\partial} H^1(\mathfrak{g}, \Gamma) \to 1,$$

which describes the fixpoint set $(X/\Gamma)^{\mathfrak{g}}$ completely. This can be applied to actions of $\mathfrak{g}$ on symmetric spaces, their compactifications and to actions on Bruhat-Tits buildings.

ii) If Γ does not act freely, the above method to construct fixpoints works as well. However there may be fixpoints which do not arise from that construction.

For every cocycle $(1, b)$, $b \in \Gamma$, we now determine $X(b)$.

LEMMA 1.4. *The natural inclusion* $\mathrm{SO}_n(\mathbf{R}) \hookrightarrow \mathrm{Sl}_n(\mathbf{R})$ *induces a bijection* $H^1(\mathfrak{g}, \mathrm{SO}_n(\mathbf{R})) \xrightarrow{i} H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R}))$.

Proof. If $(1, \mathfrak{g})$ is a cocycle for $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R}))$ we have a twisted operation given by $x \mapsto {}^{\tau}x \cdot g^{-1}$ on X. This operation is isometric and thus [8: I, 13.5] has a fixpoint $y \in X$. If $x_0 \in X$ is the point corresponding to $\mathrm{SO}_n(\mathbf{R})$, then $y = x_0 \cdot h^{-1}$ for some $h \in \mathrm{Sl}_n(\mathbf{R})$. Hence $y = {}^{\tau}y \cdot g^{-1} = x_0 \cdot h^{-1} = x_0 \cdot {}^{\tau}h^{-1} \cdot g^{-1}$ and $h^{-1} \cdot g \cdot {}^{\tau}h \in \mathrm{SO}_n(\mathbf{R})$. This means that i is surjective.

To prove the injectivity of i we consider two cocycles $(1, k_1)$ and $(1, k_2)$, $k_1, k_2 \in \mathrm{SO}_n(\mathbf{R})$, and suppose that there exists a $g \in \mathrm{Sl}_n(\mathbf{R})$ such that $k_2 = g^{-1} \cdot k_1 \cdot {}^{\tau}g$. From [8: III, 7.4] we have a unique Cartan decomposition $\mathrm{Sl}_n(\mathbf{R}) = \mathrm{SO}_n(\mathbf{R}) \cdot P$. Here P is a closed and τ-stable subset, $P = P^{-1}$, and for $k \in \mathrm{SO}_n(\mathbf{R})$ we have $k^{-1} \cdot P \cdot k = P$. Write $g = k \cdot p$ with $k \in \mathrm{SO}_n(\mathbf{R})$ and $p \in P$. Then $k_2 = p^{-1} \cdot k^{-1} \cdot k_1 \cdot {}^{\tau}k \cdot {}^{\tau}p = k^{-1} \cdot k_1 \cdot {}^{\tau}k \cdot p'$ for some $p' \in P$. From the uniqueness of the Cartan decomposition we get $k_2 = k^{-1} \cdot k_1 \cdot {}^{\tau}k$ and thus i is injective. q.e.d.

Remark. The argument just given shows that an analog of Lemma 1.4 holds more generally for any finite group $\mathfrak{g}$ acting isometrically on a symmetric space. For some other special cases see [1: 6.8] and [19: 1.6]

We now compute $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R}))$. In fact we just have to reinterpret the classical theorem of Sylvester giving the classification of real quadratic forms.

We recall the notion of a signature. If $b \in \mathrm{Gl}_n(\mathbf{R})$ is symmetric, then there exists an $A \in \mathrm{Gl}_n(\mathbf{R})$ such that ${}^{t}A \cdot b \cdot A = \eta_{r,s}$. Here ${}^{t}x \cdot \eta_{r,s} \cdot x = \sum_{i=1}^{r} x_i^2 - \sum_{j=r+1}^{n} x_j^2$ for $x \in \mathbf{R}^n$ considered as row vector. The tuple (r, s) with $r + s = n$ is called the signature of b. By Sylvester's theorem we have a bijection $H^1(\mathfrak{g}, \mathrm{Gl}_n(\mathbf{R})) \xrightarrow{\sim} \{(r, s) \in \mathbf{N} \times \mathbf{N} / r + s = n\}$ induced by $b \mapsto$ signature (b).

LEMMA 1.5. *The signature induces a bijection*

$$H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R})) \tilde{\to} \{(r, s) \in \mathbf{N} \times \mathbf{N} / r + s = n,\ s \equiv 0 \bmod 2\}.$$

Proof. From the exact sequence $1 \to \mathrm{Sl}_n(\mathbf{R}) \to \mathrm{Gl}_n(\mathbf{R}) \xrightarrow{\det} \mathbf{R}^* \to 1$ we get an exact sequence

$$1 \to \mathrm{SO}_n(\mathbf{R}) \to \mathrm{O}_n(\mathbf{R}) \to \{\pm 1\} \xrightarrow{\partial} H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R})) \xrightarrow{i} H^1(\mathfrak{g}, \mathrm{Gl}_n(\mathbf{R})) \to H^1(\mathfrak{g}, \mathbf{R}^*).$$

In order to show that i is injective we have to twist the whole situation with a cocycle $(1, b)$ for $H^1(\mathfrak{g}, \mathrm{Gl}_n(\mathbf{R}))$ and to verify that the new boundary map ∂_b is trivial. But this is obvious because every orthogonal group $\mathrm{O}(b)(\mathbf{R})$ contains a reflection of determinant -1. Thus i is injective and the result follows immediately. q.e.d.

By $\mathrm{SO}(r, s)(\mathbf{R})$ we denote the special orthogonal group over $\mathbf{R}$ belonging to the bilinear form $\eta_{r,s}$. Let $\mathrm{SO}(r, s)^0(\mathbf{R})$ be the connected component of the identity in $\mathrm{SO}(r, s)(\mathbf{R})$. We have $\mathrm{SO}(r, s)^0(\mathbf{R}) \cap \mathrm{SO}_n(\mathbf{R}) = \mathrm{SO}_r(\mathbf{R}) \times \mathrm{SO}_s(\mathbf{R})$. The inclusion $\mathrm{SO}(r, s)^0(\mathbf{R}) \hookrightarrow \mathrm{Sl}_n(\mathbf{R})$ induces a diffeomorphism of $\mathrm{SO}_r(\mathbf{R}) \times \mathrm{SO}_s(\mathbf{R}) \backslash \mathrm{SO}(r, s)^0(\mathbf{R})$ onto a closed submanifold of X.

PROPOSITION 1.6. *If $(1, b)$ is a cocycle for $H^1(\mathfrak{g}, \Gamma)$ and if (r, s) is the signature of b, then there exists a $g \in \mathrm{Sl}_n(\mathbf{R})$ such that*

$$X(b) \cong \mathrm{SO}_r(\mathbf{R}) \times \mathrm{SO}_s(\mathbf{R}) \backslash \mathrm{SO}(r, s)^0(\mathbf{R}) \cdot g.$$

Proof. We consider $(1, b)$ as cocycle for $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R}))$ and using 1.4 we suppose that $b \in \mathrm{SO}_n$. Then we have an exact sequence of pointed sets

$$1 \to {}^b\mathrm{SO}_n(\mathbf{R}) \to {}^b\mathrm{Sl}_n(\mathbf{R}) \to {}^bX \to 1,$$

where the upper index b indicates that we consider the $\mathfrak{g}$-operation obtained by twisting with b. We get an exact sequence $1 \to {}^b\mathrm{SO}_n(\mathbf{R})^\tau \to {}^b\mathrm{Sl}_n(\mathbf{R})^\tau \to {}^bX^\tau \xrightarrow{\partial} H^1(\mathfrak{g}, {}^b\mathrm{SO}_n(\mathbf{R})) \xrightarrow{i} H^1(\mathfrak{g}, {}^b\mathrm{Sl}_n(\mathbf{R}))$. Now ${}^bX^\tau = X(b)$, ${}^b\mathrm{Sl}_n(R)^\tau = \mathrm{SO}(b)(\mathbf{R})$ and ${}^b\mathrm{SO}_n(\mathbf{R})^\tau = \mathrm{SO}_n(\mathbf{R}) \cap \mathrm{SO}(b)(\mathbf{R})$. From 1.4 we deduce that i is a bijection. Hence ∂ is trivial. We know that $X(b)$ is connected. The connected component $\mathrm{SO}(b)^0(\mathbf{R})$ of the identity in $\mathrm{SO}(b)(\mathbf{R})$ operates transitively on $X(b)$. By 1.5 there is a $g \in \mathrm{Sl}_n(\mathbf{R})$ such that $b = {}^t g \cdot \eta_{r,s} \cdot g$. The desired result now follows. q.e.d.

Remark. Submanifolds like $X(b) \subset X$ occur also naturally in the reduction theory of indefinite quadratic forms [2: §5].

COROLLARY 1.7. *If b is a cocycle for $H^1(\mathfrak{g}, \Gamma)$ with signature (r, s) then the dimension of $F(b)$ is $r \cdot s$.*

Corollary 1.7 implies that $F(b)$ consists of one point if and only if b considered as a quadratic form is definite. From the compactness criterion [2: §8] we deduce:

COROLLARY 1.8. *The fixpoint set $F(b)$ is compact and different from a point if and only if b considered as a quadratic form is indefinite and anisotropic over $\mathbf{Q}$.*

If $n \geqslant 5$ then by the theorem of Minkowski-Hasse an indefinite quadratic form in n variables is isotropic over $\mathbf{Q}$ [22: p. 77].

It is well known that the abstract group cohomology $H^*(\Gamma, \mathbf{Q})$ coincides with the cohomology $H^*(X/\Gamma, \mathbf{Q})$ of the topological space X/Γ. The space X/Γ has the homotopy type of a compact manifold with boundary [7: 1.3.3.]. Therefore the $\mathbf{Q}$-dimension of $H^*(\Gamma, \mathbf{Q})$ is finite. The map $\tau^i: H^i(\Gamma, \mathbf{Q}) \to H^*(\Gamma, \mathbf{Q})$ induced by τ has a trace denoted by $\operatorname{tr}(\tau^i)$. We define the Lefschetz number of τ and Γ by $L(\tau, \Gamma) = \sum_{i=0}^{\infty} (-1)^i \operatorname{tr}(\tau^i)$. We can (and we will) consider this Lefschetz number as the Lefschetz number $L(\tau, X/\Gamma)$ of the continuous map $\tau: X/\Gamma \to X/\Gamma$. It is well known that $L(\tau, \Gamma) \in \mathbf{Z}$. The Euler–Poincaré characteristic of a topological space Y is enoted by $\chi(Y)$.

PROPOSITION 1.9 (*Lefschetz fixpoint formula*). *With the notation introduced above we have*

$$L(\tau, X/\Gamma) = \chi((X/\Gamma)^\tau).$$

Proof. Let $N = (n-1)(n+2)/2$ be the dimension of X. The involution τ fixes the point $x_0 \in X$ corresponding to $\mathrm{SO}_n(\mathbf{R})$ and induces the multiplication with -1 in the tangent space in x_0 at X. Clearly X/Γ is orientable. Denote by $H_c^*(\ , \mathbf{Q})$ cohomology with compact supports and by $[X/\Gamma] \in H_c^N(X/\Gamma, \mathbf{Q})$ an orientation class. We write the isomorphism $H_c^N(X/\Gamma, \mathbf{Q}) \to \mathbf{Q}$ determined by $[X/\Gamma]$ in the form $\alpha \mapsto \langle \alpha, [X/\Gamma] \rangle$. We have $\tau^N([X/\Gamma]) = (-1)^N [X/\Gamma]$. By Poincaré duality [26: p. 139, p. 149] we have a non degenerate pairing

$$H_c^i(X/\Gamma, \mathbf{Q}) \times H^{n-i}(X/\Gamma, \mathbf{Q}) \to \mathbf{Q}$$

given by $(\alpha, \beta) \mapsto \langle \alpha \cup \beta, [X/\Gamma] \rangle$. Here $\alpha \cup \beta$ is the cup product of α and β. If $\tau_c^i: H_c^i(X/\Gamma, \mathbf{Q}) \to H_c^i(X/\Gamma, \mathbf{Q})$ is the map induced by τ then for its trace $\operatorname{tr}(\tau_c^i)$ we have $\operatorname{tr}(\tau_c^i) = (-1)^N \operatorname{tr}(\tau^{N-i})$. By the classical Lefschetz fixpoint formula

$L_c(\tau, X/\Gamma) = \chi_c((X/\Gamma)^\tau)$, see [28]. The index c on both sides of the equation indicates that we are using cohomology with compact supports. We know from 1.3 and 1.7 that $(X/\Gamma)^\tau$ is a disjoint union of even dimensional manifolds. Thus again by Poincaré duality $\chi_c((X/\Gamma)^\tau) = \chi((X/\Gamma)^\tau)$ and the desired formula follows. q.e.d.

§2. Equivalence classes of unimodular integral bilinear forms

In this chapter we describe the Γ-equivalence classes of certain unimodular integral bilinear forms in terms of their local data and the genera fixed by the local data. Again, as in §1, we use the language of non commutative cohomology.

2.1. For all places v of $\mathbf{Q}$ we consider $\mathrm{SL}_n(\mathbf{Q}_v)$ as a locally compact group in the topology induced by the topology of $\mathbf{Q}_v$. Suppose that for all finite places v of $\mathbf{Q}$ an open compact subgroup $\Gamma_v \subset \mathrm{Sl}_n(\mathbf{Q}_v)$ is given and that $\Gamma_v = \mathrm{Sl}_n(\mathbf{Z}_v)$ for almost all v. If $v = \infty$ set $\Gamma_v = \mathrm{Sl}_n(\mathbf{R})$. Then $\Gamma = \bigcap_v (\Gamma_v \cap \mathrm{Sl}_n(\mathbf{Q}))$ is a congruence subgroup of $\mathrm{Sl}_n(\mathbf{Q})$ and by the strong approximation property of $\mathrm{Sl}_n/\mathbf{Q}$ the closure of Γ in $\mathrm{Sl}_n(\mathbf{Q}_v)$ is Γ_v for all finite places v.

If Γ is τ-stable then all Γ_v are τ-stable and the inclusion $\Gamma \to \Gamma_v$ induces a map $h_v: H^1(\mathfrak{g}, \Gamma) \to H^1(\mathfrak{g}, \Gamma_v)$. We will see in 3.1 that $H^1(\mathfrak{g}, \Gamma_v) = \{1\}$ for almost all v. Therefore $\prod_v h_v = h$ induces a map

$$h: H^1(\mathfrak{g}, \Gamma) \to \coprod_v H^1(\mathfrak{g}, \Gamma_v).$$

For $b_v \in H^1(\mathfrak{g}, \Gamma_v)$ the Hasse–Witt invariant $\varepsilon_v(b_v) \in \{\pm 1\}$ is defined [22: p. 64]. If $\{b_v\} \in \coprod_v H^1(\mathfrak{g}, \Gamma_v)$ we define $\varepsilon(\{b_v\}) = \prod_v \varepsilon_v(b_v)$. This makes sense because almost all factors of the infinite product are 1.

Let $\mathbf{A}$ be the ring of adeles of $\mathbf{Q}$. If b is a nondegenerate bilinear form over $\mathbf{Q}$ then $\mathrm{SO}(b)(\mathbf{A})$ denotes the group of $\mathbf{A}$-valued points of $\mathrm{SO}(b)/\mathbf{Q}$. This group contains diagonally the group $\mathrm{SO}(b)(\mathbf{Q})$ and also the subgroup $\prod_v ({}^b\Gamma_v^\tau)$.

PROPOSITION 2.2. *Suppose that Γ is a τ-stable congruence subgroup of* $\mathrm{Sl}_n(\mathbf{Q})$ *and use the notation introduced in* 2.1. *Then there is an exact sequence of pointed sets*

$$1 \to \mathrm{SO}_n(\mathbf{Q}) \backslash \mathrm{SO}_n(\mathbf{A}) / \prod_v \Gamma_v^\tau \xrightarrow{\partial} H^1(\mathfrak{g}, \Gamma) \xrightarrow{h} \coprod_v H^1(\mathfrak{g}, \Gamma_v) \xrightarrow{\varepsilon} \{\pm 1\} \to 1.$$

The map ∂ is injective.

Proof. Consider the diagram

$$\begin{array}{ccccccccc} 1 & \to & \Gamma & \longrightarrow & \mathrm{Sl}(\mathbf{Q}) & \longrightarrow & \mathrm{Sl}_n(\mathbf{Q})/\Gamma & \to & 1 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 1 & \to & \prod_v \Gamma_v & \to & \mathrm{Sl}_n(\mathbf{A}) & \to & \mathrm{Sl}_n(\mathbf{A})\Big/\prod_v \Gamma_v & \to & 1 \end{array}$$

From this we get the following commutative diagram with exact rows and columns

$$\begin{array}{ccccccccc} & & 1 & & 1 & & F & & \\ & & \downarrow & & \downarrow & & \wr & & \\ 1 & \to & \mathrm{SO}_n(\mathbf{Q})/\Gamma^\tau & \longrightarrow & (\mathrm{Sl}_n(\mathbf{Q})/\Gamma)^\tau & \longrightarrow & H^1(\mathfrak{g}, \Gamma) & \longrightarrow & H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Q})) \\ & & \downarrow & & \downarrow & & \downarrow h' & & \downarrow h'' \\ 1 & \to & \mathrm{SO}_n(\mathbf{A})\Big/\prod_v \Gamma_v^\tau & \longrightarrow & (\mathrm{Sl}_n(\mathbf{A})\Big/\prod_v \Gamma_v)^\tau & \longrightarrow & H^1(\mathfrak{g}, \prod \Gamma_v) & \to & H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{A})) \\ & & \downarrow & & \downarrow & & & & \\ \to & & \mathrm{SO}_n(\mathbf{Q})\backslash\mathrm{SO}_n(\mathbf{A})\Big/\prod_v \Gamma_v^\tau & \longrightarrow & (\mathrm{Sl}_n(\mathbf{Q})\backslash\mathrm{Sl}_n(\mathbf{A})\Big/\prod_v \Gamma_v)^\tau. & & & & \\ & & \downarrow & & & & & & \\ & & 1 & & & & & & \end{array}$$

Using $H^1(\mathfrak{g}, \Gamma_v) = \{1\}$ for almost all v we have $H^1(\mathfrak{g}, \prod_v \Gamma_v) = \coprod_v H^1(\mathfrak{g}, \Gamma_v)$ and $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{A})) = \coprod_v H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Q}_v))$. Therefore h' coincides up to isomorphy with h. From the theorem of Hasse–Minkowski [22: p. 77] we get that h'' is injective. By strong approximation for $\mathrm{Sl}_n/\mathbf{Q}$ the set $\mathrm{Sl}_n(\mathbf{Q})\backslash\mathrm{Sl}_n(\mathbf{A})/\prod_v \Gamma_v$ consists of one point. If we denote by F the kernel of h' the existence and bijectivity of the snaked arrow follows by diagram chase.

We have to determine the cokernel of h'. If $\gamma \in H^1(\mathfrak{g}, \Gamma)$ then $\varepsilon h(\gamma) = 1$ by [22: p. 73]. Consider $\{\gamma_v\} \in \coprod_v H^1(\mathfrak{g}, \Gamma_v)$ with the property that $\prod_v \varepsilon_v(\gamma_v) = 1$. We have maps $\coprod_v H^1(\mathfrak{g}, \Gamma_v) \tilde{\to} H^1(\mathfrak{g}, \prod_v \Gamma_v) \to H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{A}))$. Thus $\{\gamma_v\}$ determines a class γ for $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{A})) \tilde{\to} \coprod_v H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Q}_v))$ with trivial Hasse–Witt invariant. From [22: p. 78] then γ is in the image of $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Q})) \to H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{A}))$. If $(1, \gamma_v)$ is a cocycle representing γ_v we therefore have a cocycle $(1, \xi)$, $\xi \in \mathrm{Sl}_n(\mathbf{Q})$, and an $a = \{a_v\} \in \mathrm{Sl}_n(\mathbf{A})$ such that $a_v^{-1} \cdot \xi \cdot {}^\tau a_v = \gamma_v$ for all v. By strong approximation $a = b \cdot c^{-1}$ with $b \in \mathrm{Sl}_n(\mathbf{Q})$, $c \in \prod_v \Gamma_v$. Hence $\eta = b^{-1} \cdot \xi \cdot {}^\tau b \in \Gamma$ represents a class η with $h(\eta) = \{\gamma_v\}$. The surjectivity of ε is easy to see. q.e.d.

Remark. i) This sequence determines the image of h and the fiber of h through the distinguished element of $H^1(\mathfrak{g}, \Gamma)$. If $(1, b)$ is a cocycle for $H^1(\mathfrak{g}, \Gamma)$ then the fiber of h through the element $\bar{b}$ represented by $(1, b)$ is obtained by

twisting [21: I, 5.4]. We get $h^{-1}(\{h(\bar{b})\}) \cong \mathrm{SO}(b)(\mathbf{Q})\backslash \mathrm{SO}(b)(\mathbf{A})/\prod \Gamma_v(b)$. Here we abbreviate $\Gamma_v(b) = {}^b\Gamma_v^\tau$.

ii) It is well known that the class number of special orthogonal groups is finite. In the next chapter we see that $\coprod_v H^1(\mathfrak{g}, \Gamma_v)$ is finite. Hence $H^1(\mathfrak{g}, \Gamma)$ is finite. In [1: 3.8] there is a different and simpler proof of the finiteness of $H^1(\mathfrak{g}, \Gamma)$.

iii) Results similar to Proposition 2.2 hold more generally. For galois actions see [19: 3.1].

§3. Local equivalence classes of unimodular bilinear forms

Let p be a prime number, j a non negative integer and $n \geq 2$ an integer. We denote by $\Gamma_p(j)$ the kernel of the natural surjection $\mathrm{Sl}_n(\mathbf{Z}_p) \to \mathrm{Sl}_n(\mathbf{Z}/p^j\mathbf{Z})$ and observe that $\Gamma_p(j)$ is τ-stable. In this section we compute $H^1(\mathfrak{g}, \Gamma_p(j))$.

At first we consider $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_p))$. This set has been determined by Minkowski in a different language. To state the result we need some notation.

Denote by E, L, S the symmetric bilinear forms on $\mathbf{Z}_p^n$ given for $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbf{Z}_p^n$ by $E(x, y) = \sum_{i=1}^n x_i y_i$, $L(x, y) = -x_1 y_1 - x_2 y_2 + \sum_{i=3}^n x_i y_i$, $S(x, y) = \sum_{i=1}^{n/2} x_{2i-1} y_{2i} + x_{2i} y_{2i-1}$ if n is even. The symmetric matrices representing E, L, S with respect to the standard basis are again denoted by E, L, S and E, L, S also denote the cohomology classes represented by E, L, S considered as cocycles.

PROPOSITION 3.1. *If $p \neq 2$ then $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_p)) = \{1\}$. If $p = 2$ and $n \not\equiv 0$ mod 4 then $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_2))$ consists of two classes which can be represented by E and L. If $p = 2$ and $n \equiv 0$ mod 4 then $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_2))$ consists of three classes which can be represented by E, L and S.*

Proof. The set $H^1(\mathfrak{g}, \mathrm{Gl}_n(\mathbf{Z}_p))$ has been described in [12, p. 91]. From the exact sequence $1 \to \mathrm{Sl}_n(\mathbf{Z}_p) \to \mathrm{Gl}_n(\mathbf{Z}_p) \to \mathbf{Z}_p^* \to 1$ we get the exact sequence

$$1 \to H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_p)) \xrightarrow{i} H^1(\mathfrak{g}, \mathrm{Gl}_n(\mathbf{Z}_p)) \to \mathbf{Z}_p^*/(\mathbf{Z}_p^*)^2 \to 1.$$

We twist the situation and show that i is actually injective. The result now follows. q.e.d.

We collect some simple facts. Let $\mathrm{M}_n(A)$ be set of $n \times n$-matrices with coefficients in a ring A. For $B \in M_n(A)$ define ${}^\tau B = -{}^t B$. Denote by $G_p(j)$ the kernel of the natural surjection $\mathrm{Gl}_n(\mathbf{Z}_p) \to \mathrm{Gl}_n(\mathbf{Z}/p^j\mathbf{Z})$. Then the following hold.

i) For all primes p and all $j \geq 1$ there is an isomorphism of commutative

groups with τ-operation

$$M_n(\mathbf{Z}/p\mathbf{Z}) \tilde{\to} G_p(j)/G_p(j+1)$$

given by $\bar{B} \mapsto 1+p^jB \bmod G_p(j+1)$. Here $B \in M_n(\mathbf{Z}_p)$ is a matrix with $B \equiv \bar{B} \bmod p \cdot M_n(\mathbf{Z}_p)$.

ii) If $B \in p^j \cdot M_n(\mathbf{Z}_p)$ and if $p \neq 2$ and $j \geq 1$ or $p=2$ and $j \geq 2$ then $B^n/n! \in p^{2j-1} \cdot M_n(\mathbf{Z}_p)$. The series $\exp(B) = \sum_{n=0}^{\infty} B^n/n!$ converges and defines a τ-equivariant bijection

$$\exp : p^j \cdot M_n(\mathbf{Z}_p) \to G_p(j).$$

PROPOSITION 3.2. *Let b be either the cocycle E or the cocycle L. Then for all $j \geq 1$ the natural map $\Gamma_2(j) \to \Gamma_2(j)/\Gamma_2(j+1)$ induces a bijection*

$$H^1(\mathfrak{g}, {}^b\Gamma_2(j)) \tilde{\to} H^1(\mathfrak{g}, {}^b(\Gamma_2(j)/\Gamma_2(j+1))) = (\mathbf{Z}/2\mathbf{Z})^{n-1}.$$

The cocycles for $H^1(\mathfrak{g}, {}^b\Gamma_2(j))$ representing the different classes can be chosen as diagonal matrices with only integers out of $\{u, u(1+2^j)\}$, $u \in (1+2^j \cdot \mathbf{Z}_2)^2$ on the diagonal and $(1+2^j)u$ occurring an even number of times. For $p \neq 2$ and all $j \geq 1$ we have

$$H^1(\mathfrak{g}, \Gamma_p(j)) = \{1\}.$$

Proof. We have the exact sequence $1 \to G_p(j+1) \to G_p(j) \to G_p(j)/G_p(j+1) \to 1$ and thus for every cocycle d for $H^1(\mathfrak{g}, \mathrm{Gl}_n(\mathbf{Z}_p))$ an exact sequence

$$H^1(\mathfrak{g}, {}^dG_p(j+1)) \xrightarrow{u} H^1(\mathfrak{g}, {}^dG_p(j)) \xrightarrow{i} H^1(\mathfrak{g}, {}^d(G_p(j)/G_p(j+1))).$$

Now ${}^d(G_p(j)/G_p(j+1)) \tilde{\to} {}^dM_n(\mathbf{Z}/p\mathbf{Z}) \tilde{\to} M_n(\mathbf{Z}/p\mathbf{Z})$ as groups with τ-operation. The last isomorphism is given by $B \mapsto B \cdot d$. Cocycles for $H^1(\mathfrak{g}, M_n(\mathbf{Z}/p\mathbf{Z}))$ are matrices $A \in M_n(\mathbf{Z}/p\mathbf{Z})$ with $A - {}^tA = 0$ and elements of the form $C + {}^tC$ are coboundaries. We get.

$$H^1(\mathfrak{g}, {}^d(G_p(j)/G_p(j+1)) = \begin{cases} (\mathbf{Z}/2\mathbf{Z})^n & \text{if } p=2 \\ 0 & \text{if } p \neq 2. \end{cases}$$

If $d=E$ or $d=L$ all cocycles can be represented by diagonal matrices where only integers out of $\{1, 1+2^j\}$ occur on the diagonal.

If $p \neq 2$ and $j \geq 1$ or if $p = 2$ and $j \geq 2$ the map u is trivial. To see this we take a cocycle for $H^1(\mathfrak{g}, {}^d(G_p(j+1)))$ represented by a matrix $A \in G_p(j+1)$ with $A = d \cdot {}^tA \cdot d^{-1}$. If $A = \exp(B)$, $B \in p^{j+1} \cdot \mathrm{M}_n(\mathbf{Z}_p)$, then $B = d \cdot {}^tB \cdot d^{-1}$. For $C = \exp(-B/2)$ then $C \in G_p(j)$ and $C^{-1} \cdot d \cdot {}^\tau C \cdot d^{-1} = A$. Similarily the natural map $H^1(\mathfrak{g}, {}^d\Gamma_p(j+1)) \to H^1(\mathfrak{g}, {}^d\Gamma_p(j))$ is trivial.

For $p \neq 2$ and $j \geq 1$ we get $H^1(\mathfrak{g}, G_p(j)) = \{1\}$. We write $U(j)$ instead of $1 + p^j\mathbf{Z}_p$. Then from the exact sequence $1 \to \Gamma_p(j) \to G_p(j) \to U(j) \to 1$ we get the exact sequence

$$G_p(j)^\tau \to U(j)^\tau \xrightarrow{\partial} H^1(\mathfrak{g}, \Gamma_p(j)) \to 1.$$

We observe that ∂ is trivial. The result for $p \neq 2$ and $j \geq 1$ follows.

We have an exact sequence of commutative groups

$$0 \to \Gamma_2(j)/\Gamma_2(j+1) \to G_2(j)/G_2(j+1) \to U(j)/U(j+1) \to 0.$$

and $U(j)/U(j+1) = \mathbf{Z}/2\mathbf{Z}$ with trivial τ-action. We get an exact sequence

$$(G_2(j)/G_2(j+1))^\tau \to \mathbf{Z}/2\mathbf{Z} \xrightarrow{\partial} H^1(\mathfrak{g}, \Gamma_2(j)/\Gamma_2(j+1)) \to$$
$$H^1(\mathfrak{g}, G_2(j)/G_2(j+1)) \to H^1(\mathfrak{g}, \mathbf{Z}/2\mathbf{Z}).$$

Now ∂ is trivial $H^1(\mathfrak{g}, \mathbf{Z}/2\mathbf{Z}) = \mathbf{Z}/2\mathbf{Z}$ and the last arrow is non trivial. Therefore $H^1(\mathfrak{g}, \Gamma_2(j)/\Gamma_2(j+1)) = H^1(\mathfrak{g}, {}^L(\Gamma_2(j)/\Gamma_2(j+1))) \cong (\mathbf{Z}/2\mathbf{Z})^{n-1}$. If $b = E$ or $b = L$ then from the description of cocycles for $H^1(\mathfrak{g}, \Gamma_2(j)/\Gamma_2(j+1))$ we see that the natural map

$$H^1(\mathfrak{g}, {}^b\Gamma_2(j)) \to H^1(\mathfrak{g}, {}^b(\Gamma_2(j)/\Gamma_2(j+1)))$$

is surjective and our assertion follows for $p = 2$ and $j \geq 2$.

From the exact sequence $1 \to \Gamma_2(2) \to \Gamma_2(1) \to \Gamma_2(1)/\Gamma_2(2) \to 1$ we get a map ${}^d(\Gamma_2(1)/\Gamma_2(2))^\tau \xrightarrow{\partial} H^1(\mathfrak{g}, {}^d\Gamma_2(2))$. We show that ∂ is surjective. If $A \in \Gamma_2(2)$, $A = d \cdot {}^tA \cdot d^{-1}$, $A = \exp(B)$, $B \in 4 \cdot \mathrm{M}_n(\mathbf{Z}_2)$ then $B = d \cdot {}^tB \cdot d^{-1}$ and $C = 1 - B/2$ represents an element in ${}^d(G_2(1)/G_2(2))^\tau$ which lies in the image of ${}^d(\Gamma_2(1)/\Gamma_2(2))^\tau$. Now $C^{-1} \cdot d \cdot {}^tC^{-1} \cdot d^{-1} \equiv (1 + B/2)d(1 + {}^tB/2)d^{-1} \equiv 1 + B + B^2/4 \equiv 1 + B + B^2/2 \equiv \exp(B) \bmod 8\,\mathrm{M}_n(\mathbf{Z}_2)$. The natural map $H^1(\mathfrak{g}, {}^d\Gamma_2(2)) \to H^1(\mathfrak{g}, {}^d(\Gamma_2(2)/\Gamma_2(3)))$ is injective. Therefore the last equations mean that A is in the image of ∂. We get that the natural map $H^1(\mathfrak{g}, {}^d\Gamma_2(1)) \to H^1(\mathfrak{g}, {}^d(\Gamma_2(1)/\Gamma_2(2)))$ is injective. Thus the claimed result holds also for $p = 2$ and $j = 1$. q.e.d.

Later on we need information on the fibers of the natural map $H^1(\mathfrak{g}, \Gamma_2(1)) \to H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_2))$. We use the following simple combinatorial facts.

The number of subsets of even cardinality out of n elements is 2^{n-1}. This follows for example from the equation $2^n=(1+1)^n$ and $0=(1-1)^n$. We set

$$e(E)=\sum_j \binom{n}{4j} \quad \text{and} \quad e(L)=\sum_j \binom{n}{4j+2}.$$

From consideration of $(1-i)^n$, $i=(-1)^{1/2}$, we see that $e(E)=\frac{1}{2}(2^{n-1}+2^{n/2}\cdot\cos(\pi n/4))$ and $e(L)=\frac{1}{2}(2^{n-1}-2^{n/2}\cdot\cos(\pi n/4))$.

COROLLARY 3.3. *The classes represented by E and by L form the image of the natural map $H^1(\mathfrak{g}, \Gamma_2(1))\to H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_2))$. In $H^1(\mathfrak{g}, \Gamma_2(1))$ there are $e(E)$ classes with image E and $e(L)$ classes with image L.*

Proof. Clearly E and L are in the image. The bilinear form S has the property that $S(x,x)/2\in\mathbf{Z}_2$ for all $x\in\mathbf{Z}_2^n$. As no bilinear form in the image of $H^1(\mathfrak{g}, \Gamma_2(1))$ has this property S is not in the image. We have Hasse–Witt invariants $\varepsilon_2(E)=1$ and $\varepsilon_2(L)=-1$. If $2\cdot t$ is the number of integers $1+2$ occurring in the diagonal representation of a cocycle $A\in\Gamma_2(1)$ for $H^1(\mathfrak{g}, \Gamma_2(1))$ then $\varepsilon_2(A)=(-1)^t$. The result now follows from the combinatorial meaning of $e(E)$ and $e(L)$. q.e.d.

§4. Volume computations

Let $b\in\mathrm{Sl}_n(\mathbf{Z})$, $n\geq 2$, be a symmetric matrix. In this chapter we explicitly chose biinvariant measures ω_v on $\mathrm{SO}(b)(\mathbf{Q}_v)$ for all places v of $\mathbf{Q}$ such that $\prod_v \omega_v$ is the Tamagawa measure [7] on $\mathrm{SO}(b)(\mathbf{A})$ with $\mathbf{A}$ the ring of adeles over $\mathbf{Q}$. We compare ω_∞ with the Euler-Poincaré measure e_χ in the sense of Serre [23: §3] on $\mathrm{SO}(b)(\mathbf{R})$.

The volumes of congruence subgroups of $\mathrm{SO}(b)(\mathbf{Q}_p)$ are computed. We will explain in the beginning of the next chapter how such results are useful.

4.1. We recall some facts on Lie algebras of orthogonal groups. Let $\mathrm{so}(b)(\mathbf{Q})=\{A\in\mathrm{M}_n(\mathbf{Q})/A+b{}^tAb^{-1}=0\}$. If $X, Y\in\mathrm{so}(b)(\mathbf{Q})$ then by $e(X,Y)=-\mathrm{tr}(X\cdot Y)/2$ we have a non degenerate bilinear form on $\mathrm{so}(b)(\mathbf{Q})$ and it is well known that $(X,Y)\mapsto -2(n-2)e(x,y)$ is the Killing form on $\mathrm{so}(b)(\mathbf{Q})$. If $b=E$ we write $\mathrm{so}_n(\ )$ instead of $\mathrm{so}(E)(\ )$. Clearly $\mathrm{so}_n(\mathbf{Q})$ consists of skew symmetric matrices. We have a natural basis $\{u_{i,j}\}_{i<j}$ for $\mathrm{so}_n(\mathbf{Q})$ consisting of skew symmetric matrices whose coefficients above the diagonal different from the one coefficient in the j-th place of the i-th column, which is one, are zero. One has $\mathrm{so}_{n-1}(\mathbf{Q})\subset\mathrm{so}_n(\mathbf{Q})$ in a natural way. By induction on n we see that $\det(e(u_{i,j}, u_{k,l})_{i<j,k<l})=1$. This property is one reason for the choice of e. We have an isomorphism of vector

spaces $\mathrm{so}(b)(\mathbf{Q}) \xrightarrow{\sim} \mathrm{so}_n(\mathbf{Q})$ given by $A \mapsto A \cdot b$. Therefore we have a natural basis $e_{i,j} := u_{i,j} \cdot b^{-1}$, $i<j$, of $\mathrm{so}(b)(\mathbf{Q})$ and because $\det(b)=1$ we find $\det(e(e_{i,j}, e_{k,l})_{i<j,k<l}) = \pm 1$. We denote the dual basis of $e_{i,j}$, $i<j$, by $de_{i,j}$, $i<j$.

4.2. Set $d = n(n-1)/2$ and choose some order $e_1, \ldots, e_d$ of the basis elements of $\mathrm{so}(b)(\mathbf{Q})$. Then $\omega = de_1 \wedge \cdots \wedge de_d$ is a form of highest degree on $\mathrm{so}(b)(\mathbf{Q})$ which can be extended by left translation to a differential form on the algebraic group $\mathrm{SO}(b)/\mathbf{Q}$. If $X_1, \ldots, X_d \in \mathrm{so}(b)(\mathbf{Q})$ are linear independent, then

$$\omega(X_1, \ldots, X_d) = \mathrm{sign}(X_1, \ldots, X_d)(+|\det(e(X_i, X_j)_{i,j=1,\ldots,d}|^{1/2}).$$

Here $\mathrm{sign}(X_1, \ldots, X_d) = 1$ if $X_1, \ldots, X_d$ determine the same orientation as $e_1, \ldots, e_d$ and $\mathrm{sign}(X_1, \ldots, X_d) = -1$ otherwise. This formula has to be checked only for the basis $e_1, \ldots, e_d$ and follows then from 4.1. From the last equation we see that ω is invariant under the adjoint action of $\mathrm{SO}(b)/\mathbf{Q}$. Therefore ω is a biinvariant differential form.

4.3. The differential form ω determines for each place v of $\mathbf{Q}$ a biinvariant measure ω_v on the locally compact group $\mathrm{SO}(b)(\mathbf{Q}_v)$. This is explained in [7]. We can describe the measures ω_v as follows.

i) If $v = \infty$ then on $\mathrm{SO}(b)(\mathbf{R})$ the measure ω_∞ is the measure which in the sense of differential geometry belongs to the (possibly non positive) metric induced by $e(,)$.

ii) Suppose that $v = p$. The exponential map exp defines a bianalytic map of a neighbourhood V of 0 in $\mathrm{so}(b)(\mathbf{Q}_v)$ onto a neighbourhood U of the identity element in $\mathrm{SO}(b)(\mathbf{Q}_v)$. For $Y \in V$ we identify the tangentspace in $\exp(Y)$ at $\mathrm{SO}(b)(\mathbf{Q}_v)$ by left translation with the Lie algebra. Therefore the determinant of the differential of exp in the point Y $\det(\mathrm{Dexp}|_Y)$ makes sense. If V is small enough then $\|\det(\mathrm{Dexp}|_Y)\|_p = 1$ because $\mathrm{Dexp}|_0 = E$. We consider ω as constant differential form on V. Then by the transformation rule of integrals

$$\int_U \omega_v = \int_V \omega.$$

If $X_1, \ldots, X_d \in \mathrm{so}(b)(\mathbf{Q}_v)$ are linear independent and if $P = \mathbf{Z}_p \cdot X_1 + \cdots + \mathbf{Z}_p \cdot X_d \subset V$ then $\int_{\exp(P)} \omega_v = \|\det e(X_i, X_j)_{i,j}\|_p^{1/2}$.

4.4. Let $\mathbf{A}$ be the adele ring of $\mathbf{Q}$. Then $\mathrm{SO}(b)(\mathbf{A})$ is a locally compact group and ω determines a biinvariant measure denoted by $\omega = \prod \omega_v$ on $\mathrm{SO}(b)(\mathbf{A})$. This and the notation $\omega = \prod_v \omega_v$ is explained in [7]. On $\mathrm{SO}(b)(\mathbf{Q}) \backslash \mathrm{SO}(b)(\mathbf{A})$ there is a

measure induced by ω and by the formula of Minkowski–Siegel we have the volume $\mathrm{vol}_\omega(\mathrm{SO}(b)(\mathbf{Q})\backslash\mathrm{SO}(b)(\mathbf{A})) = 2$. If $n \geq 3$ see [27] if $n = 2$ see [18: p. 128]. This result is for $n = 2$ nothing else but a translation of the analytical class number formula for quadratic extensions of $\mathbf{Q}$.

We now compare the Gauß-Bonnet form with ω_∞. In principle such a comparison has been done by Siegel [24: III, 79, §3, §4].

PROPOSITION 4.5. *Suppose that b is a rational non degenerate bilinear form of signature $(r, 2j)$, $r+2j=n$. Set $l=[n/2]$ and define a biinvariant measure e_χ on* $\mathrm{SO}(b)(\mathbf{R})$ *by*

$$e_\chi = (-1)^{r \cdot j}\binom{l}{j}\mathrm{vol}_{\omega_\infty}(\mathrm{SO}_n(\mathbf{R}))^{-1} \cdot \omega_\infty$$

Then for every torsion free arithmetic subgroup Γ of $\mathrm{SO}(b)(\mathbf{Q})$ *the following holds*

$$\int_{\mathrm{SO}(b)(\mathbf{R})/\Gamma} e_\chi = \chi(\Gamma).$$

Proof. If $j=0$ or $2j=n$ then $\mathrm{SO}(b)(\mathbf{R})$ is compact, $\Gamma=\{1\}$, and the assertion is trivially true. Suppose now that $j \neq 0$ and $2j \neq n$. Then $\mathrm{SO}(b)(\mathbf{R})$ is non compact and has two connected components [8: p. 346]. We consider the symmetric space $X = K\backslash\mathrm{SO}(b)(\mathbf{R})$ where K is a maximal compact subgroup of $\mathrm{SO}(b)(\mathbf{R})$. The group K meets both components of $\mathrm{SO}(b)(\mathbf{R})$ and the connected components of the identity K^0 of K is isomorphic to $\mathrm{SO}_r(\mathbf{R}) \times \mathrm{SO}_{2j}(\mathbf{R})$. This group contains a compact torus of maximal rank $[n/2]=[r/2]+j$. From Harder's Gauß–Bonnet theorem [7] we see that there exists a $\mathrm{SO}(b)(\mathbf{R})$-right invariant differential form ω_χ of highest degree on X such that

$$\int_{X/\Gamma} \omega_\chi = \chi(\Gamma).$$

for all arithmetic torsion free subgroups $\Gamma \subset \mathrm{SO}(b)(\mathbf{Q})$. Moreover it is well known that ω_χ can be computed in a universal way from the curvature tensor R of X [13: p. 318].

To compute ω_χ we at first recall the Hirzebruch proportionality principle [9: p. 134]. Let $\mathfrak{l} = \mathfrak{so}(b)(\mathbf{R})$ be the Lie algebra of $\mathrm{SO}(b)(\mathbf{R})$, let $\mathfrak{k} \subset \mathfrak{l}$ be the subalgebra corresponding to K and let $\mathfrak{l} = \mathfrak{k} \oplus \mathfrak{p}$ be the Cartan decomposition. If $x_0 \in X$ is the point corresponding to K then $\mathfrak{p}$ is the tangent space of X at x_0. The subspace $\mathfrak{k}$ is orthogonal to $\mathfrak{p}$ with respect to the bilinear form $e(,)$. Thus $-e(,)$ induces by

right translation a Riemannian metric on X. The connected real Lie subgroup of $SO(b)(\mathbf{C})$ corresponding to the real subalgebra $\mathfrak{k} \oplus i\mathfrak{p}$ is denoted by G_0. Consider $\mathfrak{l} \oplus_{\mathbf{R}} \mathbf{C}$ as a real Lie algebra and $\mathfrak{l} \subset \mathfrak{l} \oplus_{\mathbf{R}} \mathbf{C}$ as a subalgebra. Then there is a natural extension of $e(,)$ to a real bilinear form on $\mathfrak{l} \oplus_{\mathbf{R}} \mathbf{C}$ which we again denote by $e(,)$. This e induces a Riemannian metric on the compact symmetric space $X_0 = K^0 \backslash G_0$ and the map $Y \mapsto i \cdot Y$ from $\mathfrak{p}$ to $i\mathfrak{p}$ induces an open immersion $j: X \to X_0$ which is an isometry on the tangent spaces at the point corresponding to K^0.

It is known that the curvature tensors R resp. R_0 on X resp. X_0 are $SO(b)(\mathbf{R})$-resp. G_0-right invariant and that for $U, V, W \in \mathfrak{p}$ we have $R(U, V, W) = -[[U, V], W]$ resp. $R_0(iU, iV, iW) = -[[iU, iV], iW]$. Hence $-j^*|_{x_0} R_0 = R|_{x_0}$. The G_0-right invariant Gauß–Bonnet form ω_{0_x} on X_0 has the form $\omega_{0_x} = c \cdot v_0$ where $c \in \mathbf{R}$ and where v_0 is the right invariant volume form determined by e on X_0. On the other hand ω_x and ω_{0_x} are given as universal polynomials in the curvature tensors. Using $-j^*|_{x_0} R_0 = R|_{x_0}$ we see that $\omega_x|_{x_0} = (-1)^{r \cdot j} \cdot c \cdot v|_{x_0}$. Here v is the volume form on X determined by e. Hopf and Samelson proved that $\chi(X_0) = |W_{G_0}| \cdot |W_{K^0}|^{-1}$ where $|W_{G_0}|$ and $|W_{K^0}|$ are the orders of the Weyl groups of G_0 and K^0. Therefore $c = |W_{G_0}| \, |W_{K^0}|^{-1} \operatorname{vol}_{v_0}(K^0 \backslash G_0)^{-1}$. Now

$$\int_{X/\Gamma} v = \operatorname{vol}(K)^{-1} \int_{SO(b)(\mathbf{R})/\Gamma} \omega_\infty .$$

We get

$$|W_{G_0}| \cdot |W_{K^0}|^{-1} \cdot \operatorname{vol}_{v_0}(K^0 \backslash G_0)^{-1} \int_{SO(b)(\mathbf{R})/\Gamma} \omega_\infty = \chi(\Gamma).$$

We have to compute the factors on the left hand side. Obviously $\mathfrak{k} \oplus i\mathfrak{p}$ is a real compact form for $\mathfrak{l} \oplus_{\mathbf{R}} \mathbf{C}$. Therefore [8: III, 7.3] the group G_0 is conjugate in $SO(b)(\mathbf{C})$ to $SO_n(\mathbf{R})$ and $K^0 \cong SO_r(\mathbf{R}) \times SO_{2j}(\mathbf{R})$. Thus $|W_{G_0}|$ and $|W_{K^0}|$ are known. We have $\operatorname{vol}(K) = 2 \cdot \operatorname{vol}(K^0)$. Because G_0 and $SO_n(\mathbf{R})$ are conjugate in $SO(b)(\mathbf{C})$ and because the volume elements on both groups are induced from the same $SO(b)(\mathbf{C})$-invariant bilinear form on $\mathfrak{l} \oplus_{\mathbf{R}} \mathbf{C}$ we have $\operatorname{vol}(G_0) = \operatorname{vol}(SO_n(\mathbf{R}))$. The desired formula follows now easily. q.e.d.

If $K_v \subset SO(b)(\mathbf{Q}_v)$ is a compact subset we write $\operatorname{vol}(K_v)$ for the volume of K_v with respect to the measures introduced in 4.3.

LEMMA 4.6. *Denote by* $\operatorname{vol}(S^j)$, $j \geq 1$, *the volume of the standard* j-

dimensional sphere of radius 1 *with respect to the euclidean volume element. Then*

$$\mathrm{vol}\,(\mathrm{SO}_n\,(\mathbf{R})) = \mathrm{vol}\,(S^1)\cdot\,\cdots\,\cdot\,\mathrm{vol}\,(S^{n-1}).$$

We have

$$\mathrm{vol}\,(\mathrm{SO}_{2l}\,(\mathbf{R})) = \frac{(2\pi)^l}{(l-1)!}\prod_{j=1}^{l-1}\frac{(2\pi)^{2j}}{(2j-1)!}$$

$$\mathrm{vol}\,(\mathrm{SO}_{2l+1}(\mathbf{R})) = 2l\prod_{j=1}^{l}\frac{(2\pi)^{2j}}{(2j-1)!}.$$

Proof. We have a fibration $\mathrm{SO}_{n-1}\,(\mathbf{R}) \to \mathrm{SO}_n\,(\mathbf{R}) \xrightarrow{f} S^{n-1}$. If $N = (1, 0, \ldots, 0) \in \mathbf{R}^n$ then for $A \in \mathrm{SO}_n\,(\mathbf{R})$ we define $f(A) = A(N) = A \cdot {}^tN$. On $\mathrm{so}_n\,(\mathbf{R})/\mathrm{so}_{n-1}\,(\mathbf{R})$ we have an induced scalar product and f induces an isometry of $\mathrm{so}_n\,(\mathbf{R})/\mathrm{so}_{n-1}\,(\mathbf{R})$ with the tangent space of S^{n-1} at N. The tangent space here is equipped with the euclidean scalar product. The result now follows by induction and by well known formulas for $\mathrm{vol}\,(S^i)$. q.e.d.

PROPOSITION 4.7. *With the notation introduced in* §3 *the following formulas hold.*

i)

$$\mathrm{vol}\,(\mathrm{SO}_n\,(\mathbf{Z}_2) = e(E)^{-1}\cdot\prod_{j=1}^{[(n-1)/2]}(1-2^{-2j})$$

If $p \neq 2$ *then*

$$\mathrm{vol}\,(\mathrm{SO}_n\,(\mathbf{Z}_p)) = \begin{cases} \displaystyle\prod_{j=1}^{[(n-1)/2]}(1-p^{-2j}) & \text{if} \quad n \equiv 1 \bmod 2 \\ (1-p^{-n/2})\displaystyle\prod_{j=1}^{[(n-1)/2]}(1-p^{-2j}) & \text{if} \quad n \equiv 0 \bmod 4 \\ (1-\chi(p)p^{-n/2})\displaystyle\prod_{j=1}^{[(n-1)/2]}(1-p^{-2j}) & \text{if} \quad n \equiv 2 \bmod 4 \end{cases}$$

Here we have $\chi(p) = (-1)^{(p-1)/2}$. *For all primes* p

$$\mathrm{vol}\,(\Gamma_p(j)^\tau) = p^{-j\cdot n(n-1)/2} \quad \text{if} \quad j \geq 1.$$

ii)

$$\text{vol}\,(\mathrm{SO}\,(L)(\mathbf{Z}_2)) = e(L)^{-1} \prod_{j=1}^{[(n-1)/2]} (1-2^{-2j})$$

$$\text{vol}\,({}^{L}\Gamma_2(1)^{\tau}) = 2^{-n(n-1)/2}$$

iii)

$$\text{vol}\,(\mathrm{SO}\,(S)(\mathbf{Z}_2)) = (1-2^{-n/2}) \prod_{j=1}^{[(n-1)/2]} (1-2^{-2j}) \quad \textit{if} \quad n \equiv 0 \bmod 4$$

Proof. Let b be one of the matrices E, L, S. Suppose that $p \neq 2$ and $j \geq 1$ or $p = 2$ and $j \geq 2$. Then we have a bijection exp: $p^j \cdot M_n(\mathbf{Z}_p) \tilde{\to} \Gamma_p(j)$ and ${}^{b}\Gamma_p(j)^{\tau} \cong \{A \in p^j \cdot \mathrm{M}_n(\mathbf{Z}_p)/A \cdot b + {}^{t}(A \cdot b) = 0\} \tilde{\to} \{B \in p^j \cdot \mathrm{M}_n(\mathbf{Z}_p)/B + {}^{t}B = 0\}$. The last isomorphism is given by $A \mapsto A \cdot b$ and is an isomorphism of free $\mathbf{Z}_p$-modules. By definition of our measure and 4.3 we get $\text{vol}\,({}^{b}\Gamma_p(j)^{\tau}) = p^{-j \cdot n(n-1)/2}$.

We consider now the case that $p = 2$ and $j = 1$ and $b = E$ or $b = L$. Then we have an exact sequence

$$1 \to {}^{b}\Gamma_2(1)^{\tau}/{}^{b}\Gamma_2(2)^{\tau} \to {}^{b}(\Gamma_2(1)/\Gamma_2(2))^{\tau} \xrightarrow{\partial} H^1(\mathfrak{g}, {}^{b}\Gamma_2(2)) \to H^1(\mathfrak{g}, {}^{b}\Gamma_2(1))$$

From 3.2 we know that ∂ is surjective and that $|H^1(\mathfrak{g}, {}^{b}\Gamma_2(2))| = 2^{n-1}$. It is easy to see that $|{}^{b}(\Gamma_2(1)/\Gamma_2(2))^{\tau}| = 2^{n-1} \cdot 2^{n(n-1)/2}$ and we find that the index of ${}^{b}\Gamma_2(2)^{\tau}$ in ${}^{b}\Gamma_2(1)^{\tau}$ is $2^{n(n-1)/2}$. Together with the above computation we now have proven the last formulas of i) and ii).

If $p \neq 2$ then from 3.1 we have an isomorphism $\mathrm{SO}_n(\mathbf{Z}_p)/\Gamma_p(1)^{\tau} \tilde{\to} \mathrm{SO}_n(\mathbf{Z}/p\mathbf{Z})$. The order of $\mathrm{SO}_n(\mathbf{Z}/p\mathbf{Z})$ is known [6: 25.4] and the second formula of i) follows.

We now prove the first formulas of i) and ii). Let b be the matrix E or L. We consider the diagram

$$\begin{array}{l} 1 \to {}^{b}(\Gamma_2(1)/\Gamma_2(2))^{\tau} \to {}^{b}(\Gamma_2/\Gamma_2(2))^{\tau} \to {}^{b}(\Gamma_2/\Gamma_2(1))^{\tau} \\ \qquad\qquad \xrightarrow{\partial} H^1(\mathfrak{g}, {}^{b}(\Gamma_2(1)/\Gamma_2(2))) \to H^1(\mathfrak{g}, {}^{b}(\Gamma_2/\Gamma_2(2))) \\ \qquad\qquad\qquad \uparrow \wr \qquad\qquad\qquad\qquad \uparrow \wr \\ \qquad\qquad\qquad H^1(\mathfrak{g}, {}^{b}\Gamma_2(1)) \qquad\qquad H^1(\mathfrak{g}, {}^{b}\Gamma_2). \end{array}$$

From 3.3 we know that $\mathrm{im}\,(\partial)$ has $e(b)$ elements. Now ${}^{b}(\Gamma_2(1)/\Gamma_2(2))^{\tau}$ has $2^{n-1} \cdot 2^{n(n-1)/2}$ elements. Therefore ${}^{b}(\Gamma_2/\Gamma_2(2))^{\tau}$ has $2^{n-1} \cdot 2^{n(n-1)/2} |(\Gamma_2/\Gamma_2(1))^{\tau}|\, e(b)^{-1}$

elements. From [5: p. 63] we deduce that

$$|(\Gamma_2/\Gamma_2(1))^\tau| = 2^{n(n-1)/2} \prod_{j=1}^{[(n-1)/2]} (1-2^{-2j}).$$

With the exact sequence

$$1 \to {}^b\Gamma_2^\tau/{}^b\Gamma_2(2)^\tau \to {}^b(\Gamma_2/\Gamma_2(2))^\tau \to H^1(\mathfrak{g}, {}^b\Gamma_2(2)) \to 1,$$

the formulas $|H^1(\mathfrak{g}, {}^b\Gamma_2(2))| = 2^{n-1}$ and $\mathrm{vol}\,({}^b\Gamma_2(2)^\tau) = 2^{-n(n-1)}$ we find the desired result.

We define the quadratic form $q_S(x) = S(x, x)/2$, $x \in \mathbf{Z}_2^n$. The orthogonal group $\mathrm{O}(q_S)/\mathbf{Z}_2$ then is smooth. By Hensels lemma we find

$$\mathrm{O}(q_S)/{}^S\Gamma_2(2)^\tau \xrightarrow{\sim} \mathrm{O}(q_S)(\mathbf{Z}/4\mathbf{Z}).$$

From [6: 25.1] we have $|\mathrm{O}(q_s)(\mathbf{Z}/4\mathbf{Z})| = 2 \cdot 2^{n(n-1)} \cdot (1-2^{-l}) \cdot \prod_{j=1}^{l-1} (1-2^{-2j})$, $l = n/2$. In the beginning of this proof we saw $\mathrm{vol}\,({}^S\Gamma_2(2)^\tau) = 2^{-n(n-1)}$. Therefore iii) follows. q.e.d.

Remark. We compare our local volumes with the local factors occurring in the classical Minkowski-Siegel formula. We use Siegels notation [24: I p. 367] and get.

$$A_\infty(E, E) = 2^{-n+1} \mathrm{vol}\,(\mathrm{SO}_n(\mathbf{R})).$$

If b is one of the matrices E, L, S then

$$\alpha_2(b, b) = 2^{n-1} \mathrm{vol}\,(\mathrm{SO}(b)(\mathbf{Z}_2))$$
$$\alpha_p(b, b) = \mathrm{vol}\,(\mathrm{SO}(b)(\mathbf{Z}_p)) \quad \text{for} \quad p \neq 2.$$

§5. The Euler–Poincaré characteristics of the fixpoint set

We know from 1.3 that for each class $b \in H^1(\mathfrak{g}, \Gamma)$ we have a fixpoint set $F(b) \subset (X/\Gamma)^\tau$. If $(1, b)$ is a cocycle representing b, then

$$\chi(F(b)) = \int_{\mathrm{SO}(b)(\mathbf{R})/\Gamma(b)} e_\chi$$

by 4.5. In order to apply 4.5 we have to know vol $(SO(b)(\mathbf{R})/\Gamma(b))$. It turns out to be easier (and sufficient) to compute the volume of certain unions of such spaces. To do this we use the Minkowski–Siegel formula, i.e., the fact that the Tamagawa number of an orthogonal group is 2. Using all the local results we have obtained and the classification of §2 we finish this paper with an explicit formula for the Lefschetz number $L(\tau, \Gamma(m))$.

5.1. If $(1, b)$ is a cocycle for $H^1(\mathfrak{g}, \Gamma)$ we get by twisting from 2.2 an exact sequence

$$1 \to SO(b)(\mathbf{Q})\backslash SO(b)(\mathbf{A}) \Big/ \prod_v \Gamma_v(b) \xrightarrow{i} H^1(\mathfrak{g}, {}^b\Gamma) \xrightarrow{h_b} \coprod_v H^1(\mathfrak{g}, {}^b\Gamma_v).$$

Here we use the notation $\Gamma_v(b) = {}^b\Gamma_v^\tau$.

We describe the fixpoint set corresponding to a class c in the kernel of h_b. If $x \in SO(b)(\mathbf{A})$ represents a class c then x also represents an element in ${}^b(Sl_n(\mathbf{A})/\prod_v \Gamma_v)^\tau$. By strong approximation write $x = g \cdot \eta$ with $g \in Sl_n(\mathbf{Q})$, $\eta \in \prod_v \Gamma_v$. Then $\gamma = g^{-1} \cdot b \cdot {}^\tau g \cdot b^{-1} \in \Gamma$ gives a cocycle for $H^1(\mathfrak{g}, {}^b\Gamma)$ which by definition of i represents $i(c)$. We get

$${}^b\Gamma(\gamma) = \{u \in \Gamma / b \cdot {}^\tau(g \cdot u \cdot g^{-1})b^{-1} = g \cdot u \cdot g^{-1}\} =$$
$$= g^{-1} \cdot SO(b)(\mathbf{Q}) \cdot g \cap \Gamma = g^{-1}(g\Gamma g^{-1} \cap SO(b)(\mathbf{Q}))g.$$

Clearly $\chi({}^b\Gamma(\gamma)) = \chi(g \cdot {}^b\Gamma(\gamma) \cdot g^{-1})$. Therefore it suffices to compute vol $(SO(b)(\mathbf{R})/g \cdot \Gamma \cdot g^{-1} \cap SO(b)(\mathbf{R}))$.

5.2. We get a sum of such volumes as follows. Write $SO(b)(\mathbf{A}) = \bigcup_{i=1}^{n(b)} \prod_v \Gamma_v(b) \cdot x_i^{-1} \cdot SO(b)(\mathbf{Q})$ as a disjoint finite union with $x_i \in SO(b)(\mathrm{A})$, $i = 1, \ldots, n(b)$, representing the different classes in $SO(b)(\mathrm{Q})\backslash SO(b)(A)/\prod_v \Gamma_v(b)$. Now $\prod_v \Gamma_v(b)x_i^{-1} \cdot SO(b)(\mathbf{Q}) = x_i^{-1} \cdot x_i \prod_v \Gamma_v(b) \cdot x_i^{-1} \cdot SO(b)(\mathbf{Q})$ and $x_i(\prod_v \Gamma_v(b))x_i^{-1} \cap SO(b)(\mathbf{Q}) = x_i(\prod_v \Gamma_v)x_i^{-1} \cap SO(b)(\mathbf{Q})$. As above we write $x_i = g_i \cdot u_i$ with $g_i \in Sl_n(\mathbf{Q})$ and $u_i \in \prod_v \Gamma_v$. We get $SO(b)(\mathbf{Q}) \cap x_i \prod_v \Gamma_v(b)x_i^{-1} = SO(b)(\mathbf{Q}) \cap g_i\Gamma g_i^{-1}$ and have a decomposition

$$\prod_v \Gamma_v(b) \cdot x_i^{-1} \cdot SO(b)(\mathbf{Q}) = x_i^{-1}(SO(b)(\mathbf{R}) \times \prod_p x_{i,p} \cdot \Gamma_p(b) \cdot x_{i,p}^{-1}/SO(b)(\mathbf{Q})$$
$$\cap g_i\Gamma g_i^{-1}) \times SO(b)(\mathbf{Q}).$$

Here $x_{i,p}$ is the component at the place p of x_i. If ω is the Tamagawa measure then $\mathrm{vol}_\omega(\prod_v \Gamma_v(b) \cdot x_i^{-1} \cdot SO(b)(\mathbf{Q})/SO(b)(\mathbf{Q})) = \mathrm{vol}_\omega(SO(b)(\mathbf{R}) \times \prod_p x_{i,p} \cdot \Gamma_p(b) \cdot x_{i,p}^{-1}/SO(b)(\mathbf{Q}) \cap g_i\Gamma g_i^{-1})$. From the natural fibration

$$SO(b)(\mathbf{R}) \times \prod_p x_{i,p} \cdot \Gamma_p(b) \cdot x_{i,p}^{-1}/SO(b)(\mathbf{Q}) \cap g_i\Gamma g_i^{-1}$$
$$\to SO(b)(\mathbf{R})/SO(b)(\mathbf{Q}) \cap g_i\Gamma g_i^{-1}$$

and the invariance of the factors ω_v of ω we get

$$\mathrm{vol}_\omega\left(\prod_v \Gamma_v(b)x_i^{-1}\,\mathrm{SO}\,(b)(\mathbf{Q})/\mathrm{SO}\,(b)(\mathbf{Q})\right)=\mathrm{vol}_{\omega_\infty}(\mathrm{SO}\,(b)(\mathbf{R})/\mathrm{SO}\,(b)(\mathbf{Q})\cap g_i\Gamma g_i^{-1})\cdot\prod_p \mathrm{vol}_{\omega_p}(\Gamma_p(b)).$$

Now $\mathrm{vol}_\omega(\mathrm{SO}\,(b)(\mathbf{A})/\mathrm{SO}\,(b)(\mathbf{Q}))=2$ and we obtain finally

$$\sum_{i=1}^{n(b)}\mathrm{vol}_{\omega_\infty}(\mathrm{SO}\,(b)(\mathbf{R})/\mathrm{SO}\,(b)(\mathbf{Q})\cap g_i\Gamma g_i^{-1})=2\prod_p \mathrm{vol}_{\omega_p}(\Gamma_p(b))^{-1}.$$

5.3. We suppose now that $\Gamma=\Gamma(m)$, $m\geq 3$, is a full congruence subgroup mod m of $\mathrm{Sl}_n(\mathbf{Z})$. From 5.1, 5.2 and 4.4 we see that

$$\sum_{\substack{\gamma\in H^1(\mathfrak{g},\,\Gamma)\\ h(\gamma)=h(b)}}\chi(F(\gamma))=2(-1)^{d(b)}\binom{l}{s(b)/2}\mathrm{vol}\,(\mathrm{SO}_n(\mathbf{R}))^{-1}\prod_p \mathrm{vol}\,(\Gamma_p(b))^{-1}.$$

Here $(r(b), s(b))$ is the signature of b and $d(b)=r(b)\cdot s(b)/2$. The right hand side of the equation depends only on the image of b in $H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R}))\times\prod_p H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_p))$ and we know from §3 that only the places 2 and ∞ of $\mathbf{Q}$ contribute. Therefore we write

$$L(\tau,\Gamma)=\sum_{r_2(\gamma)=E}\chi(F(\gamma))+\sum_{r_2(\gamma)=L}\chi(F(\gamma))+\sum_{r_2(\gamma)=S}\chi(F(\gamma))$$

Here $r_2: H^1(\mathfrak{g},\Gamma_2)\to H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{Z}_2))$ is the map induced by the inclusion $\Gamma_2\to \mathrm{Sl}_n(\mathbf{Z}_2)$.

If $r_2(\gamma)=E$ then the Hasse–Witt invariants are $\varepsilon_2(\gamma)=1$ and $\varepsilon_\infty(\gamma)=(-1)^{s(\gamma)(s(\gamma)-1)/2}=1$. Therefore $s(\gamma)\equiv 0 \bmod 4$ and $(-1)^{d(\gamma)}=1$. We have

$$\sum_\gamma\binom{l}{s(\gamma)/2}=2^{l-1}$$

for $\gamma\in H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R}))$ and $\varepsilon_\infty(\gamma)=1$. We get

$$\sum_{r_2(\gamma)=E}\chi(F(\gamma))=2^l\,\mathrm{vol}\,(\mathrm{SO}_n(\mathbf{R}))^{-1}\prod_p \mathrm{vol}\,(\Gamma_p(E))^{-1}\,|r_2^{-1}(\{E\})|.$$

If $r_2(\gamma)=L$ then the Hasse–Witt invariants are $\varepsilon_2(\gamma)=-1$ and $\varepsilon_\infty(\gamma)=(-1)^{s(\gamma)(s(\gamma)-1)/2}=-1$. Therefore $s(\gamma)\equiv 2 \bmod 4$ and $(-1)^{d(\gamma)}=(-1)^n$. We have

$\sum_\gamma \binom{l}{s(\gamma)/2} = 2^{l-1}$ for $\gamma \in H^1(\mathfrak{g}, \mathrm{Sl}_n(\mathbf{R}))$ and $\varepsilon_\infty(\gamma) = -1$. We get

$$\sum_{r_2(\gamma)=L} \chi(F(\gamma)) = (-1)^n 2^l \operatorname{vol}(\mathrm{SO}_n(\mathbf{R}))^{-1} \prod_p \operatorname{vol}(\Gamma_p(L))^{-1} |r_2^{-1}(\{L\})|.$$

If $r_2(\gamma) = S$ then $n \equiv 0 \bmod 4$ and $\Gamma_2 = \mathrm{Sl}_n(\mathbf{Z}_2)$. We have $\varepsilon_2(\gamma) = 1$, $(-1)^{d(\gamma)} = 1$ and $s(\gamma) \equiv 0 \bmod 4$. Therefore we get

$$\sum_{r_2(\gamma)=S} \chi(F(\gamma)) = 2^l \operatorname{vol}(\mathrm{SO}_n(\mathbf{R}))^{-1} \prod_p \operatorname{vol}(\Gamma_p(S))^{-1}.$$

5.4. Using the local results of §3 we see that the number $L(\tau, \Gamma(m))$ has the following form. It is a product of a rational number depending elementary on n and m with some power of π and with factors of the form $\zeta(2j)$, $j \in N$, and a factor $L(\chi, l)$ if $n \equiv 2 \bmod 4$. Here $\zeta(2j)$ is the Riemann zeta function at $2j$ and $L(\chi, l)$ is the value of the L-function belonging to the non trivial character of $\mathbf{Q}(i)/\mathbf{Q}$ at l. In order to avoid irrationalities in the final formula for $L(\tau, \Gamma(m))$ we use the functional equations

$$\zeta(s) = \frac{(2\pi)^s \zeta(1-s)}{2\Gamma(s)\cos(\pi s/2)} \quad \text{and} \quad \mathrm{L}(\chi, s) = \frac{(2\pi)^s \mathrm{L}(\chi, 1-s)}{2^{2s}\Gamma(s)\sin(\pi s/2)}$$

for $s \in \mathbf{C}$. Here $\Gamma(\ )$ denotes the gamma function [11: §1]. I hope no confusion with the notation of congruence subgroups will arise.

5.5. To shorten the notation we set $l = [n/2]$ and introduce the following abbreviations. For all $2 \leq n \in \mathbf{N}$, $n \not\equiv 2 \bmod 4$ and $3 \leq m \in \mathbf{N}$ we define an integer $u(n, m)$.

$$u(n, m) = \begin{cases} 1 & \text{if } 2 \mid m \text{ and } 4 \nmid m \text{ and } n \equiv 1 \bmod 2 \text{ or} \\ & \text{if } (2, m) = 1 \text{ and } n \equiv 1 \bmod 2 \\ (-1)^{l(l+1)/2} 2^l & \text{if } 4 \mid m \text{ and } n \equiv 1 \bmod 2 \\ 2^{2l-1} - 2^{l-1} + 1 & \text{if } (2, m) = 1 \text{ and } n \equiv 0 \bmod 4 \\ 2^{n-1} & \text{if } 2 \mid m \text{ and } n \equiv 0 \bmod 4 \end{cases}$$

We define

$$(d_1, \ldots, d_l) = \begin{cases} (2, 4, \ldots, 2l) & \text{if } n = 1 + 2l \\ (2, 4, \ldots, 2l-2, l) & \text{if } n = 2l. \end{cases}$$

By $\chi: \mathbf{Z} \to \{\pm 1, 0\}$ we denote the number theoretical character with $\chi(n) = 0$ if n is even and $\chi(n) = (-1)^{(n-1)/2}$ if n is odd. We define

$$c(n, m) = \begin{cases} m^{n(n-1)/2} \prod_{p \mid m} \prod_{i=1}^{l} (1 - p^{-d_i}) & \text{if} \quad n \not\equiv 2 \bmod 4 \\ m^{n(n-1)/2} \prod_{p \mid m} (1 - \chi(p) p^{-1}) \prod_{i=1}^{l-1} (1 - p^{-d_i}) & \text{if} \quad n \equiv 2 \bmod 4. \end{cases}$$

THEOREM 5.6. *If $n \geq 2$ and $m \geq 3$ and if $\Gamma(m)$ is the full congruence subgroup mod m of $\mathrm{Sl}_n(\mathbf{Z})$ then the following hold*

$$L(\tau, \Gamma(m)) = c(n, m) \mathrm{L}(\chi, 1 - d_l) \prod_{i=1}^{l-1} \zeta(1 - d_i) \quad \textit{if} \quad n \equiv 2 \bmod 4$$

resp.

$$L(\tau, \Gamma(m)) = u(n, m) c(n, m) \prod_{i=1}^{l} \zeta(1 - d_i) \quad \textit{if} \quad n \not\equiv 2 \bmod 4.$$

The right hand side of the formula for $L(\tau, \Gamma(m))$ is computable in terms of Bernoulli numbers if $n \not\equiv 2 \bmod 4$ and in terms of generalized Bernoulli numbers if $n \equiv 2 \bmod 4$ [11: §2]. We give a list of the first few values of $L(\tau, \Gamma(m))$.

n	$L(\tau, \Gamma(3))$	$L(\tau, \Gamma(4))$
2	2	2
3	-2	2^3
4	$7 \cdot 2^2$	2^7
5	$-3^2 \cdot 2^2$	2^{11}
6	$2^3 \cdot 3^4 \cdot 7$	2^{18}
7	$2^3 \cdot 3^5 \cdot 13$	2^{26}
8	$2^4 \cdot 3^7 \cdot 11^2 \cdot 13$	2^{37}
9	$2^4 \cdot 3^{11} \cdot 13 \cdot 41$	$2^{45} \cdot 17$
10	$2^5 \cdot 3^{15} \cdot 5 \cdot 13 \cdot 41 \cdot 61$	$2^{58} \cdot 5 \cdot 17$

Concluding remark. Obviously we have an estimation $\dim_{\mathbf{Q}} H^*(\Gamma, \mathbf{Q}) \geq |L(\tau, \Gamma)|$. Together with our formula for the Lefschetz number this gives some idea how big $H^*(\Gamma, \mathbf{Q})$ is.

Using information on the cohomology of the boundary $\partial(\bar{X}/\Gamma)$ of the Borel–Serre compactification $\bar{X}/\Gamma$ of X/Γ one can try to do more. If the

Lefschetz number of τ acting on the image of the natural map $H^*(\bar{X}/\Gamma, \mathbf{Q}) \xrightarrow{\text{res}} H^*(\partial(\bar{X}/\Gamma)), \mathbf{Q})$ is known – here one has to use Schwermer's methods [20] – one can compute the Lefschetz number of τ acting on the image $H^*_!(X/\Gamma, \mathbf{Q})$ of the cohomology with compact supports $H^*_c(X/\Gamma, \mathbf{Q})$ in $H^*(X/\Gamma, \mathbf{Q})$. Now B. Speh has determined all the irreducible representations of Sl_n $(\mathbf{R})$ having non trivial cohomology [B. Speh: Bonn 1980, private communication]. Thus hopefully one has a chance to get information on multiplicities of cuspidal representations having non trivial cohomology. Naturally Langland's lifting principle is connected with all of this.

For $\mathrm{Sl}_3/\mathbf{Q}$ R. Lee and J. Schwermer are pursuing these questions.

REFERENCES

[1] BOREL, A. and SERRE, J.-P., *Théorèms de finitude en cohomologie galoisienne.* Comment. Math. Helv. *39* (1964), 111–169.

[2] BOREL, A., *Introduction aux groupes arithmétiques.* Hermann Paris: 1969.

[3] BOREL, A. and SERRE, J.-P., *Corners and arithmetic groups.* Comment. Math. Helv. *48* (1973), 436–491.

[4] BOREL, A., *Stable real cohomology of arithmetic groups.* Ann. Sci. Ecole Norm. Sup. (4). *7* (1974), 235–272.

[5] DIEUDONNÉ, J., *Sur les groupes classiques.* Hermann Paris: 1948.

[6] EICHLER, M., *Quadratische Formen und orthogonale Gruppen.* Grundl. d. Math. Wiss. Bd. LXIII Springer Berlin, Heidelberg, New York: 1952.

[7] HARDER, G., *A Gauß Bonnet formula for discrete arithmetically defined groups.* Ann. Sci. Ecole Norm. Sup. (4). *4* (1971), 409–455.

[8] HELGASON, S., *Differential geometry and symmetric spaces.* Academic Press New York, London: 1962.

[9] HIRZEBRUCH, F., *Automorphe Formen und der Satz von Riemann* Roch. Symp. Intern. Top. Alg., Universidad de Mexico, (1958), 129–144.

[10] HOPF, H. and SAMELSON, H., *Ein Satz über die Wirkungsräume geschlossener Lie'scher Gruppen.* Comment. Math. Helv. 13, (1941), 240–251.

[11] IWASAWA, K., *Lectures on p-adic L-functions.* Ann. Math. Stud. 74 Princeton: 1972

[12] JONES, B., W., *The arithmetic theory of quadratic forms.* Wiley and Sons New York: 1950.

[13] KOBAYASHI, S. and NOMIZU, K., *Foundations of differential geometry II.* Wiley (Interscience) New York, London, Sydney: 1969.

[14] LEE, R. and SZCZARBA, R. H., *On the homology and cohomology of congruence subgroups.* Inventiones math. *33* (1976), 15–63.

[15] LEE, R. and SZCZARBA, R. H., *On the torsion in $K_4(\mathbf{Z})$ and $K_5(\mathbf{Z})$.* Duke Math. J. *45* (1978), 101–129.

[16] LEE, R., *Unstable cohomology classes of* Sl_n $(\mathbf{Z})$. Proc. Nat. Acad. Sci. USA *75* (1978), 43–44.

[17] MINKOWSKI, H., *Gesammelte Abhandlungen I.* Teubner Leipzig, Berlin: 1911.

[18] ONO, T., *On Tamagawa numbers.* Proc. Symp. pure Math. IX. Amer. Math. Soc., (1966), 122–132.

[19] ROHLFS, J., *Arithmetisch definierte Gruppen mit Galoisoperation.* Inventiones math. *48* (1978), 185–205.

[20] SCHWERMER, J., *Eisensteinreihen und die Kohomologie von Kongruenzuntergruppen von* Sl_n $(\mathbf{Z})$. Bonner Mathematische Schriften 99 Bonn (1977).

[21] SERRE, J.-P., *Cohomologie Galoisienne*. Lecture Notes in Mathematics 5. Springer Berlin, Heidelberg, New York: 1965.
[22] SERRE, J.-P., *Cours d'arithmétique*. Presses Universitaires de France Paris: 1970.
[23] SERRE, J.-P., *Cohomologie des groupes discrets*. Ann. Math. Stud. *70* (1971), 77–169.
[24] SIEGEL, C., L., *Gesammelte Abhandlungen I–III*. Springer Berlin, Heidelberg, New York: 1966.
[25] SOULÉ, C., *The cohomology of* $Sl_3(\mathbf{Z})$. Topology *17* (1978), 1–22.
[26] SWAN, R., G., *The Theory of Sheaves*. Chicago Lectures In Mathematics: 1964.
[27] TAMAGAWA, T., *Adeles*. Proc. Symp. pure Math. IX. Amer. Math. Soc. (1966), 113–121.
[28] VERDIER, J. L., *Characteristique d'Euler-Poincaré*. Bull. Soc. Math. France, *101* (1973), 441–445.

Katholische Universität
Mathematisch-Geographische Fakultät
Ostenstr. 26–28
D-8078 *Eichstätt*

Received March 20, 1980/November 29, 1980

Comment. Math. Helvetici **56** (1981) 297–306 0010-2571/81/002297-10$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Funktionalabschätzungen bei quasikonformen Abbildungen mit Fredholmschen Eigenwerten

Reiner Kühnau

§1. Einleitung

In [9] wurden Abschätzungen für gewisse Funktionale bei quasikonformen Abbildungen mit ortsabhängiger Dilatationsbeschränkung hergeleitet. Diese Abschätzungen waren zwar unscharf, jedoch asymptotisch scharf für den Fall, die Dilatationsschranke $p(z) \geqq 1$ weicht wenig von 1 ab. Für $p(z) \to \infty$ degenerierten die Ungleichungen zu trivialen solchen. Im folgenden werden wir diese Ungleichungen so verbessern, daß dieser Übelstand nicht mehr auftritt. Dabei gehen jetzt zusätzlich die Fredholmschen Eigenwerte (vgl. hierzu die Darstellung in [3]) ein, da wir entscheidend neben dem Extremalprinzip in [10] die Variationscharakterisierung dieser Eigenwerte benutzen werden. Wir werden die Überlegungen an einem Beispiel dartun.

Es sei G ein Gebiet der z-Ebene ($z = x + iy$) mit dem inneren Punkt $z = \infty$ und dem Rand C. Die in G erklärte Funktion $p(z)$, die später als Dilatationsschranke auftreten wird, erfülle nebst

$$0 < m \leqq p(z) \leqq M < \infty$$

(wobei zunächst nicht notwendig $m \geqq 1$ sein muß) der Einfachheit halber einschlägige, die Anwendung des Gauß-Greenschen Integralsatzes unmittelbar ermöglichende Glattheitsvoraussetzungen, ebenso C und die unten auftretenden Funktionen. In Umgebung von $z = \infty$ sei $p(z) \equiv 1$. $\mathfrak{R}$ sei ein großer Kreis $|z| = R$, G_R der ins Innere fallende Teil von G, $\mathfrak{n}$ die Innennormale von $\mathfrak{R}$ und Außennormale bei C.

Nun sei $g(z)$ diejenige schlichte quasikonforme Abbildung von G, die die Differentialgleichung

$$g_{\bar{z}} = \nu \bar{g}_z \quad \text{mit} \quad \nu = \frac{p-1}{p+1} \tag{1}$$

erfüllt, die in $z=\infty$ durch

$$g(z)=z+\frac{a_1}{z}+\cdots \tag{2}$$

hydrodynamisch normiert ist, und die Bildrandkomponenten erzeugt, die Strecken parallel zur reellen Achse sind. Entsprechend sei $\mathfrak{g}(z)$ die in $z=\infty$ durch

$$\mathfrak{g}(z)=z+\frac{\mathfrak{a}_1}{z}+\cdots \tag{3}$$

normierte durchweg konforme Abbildung von G, bei der Strecken parallel zur reellen Achse entstehen.

Speziell für $p(z)\geqq 1$ löst nach [4] $g(z)$ das folgende Extremalproblem. Unter allen schlichten quasikonformen Abbildungen von G, deren Dilatation (= Achsenverhältnis $\geqq 1$ der Ellipsen, die Bilder der zum Punkte z konzentrischen infinitesimalen Kreise sind) stets $\leqq p(z)$ ist, und die in $z=\infty$ durch

$$z+\frac{\alpha_1}{z}+\cdots \tag{4}$$

hydrodynamisch normiert sind, wird genau für $g(z)$ das Funktional $\Re\mathfrak{e}\,\alpha_1$ maximal. $\mathfrak{g}(z)$ löst nach H. Grötzsch und R. de Possel bekanntlich das entsprechende Extremalproblem in der Klasse der konformen Abbildungen von G.

Wir wollen in vorliegender Mitteilung Ungleichungen für $\Re\mathfrak{e}\,a_1$ angeben, so daß damit explizite Abschätzungen für $\Re\mathfrak{e}\,\alpha_1$ möglich werden, was bisher nur in Sonderfällen gelang.

§2. Ein Extremalprinzip

Wir setzen

$$\varphi_0(z)=\Re\mathfrak{e}\,\mathfrak{g}(z), \tag{5}$$

$$\phi^*(z)=\Re\mathfrak{e}\,g(z)=\varphi_0(z)+\varphi^*(z). \tag{6}$$

ϕ^* erfüllt dann die Differentialgleichung

$$\operatorname{div}\left(\frac{1}{p}\operatorname{grad}\phi^*\right)=0. \tag{7}$$

Daneben betrachten wir alle "zulässigen Vergleichsfunktionen" $\varphi(z)$, für die

$$|\operatorname{grad}\varphi(z)| \leqq c\,|z|^{-2} \quad \text{für} \quad z \to \infty \tag{8}$$

mit einer Konstanten c gilt.

Entscheidendes Hilfsmittel ist für uns das folgende mit dem Dirichletschen Prinzip zusammenhängende Extremalprinzip nach [10] (in leichter Abschwächung der Voraussetzungen).

SATZ 1. *Für alle zulässigen* $\varphi(z)$ *gilt*

$$\iint\limits_G \left[\left(1-\frac{1}{p}\right)\operatorname{grad}^2(\varphi_0+\varphi)-\operatorname{grad}^2\varphi\right] dx\,dy \leqq 2\pi\,\mathfrak{Re}\,(a_1-\mathfrak{a}_1) \tag{9}$$

mit Gleichheit genau für $\varphi \equiv \varphi^* + \text{const}$.

Wir geben hier einen gegenüber [10] etwas vereinfachten *Beweis* an, der Umformungen ähnlich wie in [17] benutzt. Die linke Seite von (9) ist zunächst bei Beschränkung auf den Anteil von G_R

$$\begin{aligned}
&= \iint\limits_{G_R} \left[\left(1-\frac{1}{p}\right)\operatorname{grad}^2(\varphi_0+\varphi^*)-\operatorname{grad}^2\varphi^*\right] dx\,dy - \iint\limits_{G_R}\frac{1}{p}\operatorname{grad}^2(\varphi-\varphi^*)\,dx\,dy \\
&\quad -2\iint\limits_{G_R}\frac{1}{p}\operatorname{grad}(\varphi_0+\varphi^*)\operatorname{grad}(\varphi-\varphi^*)\,dx\,dy \\
&\quad +2\iint\limits_{G_R}\operatorname{grad}\varphi_0\operatorname{grad}(\varphi-\varphi^*)\,dx\,dy.
\end{aligned} \tag{10}$$

Nach dem Gauß-Greenschen Satze wird hier der dritte Term (ds = Bogenlänge, Durchlaufung der Kurven so, daß G_R zur Linken)

$$\begin{aligned}
&= 2\int_{\mathfrak{R}+C}\frac{1}{p}(\varphi-\varphi^*)\frac{\partial(\varphi_0+\varphi^*)}{\partial n}\,ds = -2\int_{\mathfrak{R}+C}(\varphi-\varphi^*)\,d\,\mathfrak{Im}\,g \\
&= 2\int_{\mathfrak{R}}\mathfrak{Im}\,g\,d(\varphi-\varphi^*),
\end{aligned}$$

da hier zuletzt der Anteil über C verschwindet, weil $\Im\mathfrak{m}\, g$ dort konstant ist. Ebenso wird der letzte Term in (10)

$$= -2\int_{\Re+C} (\varphi-\varphi^*)\frac{\partial\varphi_0}{\partial n}\,ds = 2\int_{\Re+C} (\varphi-\varphi^*)\,d\,\Im\mathfrak{m}\,\mathfrak{g} = -2\int_{\Re} \Im\mathfrak{m}\,\mathfrak{g}\,d(\varphi-\varphi^*).$$

In der Zusammenfassung verschwinden beide Terme für $R\to\infty$. Dies liefert schon (9), da sich in (10) der erste Term zur rechten Seite von (9) analog umrechnet, z.B. auch wie in [10].

Setzt man nun in (9) $\varphi=\gamma\psi$ und bestimmt von der linken Seite das Maximum des entstehenden Ausdruckes als Funktion von γ (bei festem ψ), dann entsteht

SATZ 2. *Für alle zulässigen ψ gilt*

$$\left[\iint_G \left(1-\frac{1}{p}\right)\operatorname{grad}\varphi_0\operatorname{grad}\psi\,dx\,dy\right]^2 \Big/ \iint_G \frac{1}{p}\operatorname{grad}^2\psi\,dx\,dy$$

$$\leqq 2\pi\,\Re\mathfrak{e}\,(a_1-\mathfrak{a}_1) - \iint_G \left(1-\frac{1}{p}\right)\operatorname{grad}^2\varphi_0\,dx\,dy \qquad (11)$$

mit Gleichheit genau für $\psi\equiv\alpha\varphi^*+\beta$, α, β Konstanten.

Insbesondere ist die rechte Seite von (11) stets $\geqq 0$ in Übereinstimmung mit [10].

Im folgenden wollen wir uns mit dem Spezialfalle beschäftigen, es ist G die ganze Ebene und es wird definiert $p(z)\equiv Q$ im Innern endlich vieler sich untereinander nicht treffender oder umschlingender fixierter geschlossener analytischer Jordankurven $\mathfrak{C}_k$, $p(z)\equiv 1$ im Komplement $\mathfrak{G}$ ($=$ Gebiet $\ni\infty$). Das Innere von $\mathfrak{C}_k$ sei $\mathfrak{G}_k$, $\mathfrak{C}=\bigcup\mathfrak{C}_k$, $\bigcup\mathfrak{G}_k=\mathfrak{G}^*$. Die Abbildung $g(z)$ hängt dann vom Parameter Q ab und wir schreiben $g(z)=g_Q(z)$, $a_1=a_1(Q)$. Für $Q\to\infty$ entsteht in bekannter Weise die konforme Parallelschlitzabbildung von $\mathfrak{G}$, die wir

$$g_\infty(z)=z+\frac{\alpha_1(\infty)}{z}+\cdots \qquad (12)$$

schreiben. $\mathfrak{G}^*$ habe den Inhalt I. Der kleinste nichttriviale Fredholmsche Eigenwert >1 zum Kurvensystem $\mathfrak{C}$ sei λ. Wir setzen noch

$$\Lambda=\frac{\lambda+1}{\lambda-1}>1. \qquad (13)$$

Hier ist jetzt $\mathfrak{g}(z) \equiv z$, $\varphi_0(z) \equiv x$ und die zulässigen ψ sind in der ganzen z-Ebene erklärt und insbesondere längs $\mathfrak{C}$ stetig und klingen entsprechend (8) ab. Aus Satz 2 wird

SATZ 2'. *Für alle zulässigen ψ gilt*

$$\left(1-\frac{1}{Q}\right)^2\left[\int_{\mathfrak{C}} \psi\, dy\right]^2 \Big/ \left[\iint_{\mathfrak{G}} \operatorname{grad}^2 \psi\, dx\, dy + \frac{1}{Q}\iint_{\mathfrak{G}^*} \operatorname{grad}^2 \psi\, dx\, dy\right]$$

$$\leqq 2\pi\, \Re\mathrm{e}\, a_1(Q) - \left(1-\frac{1}{Q}\right) I \tag{14}$$

mit Gleichheit genau für $\psi = \alpha(\Re\mathrm{e}\, g - x) + \beta$, α, β Konstanten.

Insbesondere ist die rechte Seite von (14) stets $\geqq 0$ in Übereinstimmung mit [10]. Im Grenzfalle $Q \to \infty$ entsteht ein bekanntes Extremalprinzip von Diaz und Weinstein – vgl. [14], [2] – zur Charakterisierung der dann entstehenden konformen Parallelschlitzabbildung als Extremalfunktion bzw. der virtuellen Masse (als rechter Seite von (14)).

Ersetzt man $g(z)$ durch

$$-i \cdot g_Q(iz) = z - \frac{a_1(Q)}{z} + \cdots,$$

so entsteht aus Satz 2', da dann Q durch $1/Q$ zu ersetzen ist, noch

SATZ 2''. *Für alle zulässigen ψ gilt*

$$(Q-1)^2\left[\int_{\mathfrak{C}} \psi\, dy\right]^2 \Big/ \left[\iint_{\mathfrak{G}} \operatorname{grad}^2 \psi\, dx\, dy + Q\iint_{\mathfrak{G}^*} \operatorname{grad}^2 \psi\, dx\, dy\right]$$

$$\leqq (Q-1) I - 2\pi\, \Re\mathrm{e}\, a_1(Q) \tag{15}$$

mit Gleichheit genau für $\psi \equiv \alpha(\Im\mathrm{m}\, g_Q - y) + \beta$, α, β Konstanten.

§3. Abschätzungen für das Gebietsfunktional $\Re\mathrm{e}\, a_1(Q)$

Um zu handlichen Abschätzungen für $\Re\mathrm{e}\, a_1(Q)$ zu gelangen, verwenden wir neben (14) und (15) noch die folgende Ungleichung, die sich aus der bekannten

Extremalcharakterisierung der Fredholmschen Eigenwerte (vgl. hierzuz. B. [1], [16]) ergibt:

$$\left[\iint_{\mathfrak{G}} \operatorname{grad}^2 \chi\, dx\, dy + \frac{1}{Q}\iint_{\mathfrak{G}^*} \operatorname{grad}^2 \chi\, dx\, dy\right] \Big/ \left[\iint_{\mathfrak{G}} \operatorname{grad}^2 \chi + \frac{1}{Q'}\iint_{\mathfrak{G}^*} \operatorname{grad}^2 \chi\, dx\, dy\right]$$

$$\leqq (1+Q^{-1}\Lambda^{-1})/(1+Q'^{-1}\Lambda^{-1}) \quad \text{für} \quad 0<Q'\leqq Q, \quad \text{bzw.}$$

$$\leqq (1+Q^{-1}\Lambda)/(1+Q'^{-1}\Lambda) \qquad \text{für} \quad 0<Q\leqq Q'. \tag{16}$$

Dabei sind die zulässigen χ in der ganzen z-Ebene stetig, sowohl in $\mathfrak{G}$ als auch in $\mathfrak{G}^*$ harmonisch, wobei die harmonisch Konjugierte in $\mathfrak{G}$ eindeutig ist, und die χ sind noch in $z=\infty$ regulär.

Setzen wir nun $\psi = \chi = \Re e\, g_{Q'} - x$, so kommt nach (14) und (16) für $Q' \leqq Q$

$$2\pi\, \Re e\, a_1(Q) - \left(1-\frac{1}{Q}\right)I$$

$$\geqq \left(1-\frac{1}{Q}\right)^2 \frac{1+Q'^{-1}\Lambda^{-1}}{1+Q^{-1}\Lambda^{-1}} \left[\int_{\mathfrak{C}} \psi\, dy\right]^2 \Big/ \left[\iint_{\mathfrak{G}} \operatorname{grad}^2 \psi + \frac{1}{Q'}\iint_{\mathfrak{G}^*} \operatorname{grad}^2 \psi\, dx\, dy\right]$$

$$= \frac{1+Q'^{-1}\Lambda^{-1}}{1+Q^{-1}\Lambda^{-1}} \left(\frac{1-Q^{-1}}{1-Q'^{-1}}\right)^2 \cdot \left[2\pi\, \Re e\, a_1(Q') - \left(1-\frac{1}{Q'}\right)I\right] \tag{17}$$

bzw. für $Q \leqq Q'$

$$2\pi\, \Re e\, a_1(Q) - \left(1-\frac{1}{Q}\right)I \geqq \frac{1+Q'^{-1}\Lambda}{1+Q^{-1}\Lambda} \left(\frac{1-Q^{-1}}{1-Q'^{-1}}\right)^2 \cdot \left[2\pi\, \Re e\, a_1(Q') - \left(1-\frac{1}{Q'}\right)I\right].$$

Damit haben wir den

SATZ 3. *Es ist die Funktion von* Q

$$\left(1+\frac{1}{\Lambda Q}\right)\left(1-\frac{1}{Q}\right)^{-2}\left[2\pi\, \Re e\, a_1(Q) - \left(1-\frac{1}{Q}\right)I\right] \quad \textit{monoton steigend,}$$

$$\left(1+\frac{\Lambda}{Q}\right)\left(1-\frac{1}{Q}\right)^{-2}\left[2\pi\, \Re e\, a_1(Q) - \left(1-\frac{1}{Q}\right)I\right] \quad \textit{monoton fallend.}$$

Insbesondere gilt

$$\left(1-\frac{1}{Q}\right)I+[2\pi\,\mathfrak{Re}\,a_1(\infty)-I]\left(1-\frac{1}{Q}\right)^2\left(1+\frac{\Lambda}{Q}\right)^{-1}\leqq 2\pi\,\mathfrak{Re}\,a_1(Q)$$

$$\leqq\left(1-\frac{1}{Q}\right)I+[2\pi\,\mathfrak{Re}\,a_1(\infty)-I]\left(1-\frac{1}{Q}\right)^2\left(1+\frac{1}{\Lambda Q}\right)^{-1}$$

$$\leqq\left(1-\frac{1}{Q}\right)I+[2\pi\,\mathfrak{Re}\,a_1(\infty)-I]\left(1-\frac{1}{Q}\right)^2. \tag{18}$$

Arbeiten wir analog mit (15), so kommt

SATZ 4. *Es ist die Funktion von* Q

$$\left(1+\frac{Q}{\Lambda}\right)(Q-1)^{-2}[(Q-1)I-2\pi\,\mathfrak{Re}\,a_1(Q)] \quad \textit{monoton fallend,}$$

$$(1+\Lambda Q)(Q-1)^{-2}[(Q-1)I-2\pi\,\mathfrak{Re}\;a_1(Q)] \textit{ monoton steigend.}$$

Insbesondere gilt

$$(Q-1)I-I(Q-1)^2\left(\frac{1}{\Lambda}+Q\right)^{-1}\leqq 2\pi\,\mathfrak{Re}\,a_1(Q)$$

$$\leqq(Q-1)I-I(Q-1)^2(\Lambda+Q)^{-1}. \tag{19}$$

§4. Bemerkungen

a) Mit Hilfe einer der vielen bekannten Abschätzungen für die Fredholmschen Eigenwerte kann man im konkreten Einzelfalle zu konkreten Ungleichungen für $\mathfrak{Re}\,a_1(Q)$ vordringen. Besteht $\mathfrak{C}$ z.B. aus einer Cassinischen Kurve, sind die Fredholmschen Eigenwerte sogar explizit bekannt [11]. Die für diesen Fall entstehenden Ungleichungen (18), (19) liefern nach der bekannten Quadratwurzeloperation dann Abschätzungen für den Wertebereich von $w(z)$ bei festem z, falls $w(z)$ die Klasse $\Sigma(Q)$ durchläuft. Dabei muß die so aus (19) (rechts) bzw. (18) (links) entstehende Ungleichung schärfer werden (die aus (18) rechts entstehende zumindest für große Q noch schärfer – vgl. unten Bemerkung d) als die schon in [9] unter (33) hierzu angegebene, da letztere aus (29) dort hervorgeht. Der genaue Wertebereich von $w(z)$ wurde in [6], [13] angegeben, was aber formelmäßig recht aufwendig war, so daß unscharfe Abschätzungen sinnvoll

sind. Diese Bemerkung liefert auch über den durch lineare Transformation zu erhaltenden Wertebereich von $w''(z)/w'(z)$ Verbesserungen der in [12] angegebenen Abschätzungen der in [13] durch eine transzendente Gleichung charakterisierten Rundungsschranke der Klasse $\Sigma(Q)$.

b) Aussagen ähnlicher Art wie in §3 unter Einbeziehung des Inhaltes des Bildes von $\mathfrak{G}^*$ finden sich in [10].

c) Aus (19) folgt für $Q \to \infty$

$$I/\Lambda \leqq 2\pi \,\mathfrak{Re}\, a_1(\infty) - I \leqq I\Lambda. \tag{20}$$

Da sich hier der mittlere Term als virtuelle Masse deuten läßt, entspricht diese Ungleichung einer solchen von M. Schiffer [15] (dort im räumlichen Analogon).

d) Für manche Fälle ist (19) günstiger als (18), für manche gilt das Umgekehrte. (18) ist genau dann z.B. im rechten Teil schärfer als (19), wenn

$$[2\pi \,\mathfrak{Re}\, a_1(\infty) - I]/I < \Lambda - (\Lambda^2 - 1)(Q + \Lambda)^{-1}. \tag{21}$$

Dies ist wegen (20) i.allg. für große Q der Fall.

e) Genau für $\lambda = \infty$, $\Lambda = 1$ schrumpft das durch (18) (ohne letzten Teil) bzw. (19) gelieferte Intervall für $2\pi \,\mathfrak{Re}\, a_1(Q)$ zu einem Punkt, so daß dann (etwas allgemeiner als in [4])

$$2\pi \,\mathfrak{Re}\, a_1(Q) = 2I(Q-1)/(Q+1), \tag{22}$$

und dies ist dann $= \pi S$, wobei S die in [8] eingeführte "quasikonforme Spanne" ist. Die Kurvensysteme mit $\lambda = \infty$ sind einfach charakterisierbar – vgl. [7]. Besteht $\mathfrak{C}$ aus einer Kurve, ist $\lambda = \infty$ genau dann, wenn diese eine Kreislinie ist. Man kann leicht noch in (18), (19) eine ins Einzelne gehende Diskussion des Gleichheitszeichens vornehmen, da bei den verwendeten Ungleichungen die Extremalfunktionen bekannt sind.

f) Setzt man in (9) allgemeiner $\varphi = \sum \gamma_k \psi_k$ (Ritz-Ansatz) und bestimmt dann das Maximum des entstehenden Ausdruckes (quadratische Form) als Funktion der γ_k (bei festen ψ_k), so entstehen entsprechend schärfere Formen von Satz 2.

g) Durch Einsetzen günstigerer und komplizierterer "Testfunktionen" ψ in (14), (15) kann man natürlich Satz 3 und 4 verschärfen. Dies bietet sich insbesondere für sternförmige o.ä. $\mathfrak{C}$ an entsprechend dem Vorgehen z.B. in [1], [16].

h) Wenn $\mathfrak{G}$ ein $z = \infty$ im Innern enthaltendes Gebiet ist und man kennt lediglich den Inhalt I des Komplementes, so ist es natürlich unmöglich, eine Abschätzung für den Koeffizienten α_1 in der Entwicklung (4) bei den wie dort normierten schlichten konformen Abbildungen von $\mathfrak{G}$ anzugeben. Es kann

nämlich $|\alpha_1|$ beliebig groß werden, wie schon das Äußere von Ellipsen zeigt. Es sei hier die Gelegenheit wahrgenommen für den Hinweis, daß eine solche Abschätzung möglich wird, wenn man zusätzlich Q-quasikonforme Fortsetzbarkeit ins Komplement von $\mathfrak{G}$ fordert. Denn nach (19) folgt

$$|\alpha_1| < \frac{1}{2\pi}(Q-1)I. \tag{23}$$

Diese Abschätzung ist sogar scharf, da $|\alpha_1|$ beim Äußeren geeigneter Ellipsen der rechten Seite von (23) beliebig nahe kommt – vgl. z.B. [5], Zusatzbemerkung 1.

i) Die obige Betrachtung von §2 und 3 läßt sich ganz entsprechend bei Zugrundelegung anderer Extremalprobleme (statt wie oben $\Re\mathfrak{e}\,\alpha_1 \to \max$) durchführen. Man halte sich hierbei etwa an die in [9] betrachteten Beispiele, z.B. die verallgemeinerten Grunskyschen Koeffizientenbedingungen.

j) Die wie in §2 und 3 oder allgemeiner nach der vorigen Bemerkung i) bei irgendeinem Funktional entstehenden Ungleichungen lassen sich natürlich auch umgekehrt deuten als Abschätzungen für den Fredholmschen Eigenwert λ. Benutzt man z.B. Satz 4, rechte Seite von (19), so ergibt sich mit Hilfe der am Ende von §1 erwähnten Extremaleigenschaft der

SATZ 5. *Ist $w(z)$ eine schlichte konforme Abbildung von $\mathfrak{G}$ mit der Entwicklung* (4) *in $z=\infty$, die ins Komplement von $\mathfrak{G}$ noch Q-quasikonform fortsetzbar ist, so gilt*

$$\Lambda \geqq \frac{Q \cdot 2\pi\,\Re\mathfrak{e}\,\alpha_1 - (Q-1)I}{(Q \cdot 1)I - 2\pi\,\Re\mathfrak{e}\,\alpha_1} \tag{24}$$

Arbeitet man mit den Grunskyschen Koeffizientenbedingungen, erhält man eine allgemeinere Ungleichung, in der man noch die auftretenden freien komplexen Parameter optimieren kann.

SCHRIFTTUM

[1] AHLFORS, L. V., *Remarks on the Neumann–Poincaré integral equation.* Pacif. J. Math. 2, 271–280 (1952).

[2] BERGMAN, S. and M. SCHIFFER, *Kernel functions and elliptic differential equations in mathematical physics.* New York 1953.

[3] GAIER, D., *Konstruktive Methoden der konformen Abbildung.* Berlin–Göttingen–Heidelberg 1964.

[4] KÜHNAU, R., *Wertannahmeprobleme bei quasikonformen Abbildungen mit ortsabhängiger Dilatationsbeschränkung.* Math. Nachr. *40*, 1–11 (1969).

[5] —— *Bemerkungen zu den GRUNSKYschen Gebieten.* Math. Nachr. *44*, 285–293 (1970).
[6] —— *Eine funktionentheoretische Randwertaufgabe in der Theorie der quasikonformen Abbildungen.* Indiana Univ. Math. J. *21*, 1–10 (1971).
[7] —— *Eine Integralgleichung in der Theorie der quasikonformen Abbildungen.* Math. Nachr. *76*, 139–152 (1977).
[8] ——*Die Spanne von Gebieten bei quasikonformen Abbildungen.* Arch. Rat. Mech. Anal. *65*, 299–303 (1977).
[9] ——, *Verzerrungsaussagen bei quasikonformen Abbildungen mit ortsabhängiger Dilatationsbeschränkung und ein Extremalprinzip der Elektrostatik in inhomogenen Medien.* Comm. Math. Helv. *53*, 408–428 (1978).
[10] —— *Einige Verzerrungsaussagen bei quasikonformen Abbildungen endlich vielfach zusammenhängender Gebiete.* L'Enseign. math. *24*, 189–201 (1978).
[11] —— *Bemerkung zu einer klassischen Arbeit von J. Blumenfeld und W. Mayer.* Mathematische Forschung, Band 1, Berlin 1979; hier S. 161–164.
[12] KÜHNAU, R. und E. HOY, *Abschätzung des Wertebereichs einiger Funktionale bei quasikonformen Abbildungen.* Bull. Soc. Sciences Lettres Lódź *29*, 4 (1979).
[13] KÜHNAU, R. und B. THÜRING, *Berechnung einer quasikonformen Extremalfunktion.* Math. Nachr. *79*, 99–113 (1977).
[14] PÓLYA, G. and G. SZEGÖ, *Isoperimetric inequalities in mathematical physics.* Princeton 1951.
[15] SCHIFFER, M., *Problèmes aux limites et fonctions propres de l'équation intégrale de Poincaré et Fredholm.* Comptes rendus Paris *245*, *18–21 (1957).*
[16] SPRINGER, G., *Fredholm eigenvalues and quasiconformal mapping.* Acta Math. *111*, 121–142 (1964).
[17] WEISEL, J., *Numerische Ermittlung quasikonformer Abbildungen mit finiten Elementen.* Numer. Math. (im Druck).

Sektion Mathematik der Martin-Luther-Universität Halle-Wittenberg
DDR-402 Halle an der Saale

Eingegangen den 17. Oktober 1980

Comment. Math. Helvetici **56** (1981) 307–326 0010–2571/81/002307–20$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

On the secondary classes of foliations with trivial normal bundles

STEVEN HURDER[1]

The secondary classes of a smooth foliation on a manifold are invariants of the concordance class of the foliation. A central problem in the theory of foliations is to determine which of these invariants are non-zero, and also to determine their geometric significance. In this paper we examine one aspect of how the secondary classes of a foliation depend upon the topology of the foliated manifold. This dependence is used to construct many examples of foliated manifolds with framed normal bundles having non-trivial secondary invariants. This implies the non-triviality of the corresponding universal secondary classes in $H^*(F\Gamma_G)$.

There are two basic ideas which we exploit to deduce our theorems. The first regards the construction of a universal map $k_*: H^*(W(\mathfrak{G})_q) \to H^*(F\Gamma_G)$ using the theory of foliation invariants as given by Kamber–Tondeur [17]. There is a map of spectral sequences converging to k_* which, when combined with a knowledge of the connectivity of $F\Gamma_G$, leads to the proofs of Theorems 1 and 3 below. The second idea is a type of permanence theorem which is related to a concept first explored by Kamber–Tondeur [18] and Lazarov [19]. Coupled with the results on evaluating the secondary classes via residues of Baum–Bott [1], Lazarov–Pasternack [21] and Heitsch [13], this yields Theorems 2, 4, and 5.

In a previous paper [14], Sullivan's theory of minimal models was used to show that many of the secondary classes of foliated manifolds are non-trivial. This was done by an analysis of the homotopy properties of $F\Gamma_G$; more precisely, the Whitehead product action of $\pi_*(BG)$ on $\pi_*(F\Gamma_G)$ was used to construct spherical classes in $H_*(F\Gamma_G)$ which support universal secondary classes. Then following an idea of Haefliger [11], this yields much information on the higher homotopy groups of $F\Gamma_G$. The present paper significantly extends the results of [14] on the independence of the secondary cohomology invariants. However, it is no longer possible to assert that the classes in $H^*(F\Gamma_G)$ are spherically supported. The contrast between the two approaches is seen as follows. There is a sequence of fibrations $G \to F\Gamma_G \to B\Gamma_G \to BG$, and the first map defines a G-action on the space $F\Gamma_G$. Assume that $F\Gamma_G$ is N-connected. Then for $n \leq N$ there is a smash

[1] Supported in part by NSF grant MCS 77-18723 A03.

product operation

$$\pi_n(G) \times \pi_m(F\Gamma_G) \xrightarrow{\wedge} \pi_{n+m}(F\Gamma_G)$$

which is induced from this G-action. The minimal model techniques of [14] essentially analyze this smash product, and establish that it is non-trivial for many n and m. Our present method is to study the more general G-action, which defines for all n a map

$$H_n(G) \times H_m(F\Gamma_G) \to H_{n+m}(F\Gamma_G).$$

This $H_*(G)$-action is shown to be non-trivial for many n and m by the use of the Permanence Theorems 3.3 and 3.5 of §3.

Our main results are stated below as Theorems 1 to 5. The terminology is explained in §1. The techniques required for the proofs of these theorems are developed in §§2 and 3; the proofs are then given in §4. The author would like to thank James Heitsch for pointing out a gap in the original proof of Theorem 3.3, and whose questions led to a general reformulation of the results of this paper.

The first two theorems deal with the invariants of a smooth codimension q framed foliation on a manifold M. Recall that the secondary invariants in this case form a real vector space $H^*(W_q)$. A class in $H^*(W_q)$ is said to be rigid if its value in $H^*(M)$ must remain constant under deformations of the given foliation on M; it is said to be variable otherwise [12].

THEOREM 1. (a) *Let $q \geq 4$ be even. Then there exists a framed codimension q foliation on a compact manifold M such that the set of rigid secondary classes*

$$\mathcal{R}_q = \{y_2 y_{2i_2} \cdots y_{2i_s} c_q \mid 1 < i_2 < \cdots < i_s\}$$

is independently realized in $H^*(M)$.

(b) *Let $q = 4k-2$ with $k \geq 2$. Then there exists a framed codimension q foliation on a compact manifold M such that the set of rigid secondary classes*

$$\mathcal{R}'_q = \{y_{2k} y_{2i_2} \cdots y_{2i_s} c_{2k} \mid k < i_2 < \cdots < i_s\}$$

is independently realized in $H^*(M)$. Further, the set $\mathcal{R}_q$ is mapped to zero in $H^*(M)$ for this foliation.

(c) *For q even, the set $k_*(\mathcal{R}_q) \subseteq H^*(F\Gamma_q)$ is linearly independent. For $q \equiv 2 \bmod (4)$, the set $k_*(\mathcal{R}_q \cup \mathcal{R}'_q) \subseteq H^*(F\Gamma_q)$ is linearly independent.*

It can be shown that many of these rigid classes in $H^*(F\Gamma_q)$ are supported on spherical cycles in $H_*(F\Gamma_q)$, [16]. It is also interesting to contrast Theorem 1 with the results of Pittie [23].

Our second theorem considers the variable secondary classes. This uses the results of Heitsch [13].

THEOREM 2. *For each $q \geq 3$, there exists a set $\mathcal{V}_q \subseteq H^{2q+1}(W_q)$, having at least 3 elements, and a family of framed codimension q foliations on a compact manifold M such that the set of variable classes*

$$\overline{\mathcal{V}_q} = \{y_I c_J = y_i y_{2i_2} \cdots y_{2i_s} c_J \mid y_i c_J \in \mathcal{V}_q \text{ and } i < 2i_2 < \cdots < 2i_s\}$$

is mapped to a set of independent and variable classes in $H^*(M)$. Consequently, the set $k_*(\overline{\mathcal{V}}_q) \subseteq H^*(F\Gamma_q)$ is independently variable.

Note that Theorems 1 and 2 give an alternate proof of some of the results of Fuks [6] on $H^*(F\Gamma_q)$.

A foliation which has a metric on its normal bundle that is parallel along the leaves is said to be Riemannian [24]. The characteristic classes of these foliations have been studied by Kamber–Tondeur [18] and Lazarov–Pasternack [20], [21].

THEOREM 3. *There exists a Riemannian foliation with SO_q-framed normal bundle on a manifold M such that all of the secondary classes are independently realized in $H^*(M)$. Thus, the universal map $k_*: H^*(W(so(q))_{[q/2]}) \to H^*(F\Gamma_{SO_q})$ is injective.*

It was shown by Lazarov and Pasternack that for $q = 2, 3 \bmod (4)$, there are secondary classes for Riemannian foliations of codimension q which can vary under deformations of the foliation. These classes are called variable, and using the results of [21] we show:

THEOREM 4. *If $q = 2, 3 \bmod 4$, then there exists a family of Riemannian foliations of codimension q with SO_q-framed normal bundles on a compact manifold M such that all of the variable secondary classes are independent and variable in $H^*(M)$. Hence, they are independently variable in $H^*(F\Gamma_{SO_q})$.*

. A foliation which is locally defined by submersions into $\mathbf{C}^n$, with biholomorphic transition functions, is called a complex foliation. The secondary invariants of a complex foliation with trivial normal bundle form a complex vector space

$H^*(W(Gl(n, \mathbf{C}))_n)$. Examples of complex foliations on S^{2n+1} with variable secondary classes have been constructed by Baum–Bott [1]; their results are used to show:

THEOREM 5. *For $n \geq 2$, there exists a family of complex foliations with $Gl_n\mathbf{C}$-framed normal bundles on a compact manifold M such that all of the variable secondary classes in $H^*(W(Gl(n, \mathbf{C}))_n)$ are independent and variable in $H^*(M; \mathbf{C})$. Thus, they are independently variable in $H^*(F\Gamma_n^C; \mathbf{C})$.*

§1. Definitions and basic constructions

All manifolds are assumed to be smooth (C^∞) and paracompact; all foliations and maps between manifolds are assumed to be smooth. For a topological space X we use $H^*(X)$ to denote the singular cohomology of X with real coefficients. For a real number x, let $[x]$ denote the greatest integer $\leq x$.

We work with the general notion of a G-foliation as this encompasses the usual idea of a foliation, when $G = Gl_q$, and the special cases of Riemannian and complex foliations when $G = O_q$ and $G = Gl_n\mathbf{C}$ respectively. Recall that a foliation of codimension q on a manifold M is an integrable subbundle $\mathcal{F}$ of the tangent bundle TM, where the normal bundle $Q = TM/\mathcal{F}$ has rank q. Let G be a closed subgroup of Gl_q. We say that $\mathcal{F}$ is a G-foliation if Q has a G-structure which is invariant under the natural parallelism along the leaves of $\mathcal{F}$. This is equivalent to the existence of the following data [10], [17]:

(i) a model manifold B of dimension q with a G-structure on TB;

(ii) an open covering $\{U_\alpha\}$ of M and submersions $\phi_\alpha : U_\alpha \to B$ with $\mathcal{F}|_{U_\alpha} = \ker d\phi_\alpha$ such that the local transition functions $\gamma_{\alpha\beta}$ preserve the G-structure on B.

An O_q-foliation is one with a bundle-like metric in the sense of Reinhart [24] on the normal bundle Q, and is also called a Riemannian foliation.

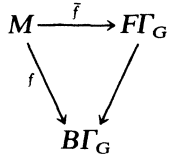
If the model manifold B can be taken to be $\mathbf{R}^q$ with the natural flat G-structure, then the G-foliation is said to be G-integrable. For example, a complex foliation is a $Gl_n\mathbf{C}$-integrable $Gl_n\mathbf{C}$-foliation of real codimension $q = 2n$. In a similar way, by an appropriate choice of group G and model manifold B, one obtains the notion of conformal, flag or symplectic foliations.

Associated to a G-foliation $\mathcal{F}$ on M is a principal G-bundle $\pi : P \to M$. This is constructed using the local defining charts $\{\phi_\alpha : U_\alpha \to B\}$ of $\mathcal{F}$ and G-bundle $P' \to B$ associated to the G-structure on TB. Locally, $P|_{U_\alpha}$ is defined to be the pullback of P' via ϕ_α. These local bundles are then patched together using

condition (ii) above. The bundle $P \to M$ is trivial if it admits a section $s: M \to P$; from a section s one can construct a bundle isomorphism $P \cong M \times G$.

A section s of the associated bundle $P \to M$ of $\mathcal{F}$ is called a G-framing of the normal bundle of the foliation. The pair $(\mathcal{F}, s)$ is called a G-framed G-foliation.

Given a group G and manifold B with a G-structure on TB, Haefliger shows in [10] that there exists a classifying space for G-foliations modeled on B. In our examples the model B is dictated by the choice of G, so by abuse of notation this classifying space is denoted $B\Gamma_G$. Let $\nu: B\Gamma_G \to BG$ classify the G-structure on $B\Gamma_G$. If $f: M \to B\Gamma_G$ classifies a G-foliation on M, then $\nu \circ f$ classifies the associated G-bundle $P \to M$. Let $F\Gamma_G$ denote the homotopy theoretic fiber of ν. If a G-foliation on M has trivial associated bundle $P \to M$, then the map $\nu \circ f$ is homotopic to a constant. Hence there is induced a map $\bar{f}$ giving a homotopy commutative diagram



In this sense we say that $F\Gamma_G$ classifies G-foliations with trivial associated G-bundle.

For the group Gl_q with model $B = \mathbf{R}^q$, we adopt the standard notation $B\Gamma_q = B\Gamma_{Gl_q}$ and $F\Gamma_q = F\Gamma_{Gl_q}$. For complex foliations, we use $F\Gamma_n^{\mathbf{C}} = F\Gamma_{Gl_n\mathbf{C}}$. The space $F\Gamma_{SO_q}$ is sometimes denoted in the literature by $FR\Gamma_q$.

Next assume that M has a fixed G-foliation $\mathcal{F}$ of codimension q and there is given a section s of $P \to M$. We outline the construction of the secondary classes for $\mathcal{F}$ using truncated Weil algebras. This approach is the most useful for our purposes; it is developed in detail by Kamber and Tondeur in [17]. When G is either Gl_q or $Gl_n\mathbf{C}$, the secondary classes which are obtained coincide with those constructed by Bernstein and Rosenfel'd [2] or Bott and Haefliger [3], [4].

Let $\mathfrak{G}$ denote the Lie algebra of G. Recall that the Weil algebra $W(\mathfrak{G}) = \Lambda\mathfrak{G}^* \otimes S\mathfrak{G}^*$ is a differential graded algebra (dga), where $\Lambda\mathfrak{G}^*$ is the exterior algebra on the dual space $\mathfrak{G}^*$ of $\mathfrak{G}$ (each $x \in \mathfrak{G}^*$ has degree 1) and $S\mathfrak{G}^*$ is the symmetric algebra on $\mathfrak{G}^*$ (each $x \in \mathfrak{G}^*$ has degree 2). This algebra has a decreasing filtration by differential ideals

$$F^{2p}W(\mathfrak{G}) = \bigoplus_{r \geq 2p} \Lambda\mathfrak{G}^* \otimes S^r\mathfrak{G}^*.$$

The quotient dga $W(\mathfrak{G})_q = W(\mathfrak{G})/F^{2q+2}W(\mathfrak{G})$ is called a truncated Weil algebra.

The secondary invariants of the G-foliation $\mathcal{F}$ are defined as follows. Choose a Bott connection ω on the G-bundle $P \to M$, [17]. If $\mathcal{E}(P)$ denotes the deRham complex of P, then the Chern–Weil construction yields a *dga* homomorphism $k(\omega): W(\mathfrak{G})_q \to \mathcal{E}(P)$. The fundamental result is that the induced map in cohomology $k(\omega)_*$ is independent of the choice of the Bott connection ω and depends only on the G-concordance class of $\mathcal{F}$ (see Theorem 4.43 of [17]). Further, the map $k(\omega)_*$ is functorial with respect to pull-backs along maps $f: M' \to M$ which are transverse to $\mathcal{F}$.

Recall that a section $s: M \to P$ is given. Set

$$\Delta^s_* = s^* \circ k(\omega)_* : H^*(W(\mathfrak{G})_q) \to H_{DR}(M).$$

The elements in the image of Δ^s_* are defined to be the secondary classes for the G-framed G-foliation $\mathcal{F}$ on M.

For certain groups G, the *dga* homomorphism $k(\omega)$ annihilates the ideal $F^{2q'+2}W(\mathfrak{G})$ where $q' = [q/2]$. The corresponding map $k(\omega): W(\mathfrak{G})_{q'} \to \mathcal{E}(P)$ is then used to construct the characteristic homomorphism Δ^s_*. This is the case for both Riemannian and complex foliations [17].

An additional useful property of the Chern–Weil homomorphism $k(\omega): W(\mathfrak{G})_q \to \mathcal{E}(P)$ is that it preserves the basic filtration on both algebras and hence induces a map of spectral sequences. The basic filtration on $\mathcal{E}(P)$ is that given by

$$F^p\mathcal{E}(P) = \mathcal{E}(P) \wedge \pi^*\mathcal{E}^p(M),$$

where $\pi: P \to M$. The curvature of any connection is a horizontal form so we have $k(\omega)F^{2p}W(\mathfrak{G})_q \subseteq F^{2p}\mathcal{E}(P)$. The induced map of the spectral sequences associated to these filtrations is denoted by

$$E_r k(\omega): E_r^{2p,t}W(\mathfrak{G})_q \to E_r^{2p,t}(P).$$

We conclude this section by exhibiting a set of cocycles in $W(\mathfrak{G})_q$ which span $H^*(W(\mathfrak{G})_q)$. Let $A(\mathfrak{G})_q$ be the standard model for $W(\mathfrak{G})_q$. More precisely, define $A(\mathfrak{G})_q = \Lambda(y_1, \ldots, y_r) \otimes I(\mathfrak{G})_q$ where $I(\mathfrak{G})$ denotes the *dga* of $ad(\mathfrak{G})$-invariant polynomials on $\mathfrak{G}$. This is identified with a symmetric algebra $I(\mathfrak{G}) \cong \mathbf{R}[c_1, \ldots, c_r]$; each c_j is homogeneous and of even degree. We assume that $\deg c_i \leq \deg c_j$ if $i \leq j$. The element $y_j \in W(\mathfrak{G})$ is defined to be the suspension of c_j and satisfies $dy_j = c_j$. There is an inclusion $A(\mathfrak{G})_q \subseteq W(\mathfrak{G})_q$ which induces an isomorphism in cohomology [8], [17]. In fact, if we give $A(\mathfrak{G})_q$ the basic filtration from $W(\mathfrak{G})_q$, then there are natural isomorphisms

$$A(\mathfrak{G})_q \cong E_2A(\mathfrak{G})_q \cong E_2W(\mathfrak{G})_q.$$

For the group Gl_q, the algebra $A(gl(q,\mathbf{R}))_q$ is commonly denoted by W_q.

We adopt the standard notation for the elements of $A(\mathfrak{G})_q$: Given index sets $I=(i_1,\ldots,i_s)$ and $J=(j_1,\ldots,j_r)$ with each $j_k \geq 0$, let

$$y_I c_J = y_{i_1} \cdots y_{i_s} \otimes c_1^{j_1} \cdots c_r^{j_r}.$$

According to Vey [7], the following set of cocycles (which are called admissible) gives a basis for $H^*(W(\mathfrak{G})_q)$:

$$\{y_I c_J \mid 1 \leq i_1 < \cdots < i_s \leq r;\ \deg y_{i_1} c_J > 2q;\ l < i_1 \Rightarrow j_l = 0\}.$$

For a complex $H^*(W(\mathfrak{G})_{[q/2]})$ one modifies this by requiring that $\deg y_{i_1} c_J > 2[q/2]$.

The admissible cocycles are divided into two types. The classes in the image of the natural map $H^*(W(\mathfrak{G})_{q+1}) \to H^*(W(\mathfrak{G})_q)$ are said to be *rigid*. This terminology is used because the rigid secondary classes give invariants for a foliation $\mathcal{F}$ which are constant under deformations of $\mathcal{F}$ [12]. With respect to the Vey basis, the rigid classes correspond to the set

$$\{y_I c_J \mid y_I c_J \text{ admissible and } \deg y_{i_1} c_J > 2q+2\}.$$

The remaining admissible cocycles, for which $\deg y_{i_1} c_J = 2q+1$, are said to be *variable*.

§2. The universal characteristic map

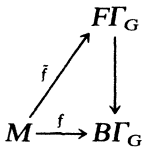
In this section we review the construction of the universal map $k_*: H^*(W(\mathfrak{G})_q) \to H^*(F\Gamma_G)$, and also analyze the geometric significance of the lifts $\tilde{f}: M \to F\Gamma_G$ of a given map $f: M \to B\Gamma_G$.

First recall that there is a fibration sequence $G \to F\Gamma_G \to B\Gamma_G \to BG$. We therefore can consider $F\Gamma_G$ as the total space of the G-bundle over $B\Gamma_G$ constructed as follows. Let $EG \to BG$ denote the universal G-bundle. Define $F\Gamma_G$ to be the pullback in the diagram

$$\begin{array}{ccc} G & & G \\ \downarrow & & \downarrow \\ F\Gamma_G & \xrightarrow{\tilde{\nu}} & EG \\ \downarrow & & \downarrow \\ B\Gamma_G & \xrightarrow{\nu} & BG \end{array}$$

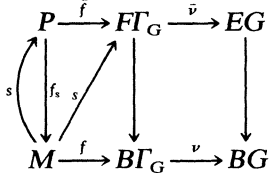
It is then standard that there exists a weak homotopy equivalence from $F\Gamma_G$ to the fiber of the fibration induced from $\nu: B\Gamma_G \to BG$.

Given a map $f: M \to B\Gamma_G$, the set of lifts $\tilde{f}$ of f



corresponds to the set of maps $M \to G$: Given one lift $\tilde{f}_0$, every other lift $\tilde{f}$ is obtained by choosing some function $\varphi: M \to G$ and setting $\tilde{f} = \tilde{f}_0 \cdot \varphi$, where for $m \in M$, the element $\varphi(m) \in G$ acts on the fiber over $f(m) \in B\Gamma_G$. Note that the composition $\nu \circ f: M \to BG$ fixes a representation of the bundle $P \to M$, and the lift $\bar{\nu} \circ \tilde{f}_0$ defines a G-framing of P. Hence a different lift $\tilde{f} = \tilde{f}_0 \cdot \varphi$ corresponds to a change in the framing $\bar{\nu} \circ \tilde{f}_0$.

Conversely, given a map $f: M \to B\Gamma_G$ we construct the associated G-bundle as the pullback via $\nu \circ f$ of $EG \to BG$. This gives a commutative diagram



(2.1)

A section s of $P \to M$ then gives rise to a lift $f_s = \bar{f} \circ s$ of the map f.

Remark 2.2. The above does not show that $F\Gamma_G$ is the classifying space of G-framed G-foliations; it only shows that given any pair $(\mathcal{F}, s)$ on M there exists a natural map $f_s: M \to F\Gamma_G$. The problem arises with the following type of example. Consider the foliation by points on $M = (\mathbf{R}^{n+1} - \{0\}) \times \mathbf{R}^{q-n-1}$ with the canonical flat G-structure. This is classified by a unique homotopy class of maps $M \to F\Gamma_G$. However, the set of homotopy classes of G-framings corresponds to $\pi_n(G)$ which may not be trivial.

In spite of the above remark, we still have the following basic result.

THEOREM 2.3. *There exists a universal map*

$$k_*: H^*(W(\mathfrak{G})_q) \to H^*(F\Gamma_G).$$

That is, given a G-framed G-foliation $(\mathcal{F}, s)$ on M, there exists a classifying map $f_s: M \to F\Gamma_G$ such that $\Delta^s_* = f_s^* \circ k_*$.

If $\mathcal{F}$ is either a Riemannian or complex foliation on M, then the truncation index is taken to be $[q/2]$ instead of q.

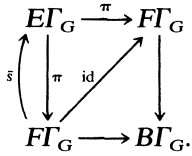
Proof. Given a G-foliation $\mathcal{F}$ on M with associated G-bundle $P \to M$, the Chern–Weil homomorphism $k(\omega)_*: H^*(W(\mathfrak{G})_q) \to H^*(P)$ depends only on the concordance class of $\mathcal{F}$ and is functorial with respect to pullbacks along maps $f: N \to M$ transverse to $\mathcal{F}$. Thus, there exists a universal Chern–Weil homomorphism for the universal bundle $F\Gamma_G \to B\Gamma_G$. We denote this map by $k_*: H^*(W(\mathfrak{G})_q) \to H^*(F\Gamma_G)$, and show that it satisfies the conditions of the theorem.

Consider a G-framed G-foliation $(\mathcal{F}, s)$ on M. Let $\mathcal{F}$ be classified by a map $f: M \to B\Gamma_G$. This defines a pullback square as in (2.1), where the section s yields a map $f_s = \bar{f} \circ s$. For a Bott connection ω on $P \to M$ we then have

$$\Delta^s_* = s^* \circ k(\omega)_* = s^* \circ \bar{f}^* \circ k_* = (f_s)^* \circ k_*.$$

Thus k_* is universal as claimed. ■

Remark 2.4. If one chooses a section s of the G-bundle $E\Gamma_G \to F\Gamma_G$ induced by the composition $F\Gamma_G \to B\Gamma_G \to BG$, then there is a corresponding map $\Delta^s_*: H^*(W(\mathfrak{G})_q) \to H^*(F\Gamma_G)$. However, this map depends upon the choice of the section s. There exists a canonical G-framing $\bar{s}$ of $E\Gamma_G \to F\Gamma_G$, which is defined by the diagram



The map k_* then corresponds to $\Delta^{\bar{s}}_*$, for $\Delta^{\bar{s}}_* = \bar{s}^* \circ k(\bar{\omega})_* = \bar{s}^* \circ \pi^* \circ k_* = \mathrm{id}^* \circ k_* = k_*$.

The map $\Delta^{\bar{s}}_*$ is the universal secondary map with respect to the canonical G-framing on $F\Gamma_G$ that is constructed in [3], [4], [7], [20]. The advantage of using k_* is that it is a bundle map and has a map of spectral sequences which converge to it. This property is not so obvious for $\Delta^{\bar{s}}_*$.

The next concept is needed in order to formulate the permanence theorem for variable classes in the next section.

DEFINITION 2.5. Let $\{(\mathcal{F}_\alpha, s_\alpha) \mid \alpha \in \mathcal{A}\}$ be a family of framed foliations on M. A set of secondary classes $\{z_1, \ldots, z_d\} \subseteq H^n(W(\mathfrak{G})_q)$ is said to be independent and variable (with respect to this family) if the union over $\alpha \in \mathcal{A}$ of the images of the evaluation homomorphisms

$$\bigoplus_{j=1}^{d} \Delta^{s_\alpha}(z_j): H_n(M; Z) \to \mathbf{R}^d \tag{2.6}$$

generates $\mathbf{R}^d$ additively.

The point of this definition is that if $\{z_1, \ldots, z_d\}$ is an independent and variable set with respect to some family of framed G-foliations, then the corresponding universal classes $k_*\{z_1, \ldots, z_d\}$ in $H^n(F\Gamma_G)$ are independently variable. That is, the evaluation homomorphism

$$\bigoplus_{j=1}^{d} k_*(z_j): H_n(F\Gamma_G; Z) \to \mathbf{R}^d \tag{2.7}$$

is an epimorphism. This follows because k_* is universal, so the image of (2.7) contains an additive basis for $\mathbf{R}^d$ and thus is all of $\mathbf{R}^d$.

A useful criterion for when a set is independent and variable is given by the next result, which from the Appendix to Baum–Bott [1].

PROPOSITION 2.8. *Suppose that $V \subseteq \mathbf{R}^d$ is a connected subvariety of dimension ≥ 1 which is not contained in any codimension one hyperplane of $\mathbf{R}^d$. Then V generates $\mathbf{R}^d$ additively.*

For example, if $F = (f_1, \ldots, f_d): \mathbf{R}^\lambda \to \mathbf{R}^d$ is a polynomial mapping such that $\{1, f_1, \ldots, f_d\}$ is a linearly independent set of functions, then the image of F generates $\mathbf{R}^d$ additively. For each of the foliations constructed in [1], [13], [21], a parameter space $\mathcal{A} = \mathbf{R}^\lambda$ and a set $\{z_1, \ldots, z_d\}$ is given such that the corresponding evaluation map (2.6) contains in its image a variety V as in (2.8). So it follows from Proposition 2.8 that $\{z_1, \ldots, z_d\}$ is an independent and variable set.

§3. Permanence theorems

It was seen in section 2 that given a map $f: M \to B\Gamma_G$, the various framings s of the normal bundle give rise to maps $f_s: M \to F\Gamma_G$ which lift f. It is natural to ask how the image of the map

$$\Delta^s_* = f_s^* \circ k_*: H^*(W(\mathfrak{G})_q) \to H^*(M)$$

depends upon the choice of G-framing s. The answer to this question is implicit in

the literature [5], [12]; the formulation we use appears in [20]. Our Theorems 3.3 and 3.5 then follow from an application of this result.

For each $1 \leq j \leq r$, the cocycle $c_j \in I(G)$ determines a cohomology class in $H^{m_j+1}(BG)$. Let $\tau_j \in H^{m_j}(G)$ denote the suspension of this class. Next suppose there is given a G-foliation $\mathcal{F}$ on M with two sections s, s' of $P \to M$. Define a mapping $\varphi: M \to G$ by the relation $s = s' \cdot \varphi$, where for $m \in M$, $\varphi(m)$ acts on the fiber over m. Then for each j there is a cohomology class $\varphi^*(\tau_j) \in H^{m_j}(M)$. The next theorem was proved for SO_q-foliations in [Theorem 4.2; 20]. The proof carries over directly to G-foliations to give:

THEOREM 3.1. *Let $\mathcal{F}$ be a G-foliation on M and suppose that two sections s, s' of $P \to M$ are given. Given a cocycle $y_I c_J = y_{i_1} \cdots y_{i_n} c_J \in W(\mathfrak{G})_q$, then*

$$\Delta^{s'}_*(y_I c_J) = \prod_{j=1}^{n} \{\Delta^s(y_{i_j}) - \varphi^*(\tau_{i_j})\} \Delta^s(c_J).$$

The meaning of the right-hand side of this equation is that it is to be expanded multiplicatively and the terms collected to get a well-defined class in $H^*(M)$.

COROLLARY 3.2. *Let $V_G \subseteq H^*(W(\mathfrak{G})_q)$ be the subspace spanned by the following admissible cocycles:*

(i) $y_i c_J \in W(\mathfrak{G})_q$, *or*

(ii) $y_I c_J \in W(\mathfrak{G})_q$ *such that $\tau_i \in H^{m_i}(G)$ is zero for each $i \in I$. Then for any G-foliated manifold M with trivial associated G-bundle, the characteristic map $\Delta^s_*|_{V_G}: V_G \to H^*(M)$ is independent of the choice of G-framing s.*

Proof. For a class $y_i c_J$ of type (i), by (3.1) we have $\Delta^{s'}_*(y_i c_J) = \Delta^s_*(y_i c_J) - \varphi^*(\tau_i)\Delta^s_*(c_J)$ in $H^*(M)$. The normal bundle of $\mathcal{F}$ is trivial so $\Delta^s_*(c_J) = 0$, and thus the difference term $\varphi^*(\tau_i) \cdot \Delta^s_*(c_J) = 0$.

For a class $y_I c_J$ of type (ii), use (3.1) to expand $\Delta^{s'}_*(y_I c_J)$, then note that each $\varphi^*(\tau_{i_j}) = 0$ by assumption. ■

It is useful to know which of the classes $\tau_i \in H^{m_i}(G)$ are non-zero. Let $K \subseteq G$ be a maximal compact subgroup and assume that G is connected. There is a natural restriction map $r: I(G) \to I(K)$ for which there is a commutative diagram, where σ is a map of vector spaces:

$$\begin{array}{ccc} I(K) & \xrightarrow{\sigma} & H^*(K) \\ \uparrow{\scriptstyle r} & & \uparrow{\scriptstyle \cong} \\ I(G) & \xrightarrow{\sigma} & H^*(G). \end{array}$$

The kernel of the map $\sigma: I(K) \to H^*(K)$ is exactly the ideal of products $I^+(K) \cdot I^+(K) \subseteq I(K)$. Thus $\tau_i \neq 0$ precisely when c_i is not mapped to zero under the composition $I(G) \xrightarrow{r} I(K) \to QI(K)$, where $QI(K) = I(K)/\{I^+(K) \cdot I^+(K)\}$. For example, when G is compact each $\tau_i \neq 0$. When $G = Gl_q$, then $K = O_q$ and we have $\{\tau_2, \tau_4, \ldots, \tau_{2[q-1/2]}\}$ is part of an algebra basis for $H^*(Gl_q)$. Also, in this case $\tau_{\text{odd}} = 0$.

The next two theorems are examples of a phenomenon called permanence in [19]. There are a few related remarks to make. Many of the first examples of foliated manifolds with non-trivial secondary classes were constructed using generalized flat bundles. There is a relationship between the secondary classes for these foliations and the characteristic classes of the associated flat bundle. This relationship can be used to show that if one secondary class is non-trivial, then a whole family must be non-trivial. This principle was used by Kamber–Tondeur [18] for locally homogeneous foliations; it was further explored by Lazarov [19], and also used by Heitsch [13] to obtain many new results on the variation of the secondary classes. The theorems we give apply to an arbitrary foliation $\mathcal{F}$ on a manifold M. In one sense, we use the natural flat bundle structure on the leaves of the foliation $\mathcal{F}$. In order to assure the leaves have enough cohomological complexity to carry the classes we construct, we extend them by the group G. This forces us to lift the foliation $\mathcal{F}$ to the product $M \times G$ where we obtain our results.

The following notation is very convenient. Given a subset $\mathcal{S} \subseteq H^*(W(\mathfrak{G})_q)$ of admissible cocycles, define the *extension* of $\mathcal{S}$ to be the set of admissible cocycles

$$\bar{\mathcal{S}} = \{y_I c_J \mid I = I' \cup I'' \text{ with } y_{I'} c_J \in \mathcal{S},\ \tau_{I''} \neq 0,$$
$$\text{and if } i \in I' \text{ has } \tau_i \neq 0 \text{ then } \deg y_j > \deg y_i \text{ for all } j \in I''\}.$$

Recall that V_G is the set defined in Corollary 3.2. If $\mathcal{S} \subseteq V_G$, then a typical element of $\bar{\mathcal{S}}$ is either of the form $y_I c_J = y_{i_1} \cdots y_{i_s} c_J$ with $\tau_{i_1} \neq 0$ and $\deg y_{i_1} > \deg y_{i_j}$ for $2 \leq j \leq s$, or of the form $y_I c_J = y_{I'} y_{I''} c_J$ with $\tau_i = 0$ for each $i \in I'$, $y_{I'} c_J \in \mathcal{S}$ and $\tau_{I''} \neq 0$. It follows that if $y_I c_J \in \bar{\mathcal{S}}$ for $\mathcal{S} \subseteq V_G$, then there is a unique decomposition $I = I' \cup I''$ with $y_{I'} c_J \in \mathcal{S}$, $\tau_{I''} \neq 0$ and the degree of $y_{I''}$ maximal. For this decomposition we define $D(y_I c_J) = \deg y_{I''}$. This gives a function $D: \bar{\mathcal{S}} \to \{0, 1, 2, \ldots\}$ which is used in the proofs of the permanence theorems:

THEOREM 3.3. *Let $(\mathcal{F}, s)$ be a G-framed G-foliation on M and $\mathcal{S} \subseteq V_G$ a subset such that $\Delta^s_*(\mathcal{S}) \subseteq H^*(M)$ is a linearly independent set. Then there exists a G-framed G-foliation $(\mathcal{F}', s')$ on $N \stackrel{\text{def}}{=} M \times G$ such that the extended set $\Delta^{s'}_*(\bar{\mathcal{S}}) \subseteq H^*(N)$ is linearly independent.*

By Theorem 2.3 this has the consequence:

COROLLARY 3.4. *If $k_*(\mathcal{S}) \subseteq H^*(F\Gamma_G)$ is a linearly independent set for $\mathcal{S} \subseteq V_G$, then the extended set $k_*(\bar{\mathcal{S}}) \subseteq H^*(F\Gamma_G)$ is also linearly independent.*

In particular, in order to conclude that $k_*: H^*(W(\mathfrak{G})_q) \to H^*(F\Gamma_G)$ is injective, it suffices to show that the set V_G is realized independently for some G-framed G-foliation.

THEOREM 3.5. *Let $\{(\mathcal{F}_\alpha, s_\alpha) \mid \alpha \in \mathcal{A}\}$ be a family of G-framed G-foliations on M and $\mathcal{V} \subseteq V_G$ a subset which is independent and variable with respect to these foliations. Then there exists a family of G-framed G-foliations $\{(\mathcal{F}'_\alpha, s'_\alpha) \mid \alpha \in \mathcal{A}\}$ on $N = M \times G$ such that the extended set $\bar{\mathcal{V}}$ is independent and variable. Consequently, the set $k_*(\bar{\mathcal{V}}) \subseteq H^*(F\Gamma_G)$ is independently variable.*

Given that the set $\mathcal{V}$ is independently variable for the foliations on M, one can conclude, in addition, that for a suitable integer m with $1 \leq m \leq \dim H^*(G)$, the set $\bar{\mathcal{V}}$ is independently variable for an $\mathbf{X}_{j=1}^{m} \mathcal{A}$-parameter family of framed foliations on $\dot{\bigcup}_{j=1}^{m} N$.

Proof of 3.3. Define the G-foliation $\mathcal{F}'$ on $N = M \times G$ to be the pullback of $\mathcal{F}$ via the projection $p: N \to M$. A leaf of $\mathcal{F}'$ is thus a leaf of $\mathcal{F} \times G$. Define $\varphi: N \to G$ to be the projection, so that $\varphi^*: H^*(G) \to H^*(N) \cong H^*(M) \otimes H^*(G)$ is the inclusion. The associated G-bundle to $\mathcal{F}'$ is defined by the pullback

$$\begin{array}{ccc} P' & \longrightarrow & P \\ \downarrow & & \downarrow \\ N & \xrightarrow{p} & M \end{array} \qquad s$$

We construct two sections of $P' \to N$. The first, $\bar{s}$, is that induced by $s \circ p$. The second, s', is defined by the relation $s' \cdot \varphi = \bar{s}$. For the pair $(\mathcal{F}', s')$ we show that $\Delta^{s'}_*(\bar{\mathcal{S}}) \subseteq H^*(N)$ is a linearly independent set.

For the section $\bar{s}$, note that $\Delta^{\bar{s}}_* = p^* \circ \Delta^s_*$ by functoriality. Then from Theorem 3.1 we have

$$\begin{aligned} \Delta^{s'}_*(y_I c_J) &= \sum_{I' \cup I'' = I} \pm \Delta^{\bar{s}}_*(y_{I'} c_J) \cdot \tau_{I''} \\ &= \sum_{I' \cup I'' = I} \pm p^*(\Delta^s_*(y_{I'} c_J)) \cdot \tau_{I''}. \end{aligned} \tag{3.6}$$

Now suppose that $\sum_{(I,J) \in \mathcal{I}} \lambda_{(I,J)} y_I c_J \neq 0$ is an element in the span of $\bar{\mathcal{S}}$ which is

in the kernel of $\Delta^{s'}_*$. We show this yields a contradiction to the hypothesis of (3.3). We may assume that the elements $y_I c_J$ for $(I, J) \in \mathcal{J}$ are all distinct and that each $\lambda_{(I,J)} \neq 0$. Then by (3.6) we obtain

$$0 = \sum_{\substack{(I,J)\in\mathcal{J} \\ I'\cup I''=I}} \pm \lambda_{(I,J)} p^* \Delta^s_*(y_{I'} c_J) \cdot \tau_{I''}. \tag{3.7}$$

Define a non-negative integer $n_0 = \max\{D(y_I c_J) \,|\, (\mathcal{J}, J) \in \mathcal{J}\}$ where $D(y_I c_J)$ is the integer defined above for a class $y_I c_J \in \bar{\mathcal{V}}$. Let $\mathcal{J}_0 \subseteq \mathcal{J}$ be the subset

$$\mathcal{J}_0 = \{(I, J) \in \mathcal{J} \mid D(y_I c_J) = n_0\}.$$

Note that if $y_I c_J \in \bar{\mathcal{V}}$ and $I = I' \cup I''$ has $\deg y_{I''} = n_0$ and $y_{I'} c_J \in \mathcal{V}$, then by the definition of n_0, either $\tau_{I''} = 0$ or $I' = \emptyset$ and $\Delta^s_*(y_{I'} c_J) = 0$.

Thus, in the sum (3.7) we can split off a factor

$$0 = \sum_{(I,J)\in\mathcal{J}_0} \pm \lambda_{(I,J)} p^* \Delta^s_*(y_{I'_0} c_J) \cdot \tau_{I''_0}. \tag{3.8}$$

We may assume that $\tau_{I''_0}$ is the same class in $H^{n_0}(G)$ for all $(I, J) \in \mathcal{J}_0$, for otherwise we decompose the sum in (3.8) further. Hence we conclude that

$$0 = \sum_{(I,J)\in\mathcal{J}_0} \pm \lambda_{(I,J)} \Delta^s_*(y_{I'_0} c_J)$$

in $H^*(M)$. But the classes $y_{I'_0} c_J$ are all distinct for $(I, J) \in \mathcal{V}_0$, since the classes $y_I c_J = y_{I'_0} c_J \cdot y_{I''_0}$ are assumed to be distinct. This implies that $\Delta^s_*(\mathcal{V})$ is not an independent set, contrary to hypothesis. ■

Proof of (3.5). Define a family of G-framed G-foliations $\{(\mathcal{F}'_\alpha, s'_\alpha) \mid \alpha \in \mathcal{A}\}$ on $N = M \times G$ as in the proof of Theorem 3.3. We must show that the union over $\alpha \in \mathcal{A}$ of the images of $\Delta^{s'_\alpha}_*(\bar{\mathcal{V}}) : H_*(N; Z) \to \mathbf{R}^{\#\bar{\mathcal{V}}}$ generates $\mathbf{R}^{\#\bar{\mathcal{V}}}$ additvely. The proof will be by induction on the function $D : \bar{\mathcal{V}} \to \{0, 1, 2, \ldots\}$ which defines the "excess" degree of a class $y_I c_J \in \bar{\mathcal{V}}$ from being in $\mathcal{V}$. For each $n \geq 0$, set $\bar{\mathcal{V}}_n = \{y_I c_J \in \bar{\mathcal{V}} \mid D(y_I c_J) \leq n\}$. Note that $\bar{\mathcal{V}}_0 = \mathcal{V}$ and $\bar{\mathcal{V}}_d = \bar{\mathcal{V}}$ where $d = \dim H^*(G)$. In the following, we view $H_*(M; Z) \otimes H_*(G; Z)$ as a subgroup of $H_*(N; Z)$ and will consider $\Delta^{s'_\alpha}_* \bar{\mathcal{V}}$ restricted to it.

It is given that the union over $\alpha \in \mathcal{A}$ of the images of

$$\Delta^{s'_\alpha}_* \bar{\mathcal{V}}_n : \bigoplus_{m\leq n} H_*(M; Z) \otimes H_m(G; Z) \to \mathbf{R}^{\#\bar{\mathcal{V}}_n} \tag{3.9}$$

generates $\mathbf{R}^{\#\bar{\mathcal{V}}_n}$ when $n=0$. Now induct on n: assume this holds for some $n \geq 0$; then we show it holds for $n+1$.

Let $\mathcal{D}_n = \bar{\mathcal{V}}_{n+1} - \bar{\mathcal{V}}_n$ and assume $\mathcal{D}_n$ is not empty. First, note that

$$\Delta^{s'\alpha}_* \bar{\mathcal{V}}_n : H_*(M; Z) \otimes H_{n+1}(G; Z) \to \mathbf{R}^{\#\bar{\mathcal{V}}_n} \tag{3.10}$$

is the zero map. For any $y_I c_J \in \bar{\mathcal{V}}$ we can use Theorem 3.1 to obtain the expansion

$$\Delta^{s'\alpha}_*(y_I c_J) = \sum_{I = I' \cup I''} \pm \Delta^{s\alpha}_*(y_{I'} c_J) \otimes \tau_{I''}. \tag{3.11}$$

Given any $y_I c_J \in \bar{\mathcal{V}}_n$ such that in the sum (3.11) there is a term with $\deg y_{I''} = n+1$, then by the definition of $D(y_I c_J)$ either $\tau_{I''} = 0$ or $\Delta^{s\alpha}_*(y_{I'} c_J) = 0$. Therefore, $\Delta^{s\alpha}_*(y_I c_J)$ vanishes on the domain of (3.10).

To complete the inductive step, it suffices to show the union over $\alpha \in \mathcal{A}$ of the images of

$$\Delta^{s'\alpha}_* \mathcal{D}_n : H_*(M; Z) \otimes H_{n+1}(G; Z) \to \mathbf{R}^{\#\mathcal{D}_n} \tag{3.12}$$

generates $\mathbf{R}^{\#\mathcal{D}^n}$. The assumption $\mathcal{V} \subseteq V_G$ implies that for any $y_I c_J \in \mathcal{D}_n$, there is a unique decomposition $I = I'_0 \cup I''_0$ such that $\tau_{I''_0} \neq 0$, $D(y_I c_J) = \deg y_{I''_0} = n+1$ and $y_{I'_0} c_J \in \mathcal{V}$. So Formula (3.11) implies

$$\Delta^{s'\alpha}_*(y_I c_J) = \pm \Delta^{s\alpha}_*(y_{I'_0} c_J) \otimes \tau_{I''_0}$$

when restricted to the domain of the map (3.12). Since the classes $\tau_{I''_0} \in H_{n+1}(G; Z)$ are as independent as the $y_{I''_0}$ and it is given that $\mathcal{V}$ is independent and variable for the foliations on M, it follows that the images of (3.12) generate the range. ■

§4. Proofs of the main theorems

In this section we combine some known properties of $F\Gamma_G$ with the machinery and results of sections 1 to 3 to prove Theorems 1 through 5.

Proof of Theorem 1. Our construction of foliated manifold with non-zero rigid classes begins with a theorem of Mather and Thurston [27], [22], [25]:

THEOREM 4.1. *For $q \geq 1$ the space $F\Gamma_q$ is $(q+1)$-connected.*

For $q=2$ this implies the map $\nu : B\Gamma_2 \to BO_2$ is 4-connected. Let $g : S^2 \times S^2 \to BO_2$ be a map such that $g_* : H_4(S^2 \times S^2) \to H_4(BO_2)$ is an isomorphism. Since ν is

4-connected, we can lift g to a map $\bar{g}: S^2 \times S^2 \to B\Gamma_2$. Note that $(\nu \circ \bar{g})^*(p_1) \neq 0$, where $p_1 \in H^4(BO_2)$ is the first pontrjagin class.

Part (a). Given $q = 2q' \geq 4$, set $X = (S^2 \times S^2) \times \cdots \times (S^2 \times S^2)$, q' factors. The map $\bar{g}$ induces a natural map $f: X \to B\Gamma_q$ with $(\nu \circ f)^*(p_1^{q'}) \neq 0$. To obtain a foliated manifold, we let $M' = X \times S^n$ with $\alpha: M' \to X$ the projection, where $n > 2q$ is chosen large enough so that $\nu \circ f \circ \alpha$ determines a splitting of TM' into $F' \oplus Q$. Note that $\alpha^*: H^m(X) \to H^m(M')$ is an isomorphism for $m \leq 2q$. By Theorem 1 of Thurston [28], the data consisting of the splitting $TM' = F' \oplus Q$ and the map $f \circ \alpha$ determines a foliation $\mathcal{F}'$ on M' which is classified by a map homotopic to $f \circ \alpha$.

Let $P' \to M'$ be the Gl_q-bundle associated to $\mathcal{F}'$. Choose a reduction of P' to an O_q-bundle with total space M:

$$\begin{array}{ccc} M & \subseteq & P' \\ & {\scriptstyle \pi} \searrow & \swarrow \\ & M' & \end{array}$$

The projection π induces a foliation $\mathcal{F}$ on M. Let $P \to M$ be the Gl_q-bundle associated to $\mathcal{F}$, and define a section s of this bundle by

$$\begin{array}{ccc} P & \xrightarrow{\bar{\pi}} & P' \\ {\scriptstyle s}\uparrow\downarrow & \subseteq \nearrow & \downarrow \\ M & \xrightarrow{\pi} & M' \end{array}$$

Choose a Bott connection ω' on $P' \to M'$. Let ω be the pull-back of ω' to $P \to M$. Then

$$\Delta^s_* = s^* \circ k(\omega)_* = s^* \circ \bar{\pi}^* \circ k(\omega')_* = k(\omega')_*,$$

where we use the isomorphism $H^*(M) \cong H^*(P')$. By Theorem 3.3, it suffices to show that $k(\omega')_*(y_2 c_q) \neq 0$.

Consider the map of spectral sequences of §1 associated to $k(\omega')$:

$$E_2 k(\omega'): W_q^{2p,t} \to E_2^{2p,t}(P') \cong H^{2p}(X) \otimes H^t(O_q),$$

where the last isomorphism holds for $2p \leq 2q$. Denote the elements in the image $E_2 k(\omega')$ by:

$$E_2 k(\omega')(y_i) = \bar{y}_i \in H^{2i-1}(O_q)$$
$$E_2 k(\omega')(c_i) = \bar{c}_i \in H^{2i}(X).$$

We want to show that the cocycle $\bar{y}_2\bar{c}_q$ in $E_2^{2q,3}(P')$ survives to $E_\infty(P')$. To see this, first not that $d_{2r}\bar{y}_{2r} = \bar{c}_{2r}$, and each class $\bar{c}_{2j}$ is a multiple of $\bar{c}_2^j$. Thus $E_5^{2p,t}(P') \cong E_\infty^{2p,t}(P')$ for $2p \leq 2q$, and $E_5^{2p,t}(P') \cong E_2^{2p,t}(P')/\bar{c}_2 \cdot E^{2p-4,t}$. The class $\bar{y}_2\bar{c}_q$ survives in this quotient, so we are done.

Part (b). For $q = 4k-2$, choose a map $g: S^{4k} \to BO_q$ such that $g^*(p_k) \neq 0$. Then by Theorem 4.1 there exists a lift $\bar{g}: S^{4k} \to B\Gamma_q$ of g. Choose a map $\alpha: M' \to S^{4k}$ as in the proof of (a) with $\alpha^*: H^m(S^{4k}) \to H^m(M')$ an isomorphism for $m \leq 2q$. The proof then proceeds exactly as in (a). Note that $\Delta_*^s(y_2 y_I c_q) = 0$ in $H^*(M)$ since this class comes from the cocycle

$$\bar{y}_2\bar{y}_I\bar{c}_q \in E_2^{2q,t}(P) \cong H^{2q}(S^{4k}) \otimes H^t(O_q) = 0.$$

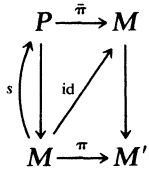
Part (c). This follows directly from Theorem 2.3. ■

The proof of Theorem 3 is very similar to that of Theorem 1 so we give it next.

Proof of Theorem 3. Let X be a connected, finite CW complex of dimension q with $g: X \to BSO_q$ a map such that $g^*: H^m(BSO_q) \to H^m(X)$ is an isomorphism for $0 \leq m \leq q$. Let U be an open manifold which retracts to X; let M' be the total space of the $\mathbf{R}^q$-vector bundle over U induced by $g: U \to BSO_q$. We need the following analog of Theorem 4.1 which is proved in [15]:

THEOREM 4.2. *For $q \geq 1$ the space $F\Gamma_{SO_q}$ is $(q-1)$-connected.*

By this theorem the map g lifts to a map $f: M' \to B\Gamma_{SO_q}$. The bundle TM' has a canonical q-dimensional subbundle tangent to the fibers of $M' \to U$. Thus by the Gromov–Phillips Theorem as applied by Haefliger [10], these data determine a Riemannian foliation $\mathcal{F}'$ on M'. Let M be the total space of the principal SO_q-bundle associated to $\mathcal{F}'$. The bundle map $\pi: M \to M'$ induces a Riemannian foliation $\mathcal{F}$ on M. There is a canonical framing s of the normal bundle defined by the diagram



Let ω' be a Bott connection on $M \to M'$. Then as we have seen before,

$$\Delta^s_* = s^* \circ k(\omega) = s^* \circ \bar{\pi} \circ k(\omega')_* = k(\omega')_*.$$

Now the map of spectral sequences associated to $k(\omega')$ gives

$$E_2 k(\omega') : A(so(q))_{[q/2]} \to E_2^{2p,t}(M) \cong H^{2p}(X) \otimes H^t(SO_q).$$

By the construction of the foliation $\mathcal{F}'$ on $M' \simeq X$, this map is an isomorphism. Thus $k(\omega')_*$ is also an isomorphism. ■

The proofs of the remaining theorems all use the permanence result, Theorem 3.5.

Proof of Theorem 2. For $q \geq 3$ and odd, Heitsch constructs in [Theorem 6.3; 13] a family of framed foliations $\{(\mathcal{F}_\alpha, s_\alpha) \mid \alpha \in \mathcal{A}\}$ of codimension q on a compact manifold M. Also, a set $\mathcal{V}_q \subseteq H^{2q+1}(W_q)$ of admissible cocycles is given such that $\Delta^s_*(\mathcal{V}_q) \subseteq H^*(M)$ varies independently.

Note that a class of degree $2q+1$ in $H^*(W_q)$ must be of the form $y_i c_J$ and thus lies in the set V_{Gl_q}. So we can apply Theorem 3.5 to get a family of foliations $\{(\mathcal{F}'_\alpha, s'_\alpha) \mid \alpha \in \mathcal{A}\}$ on $N = M \times Gl_q$ with the desired properties. Finally, we note that the inclusion $M \times O_q \subseteq M \times Gl_q$ is transverse to the $\mathcal{F}'_\alpha$, so we can use the compact manifold $M \times O_q = N$.

For $q \geq 2$ and even, let $\mathcal{V}_q \subseteq H^{2q+1}(WO_q) \subseteq H^{2q+1}(W_q)$ be the subset which is realized independently in [Theorem 6.2; 13]. It is shown in [16] that this set can be realized independently on a compact manifold M with a family of framed foliations. Then proceed as in the case q odd. ■

Proof of Theorem 4. For $q = 4k-2$, let $\mathcal{V}_q$ be the basis of $H^{q+1}(W(so_q)_{[q/2]})$ which consists of admissible classes. In Section 3 of [21], Lazarov and Pasternack construct a family of SO_q-framed Riemannian foliations on the compact Lie group U_{2k} for which the set $\mathcal{V}_q$ is independent and variable. By Theorem 3.5, the extended set $\bar{\mathcal{V}}_q$ is independent and variable with respect to a family of SO_q-framed foliations on $M = U_{2k} \times SO_q$. The set $\bar{\mathcal{V}}_q$ is a basis of the variable classes, so this establishes the theorem in this case.

For $q = 4k-1$, start with the family of SO_q-framed Riemannian foliations on $U_{2k} \times S^1$ which are obtained from the family on U_{2k} of [21] by letting S^1, with the standard metric, be normal to the leaves. Then take $M = U_{2k} \times S^1 \times SO_q$ and apply Theorem 3.5. ■

Proof of Theorem 5. Let $\mathcal{V}_n$ be the basis of $H^{2n+1}(W(Gl_n\mathbf{C})_n)$ consisting of admissible cocycles. Let $\{(\mathcal{F}_\alpha, s_\alpha) \mid \alpha \in \mathcal{A}\}$ be the family of $Gl_n\mathbf{C}$-framed complex

foliations on the compact Lie group U_{n+1} which is derived from Example 11.1 of [1]. It follows from [1] that $\mathcal{V}_n$ is an independent and variable set with respect to this family. Set $M = U_{n+1} \times U_n$. By Theorem 3.5 the extended set $\bar{\mathcal{V}}_n$ is independent and variable with respect to a family $\{(\mathcal{F}'_\alpha, s'_\alpha) | \alpha \in \mathcal{A}\}$ on M. The set $\bar{\mathcal{V}}_n$ is a basis for the variable classes in $H^*(W(Gl_n\mathbf{C})_n)$, so this finishes the proof. ■

REFERENCES

[1] P. BAUM and R. BOTT, Singularities of holomorphic foliations, *J. Differential Geometry 7* (1972), 279–342.

[2] I. N. BERNSTEIN and B. I. ROZENFEL'D, Characteristic classes of foliations, *Funkcional Anal. i Priložen 6* (1972), 68–69.

[3] R. BOTT, Lectures on characteristic classes and foliations, *Springer Lecture Notes 279* (1972), 1–94.

[4] —— and A. HAEFLIGER, On characteristic classes of Γ-foliations, *Bull. A.M.S. 78* (1972), 1038–1044.

[5] S. S. CHERN and J. SIMONS, Characteristic forms and geometric invariants, *Annals of Math. 99* (1974), 48–69.

[6] D. B. FUKS, Non-trivialité des classes caracteristiques de $\mathfrak{G}$-structures, *C.R. Acad. Sci. Paris 284* (1977), 1017–1019 and 1105–1107.

[7] C. GODBILLON, Cohomologie d'algèbres de Lie de champs de vecteurs formels, *Séminaire Bourbaki* 421.01–421.19 (1972/73).

[8] W. GREUB, S. HALPERIN and R. VANSTONE, *Connections, Curvature and Cohomology*, vol. III, Academic Press (1976).

[9] A. HAEFLIGER, Feuilletages sur les variétés ouvertes, *Topology 9* (1970), 183–194.

[10] ——, Homotopy and integrability, *Springer Lecture Notes 197* (1971), 133–166.

[11] ——, Whitehead products and differential forms, *Springer Lecture Notes 652* (1978), 13–24.

[12] J. HEITSCH, Deformations of secondary characteristic classes. *Topology 12* (1973), 381–388.

[13] ——, Independent variation of secondary classes, *Annals of Math. 108* (1978), 421–460.

[14] S. HURDER, Dual homotopy invariants of G-foliations, *Topology* (to appear).

[15] ——, On the homotopy and cohomology of the classifying space of Riemannian foliations, *Proceedings A.M.S. 81* (1981), 485–489.

[16] ——, On the classifying space of smooth foliations, (preprint).

[17] F. KAMBER and P. TONDEUR, Foliated Bundles and Characteristic Classes, *Springer Lecture Notes 493* (1975).

[18] ——, Non-trivial characteristic invariants of homogeneous foliated bundles, *Ann. Sci. Ecole Norm. Sup. 8* (1975), 433–486.

[19] C. LAZAROV, A permanence theorem for exotic classes, *J. Differential Geometry 14* (1979), 475–486.

[20] —— and J. PASTERNACK, Secondary characteristic classes for Riemannian foliations, *J. Differential Geometry 11* (1976), 365–385.

[21] ——, Residues and characteristic classes for Riemannian foliations, *J. Differential Geometry 11* (1976), 599–612.

[22] D. MCDUFF, Foliations and monoids of embeddings, *Geometric Topology*, ed. by J. C. Cantrell, Academic Press (1979), 429–444.

[23] H. PITTIE, The secondary characteristic classes of parabolic foliations, *Comment. Math. Helv. 53* (1979), 601–614.

[24] B. REINHART, Foliated manifolds with bundle-like metrics, *Annals of Math. 69* (1959), 119–132.

[25] G. SEGAL, Classifying spaces related to foliations, *Topology 17* (1978), 367–382.
[26] H. SHULMAN, Secondary obstructions to foliations, *Topology 13* (1974), 177–183.
[27] W. THURSTON, Foliations and groups of diffeomorphisms, *Bull. A.M.S. 80* (1974), 304–307.
[28] ——, The theory of foliations of codimension greater than one, *Comment. Math. Helv. 49* (1974), 214–231.

The Institute for Advanced Study,
School of Mathematics
Princeton, N.J. 08540
U.S.A.

Received December 22, 1980/April 28, 1981.

Comment. Math. Helvetici **56** (1981) 327–338 0010–2571/81/003327–12$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Heat kernel estimates and lower bound of eigenvalues

SHIU-YUEN CHENG* and PETER LI

§0. Introduction

Let M^n be a n-dimensional compact Riemannian manifold with possibly empty boundary, ∂M. We consider the eigenvalues of the Laplace operator on M with either Dirichlet or Neumann boundary condition. The famous asymptotic formulas of H. Weyl give

$$\lambda_k^{n/2} \sim A(n)k/V \tag{0.1}$$

and

$$\mu_k^{n/2} \sim A(n)k/V \tag{0.2}$$

where λ_k is the kth non-zero eigenvalue on M (with Neumann boundary condition if $\partial M \neq \varnothing$) and μ_k is the kth eigenvalue on M with Dirichlet boundary condition. In view of this, it is interesting to investigate if the eigenvalues λ_k and μ_k possess lower estimates which are compatible with the Weyl formulas.

When M is assumed to be a domain in a simply connected non-positively curved manifold, the answer was obtained in [7, 3], where they showed

$$\mu_k^{n/2} \geq B(n)k/V$$

for some constant $B(n)$ which only depend on n. Lower estimates for λ_k on general Riemannian manifold were derived in [5]. However the order of λ_k was not optimal, instead he proved

$$\lambda_k^{n-1} \geq C(M)kV^{-2(n-1)/n}$$

where $C(M)$ is a constant depending on M. More precisely, $C(M)$ depends on the Sobolev constant of M which in term can be bounded from below by the

* Research partially supported by an NSF grant and a Sloan Fellowship

diameter, d; lower bound of the Ricci curvature, $(n-1)K$; and the volume of M, V.

Later in [6], using a completely different method Li and Yau showed that if M is non-negatively curved then

$$\lambda_k^{n/2} \geq B'(n)k/V$$

for k not less than a constant that depends on the geometry of M. Also if the Ricci curvature of M is non-negative, then

$$\lambda_k^{n/2} \leq B''(n)(k+1)/V$$

for all k, $B'(n)$ and $B''(n)$ are constants depending only on n. For general manifolds, lower bounds for small eigenvalues are more complicated. They proved that

$$\lambda_k^{n/2} \geq C'(M)k/V, \text{ for all } k,$$

where $C'(M)$ depends on d, V, n, lower bound of the sectional curvature, and radius of convexity of M. Upper bound in the form

$$\lambda_k^{n/2} \leq C(K, d, n)[(k+1)/V] + C(K, n)$$

was also derived. The constants in this case depends on the quantities described.

In view of the upper estimates for λ_k and previous results, it is not surprising to speculate that

$$\lambda_k^{n/2} \geq C(K, d, n, V) \cdot k$$

for all k. In fact, the purpose of the work is to show the above inequality. More precisely, we will prove that

$$\lambda_k^{n/2} \geq \alpha(n) C_1^{n/2} k/V$$

where $\alpha(n)$ is a constant depending only on n, and C_1 is the Sobolev constant from Lemma 1. Together with the estimates of the isoperimetric constant obtained in [2], we conclude that C_1 has a lower bound depending on n, K, d, and V.

In a case where M is a manifold with boundary, we showed that

$$\mu_k^{n/2} \geq \alpha(n) C_2^{n/2} k/V \quad \text{and} \quad \lambda_k^{n/2} \geq \alpha(n) C_1^{n/2} k/V$$

where μ_k and λ_k are eigenvalues of the Dirichlet and Neumann boundary problems respectively. Since C_2 is equivalent to the isoperimetric constant, when M is a domain in a simply connected non-positively curved manifold, we recovered the results in [3].

It is interesting to consider the special case where M is a compact manifold without boundary of constant -1 curvature. If $n \geq 4$, we showed that

$$\lambda_k^{n/2} \geq \alpha(n) k / V^{n^2+n+1}$$

where $\alpha(n)$ is a constant depending only on n. This inequality is known to be false for the case $n = 2$ and $k \leq 2g-3$ (see [8]).

§1. Sobolev inequalities

The standard Sobolev inequality for compact manifold M with or without boundary asserts that there exists a constant C_0 depending only on M, such that

$$\int_M |\nabla f| \geq C_0 \left(\inf_{a \in \mathbf{R}} \int_M |f-a|^{n/(n-1)} \right)^{(n-1)/n} \tag{1.1}$$

for any $f \in H_{1,1}(M)$.

In this section we will derive a weaker form of (1.1) which is more suitable for our application.

LEMMA 1. *Let M^n be a compact Riemannian manifold with possibly empty boundary. Then there exists a constant C_1 depending on M such that*

$$\int_M |\nabla f|^2 \geq C_1 \left(\int_M f^2 \right)^{(2+n)/n} \left(\int_M |f| \right)^{-4/n}$$

for any $f \in H_{1,2}(M)$ *which satisfies*

$$\int_M f = 0.$$

Proof. The proof is divided into two cases.

Case 1. When $n \geq 3$, in [5] we derived the following weaker form of (1.1), namely

$$\int_M |\nabla f|^2 \geq C_0^2 \left[\alpha(n) \left(\int_M |f|^{2n/(n-2)} \right)^{(n-2)/n} - \beta(n) V^{-2/n} \int_M f^2 \right] \tag{1.2}$$

for $f \in H_{1,2}(M)$, where $\alpha(n)$ and $\beta(n)$ are constants depending only on n and $V =$ volume of M. However if $\int_M f = 0$, then the Poincaré inequality gives

$$\int_M |\nabla f|^2 \geq \lambda_1 \int f^2 \tag{1.3}$$

On the other hand, it is known that (see [5])

$$\lambda_1 \geq \gamma(n) C_0^2 V^{-2/n} \tag{1.4}$$

for $\gamma(n) =$ constant depending on n alone. Combining (1.2), (1.3), and (1.4) we conclude that

$$\int_M |\nabla f|^2 \geq C_0^2 \alpha(n) \left(\int_M |f|^{2n/(n-2)} \right)^{(n-2)/n} \tag{1.5}$$

for all $f \in H_{1,2}(M)$ satisfying $\int_M f = 0$. The lemma for $n \geq 3$ then follows by setting $C_1 = C_0^2 \cdot \alpha(n)$ and using Hölder inequality.

Case 2. When $n = 2$, we will prove the lemma directly from (1.1). The condition for the constant $k \in \mathbf{R}$ such that

$$\int_M (g-k)^2 = \inf_{a \in \mathbf{R}} \int_M |g-a|^2 \tag{1.6}$$

is equivalent to

$$\int_M (g-k) = 0 \tag{1.7}$$

Hence (1.1) can be written in the form

$$\left(\int_M |\nabla g| \right)^2 \geq C_0^2 \int_M (g-k)^2 \tag{1.8}$$

for any $g \in H_{1,1}(M)$ with $kV = \int_M g$. If we let $g = (\operatorname{sgn} f)\,|f|^{3/2}$, then we have

$$\left(\frac{3}{2} \int_M |f|^{1/2}\, |\nabla f| \right)^2 \geq C_0^2 \left[\frac{1}{2} \int |f|^3 - Vk^2 \right]. \tag{1.9}$$

However, since

$$\int_M |f| \int_M |\nabla f|^2 \geq \int_M |f|^{1/2} |\nabla f| \tag{1.10}$$

and

$$Vk^2 = \frac{(\int_M g)^2}{V} \leq \frac{(\int_M |g|)^2}{V} = V^{-1}\left(\int_M |f|^{3/2}\right)^2 \leq V^{-1} \int_M |f| \int_M f^2 \tag{1.11}$$

we have

$$\frac{9}{4} \int_M |f| \int_M |\nabla f|^2 + \frac{C_0^2}{V} \int_M |f| \int_M f^2 \geq \frac{C_0^2}{2} \int_M |f|^3 \geq \frac{C_0^2}{2} \frac{(\int_M f^2)^2}{\int_M |f|}$$

On the other hand, Poincaré inequality and (1.4) yield

$$\left(\frac{9}{4} + \frac{1}{\gamma(2)}\right) \int_M |f| \int_M |\nabla f|^2 \geq \frac{C_0^2}{2} \left(\int_M f^2\right)^2 \left(\int_M |f|\right)^{-1}.$$

Hence $C_1 = C_0^2 \left(\frac{9}{2} + 2/\gamma(2)\right)^{-1}$ for the case $n = 2$.

LEMMA 2. *Let M^n be a compact Riemannian manifold with boundary ∂M. Then there exists a constant C_2 depending only on M such that*

$$\int_M |\nabla f|^2 \geq C_2 \left(\int_M f^2\right)^{(2+n)/n} \left(\int_M |f|\right)^{-4/n}$$

for all $f \in H_{1,2}(M)$ with $f/\partial M \equiv 0$.

Proof. The lemma follows from the appropriate Sobolev inequality for Dirichlet boundary condition, namely

$$\int_M |\nabla f| \geq \tilde{C}_0 \left(\int_M |f|^{n/(n-1)}\right)^{(n-1)/n} \tag{1.12}$$

for $f \in H_{1,1}(M)$ satisfying $f/\partial M \equiv 0$. Here we use Lemma 6 of [5] instead, which gives for the case when $n \geq 3$,

$$\int_M |\nabla f|^2 \geq \tilde{C}_0 \alpha(n) \left(\int_M |f|^{2n/(n-2)}\right) \tag{1.13}$$

By setting $C_2 = \alpha(n)\tilde{C}_0^2$ and using the Hölder inequality, the lemma follows for $n \geq 3$.

When $n = 2$, we substitute the function $(\operatorname{sgn} f)|f|^{3/2}$ instead of f into (1.12), and the rest of the proof follows similarly to that of Lemma 1.

Remarks. It is known that [1], the constants C_0 and $\tilde{C}_0$ are equivalent to the following isoperimetric inequalities.

(i) Let M be a compact manifold with or without boundary. We define the constant

$$I(M) = \inf_N \frac{A(N)^{n/(n-1)}}{\operatorname{Min}\{V(M_1), V(M_2)\}}, \tag{1.14}$$

where the infemum is taken over all codimension-1 submanifolds N which divide M into two parts, M_1 and M_2. Then

$$I(M) \leq C_0 \leq 2^{1/(n-1)} I(M). \tag{1.15}$$

(ii) Let M be a compact manifold with boundary ∂M. We define another isoperimeter constant

$$\tilde{I}(M) = \inf_N \frac{A(N)^{n(n-1)}}{V(\tilde{M})} \tag{1.16}$$

where infemum is taken over all codimension-1 submanifolds N which divide M into two parts, and $\tilde{M}$ is the part such that $\partial\tilde{M} \cap \partial M = \varnothing$. Then

$$\tilde{C}_0 = \tilde{I}(M) \tag{1.17}$$

Together with the proofs of Lemmas 1 and 2, we have

$$C_1 \geq \alpha(n) I^2(M) \tag{1.18}$$

and

$$C_2 \geq \alpha(n) \tilde{I}^2(M).$$

When $\partial M = \varnothing$, it was shown in [2] that $I(M)$ can be estimated from below by a constant depending only on the diameter $= d$, volume of $M = V$, and the lower

bound of the Ricci curvature of M, $(n-1)K$. In fact,

$$I(M) \geq \alpha(n) \left[\frac{V}{\int_0^d F_k(r)\, dr} \right]^{(n+1)/(n-1)} \tag{1.19}$$

where

$$F_K(r) = \begin{cases} (-K)^{-1/2} (\sinh \sqrt{-K} r)^{n-1} & \text{if } K < 0 \\ r^{n-1} & \text{if } K = 0 \\ (K)^{-1/2} (\sinh \sqrt{K} r)^{n-1} & \text{if } K > 0 \end{cases}$$

§2. Eigenvalue estimates

We are now ready to derive lower estimates for the eigenvalues. In view of eigenfunction expansion it is natural to consider the fundamental solutions of the heat equations with either Dirichlet or Neumann boundary condition. In fact, we utilize the Sobolev inequalities, i.e. Lemmas 1 and 2, to obtain an upper bound for the heat kernel and hence lower bounds for the eigenvalues.

THEOREM 1. *Let M be a compact manifold with or without boundary. Suppose λ_k is the kth non-zero eigenvalue (with Neumann boundary condition if $\partial M \neq \varnothing$) on M. Then there exists a constant $\alpha(n)$ depending only on n, such that*

$$\lambda_k^{n/2} \geq \alpha(n) C_1^{n/2} k / V$$

In particular if $\partial M = \varnothing$, then $C_1^{n/2}$ has a lower bound depending only on d, V, and K. In which case

$$\lambda_k^{n/2} \geq \alpha(n) \frac{k}{V} \left[\frac{V}{\int_0^d F_k(r)\, dr} \right]^{n(n+1)/(n-1)}$$

where $F_K(r)$ has been defined by (1.19).

Proof. Let $H(x, y, t)$ be the fundamental solution of the heat equation

$$\square_y H(x, y, t) = \left(\Delta_y - \frac{\partial}{\partial t} \right) H(x, y, t) = 0$$

When $\partial M \neq \varnothing$, $H(x, y, t)$ is taken to be the heat kernel with Neumann boundary condition.

We consider the function $G(x, y, t)$ defined on $M \times M \times [0, \infty]$ by

$$G(x, y, t) = H(x, y, t) - \frac{1}{V}$$

Since $H(x, y, t)$ has the eigenfunction expansion

$$H(x, y, t) = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(x) \phi_i(y) \tag{2.1}$$

and also because $\lambda_0 = 0$ with $\phi_0(x) = V^{-1/2}$,

$$G(x, y, t) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \phi_i(x) \phi_i(y). \tag{2.2}$$

It is easy to check that G satisfies the semi-group property

$$G(x, y, t+s) = \int_M G(x, z, s) G(z, y, t)\, dz \tag{2.3}$$

for $x, y \in M$, and $t, s \in [0, \infty]$. Moreover

$$\int_M G(x, y, t)\, dy = 0 \tag{2.4}$$

because $\int_M \phi_i(y)\, dy = 0$, for $i \geq 1$. Also

$$\int_M |G(x, y, t)|\, dy \leq \int_M |H(x, y, t)|\, dy + 1 = 2 \tag{2.5}$$

because $H(x, y, t) \geq 0$ and $\int_M H(x, y, t) = 1$.

To prove the theorem, we first estimate $G(x, x, t)$ from above. By (2.3)

$$G'(x, x, t) = \int_M G'(x, y, t/2) G(x, y, t/2)\, dy = \int_M \Delta_y G(x, y, t/2) G(x, y, t/2)\, dy \tag{2.6}$$

because $0 = \square_y H(x, y, t/2) = \square_y G(x, y, t/2)$. Integration by parts and the boundary

condition of G yields

$$-G'(x, x, t) = \int_M |\nabla_y G(x, y, t/2)|^2 \, dy \geq 2^{-4/n} C_1 \left(\int_M G^2(x, y, t/2) \right)^{(2+n)/n}, \tag{2.7}$$

by Lemma 1 and (2.5). However the semi-group property (2.3) of G implies

$$-G'(x, x, t)(G(x, x, t))^{-(2+n)/n} \geq 2^{-4/n} C_1. \tag{2.8}$$

Integrating both sides of (2.8) with respect to t, and using the fact that

$$G(x, x, t) = (H(x, x, t) - 1/V) \to \infty$$

as $t \to 0$, we have

$$\frac{n}{2} G^{-2/n}(x, x, t) \geq 2^{-4/n} C_1 t \tag{2.9}$$

Hence

$$G(x, x, t) \leq 4 \left(\frac{2C_1}{n} \right)^{-n/2} t^{-n/2}$$

Integrating both sides with respect to x and using the expansion (2.2), we have

$$\sum_{i=1}^{\infty} e^{-\lambda_i t} \leq \alpha(n)(C_1 t)^{-n/2} V.$$

Let $t = 1/\lambda_k$, and observe that $\lambda_1/\lambda_k \leq 1$ for $i \leq k$, (2.10) becomes

$$\alpha(n) \left(\frac{\lambda_k}{C_1} \right)^{n/2} V \geq \sum_{i=1}^{\infty} e^{-\lambda_i/\lambda_k} \geq \sum_{i=1}^{k} e^{-\lambda_i/\lambda_k} \geq k e^{-1}$$

This proves the theorem.

THEOREM 2. *Let M be a compact manifold with boundary ∂M. Then the kth eigenvalue μ_k for Dirichlet boundary condition satisfies*

$$\mu_k^{n/2} \geq \alpha(n) C_2^{n/2} k / V$$

Proof. The proof here is similar to the above theorem. Instead, we use the heat kernel for Dirichlet boundary condition and observe the fact that it satisfies the semi-group property, its L^1 norm is identically 1, and it has the eigenfunction expansion. Also Lemma 2 will be used in terms of Lemma 1 for the Sobolev inequality.

COROLLARY 1. *Let M be a compact manifold. Suppose ϕ satisfies $\Delta\phi = -\lambda\phi$, and*

(i) *If $\partial M \neq \varnothing$, and $\partial\phi/\partial\nu \equiv 0$ on ∂M, then*

$$\|\phi\|_\infty^2 \leq \alpha(n)\left(\frac{\lambda}{C_1}\right)^{n/2}$$

(ii) *If $\partial M = \varnothing$, then*

$$\|\phi\|_\infty^2 \leq \alpha(n)\left(\frac{\lambda}{C_1}\right)^{n/2}$$

(iii) *If $\partial M \neq \varnothing$, and if $\phi/\partial M \equiv 0$, then*

$$\|\phi\|_\infty^2 \leq \alpha(n)\left(\frac{\lambda}{C_2}\right)^{n/2}$$

Proof. Clearly the proof follows from (2.9) by letting $t = 1/\lambda$, we have

$$e^{-1}\phi^2(x) \leq \sum_{i=1}^{\infty} e^{-\lambda_i/\lambda}\phi_i^2(x) \leq \alpha(n)\left(\frac{\lambda}{C_1}\right)^{n/2} \tag{2.12}$$

for all $x \in M$.

COROLLARY 2. *Let M be a compact manifold without boundary. Suppose $n \geq 4$ and the curvature of M is constant -1, then*

$$\lambda_k^{n/2} \geq \alpha'(n)\left[\frac{V^{(n^2+1)/(n-1)}}{(\sinh \beta(n) V)^{n(n+1)}}\right]k,$$

where $\alpha'(n)$, $\beta(n)$ are positive constants depending only on n.

Proof. By Theorem 1, (1.18), and (1.19), we have

$$\lambda_k^{n/2} \geq \alpha(n)\left[\frac{V^{n(n+1)/(n-1)-1}}{(\int_0^d F_{-1}(r)\,dr)^{n(n+1)/(n-1)}}\right]k$$

However, since

$$\sinh r \leq \cosh r$$

this gives

$$\lambda_k^{n/2} \geq \alpha'(n)\left[\frac{V^{(n^2+1)/(n-1)}}{(\sinh d)^{n(n+1)}}\right]k \tag{2.14}$$

On the other hand, a theorem of Gromov [4] showed that

$$d \leq \beta(n)V - 1 < \beta(n)V \tag{2.15}$$

Hence

$$\lambda_k^{n/2} \geq \alpha'(n)\left[\frac{V^{(n^2+1)/(n-1)}}{(\sinh \beta(n)V)^{n(n+1)}}\right]k$$

has to be proved.

COROLLARY 3 (Lieb, Donnelly–Li). *Let M be a compact domain on a simply connected non-positively curved manifold. Then there exists a constant $\alpha(n)$ depending only on n, such that*

$$\mu_k^{n/2} \geq \alpha(n)k/V$$

for all $k \geq 1$.

Proof. This follows trivially from Theorem 2, (1.16) and (1.18). We also observe that by a standard comparison theorem

$$\tilde{I}(M) \geq \frac{A(S^{n-1})^{n/(n-1)}}{V(D^n)} = \text{constant}$$

where $A(S^{n-1})$ and $V(D^n)$ stand for the area of the unit $(n-1)$-sphere in $\mathbf{R}^n$ and the volume of the unit n-disk in $\mathbf{R}^n$, respectively.

REFERENCES

[1] BOMBIERI, E., *Theory of minimal surfaces and a counter-example to the Bernstein conjecture in high dimension*, Lecture Notes, Courant Institute (1970).
[2] CROKE, C., *Some isoperimetric inequalities and eigenvalue estimates*, preprint.
[3] DONNELLY H. and LI, P., *Lower bounds for the eigenvalues of negatively curved manifolds*, to appear in Math. Z.
[4] GROMOV, M., *Manifolds of negative curvature*, preprint.
[5] LI, P., *On the Sobolev constant and the p-spectrum of a compact Riemannian manifold*, preprint.
[6] LI, P. and YAU, S. T., *Estimates of eigenvalues of a compact Riemannian manifold*, Proc. Symp. Pure Math. *36* (1980), 205–240.
[7] LIEB, E., *The number of bound states of one body Schroedinger operators and the Weyl problem*, Proc. Symp. Pure Math. *36* (1980), 241–252.
[8] SCHOEN, R., WOLPERT, S. and YAU, S. T., *Geometric bounds on the low eigenvalues of a compact surface*, Proc. Symp. Pure Math. *36* (1980), 279–286.

Princeton University
Dept. of Mathematics
Fine Hall – Box 37
Princeton, N.J. 08544

Received May 14, 1980

Comment. Math. Helvetici **56** (1981) 339–351 0010-2571/81/003339-13$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

All knots are algebraic

S. AKBULUT and H. KING

In this paper we study local knottedness of real algebraic sets. For instance we show that any knot in S^3 is "algebraic" i.e.

THEOREM 0.1. *Given a knot $K \subset S^3$ there is a real algebraic set $Z \subset R^4$ with* Sing $(Z) = 0$ *so that for all sufficiently small $\varepsilon > 0$, $(\varepsilon S^3, \varepsilon S^3 \cap Z)$ is diffeomorphic to (S^3, K). (Here εS^3 is the sphere of radius ε around 0.)*

In fact we have more generally the following theorem:

THEOREM 0.2. *Suppose $U \subset S^{k-1}$ is a compact smooth submanifold with codimension ≥ 1 and trivial normal bundle. Then there is a real algebraic set $Z \subset R^k$ with* Sing $(Z) = 0$ *so that for all sufficiently small $\varepsilon > 0$, $(\varepsilon S^{k-1}, \varepsilon S^{k-1} \cap Z)$ is diffeomorphic to $(S^{k-1}, \partial U)$. (In fact, $\varepsilon \partial U$ will be isotopic in εS^{k-1} to $\varepsilon S^{k-1} \cap Z$ where $\varepsilon \partial U = \{x \in \varepsilon S^{k-1} \mid x/\varepsilon \in \partial U\}$.)*

To deduce Theorem 0.1 from Theorem 0.2 we may let U be a Seifert surface for the knot K. However, we give a direct proof of Theorem 0.1 in Section 2 since it is simpler and illustrates the main ideas in the proof of Theorem 0.2.

§1. Preliminary lemmas

Definitions are as in [1]. For instance a polynomial $p: R^n \to R$ is *overt* if $x \in R^n - 0$ implies $p^*(0, x) \neq 0$ where $p^*: R \times R^n \to R$ is the homogenization of p, $(p^*(t, x) = t^d p(x/t)$ where $d =$ degree of $p)$. An algebraic set $V \subset R^n$ is *projectively closed* if $V = p^{-1}(0)$ for some overt polynomial p.

DEFINITION. Let W be a topological space, $M \subset W$. We say that M *compactly separates* W if there are closed sets W_0 and W_1 so that $W = W_0 \cup W_1$ and $M = W_0 \cap W_1$ and W_1 is compact.

We first give a few useful lemmas about algebraic sets.

LEMMA 1.1. *Suppose V and W are nonsingular algebraic sets, $W \subset V$ and* dim W = dim V. *Then $V - W$ is a nonsingular algebraic set.*

Proof. Lemma 1.6 of [1].

LEMMA 1.2 (Imprecise version). *The boundary of a submanifold with trivial normal bundle is isotopic to an algebraic set. This isotopy can fix nice subsets.*

LEMMA 1.2 (Precise statement). *Suppose W, L_i and V_i $i = 1, \ldots, k$ are algebraic sets, $U \subset$* Nonsing W *is a smooth compact submanifold with trivial normal bundle, the V_i are pairwise disjoint, $L_i \subset$* (Nonsing V_i) $\cap \partial U$ $i = 1, \ldots, k$ and $V_i \cap \partial U$ *contains a neighborhood of L_i in V_i $i = 1, \ldots, k$. Then there are arbitrarily small isotopies of ∂U which fix $\bigcup_{i=1}^{k} L_i$ and take ∂U to a nonsingular algebraic set. If each L_i is projectively closed then we may take this algebraic set to be projectively closed also.*

Proof. We will prove this by induction on the codimension of U in W. In case U has codimension 0, then ∂U compactly separates W so the result is essentially Lemma 2.2 of [1]. (The statement of Lemma 2.2 assumes $k = 1$ but its proof works for arbitrary k.)

Now let us prove the lemma for arbitrary codimension of U. Let $m > 0$ be the codimension of U in Nonsing W. Then we have a smooth open imbedding $\beta : U \times R^m \to$ Nonsing W so that $\beta(U \times 0) = U$. By induction $\beta(\partial(U \times B^m))$ is isotopic fixing $\bigcup_{i=1}^{k} L_i$ to a nonsingular algebraic set Y. Also there is a smooth compact submanifold $M \subset$ Nonsing W so that U is a codimension 0 submanifold of ∂M, M has trivial normal bundle, ∂M intersects $\beta(\partial(U \times B^m))$ transversely and $\partial M \cap \beta(\partial(U \times B^m)) = \partial U$ (see Figure 1). Then by induction we may isotop ∂M fixing $\bigcup_{i=1}^{k} L_i$ to a nonsingular algebraic set Z. Then since these isotopies can be C^1 small, there is an isotopy of ∂U to $Y \cap Z$ fixing $\bigcup_{i=1}^{k} L_i$.

COROLLARY 1.3. *Let $K \subset R^2$ be a link (i.e. a union of disjoint imbedded circles). Then there is a small isotopy of R^3 taking K to a projectively closed nonsingular algebraic set.*

Proof. Let U be a Seifert surface for K, i.e. U is a smoothly imbedded compact surface with trivial normal bundle so that $\partial U = K$. The result now follows from Lemma 1.2 setting $W = R^3$, $L_i = \phi = V_i$ and $U = U$.

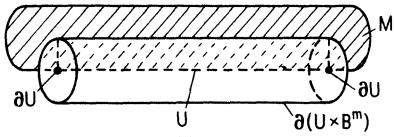


Fig. 1.

LEMMA 1.4. *Let W be a smooth n-dimensional manifold and $M \subset W$ a closed codimension 1 submanifold. Then M compactly separates W if and only if M represents 0 in $H_{n-1}(W, \partial W; \mathbf{Z}/2\mathbf{Z})$.*

Proof. Triangulate W so that M is a subpolyhedron. Then M represents 0 in $H_{n-1}(W, \partial W; \mathbf{Z}/2\mathbf{Z})$ if and only if in simplicial homology $[M] = \partial(\sum_{\sigma \in K} a_\sigma \cdot \sigma)$ where K is the set of n-simplices in W and $a_\sigma \in \mathbf{Z}/2\mathbf{Z}$ and only a finite number of a_σ are non-zero which happens if and only if $W = W_0 \cup W_1$ and $M = W_0 \cap W_1$ with W_1 compact where $W_i = \bigcup_{\substack{\sigma \in K \\ a_\sigma = i}} C1(\sigma)$.

§2. All knots are algebraic

This section is devoted to a proof of Theorem 0.1 which shows all knots are algebraic. Note that a very slight modification of the proof shows that all links are algebraic. In fact, the proof for links with an even number of components is a good deal easier since it is not necessary to add the extra one handle. Let us now proceed with the proof of Theorem 0.1.

Pick a point $z \in S^3 - K$ and a diffeomorphism $h : S^3 - z \to R^3$. We show below that there is a projectively closed nonsingular algebraic set $W \subset R^3$ and an algebraic subset $L \subset W$ so that the boundary of a smooth regular neighborhood of L in W is isotopic in R^3 to $h(K)$. Supposing this, let $W = p^{-1}(0)$ and $L = q^{-1}(0)$ for overt polynomials p and q. Let p^* and q^* be the homogenizations of p and q. Define an algebraic set

$$Z = \{(t, x) \in R \times R^3 \mid t^{2\deg(q)+1} = (q^*(t, x))^2 \quad \text{and} \quad p^*(t, x) = 0\}.$$

Then if $(t, x) \in Z$, $t \geq 0$. Also if $t = 0$ then $p^*(0, x) = 0$ so $x = 0$ by overtness of p. Hence $\varepsilon S^3 \cap Z = \varepsilon H^3 \cap Z$ where $\varepsilon H^3 = \{(t, x) \in R \times R^3 \mid t^2 + |x|^2 = \varepsilon^2$ and $t > 0\}$. If $\varphi_\varepsilon : R^3 \to \varepsilon H^3$ is the diffeomorphism $\varphi_\varepsilon(y) = (\varepsilon, \varepsilon y)/(1 + |y|^2)^{1/2}$ then

$$\varphi_\varepsilon^{-1}(\varepsilon H^3 \cap Z) = \{y \in R^3 \mid p(y) = 0$$

and

$$q^4(y)(1 + |y|^2) = \varepsilon^2\} = \{y \in W \mid q^4(y)(1 + |y|^2) = \varepsilon^2\}.$$

Since $q^4(y)(1 + |y|^2)$ is 0 only on L, [2] implies that for small enough ε, $\varphi_\varepsilon^{-1}(\varepsilon H^3 \cap Z)$ is the boundary of a smooth regular neighbourhood of L in W. Thus $\varepsilon S^3 \cap Z$ is isotopic in εS^3 to $\varphi_\varepsilon(h(K))$ so $(\varepsilon S^3, \varepsilon S^3 \cap Z)$ is diffeomorphic to (S^3, K) as we desired.

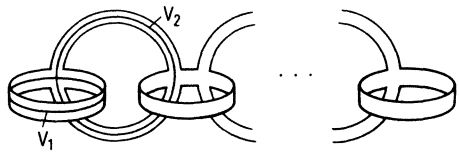


Fig. 2.

It remains to show the existence of W and L. Let $U \subset R^3$ be a connected Seifert surface for $h(K)$, that is U is an orientable compact surface with $\partial U = h(K)$. For a technical reason we must add a special 1-handle to U. In particular, let $V \subset R^3$ be some translation of the torus {points in R^3 of distance 1 from the circle $x=0,\ y^2+z^2=4\} = \{(x, y, z) \in R^3 \mid (x^2+y^2+z^2+3)^2 = 16(y^2+z^2)\}$ so that V is disjoint from U. We have a meridian V_1 and longitude V_2 of V corresponding to $\{z=0,\ x^2+(y-2)^2=1\}$ and $\{x=0,\ y^2+z^2=1\}$. Notice V, V_1 and V_2 are nonsingular projectively closed algebraic sets. We let U' be a connected sum of U and V, i.e. we run a.tube from U to V, being careful to connect the tube to V somewhere in $V-(V_1 \cup V_2)$. Hence $V_1 \subset U'$ and $V_2 \subset U'$ and the germ of U' at $V_1 \cup V_2$ is the germ of V at $V_1 \cup V_2$.

Let U'' be U' with a disc deleted from its interior. If we took U'' out of R^3 and unknotted it, it would look like Figure 2.

We will consider two transversely intersecting submanifolds of U'', L_1 and L_2 as shown in Figure 3. We want V_i to be a component of L_i $i=1, 2$. Notice that $U''-(L_1 \cup L_2)$ is diffeomorphic to $\partial U'' \times [0, 1)$. Also, $U'-((L_1-V_1) \cup L_2)$ is diffeomorphic to $\partial U' \times [0, 1)$.

We now let W' be the double of U'', i.e. since U'' has trivial normal bundle we may find a diffeomorphism $g: R^3 \to R^3$ so $g(U'') \cap U'' = \partial U''$ and $g(U'') \cup U''$ is a smooth submanifold W' of R^3. Note that $g(L_1) \cup (L_2 - V_2)$ is a union of disjoint circles. Thus by Corollary 1.3 we may isotop $g(L_1) \cup (L_2 - V_2)$ to a nonsingular projectively closed algebraic subset of R^3 so we may as well assume (after isotoping U'' a bit) that $g(L_1) \cup L_2$ is a nonsingular algebraic set. Note that W' compactly separates R^3 so by Lemma 1.2 we may isotop W', fixing $V_1 \cup L_2 \cup g(L_1)$ to a nonsingular projectively closed algebraic set W. Let $f: W' \to W$ be the time one map of this isotopy. Since $L_1 \cup g(L_1)$ compactly separates W', we may

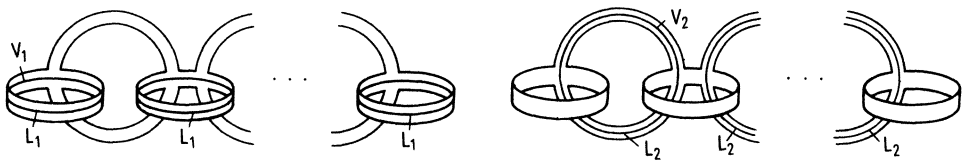


Fig. 3.

assume by Lemma 1.2 that $f(L_1 \cup g(L_1))$ is a nonsingular algebraic set. (We might have to do a little isotopy of W to itself to do this.) But then $f(L_1 \cup g(L_1)) - (V_1 \cup g(L_1)) = f(L_1 - V_1)$ is a nonsingular algebraic set by Lemma 1.1. Hence $f(L_2 \cup (L_1 - V_1)) = L_2 \cup f(L_1 - V_1)$ is an algebraic subset L of W. Notice that a regular neighborhood of L in W is isotopic to a regular neighborhood of $L_2 \cup (L_1 - V_1)$ in U'' which is isotopic to $\partial U' = h(K)$. So we are done.

§3. Obtaining nice spines

An important feature in the proof of Theorem 0.1 was that a Seifert surface for the knot had a spine of transversely intersecting circles. This section is devoted to the corresponding result we will need to prove Theorem 0.2. We will show that after adding one-handles to U we will have a spine of codimension one spheres and circles in general position. This point is used to great advantage in other papers of ours. We are indebted to Lowell Jones for discovering Lemma 3.2 for us.

EXAMPLE 3.1. At this point we need examples of nonsingular algebraic sets. Let

$$V = \left\{(x_1, x_2, \ldots, x_{n+1}) \in \mathbf{R}^{n+1} \;\middle|\; 16 \sum_{i=2}^{n+1} x_i^2 = \left(3 + \sum_{i=1}^{n+1} x_i^2\right)^2\right\},$$

$$V_1 = \{(x_1, x_2, 0, 0, \ldots, 0) \in \mathbf{R}^{n+1} \mid x_1^2 + (2 - x_2)^2 = 1\},$$

$$V_2 = \left\{(0, x_2, \ldots, x_{n+1}) \in \mathbf{R}^{n+1} \;\middle|\; \sum_{i=2}^{n+1} x_i^2 = 1\right\}.$$

A moment of examination shows that V, V_1 and V_2 are nonsingular projectively closed algebraic sets, V_1 is a circle, V_2 is an $n-1$ sphere and V_1 and V_2 intersect transversely in V. V is diffeomorphic to $S^1 \times S^{n-1}$ and can be described geometrically as the set of points in $\mathbf{R}^{n+1}$ of distance 1 from the $n-1$ sphere of radius 2 in $0 \times \mathbf{R}^n \subset \mathbf{R}^{n+1}$.

LEMMA 3.2. *Let W be a compact connected smooth manifold with boundary. Then there is a finite collection D_α, $\alpha \in A$ of imbedded discs in* int W *so that*

(1) *The boundaries $S_\alpha = \partial D_\alpha$, $\alpha \in A$ are in general position.*
(2) *W minus a finite disjoint collection of discs is a smooth regular neighborhood of $\bigcup_{\alpha \in A} S_\alpha$ in W.*

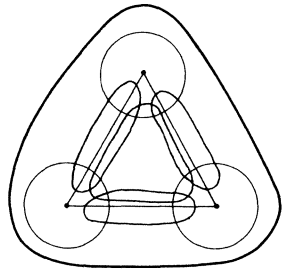


Fig. 4.

Proof. Take a smooth triangulation of W and let X be the subpolyhedron consisting of all simplices which do not touch ∂W. There will at the end be one disc D_α for each simplex of X. One way to describe these discs is as follows. For a simplex σ of dimension k, let E_σ be the union of all simplices in a $k+1$-th barycentric subdivision of σ which do not touch $\partial\sigma$. Now let D_σ be the union of all simplices in a $k+2$-th barycentric subdivision of W which intersect E_σ. Of course we must take the subdivisions in such a manner that D_σ is smooth. Figure 4 illustrates the case where X is a 2-simplex.

Let T be the closure of some component of $W - U_{\alpha\in A} S_\alpha$. If T contains a component Y of ∂W, then T is homeomorphic to $Y\times[0, 1]$, since $\bigcup_{\alpha\in A} D_\alpha$ is a regular neighborhood of X and W is a regular neighborhood of X, so $W - \bigcup_{\alpha\in A} D_\alpha$ is diffeomorphic to $\partial W\times[0, 1]$ and W is a regular neighborhood of $\bigcup_{\alpha\in A} D_\alpha$. If our component T lies in $\bigcup_{\alpha\in A} D_\alpha$ then it is of the form $\bigcap_{\alpha\in A'} D_\alpha - \bigcup_{\alpha\notin A'} \operatorname{int} D_\alpha$ for some nonempty $A'\subset A$. Let $A' = \{\alpha_0, \alpha_1, \ldots, \alpha_k\}$ where α_j is a face of α_{j+1} for all j. We can do this because $D_\alpha \cap D_{\alpha'} \neq \varnothing$ implies α is a face of α' or vice versa. Then note that $\bigcap_{\alpha\in A'} D_\alpha - \bigcup_{\alpha\notin A'} \operatorname{int} D_\alpha$ collapses to $E_{\alpha_k} \cap (\bigcap_{\alpha\in A'} D_\alpha - \bigcup_{\alpha\notin A'} \operatorname{int} D_\alpha)$ which is always a disc in the simplex α_k.

Thus a smooth regular neighborhood of $\bigcup_{\alpha\in A} S_\alpha$ is obtained by deleting a disc from each component of $\bigcup_{\alpha\in A} D_\alpha - \bigcup_{\alpha\in A} S_\alpha$. (We in fact delete one disc for each simplex in the first barycentric subdivision of X.)

LEMMA 3.3. *Suppose W is a compact connected n-dimensional manifold with nonempty boundary. Then there is a compact manifold $W'\subset W\times R$ with trivial normal bundle in $W\times R$ and a finite collection of closed submanifolds in general position S_α, $\alpha\in A$, R_β and T_β, $\beta\in B$ and a collection of pairwise disjoint imbeddings $\varphi_\beta : R^{n+1}\to W\times R$ so that*

(a) *$\partial W'$ is isotopic to $\partial W\times 0$ in $W\times R$.*
(b) *W' is a smooth regular neighborhood of $\bigcup_{\alpha\in A} S_\alpha \cup \bigcup_{\beta\in B} R_\beta$ in W'.*
(c) *For each $\alpha\in A$ there is a $B_\alpha\subset B$ so that $S_\alpha \cup \bigcup_{\beta\in B_\alpha} T_\beta$ is a manifold compactly separating W'.*

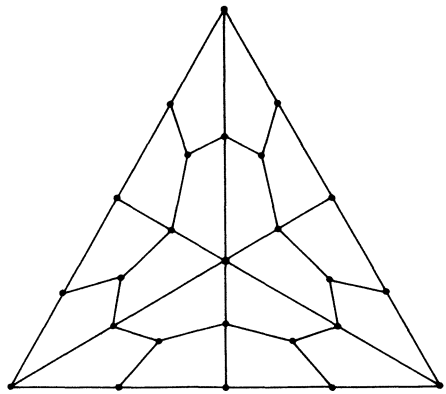


Fig. 5.

(d) *$T_\beta = \varphi_\beta(V_2)$ and $R_\beta = \varphi_\beta(V_1)$ and the germ of W' at $T_\beta \cup R_\beta$ is the germ of $\varphi_\beta(V)$ at $T_\beta \cup R_\beta$ where V_1, V_2 and V are as in Example* 3.1.

(*see Figure* 11).

Proof. By Lemma 3.2 we have a collection D_α, $\alpha \in A$ of n-discs in W satisfying the conclusion of Lemma 3.2. Let K be the one-complex whose vertices are components of $\bigcup_{\alpha\in A} D_\alpha - \bigcup_{\alpha\in A} S_\alpha$ and so that there is a one-simplex between two vertices if and only if there is an $\alpha' \in A$ so that the components of $\bigcup_{\alpha\in A} D_\alpha - \bigcup_{\alpha\in A} S_\alpha$ corresponding to the two vertices lie in the same component of $\bigcup_{\alpha\in A} D_\alpha - \bigcup_{\alpha\in A-\alpha'} S_\alpha$. (This K is a subcomplex of the second barycentric subdivision of the X in the proof of Lemma 3.2. For instance, for the two simplex of Figure 4, K is as in Figure 5.) Notice that K is connected, for if p and q are any two points of $\bigcup_{\alpha\in A} D_\alpha - \bigcup_{\alpha\in A} S_\alpha$ we may draw a smooth path in $\bigcup_{\alpha\in A} D_\alpha$ from p to q in general position with the collection of S_α. This path gives us a path in K for whenever the path crosses some S_α, it passes between two components of $\bigcup D_\alpha - \bigcup S_\alpha$ which have a one simplex between them.

Now pick a vertex v of K so that for some $\alpha' \in A$, v lies in the same component of $W - \bigcup_{\alpha\in A-\alpha'} S_\alpha$ as some component * of $W - \bigcup_{\alpha\in A} D_\alpha$. Now define a one complex K^* to be the complex K with one additional vertex * and one additional 1-simplex which lies between * and v. For each vertex u of K let B_u be a smooth disc in the component u of $\bigcup D_\alpha - \bigcup S_\alpha$. Denote $U = W - \bigcup_{u\in K} \operatorname{int} B_u$. Then for each vertex of K^* we have $\kappa(u)$, an associated component of ∂U. Namely $\kappa(u) = \partial B_u$ if $u \in K$ and $\kappa(*) = \partial^* \subset \partial W$.

Notice that U is a smooth regular neighborhood of $\bigcup_{\alpha\in A} S_\alpha$. Let C be a maximal tree in K^*, (that is C is a contractible subcomplex containing all the vertices of K^*). Let B be the set of one-simplices of C. For each $\beta \in B$ we may

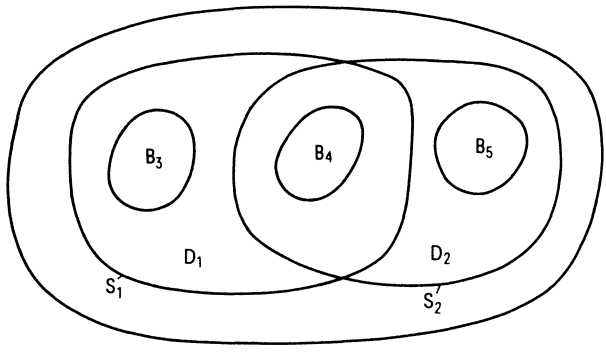


Fig. 6.

associate a smooth arc A_β in U so that

(1) The A_β's are pairwise disjoint.
(2) If β has vertices u and u' then $A_\beta \cap \partial U = \{p_\beta, p'_\beta\} = \partial A_\beta$ with $p_\beta \in u$ and $p'_\beta \in u'$ and A_β is transverse to ∂U.
(3) A_β is in general position with the collection S_α, $\alpha \in A$ and $A_\beta \cap \bigcup_{\alpha \in A} S_\alpha =$ one point q_β.

For instance, suppose W is a 2-disc and $A = \{1, 2\}$ so the discs D_α and B_u are as in Figure 6.

Then K and K^* are as in Figure 7 and $C = K^*$ because K^* is contractable. Denote the simplices of C as in Figure 7.

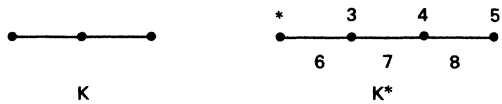


Fig. 7.

We now show the A_β in Figure 8. Pick an imbedding $g: (\bigcup A_\beta) \times \mathbf{R}^{n-1} \to U$ with $g^{-1}(\partial U) = \bigcup (\partial A_\beta) \times \mathbf{R}^{n-1}$ and $g(x, 0) = x$ for all x. Now we define $W' \subset W \times \mathbf{R}$ by

$$W' = U \times 0 \cup g(\bigcup \partial A_\beta \times B^{n-1}) \times [0, 1] \cup g(\bigcup A_\beta \times B^{n-1}) \times 1$$

with the corners rounded. We define $R_\beta \subset W'$ by

$$R_\beta = \partial A_\beta \times [0, 1] \cup A_\beta \times \{0, 1\}.$$

Thus, W' is obtained from U by attaching 1-handles to the boundary of U.

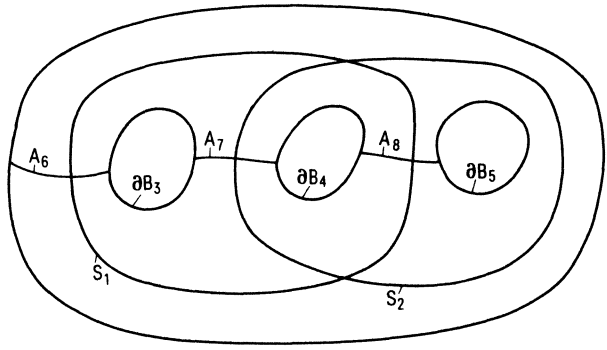


Fig. 8.

Each R_β is A_β union the core of one of these 1-handles. See Figure 9 for the W' constructed via Figures 6, 7 and 8.

It is easy to see (b) from the fact that U is a smooth regular neighborhood of $\bigcup_{\alpha\in A} S_\alpha$. Likewise, (a) is easily seen since $\partial W'$ is obtained from ∂W by "growing" the tree C, see Figure 10.

We now construct the T_β's. For each $\beta\in B$, let Y_β = the component of $\partial W' - g(q_\beta\times S^{n-2})\times 1$ which does not contain points of $\partial W\times 0$ (i.e. Y_β is the limb we get by sawing the branch β). Notice that each Y_β is an open $n-1$ disc. Let $h:\partial W'\times[0,1)\to W'-(\bigcup_{\alpha\in A} S_\alpha\cup\bigcup_{\beta\in B} R_\beta)$ be a collar, i.e. an open imbedding so that $h(x,0)=x$ for all $x\in\partial W'$. By extending collars we may assume that $h((g(x,y),t),s)=(g(x,(1-s)y),t)$ for all $y\in S^{n-2}$ and $(x,t)\in\bigcup_\beta \partial A_\beta\times[0,1]\cup\bigcup_\beta A_\beta\times 1$. We may put an order $<$ on B by $\beta'<\beta$ if $Y_{\beta'}\subsetneq Y_\beta$. Now let $\gamma:B\to(0,1)$ be any $1-1$ order preserving function. Then define $T_\beta = h(Y_\beta\times(1-\gamma(\beta)))\cup g(q_\beta\times\gamma(\beta)B^{n-1})\times 1$ after smoothing out the corner. (See Figure 11.)

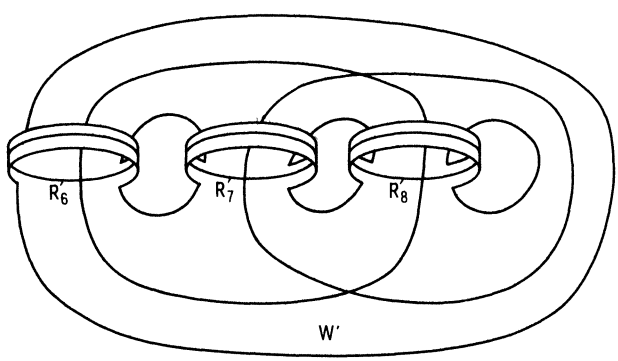


Fig. 9.

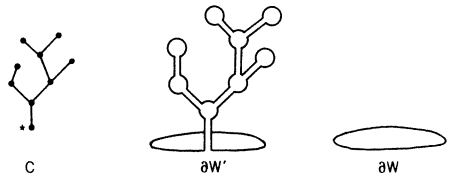


Fig. 10.

Now we show that (c) is satisfied. For any $\alpha \in A$, let $B_\alpha = \{\beta \in B \mid R_\beta \cap S_\alpha \neq \varnothing\}$. Then we claim $S_\alpha \cup \bigcup_{\beta \in B_\alpha} T_\beta$ represents 0 in $H_{n-1}(W', \mathbf{Z}/2\mathbf{Z})$. Lemma 1.4 then insures that $S_\alpha \cup \bigcup_{\beta \in B_\alpha} T_\beta$ is a manifold compactly separating W'. Let E be the set of vertices of K. For each $\beta \in B$ let

$$E_\beta = \{u \in E \mid Y_\beta \cap \kappa(u) \text{ is nonempty}\}$$

(recall $\kappa(u)$ is a component of ∂U) and for each $\alpha \in A$ let

$$E_\alpha = \{u \in E \mid \text{so that } u \subset D_\alpha\}.$$

Then if [ ] denotes the homology class in $H_{n-1}(W', \mathbf{Z}/2\mathbf{Z})$ we easily see that

$$[T_\beta] = \sum_{u \in E_\beta} [\kappa(u)] \quad \text{and} \quad [S_\alpha] = \sum_{u \in E_\alpha} [\kappa(u)].$$

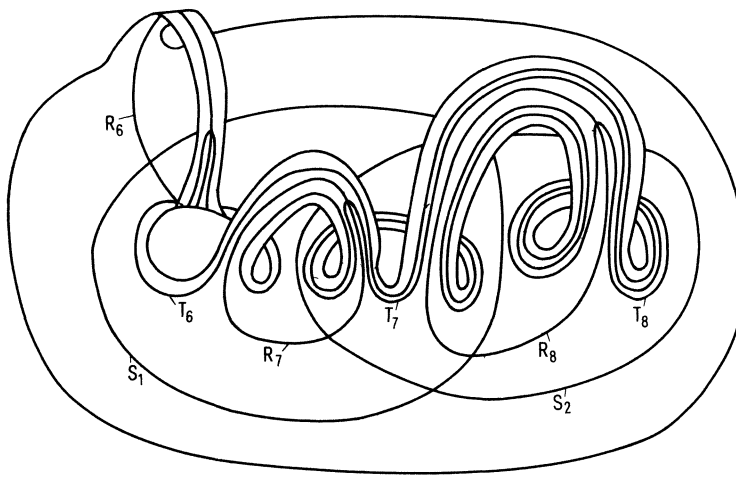


Fig. 11.

Now for $u \in E$, let C_u be the unique subcomplex of C which is homeomorphic to a line with one end u and the other end *. The 1-simplices of C_u are those β for which $u \in E_\beta$. Let $B_{\alpha u} = \{\beta \in B_\alpha \mid u \in E_\beta\}$ and notice that $B_{\alpha u}$ is just the set of 1-simplices of C_u one of the vertices of which is in E_α and the other vertex of which is not in E_α. Hence if $u \notin E_\alpha$, $B_{\alpha u}$ has even cardinality and if $u \in E_\alpha$ then $B_{\alpha u}$ has odd cardinality. Thus we have

$$\left[S_\alpha \cup \bigcup_{\beta \in B_\alpha} T_\beta\right] = \sum_{u \in E_\alpha} [\kappa(u)] + \sum_{u \in E} (\#B_{\alpha u})[\kappa(u)] = 0$$

where $\#B_{\alpha u}$ is the cardinality of $B_{\alpha u}$.

It remains to show the existence of the φ_β's. Notice that each R_β bounds a 2-disc $R'_\beta = A_\beta \times [0, 1]$ in $W \times R$. Likewise, each T_β bounds an n-disk T'_β in $W \times R$ which is a union of B_u's and part of the 1-handles, more precisely it is given as follows: Let $f: W' \times R \to W \times R$ be an imbedding so that $f(x, 0) = x$ all $x \in W'$ and $R'_\beta \cap f(W' \times (0, 1))$ is empty and $f((x, 0), t) = (x, -t)$ for $(x, t) \in U \times [0, 1]$. Let

$$H_\beta = h((Y_\beta \cap \partial U) \times [0, 1 - \gamma(\beta)]) \cup \text{closure}\,(h((Y_\beta - \partial U) \times [1 - \gamma(\beta), 1))).$$

We now define

$$T'_\beta = f(T_\beta \times [0, \gamma(\beta)] \cup H_\beta \times \gamma(\beta)) \cup \bigcup_{u \in E_\beta} B_u \times -\gamma(\beta),$$

i.e. we push out a little bit and then fill in the holes we originally took out of W to make U (see Figure 12).

Now $R'_\beta \cup T'_\beta$ is collapsible so its smooth regular neighborhood is a ball, which we make the image of φ_β, then φ_β exists since R'_β and T'_β can be isotoped inside the ball to a standard pair of discs.

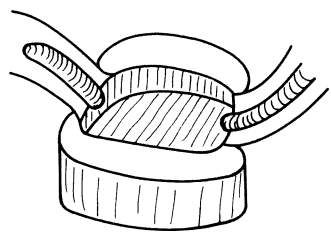


Fig. 12.

§4. Generalized knots are algebraic

This section is devoted to a proof of Theorem 0.2. We may assume each component of U has non-empty boundary and after applying an isotopy to S^{k-1} we may also assume that U lies in just one hemisphere $H^{k-1} = \{(x_1, \ldots, x_k) \in S^{k-1} \mid x_k > 0\}$. Let $n < k-1$ be the dimension of U. Let $\pi : R^k \to R^{k-1}$ be projection onto the first $k-1$ coordinates and let $W = \pi(U)$. Then W is a submanifold of R^{k-1} diffeomorphic to U. Let $h : W \times \mathbf{R} \times \mathbf{R}^{k-2-n} \to R^{k-1}$ be an imbedding so $h(x, 0, 0) = x$ for all $x \in W$. By Lemma 3.3 there is a $W' \subset W \times R$ and S_α, $\alpha \in A$ and R_β, T_β and φ_β satisfying the conclusions of Lemma 3.3. Notice that $h(W' \times 0)$ has trivial normal bundle in R^{k-1} and $h(\partial W' \times 0)$ is isotopic to ∂W. Pick injective linear transformations $\lambda_\beta : R^{n+1} \to R^{k-1}$ for each $\beta \in B$ so that $\lambda_\beta(V) \cap \lambda_{\beta'}(V)$ is empty for $\beta \neq \beta'$. In addition, if $n+1 = k-1$ we require λ_β to preserve orientation if $h\varphi_\beta$ preserves orientation and we require λ_β to reverse orientation if $h\varphi_\beta$ reverses orientation. Then since any two orientation preserving imbeddings of a disc are isotopic, we may (after isotoping R^{k-1}) assume that $h(\varphi_\beta(x), 0) = \lambda_\beta(x)$ for all x near V in R^{n+1}. Let $U'' = h(W' \times 0)$ and $L = \bigcup_{\beta \in B} \lambda_\beta(V_1 \cup V_2)$. Notice that each $\lambda_\beta(V)$ and $\lambda_\beta(V_i)$ is a nonsingular projectively closed algebraic set, $L \subset U''$ and the germ of U'' at L is the germ of $\bigcup \lambda_\beta(V)$ at L.

Now let Y be a smooth closed n dimensional submanifold of R^{k-1} so that $U'' \subset Y$ and Y bounds a compact smooth submanifold of R^{k-1} with trivial normal bundle. For instance, if $h' : U'' \times R \times R^{k-2-n} \to R^{k-1}$ is a trivialization of the normal bundle of U'', let Y be $h'(\partial(U'' \times [0, 1] \times 0))$ with corners smoothed.

By Lemma 1.2 there is a projectively closed nonsingular algebraic set $X \subset R^{k-1}$ so that $L \subset X$ and Y is isotopic to X fixing L. Let $g : Y \to X$ be the time one map of this isotopy.

Notice that for each $\alpha \in A$, $g(S_\alpha) \cup \bigcup_{\beta \in B_\alpha} \lambda_\beta(V_2)$ compactly separates X. Hence by isotoping a little more we may assume by Lemma 1.2 that each $g(S_\alpha) \cup \bigcup_{\beta \in B_\alpha} \lambda_\beta(V_2)$ is a nonsingular algebraic set. But then by Lemma 1.1 each $g(S_\alpha)$ is a nonsingular algebraic set.

Pick polynomials $p : R^{k-1} \to R$ and $q : R^{k-1} \to R$ so that p is overt, $X = p^{-1}(0)$ and $q^{-1}(0) = \bigcup_{\alpha \in A} g(S_\alpha) \cup \bigcup_{\beta \in B} \lambda_\beta(V_1)$. Let p^* and q^* be the homogenizations of p and q.

Define an algebraic set $Z \in R^{k-1} \times R$ by

$$Z = \{(x, t) \in R^{k-1} \times R \mid t^{2\deg(q)+1} = (q^*(t, x))^2 \text{ and } p^*(t, x) = 0\}.$$

Notice that $(x, t) \in Z$ implies $t \geq 0$ and if $t = 0$ then $x = 0$ also since p is overt. Hence $\varepsilon S^{k-1} \cap Z = \varepsilon H^{k-1} \cap Z$ where $\varepsilon H^{k-1} = \{(x_1, \ldots, x_k) \in \varepsilon S^{k-1} \mid x_k > 0\}$. If

$\varphi_\varepsilon : \mathbf{R}^{k-1} \to H^{k-1}$ is the diffeomorphism $\varphi_\varepsilon(y) = (\varepsilon y, \varepsilon)/(1+|y|^2)^{1/2}$ then $\varphi_\varepsilon^{-1}(\varepsilon H^{k-1} \cap Z) = \{y \in \mathbf{R}^{k-1} \mid p(y) = 0,\ q^4(y)(1+|y|^2) = \varepsilon^2\}$ which for small enough $\varepsilon > 0$ is the boundary of an algebraic regular neighborhood of $q^{-1}(0)$ in X which by [2] is isotopic to ∂W. Hence $\varepsilon S^{k-1} \cap Z$ is isotopic in εS^{k-1} to $\varepsilon \partial U$.

Since [2] has not yet appeared, we indicate to the reader how a proof would go for this special case. Norice that $q^{-1}(0)$ is a union of nonsingular algebraic sets in general position, so we could have taken q to be a product of polynomials q_i such that for small $\varepsilon > 0$; $X \cap q_i^{-1}([-\varepsilon, \varepsilon])$ is a tubular neighborhood in X of the nonsingular algebraic set $q_i^{-1}(0)$. Now (using the curve selection lemma, say) one sees that for small ε; $\{y \in X \mid \Pi q_i^4(y)(1+|y|^2) \leq \varepsilon\}$ is the union of tubular neighborhoods of these nonsingular algebraic sets, rounded off where they intersect; i.e. it is a smooth regular neighborhood of $q^{-1}(0)$. Since ∂W is the boundary of a smooth regular neighborhood of $q^{-1}(0)$, uniqueness of smooth regular neighborhoods gives an isotopy from ∂W to $\varphi_\varepsilon^{-1}(\varepsilon H^{k-1} \cap Z)$ as we desired.

REFERENCES

[1] S. AKBULUT and H. KING, *The topology of real algebraic sets with isolated singularities*, Ann. of Math. 113, (1981).

[2] H. KING, *Algebraic regular neighborhoods*, to appear.

The Institute for Advanced Study
Princeton, New Jersey 08540 *U.S.A.*
University of Maryland
College Park, Maryland 20742 U.S.A.

Received October 21, 1980/March 23, 1981.

Comment. Math. Helvetici **56** (1981) 352–365 0010-2571/81/003352-14$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Arcs omitted by support points of univalent functions

PETER L. DUREN

This paper develops a general method for investigating certain geometric properties of the analytic arc omitted by a support point in the class S of univalent functions. The main result is that for every support point arising from point-evaluation of the derivative, the omitted arc has monotonic argument.

§1. Introduction

The class S consists of all functions of the form

$$f(z) = z + a_2 z^2 + a_3 z^3 + \cdots$$

analytic and univalent in the unit disk $|z| < 1$. A *support point* of S is a function $f \in S$ which maximizes $\operatorname{Re}\{L\}$ for some complex-valued continuous linear functional L not constant on S. For instance, the Koebe function $k(z) = z(1-z)^{-2}$ and its rotations are support points of S because they maximize $\operatorname{Re}\{e^{i\theta}a_2\}$. There are other examples, but a complete description is not yet available.

Some general properties are well known (see *e.g.* [10], [4], or [5]). Each support point f maps the disk onto the complement of an analytic arc Γ which extends to infinity with increasing modulus and satisfies the differential equation

$$\Phi(w)\frac{dw^2}{w^2} > 0, \qquad \Phi(w) = L\left(\frac{f^2}{f-w}\right). \tag{1}$$

This function Φ is analytic on Γ and has the property $\operatorname{Re}\{\Phi(w)\} > 0$ there, except perhaps at the endpoints. At each point of Γ, the angle measured from the radial line to the tangent line is called the *radial angle* of Γ. In view of (1), the property $\operatorname{Re}\{\Phi(w)\} > 0$ is equivalent to $\operatorname{Re}\{(dw/w)^2\} > 0$, which says geometrically that the radial angle of Γ is less than $\pi/4$ in magnitude at each interior point. This is known as the $\pi/4$-*property*.

Pfluger [8] showed that $L(f^2) \neq 0$, or equivalently that the quadratic differential (1) has a simple pole at infinity. Brickman and Wilken [1] observed that Γ is tangent to the half-line

$$w = \frac{L(f^3)}{3L(f^2)} - L(f^2)t, \qquad t \geq 0 \tag{2}$$

at infinity. This is called the *asymptotic half-line.* It should be remarked that if the asymptotic half-line is a trajectory of the quadratic differential (1) near infinity, then Γ is itself a half-line which coincides with the asymptotic half-line near infinity. This is true because the quadratic differential has a unique trajectory extending to its simple pole at infinity.

For the support points associated with point-evaluation functionals $L(f) = f(\zeta)$, Brown [2] found that the omitted arc Γ has monotonic argument and monotonic radial angle. No rotation of the Koebe function can maximize $\operatorname{Re}\{f(\zeta)\}$ unless ζ lies on a certain segment of the real axis (*cf.* Schober [10], p. 84). For a certain choice of ζ on the negative real axis, Brown found numerically that the radial angle at the tip of Γ is approximately equal to $\pi/4$. This was a "numerical proof" that the upper bound $\pi/4$ in the general description of support points is best possible. For the functionals

$$L(f) = a_3 + \lambda a_2, \qquad \lambda \in \boldsymbol{C},$$

which typically exclude all rotations of the Koebe function as support points, Brown [3] again showed that Γ has monotonic argument.

More recently, Pearce [7] made the observation that for $0 < \zeta \leq \sin \pi/8$, the nonlinear problem of maximizing (or minimizing) $\arg f'(\zeta)$ is equivalent to a certain linear problem of the form

$$\max_{f \in S} \operatorname{Re}\{e^{-i\sigma} f'(\zeta)\}, \tag{3}$$

since the set of values $\{f'(\zeta) : f \in S\}$ then lies in the right half-plane and is supported by the radial line of maximal argument. Thus the functions

$$f(z) = \frac{z - \zeta z^2}{(1 - e^{i\theta} z)^2}, \qquad \theta = \cos^{-1} \zeta, \tag{4}$$

which are known to maximize $\arg f'(\zeta)$ over S for each $\zeta \leq 1/\sqrt{2}$ (see [6], p. 115), are support points of S if $\zeta \leq \sin \pi/8$. The arc Γ omitted by a function f of the form (4) is a half-line which for $\zeta = \sin \pi/8$ has a radial angle of exactly $\pi/4$ at its tip. In all cases, this arc Γ has monotonic argument and monotonic radial angle.

We shall generalize this result by proving that every support point arising from a problem of the type (3), for arbitrary real σ and $0<\zeta<1$, has an omitted arc Γ with monotonic argument. (There is no loss of generality in assuming that ζ is real and positive.) Without appeal to the rotation theorem, we shall observe independently that Γ is a nonradial half-line if and only if the extremal problem (3) is equivalent to that of maximizing or minimizing $\arg f'(\zeta)$. In a sense these results complement the work of Grad [9], who described the region of values of $f'(\zeta)$ as f ranges over S. We are concerned with the geometric properties of those functions f for which $f'(\zeta)$ lies on the boundary of the convex hull of this region.

Before turning to the problem (3), we shall develop some general techniques for using the differential equation (1) to deduce certain geometric properties of the arc Γ omitted by a support point. This method is then applied to establish the monotonic argument result for support points arising from both the point-evaluation problem and the derivative problem (3). In all cases it is the asymptotic half-line which distinguishes the trajectory Γ and makes this approach effective.

§2. General criteria for monotonicity

Let $f\in S$ be a support point which maximizes $\operatorname{Re}\{L\}$. Then f omits an analytic arc Γ which extends with increasing modulus from a point w_0 to infinity and satisfies the differential equation (1). Let Γ have a nonsingular parametrization $w=w(t)$, $0\le t\le\infty$, with $w(0)=w_0$ and $w(\infty)=\infty$. Then the relation

$$\frac{d}{dt}\arg w(t)=\operatorname{Im}\left\{\frac{w'(t)}{w(t)}\right\}$$

shows that $\arg w$ is increasing where $\operatorname{Im}\{w'/w\}>0$ and decreasing where $\operatorname{Im}\{w'/w\}<0$. On the other hand, $|\arg\{w'/w\}|<\pi/4$ by the $\pi/4$-property of Γ. Thus $\arg w$ is increasing or decreasing according as $\operatorname{Im}\{(w'/w)^2\}$ is positive or negative. In view of (1), this says that $\arg w$ increases where $\operatorname{Im}\{\Phi(w)\}<0$ and decreases where $\operatorname{Im}\{\Phi(w)\}>0$. If $\arg w$ ever reverses direction, it must do so on the set where $\Phi(w)>0$.

It is equivalent to the $\pi/4$-property that Γ lies entirely in the region where $\operatorname{Re}\{\Phi(w)\}>0$, except perhaps for the endpoint w_0. This will be called the *accessible region*. The set where $\operatorname{Re}\{\Phi(w)\}<0$ will be called the *forbidden region*. The range of f contains the forbidden region. The set where $\Phi(w)$ is real will be called the *critical set*.

It is clear that the intersection of the critical set with the accessible region (*i.e.*, the set of points where $\Phi(w)>0$) coincides with the set where trajectories are

tangent to radial lines; while the part of the critical set in the forbidden region (the set of points where $\Phi(w)<0$) is the locus of points where trajectories are tangent to circles centered at the origin. The critical set is a collection of curves which divide the accessible region into subregions where the trajectories have strictly monotonic argument. In order to show that the trajectory Γ omitted by a support point has strictly monotonic argument, it is sufficient (in fact, equivalent) to show that Γ lies entirely in one of these subregions. It is here that the explicit knowledge of the asymptotic half-line becomes a powerful tool.

The monotonicity of the radial angle

$$\alpha(t)=\arg\left\{\frac{w'(t)}{w(t)}\right\}$$

can be discussed in a similar way. The differential equation (1) shows that

$$\arg\Phi(w(t))+2\alpha(t)=0,$$

and so

$$\alpha'(t)=-\tfrac{1}{2}\operatorname{Im}\left\{\frac{\Phi'(w)w'(t)}{\Phi(w)}\right\}.$$

Thus $\alpha'(t)=0$ if and only if

$$\left\{\frac{\Phi'(w)w'(t)}{\Phi(w)}\right\}^2\geq 0.$$

Comparing this with (1), we see that if the radial angle α ever reverses direction, it must do so on the set where

$$\frac{[\Phi'(w)]^2w^2}{[\Phi(w)]^3}\geq 0. \tag{5}$$

It should be remarked that for the trajectory Γ omitted by a support point, the monotonicity of the radial angle implies the monotonicity of the argument. This is true because the radial angle is bounded by $\pi/4$ and tends to zero at infinity as Γ approaches its asymptotic half-line (*cf.* Brown [2]).

§3. Applications to the point-evaluation problem

As a first illustration of the method, we now consider the point-evaluation functional $L(f) = f(\zeta)$ for fixed ζ with $|\zeta| < 1$. Let $f \in S$ be a corresponding support point, and let Γ be its omitted arc. Then Γ satisfies the differential equation (1) with $\Phi(w) = B^2(B-w)^{-1}$ and $B = f(\zeta)$. Brown [2] integrated the differential equation to show that Γ is essentially the image under the Koebe function of a logarithmic spiral. This and the implicit determination of B as a function of ζ enabled him to show that Γ has monotonic argument and monotonic radial angle, and that Γ lies in a certain half-strip excluding the origin. We shall now use the tools developed in §2 to recapture some of these qualitative properties directly from the differential equation.

The asymptotic half-line (2) reduces to

$$w = \frac{B}{3} - B^2 t, \qquad 0 \le t < \infty.$$

The accessible region is a half-plane bounded by the line through B orthogonal to the asymptotic half-line. The critical set is the line

$$w = B - B^2 t, \qquad -\infty < t < \infty,$$

through B parallel to the asymptotic half-line (see Figure 1). We may assume that $\operatorname{Im}\{B\} > 0$, since the class S is preserved under conjugation and f is the Koebe function if B is real. The asymptotic half-line lies between the origin and the critical line. If Γ were to enter the quarter-plane I where $\operatorname{Im}\{\Phi(w)\} < 0$, it would be forced to cross the critical line radially, violating its monotonic modulus property, in order to approach the asymptotic half-line. Thus Γ is confined to the quarter-plane II where $\operatorname{Im}\{\Phi(w)\} > 0$, which contains the asymptotic half-line. In this region, $\arg w$ must decrease as $|w|$ increases. This proves that Γ has monotonic argument.

If $\operatorname{Re}\{B\} \le 0$, then the origin lies in the forbidden half-plane or on its boundary (because $\Phi(0) = B$), and Γ is restricted to a half-plane bounded by the line

$$w = iB^2 t, \qquad -\infty < t < \infty,$$

through the origin. The monotonicity of $\arg w$ then restricts Γ to the quarter-plane bounded by the rays $w = -B^2 t$ and $w = -iB^2 t$, $t \ge 0$. This quarter-plane

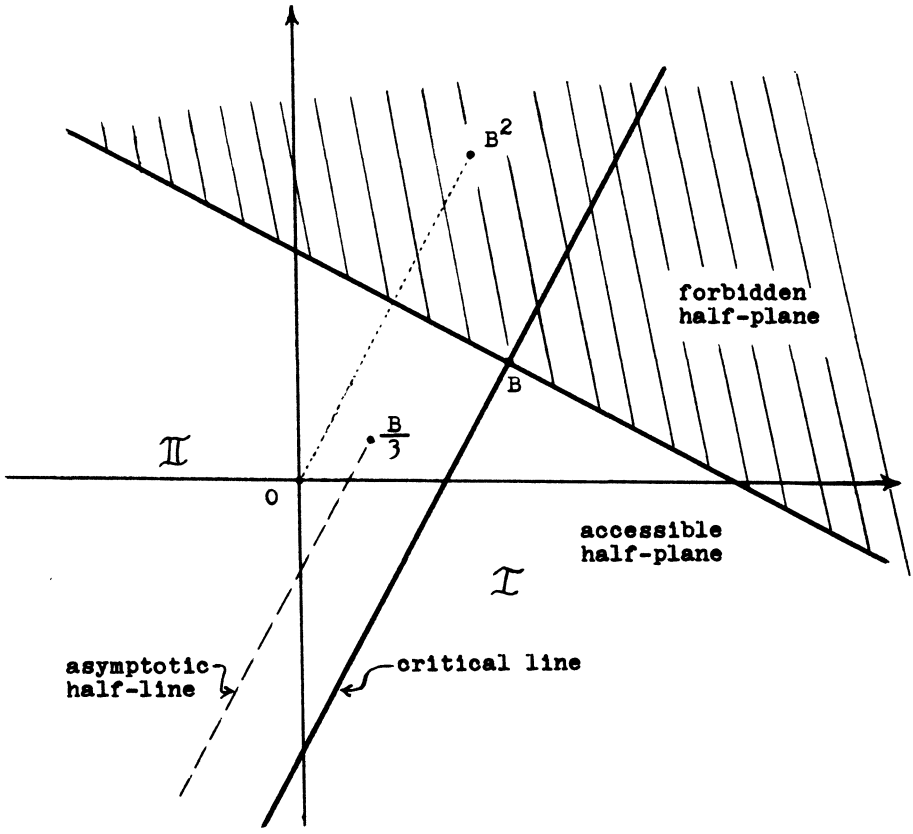


Figure 1

contains the asymptotic half-line near infinity. The case where $\mathrm{Re}\{B\}>0$ is more difficult.

Now consider the radial angle. The equation (5), which determines the locus of points where the radial angle may be stationary, takes the form

$$B^2(B-w)w^{-2}\geq 0.$$

This is a curve γ from B to ∞, passing through the origin. It may be analyzed by setting

$$B^2(B-w)w^{-2}=t>0$$

and observing that w is then determined by the quadratic equation

$$tw^2+B^2w-B^3=0,$$

with solutions

$$w=-\frac{B^2}{2t}\left\{1\pm\left(1+\frac{4t}{B}\right)^{1/2}\right\}.$$

For small t these two solutions are

$$w=B-t+O(t^2)$$

and

$$w=-\frac{B^2}{t}-B+t+O(t^2).$$

As $t\to 0$, the first solution approaches B asymptotic to the half-line $w=B-t$, while the second tends to infinity asymptotic to the half-line $w=-B-B^2/t$. For large t the two solutions are

$$w=\pm B^{3/2}t^{-1/2}+O(t^{-1}), \qquad t\to\infty,$$

which approach the origin in opposite directions. Because Γ and γ have different (but parallel) asymptotes near infinity, it follows that they have no intersections outside a circle of sufficiently large radius. Thus the radial angle of Γ is eventually monotonic.

§4. Applications to the derivative problem

We now return to the linear problem (3) of maximizing $\operatorname{Re}\{e^{-i\sigma}f'(\zeta)\}$ over all functions $f\in S$, where σ is an arbitrary real number and ζ is an arbitrary point in the unit disk. We shall prove the following theorem.

THEOREM. *Let $f\in S$ be a function which maximizes $\operatorname{Re}\{e^{-i\sigma}g'(\zeta)\}$ among all $g\in S$, and let Γ be its omitted arc. Then Γ has monotonic argument.*

Proof. Since S is preserved under rotations, we may assume without loss of generality that $0<\zeta<1$. Specializing the general result for support points, we see

that an extremal function f must map the disk onto the complement of an analytic arc Γ which satisfies

$$\Phi(w)\frac{dw^2}{w^2}>0, \tag{6}$$

where

$$\Phi(w)=\frac{e^{-i\sigma}CB(B-2w)}{(B-w)^2}, \qquad B=f(\zeta), \qquad C=f'(\zeta).$$

The asymptotic half-line (2) takes the form

$$w=\frac{B}{2}-e^{-i\sigma}CBt, \qquad t\geq 0.$$

If $e^{-i\sigma}C$ is real, the asymptotic half-line is radial and is a trajectory of the quadratic differential. This implies that Γ is a radial half-line and that f is a rotation of the Koebe function. Thus we may assume that $e^{-i\sigma}C$ is not real. In fact, we may assume without essential loss of generality that $\operatorname{Im}\{e^{-i\sigma}C\}>0$, since S is preserved under conjugation and so the region of values of $g'(\zeta)$ is symmetric with respect to the real axis. More specifically, f maximizes $\operatorname{Re}\{e^{-i\sigma}g'(\zeta)\}$ if and only if its conjugate $\bar{f}$ (defined by $\bar{f}(z)=\overline{f(\bar{z})}$) maximizes $\operatorname{Re}\{e^{i\sigma}g'(\zeta)\}$. Having made these reductions, we may set

$$e^{-i\sigma}C=e^{2i\gamma}|C|, \qquad 0<\gamma<\frac{\pi}{2}.$$

We also let

$$\omega=\xi+i\eta=\frac{e^{i\gamma}w}{B-w}, \tag{7}$$

whereupon a simple calculation gives

$$\Phi(w)=|C|\{e^{2i\gamma}-\omega^2\}.$$

The boundary of the accessible region is the curve $\operatorname{Re}\{\Phi(w)\}=0$, which in the ω-plane is the (possibly degenerate) hyperbola

$$\xi^2-\eta^2=\cos 2\gamma. \tag{8}$$

The critical curve $\operatorname{Im}\{\Phi(w)\} = 0$ transforms to the hyperbola

$$2\xi\eta = \sin 2\gamma \tag{9}$$

in the ω-plane.

The variables w and ω are related by the linear fractional transformation (7) with the following table of values:

w	0	∞	$B/2$	B
ω	0	$-e^{i\gamma}$	$e^{i\gamma}$	∞

The line in the w-plane through 0 and B is mapped onto the line in the ω-plane through $e^{i\gamma}$ and $-e^{i\gamma}$. More generally, an arbitrary radial half-line in the w-plane is mapped onto a circular arc joining 0 and $-e^{i\gamma}$ in the ω-plane. Circles in the w-plane centered at the origin are mapped onto circles of Apollonius in the ω-plane with 0 and $-e^{i\gamma}$ as inverse points. The asymptotic half-line in the w-plane is mapped onto a circular arc in the ω-plane joining $e^{i\gamma}$ and $-e^{i\gamma}$. The typical situation is illustrated in Figures 2 and 3.

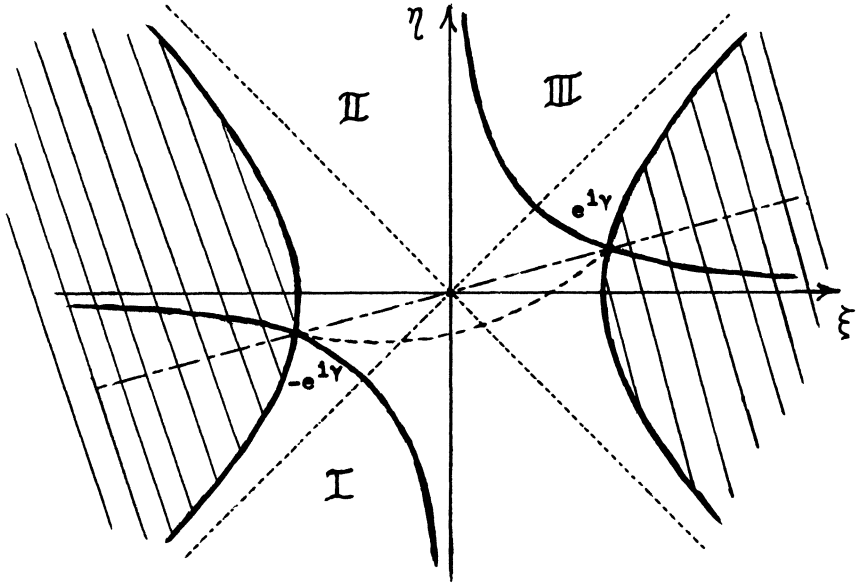


Figure 2. ω-plane, $0 < \gamma < \pi/4$.

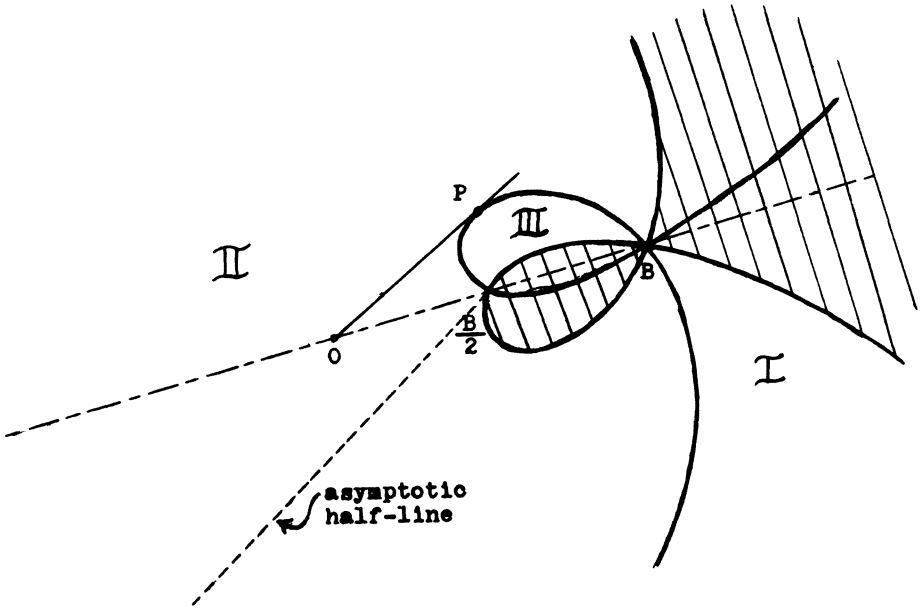


Figure 3. w-plane, $0<\gamma<\pi/4$.

Suppose first that $0<\gamma<\pi/4$, so that $\cos 2\gamma>0$ and $\sin 2\gamma>0$. The hyperbolas (8) and (9) are then as shown in Figure 2. Observe that the two hyperbolas intersect at the points $\pm e^{i\gamma}$. The image of the forbidden region in the ω-plane is the shaded region determined by the inequality

$$\cos 2\gamma+\eta^2<\xi^2.$$

The critical curve divides the accessible region into three subregions which are labeled I, II, and III in Figure 3. Observe that $\operatorname{Im}\{\Phi(w)\}<0$ in Regions I and III, while $\operatorname{Im}\{\Phi(w)\}>0$ in Region II. According to our previous remarks (§2), this implies that as Γ is traversed with increasing modulus, its argument must decrease in Region II and increase in Regions I and III. Should Γ ever pass from one region to another, it must have radial direction as it crosses the critical curve.

Our first observation is that the asymptotic half-line

$$w=\frac{B}{2}-e^{2i\gamma}Bt, \qquad t\geq 0, \tag{10}$$

lies entirely in Region II, except perhaps for an initial segment near $B/2$. Indeed,

its image in the ω-plane is a circular arc from $e^{i\gamma}$ to $-e^{i\gamma}$ which makes an angle of 2γ with the line joining these two points, while the hyperbola (9) has slope $d\eta/d\xi = -\tan\gamma$ at $-e^{i\gamma}$. Referring to Figure 2, we conclude that the circular arc is tangent to (9) at $-e^{i\gamma}$, which implies that the critical curve is tangent to the asymptotic half-line at ∞.

We claim that Γ lies entirely in Region II and so has decreasing argument. Suppose first that Γ has points in Region I. Then, in order to reach the asymptotic half-line at infinity, it must either cross the critical curve radially into Region II or approach the critical curve asymptotically. If Γ crosses the critical curve from Region I to Region II, it must do so with decreasing modulus, because each ray intersects this portion of the critical set at most once. (This follows from the fact that the image of such a ray is a circular arc in the ω-plane from 0 to $-e^{i\gamma}$, which can have at most one other intersection with the lower branch of the hyperbola (9).) If Γ approaches the critical curve asymptotically at infinity within Region I, it must do so with decreasing argument. (This is again seen by considering its image in the ω-plane, which would approach the point $-e^{i\gamma}$ within the image of Region I and would be tangent to the critical hyperbola there.) In Region I, however, we know that Γ must have *increasing* argument. This shows that no part of Γ can lie in Region I.

Suppose next that Γ has points in Region III. Then, in order to approach the asymptotic half-line at infinity, it must eventually cross the critical curve (radially) into Region II. To see that this is impossible, we observe that the critical curve which separates Region II from Region III is divided into two subarcs by a point P where the ray from the origin is tangent to the curve. At every point on the open subarc from $B/2$ to P, rays from the origin pass from Region II to Region III; while on the subarc from P to B, rays pass from Region III to Region II. (Again consider the image of such a ray in the ω-plane, which is a circular arc from 0 to $-e^{i\gamma}$. This circle intersects the upper branch of the hyperbola (9) at most twice.) If Γ crosses the critical curve from Region III to Region II, it must cross radially along the subarc from P to B (possibly at P), because of its increasing modulus. In Region II, however, the argument of Γ must decrease. Therefore, after Γ passes into Region II, it is blocked away from the asymptotic half-line by a curve which extends along the critical set from P to B, then radially from B to ∞. The radial half-line from B to ∞ lies in the forbidden region. This shows that no part of Γ can lie in Region III. Thus Γ lies entirely in Region II, where its argument is strictly decreasing. This concludes the proof that Γ has monotonic argument if $0 < \gamma < \pi/4$.

The case $\pi/4 < \gamma < \pi/2$ is somewhat similar. The hyperbola (9) remains in the same position, but the hyperbola (8) now has a vertical axis. The image in the ω-plane of the forbidden region now lies between the two branches of the

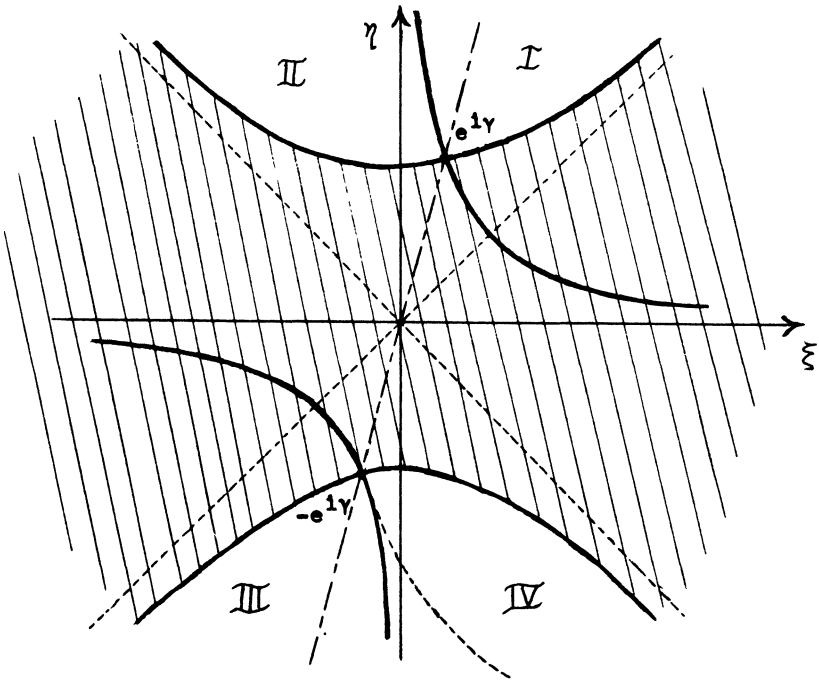


Figure 4. ω-plane, $\pi/4 < \gamma < \pi/2$.

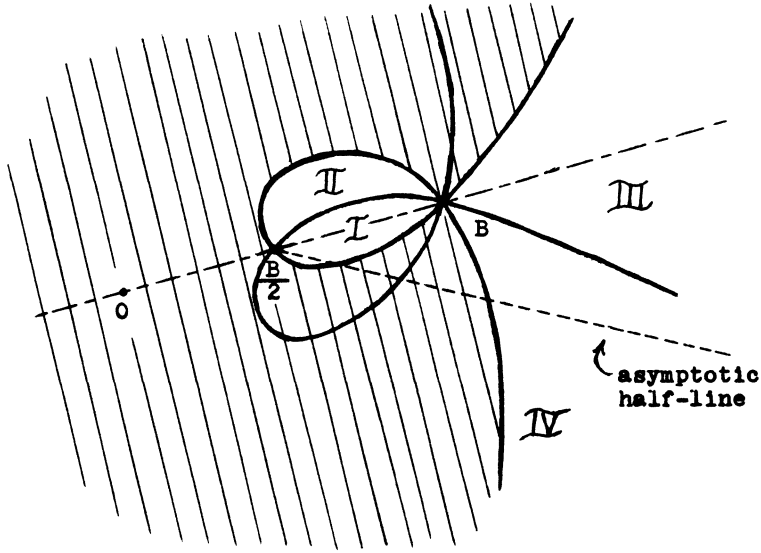


Figure 5. w-plane, $\pi/4 < \gamma < \pi/2$.

hyperbola (8). The critical curve corresponds to the hyperbola (9) and divides the accessible region into four subregions. These are called Regions I, II, III, and IV, as indicated in Figures 4 and 5. Now $\operatorname{Im}\{\Phi(w)\}<0$ in Regions I and III, while $\operatorname{Im}\{\Phi(w)\}>0$ in Regions II and IV. The asymptotic half-line lies eventually in Region IV and is tangent to the critical curve at infinity. (This is seen by considering the image of the asymptotic half-line in the ω-plane, an arc of a circle through $e^{i\gamma}$ and $-e^{i\gamma}$ which makes an angle of $\pi-2\gamma$ with the line through these two points. Geometric considerations reveal that this circular arc is tangent to the lower branch of the hyperbola (9) at $-e^{i\gamma}$.) No part of Γ can lie in Region I or II, because these are bounded regions entirely surrounded by the forbidden region. (It is clear that $B\notin\Gamma$, since $f(\zeta)=B$.)

If any part of Γ lies in Region III, then eventually either Γ crosses the critical curve (radially) into Region IV, or Γ approaches the asymptotic half-line within Region III. However, a ray from the origin may meet this part of the critical curve at most once, where it crosses from Region IV into Region III. (This is seen by observing that its image in the ω-plane is a circular arc through 0 and $-e^{i\gamma}$ which meets the lower branch of the hyperbola (9) at $-e^{i\gamma}$ and at one other point at most, where it must cross.) Thus if Γ crosses from Region III to Region IV, it must do so with decreasing modulus, which is impossible. Finally, it is clear from Figure 5 that if Γ approaches the asymptotic half-line within Region III, it must eventually have decreasing argument. But this is impossible, since $\operatorname{Im}\{\Phi(w)\}<0$ in Region III, and so Γ has increasing argument there. Thus Γ is confined entirely to Region IV, where it has decreasing argument.

The final case $\gamma=\pi/4$ is quite special. Here $e^{-i\sigma}C=i|C|$, and it is easily verified that the asymptotic half-line

$$w=\frac{B}{2}-iBt, \qquad t>0,$$

satisfies the differential equation (6) for the omitted arc Γ. As we have already remarked (§1), this implies that Γ is itself a half-line which coincides with the asymptotic half-line near infinity. Since a half-line clearly has monotonic argument, this completes the proof of the theorem.

The last case merits closer inspection. What is the most general situation in which Γ is a half-line? This occurs if and only if the asymptotic half-line (10) is a trajectory (near infinity) of the quadratic differential (6). A direct calculation shows that this is true if and only if $e^{4i\gamma}$ is real. If $e^{4i\gamma}=1$, then $e^{2i\gamma}=\pm1$ and Γ is a radial half-line, which implies that f is a rotation of the Koebe function. If $e^{4i\gamma}=-1$, then $e^{2i\gamma}=\pm i$ and $e^{-i\sigma}C=\pm i|C|$. On the other hand,

$$\max_{g\in S}\operatorname{Re}\{e^{-i\sigma}g'(\zeta)\}=\max_{g\in S}\operatorname{Im}\{ie^{-i\sigma}g'(\zeta)\}=0$$

if and only if $C = -ie^{i\sigma}|C|$, and it is clear geometrically that this occurs if and only if the extremal problem is equivalent to that of finding the maximum of $\arg g'(\zeta)$. Similarly, $C = ie^{i\sigma}|C|$ if and only if the extremal problem is equivalent to the minimum argument problem. In other words, the only cases (aside from rotations of the Koebe function) in which Γ is linear are those identified by Pearce [7], where the extremal problem (3) is equivalent to the maximum or minimum argument problem. It should be remarked that we have been led to this result without appeal to the rotation theorem and without prior knowledge of its specific extremal functions in the case $|\zeta| \leq 1/\sqrt{2}$. In fact, our approach demonstrates *a priori* that whenever ζ is chosen so that $\operatorname{Re}\{g'(\zeta)\} \geq 0$ for all $g \in S$, the maximum and minimum of $\arg g'(\zeta)$ are attained *only* by the elementary rational functions which map the disk onto the complement of a half-line.

Acknowledgements. This work was supported in part by a grant from the National Science Foundation. The author wishes to thank Johnny Brown for helpful criticism and suggestions.

REFERENCES

[1] L. BRICKMAN and D. WILKEN, "Support points of the set of univalent functions", Proc. Amer. Math. Soc. *42* (1974), 523–528.
[2] J. E. BROWN, "Geometric properties of a class of support points of univalent functions", Trans. Amer. Math. Soc. *256* (1979), 371–382.
[3] J. E. BROWN, "Univalent functions maximizing $\operatorname{Re}\{a_3 + \lambda a_2\}$", Illinois J. Math., to appear.
[4] P. L. DUREN, "Extreme points of spaces of univalent functions", in *Linear Spaces and Approximation* (Birkhäuser Verlag, Basel, 1978); pp. 471–477.
[5] P. L. DUREN, *Univalent Functions* (Springer-Verlag, Heidelberg and New York, to appear).
[6] G. M. GOLUZIN, *Geometric Theory of Functions of a Complex Variable* (Moscow, 1952; German transl., Deutscher Verlag, Berlin, 1957; 2nd ed., Moscow, 1966; English transl., Amer. Math. Soc., 1969).
[7] K. PEARCE, "New support points of S and extreme points of HS", Proc. Amer. Math. Soc., 81 (1981), 425–428.
[8] A. PFLUGER, "Lineare Extremalprobleme bei schlichten Funktionen", Ann. Acad. Sci. Fenn. Ser. AI, no. 489 (1971), 1–32.
[9] A. C. SCHAEFFER and D. C. SPENCER, *Coefficient Regions for Schlicht Functions* (with a chapter on the region of values of the derivative of a schlicht function by Arthur Grad) (Amer. Math. Soc. Colloq. Publ., vol 35, 1950).
[10] G. SCHOBER, *Univalent Functions – Selected Topics*, Lecture Notes in Mathematics 478 (Springer-Verlag, Heidelberg and New York, 1975).

Department of Mathematics
University of Michigan
Ann Arbor, Michigan 48109
U.S.A.

Received November 25, 1980

Comment. Math. Helvetici **56** (1981) 366–403 0010-2571/81/003366-38$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Level sets of univalent functions

W. K. HAYMAN and J.-M. G. WU*

1. Introduction

Let $w=f(z)$ be a univalent (one to one analytic) map from $\Delta: |z|<1$ onto a domain Ω in the closed complex plane. By a level set of f we mean the preimage of $f^{-1}(\Omega \cap L)$ for some straight line or circle L. Piranian and Weitsman asked in [7] if every level set of f has finite length. Here we give an affirmative answer. We shall denote by $A_0, A_1, \ldots$ positive absolute constants. If E is any set, $\delta(E)$ is the diameter and $|E|$ is the length or one-dimensional Hausdorff measure of E. Then our result is

THEOREM 1. *If E is a subset of a level set of a univalent function, then*

$$|E| \leq A_0 \delta(E) \leq 2A_0 \tag{1.1}$$

where $A_0 < 10^{35}$.

A special case of the theorem when Ω is a Lipschitz domain was proved in [8].

Our argument is rather lengthy and falls naturally into three parts. In the first part (Sections 2 to 5) we assume that E is the full level set γ. The components, i.e., maximal connected subsets of γ, will be called level curves and denoted by γ_k. We shall then prove in Section 5,

$$\sum \delta(\gamma_k) \leq A_1 = 2.1 \times 10^{16}. \tag{1.2}$$

About three weeks after we obtained this result a proof of (1.2) was obtained independently by Gehring and Jones [1]. It is also worth noting that Jones [4] has constructed a bounded analytic function on Δ for which every level set $|w|=R$ is either empty or has infinite length. Thus the analogue of Theorem 1 for bounded functions is false.

In the second part (Sections 6 to 10), we prove that for any individual level

* Research of the second author is partially supported by the National Science Foundation.

curve γ_k, we have

$$|\gamma_k| \leq A_2 < 10^{18}. \tag{1.3}$$

These two parts make up the bulk of the paper.

In the third part (Section 11) we complete the proof of Theorem 1. We show first by an elementary transformations that (1.3) leads very simply to

$$|\gamma_k| \leq 2A_2\delta(\gamma_k). \tag{1.4}$$

Using this and (1.2) we deduce at once that

$$|\gamma| = \sum |\gamma_k| \leq 2A_2 \sum \delta(\gamma_k) \leq 2A_1A_2. \tag{1.5}$$

A similar argument to that leading from (1.3) to (1.4) then shows that if E is any subset of γ we have

$$|E| \leq 4A_1A_2\delta(E)$$

and this gives Theorem 1. This part is relatively short.

Our arguments in parts I and II are based on harmonic measure. If D is a domain and E is a subset of the closure $\bar{D}$ of D, we denote by $\omega(z, E, D)$ a function which is harmonic in $D \setminus E$ and has boundary value 1 on E and zero on $\partial D \setminus E$. Thus if E is a subset of ∂D, $\omega(z, E, D)$ is the harmonic measure of E with respect to D at z. Clearly $\omega(z, E, D)$ increases with expanding E for fixed z and D and with expanding D for fixed z and E. If no confusion is likely because D is fixed we sometimes write simply $\omega(z, E)$.

2. Elementary lemmas on harmonic measure

The only property of the γ_k which we shall use in order to prove (1.2) is that the γ_k are arcs in Δ with end points on the boundary of Δ, which satisfy the separation condition involving harmonic measure which is described in Lemma 1. However, in order to use Lemma 1 we need various other properties of harmonic measure.

LEMMA 1. *If γ_k are the components of the level set γ and $\gamma_k' = \gamma \setminus \gamma_k$, then*

$$\omega(z, \gamma_k', \Delta) < \tfrac{1}{2}, \qquad z \in \gamma_k. \tag{2.1}$$

We recall that $w = f(z)$ maps Δ onto the domain Ω. Then γ_k is mapped onto an arc l of L and γ_k' is mapped onto the remainder l' of $L \cap \Omega$. Clearly we may

assume that L is the real axis, and (2.1) is equivalent to

$$\omega(w, l', \Omega) < \tfrac{1}{2}, \qquad w \in l \tag{2.2}$$

in view of the invariance of harmonic measure under conformal mapping. To prove (2.2) we construct the reflection Ω^* of Ω in L and define U to be that component of $\Omega \cap \Omega^*$, which contains l. Write

$$\omega(w) = \omega(w, l', \Omega), \qquad \omega^*(w) = \omega(w) + \omega(\bar{w});$$

then $\omega^*(w)$ is harmonic in U. Let ξ be a boundary point of U. Then, since Ω is simply-connected, ξ cannot lie in l'. Thus ξ is either a boundary point of Ω^*, in which case $\omega(\xi) = 0$, $\omega(\bar{\xi}) < 1$, or a boundary point of Ω^*, in which case $\omega(\bar{\xi}) = 0$, $\omega(\xi) < 1$. Thus $\omega^*(\xi) < 1$ and so $\omega^*(w) < 1$ in U. Taking for w a point on l, we deduce (2.2) and hence Lemma 1.

If Ω is the region $-\theta < \arg z < 2\pi - \theta$, where $0 < \theta < \pi$, and l and l' are the negative and positive real axes, we have $\omega(w) = (\pi - \theta)/(2\pi - \theta)$ on l, so that (2.1) is sharp.

Our next lemma is a special case of the Milloux–Schmidt inequality [3, p. 109] and [6, p. 107].

LEMMA 2. *Let η be an arc in $\Delta_R = \{z \mid |z| < R\}$, which passes through the origin and has one end point on $|z| = R$. Then*

$$\omega(z, \eta, \Delta_R) \geq 1 - \frac{4}{\pi} \arctan\, (|z|/R)^{1/2}.$$

We also need a variant of Hall's Lemma [2]. It is pointed out by David Drasin that a theorem which bears some resemblance to Lemma 3 in disks was proved by Maitland [5].

LEMMA 3. *Suppose that $H = \{z \mid z = x + iy, y > 0\}$ is the upper half plane. Let E be a relatively closed set in $\{0 < y < a/100\}$ and let $E^* = \{x \mid x + iy \in E\}$ be the projection of E on the real axis. Then for* $\operatorname{Im} z \geq a$,

$$\omega(z, E, H) \geq \tfrac{2}{3}\omega(z, E^*, H).$$

As in the proof of Hall's Lemma we may assume that E is the union of a finite or countable set of line segments l_k, which are parallel to the real axis and whose projections have at most endpoints in common. This is also the only case of

Lemma 3 which we use. Further we assume, as we may, that $a = 1$. If

$$l_k = \{t + ib_k,\ a_k \le t \le a_k'\}$$

we define for $z \in H$

$$U(z) = \frac{1}{2\pi} \sum_k \frac{1}{b_k} \int_{l_k} G(z, \zeta)\, |d\zeta|,$$

where

$$G(z, \zeta) = \log \left| \frac{\bar{z} - \zeta}{z - \zeta} \right|$$

is the Green function in H. When $\zeta = t + ib$, $z = x + iy$, we set $|z - \zeta| = r$, $|z - t| = \rho$ and assume that $y \ge 1 \ge 100b$. Then

$$\frac{1}{b} \log \left| \frac{\bar{z} - \zeta}{z - \zeta} \right| = \frac{1}{2b} \log \left(1 + \frac{4by}{r^2}\right) \ge \frac{1}{2b} \log \left(1 + \frac{4by}{\rho^2}\right)$$

$$\ge \frac{2y}{\rho^2} \Big/ \left(1 + \frac{4by}{\rho^2}\right) \ge \frac{2y}{\rho^2} \Big/ \left(1 + \frac{1}{25}\right) = \frac{25}{26} \frac{2y}{\rho^2}.$$

Thus

$$U(z) \ge \frac{25}{26} \frac{1}{\pi} \sum_k \int_{a_k}^{a_k'} \frac{y}{\rho^2}\, dt = \frac{25}{26}\, \omega(z, E^*). \tag{2.3}$$

Next, we prove, following Hall, that

$$U(z) \le \frac{\pi}{4} + \frac{2}{\pi}, \qquad z \in H. \tag{2.4}$$

We recall that

$$\frac{1}{b} \log \left| \frac{\bar{z} - \zeta}{z - \zeta} \right| = \frac{1}{2b} \log \left(1 + \frac{4by}{r^2}\right).$$

The right hand side decreases with increasing b when z and r are fixed and so

assumes its maximum value for $b = y - r$ if $r < y$ and for $b = 0$ when $r \geq y$. Thus

$$\frac{1}{b} \log \left| \frac{\bar{z} - \zeta}{z - \zeta} \right| \leq M(r, z),$$

where

$$M(r, z) = \frac{1}{y - r} \log \left(\frac{2y}{r} - 1 \right), \qquad r < y$$

$$M(r, z) = \frac{2y}{r^2}, \qquad r \geq y.$$

Also $M(r, z)$ decreases with increasing r and so

$$\frac{1}{b} \log \left| \frac{\bar{z} - \zeta}{z - \zeta} \right| \leq M(r, z) \quad \text{if} \quad |z - \zeta| \geq r.$$

In particular this inequality holds if $\zeta = t + ib$, where $t - x = \mp r$. Thus

$$U(z) \leq \frac{1}{\pi} \int_0^\infty M(r, z)\, dr$$

$$= \frac{1}{\pi} \int_0^y \frac{1}{y - r} \log \left(\frac{2y}{r} - 1 \right) dr + \frac{1}{\pi} \int_y^\infty \frac{2y}{r^2}\, dr.$$

The second integral is $2/\pi$. To evaluate the first integral set $y - r = ty$, then

$$\frac{1}{\pi} \int_0^y \frac{1}{y - r} \log \left(\frac{2y - r}{r} \right) dr = \frac{1}{\pi} \int_0^1 \log \left(\frac{1 + t}{1 - t} \right) \frac{dt}{t}$$

$$= \frac{1}{\pi} \sum_0^\infty \frac{2}{(2n + 1)^2} = \frac{\pi}{4}.$$

This proves (2.4). Now the maximum principle yields

$$U(z) \leq \left(\frac{\pi}{4} + \frac{2}{\pi} \right) \omega(z, E), \qquad z \in H. \tag{2.5}$$

Combining (2.3) and (2.5), we conclude for $y \geq 1$,

$$\omega(z, E) \geq \left(\frac{\pi}{4} + \frac{2}{\pi} \right)^{-1} \tfrac{25}{26} \omega(z, E^*) \geq \tfrac{2}{3} \omega(z, E^*)$$

and this proves Lemma 3.

LEMMA 4. *Suppose that E is a subset of the real axis, that $a>0$ and that I is the interval $[-10a, 10a]$. Then if*

$$\omega(ia, E, H)<7/8$$

we have

$$|I\setminus E|\geq \tfrac{1}{120}|I|.$$

Write $E'=I\setminus E$, $|E'|=2\eta$. We suppose without loss of generality that $a=1$. Then if $\lambda(t)$ denotes the measure of $E'\cap[-t, t]$ we have

$$\omega(i, E')=\frac{1}{\pi}\int_0^{10}\frac{d\lambda(t)}{1+t^2}.$$

Our hypotheses imply that $\lambda(t)\leq 2\min(t, \eta)$. Thus

$$\omega(i, E')\leq\frac{2}{\pi}\int_0^{\eta}\frac{dt}{1+t^2}=\frac{2}{\pi}\tan^{-1}\eta.$$

Also

$$\tfrac{7}{8}>\omega(i, E)\geq\omega(i, I)-\omega(i, E')\geq\frac{2}{\pi}(\tan^{-1}10-\tan^{-1}\eta).$$

Thus

$$\frac{2}{\pi}\tan^{-1}\eta\geq\frac{2}{\pi}\tan^{-1}10-\tfrac{7}{8}=\tfrac{1}{8}-\frac{2}{\pi}\tan^{-1}\tfrac{1}{10}>\tfrac{1}{8}-\frac{1}{5\pi},$$

$$\eta>\frac{\pi}{16}-\tfrac{1}{10}>0.09>\tfrac{1}{12}.$$

This proves Lemma 4.

LEMMA 5. *Suppose that η is an arc in $\{z\,|\,|x|\leq 32, 0<y<1\}$ which connects the lines $x=\mp 32$. Then if $1<y_0\leq 2$, we have*

$$\omega(iy_0, \eta, H)>\tfrac{18}{19}.$$

Let $l=32$. Let η_1, η_2 denote the segments $\{-l+iy, 0\le y\le 1\}$, $\{l+iy, 0\le y\le 1\}$ respectively. Let η_3, η_4 denote the segments $\{-l-1\le x\le -l+1, y=0\}$ and $\{l-1\le x\le l+1, y=0\}$. Then

$$\omega(z,\eta_3)\ge\tfrac{1}{2} \text{ on } \eta_1 \quad \text{and} \quad \omega(z,\eta_4)\ge\tfrac{1}{2} \text{ on } \eta_2.$$

Thus for $z\in H$

$$\omega(z,\eta_1)+\omega(z,\eta_2)\le 2(\omega(z,\eta_3)+\omega(z,\eta_4)).$$

Finally if η_5 denote the segment $\{-l\le x\le l, y=0\}$ then the maximum principle shows that for $z=x_0+iy_0$, where $y_0>1$,

$$\omega(z,\eta_5)\le\omega(z,\eta)+\omega(z,\eta_1)+\omega(z,\eta_2),$$

so that

$$\omega(z,\eta)\ge\omega(z,\eta_5)-2(\omega(z,\eta_3)+\omega(z,\eta_4)).$$

Setting $z=iy_0$ where $1\le y_0\le 2$, we obtain

$$\begin{aligned}\omega(z,\eta)&\ge\frac{2}{\pi}\left(\tan^{-1}\frac{l}{2}-2\tan^{-1}\frac{l+1}{2}+2\tan^{-1}\frac{l-1}{2}\right)\\&=1-\frac{2}{\pi}\left(\tan^{-1}\frac{2}{l}+2\tan^{-1}\frac{4}{l^2+3}\right)\\&>1-\frac{4}{\pi}\left(\frac{1}{l}+\frac{4}{l^2+3}\right)>\tfrac{18}{19}.\end{aligned}$$

This proves Lemma 5.

3. Construction of segments

The inequality (2.1) is the only property of the level curves which we use in order to complete our results. It is convenient to work in H rather than Δ since the noneuclidean metric is easier to visualize in H. We consider in the first instance only those level curves which lie entirely in the square

$$R_0: 0\le x\le 1, \qquad 0\le y\le 1. \tag{3.1}$$

More precisely we assume that $w = F(z)$ is a univalent map from H into the complex plane, whose image does not contain the whole real axis L in the ω plane, and denote by l_j those components of $F^{-1}(L)$ which lie entirely in R_0. (The simpler case $L \subseteq F(H)$ will be considered at the end of Section 11.) If

$$l_k' = \bigcup_{j \neq k} l_j,$$

it follows from Lemma 1, the invariance of harmonic measure and the fact that $\omega(z, E, H)$ increases with expanding E that

$$\omega(z, l_k', H) \leq \tfrac{1}{2} \quad \text{on} \quad l_k. \tag{3.2}$$

Apart from (3.2) we only use the fact that the l_k are arcs which lie in R_0 and have both end points on the segment $0 \leq x \leq 1$, $y = 0$.

It proves convenient to replace each l_k by certain horizontal line segments q_k of diameter comparable to the diameter of l_k. We can do this at the cost of replacing $\frac{1}{2}$ by $\frac{7}{12}$ in (3.2). Using Lemma 3 we shall deduce that the projection of a suitable subset of $\bigcup_{j \neq k} q_j$ has harmonic measure less than $\frac{7}{8}$ on q_k and so by Lemma 4 must leave uncovered $\frac{1}{120}$ of a neighborhood of the projection of q_k. From this we can deduce that the sum of the diameters of the q_k and so that of the l_k is bounded by a (very large) absolute constant and (1.2) will follow.

Let m, n be integers such that

$$n \geq 10, \qquad 0 < m \leq 2^n.$$

We shall call a dyadic *segment* the set

$$q(m, n): (m-1)2^{-n} < x < m2^{-n}, \qquad y = 600 \cdot 2^{-n}. \tag{3.3}$$

The projection

$$J(m, n): (m-1)2^{-n} < x < m2^{-n}, \qquad y = 0 \tag{3.4}$$

of $q(m, n)$ on the real axis will be called a dyadic *interval*. We note that two dyadic segments are disjoint, unless they are identical. Two dyadic intervals are disjoint unless one is contained in the other. Also all dyadic segments lie in R_0.

LEMMA 6. *If l_k is a level curve in R_0 there exists a dyadic segment q_k such that*

$$\delta(q_k) > 10^{-5} \delta(l_k) \tag{3.5}$$

and

$$\omega(z, l_k, H) > \tfrac{18}{19} \quad \textit{for all } z \textit{ on } q_k. \tag{3.6}$$

Let x_1, x_2 be the lower and upper bounds of x on l_k and let h be the upper bound of y on l_k. We write $x_2 - x_1 = 2d$ and distinguish two cases.

(i) Suppose first that $d > 40h$. In this case we define n to be the largest integer such that

$$2^{-n} > \frac{d}{24{,}000},$$

and then define m to be the largest integer such that

$$m2^{-n} \geq \tfrac{1}{2}(x_1 + x_2).$$

Let $q(m, n)$ be the corresponding segment given by (3.3), and suppose that $\zeta = \xi + i\eta \in q(m, n)$. Then

$$\frac{d}{40} < \eta \leq \frac{d}{20}.$$

Also if $x_0 = \frac{1}{2}(x_1 + x_2)$, we have for $\zeta = \xi + i\eta$ on $q(m, n)$

$$|\xi - x_0| \leq 2^{-n} < (12{,}000)^{-1}\, d < \frac{d}{5}.$$

Thus l_k has a subarc l', lying in the rectangle

$$|x - \xi| \leq \tfrac{32}{40} d, \qquad 0 < y < \frac{d}{40},$$

and joining the sides $x = \xi \mp 32d/40)$ of this rectangle, since by hypothesis $h \leq d/(40)$. Now it follows from Lemma 5, $d/40 < \eta \leq d/20$ and conformal invariance that

$$\omega(\zeta, l_k) \geq \omega(\zeta, l') > \tfrac{18}{19},$$

which gives (3.6). Further, using our choice of n we see

$$\delta(l_k) \leq (h^2 + (2d)^2)^{1/2} \leq 2d(1 + (\tfrac{1}{80})^2)^{1/2} < 50{,}000 \cdot 2^{-n},$$

which yields (3.5). Thus Lemma 6 is proved in this case.

(ii) Suppose next that $d \leq 40h$. We choose for n the least integer such that

$$600 \cdot 2^{-n} \leq h,$$

and set $\eta = 600 \cdot 2^{-n}$, so that

$$h/2 < \eta \leq h.$$

Since l_k has endpoints on the positive axis we see that l_k contains a point $\zeta_0 = \xi_0 + i\eta$. If $\xi_0 = 0$ we choose $m = 1$ and otherwise we choose for m the smallest integer such that $m \cdot 2^{-n} \geq \xi_0$. Then if ζ is any point on $q(m, n)$, defined by (3.3), we have

$$|\zeta - \zeta_0| \leq 2^{-n} = \eta/600 < \eta \tan^2\left(\frac{\pi}{76}\right).$$

Now l_k contains a subarc l' lying in the disk $D_0 : |z - \zeta_0| < \eta$ and joining the point ζ_0 to the circumference of D_0. Hence Lemma 2 shows that for $\zeta \in q(m, n)$,

$$\omega(\zeta, l_k, H) \geq \omega(\zeta, l', H) \geq \omega(\zeta, l', D_0)$$
$$\geq 1 - \frac{4}{\pi} \arctan \sqrt{\frac{|\zeta - \zeta_0|}{\eta}} \geq 1 - \frac{4}{\pi} \cdot \frac{\pi}{76} = \frac{18}{19}.$$

Also $\delta(l_k) \leq ((2d)^2 + h^2)^{1/2} \leq h(1 + (80)^2)^{1/2}$

$$< 81h < 81 \cdot 1200 \cdot 2^{-n}.$$

Thus (3.5) and (3.6) are satisfied and Lemma 6 is proved.

We now prove the required separation property for the segments q_k, which is the analogue of Lemma 1.

LEMMA 7. *Let q_k be the segments defined in Lemma 6 and write*

$$q_k' = \bigcup_{j \neq k} q_j.$$

Then we have

$$\omega(z, q_k', H) < \tfrac{7}{12}, \qquad z \in q_k.$$

We note that

$$\omega(z, l_k)+\omega(z, l_k')\leq\tfrac{3}{2}, \qquad z\in H. \tag{3.7}$$

For if z lies on l_k, the inequality holds in view of (3.2) and if z lies on l_j, where $j\neq k$, we have

$$\omega(z, l_k)\leq\omega(z, l_j')\leq\tfrac{1}{2}$$

so that (3.7) still holds. Also the left-hand side of (3.3) is zero on the boundary of H and is harmonic in H except on the l_j. Thus (3.7) follows from the maximum principle.

Using (3.7) and (3.6), we deduce that

$$\omega(z, l_k')\leq\tfrac{3}{2}-\tfrac{18}{19}=\tfrac{21}{38} \quad \text{on} \quad q_k. \tag{3.8}$$

Further we have on q_j, where $j\neq k$, from (3.6)

$$1=\omega(z, q_k')<\tfrac{19}{18}\omega(z, l_j)\leq\tfrac{19}{18}\omega(z, l_k').$$

Thus

$$\omega(z, q_k')<\tfrac{19}{18}\omega(z, l_k') \quad \text{for} \quad z\in q_k' \tag{3.9}$$

and (3.9) trivially holds on l_k' and on the boundary of H. Thus (3.9) holds in H. Combining (3.8) and (3.9), we deduce on q_k

$$\omega(z, q_k')<\tfrac{19}{18}\cdot\tfrac{21}{38}=\tfrac{7}{12}$$

and this proves Lemma 7.

Let S be a collection of distinct dyadic intervals I on the real axis. Suppose that a_1, a_2, a_3 are positive numbers such that $a_1\geq 1$, $0<a_2\leq 1$, $0<a_3\leq 1$. If I is an interval $|x-x_0|<\frac{1}{2}|I|$, we write aI for the interval $|x-x_0|<\frac{1}{2}a|I|$. We shall say that S satisfies the hypothesis $P(a_1, a_2, a_3)$ if for every $I\in S$ there exists a set $e(I)$ such that

$$e(I)\subseteq a_1 I \tag{3.10}$$

$$|e(I)|\geq a_1 a_2 |I| \tag{3.11}$$

and if I' is any interval of S such that $|I'| \leq a_3 |I|$ we have

$$e(I) \cap I' = \phi. \tag{3.12}$$

We shall prove in Lemma 9 that such a collection S necessarily has finite total length. Thus in order to prove that $\sum \delta(l_k)$ is finite, it is sufficient, in view of Lemma 6, to prove that the projections J_k of the q_k onto the real axis satisfy $P(a_1, a_2, a_3)$ for suitable constants a_1, a_2, a_3. We proceed to deduce this result from Lemma 7.

LEMMA 8. *The projections J_k of the dyadic intervals q_k are all distinct and satisfy* $P(12{,}000, \frac{1}{120}, \frac{1}{100})$.

It follows at once from Lemma 7 that $q_k \cap q_k' = \phi$ so that the different q_k are disjoint. Thus their projections are not identical. Suppose that $q = q_k$, and let $z_0 = x_0 + 600 \cdot 2^{-n} i$ be the midpoint of q. Let J be the projection of q on the real axis and let E be the union of all those q_j whose length is less than $10^{-2} \cdot 2^{-n}$. Then it follows from Lemma that

$$\omega(z_0, E) \leq \omega(z_0, q_k') < \tfrac{7}{12}.$$

Also by our construction E lies in $0 > y > 6 \cdot 2^{-n}$. Thus if E^* is the projection of E on the real axis, we deduce from Lemma 3 and conformal invariance that

$$\omega(z_0, E^*) \leq \tfrac{3}{2}\omega(z_0, E) < \tfrac{7}{8}.$$

It now follows from Lemma 4 that if I is the interval $12{,}000\, J$, i.e. $|x - x_0| < 6000 \cdot 2^{-n}$, then

$$|I - E^*| \geq \tfrac{1}{120} |I| = 12{,}000 \cdot \tfrac{1}{120} |J|.$$

Letting $e(J) = I \setminus E^*$, we conclude Lemma 8.

We can now obtain a bound for the sum of the diameters of the l_k in R_0 by proving

LEMMA 9. *If the collection S of dyadic intervals I satisfies $P(a_1, a_2, a_3)$ then*

$$\sum_S |I| \leq K(a_1, a_2, a_3) \leq 288 \frac{a_1}{a_2} \left(5 + \log \frac{a_1}{a_2 a_3}\right). \tag{3.13}$$

Thus if l_k are the level curves in R_0, we have

$$\sum \delta(l_k) \leq 10^5 K(12{,}000, \tfrac{1}{120}, \tfrac{1}{100}) < 10^{15}. \tag{3.14}$$

If I is a dyadic interval, then $I = J(m, n)$ for some m and n as in (3.4); we call $n = n(I)$ the index of I. Thus $|I|$ decreases with increasing $n(I)$.

4. Proof of Lemma 9

We first assume that if $J(m, n)$ and $J(m', n')$ are intervals of S then

$$m = m' (\text{mod } k_1) \quad \text{and} \quad n = n' (\text{mod } k_2), \tag{4.1}$$

where k_1, k_2 are the least integers satisfying respectively

$$k_1 \geq 3a_1 \tag{4.2}$$

and

$$2^{k_2} \geq \frac{1}{a_3} + \frac{48a_1}{a_2}. \tag{4.3}$$

If S does not satisfy (4.1) we shall divide S into at most $k_1 k_2$ subclasses each of which satisfies (4.1) and sum over each subclass separately.

Let F_ν be the subset of all points of $(0, 1)$ which are covered by at least ν distinct intervals of S. Evidently F_ν is the union of a finite or countable set of intervals in S. Also $F_{\nu+1} \subseteq F_\nu$. Let I_0 be a component of F_ν. We proceed to show that

$$|I_0 \cap F_{\nu+2}| \leq \left(1 - \frac{a_2}{24}\right) |I_0|. \tag{4.4}$$

To see this, suppose that $n(I_0) = n_0$ and let n_1 be the least index greater than n_0 of intervals of S which meet $\frac{1}{2}I_0$. We call V_1 the class of all such intervals of index n_1. We note that by the first relation in (4.1) the intervals $3a_1 I_1$, where $I_1 \in V_1$, are disjoint. We now define

$$H_1 = \tfrac{1}{2}I_0 \Big\backslash \bigcup_{V_1} 3a_1 I_1.$$

More generally if the class V_k of intervals I_k with index n_k has been constructed we write

$$H_k = \tfrac{1}{2}I_0 \setminus \left(\bigcup_{V_1} 3a_1 I_1 \cup \bigcup_{V_2} 3a_1 I_2 \cup \cdots \cup \bigcup_{V_k} 3a_1 I_k \right) \tag{4.5}$$

and define V_{k+1} to be the union of all intervals of S, meeting H_k and having least index $n_{k+1} > n_0$. By our construction $n_{k+1} > n_k$.

This process continues indefinitely or stops, either because H_k is empty or because H_k meets no intervals of S with index greater than n_0. We then define

$$H = \bigcap H_k, \qquad V = \bigcup V_k,$$

where intersection and union are taken over all k for which H_k is defined. We now distinguish two cases.

(i) Suppose that $|H| \geq \frac{1}{4}|I_0|$.

We note that no point of H meets any interval I of S having index greater than n_0. For suppose contrary to this that I is such an interval of least index n. Let k be the largest integer for which $n_k < n$ and H_k is defined. Then I meets H_k and so belongs to the class V_{k+1}, and thus I is disjoint from H_{k+1}, contrary to hypothesis. Thus in case (i) H does not meet $F_{\nu+1}$ and

$$|I_0 \cap F_{\nu+1}| \leq |I_0 \setminus H| \leq \tfrac{3}{4}|I_0|.$$

Since $F_{\nu+2} \subseteq F_{\nu+1}$, we deduce (4.4) in this case.

(ii) Suppose next that $|H| < \frac{1}{4}|I_0|$.

In this case we see that

$$\left| \bigcup_V (3a_1 I) \right| > \tfrac{1}{4}|I_0|.$$

Thus

$$\sum_V |I| \geq \frac{1}{12a_1}|I_0|. \tag{4.6}$$

Suppose that I is an interval in V_k. Then by hypothesis $a_1 I$ contains a set $e(I)$ not meeting any interval of S having length less than $a_3|I|$, i.e., no interval of S having index greater than n_k in view of (4.1), (4.3) and the fact that I has index n_k. We note that the sets $e(I)$ corresponding to distinct intervals I are disjoint,

since the intervals $a_1 I$ are disjoint by (4.2), (4.5) and the fact $a_1 \geq 1$. Thus using (3.11) and (4.6) we deduce

$$\left|\bigcup_V e(I)\right| = \sum_V |e(I)| \geq a_1 a_2 \sum_V |I| \geq \frac{a_2}{12}|I_0|. \tag{4.7}$$

We have just seen that $e(I)$ corresponding to $I \in V_k$ meets no interval of S having index greater than n_k. Next suppose that $e(I)$ meets an interval $I' \in S$ having index n, $n_0 < n < n_k$. It follows from (4.1) that $n \leq n_k - k_2$, so that

$$|I'| \geq 2^{k_2}|I|.$$

On the other hand, I' cannot contain I since otherwise I' would meet H_{k-1} and have index less than n_k, contrary to hypothesis. Thus $e(I) \cap I'$ must lie in one of two subintervals of I', which adjoin the end points of I' and have total length

$$(a_1 + 1)|I| \leq 2a_1 \cdot 2^{-k_2}|I'|.$$

Thus we see that the total length of the intersections of the $e(I)$ with intervals I' of S having index n, such that $n_0 < n(I') < n(I)$ is at most

$$\sum a_1 2^{1-k_2}|I'| \leq a_1 2^{1-k_2}|I_0|, \tag{4.8}$$

where $\sum$ is taken over all the maximal intervals I' in S of index greater than n_0 and containing some point $x \in I_0$. For these intervals I' are disjoint and lie in I_0 and so have total length at most $|I_0|$. Thus if $e'(I)$ is the subset of $e(I)$ meeting no intervals of S of index greater than n_0 and different from n_k, we deduce, using (4.7) and (4.8) that

$$\begin{aligned} |\bigcup e'(I)| &\geq |\bigcup e(I)| - a_1 \cdot 2^{1-k_2}|I_0| \\ &\geq \left(\frac{a_2}{12} - a_1 2^{1-k_2}\right)|I_0|. \end{aligned} \tag{4.9}$$

Using (4.3) and (4.9) we deduce that

$$|\bigcup e'(I)| \geq \frac{a_2}{24}|I_0|.$$

Also a point x in $e'(I)$ lies in no interval of S having index $n > n_0$, $n \neq n(I)$. Thus $x \notin F_{\nu+2}$. Hence in this case (4.4) holds, so that (4.4) is true in all cases.

We write

$$\theta = 1 - \frac{a_2}{24},$$

and deduce by induction that since $|F_1| \leq 1$, we have

$$|F_{2\nu}| \leq |F_{2\nu-1}| \leq \theta^{\nu-1}, \qquad \nu \geq 1.$$

Again

$$\sum_S |I| = \sum_1^\infty |F_\nu| = \sum_1^\infty (|F_{2\nu}| + |F_{2\nu-1}|) \leq 2 \sum_1^\infty \theta^{\nu-1} = \frac{2}{1-\theta}.$$

Thus with the hypotheses of Lemma 9, together with (4.1) we have

$$\sum_S |I| \leq \frac{2}{1-\theta} = \frac{48}{a_2}. \tag{4.10}$$

Also in the general case, when (4.1) is not satisfied we can divide the intervals of S into at most $k_1 k_2$ subclasses each of which satisfies (4.10) so that we always have

$$\sum_S |I| \leq 48 k_1 k_2 / a_2 = K(a_1, a_2, a_3).$$

Using (4.2) and (4.3) we obtain

$$k_1 \leq 3a_1 + 1 \leq 4a_1,$$
$$k_2 \leq \frac{1}{\log 2} \log\left(\frac{50a_1}{a_2 a_3}\right) + 1 < \frac{3}{2}\left(\log\left(\frac{a_1}{a_2 a_3}\right) + 5\right).$$

We deduce that

$$K \leq \frac{3}{2} \cdot 4 \cdot 48 \frac{a_1}{a_2}\left(5 + \log\left(\frac{a_1}{a_2 a_3}\right)\right) = 288 \frac{a_1}{a_2}\left(5 + \log \frac{a_1}{a_2 a_3}\right).$$

This proves (3.13).

Next it follows from Lemma 8, that if the class S consists of the projections of the q_k, we may take

$$a_1 = 12{,}000, \qquad a_2 = \tfrac{1}{120}, \qquad a_3 = \tfrac{1}{100}.$$

Thus

$$\sum |q_k| \leq 288 \cdot 12{,}000 \cdot 120(5+\log(1.44\times 10^8)) < 10^{10}.$$

Using (3.5) we obtain

$$\sum \delta(l_k) < 10^{15}.$$

This completes the proof of Lemma 9.

5. Proof of (1.2)

Let γ_k now be the level curves of a univalent function in Δ. We divide γ_k into 3 subclasses. Consider first those γ_k, which lie entirely in

$$\Delta_1 = \Delta \cap (x > -\tfrac{1}{10}). \tag{5.1}$$

We consider the transformation

$$w = u + iv = \frac{i}{5}\left(\frac{1-z}{1+z}\right) + \frac{1}{2} = \frac{i}{5}\left(\frac{(1-|z|^2)-2iy}{|1+z|^2}\right) + \frac{1}{2}.$$

Clearly Δ corresponds to the upper half-plane H in the w plane and the subset (5.1) maps into the unit square R_0 given by (3.1) with (u, v) instead of (x, y). The level curves γ_k in Δ_1 correspond to level curves l_k in R_0. Also

$$\left|\frac{dw}{dz}\right| = \frac{2}{5|1+z|^2} > \tfrac{1}{10} \text{ in } \Delta_1,$$

and so

$$\left|\frac{dz}{dw}\right| \leq 10$$

in the image of Δ_1. We deduce that

$$\delta(\gamma_k) \leq 10\delta(l_k).$$

Using Lemma 9, we obtain

$$\sum\nolimits_1 \delta(\gamma_k) \leq 10^{16}. \tag{5.2}$$

Similarly if $\Delta_2 = \Delta \cap (x < \frac{1}{10})$ and $\sum_2$ denotes the sum over those γ_k which lie entirely in Δ_2 we obtain

$$\sum\nolimits_2 \delta(\gamma_k) \leq 10^{16}. \tag{5.3}$$

Consider now the remaining level curves γ_k. Each of them must contain an arc $\tilde{\gamma}_k$ with end points on $x = \mp\frac{1}{10}$ in Δ and hence meets the imaginary axis at finitely many points iy_k with $|y_k| < 1$. We choose the least such y_k, and enumerate the γ_k in order of increasing y_k. We proceed to prove that if $y_k \leq 0$ then

$$y_{k+1} - y_k > \tfrac{1}{70}. \tag{5.4}$$

Suppose that (5.4) is false for some k. We note that $\tilde{\gamma}_k$ separates iy_{k+1} from the arc.

$$-\tfrac{1}{10} < x < \tfrac{1}{10}, \qquad y = -\sqrt{1-x^2} \tag{5.5}$$

in $-\frac{1}{10} < x < \frac{1}{10}$. Let Δ_k be the disk

$$|z - iy_k| < \tfrac{1}{10}$$

and let Δ_k' be that component of $\Delta_k \setminus \tilde{\gamma}_k$ which contains iy_{k+1}. Then Δ_k' cannot contain any point of the arc (5.5). Thus since $y_k \leq 0$, we deduce that $\Delta_k' \subseteq \Delta$. Hence if $\tilde{\tilde{\gamma}}_k = \tilde{\gamma}_k \cap \Delta_k$, we deduce that

$$\omega(iy_{k+1}, \tilde{\tilde{\gamma}}_k, \Delta_k') \leq \omega(iy_{k+1}, \tilde{\tilde{\gamma}}_k, \Delta) \leq \omega(iy_{k+1}, \gamma_k, \Delta) \leq \tfrac{1}{2}$$

in view of Lemma 1. On the other hand $\tilde{\tilde{\gamma}}_k$ contains an arc η_k joining $z = iy_k$ to the boundary of Δ_k and so we deduce that

$$\omega(iy_{k+1}, \eta_k, \Delta_k) \leq \omega(iy_{k+1}, \tilde{\tilde{\gamma}}_k, \Delta_k) = \omega(iy_{k+1}, \tilde{\tilde{\gamma}}_k, \Delta_k') \leq \tfrac{1}{2}.$$

Now we apply Lemma 2 and deduce that

$$y_{k+1} - y_k \geq \tfrac{1}{10} \tan^2 \frac{\pi}{8} > \tfrac{1}{70}.$$

Thus (5.4) is true after all and less than 70 different γ_k can meet the negative imaginary axis. Thus there are at most $140\gamma_k$ in our remaining group and if $\sum_3$ denotes the sum over these, we have

$$\sum_3 \delta(\gamma_k) < 300.$$

On combining this with (5.2) and (5.3) we deduce (1.2).

6. Preliminary reductions

We now embark on the proof of (1.3). We confine ourselves to the following special case to which the general result can easily be reduced. We assume that Ω is the interior of an analytic Jordan curve Γ and that I is a segment $[b_1, b_2]$ of the real axis in Ω whose endpoints b_1, b_2 lie on Γ. We denote by γ the image of I under the conformal map

$$z = F(w) = f^{-1}(w)$$

of Ω onto Δ and shall show that $\gamma = \gamma_k$ satisfies (1.3). In this part we work with the geometry of Ω rather than that in H or Δ. If w is a point of Ω we write

$$d(w) = \inf_{\zeta \in \Gamma} |w - \zeta| \tag{6.1}$$

for the distance from w to Γ. We start by constructing a function $\phi(u)$ which is comparable to $d(u)$ on I but behaves in a smooth manner. We shall then dissect the interval I into a sequence of intervals $I_{j,k}$, and with each $I_{j,k}$ we associate an arc $\Gamma_{j,k}$ of Γ, such that length of the image of $I_{j,k}$ by $F(w)$ is comparable with that of $\Gamma_{j,k}$. The $\Gamma_{j,k}$ will be disjoint and so their images have total length at most 2π and from this (1.3) will follow.

We set $w = u + iv$.

By our construction the line $u = c$ meets Γ for $b_1 \leq c \leq b_2$ and we write

$$\phi_0(u) = \inf\{|v| \,|\, u + iv \in \Gamma\}, \qquad b_1 \leq u \leq b_2. \tag{6.2}$$

Further we define

$$\phi(u) = \inf_{u_1 \in I} \{\phi_0(u_1) + \tfrac{1}{12}|u_1 - u|\}. \tag{6.3}$$

Clearly $\phi(u) \geq 0$, with equality only at the endpoints b_1, b_2 of I. A point $u \in I$ will be called a *spike-point* if

$$\phi(u) = \phi_0(u).$$

Evidently the endpoints b_1, b_2 are spike points.

In order to establish our results we proceed to subdivide I into intervals bounded by spike points. We prove first

LEMMA 10. *If u is a point of I and $d(u)$, $\phi(u)$ are defined as above then*

$$d(u) \geq \tfrac{24}{25}\phi(u).$$

Suppose that $w_1 = u_1 + iv_1$ is a point of Γ such that

$$|w_1 - u| = d(u).$$

Then

$$\begin{aligned}\phi(u) &\leq \phi_0(u_1) + \tfrac{1}{12}|u_1 - u| \leq |v_1| + \tfrac{1}{12}|u_1 - u| \\ &\leq [1 + (\tfrac{1}{12})^2)^{1/2}\,|w_1 - u| \leq \tfrac{25}{24}d(u).\end{aligned}$$

This proves Lemma 10.

LEMMA 11. *We have for $u_1, u_2 \in I$*

$$|\phi(u_1) - \phi(u_2)| \leq \tfrac{1}{12}|u - u_2|.$$

Suppose that $u_3 \in I$. Then

$$\phi_0(u_3) + \tfrac{1}{12}|u_1 - u_3| \leq \phi_0(u_3) + \tfrac{1}{12}|u_2 - u_3| + \tfrac{1}{12}|u_1 - u_2|.$$

Taking lower bounds for varying u_3 we obtain

$$\phi(u_1) \leq \phi(u_2) + \tfrac{1}{12}|u_1 - u_2|.$$

Interchanging u_1 and u_2 we obtain

$$\phi(u_2) \leq \phi(u_1) + \tfrac{1}{12}|u_1 - u_2|$$

and so we deduce Lemma 11.

Lemmas 10 and 11 are not true if we use ϕ_0 in place of ϕ.

LEMMA 12. *The set of spike points is a closed non-empty subset of I.*

Since b_1, b_2 are spike points, the set is certainly non-empty. Next we note that since Γ is closed the function ϕ_0 defined by (6.2) is lower semi-continuous. Also we see by Lemma 11 that $\phi(u)$ is continuous. Thus $h(u)=\phi_0(u)-\phi(u)$ is a lower semicontinuous nonnegative function. To see that $h(u)\geq 0$ we just set $u_1=u$ in (6.3). Hence the set where $h(u)\leq 0$ is closed and this is the set of spike points.

It follows from Lemma 12, that the complement in I of the set of spike points consists of a finite or countable set of open intervals J. In each of these intervals $\phi(u)$ has a particularly simple form.

LEMMA 13. *Suppose that a, a' are spike points in I such that $a<a'$, and that the interval (a, a') contains no spike points. Then for $a<u<a'$,*

$$\phi(u)=\min\{\phi_0(a)+\tfrac{1}{12}(u-a),\ \phi_0(a')+\tfrac{1}{12}(a'-u)\}.$$

Since a is a spike point we have for $a_1<a$, $a_1\in I$,

$$\phi_0(a_1)+\tfrac{1}{12}(a-a_1)\geq\phi_0(a).$$

Thus for $a<u<a'$, we deduce that

$$\begin{aligned}\phi_0(a_1)+\tfrac{1}{12}(u-a_1)&=\phi_0(a)+\tfrac{1}{12}(u-a)+\phi_0(a_1)-\phi_0(a)+\tfrac{1}{12}(a-a_1)\\&\geq\phi_0(a)+\tfrac{1}{12}(u-a).\end{aligned}\tag{6.4}$$

Similarly if $a_1>a'$, we deduce that

$$\phi_0(a_1)+\tfrac{1}{12}(a_1-u)\geq\phi_0(a')+\tfrac{1}{12}(a'-u).\tag{6.5}$$

Next we note that in the definition (6.3) of $\phi(u)$ we may allow u_1 to range only over spike points in I. For since $\phi_0(u)$ is lower semicontinuous the infimum in (6.3) is attained for some u_1 in I. If u_1 is not a spike point we can find u_2 such that

$$\phi_0(u_2)+\tfrac{1}{12}|u_2-u_1|<\phi_0(u_1).$$

Thus

$$\begin{aligned}\phi_0(u_2)+\tfrac{1}{12}|u-u_2|&\leq\phi_0(u_2)+\tfrac{1}{12}|u_2-u_1|+\tfrac{1}{12}|u-u_1|\\&<\phi_0(u_1)+\tfrac{1}{12}|u-u_1|=\phi(u)\end{aligned}$$

and this contradicts the definition of $\phi(u)$. Hence if $a<u<a'$ there exists a spike point u_1, such that

$$\phi(u)=\phi_0(u_1)+\tfrac{1}{12}|u_1-u|. \tag{6.6}$$

In view of (6.4) and (6.5) we may suppose that $a\leq u_1\leq a'$, so that $u_1=a$ or $u_1=a'$, since (a, a') contains no spike points. Using Lemma 11, we see that u_1 is that one of a, a' which gives the smaller value of $\phi(u)$ in (6.6).

7. A dissection of the interval I

Suppose first that the interval I contains no spike point other than the end points b_1, b_2. In this case we write $a_1=b_1$, $a_2=b_2$ and deduce from Lemma 13 that

$$\phi(u)=\tfrac{1}{12}\min(a-a_1, a_2-a), \qquad a_1<a<a_2. \tag{7.1}$$

Thus we deduce from Lemma 10 that in this case

$$\tfrac{2}{25}\min(a-a_1, a_2-a)\leq d(a), \qquad a_1<a<a_2. \tag{7.2}$$

Suppose next that I contains at least one spike point a_0, such that $b_1<a_0<b_2$. Having chosen a_0 we define other spike points a_j inductively as follows. If a_j has been defined, $j\geq 0$, we define a_{j+1} to be the smallest spike point such that

$$a_{j+1}\geq a_j+6\phi(a_j). \tag{7.3}$$

We deduce from (6.3)

$$\phi_0(b_2)+\tfrac{1}{12}|b_2-a_j|\geq\phi(a_j)$$

i.e.

$$b_2-a_j\geq 12\phi(a_j),$$

so that $a_{j+1}\leq b_2$. Thus either at some stage $a_{j+1}=b_2$, in which case we stop the procedure, or else the a_j are defined for all positive j and

$$a_j\to b_2, \quad \text{as} \quad j\to+\infty.$$

Similarly if, for some nonpositive j, a_j has been defined we define a_{j-1} to be the largest spike point such that

$$a_{j-1} \leq a_j - 6\phi(a_j). \tag{7.4}$$

If $a_j > b_1$, we deduce again that $a_{j-1} \geq b_1$.

The relevant properties of our subdivision are given in

LEMMA 14. *The interval I can be divided into a finite or countable set of subintervals $[a_j, a_{j+1}]$, where the a_j are spike points with the following properties*

$$a_{j+1} - a_j \geq 4 \max\{\phi(a_j), \phi(a_{j+1})\}. \tag{7.5}$$

Further, if $d(a)$ denotes the distance of a from the boundary Γ of D, we have for $a_j < a < a_{j+1}$

$$d(a) \geq \tfrac{2}{5} \min\{\max[\phi(a_j), \tfrac{1}{5}(a - a_j)], \max[\phi(a_{j+1}), \tfrac{1}{5}(a_{j+1} - a)]\}. \tag{7.6}$$

If $a_j = b_1$, $a_{j+1} = b_2$, (7.5) is trivial and (7.6) follows from (7.2). Thus we may assume that a_0 is a spike point in I and that the remaining a_j are defined by (7.3) and (7.4). We concentrate on (7.3) and $j \geq 0$ for definiteness. The case $j < 0$ is similar.

We first prove (7.5). Suppose first that $\phi(a_{j+1}) \leq \frac{3}{2}\phi(a_j)$. Then (7.5) follows from (7.3). On the other hand if

$$\phi(a_j) < \tfrac{2}{3}\phi(a_{j+1}) \quad \text{so that} \quad |\phi(a_{j+1}) - \phi(a_j)| > \tfrac{1}{3}\phi(a_{j+1})$$

we deduce from Lemma 11 that

$$|a_j - a_{j+1}| \geq 12\,|\phi(a_{j+1}) - \phi(a_j)| > 4\phi(a_{j+1}).$$

Thus (7.5) holds in all cases.

Next we prove (7.6). Suppose first that

$$a_j < a \leq a_j + 6\phi(a_j).$$

Then Lemma 11 shows that

$$|\phi(a) - \phi(a_j)| \leq \tfrac{1}{12}(a - a_j) \leq \tfrac{1}{2}\phi(a_j).$$

Thus in this case

$$\phi(a) \geq \tfrac{1}{2}\phi(a_j) \geq \tfrac{1}{12}(a - a_j). \tag{7.7}$$

Using Lemma 10, we deduce (7.6).

Suppose next that $a_j + 6\phi(a_j) < a < a_{j+1}$. In this case it follows from (7.3) that if b is the largest spike point such that $b \leq a$, we have

$$a_j \leq b < a_j + 6\phi(a_j).$$

Also the interval (b, a_{j+1}) contains no spike point and so by Lemma 13, we have

$$\phi(a) = \min\{\phi(b) + \tfrac{1}{12}(a - b),\ \phi(a_{j+1}) + \tfrac{1}{12}(a_{j+1} - a)\}. \tag{7.8}$$

In view of (7.7) applied to b instead of a, we have

$$\phi(b) \geq \tfrac{1}{2}\phi(a_j) \geq \tfrac{1}{12}(b - a_j).$$

Thus

$$\phi(b) + \tfrac{1}{12}(a - b) \geq \tfrac{1}{2}\phi(a_j) + \tfrac{1}{12}(a - b) \geq \tfrac{1}{12}(a - a_j).$$

Hence (7.8) yields in this case

$$\phi(a) \geq \min\{\max[\tfrac{1}{2}\phi(a_j),\ \tfrac{1}{12}(a - a_j)],\ \max[\tfrac{1}{2}\phi(a_{j+1}),\ \tfrac{1}{12}(a_{j+1} - a)]\}.$$

Because of (7.7), this inequality also holds for $a_j < a < a_{j+1}$. Using Lemma 10, we deduce (7.6). This completes the proof of Lemma 14.

Having obtained the points a_j satisfying the conditions of Lemma 14, we now proceed to a further subdivision as follows. Let (a_j, a_{j+1}) be one of the intervals defined in Lemma 14. We write

$$a_{j,0} = \tfrac{1}{2}(a_j + a_{j+1}). \tag{7.9}$$

We then define $a_{j,k}$ for positive k inductively by

$$a_{j,k+1} = \tfrac{1}{2}(a_{j+1} + a_{j,k}), \tag{7.10}$$

provided that $a_{j,k+1}$, so defined, satisfies

$$a_{j,k+1} \leq a_{j+1} - \phi(a_{j+1}). \tag{7.11}$$

Otherwise we set this value of $k+1$ equal to k_2 and define

$$a_{j,k_2} = a_{j+1}.$$

If $a_{j+1} = b_2$, so that $\phi(a_{j+1}) = 0$, the process continues indefinitely and we set $k_2 = \infty$. Thus (7.10) defines $a_{j,k+1}$ for $0 < k+1 < k_2$.

Similarly we define $a_{j,k}$ for negative k inductively by

$$a_{j,k-1} = \tfrac{1}{2}(a_j + a_{j,k}). \tag{7.12}$$

The process continues as long as $a_{j,k-1}$ so defined satisfies

$$a_{j,k-1} \geq a_j + \phi(a_j).$$

Otherwise we set $k-1 = k_1$, and define

$$a_{j,k_1} = a_j. \tag{7.13}$$

In this way (7.9)–(7.12) define $a_{j,k}$ for $k_1 < k < k_2$. We deduce from (7.5) that

$$-\infty \leq k_1 \leq -2 \tag{7.14}$$

and

$$2 \leq k_2 \leq +\infty \tag{7.15}$$

We define

$$I_{j,k} = (a_{j,k}, a_{j,k+1}), \qquad k_1 + 1 < k < k_2 - 2. \tag{7.16}$$

If k_1 is finite we set

$$I_{j,k} = (a_{j,k-1}, a_{j,k+1}), \qquad k = k_1 + 1, \tag{7.17}$$

and if k_2 is finite we set

$$I_{j,k} = (a_{j,k}, a_{j,k+2}), \qquad k = k_2 - 2. \tag{7.18}$$

Thus $I_{j,k}$ is defined for $k_1 < k < k_2 - 1$, i.e. certainly for $k = -1, 0$. Also the $I_{j,k}$ cover I apart from isolated points.

8. An association of arcs $\Gamma_{j,k}$ of Γ with $I_{j,k}$

Suppose that $k_1 < k \leq 0$. We construct an arc of the circle

$$|w - a_j| = |a_{j,k} - a_j| \tag{8.1}$$

starting from the point $a_{j,k}$. If a_j is the left end point of I we start anticlockwise into the upper half plane. Otherwise $\phi_0(a_j) > 0$ and one of the two points

$$a_j \pm i\phi_0(a_j) \tag{8.2}$$

lies on Γ. If $a_j + i\phi_0(a_j)$ lies on Γ, we start the arc of (8.1) by moving in the anticlockwise sense into the upper half plane; otherwise we move in the clockwise sense into the lower half plane. In either case we continue along the circle (8.1) until we first meet a point of Γ which we denote by $b_{j,k}$. Since Γ contains points inside or on the circle (8.1) namely one of the points (8.2), and since $a_{j,k}$ lies inside Γ the point $b_{j,k}$ certainly exists.

If $k_1 < k < 0$, the points $b_{j,k}$, $b_{j,k+1}$ determine two arcs of Γ. We choose that arc $\Gamma_{j,k}$ which we reach first when going along the circles

$$|w - a_j| = r, \qquad |a_{j,k} - a_j| < r < |a_{j,k+1} - a_j|$$

from the point $a_j + r$ in the anticlockwise or clockwise sense according as $a_j + i\phi_0(a_j)$ does or does not lie on Γ. Thus we have associated with each interval $I_{j,k}$ defined by (7.15) or (7.16) an arc $\Gamma_{j,k}$ of Γ if $k_1 < k < 0$. It follows from (7.13) that at least one such k exists. The corresponding intervals $I_{j,k}$ cover the interval $[a_j, a_{j,0}]$.

We proceed in an exactly analogous manner with the intervals $I_{j,k}$, $0 \leq k < k_2 - 1$. We go along the circle

$$|w - a_{j+1}| = |a_{j,k} - a_{j+1}| \tag{8.3}$$

where $0 \leq k < k_2$ starting at the point $a_{j,k}$ until we meet Γ at $b'_{j,k}$. We then associate with the interval $I_{j,k}$, $0 \leq k < k_2 - 1$, one of the arcs $[b'_{j,k}, b'_{j,k+1}]$ of Γ, which we denote by $\Gamma_{j,k}$. We move along circles $|w - a_{j+1}| = r$, $|a_{j,k+1} - a_{j+1}| < r < |a_{j,k} - a_{j+1}|$ in the clockwise sense if $a_{j+1} + i\phi(a_{j+1})$ lies on Γ and

the anticlockwise sense otherwise and $\Gamma_{j,k}$ contains the first point of Γ we meet in this way. In general $\bigcup \Gamma_{j,k}$ is a proper subset of Γ.

It follows from the construction that our interval I has been subdivided into a finite or countable set of subintervals $I_{j,k}$ which are associated with arcs $\Gamma_{j,k}$ of Γ, and no arc of Γ is associated with more than one distinct interval of I. We complete this section by proving that in this association distinct arcs are disjoint except for endpoints. We show in the next section that the length of the image of $I_{j,k}$ is not much greater than that of $\Gamma_{j,k}$. From these two facts (1.3) will follow.

LEMMA 15. *The arcs $\Gamma_{j,k}$ defined as above are pairwise disjoint except for endpoints.*

Let $\beta_{j,k}$ denote the circular arc from $a_{j,k}$ to $b_{j,k}$, or $b'_{j,k}$ defined as above. We show that distinct arcs $\beta_{j,k}$ are disjoint except for endpoints.

Let β be one of these circular arcs starting at $a = a_{j,k}$. Then β can contain a semicircle s starting at a only if the other endpoint a' of s lies outside I. For if a' lies in I, then the segment aa' together with s constitutes a closed Jordan curve c in Ω. If a_j is the midpoint of aa', then our construction ensures that one of the points $a_j \mp i\phi(a_j)$ lies on Γ and so is in the same halfplane as c and so by (7.11) inside c, which is impossible since Ω is simply connected.

Suppose now that β_1, β_2 are two distinct circular arcs starting at P_1, P_2 and first meeting at a point P of Ω. Consider the curve γ formed by going along β_1 from P_1 to P then along β_2 from P to P_2 and returning along the segment P_1, P_2. By construction γ lies in Ω and the arcs P_1P, PP_2 have only endpoints in common. Neither β_1 not β_2 can meet the segment P_1P_2 again since β_1, β_2 have no points in I other than P_1, P_2 respectively. Thus γ is a Jordan curve in Ω.

We shall show that γ contains a point of Γ in its interior and this leads to a contradiction since Ω is simply connected. Suppose that P_1 lies to the left of P_2. It is not possible for the centres of both circular arcs to lie outside the segment P_1P_2, for if the centres are on the same side, the circles are concentric and distinct[1] and can certainly not meet, and if they are on opposite sides, P_1P_2 is the shortest distance between the circles. Suppose then that at least one centre, say a_j, the centre of the arc P_1P, lies on P_1P_2. Suppose also that $z_j = a_j + i\phi(a_j)$ lies on Γ. We distinguish a number of cases.

(i) Suppose first that neither β_1 nor β_2 contains a semicircle. Then β_1, β_2 both lie in the upper half plane. We show that in this case the segment (a_j, z_j) lies inside γ. Suppose first that P_2 lies inside the circle s_1, of which β_1 is an arc. Then the

[1] If $P = a_j$ is the centre of the arc at $P_2 = a_{j,k}$ and P lies to the left of P_1, then, by (8.1), $P_1 = a_{j,k'}$, where $k' < k < 0$.

centre of the circle having β_2 as an arc must lie to the right of a_j and so to the right of P_2 since otherwise β_2 would lie inside s_1. Hence the segment $[a_j, z_j]$ does not meet β_2 and so z_j lies inside γ.

Thus P_2 must lie outside s_1 and so does the whole arc β_2, since if β_2 went inside s_1, β_2 would contain a semi-circle. Hence in this case the interior of γ includes all points in the upper half plane and inside s_1 and so in particular z_j because of (7.11).

(ii) Suppose that β_1 contains a semi-circle, but β_2 does not. Then β_2 lies entirely in the lower half-plane and again the interior of γ contains all points in the upper half-plane and inside s_1, and in particular z_j.

(iii) Suppose that β_2 contains a semi-circle but that β_1 does not. If the centre a_k of the circle s_2 containing β_2 lies to the right of P_2, then the whole of s_2 lies to the right of P_2, and so cannot meet the segment $[a_j, z_j]$. Thus in this case z_j again lies inside γ. If on the other hand a_k lies on P_1, P_2 then we have the case (ii) with P_1, P_2 interchanged. Finally a_k cannot lie to the left of P_1, since otherwise β_1, β_2 would be arcs of concentric circles which cannot meet. For all circular arcs starting from a point between a_k and P_2 have centre a_k.

(iv) If β_1, β_2 both contain semicircles, they must reduce to semi-circles, since otherwise they would have two distinct points of intersection. In this case β_1, β_2 meet at P, which is to the right of P_2 and again the point z_j lies inside γ.

Thus in all cases β_1, β_2 can have at most end points in common, since otherwise γ contains a point of Γ in its interior, which contradicts the fact that γ lies inside Γ.

Suppose now that $\Gamma_{j,k}$ is an arc corresponding to an interval $\mathrm{I}_{j,k}$. Let β_k, β_{k+1} be the arcs of the circles (8.1) if $k<0$, or (8.3) if $k\geq 0$, to the points $a_{j,k}$, $a_{j,k+1}$. Then β_k, β_{k+1}, and the interval $[a_{j,k}, a_{j,k+1}]$ determine a crosscut $\delta_{j,k}$ in Ω. In view of what we have just proved distinct crosscuts $\delta_{j,k}$ may have a common arc β_k or a common point $a_{j,k}$, but cannot cross each other. Thus if $D_{j,k}$ is the interior of the Jordan curve formed by $\delta_{j,k}$ and $\Gamma_{j,k}$ then two distinct domains $D_{j,k}$ are disjoint.

In fact otherwise one of these domains would lie inside the other. However our construction ensures that the interval I lies outside all the $D_{j,k}$, since none of the β_k meet I again. Hence points near $I_{j,k}$ inside $D_{j,k}$ are exterior to all the other domains $D_{j',k'}$. Thus $D_{j,k}$ cannot lie in $D_{j',k'}$. Now it follows that two distinct arcs $\Gamma_{j,k}$ are disjoint except for end points and this proves Lemma 15.

9. Images of intervals and associated arcs

Suppose that $I_{j,k}$ is an interval of the real axis and that $\Gamma_{j,k}$ is the associated arc of Γ. We assume for definiteness that $k<0$. If $k\geq 0$ the argument is similar. We

recall from (7.9) to (7.11) that

$$a_{j,k} = a_j + 2^{(k-1)}(a_{j+1} - a_j), \qquad k_1 < k \leq 0. \tag{9.1}$$

We denote by $\omega(w)$ the harmonic measure of $\Gamma_{j,k}$ with reference to the full open set Ω. We write

$$r = 2^{k-1/2}(a_{j+1} - a_j) \tag{9.2}$$

and prove

LEMMA 16. *If* $w_1 = a_j + 2^{\delta} r$, *where* $-\frac{7}{16} \leq \delta \leq \frac{7}{16}$, *then* $\omega(w_1) > \exp\{-30.9\}$.

Let β be the arc of the circle (8.1), joining $a_{j,k}$ to an endpoint $\beta_{j,k}$ of $\Gamma_{j,k}$. Let β' be the corresponding arc of (8.1) with $k+1$ instead of k, which joins $a_{j,k+1}$ to the other endpoint $\beta_{j,k+1}$ of $\Gamma_{j,k}$. Then β, $\Gamma_{j,k}$, β' and the segment $[a_{j,k}, a_{j,k+1}]$ form a subdomain $D_{j,k}$ of Ω. We now define a domain $D' = D'_{j,k}$ as follows. We define θ_0 by

$$\sin(\tfrac{1}{2}\theta_0) = \tfrac{1}{25}. \tag{9.3}$$

Then it follows from (7.6) of Lemma 14, that the sectorial region

$$2^{-1/2} r < |w - a_j| < 2^{1/2} r, \; |\arg(w - a_j)| < \theta_0 \tag{9.4}$$

does not meet Γ and so lies inside Ω. Next it follows from our construction that $\Gamma_{j,k}$ contains an arc $\eta_{j,k}$ joining the circles $|w - a_j| = 2^{\mp 1/2} r$ in the annulus

$$2^{-1/2} r < |w - a_j| < 2^{1/2} r. \tag{9.5}$$

We now define D' to be a subdomain of the annulus (9.5) determined by such an arc $\eta_{j,k}$ and one of the rays $\arg(w - a_j) = \mp\theta_0$ and having the arcs β, β' as part of its boundary. In other words if β, β' start off in the clockwise sense we choose the ray $\arg(w - a_j) = +\theta_0$ and otherwise the ray $\arg(w - a_j) = -\theta_0$. We continue along $|w - a_j| = 2^{\mp 1/2} r$ until we meet the first arc $\eta_{j,k}$.

We note that D' constructed as above lies in the annulus (9.5) and contains the region (9.4). Let $\omega'(w)$ be the harmonic measure of $\eta_{j,k}$ at w w.r.t. D', and let D_1 be that component of $D' \cap \Omega$ which contains the segment $(a_{j,k}, a_{j,k+1})$. Then

$$\omega(w_1) \geq \omega'(w_1), \qquad w_1 \in D_1. \tag{9.6}$$

To see this let ζ be any boundary point of D_1. If ζ lies on one of the circles $|w-a_j|=2^{\mp 1/2}r$ or on the ray $\arg(w-a_j)=\mp\theta_0$, we have

$$\omega'(\zeta)=0\leq\omega(\zeta)$$

Any other boundary point ζ of D_1 lies on $\Gamma_{j,k}$ so that

$$1=\omega(\zeta)\geq\omega'(\zeta).$$

Thus $\omega(w)-\omega'(w)\geq 0$ on the boundary of D_1 and so in the interior of D_1 by the maximum principle. This proves (9.6).

A further application of the maximum principle now shows that $\omega'(w_1)$ assumes its lower bound for variable $\eta_{j,k}$ in the (limiting) case, when D' reduces to the subdomain

$$-\theta_0<\arg(w-a_j)<2\pi-\theta_0, \qquad 2^{-1/2}r<|w-a_j|<2^{1/2}r \tag{9.7}$$

of the annulus (9.5) and $\eta_{j,k}$ to the arc $\arg(w-a_j)=2\pi-\theta_0$. It remains to estimate the corresponding harmonic measure. To do this we use the invariance of harmonic measure and set

$$s=\sigma+i\tau=\log(w-a_j). \tag{9.8}$$

We write

$$\sigma_0=\log r,$$

and note that the cut annulus (9.7) corresponds by (9.8) to the rectangle

$$\Delta_1:\sigma_0-\tfrac{1}{2}\log 2<\sigma<\sigma_0+\tfrac{1}{2}\log 2, \qquad -\theta_0<\tau<2\pi-\theta_0.$$

We have to estimate the harmonic measure w.r.t. Δ_1 of $\tau=2\pi-\theta_0$ at $s_1=\log w_1=\sigma_0+\delta\log 2$.

By the maximum principle this harmonic measure is greater than that of the two rays

$$\tau\geq 2\pi-\theta_o, \qquad \sigma=\sigma_0\mp\tfrac{1}{2}\log 2 \tag{9.9}$$

w.r.t. the half strip

$$\Delta_2:-\theta_0<\tau<+\infty, \qquad \sigma_0-\tfrac{1}{2}\log 2<\sigma<\sigma_0+\tfrac{1}{2}\log 2$$

at s_1. This latter harmonic measure is calculated explicitly as follows. We set

$$z = \exp\left\{\frac{\pi i(s + i\theta_0 - \sigma_0)}{\log 2}\right\} = x + iy.$$

This maps Δ_2 onto the semidisk

$$T_2: |z| < 1, \qquad x > 0,$$

the pair of rays (9.9) onto the segment

$$x = 0, \qquad |y| \leq \eta = \exp(-2\pi^2/\log 2), \tag{9.10}$$

and s_1 onto

$$z_1 = \exp\left\{\frac{-\pi\theta_0}{\log 2} + i\delta\pi\right\} = x_1 + iy_1 = r_1 e^{i\delta\pi},$$

say. The harmonic measure of the segment (9.10) at $z = x + iy$ w.r.t. T_2 is

$$\omega_1(z) = \frac{1}{\pi}\left\{\tan^{-1}\frac{y+\eta}{x} + \tan^{-1}\frac{\eta - y}{x} - \tan^{-1}\frac{x\eta}{1-\eta y} - \tan^{-1}\frac{x\eta}{1+\eta y}\right\}.$$

For clearly ω_1 is harmonic and bounded in T_1, and $\omega_1(z) = 1$ on the segment (9.10) and $\omega_1(z) = 0$, elsewhere on the boundary of T_1. To see this when $|z_1| = 1$, we use the fact that the triangles $0\eta z$ and $0z\eta^{-1}$ are similar in this case. Using the addition formula for the inverse tangent we obtain finally

$$\omega_1(z_1) = \frac{1}{\pi}\tan^{-1}\left\{\frac{2\eta x_1(1+\eta^2)(1-r_1^2)}{(r_1^2 - \eta^2)(1-\eta^2 r_1^2) + 4x_1^2\eta^2}\right\}$$

since $\eta < r_1 < 1$. In our case we have, using (9.3),

$$r_1 = \exp\left(\frac{-\pi\theta_0}{\log 2}\right), \quad \text{where} \quad 0.08 < \theta_0 < 0.081,$$

while $\eta < 10^{-10}$. Also $x_1 = r_1 \cos\delta\pi \geq r_1 \cos 7\pi/16 = r_1 \sin(\pi/16)$.

$$\omega(z_1) > 0.999\,\frac{2\eta x_1(1-r_1^2)}{\pi r_1^2} \geq 0.999\,\frac{2\eta}{\pi}\left(\frac{1}{r_1} - r_1\right)\sin\frac{\pi}{16}$$

$$> \frac{\pi}{16}\,\frac{2\eta}{\pi}\cdot\frac{2\pi\theta_0}{\log 2} > \exp\{-30.9\}.$$

This completes the proof of Lemma 16.

We must extend Lemma 16 to obtain a bound for $\omega(w)$ on the whole of $I_{j,k}$. This is

LEMMA 17. We have in $I_{j,k}$

$$\omega(w_1) \geq A_3 = e^{-37}.$$

We write

$$w_1 = a_j + t.$$

It follows from (7.6) that if

$$\phi(a_j) \leq t \leq \tfrac{1}{2}(a_{j+1} - a_j) \tag{9.11}$$

then

$$d(w_1) \geq \frac{2t}{25}.$$

Also the function $\omega(w)$ is positive and harmonic in the disk $|w - w_1| < d(w_1)$. Thus Harnack's inequality [3, p. 64] yields

$$\left|\frac{d}{dt}\omega(w_1)\right| < \frac{2}{d(w_1)}\omega(w_1) \tag{9.12}$$

i.e.

$$\left|\frac{d}{dt}\log\omega(w_1)\right| \leq \frac{25}{t}$$

hence if $w_1 = a_j + t_1$, $w_1' = a_j + t_1'$ are two points in the range (9.11) we have

$$|\log\omega(w_1) - \log\omega(w_1')| \leq 25\,|\log t_1' - \log t_1|.$$

We suppose that

$$t_1 = 2^{7/16}r < t_1' \leq 2^{1/2}r \quad \text{or} \quad 2^{-1/2}r \leq t_1' < t_1 = 2^{7/16}r$$

and apply Lemma 16 to w_1. Thus we obtain

$$\log \omega(w_1') \geq -\{30.9+\tfrac{25}{16}\log 2\} > -32, \qquad a_{j,k} \leq w_1' \leq a_{j,k+1}. \tag{9.13}$$

This yields Lemma 17 if $k_1+1<k<0$.

If $k=k_1+1$, we have to estimate $\omega(w_1)$ also on the interval $I'=[a_j, a_{j,k}]$. In this case it follows from our construction that

$$a_{j,k} \leq a_j + 2\phi(a_j).$$

Thus if w_1 is any point on I' it follows from Lemmas 10 and 11 that

$$d(a) \geq \tfrac{4}{5}\phi(a_j).$$

We deduce from (9.12) in this case that

$$\left|\frac{d}{dw_1}\log \omega(w_1)\right| \leq \frac{2.5}{\phi(a_j)} \quad \text{on} \quad I'.$$

Thus if w_1 is any point on I' and $w_1' = a_{j,k}$ is the right endpoint of I', we deduce that

$$\log \omega(w_1) \geq \log \omega(w_1') - 2\phi(a_j)\frac{2.5}{\phi(a_j)} \geq \log \omega(w_1') - 5.$$

On combining this with (9.13) we deduce that the inequality of Lemma 17 holds on I' also and so on all of $I_{j,k}$. Thus Lemma 17 is proved whenever $k<0$. The case $k \geq 0$ is similar and our proof is complete.

10. Proof of (1.3)

We need a final estimate.

LEMMA 18. *If l, λ are the lengths of the images of an interval $I'=I_{j,k}$ and the associated arc $\Gamma'=\Gamma_{j,k}$ respectively in the z plane then*

$$\frac{l}{\lambda} \leq \frac{1}{\pi A_3}(5+25\log 2).$$

Suppose that w_0 is any point on I' and let $z_0 = \rho e^{i\theta}$ be the image of w_0 in Δ. Let $\gamma = [\phi_1, \phi_1 + \lambda]$ be the image of Γ' on $|z| = 1$. Then, since harmonic measure is invariant under conformal mapping,

$$\omega(w_0, \Gamma', \Omega) = \omega(z_0, \gamma, \Delta) = \frac{1}{2\pi} \int_{\phi_1}^{\phi_1 + \lambda} \frac{(1-\rho^2)\, d\phi}{1 - 2\rho \cos(\theta - \phi) + \rho^2}$$

$$\leq \frac{\lambda}{2\pi} \frac{1+\rho}{1-\rho}.$$

Thus

$$1 - \rho \leq \frac{\lambda}{\pi \omega(w_0, \Gamma', \Omega)} \leq \frac{\lambda}{\pi A_3}$$

in view of Lemma 17.

On the other hand if $z = F(w)$ maps Ω onto Δ, then $F(w)$ maps $|w - w_0| < d(w_0)$ into Δ and now we deduce from Schwarz's Lemma that

$$|F'(w_0)| \leq \frac{1 - |z_0|^2}{d(w_0)} \leq \frac{2(1-\rho)}{d(w_0)} \leq \frac{2\lambda}{\pi A_3\, d(w_0)}.$$

Thus

$$l = \int_{I'} |F'(w_0)|\, |dw_0| \leq \frac{2\lambda}{\pi A_3} \int_{I'} \frac{|dw_0|}{d(w_0)}. \tag{10.1}$$

Now we again use (7.6). If I' is the interval $[a_{j,k}, a_{j,k+1}]$ where $k_1 + 1 < k < 0$, we set

$$w_0 = a_{j,k} + t,$$

and deduce from Lemma 14 that

$$d(w_0) \geq \tfrac{2}{25} t.$$

Thus

$$\int_{I'} \frac{|dw_0|}{d(w_0)} \leq \frac{25}{2} \int_{I'} \frac{dt}{t} = \frac{25}{2} \log 2.$$

If $k = k_1 + 1$ we must add to $[a_{j,k}, a_{j,k+1}]$ the interval $[a_j, a_{j,k}]$. In this case $t \leq 2\phi(a_j)$, $d(w_0) \geq \frac{4}{5}\phi(a_j)$ and

$$\int_{a_j}^{a_{j,k}} \frac{|dw_0|}{d(w_0)} \leq \frac{5}{4\phi(a_j)} \int_0^{2\phi(a_j)} dt \leq \tfrac{5}{2}.$$

Thus in all cases

$$\int_{I'} \frac{|dw_0|}{d(w_0)} \leq \tfrac{5}{2} + \tfrac{25}{2} \log 2.$$

Hence (10.1) yields

$$\frac{l}{\lambda} \leq \frac{1}{\pi A_3}(5 + 25 \log 2).$$

This proves Lemma 18.

Now (1.3) follows at once. For the length L of the level curve is the sum of the lengths $l_{j,k}$ of the images of the $I_{j,k}$. This yields

$$L \leq \sum l_{j,k} \leq \frac{1}{\pi A_3}(5 + 25 \log 2) \sum \lambda_{j,k} \leq \frac{1}{A_3}(10 + 50 \log 2) < 10^{18}$$

which is (1.3).

11. Proof of Theorem 1

We can now put our various results together. We need a final Lemma.

LEMMA 19. *Suppose that E is a set in Δ. Then there exists a bilinear function $z = L(Z)$ mapping $|Z| < 1$ into Δ and a set E' in $|Z| < 1$ onto E, such that*

$$|L'(Z)| < 2\delta(E) \quad \text{on} \quad E'. \tag{11.1}$$

Suppose that the upper bound of $|z|$ in the closure $\bar{E}$ of E is ρ. We suppose without loss of generality in this case that $\bar{E}$ contains $z = \rho$ so that $\bar{E}$ lies in

$$|\rho - z| \leq \delta, \tag{11.2}$$

where $\delta = \delta(E)$. We define $z = L(Z)$ by

$$z = \rho\frac{Z+tr}{t+rZ}, \qquad Z = t\frac{z-\rho r}{\rho - rz}, \quad \text{where} \quad r = \frac{\rho}{\rho+\delta}, \qquad \frac{2\rho}{2\rho+\delta} < t \leq 1.$$

If $\rho = 1$, we choose $t = 1$, so that $z = L(Z)$ is a bilinear map of $|Z| < 1$ onto $|z| < 1$, and the inverse image E' of E lies in $|Z| < 1$. If $\rho < 1$, we choose t just less than 1. Then $|Z| \leq t$ corresponds to $|z| \leq \rho$ and so if t is sufficiently near 1, $|Z| < 1$ corresponds to a subset of $|z| < 1$, which contains $|z| \leq \rho$ and so E.

Then if $z \in E$, so that (11.2) holds, we compute the derivative of L^{-1} and find

$$|L'(Z)| = \frac{|\rho - rz|^2}{t\rho(1-r^2)} = \frac{|\rho - r\rho + r(\rho - z)|^2}{t\rho(1-r^2)} \leq \frac{(\rho(1-r)+\delta r)^2}{t\rho(1-r^2)} = \frac{4\delta\rho}{t(2\rho+\delta)} < 2\delta,$$

which proves (11.1). This proves Lemma 19.

We recall that we have proved (1.3). We now apply this result to

$$g(Z) = f\{L(Z)\} \tag{11.3}$$

where E is a single level curve of $f(z)$ and $E' = L^{-1}(E)$. Then E' is part of a level curve of $g(Z)$ and so

$$|E'| \leq A_2.$$

In view of (11.1) we deduce that

$$|E| \leq 2\delta\,|E'| \leq 2\delta A_2.$$

where $\delta = \delta(E)$ and this is (1.4). Using (1.2) we deduce (1.5), for the length of any level set γ.

Finally suppose that E is part of a level set of $f(z)$ and that $\delta(E) = \delta$. We again employ the subsidiary function $g(Z)$, given by (11.3) and define $E' = L^{-1}(E)$. Since E' is part of a level set γ of $g(z)$, we can now apply (1.5) and deduce that

$$|E'| \leq |\gamma| \leq 2A_1A_2.$$

Since also $E = L(E')$ we deduce from (11.1) that

$$|E| \leq 2\delta\,|E'| \leq 4A_1A_2\delta(E).$$

This proves (1.1) with

$$A_0 = 4A_1A_2 < 10^{35},$$

as stated.

We have assumed throughout that Ω is an analytic Jordan domain and that a level set E is the inverse image of the real axis by $F(z)$. If E is the inverse image by $F(z)$ of a circle or straight line L, we can find a bilinear map $W = \phi(w)$ which maps L onto the real axis so that E is also the inverse image of the real axis by $f(z) = \phi\{F(z)\}$.

Next if $F(z)$ is a general univalent function and E is the inverse image of the real axis, suppose first that the image of E does not cover the whole real axis but leaves out a point w_0. Then $(F(z) - w_0)^{-1}$ is a regular univalent function with the same level set E. Thus we may assume that $F(z)$ is regular and univalent in this case. We now apply (1.2) and (1.3) to the level sets E_ρ of $F(\rho z)$, where $0 < \rho < 1$. Clearly $F(\rho z)$ maps Δ onto an analytic Jordan domain Also $|E_\rho|$ tends to $|E|$ as $\rho \to 1$, so that (1.2) and (1.3) also hold for $F(z)$.

Finally if the image of E covers the whole real axis, then E must consist of a single closed Jordan curve in Δ. Let ρ be the upper bound of z on E. Then E has at least one point on $|z| = \rho$.

Let $z_1 = \rho e^{i\theta}$ be such a point, and write $t = \frac{1}{2}(1+\rho)$, $z_0 = (\rho - t)e^{i\theta}$ and

$$f(z) = F(z_0 + tz).$$

Then the level set E' of f consists of a single curve going from $e^{i\theta}$ to $e^{i\theta}$ in $|z| < 1$ and having length $|E|/t > |E|$. Thus we may apply (1.3) to E' and obtain

$$|E| < |E'| \leq A_2$$

in this case also. Thus (1.3) holds in all cases. Also (1.2) is trivial in this case, since E is connected. Thus (1.2) and (1.3) hold in all cases and so do (1.4), (1.5) and (1.1). This completes the proof of Theorem 1.

We are most grateful to the referee, David Drasin, for his many helpful suggestions on exposition.

REFERENCES

[1] GEHRING, F. W. and JONES, P. W., Oral communication.

[2] HALL, T., *Sur la mesure harmonique de certains ensembles.* Arkiv. Mat. Astr. Fys. *25A* (1937), No. 28, 1–8.

[3] HEINS, M., *Selected topics in the classical theory of functions of a complex variable.* Holt, Rinehart and Winston (1962).
[4] JONES, P. W., *Bounded holomorphic functions with all level sets of infinite length.* Mich. Math. J., *27* (1980), 75–80.
[5] MAITLAND, B. J., *A note on functions regular and bounded in the unit circle and small at a set of points near the circumference of the circle.* Proc. Cam. Phil. Soc. *35* (1939), 382–388.
[6] NEVANLINNA, R., *Analytic Functions*, Springer-Verlag (1970).
[7] PIRANIAN G. and WEITSMAN, A., *Level sets of infinite length.* Comment. Math. Helvetici, *53* (1978), 161–164.
[8] WU, J.-M., *Harmonic measure and level sets of conformal mapping.* J. London Math. Soc., (2), *22* (1980), 263–268.

Imperial College
180 Queens Gate.
London SW7 2BZ

University of Illinois,
Urbana, Illinois 61801.

Received March 19, 1981

Comment. Math. Helvetici **56** (1981) 404–430 0010-2571/81/003404-27$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Whitehead transfers for S^1-bundles, an algebraic description

HANS J. MUNKHOLM and ERIK KJAER PEDERSEN

§0. Introduction

Let $F \to E \xrightarrow{f} B$ be a bundle of finite complexes. D. Anderson [1] defines a geometric transfer homomorphism $f^*_{\mathrm{Wh}}: \mathrm{Wh}(\pi_1 B) \to \mathrm{Wh}(\pi_1 E)$ which maps the Whitehead torsion of a homotopy equivalence $g: K \to B$ to the Whitehead torsion of the induced homotopy equivalence $g^*(E) \to E$. (Actually Anderson works with PL bundles, but by topological invariance of Whitehead torsion [4] this is no longer necessary.) Similarly there is a homomorphism $f^*_{K_0}: \tilde{K}_0(\mathbf{Z}\pi_1 B) \to \tilde{K}_0(\mathbf{Z}\pi_1 E)$ [6] relating the finiteness obstruction of a complex dominated by B to the finiteness obstruction of the total space of the pullback. The homomorphisms f^*_{Wh} and $f^*_{K_0}$ are known in some cases e.g. if the bundle is trival [9], [8] or when some assumptions are made on the behaviour of fundamental groups [7], [11], [12]. The case $F = S^1$ turns out to be of particular interest since knowing the geometric transfer homomorphisms in this case enables one to compute it in a number of other cases (see §7). In this paper we give a complete algebraic description of the geometric transfers in the case $F = S^1$ and also compute it for other fibres such as lens spaces, $S^3/Q(8)$, tori, see Theorems 7.1, 7.2, 7.3.

We prove

THEOREM A. *Let $S^1 \to E \to B$ be an orientable S^1-bundle with fundamental group sequence $\mathbf{Z} \to \pi \to \rho \to 1$ and let $x \in \mathrm{Wh}(\rho)$ be represented by an $n \times n$ invertible matrix $\bar{A}$ over $\mathbf{Z}\rho$. Choose A and B, matrices over $\mathbf{Z}\pi$, reducing to $\bar{A}$ and $\bar{A}^{-1}$ respectively, and let $t \in \pi$ be the image of $1 \in \mathbf{Z}$. Then $AB = I - (t-1)C$ for some $n \times n$ matrix C over $\mathbf{Z}\pi$ and $f^*_{\mathrm{Wh}}(x)$ is given by the $2n \times 2n$ matrix*

$$\begin{pmatrix} A & -C \\ t-1 & B \end{pmatrix}.$$

which determines a well defined element of $\mathrm{Wh}(\pi)$.

We also compute f^*_{Wh} in case the bundle is not orientable, see Theorem 2.1 and Theorem 3.1 for the exact formulation.

The theorem has various corollaries (Corollary 5.2):

THEOREM B. *Let $S^1 \to E \to B$ be an orientable fibration with $\pi_1 B$ finite and B finitely dominated. Then E is homotopy equivalent to a finite complex.*

This theorem was known in case of a product S^1-bundle or with the assumption that the exact sequence $0 \to \mathbf{Z}/\ker i_* \to \pi_1 E \xrightarrow{f_*} \pi_1 B$ is the reduction of an integral extension $0 \to \mathbf{Z} \to \pi \to \pi_1 B$.

A corollary of Theorem B is the following (Theorem 7.2):

THEOREM C. *Let M be a differentialbe manifold with 0 Euler characteristic and $M \to E \to B$ a bundle with structure group a compact connected Lie group. If $\pi_1 B$ is finite and B is finitely dominated then E is homotopy equivalent to a finite complex.*

Given a bundle of manifolds $S^1 \to E \to B$ various authors have defined transfers $K_{-i}(\mathbf{Z}\pi_1 B) \to K_{-i}(\mathbf{Z}\pi_1 E)$, $\mathrm{Wh}(\pi_1 B) \to \mathrm{Wh}(\pi_1 E)$ using concordances. It is easy to see [14] that these transfers agree with the ones considered here, and thus our computations extend to computations of concordance transfers.

§1. The algebraic S^1-transfer map

Let S be a ring (with unit, not necessarily commutative). Assume given an automorphism $s \to s^t$ of S and an element $\sigma \in S$ with $\sigma^t = \sigma$, $\sigma s = s^t \sigma$, for all $s \in S$. Then the right ideal (σ) generated by σ is a twosided ideal. Let $p : S \to S/(\sigma)$ be the projection. If $M_n(S)$ is the ring of $n \times n$ matrices over S then $s \to s^t$ extends to an automorphism, $A \to A^t$, of $M_n(S)$ with $\sigma A = A^t \sigma$. Also $M_n(S/(\sigma)) = M_n(S)/(\sigma)$.

THEOREM 1.1. *In the above situation there is a well defined homomorphism $p^{\#} : K_1(S/(\sigma)) \to K_1(S)$ given as follows: If $\bar{A} \in \mathrm{Gl}(n, S/(\sigma))$ then*

$$p^{\#}([\bar{A}]) = \left[\begin{pmatrix} A & -C \\ \sigma I_n & B^t \end{pmatrix}\right]$$

where $A, B, C \in M_n(S)$ are chosen to satisfy

(i) $p(A) = \bar{A}$
(ii) $AB = I_n - C\sigma$

Moreover,

(iii) $p^{\#}$ *is natural, i.e. if* $f: S \to S'$ *maps* σ *to* σ' *and* $f(s^t) = f(s)^t$ *then*

$$\begin{array}{ccc} K_1(S) & \xrightarrow{f_*} & K_1(S') \\ \uparrow{\scriptstyle p^{\#}} & & \uparrow{\scriptstyle (p')^{\#}} \\ K_1(S/(\sigma)) & \xrightarrow[\bar{f}_*]{} & K_1(S'/(\sigma')) \end{array}$$

commutes

(iv) $p^{\#}p_*(x) = x - x^t$, $x \in K_1(S)$
where $x \to x^t$ *is induced by* $s \to s^t$

(v) *Let* $\bar{s} \to \bar{s}^t = (\overline{s^t})$ *be the induced automorphism of* $S/(\sigma)$, *and let* $y \to y^t$ *be the induced automorphism on* $K_1(S/(\sigma))$. *Then* $p_*p^{\#}(y) = y - y^t$ *for all* $y \in K_1(S/(\sigma))$.

(vi) *If* σ *is not a zero divisor then* $S/(\sigma)$ *has a free, finitely generated resolution over* S *and the resulting classically defined transfer map* $K_1(S/(\sigma)) \to K_1(S)$ *coincides with* $p^{\#}$.

Warning. Note that $p^{\#}$ depends on the element σ and the automorphism $s \to s^t$, not just the projection p.

Proof. Let $M(A, B, C) = \begin{pmatrix} A & -C \\ \sigma I_n & B^t \end{pmatrix}$. If we have $AB = I_n - C\sigma$ then $BA = I_n - D\sigma$ for suitable D and

$$M(A, B, C)M(-B, -A, D) = \begin{pmatrix} -I_n & -AD & +CA^t \\ 0 & -I_n & \end{pmatrix}.$$

Hence $M(A, B, C)$ is invertible.

If, for fixed A, B one chooses a different C, say C_1, then $(C - C_1)\sigma = 0$. The identity

$$\begin{pmatrix} I_n & (C-C_1)A^t \\ 0 & I_n \end{pmatrix}\begin{pmatrix} A & -C \\ \sigma I_n & B^t \end{pmatrix} = \begin{pmatrix} A & -C_1 \\ \sigma I_n & B^t \end{pmatrix}$$

shows that $[M(A, B, C)]$ is independent of the choice of C.

Any change in A and B has the form $A_1 = A + A_2\sigma$, $B_1 = B + B_2\sigma$ for

arbitrary $A_2, B_2 \in M_n(S)$. It may be compensated for by letting $C_1 = C - A_2B^t - AB_2 - A_2\sigma B_2$. The identity

$$\begin{pmatrix} I_n & A_2 \\ 0 & I_n \end{pmatrix}\begin{pmatrix} A & -C \\ \sigma I_n & B^t \end{pmatrix}\begin{pmatrix} I_n & B_2 \\ 0 & I_n \end{pmatrix} = \begin{pmatrix} A_1 & -C_1 \\ \sigma I_n & B_1^t \end{pmatrix}$$

shows that $[M(A, B, C)]$ is independent of the choice of A and B. If $\bar{E} \in \mathrm{Gl}(n, S/(\sigma))$ is an elementary matrix then $\bar{E} = p(E)$ for some elementary matrix $E \in \mathrm{Gl}(n, S)$. Also $p(AE) = \overline{AE}$ and $(AE)(E^{-1}B) = I_n - C\sigma$. The identity

$$\begin{pmatrix} I_n & 0 \\ 0 & (E^{-1})^t \end{pmatrix}\begin{pmatrix} A & -C \\ \sigma I_n & B^t \end{pmatrix}\begin{pmatrix} E & 0 \\ 0 & I_n \end{pmatrix} = \begin{pmatrix} AE & -C \\ \sigma I_n & (E^{-1}B)^t \end{pmatrix}$$

shows that $\bar{A}$ and $\overline{AE}$ give rise to the same element in $K_1(S)$. Similarly one treats $\overline{EA}$ and $\bar{A}$.

Finally, it is easy to deal with stabilisation. Altogether we have shown that $p^{\#}: K_1(S/(\sigma)) \to K_1(S)$ is a well defined map.

The homomorphism property is obvious when one uses

$$[\bar{A}_1] + [\bar{A}_2] = \left[\begin{pmatrix} \bar{A}_1 & 0 \\ 0 & \bar{A}_2 \end{pmatrix}\right]$$

to define the sum.

The naturality property is obvious. If $A \in \mathrm{Gl}(n, S)$ and $\bar{A} = p(A)$ then one may take $B = A^{-1}$, $C = 0$. Hence the formula in (iv). Finally (v) holds because

$$p_*([M(A, B, C)]) = \left[\begin{pmatrix} \bar{A} & -\bar{C} \\ 0 & (\bar{A}^{-1})^t \end{pmatrix}\right] = [\bar{A}] - [\bar{A}]^t.$$

In case σ is not a zero divisor then $S/(\sigma)$ has $S \xrightarrow{r_\sigma} S \xrightarrow{p} S/(\sigma) \to 0$ as a resolution by left S modules, so p gives rise to a classical transfer map $\mathrm{tr}_p: K_1(S/(\sigma)) \to K_1(S)$, see e.g. Bass [17] p. 451. The commutative diagram

$$\begin{array}{ccccc} S^n \oplus S^n & \xrightarrow{\begin{pmatrix} \sigma I_n & 0 \\ 0 & I_n \end{pmatrix}} & S^n \oplus S^n & \xrightarrow{\begin{pmatrix} pI_n \\ 0 \end{pmatrix}} & (S/(\sigma))^n \\ \downarrow{\scriptstyle \begin{pmatrix} A^t & -\sigma C \\ I_n & B^t \end{pmatrix}} & & \downarrow{\scriptstyle \begin{pmatrix} A & -C \\ \sigma I_n & B^t \end{pmatrix}} & & \downarrow{\scriptstyle \bar{A}} \\ S^n \oplus S^n & \xrightarrow[\begin{pmatrix} I_n & 0 \\ 0 & I_n \end{pmatrix}]{} & S^n \oplus S^n & \xrightarrow[\begin{pmatrix} pI_n \\ 0 \end{pmatrix}]{} & (S/(\sigma))^n \end{array}$$

shows that

$$\mathrm{tr}_p([\bar{A}]) = p^{\#}([\bar{A}]).$$

In fact the identity

$$\begin{pmatrix} I & 0 \\ B^t & I \end{pmatrix}\begin{pmatrix} I & -A^t \\ 0 & I \end{pmatrix}\begin{pmatrix} A^t & -\sigma C \\ I & B^t \end{pmatrix} = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$$

shows that

$$\left[\begin{pmatrix} A^t & -\sigma C \\ I_n & B^t \end{pmatrix}\right]$$

vanishes. (Note: When we deal with left modules we have matrices on the right of row vectors in the description of homomorphisms.)

We proceed to introduce the example we are most interested in.

EXAMPLE 1.2. Let

$$\mathbf{Z} \xrightarrow{i} \pi \xrightarrow{p} \rho \longrightarrow 1 \qquad (\mathscr{E})$$

be an exact sequence of groups with $t = i(1) \in \pi$. Let $\omega : \rho \to \{\pm 1\}$ be a homomorphism with $gtg^{-1} = t^{\omega(p(g))}$ for all $g \in \pi$. Also let R be any commutative ring. By abuse of language we write $p : R\pi \to R\rho$ for the homomorphism of group rings induced by $p : \pi \to \rho$. Now $S = R\pi$ and $\sigma = t - 1$ together with the automorphism given on generators $g \in \pi$, by

$$g^t = \begin{cases} -gt^{-1} & \text{if} \quad \omega(p(g)) = -1 \\ g & \text{if} \quad \omega(p(g)) = 1 \end{cases}$$

is an example of the situation described in Theorem 1.1.

The resulting homomorphism $p^{\#} : K_1(R\rho) \to K_1(R\pi)$ induces (when $R = \mathbf{Z}$) a homomorphism (still called) $p^{\#} : \mathrm{Wh}(\rho) \to \mathrm{Wh}(\pi)$ which still satisfies (iv) and (v).

In this case the naturality shows that the following diagrams commute

$$\begin{array}{ccc} K_1(R\pi) & \xrightarrow{f_*} & K_1(R'\pi) \\ \uparrow{\scriptstyle p^{\#}} & & \uparrow{\scriptstyle p^{\#}} \\ K_1(R\rho) & \xrightarrow[f_*]{} & K_1(R'\rho) \end{array} \qquad \begin{array}{ccc} K_1(R\pi) & \xrightarrow{\varphi_*} & K_1(R\pi_1) \\ \uparrow{\scriptstyle p^{\#}} & & \uparrow{\scriptstyle p^{\#}} \\ K_1(R\rho) & \xrightarrow[\psi_*]{} & K_1(R\rho_1) \end{array}$$

Here $f: R \to R'$ is a ring homomorphism and φ, ψ are group homomorphisms making

$$\begin{array}{ccccccc} \mathbf{Z} & \longrightarrow & \pi & \longrightarrow & \rho & \longrightarrow & 1 \\ \| & & \downarrow{\scriptstyle \varphi} & & \downarrow{\scriptstyle \psi} & & \\ \mathbf{Z} & \longrightarrow & \pi_1 & \longrightarrow & \rho_1 & \longrightarrow & 1 \end{array}$$

commute and having $\omega_1(\psi(g)) = \omega(g)$, $g \in \rho$. Note that the element $t = i(1) \in \pi$ can be of finite or infinite order. Call the order k. If $k > 2$, then the extension $(\mathscr{E})$ determines ω uniquely. If $k = 1$ or 2 then $p^{\#}$ depends also on ω though this fact is suppressed in the notation. The choice of a generator for $\mathbf{Z}$ is immaterial. In fact, replacing t by t^{-1} can be compensated for by changing A to At^{-1}, B to tB and C to $-Ct$. The identity

$$\begin{pmatrix} I_n & 0 \\ 0 & -I_n \end{pmatrix}\begin{pmatrix} At^{-1} & Ct \\ (t^{-1}-1)I_n & (tB)^{t^{-1}} \end{pmatrix}\begin{pmatrix} tI_n & 0 \\ 0 & -t^{-1}I_n \end{pmatrix} = \begin{pmatrix} A & -C \\ (t-1)I_n & B^t \end{pmatrix}$$

shows the desired.

In case $k < \infty$ one may change t to t^r where $(r, k) = 1$. This gives a new exact sequence

$$\mathbf{Z} \xrightarrow[ir]{} \pi \xrightarrow[p]{} \rho \longrightarrow 1 \qquad (\mathscr{E}_r)$$

and an S^1-transfer map which we shall call $p^{\#}_{(r)}: K_1(\mathbf{Z}\rho) \to K_1(\mathbf{Z}\pi)$.

§2. Whitehead torsion and linear S^1-bundles

Let $f: E \to B$ be a locally trivial S^1-bundle with structure group $O(2)$ (for short: a linear S^1-bundle) over a connected, finite CW complex B. There results a pair $(\mathscr{E}, \omega)$

$$\mathbf{Z} = \pi_1(S^1) \xrightarrow{i} \pi = \pi_1(E, e_0) \xrightarrow{p} \rho = \pi_1(B, b_0) \longrightarrow 1 \qquad (\mathscr{E})$$

$$\omega: \pi_1(B, b_0) = \rho \to \{\pm 1\} = \mathrm{Aut}\,(\pi_1(S^1))$$

where p is induced by f. Thus, from Section 1, we get $p^{\#}: \mathrm{Wh}\,(\rho) \to \mathrm{Wh}\,(\pi)$.

Next, let X be a finite CW complex containing B as a deformation retract, with retraction r, say. Then E is a deformation retract of $Y = r^*(E)$. The pairs (Y, E) and (X, B) have Whitehead torsions $\tau(Y, E) \in \mathrm{Wh}(\pi)$, $\tau(X, B) \in \mathrm{Wh}(\rho)$. We shall prove

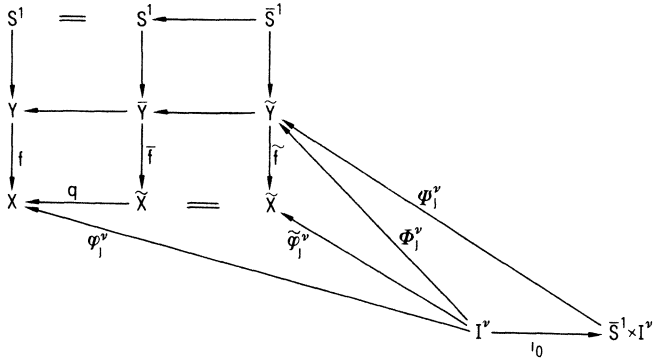
THEOREM 2.1. *In the above situation one has*

$$\tau(Y, E) = p^{\#}\tau(X, B).$$

Proof. First consider the case where (X, B) is in simplified form ((4.7) of Cohen, [5]), i.e.

$$X = B \cup \left(\bigcup_{j=1}^{n} e_j^r\right) \cup \left(\bigcup_{j=1}^{n} e_j^{r+1}\right). \tag{2.1}$$

Here we take r to be an *even* integer $> \dim B$. The cells e_j^ν have characteristic maps $\varphi_j^\nu : I^\nu \to X$ with $\varphi_j^r(\partial I^r) = b_0$. There is a commutative diagram



(2.2)

which we proceed to explain. The horizontal maps are coverings, universal ones being indicated by a $\sim$. The map $\bar{f}$ is the pull back of f. If $X \to \mathrm{BO}(2)$ classifies f then $\tilde{X} \xrightarrow{q} X \longrightarrow \mathrm{BO}(2)$ deforms into BS^1. Therefore, we can assume that $\bar{f} : \bar{Y} \to \tilde{X}$ is a *principal* S^1 bundle. Let $\mu : S^1 \times \bar{Y} \to \bar{Y}$ be the corresponding action. If, as in Section 1, k is the order of $t = i(1) \in \pi$ then $\tilde{Y} \to \bar{Y}$ is a principal $\mathbf{Z}/k\mathbf{Z}$ bundle ($\mathbf{Z}/k\mathbf{Z} = \mathbf{Z}$, if $k = \infty$) The action μ lifts to an action $\bar{\mu} : \bar{S}^1 \times \tilde{Y} \to \tilde{Y}$ where $\bar{S}^1 = \mathbf{R}/k\mathbf{Z}$ ($= \mathbf{R}$ if $k = \infty$). And $\tilde{f}$ is the corresponding principal $\bar{S}^1$-bundle. Finally $\tilde{\varphi}_j^\nu$ and Φ_j^ν lift φ_j^ν, and ψ_j^ν extends Φ_j^ν $\bar{S}^1$-equivariantly, i_0 being given by $i_0(x) = (0, x)$. Also note that $\tilde{E} = \tilde{Y} \mid \tilde{B}$ is an $\bar{S}^1$-invariant subspace of $\tilde{Y}$.

We pick basepoints $\tilde{e}_0 \in \tilde{E} \subseteq \tilde{Y}$ and $\tilde{b}_0 \in \tilde{B} \subseteq \tilde{X}$, compatibly, and above e_0 and b_0. We use them to identify π and ρ with the covering transformation groups on $\tilde{Y}$ and $\tilde{B}$ in the standard way (see e.g. §3 of Cohen, [5]).

We go on to describe the cellular chain complexes $C_*(\tilde{X}, \tilde{B})$ and $C_*(\tilde{Y}, \tilde{E})$.

2.A. *The description of* $C_*(\tilde{X}, \tilde{B})$

The cell e_j^ν lifts to a cell of $\tilde{X}$, say $\tilde{e}_j^\nu$, with characteristic map $\tilde{\varphi}_j^\nu$ (compare (2.1) and (2.2)). The translates under ρ, $g\tilde{e}_j^\nu$, form a cell decomposition of $(\tilde{X}, \tilde{B})$, i.e.

$$\tilde{X} = \tilde{B} \cup \left(\bigcup_{j,g} g\tilde{e}_j^r \right) \cup \left(\bigcup_{j,g} g\tilde{e}_j^{r+1} \right) \tag{2.3}$$

where the cell $g\tilde{e}_j^\nu$ has characteristic maps $g\tilde{\varphi}_j^\nu : I^\nu \to \tilde{X}$.

Now let $\iota_\nu \in H_\nu(I^\nu, \partial I^\nu)$ be the standard generator. Then $C_*(\tilde{X}, \tilde{B})$ has a $\mathbf{Z}\rho$-basis consisting of

$$x_j^\nu = (\tilde{\varphi}_j^\nu)_*(\iota_\nu) \in C_\nu(\tilde{X}, \tilde{B}), \qquad j = 1, 2, \ldots, n; \qquad \nu = r, r+1. \tag{2.4}$$

Using this basis we see that $C_*(\tilde{X}, \tilde{B})$ takes the form

$$(\mathbf{Z}\rho)^n \xleftarrow{\bar{A}} (\mathbf{Z}\rho)^n, \quad \text{where} \quad \bar{A} \in \mathrm{Gl}(n, \mathbf{Z}\rho).$$

Of course, $\bar{A}$ is invertible because $\tilde{X}$ deforms to $\tilde{B}$.

2.B. *The cells of* $(\tilde{Y}, \tilde{E})$

Let us fix the following ranges for the various indices used

j and l range over $1, 2, \ldots, n$,
g and h range over ρ,
α and β range over $\mathbf{Z}/k\mathbf{Z}$,
ν ranges over $r, r+1$

Also choose a set map $\sigma : \rho \to \pi$ splitting $p : \pi \to \rho$ and having $\sigma(1) = 1$.

The map $\sigma(g)\psi_j^\nu : \bar{S}^1 \times \mathrm{int}\,(I^\nu) \to \tilde{Y}$ then trivializes the restriction of $\tilde{Y}$ to the cell $g\tilde{e}_j^\nu \subseteq \tilde{X}$. We give $\bar{S}^1$ the obvious cellular structure with vertices α and 1-cells $(\alpha, \alpha+1)$, $\alpha \in \mathbf{Z}/k\mathbf{Z}$. The product structure on $\bar{S}^1 \times \mathrm{int}\,(I^\nu)$ can now be transported

into $\tilde{Y}$ by means of $\sigma(g)\psi_j^\nu$. Thus we get the cells

$$e^\nu(j, g, \alpha) = \sigma(g)\psi_j^\nu(\{\alpha\} \times \text{int}(I^\nu)) \tag{2.6}$$

$$e^{\nu+1}(j, g, [\alpha, \alpha+1]) = \sigma(g)\psi_j^\nu((\alpha, \alpha+1) \times \text{int}(I^\nu)) \tag{2.7}$$

with characteristic maps

$$\eta^\nu(j, g, \alpha) = \sigma(g)\psi_j^\nu(i_\alpha \times I^\nu) : I^\nu \to \tilde{Y} \tag{2.8}$$

$$\eta^{\nu+1}(j, g, [\alpha, \alpha+1]) = \sigma(g)\psi_j^\nu(i_{[\alpha,\alpha+1]} \times I^\nu) : I^{\nu+1} \to \tilde{Y}. \tag{2.9}$$

Here $i_\alpha : * \to \bar{S}^1$ and $i_{[\alpha,\alpha+1]} : I^1 \to \bar{S}^1$ are given by $i_\alpha(*) = \alpha$, $i_{[\alpha,\alpha+1]}(s) = s + \alpha (\in \mathbf{R}/k\mathbf{Z})$. It is clear that the cells of (2.6–7) form a cellular structure on $(\tilde{Y}, \tilde{E})$.

2.C. *The π action on the cells of $(\tilde{Y}, \tilde{E})$*

Recall, e.g. from MacLane [10], that each element of π is uniquely of the form $\sigma(g)t^\alpha$ (g, α ranging as in 2B). And the group structure is given by

$$\sigma(h)\sigma(g) = \sigma(hg)t^{\gamma(h,g)} \tag{2.10}$$

$$t\sigma(g) = \sigma(g)t^{\omega(g)} \tag{2.11}$$

where $\gamma : \rho \times \rho \to \mathbf{Z}/k\mathbf{Z}$ is a 2-cocycle representing the characteristic class of the extension $\mathbf{Z}/k\mathbf{Z} \to \pi \to \rho$.

Also, note that under the action $\bar{\mu} : \bar{S}^1 \times \tilde{Y} \to \tilde{Y}$ one has $\bar{\mu}(1, -) = t : \tilde{Y} \to \tilde{Y}$. It follows that

$$t\psi_j^\nu(s, -) = \psi_j^\nu(s+1, -), \qquad s \in \bar{S}^1 \tag{2.12}$$

and from (2.12) together with (2.10–11) and (2.8–9) one easily computes the action of the general element $\sigma(h)t^\beta \in \pi$ on the cells of $(\tilde{Y}, \tilde{E})$. One gets

$$\sigma(h)t^\beta \cdot \eta^\nu(j, g, \alpha) = \eta^\nu(j, hg, \alpha + \omega(g)\beta + \gamma(h, g)) \tag{2.13}$$

$$\sigma(h)t^\beta \cdot \eta^{\nu+1}(j, g, [\alpha, \alpha+1]) = \eta^{\nu+1}(j, hg, [\alpha + \omega(g)\beta + \gamma(h, g), \alpha + 1 + \omega(g)\beta + \gamma(h, g)]). \tag{2.14}$$

Especially

$$\sigma(h)t^{\beta} \cdot \eta^{\nu}(j, 1, 0) = \eta^{\nu}(j, h, \beta) \tag{2.15}$$

$$\sigma(h)t^{\beta} \cdot \eta^{\nu+1}(j, 1, [0, 1]) = \eta^{\nu+1}(j, h, [\beta, \beta+1]). \tag{2.16}$$

Thus there is a $\mathbf{Z}\pi$ basis for $C_*(\tilde{Y}, \tilde{E})$ consisting of the elements

$$y^{\nu}_{j,0} = \eta^{\nu}(j, 1, 0)_*(\iota_\nu), \quad \text{and} \tag{2.17}$$

$$y^{\nu+1}_{j,[0,1]} = \eta^{\nu+1}(j, 1, [0, 1])_*(\iota_{\nu+1}) \tag{2.18}$$

2.D. *The boundary in* $C_*(\tilde{Y}, \tilde{E})$

We want to prove

PROPOSITION 2.2. *Using the* $\mathbf{Z}\pi$ *basis given in* (2.17–18) $C_*(\tilde{Y}, \tilde{E})$ *takes the form*

$$(\mathbf{Z}\pi)^n \xleftarrow[(\iota,-1)]{} (\mathbf{Z}\pi)^n \oplus (\mathbf{Z}\pi)^n \xleftarrow[(t-1,-A^t)]{} (\mathbf{Z}\pi)^n$$

where $A \in M_n(\mathbf{Z}\pi)$ *has* $p(A) = \bar{A} \in \mathrm{Gl}(n, \mathbf{Z}\rho)$ (cfr. (2.5)).

Proof. Let us use superscript $\langle i \rangle$ to denote the i-skeleton of any CW complex. Also, if b_κ is a given R-basis for a free R-module M, and if $m \in M$ let us write $\langle m, b_\kappa \rangle$ for the coordinates of m relative to b_κ (here R will be $\mathbf{Z}$ or $\mathbf{Z}\pi$). The commutative diagram

$$\begin{array}{ccc}
H_{r+1}(I^1 \times I^r, \partial(I^1 \times I^r)) & \xrightarrow{\eta^{r+1}(j,1,[0,1])_*} & H_{r+1}(\tilde{Y}^{\langle r+1\rangle}, \tilde{Y}^{\langle r\rangle}) = C_{r+1}(\tilde{Y}, \tilde{E}) \\
\downarrow \partial & & \downarrow \partial \quad \downarrow \partial \\
H_r(\partial(I^1 \times I^r), (I^1 \times I^r)^{\langle r-1\rangle} \cup I^1 \times \partial I^r & \longrightarrow & H_r(\tilde{Y}^{\langle r\rangle}, \tilde{E}) = C_r(\tilde{Y}, \tilde{E}) \\
\downarrow \simeq (\mathrm{incl}_*)^{-1} & \nearrow_{(\eta^r(j,1,0) \cup \eta^r(j,1,1))_*} & \\
H_r(\{0,1\} \times I^r, \{0,1\} \times \partial I^r) & &
\end{array}$$

together with (2.15) shows that

$$\partial y^{r+1}_{j,[0,1]} = (t-1)y^r_{j,0}. \tag{2.19}$$

Similarly one proves

$$\langle \partial y^{r+2}_{j,[0,1]}, y^{r+1}_{l,0} \rangle = \delta_{jl}(t-1). \tag{2.20}$$

Let $A = (A_{jl}) \in M_n(\mathbf{Z}\pi)$ be given by

$$\partial y^{r+1}_{j,0} = \sum_l A_{jl} y^r_{l,0}. \tag{2.21}$$

Under $\tilde{f}_*$ $y^\nu_{j,0}$ maps to x^ν_j so a comparison with (2.5) reveals that

$$p(A) = \bar{A}. \tag{2.22}$$

Thus all that is now left to prove is

$$\langle \partial y^{r+2}_{j,[0,1]}, y^{r+1}_{l,[0,1]} \rangle = -(A_{jl})^t \in \mathbf{Z}\pi. \tag{2.23}$$

Let

$$A_{jl} = \sum_{h,\beta} a^{h\beta}_{jl} \sigma(h) t^\beta, \qquad a^{h\beta}_{jl} \in \mathbf{Z}. \tag{2.24}$$

Then

$$(A_{jl})^t = \sum\nolimits^+ a^{h\beta}_{jl} \sigma(h) t^\beta - \sum\nolimits^- a^{h\beta}_{jl} \sigma(h) t^{\beta-1} \tag{2.25}$$

where $\sum^\pm$ means the sum over all h, β with $\omega(h) = \pm 1$.

In order to prove (2.23) we *change to integral bases.* For $C_*(\tilde{Y}, \tilde{E})$ an integral basis consists of

$$y^\nu(j, g, \alpha) = \eta^\nu(j, g, \alpha)_*(\iota_\nu), \quad \text{and} \tag{2.26}$$

$$y^{\nu+1}(j, g, [\alpha, \alpha+1]) = \eta^{\nu+1}(j, g, [\alpha, \alpha+1])_*(\iota_{\nu+1}). \tag{2.27}$$

In terms of these the known formula, (2.21), becomes

$$\langle \partial y^{r+1}(j, 1, 0), y^r(l, h, \beta) \rangle = a^{h\beta}_{jl}. \tag{2.28}$$

And the desired one, (2.23), becomes

$$\langle \partial y^{r+2}(j, 1, [0, 1]),\ y^{r+1}(l, h, \beta)\rangle = \begin{cases} -a_{jl}^{h\beta} & \text{if} \quad \omega(h) = 1 \\ a_{jl}^{h,\beta+1} & \text{if} \quad \omega(h) = -1. \end{cases} \tag{2.29}$$

We shall prove (2.29) for the case $\omega(h) = -1$ (the other case being slightly easier). Consider the diagram (2.30).

By (2.28) the left hand vertical composition maps ι_{r+1} to $a_{jl}^{h,\beta+1}\iota_r$. Similarly (2.29) means that the right hand vertical composition maps ι_{r+2} to $a_{jl}^{h,\beta+1}\iota_{r+1}$. Thus, if we can fill in dotted arrows so that the squares commute up to the sign indicated then we are through. The action $\bar{\mu}: \bar{S}^1 \times \tilde{Y} \to \tilde{Y}$ induces a map

$$a: (I^1, \partial I^1) \wedge (\tilde{Y}^{\langle r+1\rangle}, \tilde{Y}^{\langle r\rangle}) \to (\tilde{Y}, \tilde{Y}^{\langle r+1\rangle})$$

where we include $(I^1, \partial I^1)$ into $(\bar{S}^1, \{0, 1\})$ via $i_{[0,1]}$. We put $\Sigma_1(x) = a_*(\iota_1 \wedge x)$. Since

$$\begin{array}{ccc} I^1 \times I^{r+1} & \xrightarrow{I^1 \times \eta^{r+1}(j,1,0)} & I^1 \times \tilde{Y} \\ \| & & \downarrow \bar{\mu} \\ I^{r+2} & \xrightarrow[\eta^{r+2}(j,1,[0,1])]{} & \tilde{Y} \end{array}$$

commutes, so does the upper square in (2.30).

$$\begin{array}{ccc} H_{r+1}(I^{r+1}, \partial I^{r+1}) & \xrightarrow{\iota_1 \wedge -} & H_{r+2}(I^{r+2}, \partial I^{r+2}) \\ \downarrow \eta^{r+1}(j,1,0)_* & (-1) & \downarrow \eta^{r+2}(j,1,[0,1])_* \\ H_{r+1}(\tilde{Y}^{\langle r+1\rangle}, \tilde{Y}^{\langle r\rangle}) & \xrightarrow{\Sigma_1} & H_{r+2}(\tilde{Y}^{\langle r+2\rangle}, \tilde{Y}^{\langle r+1\rangle}) \\ \downarrow \partial & (-1) & \downarrow \partial \\ H_r(\tilde{Y}^{\langle r\rangle}, \tilde{Y}^{\langle r\rangle} - e^r(l, h, \beta+1)) & \xrightarrow{\Sigma_2} & H_{r+1}(\tilde{Y}^{\langle r+1\rangle}, \tilde{Y}^{\langle r+1\rangle} - e^{r+1}(l, h, [\beta, \beta+1])) \\ \downarrow \eta^r(l,h,\beta+1)_*^{-1} & (-1) & \downarrow \eta^{r+1}(l,h,[\beta,\beta+1])_*^{-1} \\ H_r(I^r, \partial I^r) & \xrightarrow[\iota_1 \wedge -]{} & H_{r+1}(I^{r+1}, \partial I^{r+1}) \end{array}$$

Diagram 2.30.

If we let $T_1(s) = 1 - s$ for $s \in I^1 \subset \bar{S}^1$ then

$$\begin{array}{ccc} I^1 \times \tilde{Y} & \xrightarrow{\bar{\mu}} & \tilde{Y} \\ \downarrow{\scriptstyle T_1 \times \sigma(h)t^{\beta+1}} & & \downarrow{\scriptstyle t\sigma(h)t^{\beta+1} = \sigma(h)t^{\beta}} \\ I^1 \times \tilde{Y} & \xrightarrow[\bar{\mu}]{} & \tilde{Y} \end{array}$$

commutes. In fact

$$\begin{aligned} \bar{\mu}(1-s, \sigma(h)t^{\beta+1}x) &= \sigma(h)t^{\beta+1}\bar{\mu}(s-1, x) \\ &= \sigma(h)t^{\beta}\bar{\mu}(s, x). \end{aligned}$$

Hence

$$\begin{array}{ccc} I^1 \times I^r & = & I^{r+1} \\ \downarrow{\scriptstyle T_1 \times \eta^r(l, h, \beta+1)} & & \downarrow{\scriptstyle \eta^{r+1}(l, h, [\beta, \beta+1])} \\ I^1 \times \tilde{Y} & \xrightarrow[\bar{\mu}]{} & \tilde{Y} \end{array} \tag{2.31}$$

also commutes. Consequently $\bar{\mu}$ induces a map

$$b : (I^1, \partial I^1) \wedge (\tilde{Y}^{\langle r \rangle}, \tilde{Y}^{\langle r \rangle} - e^r(l, h, \beta+1)) \to (\tilde{Y}^{\langle r+1 \rangle}, \tilde{Y}^{\langle r+1 \rangle} - e^{r+1}(l, h, [\beta, \beta+1])).$$

We let $\Sigma_2(x) = b_*(\iota_1 \wedge x)$. Then the middle square in (2.30) commutes up to the sign -1 because

$$\partial(\iota_1 \wedge x) = -\iota_1 \wedge \partial x \in H_{r+1}(\tilde{Y}^{\langle r+1 \rangle}, \tilde{Y}^{\langle r+1 \rangle} - e^{r+1}(l, h, [\beta, \beta+1])).$$

And the bottom square in (2.30) commutes up to the sign -1 because of (2.31). This finishes the proof of Proposition 2.2.

2.E. *Completion of the proof of Theorem* 2.1

As in Section 1 we may choose $B, C \in M_n(\mathbf{Z}\pi)$ with $AB = I_n - C(t-1)$. Then $(t-1, -A^t)$ is split by $\begin{pmatrix} C \\ -B^t \end{pmatrix}$. Thus (from the definitions in §§15, 19 of Cohen,

[5]) we see that

$$\tau(Y, E) = \left[\begin{pmatrix} A & C \\ t-1 & -B^t \end{pmatrix}\right], \qquad \tau(B, X) = [\bar{A}].$$

This finishes the proof in the special case considered so far.

Using the special case repeatedly we see that

In the situation of Theorem 2.1, if X is a CW expansion of B then

$$\tau(Y, E) = 0. \tag{2.32}$$

Also, if the pairs (X, B) and (X_1, B) have equal torsions then X and X_1 both expand to a common expansion of B. Therefore, arguing as in Anderson's proof of 2.1 of [1], we see that

For fixed $f: E \to B$ $\tau(Y, E)$ depends only on $\tau(X, B)$. (2.33)

Since it is known (see e.g. (7.4) of Cohen's book [5]) that any $\tau \in \mathrm{Wh}(\rho)$ is realized by a pair in simplified form the proof is now complete.

§3. Relation to Anderson's geometric transfer

In this section we consider PL fibre bundles $f: E \to B$ with fiber S^1, as defined by Anderson [2]. Anderson, [1] proves that any such gives rise to a map $f^*: \mathrm{Wh}(\pi_1(B)) \to \mathrm{Wh}(\pi_1(E))$ – we shall call it a geometric transfer map – such that

For any homotopy equivalence $h: B_1 \to B$ (B_1 a finite polyhedron) one has $f^*\tau((h)) = \tau(\bar{h})$ where $\bar{h}: h^*E \to E$ covers h. (3.1)

In this context we also have $p = f_*: \pi_1(E) \to \pi_1(B)$ and the algebraic S^1-transfer map $p^{\#}: \mathrm{Wh}(\pi_1(B)) \to \mathrm{Wh}(\pi_1(E))$. As a corollary of Theorem 2.1 we get

THEOREM 3.1. *For any* PL *fiber bundle* $f: E \to B$ *with fiber* S^1 *one has*

$$f^* = p^{\#}: \mathrm{Wh}(\pi_1(B)) \to \mathrm{Wh}(\pi_1(E)).$$

In fact, Pedersen, [15], proves that f^* depends only on the exact sequence

$$\mathbf{Z} = \pi_1(S^1) \longrightarrow \pi_1(E) \xrightarrow{p} \pi_1(B) \longrightarrow 1 \tag{$\mathscr{E}$}$$

and the orientation map

$$\omega : \pi_1(B) \to \{\pm 1\}.$$

Now any pair $(\mathscr{E}, \omega)$ can be realized by a PL fiber bundle (fiber S^1) with structure group O (2). Thus we may assume that Theorem 2.1 applies to $f : E \to B$. Also, any $\tau \in \mathrm{Wh}(\pi_1(B))$ has the form $\tau(X, B)$ for some $X \supseteq B$, and if $i : B \to X$, $r : X \to B$ are the inclusion and retraction respectively then $\tau = \tau(X, B) = i_*^{-1}\tau(i) = r_*\tau(i) = -\tau(r)$.

Thus

$$f^*\tau = -f^*\tau(r) = -\tau(\bar{r}) = \tau(Y, E) = p^{\#}\tau(X, B) = p^{\#}\tau.$$

Remark 3.2. Instead of comparing $p^{\#}$ with Anderson's f^* one might also redo Anderson's work, [1], in the context of locally trivial S^1-bundles $f : E \to B$ over finite CW complexes. One proves that any such gives a well defined map $f^* : \mathrm{Wh}(\pi_1(B)) \to \mathrm{Wh}(\pi_1(E))$, and that f^* depends only on the fundamental group sequence and the orientation map (as in Pedersen, [15]). The above argument then proves that $f^* = p^{\#}$ also in such a context.

§4. The Bass–Heller–Swan homomorphism

Let ρ be a group. In [3], Bass, Heller and Swan define a split epimorphism $\varphi : K_1(\mathbf{Z}(\rho \times \mathbf{Z})) \to K_0(\mathbf{Z}\rho)$. The splitting $h : K_0(\mathbf{Z}\rho) \to K_1(\mathbf{Z}(\rho \times \mathbf{Z}))$ is given by $h([P]) = [P[s, s^{-1}], l_s]$. Here s is the generator of $\mathbf{Z}$, P is any finitely generated projective $\mathbf{Z}\rho$ module, $P[s, s^{-1}]$ its extension to a $\mathbf{Z}(\rho \times \mathbf{Z}) = \mathbf{Z}(\rho)[s, s^{-1}]$ module, and l_s is left multiplication by s. It is easily seen that there are "reduced editions" forming a commutative diagram

$$\begin{array}{ccc} \tilde{K}_0(\mathbf{Z}\rho) & \xrightarrow{h} & \mathrm{Wh}(\rho \times \mathbf{Z}) \\ & \searrow^{\mathrm{id}} & \downarrow \varphi \\ & & \tilde{K}_0(\mathbf{Z}\rho) \end{array}$$

In this section we shall give the following homotopy theoretic interpretation of φ.

THEOREM 4.1. *Let X be a finite, connected* CW *complex with fundamental group ρ. Any $\tau \in \mathrm{Wh}(\rho \times \mathbf{Z})$ is the Whitehead torsion, $\tau(g)$, of some map $g : Y \to$*

$X \times S^1$ *where*

(i) Y *is a connected, finite* CW *complex with* $Z = g^{-1}(X \times 1)$ *a connected subcomplex*
(ii) $g \mid Z : Z \to X \times 1$ *induces an isomorphism on fundamental groups.*

For such a g *let*

$$\begin{array}{ccc} \bar{Y} & \xrightarrow{g} & X \times \mathbf{R} \\ \downarrow{\scriptstyle q} & & \downarrow{\scriptstyle X \times \exp} \\ Y & \xrightarrow[g]{} & X \times S^1 \end{array}$$

be a pull-back, and put $\bar{Y}_+ = \bar{g}^{-1}(X \times [0, \infty))$. *Then*

(iii) $\bar{g}_+ : \bar{Y}_+ \to X \times [0, \infty) \to X$ *induces an isomorphism* $\bar{g}_{+*} : \pi_1(\bar{Y}_+) \to \rho$,
(iv) $\bar{Y}_+$ *is finitely dominated with finiteness obstruction given by* $\bar{g}_{+*}(w(\bar{Y}_+)) = \varphi(\tau) \in \tilde{K}_0(\mathbf{Z}\rho)$.

Proof. Certainly $\tau = \tau(g)$ for some homotopy equivalence $g : Y \to X \times S^1$ where Y is a finite complex. Replacing Y by a regular neighborhood of Y in some $\mathbf{R}^N$ we have Y a compact manifold. Then make prog: $Y \to X \times S^1 \to S^1$ smooth and transverse to 1 to ensure that $Z = g^{-1}(X \times 1)$ is a submanifold. Finally do "ambient surgery" to ensure that Z be connected and that (ii) holds, and triangulate the pair (X, Y) to obtain (i).

It follows easily from (i) and (ii) that $\bar{Y}_+$ and $\bar{Y}_- = \bar{g}^{-1}(X \times (-\infty, 0])$ are subcomplexes with $g : \bar{Y}_+ \cap \bar{Y}_- \to Z$ a homeomorphism. An application of van Kampens theorem then gives the following pushout diagram of groups

$$\begin{array}{ccccc} & & \pi_1(\bar{Y}_+) & & \\ & {\scriptstyle i_+} \nearrow & & \searrow {\scriptstyle j_+} & \\ \pi_1(\bar{Y}_+ \cap \bar{Y}_-) = \rho & & & & \rho = \pi_1(\bar{Y}). \\ & {\scriptstyle i_-} \searrow & & \nearrow {\scriptstyle j_-} & \\ & & \pi_1(\bar{Y}_-) & & \end{array}$$

Here $j_+ i_+ = j_- i_- = \mathrm{id}$ with the identification indicated. Since this implies that $\pi_1(\bar{Y}_\pm)$ is a semi direct product $K_\pm \times_{\alpha_\pm} \rho$ it follows that the inclusion $\rho \to (K_+^* K_-) \times_{\alpha_+ * \alpha_-} \rho$ is an isomorphism. Thus $K_\pm = \{1\}$ and $j_+ : \pi_1(\bar{Y}_+) \to \rho$ is an isomorphism. This proves (iii).

To prove (iv) we consider the cellular complexes of the universal coverings of $X \times \mathbf{R}$ and $\bar{Y}$ as well as their restrictions to $X \times [0, \infty)$ and $\bar{Y}_+$. They fit into a diagram with short exact rows

$$\begin{array}{ccccc} C_*(\tilde{X} \times [0,\infty)) & \longrightarrow & M(\gamma_+) & \longrightarrow & C_{*-1}(\bar{Y}_+) \\ \downarrow & & \downarrow & & \downarrow \\ C_*(\tilde{X} \times \mathbf{R}) & \longrightarrow & M(\gamma) & \longrightarrow & C_{*-1}(\tilde{Y}). \end{array}$$

Here $M(\gamma)$ is the mapping cone of $\gamma = \tilde{g}_* : C_*(\tilde{Y}) \to C_*(\tilde{X} \times \mathbf{R})$, and $M(\gamma_+)$ arises similarly. By definition τ is the Whitehead torsion, $\tau(M(\gamma))$, of the acyclic simplex $M(\gamma)$ with its $\mathbf{Z}(\rho \times \mathbf{Z})$ basis coming from the cells of Y and $X \times S^1$ lifted to $\tilde{Y}$ and $\tilde{X} \times \mathbf{R}$. Also $X \times [0, \infty)$ has the homotopy type of X so $C_*(\tilde{X} \times [0, \infty))$ is equivalent (over $\mathbf{Z}\rho$) to a finite free chain complex. Hence it suffices to prove that $M(\gamma_+)$ is dominated (over $\mathbf{Z}\rho$) by a finite, free chain complex, and that its finiteness obstruction is $\varphi(\tau)$.

Now we may choose the lifted cells in $\tilde{Y}$ and $\widetilde{X \times S^1} = \tilde{X} \times \mathbf{R}$ so that:

If $b_{i,1}, \ldots, b_{i,k(i)}$ is the preferred basis for $M_i(\gamma)$ over $\mathbf{Z}(\rho \times \mathbf{Z})$ and $s \in \mathbf{Z}$ is the generator for $\mathbf{Z}$ (written multiplicatively) then

$$\{b_{ij} s^k \mid j = 1, 2, \ldots, k(i);\ k = 0, 1, 2, \ldots\} \tag{4.1}$$

is a basis for $M_i(\gamma_+)$ over $\mathbf{Z}(\rho)$.

In fact, in each $\rho \times \mathbf{Z}$ orbit of cells in $\tilde{Y}$ or $X \times S^1$ one chooses a cell e so that $e \subseteq \tilde{Y}$ or $\tilde{X} \times [0, \infty)$ but $s^{-1}e$ is not. After this observation the following proposition finishes the proof of Theorem 4.1.

PROPOSITION 4.2. *Let*

$$D = \left(D_n \xrightarrow{d_n} D_{n-1} \xrightarrow{d_{n-1}} \cdots \xrightarrow{d_1} D_0 \right)$$

be an acyclic chain complex of finitely based $\mathbf{Z}(\rho \times \mathbf{Z})$ *modules. Let* $b_i = \{b_{i,1}, \ldots, b_{i,k(i)}\}$ *be a preferred basis for* D_i *and let* $D_i^{(+)}$ *be the* $\mathbf{Z}(\rho)$ *module generated by all* $b_{i,j} s^k$, $(j = 1, 2, \ldots, k(i);\ k = 0, 1, 2, \ldots)$. *Choose an integer* N *so big that* $s^N d_i(D_i^{(+)}) \subseteq D_{i-1}^{(+)}$ *for all* i, *and let*

$$D^{(N)} = \left(D_n^{(+)} \xrightarrow{s^N d_n} D_{n-1}^{(+)} \longrightarrow \cdots \xrightarrow{s^N d_1} D_0^{(+)} \right)$$

be the resulting chain complex of $\mathbf{Z}\rho$ *modules. Then* $D^{(N)}$ *is finitely dominated with finiteness obstruction* $w(D^{(N)}) = \varphi(\tau(D))$ *in* $\tilde{K}_0(\mathbf{Z}\rho)$.

Proof. We use induction on n. For $n = 1$ Bass, Heller and Swan show that $\mathrm{Cok}\,(s^N d_1)$ is a projective over $\mathbf{Z}\rho$, and they define φ by letting $\varphi(\tau(D)) = [\mathrm{Cok}\,(s^N d_1)]$. Since $D^{(N)}$ is equivalent to the trivial complex $\mathrm{Cok}\,(s^N d_1)$ we have the desired conclusion. Let $n > 1$ and write E for the chain complex $D_0 \xrightarrow{1} D_0$ concentrated in degrees 1 and 0, F for $D_0 \xrightarrow{-1} D_0$ in degrees 2 and 1. Also choose a splitting σ of $d_1 : D_1 \to D_0$ and let

$$\bar{D} = \left(D_n \xrightarrow{d_n} \cdots \xrightarrow{\binom{d_3}{0}} D_2 \oplus D_0 \xrightarrow{\binom{d_2 \quad \sigma}{0 \quad -1}} D_1 \oplus D_0 \xrightarrow{(d_1, 1)} D_0 \right)$$

$$\hat{D} = \left(D_n \xrightarrow{d_n} \cdots \xrightarrow{\binom{d_3}{0}} D_2 \oplus D_0 \xrightarrow{(d_2, \sigma)} D_1 \right).$$

The obvious short sequences

$$0 \to D \to \bar{D} \to F \to 0, \qquad 0 \to E \to \bar{D} \to \hat{D} \to 0$$

(which are compatibly based over $\mathbf{Z}(\rho \times \mathbf{Z})$) show that

$$\tau(D) = \tau(\bar{D}) = \tau(\hat{D}).$$

Also, for N large enough, we have short exact sequences

$$0 \to D^{(N)} \to \bar{D}^{(N)} \to F^{(N)} \to 0, \qquad 0 \to E^{(N)} \to \bar{D}^{(N)} \to \hat{D}^{(N)} \to 0$$

of $\mathbf{Z}(\rho)$ chain complexes. Now $F^{(N)}$ and $E^{(N)}$ are finitely dominated with vanishing finiteness obstruction. Also, by the inductive hypothesis, $\hat{D}^{(N)}$ is finitely dominated with $w(\hat{D}^{(N)}) = \varphi\tau(\hat{D})$. It follows that $\bar{D}$, and then D, are finitely dominated with $w(D^{(N)}) = w(\bar{D}^{(N)}) = w(\hat{D}^{(N)})$, which finishes the proof.

§5. The $\tilde{K}_0$- and Wh transfer maps related by the Bass–Heller–Swan homorphism

Let $f : E \to B$ be a Hurewicz fibration with fibre S^1 and connected, finitely dominated base B. Let

$$\mathbf{Z} = \pi_1(S^1) \to \pi = \pi_1(E) \to \rho = \pi_1(B) \to 1 \qquad (\mathscr{E})$$

$$\omega : \pi_1(B) \to \{\pm 1\}$$

be the associated pair as in Section 1. In [6], Ehrlich showed that f gives rise to a homomorphism $f^*: \tilde{K}_0(\mathbf{Z}\rho) \to \tilde{K}_0(\mathbf{Z}\pi)$ taking $w(B)$ to $w(E)$. And in [15], Pedersen proved that f^* depends only on the pair $(\mathscr{E}, \omega)$. We shall give the following algebraic description of f^*.

THEOREM 5.1. *For any Hurewicz fibration as above, the diagram commutes.*

$$\begin{array}{ccc} \tilde{K}_0(\mathbf{Z}\pi) & \xleftarrow{\varphi} & \mathrm{Wh}\,(\pi \times \mathbf{Z}) \\ \uparrow{\scriptstyle f^*} & & \uparrow{\scriptstyle (f^* \times 1)^\#} \\ \tilde{K}_0(\mathbf{Z}\rho) & \xrightarrow[h]{} & \mathrm{Wh}\,(\rho \times \mathbf{Z}) \end{array}$$

COROLLARY 5.2. *Let $f: E \to B$ be an orientable S^1-fibration with B finitely dominated and connected. If $\pi_1(B)$ is finite then E has the homotopy type of a finite complex.*

Proof. If $\pi_1(E)$ is infinite, then $\ker(p) = \mathbf{Z}$ so $\mathscr{E}$ is pseudoabelian, in the sense of Ehrlich, [7]. Hence the result follows from Ehlich, [7] (or see Munkholm and Pedersen, [12]). Compare also with the proof of Proposition 6.1.

Now let $\pi = \pi_1(E)$ be finite. Then an easy application of Milnor's Mayer–Victoris sequence shows that $p_*: \tilde{K}_0(\mathbf{Z}\pi) \to \tilde{K}_0(\mathbf{Z}\rho)$ is onto. Thus the general element of $\tilde{K}_0(\mathbf{Z}\rho)$ has the form $[p_*P]$ for some $\mathbf{Z}\pi$ module P which admits a stable inverse Q, i.e. $P \oplus Q \cong (\mathbf{Z}\pi)^n$. Now

$$\begin{aligned} h([p_*P]) &= [(p_*P)[s, s^{-1}], l_s] \\ &= (p \times 1)_*[P[s, s^{-1}], l_s]. \end{aligned}$$

But by (iv) of Theorem 1.1 $(p \times 1)^\#(p \times 1)_* = 0$ in the orientable case. It follows that $f^* = 0$ as claimed.

Proof of Theorem 5.1. Since the given data $(\mathscr{E}, \omega)$ can be realized by a PL fibre bundle, and since f^* depends only on $(\mathscr{E}, \omega)$ we may assume that f *is* a PL fibre bundle. By Theorem 3.1 it then suffices to show that

$$\begin{array}{ccc} \tilde{K}_0(\mathbf{Z}\pi) & \xleftarrow{\varphi} & \mathrm{Wh}\,(\pi \times \mathbf{Z}) \\ \uparrow{\scriptstyle f^*} & & \uparrow{\scriptstyle f^*} \\ \tilde{K}_0(\mathbf{Z}\rho) & \xleftarrow[\varphi]{} & \mathrm{Wh}\,(\rho \times \mathbf{Z}) \end{array}$$

the righthand vertical arrow is associated with the bundle crossed with S^1

commutes. And that is an easy consequence of the description of φ given in Theorem 4.1.

§6. Some computations

We have so far been unable to answer the following

Basic question. Does there exist an orientable extension $\mathbf{Z} \to \pi \xrightarrow{p} \rho \to 1$ (with ρ finitely presented, say) for which $0 \neq p^{\#}: \mathrm{Wh}(\rho) \to \mathrm{Wh}(\pi)$?

For the unoriented case the examples of Pedersen and Taylor, [16], together with Theorem 5.1 shows that one can have $(p \times 1)^{\#} \neq 0$. Since essentially nothing is known about $\mathrm{Wh}(\rho)$ when ρ is infinite we have concentrated on the case when ρ is finite.

PROPOSITION 6.1. *If* $\mathbf{Z} \to \pi \xrightarrow{p} \rho \to 1$ *is orientable and* π *is infinite,* ρ *is finite, then* $p^{\#} = 0: K_1(\mathbf{Z}\rho) \to K_1(\mathbf{Z}\pi)$.

Proof. If $|\rho| = m$ then there is a commutative diagram

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathbf{Z} & \longrightarrow & \pi & \xrightarrow{p} & \rho & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle m} & & \downarrow{\scriptstyle i} & & \| & & \\
0 & \longrightarrow & \mathbf{Z} & \longrightarrow & \rho \times \mathbf{Z} & \xrightarrow[p_1]{} & \rho & \longrightarrow & 1
\end{array}$$

where m is multiplication by m and i is an inclusion. One may realize this diagram by

$$\begin{array}{ccccc}
S^1 & \longrightarrow & E & \xrightarrow{f} & B \\
\downarrow & & \downarrow & & \| \\
S^1/(\mathbf{Z}/m\mathbf{Z}) & \longrightarrow & E/(\mathbf{Z}/m\mathbf{Z}) & \xrightarrow[f_1]{} & B
\end{array}$$

where the map f is a PL fibre bundle with fibre S^1, hence a principal S^1-bundle, and the $\mathbf{Z}/m\mathbf{Z}$-action comes from the inclusion $\mathbf{Z}/m\mathbf{Z} \to S^1$. It then follows that $f^* = i^* f_1^*$, i.e. – in view of Theorem 3.1 – $p^{\#} = i^* p_1^{\#}$. Since p_{1*} is onto a reference to (iv) in Theorem 1.1 finishes the proof.

For the rest of the paragraph assume that π is finite. Recall that $Cl_1(\mathbf{Z}\pi)$ is defined to be the kernel of

$$SK_1(\mathbf{Z}\pi) \to \bigoplus_q SK_1(\hat{\mathbf{Z}}_q\pi) \hookrightarrow \bigoplus_q K_1(\hat{\mathbf{Z}}_q\pi)$$

where $\hat{\mathbf{Z}}_q$ denotes the q-adic completion of $\mathbf{Z}$ (see Oliver, [13]).

PROPOSITION 6.2. *If* $\mathbf{Z} \to \pi \to \rho \to 1$ *is orientable, and* π *is finite then* $p^{\#}(K_1(\mathbf{Z}\rho)) \subseteq Cl_1(\mathbf{Z}\pi)$.

Proof. Since $Q\pi \to Q\rho$ splits as a map of $Q\pi$ modules, $p^{\#} = 0 : K_1(Q\rho) \to K_1(Q\pi)$. Hence, from Theorem 1.1 $p^{\#}(K_1(\mathbf{Z}\rho)) \subseteq SK_1(\mathbf{Z}\pi)$ $(= \operatorname{Ker}(K_1(\mathbf{Z}\pi) \to K_1(Q\pi)))$. Similarly $p_* : K_1(\hat{\mathbf{Z}}_q\pi) \to K_1(\hat{\mathbf{Z}}_q\rho)$ is onto so the same argument finishes the proof.

Now $Cl_1(\mathbf{Z}\pi)$ is pretty well understood when π is abelian, so one might hope to detect nontriviality of $p^{\#}$ by projecting into an abelian group. However one has

PROPOSITION 6.3. *Let* $\mathbf{Z} \to \pi \to \rho \to 1$ *be orientable and* π *finite. If* $h : \pi \to \pi_1$ *is a homomorphism into an abelian group then* $h_* p^{\#} = 0 : K_1(\mathbf{Z}\rho) \to K_1(\mathbf{Z}\pi_1)$.

Proof. Clearly we can assume that h is onto. Then, by the naturality in Theorem 1.1(iii), we can assume that π is abelian and h is the identity. But any such extension is the reduction of an integral extension so we have a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbf{Z} & \xrightarrow{i_1} & \pi' & \xrightarrow{p_1} & \rho & \longrightarrow & 0 \\ & & \| & & \downarrow{\scriptstyle \varphi} & & \| & & \\ & & \mathbf{Z} & \xrightarrow[i]{} & \pi & \xrightarrow[p]{} & \rho & \longrightarrow & 0 \end{array}$$

of abelian groups and with exact rows, and we only have to refer to Theorem 1.1(iii) and Proposition 6.1.

§7. Examples

In this section we infer consequences for the geometrically defined transfers in bundles with other fibres than S^1. We remind the reader of some facts from [15]: Let $F \xrightarrow{i} E \xrightarrow{f} B$ be a bundle (structure group Homeo (F)) with B and F finite

CW-complexes. Anderson [1] then gives a homomorphism $f^*: \mathrm{Wh}(\pi_1 B) \to \mathrm{Wh}(\pi_1 E)$ that relates Whitehead torsion at total and basespace level. By the fundamental group data of the bundle we understand the exact sequence $\pi_1 F \to \pi_1 E \to \pi_1 B$ and the orientation homomorphism $\pi_1 E \to \pi_0$ (Homeo $(F, *)$) (Homeo $(F, *)$ = basepoint preserving homeomorphisms). The existence of universal examples with given fundamental group data [15] implies that f^* above only depends on F and the fundamental group data. This means that we compute f^* in general if we can compute f^* for enough examples of bundles with F as fibre. The results of [15] also tell us how many examples we need to compute with a given fundamental group ρ in the base and a given orientation homomorphism $\rho \to \pi_0(\mathrm{Homeo}\,(F))$: Let Homeo $(F) \to F$ be the evaluation map and $G_1^{\mathrm{Top}}(F) \subset \pi_1(F)$ the image of the induced map. If it is possible to realize some fundamental group data $\pi_1 F \to \pi \to \rho$ compatible with the above orientation homomorphism and with $A = \ker(\pi_1 F \to \pi)$, then all such fundamental group data are classified by $H^2(\rho; G_1(F)/A)$ (local coefficients) in the sense that this group acts (transitively and faithfully) on the set of realizable fundamental group data. Also the image in $H^2(\rho; C(\pi_1(F)/A))$ gives the corresponding action on the exact sequence [10]. So if we are considering an orientable case the image in $H^2(\rho; C(\pi_1(F)/A))$ will be the characteristic class of the extension $\pi_1(F)/A \to \pi \to \rho$. We now restrict ourselves to orientable bundles. (Note: this does not imply that $\pi_1 E \to \pi_0(\mathrm{Homeo}\,(F, *))$ is trivial.) We use these observations to compute some examples:

Let $L = L(m; a_1, \ldots, a_n)$ be a $2n-1$ dimensional lens space given as $S^{2n-1} = \{(z_1, \ldots, z_n) \in \mathbf{C}^n \mid \sum |z_i|^2 = 1\}$ divided out by $\mathbf{Z}/m\mathbf{Z}$ thought of as m'th roots of unity and $\zeta \in \mathbf{Z}/m\mathbf{Z}$ acting as $\zeta(z_1, \ldots, z_n) = (\zeta^{a_1} z_1, \ldots, \zeta^{a_n} z_n)$ (the a_i's are relatively prime and prime to m). S^1 acts on L by the formulae

$$z \cdot [(z_1, \ldots, z_n)] = [(\xi^{a_1} z_1, \ldots, \xi^{a_n} z_n)] \quad \text{where} \quad \xi^m = z.$$

We let $s \in \pi_1(L)$ be the element obtained by letting S^1 act on the basepoint. Let $L \to E \to B$ be an orientable bundle with fundamental group data $\mathbf{Z}/m\mathbf{Z} \xrightarrow{i} \pi \xrightarrow{p} \rho \to 1$. Consider the exact sequence

$$\mathbf{Z} \to \pi \to \rho \to 1$$

where $\mathbf{Z} \to \pi$ sends 1 to $i(s)$. We then have homomorphisms $p_{(b)}^{\#}: \mathrm{Wh}(\rho) \to$

Wh (π) defined for any b prime to the order of $i(s)$ (compare with the end of Section 1).

THEOREM 7.1. *The geometric* Wh-*transfer* f^* *associated with the bundle* $L \to E \to B$ *is given by*

$$f^* = \sum p^{\#}_{(b_i)}$$

where $a_i b_i \equiv 1 \pmod{m}$.

Proof. $G_1^{\mathrm{Top}}(L) = \pi_1(L)$ since we have exhibited an S^1-action on L which applied to the basepoint gives a generator of $\pi_1(L)$. Thus the transfer is completely determined by the exact sequence $\mathbf{Z}/m\mathbf{Z} \to \pi \to \rho \to 1$ and we need to construct appropriate examples. Let $\mathbf{Z} \to \pi \to \rho \to 1$ be the exact sequence above and let $S^1 \to X \to B$ be a principal S^1-bundle realizing this fundamental group data. (This is always possible by [12] or [15].) Clearly $L \to X \times_{S^1} L \to B$ realizes the fundamental group data $\mathbf{Z}/m\mathbf{Z} \to \pi \to \rho$. (Notice we have carefully chosen a generator of $\pi_1(L)$ thus exhibiting an isomorphism to $\mathbf{Z}/m\mathbf{Z}$.) We prove the theorem by induction, using what amounts to an S^1-CW structure on L. Let K be an S^1-equivariant regular neighborhood of $L(m; a_1, \ldots, a_{n-1}) \subset L(m; a_1, \ldots, a_n)$. Then $X \times_{S^1} L \to B$ is the union of two bundles $A = X \times_{S^1} K$ and $C = X \times_{S^1} S^1 \times D^{2n-2}$ (the S^1-action on $S^1 \times D^{2n-2}$ given by $z \cdot (z_n, d) = (z^{a_n} \cdot z_n, d)$, intersecting in a bundle $H = X \times_{S^1} S^1 \times S^{2n-3}$. Siebenmanns sum formulae for Whitehead torsion (see e.g. [5]) implies in obvious notation that

$$f^* = f^*_A + f^*_C - f^*_H$$

(the sums computed in Wh (π) through the natural maps). However the S^1-equivariant map

$$S^1 \times S^{2n-3} \subset S^1 \times D^{2n-2} \to S^1$$

induces a bundle

$$X \times_{S^1} S^1 \times S^{2n-3} \to X \times_{S^1} S^1$$

with fibre S^{2n-3}, so by [11] f^*_H is 0 (because $2n-3$ is odd. Had we considered the attaching of an even S^1-cell we would get $f^*_H = 2f^*_C$ and thus f^*_C would be subtracted leading one to think of an S^1-CW-complex Euler characteristic).

To complete the proof we only need to show $f_C^* = p^{\#}_{(b_n)}$.
Since $X \times_{S^1} S^1 = X/(\mathbf{Z}/a_n\mathbf{Z})$ we get a diagram

$$\begin{array}{ccccc}
S^1 & \longrightarrow & S^1/(\mathbf{Z}/a_n\mathbf{Z}) = S^1 & \longrightarrow & L \\
\downarrow & & \downarrow & & \downarrow \\
X & \longrightarrow & X/(\mathbf{Z}/a_n\mathbf{Z}) & \longrightarrow & X \times_{S^1} L \\
\downarrow & & \downarrow & & \downarrow \\
B & \xrightarrow{1} & B & \longrightarrow & B
\end{array}$$

with a corresponding diagram of fundamental groups

$$\begin{array}{ccccc}
\mathbf{Z} & \xrightarrow{a_n} & \mathbf{Z} & \longrightarrow & \mathbf{Z}/m\mathbf{Z} \\
\downarrow & & \downarrow & & \downarrow \\
\pi & \longrightarrow & \pi_1(E/(\mathbf{Z}/a_n\mathbf{Z})) & \longrightarrow & \pi \\
\downarrow & & \downarrow & & \downarrow \\
\rho & \longrightarrow & \rho & \longrightarrow & \rho
\end{array}$$

The composite $\mathbf{Z} \xrightarrow{a_n} \mathbf{Z} \longrightarrow \mathbf{Z}/m\mathbf{Z}$ sends 1 to 1 so $\mathbf{Z} \longrightarrow \mathbf{Z}/m\mathbf{Z}$ must send 1 to b_n where $a_n \cdot b_n \equiv 1 \pmod{m}$. It follows by naturality of the algebraic S^1-transfer that $f_C^* = p^{\#}_{(b_n)}$ thus ending the induction step and the proof.

We now consider $F = S^3/Q(8)$, a simple example with nonabelian fundamental group. Let $F \to E \to B$ be an orientable bundle with fundamental group sequence $Q(8) \to \pi \to \rho \to 1$. If the kernel $Q(8) \to \pi$ is nontrivial the sequence is split and it is shown e.g. in [15] that $f^* : \mathrm{Wh}(\rho) \to \mathrm{Wh}(\pi)$ is trivial. Otherwise since $\mathbf{Z}/2\mathbf{Z}$ is the center of $Q(8)$ we may establish a pushout diagram

$$\begin{array}{ccccc}
\mathbf{Z}/2\mathbf{Z} & \longrightarrow & \bar{\pi} & \longrightarrow & \rho \\
\downarrow & & \downarrow & & \| \\
Q(8) & \longrightarrow & \pi & \longrightarrow & \rho
\end{array} \tag{7.1}$$

Let $h_s : \mathbf{Z}/4\mathbf{Z} \to Q(8)$, $s = i, j, k$ be the inclusions sending the generator to i, j and k respectively and consider the pushout diagram

$$\begin{array}{ccccc}
\mathbf{Z}/2\mathbf{Z} & \longrightarrow & \bar{\pi} & \longrightarrow & \rho \\
\downarrow{\scriptstyle 2} & & \downarrow & & \| \\
\mathbf{Z}/4\mathbf{Z} & \longrightarrow & \pi' & \xrightarrow{q} & \rho \\
\downarrow{\scriptstyle h_s} & & \downarrow{\scriptstyle \bar{h}_s} & & \| \\
Q(8) & \longrightarrow & \pi & \longrightarrow & \rho
\end{array} \tag{7.2}$$

Let $a: \mathrm{Wh}(\rho) \to \mathrm{Wh}(\pi)$ be the composite $\mathrm{Wh}(\rho) \xrightarrow{p^\#} \mathrm{Wh}(\bar{\pi}) \to \mathrm{Wh}(\pi')$ and $b_s: \mathrm{Wh}(\rho) \to \mathrm{Wh}(\pi)$, $s = i, j, k$ the composite $\mathrm{Wh}(\rho) \xrightarrow{q^\#} \mathrm{Wh}(\pi') \xrightarrow{(\bar{h}_s)_*} \mathrm{Wh}(\pi)$.

THEOREM 7.2. *The geometric* Wh-*transfer associated with the bundle* $S^3/Q(8) \to E \xrightarrow{f} B$ *is given by*

$$f^* = 2b_i + 2b_j + 2b_k - 4a$$

whenever $Q(8) \to \pi$ *is monic.*

Remark. Note that the diagram

$$\begin{array}{ccc} \mathrm{Wh}(\rho) & \xrightarrow{\quad q^\# \quad} & \mathrm{Wh}(\pi'), \\ & {\scriptstyle p^\#} \searrow \quad \swarrow {\scriptstyle \mathrm{Res}} & \\ & \mathrm{Wh}(\bar{\pi}) & \end{array}$$

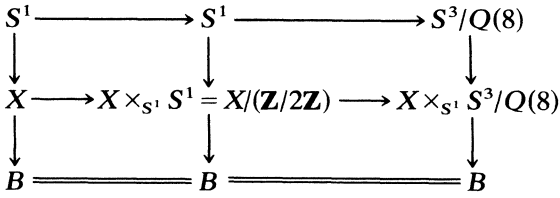
where Res is the restriction map, is commutative, so to prove $f^* = 0$ in a specific case it would suffice to prove $q^\# = 0$.

Proof of Theorem 7.2. We let S^1 act on the left of $S^3/Q(8)$ by $z \cdot [z_1, z_2] = [\xi z_1, \xi z_2]$ where $\xi^2 = z$. Letting S^1 act on the base-point gives the central element of $\pi_1(S^3/Q(8))$. The action has 6 singular orbits of the type $\mathbf{Z}/2\mathbf{Z}$ and the induced action of $S^1/(\mathbf{Z}/2\mathbf{Z})$ at the singular orbits define elements of $\pi_1(S^3/Q(8))$ up to conjugacy. We get i, j and k respectively at 2 points each. We have an S^1-CW structure on $S^3/Q(8)$ with 6 0-cells, 12 1-cells, and 8 2-cells, all 1 and 2-cells with 0 isotropy subgroup. We construct a principal S^1-bundle $S^1 \to X \to B$ with fundamental group data $\mathbf{Z} \to \mathbf{Z}/2\mathbf{Z} \to \bar{\pi} \to \rho$ and consider

$$\begin{array}{ccc} S^1 & \longrightarrow & S^3/Q(8) \\ \downarrow & & \downarrow \\ X & \longrightarrow & X \times_{S^1} S^3/Q(8) \\ \downarrow & & \downarrow \\ B & \longrightarrow & B \end{array}$$

The right hand bundle has the right fundamental group data and can thus be used to compute f^*. The decomposition of $S^3/Q(8)$ as an S^1-CW complex gives a decomposition of the bundle in S^1-bundles which we sum using Siebenmanns sum

formulae. The 0-cells give diagrams



with fundamental group diagram (7.2) which gives the contribution $2b_i + 2b_j + 2b_k$ to f^*. The contributions from the 1-cells are subtracted (see proof of Theorem 7.2) and the 2-cells added to give the result.

It has been essential to these two examples that the center of the fundamental group was "picked up" by an S^1-action so therefore has to be cyclic. To remedy this restriction slightly we consider $F = T^n$, the n-torus. Let $T^n \to E \to B$ be an orientable bundle with fundamental group data $\mathbf{Z}^n \to \pi \to \rho$. Consider $\pi_i = \mathrm{Cok}(\mathbf{Z}^i \to \pi)$ where $\mathbf{Z}^i \subset \mathbf{Z}^n$ as the first i factors. We get exact sequences $\mathbf{Z} \to \pi_{i-1} \xrightarrow{p_i} \pi_i$.

THEOREM 7.3. *The* Wh*-transfer* $f^* : \mathrm{Wh}(\rho) \to \mathrm{Wh}(\pi)$ *associated with the bundle* $T^n \to E \to B$ *is the composition*

$$\mathrm{Wh}(\rho) = \mathrm{Wh}(\pi_n) \xrightarrow{p_1^{\#} \circ p_2^{\#} \circ \cdots \circ p_n^{\#}} \mathrm{Wh}(\pi_0) = \mathrm{Wh}(\pi).$$

Proof. We may replace $T^n \to E \to B$ by a principal T^n-bundle $T^n \to X \to Y$ with the same fundamental group data and use this to compute f^*. The sequence of S^1-bundles

$$S^1 \to X/T^i \to X/T^{i+1}$$

now finishes the proof by referring to the main theorem.

If G is a compact connected Lie group and F a G-CW complex it is clear from the above that we can compute the geometric Wh-transfer for any bundle $F \to E \to B$ with G as structure group, since the inclusion of the maximal torus $T \subset G$ induces an epimorphism of fundamental groups and we thus may produce a bundle with the same fundamental group data and T as structure group, and then may proceed as in Theorems 7.1 and 7.2. This general result we prefer not to state since it uses the actual T-cellular structure of F and thus gets a very

complicated form. In case of the K_0-transfer and finite fundamental group of the base however we know the geometric S^1-transfer is 0 so in the inductive argument we keep adding zeroes to obtain e.g.

THEOREM 7.4. *Let M be a differentiable manifold with 0 Euler characteristic and $M \to E \to B$ a smooth bundle with structure group a compact connected Lie group. If $\pi_1 B$ is finite and B is finitely dominated then E is homotopy equivalent to a finite complex.*

Remark. In case $\chi(M) \neq 0$ the K_0-transfer has been computed by Ehrlich [7].

REFERENCES

[1] D. R. ANDERSON, *The Whitehead torsion of a fiber homotopy equivalence*, Mich. Math. J., *21* (1974), 171–180.
[2] ——, *The Whitehead torsion of the total space of a fiber bundle*, Topology *11* (1972), 179–194.
[3] H. BASS, A. HELLER and R. G. SWAN, *The Whitehead group of a polynomial extension*, Publ. Math. (I.H.E.S.) *22* (1964), 61–80.
[4] T. A. CHAPMAN, *Topological invariance of Whitehead torsion*, Amer. J. of Math. *96* (1974). 488–497.
[5] M. M. COHEN, A course in simple-homotopy theory, Springer Verlag, Berlin 1973.
[6] K. EHRLICH, *Fibrations and a transfer map in algebraic K-theory*, J. Pure Appl. Alg. *14* (1979), 131–136.
[7] ——, Finiteness obstructions of fiber spaces, Cornell University, Ph.D. thesis (1977).
[8] S. M. GERSTEN, *A Product Formula for Wall's obstruction*, Amer. J. Math. *88* (1966), 337–346.
[9] K. W. KWUN and R. H. SZCZARBA, *Product and Sum Theorems for Whitehead Torsion*, Ann. Math. *82* (1965), 183–190.
[10] S. MACLANE, Homology, Springer Verlag, Berlin 1963.
[11] H. J. MUNKHOLM, *Transfer on algebraic K-theory and Whitehead torsion for PL fibrations*, J. Pure Appl. Alg., to appear.
[12] H. J. MUNKHOLM and E. K. PEDERSEN, *On the Wall finiteness obstruction for the total space of certain fibrations*, Trans. Amer. Math. Soc. *261* (1980), 529–545.
[13] R. OLIVER, *SK_1 of finite grouprings*, I, Inv. Math., *57* (1980), 183–204.
[14] E. K. PEDERSEN, *Comparisons of Geometrically defined transfer homomorphisms*, in preparation.
[15] ——, *Universal Geometric Examples for Transfer maps in Algebraic K- and L-theory*, J. Pure Appl. Alg. (1981).
[16] E. K. PEDERSEN and L. TAYLOR, *The Wall finiteness obstruction for a fibration*, Amer. J. Math. *100* (1978), 887–896.
[17] H. BASS, Algebraic K-theory, W. A. Benjamin, New York 1968.

Dept. of Mathematics
Odense University
Campusvy 55
DK-5230 Odense

Received July 23, 1980/June 15, 1981

Comment. Math. Helvetici **56** (1981) 431–446 0010-2571/81/003431-16$01.50 + 0.20/0
© 1981 Birkhäuser Verlag, Basel

On J. H. C. Whitehead's aspherical question I

JOE BRANDENBURG and MICHEAL DYER

Abstract. A connected, finite two-dimensional CW-complex with fundamental group isomorphic to G is called a $[G, 2]_f$-complex. Let $L \lhd G$ be a normal subgroup of G. L has weight k if and only if k is the smallest integer such that there exists $\{l_1, \ldots, l_k\} \subset L$ such that L is the normal closure in G of $\{l_1, \ldots, l_k\}$. We prove that a $[G, 2]_f$-complex X may be embedded as a subcomplex of an aspherical complex $Y = X \cup \{e_1^2, \ldots, e_k^2\}$ if and only if G has a normal subgroup L of weight k such that $H = G/L$ is at most two-dimensional and def $G =$ def $H + k$. Also, if X is a *non-aspherical* $[G, 2]_f$-subcomplex of an aspherical 2-complex, then there exists a non-trivial superperfect normal subgroup P such that G/P has cohomological dimension ≤ 2. In this case, any torsion in G must be in P.

0. Introduction

A $[G, 2]_{(f)}$-*complex* X is any (finite) connected two dimensional CW-complex with fundamental group isomorphic to G. Sometimes we will abuse the notation and say $X \in [G, 2]_{(f)}$. Let X be a connected subcomplex of an $[H, 2]$-complex Y. J. H. C. Whitehead's question is this: if Y is aspherical, is X also aspherical? [W_1, p. 428].

The question seems very hard. We say that a group G satisfies the (finite) *Whitehead condition* ($G \in WC_{(f)}$) if any $[G, 2]$-complex X, which is the subcomplex of an aspherical (finite) 2-complex, is aspherical. Thus $G \in WC_{(f)}$ iff for any $[G, 2]_{(f)}$-complex X, either X is aspherical or, if not, then no $[H, 2]_{(f)}$-complex $Y > X$ is aspherical. The philosophy of this paper is to isolate properties of a group G which imply that $G \in WC$ or WC_f.

There are a number of results in this direction. W. Cockcroft [C, Theorem 2] showed that if G is one-relator group, then G has WC. One crucial observation he made was: *Let X be a $[G, 2]$-complex such that the Hurewicz homomorphism $h_2 : \pi_2 X \to H_2 X$ is non-zero. Then no* 2-complex $Y > X$ *is aspherical.* It follows that any group G which admits a $[G, 2]$-complex X which is the subcomplex of an aspherical 2-complex has $H_2 G$ free abelian. (Here $H_2 G$ means the homology of G with coefficients in the trivial module $\mathbf{Z}$.)

Let $\chi_{\min}(G, 2) = \min\{\chi(X) \mid X$ is a $[G, 2]_f$-complex$\}$. A complex X whose Euler characteristic $\chi(X)$ is minimal is called a *minimal* $[G, 2]_f$-complex. *For simplicity, all $[G, 2]$-complexes will have a single vertex, and this will be the base*

point for all homotopy groups. Any $[G, 2]$-complex has the (simple) homotopy type of such a complex.

We observe for later use that WC_f can be proved or disproved for a particular group G as follows: choose your favorite minimal $[G, 2]_f$-complex X and, if it is not aspherical, check that no $[H, 2]_f$-complex containing X is aspherical (see Lemma 1.4).

It follows from Cockcroft's result above that if $X < Y$, where $Y \in [H, 2]_f$ is aspherical, then X is a *minimal* $[G, 2]_f$-complex. To see this, we simply observe that if X is any non-minimal $[G, 2]_f$-complex, then $h_2 : \pi_2 X \to H_2 X$ is *not* zero. For let Y be a minimal $[G, 2]_f$-complex, with $\chi(Y) < \chi(X)$. Let $\Sigma_2(-) = \operatorname{im}\{h_2 : \pi_2(-) \to H_2(-)\}$ be the image of the Hurewicz homomorphism. Then, by a result of H. Hopf, $H_2 G \cong H_2 X/\Sigma_2 X \cong H_2 Y/\Sigma_2 Y$. Because $H_2 X$ and $H_2 Y$ are finitely generated free abelian groups with $\operatorname{rank}_Z H_2 Y < \operatorname{rank}_Z H_2 X$, we must have $\Sigma_2 X \neq 0$.

J. F. Adams' approach [A, p. 483] was to assume that a non-aspherical $X < Y = X \cup \{e_\alpha^2 \mid \alpha \in \mathcal{A}\}$, with Y aspherical, and to study $L = \ker\{\pi_1 X \to \pi_1 Y\}$. Adams proved that $H_1 L$ is a free abelian group and L is not transfinite metabelian; i.e., L has a non-trivial (normal) subgroup P which is perfect ($H_1 P = P^{ab} = 0$). This shows that *any* solvable group has WC.

In [Co], J. Cohen points out that Adams' perfect subgroup $P < L$ is actually *superperfect*; i.e., $H_2 P = 0$. He also shows [Co, Theorem 3] that if G a group of cohomological dimension 3 and type *FL* (that is, Z has a finite resolution by finitely generated free G-modules) such that $H_3 G = 0$, then G has WC_f.

In [GR, Theorem 4], M. Gutierrez and J. Ratcliffe show that if $X < Y (Y \in [H, 2]_f)$ which is aspherical, then X is aspherical if and only if the cohomological dimension of $G \leq 2$ and G has type *FL*.

In [H], J. Howie shows that any torsion element $x \in G (x^n = 1)$ is contained in a *finitely generated* perfect subgroup of L. We show that, in fact, all the torsion of G is contained in Adams' superperfect subgroup.

Finally, in his thesis [Be] W. Beckmann shows that locally finite groups have WC.

Specifically, we show the following

THEOREM 1. *Let X be a non-aspherical $[G, 2]$-subcomplex of an aspherical 2-complex. Then there is a nontrivial superperfect normal subgroup P (Adams) such that G/P has cohomological dimension ≤ 2.*

COROLLARY. *Any torsion in G must be in P.*

As a second result we characterize when one may add 2-cells to a $[G, 2]_f$-complex to obtain an aspherical complex. Let F_n denote a free group of rank n.

THEOREM 2. *Let X be a minimal $[G, 2]_f$-complex. One may add n 1-cells and k 2-cells to X to obtain an aspherical 2-complex if and only if $G * F_n$ has a normal subgroup L which is a free-crossed G-module of weight k* (see 3.2) *such that* (1) *there is an aspherical $[G/L, 2]_f$-complex and* (2) $\operatorname{def} G + n = \operatorname{def}(G/L) + k$.

COROLLARY. *Let X be a minimal $[G, 2]_f$-complex. One may add two-cells to X to obtain a finite contractible space if and only if weight $G = \operatorname{def} G$.*

The groups in the corollary are of interest because they are (higher) knot and link groups, according to a theorem of M. Kervaire [K]. These groups are all *E-groups* in the sense of [St_1], [St_2] and [B]. From this it follows that the derived series of G has many interesting properties; such as, each element G^α in the derived series for G is an E-group, and the derived length of G is severely restricted.

The paper is organized as follows. In section one we study complexes X for which the Hurewicz map $h_2 : \pi_2 X \to H_2 X$ is zero and reprove a crucial lemma of W. Cockcroft and R. Swan about minimal aspherical complexes. In section two we study necessary and sufficient conditions for the inclusion $X < Y$ to induce the zero map on the second homotopy groups. In section three we prove Theorem 2 and in section four, Theorem 1. We defer examples and applications of these results to a later paper.

To fix notation, let G be a group and let $\mathbf{Z}G$ be the integral group ring of G. Let IG denote the augmentation ideal, the kernel of the map $\epsilon : \mathbf{Z}G \to \mathbf{Z}$.

1. Cockcroft complexes and the Cockcroft–Swan lemma

DEFINITION 1.1. A connected CW-complex X is called *Cockcroft* if and only if the Hurewicz homomorphism $h : \pi_2 X \to H_2 X$ is trivial. A group G is *Cockcroft* if and only if some $[G, 2]$-complex is Cockcroft.

Note that any non-minimal $[G, 2]_f$-complex is not Cockcroft. It follows that any group G having $H_2(G; Z)$ *not* free abelian is not Cockcroft. It was shown in [C, lemma 1] that for any non-Cockcroft $[G, 2]$-complex X, *any* $Y = X \cup \{e_\alpha^2\}$ has $\pi_2 X \to \pi_2 Y$ not zero. It follows that non-*Cockcroft groups have WC.*

EXAMPLE 1.2. Any finitely generated one-relator group is Cockcroft. Let G

be the group presented by $\{x_1, \ldots, x_n; r\}$ and let X be the model associated with the presentation. Write $r = Q^q$, where Q is not a proper power in the free group $F(x_1, \ldots, x_n)$. By a theorem of Lyndon [L], $\pi_2 X \cong ZG(\bar{Q} - 1)$ as a left G-module, where $\bar{Q}$ is the image in G of Q under the natural projection $F \to G$. As the Hurewicz map $h : \pi_2 X \to H_2 X$ is given by restricting the augmentation $\epsilon : C_2 \tilde{X} = ZG \to C_2 X = Z$, we see that $h = 0$. Note that if X is a subcomplex of an aspherical two-complex Y, then X must be a Cockcroft $[G, 2]$-complex. This follows because if $[X \vee \bigvee S^1_\beta) \cup \{e^2_\alpha\} = Y$ is aspherical, then the Hurewicz homomorphism $\pi_2(X \vee \bigvee S^1_\beta) \to H_2(X \vee \bigvee S^1_\beta)$ is zero [C, lemma 1]. That X is Cockcroft is clear from the commutative diagram:

$$\begin{array}{ccc} \pi_2 X & \longrightarrow & H_2 X \\ \downarrow & & \downarrow \\ \pi_2(X \vee \bigvee S^1_\beta) & \to & H_2(X \vee \bigvee S^1_\beta). \end{array}$$

Observe that if X and X' are minimal $[G, 2]_f$-complexes, X is Cockcroft iff X is.

The following theorem characterizes in several different ways the property that X is Cockcroft.

PROPOSITION 1.3. *Let X be a $[G, 2]$-complex. The following are equivalent:*

(a) $h_2 : \pi_2 X \to H_2 X$ *is zero.*
(b) *The Hopf epimorphism* $H_2 X \to H_2 G$ *is injective.*
(c) *The natural inclusion* $H_3 G \to Z \otimes_G \pi_2 X$ *is surjective.*

Proof. This follows from the exact sequence of [D], which is just a fancy rewrite of two theorems of H. Hopf:

$$0 \to H_3 G \to Z \otimes_G \pi_2 X \xrightarrow{\bar{h}_2} H_2 X \to H_2 G \to 0,$$

where $\bar{h}_2$ is induced by h_2. □

We now prove the following key lemma of Cockcroft and Swan [CS, p. 197].

LEMMA 1.4. *Let X and X' be minimal $[G, 2]_f$-complexes. Then X is aspherical iff X' is.*

Proof. As X and X' have the same Euler characteristic, it follows from Schanuel's lemma that $\pi_2 X \oplus ZG^n \cong \pi_2 X' \oplus ZG^n$ for some integer $n > 0$. Then

$\pi_2 X = 0$ implies that

$$\pi_2 X' \rightarrowtail ZG^n \twoheadrightarrow ZG^n$$

is exact, which yields that $\pi_2 X' = 0$ by a theorem of I. Kaplansky. □

The *proof* of Lemma 1.4 clearly breaks down if X is an infinite, but Cockcroft, $[G, 2]$-complex. *We conjecture that the lemma is still true for such complexes.*

If G is finitely presented, then one may show from Lemma 1.4 that either *all* minimal $[G, 2]_f$-complexes are subcomplexes of finite aspherical 2-complexes or *none* are.

LEMMA 1.5. *Let G be a finitely presented group. Let X, X' be minimal $[G, 2]_f$-complexes and $Y = (X \vee \bigvee_{i=1}^{m} S_i^1) \cup \{e_1^2, \ldots, e_n^2\}$ be aspherical. Then one may add m one-cells and n two-cells to X' to obtain an aspherical complex.* □

2. Killing $\pi_2 X \to \pi_2 Y$ for X a subcomplex of Y

DEFINITION 2.1. For any subgroup $A < G$, let $K_A = \mathbf{Z}G \cdot IA$ be the left ideal in $\mathbf{Z}G$ generated by $\{a - 1 \mid a \in A\}$. Note that if A is a *normal* subgroup of G, then K_A is a *two-sided* ideal. In any case, $K_A = \ker\{\mathbf{Z}G \to \mathbf{Z}(G/A)\}$ induced by the coset function $G \to G/A$.

DEFINITIONS 2.2. Let M be any (left) submodule of a free G-module. The *Fox ideal of M, $F(M)$*, is the two-sided ideal in $\mathbf{Z}G$ generated by the coordinates of each element (of a generating set) of M. We say that a *subgroup $A < G$ kills M* if the Fox ideal $F(M)$ is contained in the kernel K_A. Note that $F(M)$ is independent of any chosen basis.

EXAMPLE. For a $[G, 2]$-complex X, G itself kills $\pi_2 X$ iff $F(\pi_2 X) \subset K_G = IG$. This happens iff the Hurewicz map $h_2 : \pi_2 X \to H_2 X$ is zero.

Now let $X \in [G, 2]$ be a subcomplex of an $[H, 2]$-complex Y. Let $\iota : X \to Y$ denote the inclusion map and $L = \ker \pi_1(\iota)$. If $C_* \tilde{X}$ is the cellular chain complex (considered as left G-modules) of the universal cover $\tilde{X}$ of X, let $R_X = \ker\{\partial_1 : C_1 \tilde{X} \to C_0 \tilde{X} = \mathbf{Z}G\}$ be a so-called relation module for G.

THEOREM 2.3. *The following are equivalent:*

(1) $i_\# : \pi_2 X \to \pi_2 Y$ *is zero,*
(2) *the Fox ideal* $F(\pi_2 X) \subset K_L$ (*L kills* $\pi_2 X$),

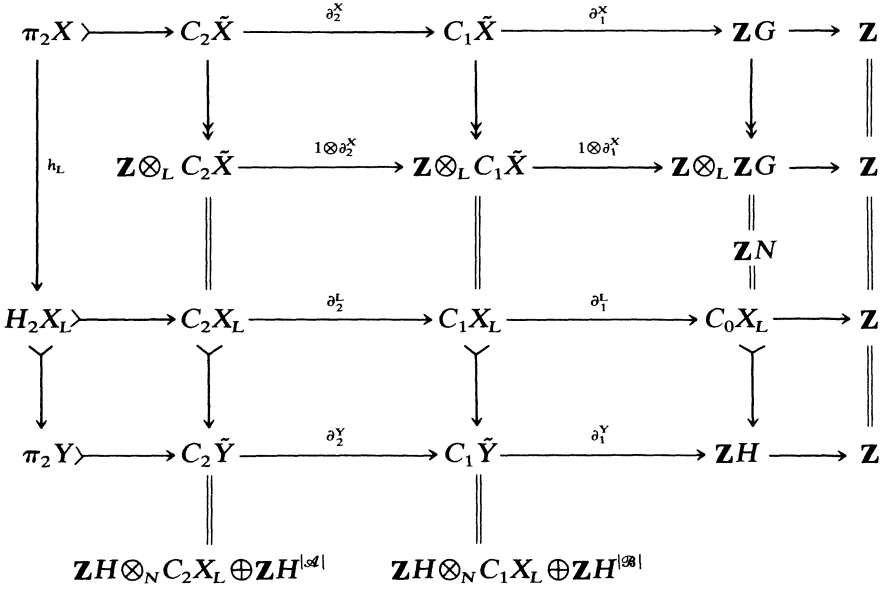
(3) *the Hurewicz homomorphism* $h_L : \pi_2 X \to H_2 X_L$ *is zero, where* X_L *is a covering of* X *corresponding to the subgroup* L.

(4) *The natural surjection* $\bar{\partial}_2 : C_2\tilde{X} \to R_X$ *induces an isomorphism* $\mathbf{Z}\otimes_L C_2\tilde{X} \to Z\otimes_L R$ *of free* G/L*-modules.*

Proof. Let $(\mathbf{Z}H)^{|\mathscr{A}|}$ denote $\oplus_{\alpha\in\mathscr{A}}(\mathbf{Z}H)_\alpha$. Consider the universal covering $\tilde{X}$ of X, the cellular chain complex $C_*(\tilde{X})$ of $\tilde{X}$ (viewed as free left G-modules and homomorphisms), and the cellular chain complexes $C_*X_L = \mathbf{Z}\otimes_L C_*\tilde{X} \to C_*\tilde{Y}$ (viewed as free G/L and H modules, respectively). Let $N = G/L$ denote the image of $\pi_1(i) : G \to H$ and $\eta : G \to G/L$ be the natural map. Also let

$$Y = \left(X \vee \bigvee_{\beta\in\mathscr{B}} S^1_\beta\right) \cup \{e^2_\alpha\}_{\alpha\in\mathscr{A}}.$$

Consider the following commutative diagram:



The chain map $i_* : C_2\tilde{X} = (\mathbf{Z}G)^m \to C_2\tilde{Y}$ factors as

$$\mathbf{Z}G^m = C_2\tilde{X} \twoheadrightarrow C_2X_L = \mathbf{Z}N^m \rightarrowtail \mathbf{Z}H^m \oplus \mathbf{Z}H^{|\mathscr{A}|} = C_2\tilde{Y}.$$

with the first map being $\oplus\mathbf{Z}\eta : \mathbf{Z}G^m \to \mathbf{Z}N^m$. Hence, the kernel of $i_\# : \pi_2 X \to \pi_2 Y$ is $\pi_2 X \cap (K_L)^m$. Also, $H_2X_L \rightarrowtail \pi_2 Y$ (it is a direct summand as Z-modules).

So $i_{\#}: \pi_2 X \to \pi_2 Y$ is zero if and only if $h_L: \pi_2 X = \pi_2 X_L \to H_2 X_L$ is zero. This happens if and only if $\pi_2 X \subset K_L^m$, which in turn is true if and only if $F(\pi_2 X) \subset K_L$.

In order to prove (4)$\Leftrightarrow$(1), consider the following exact sequences (see [D]).

$$\begin{array}{ccccc} \mathbf{Z}\otimes_L \pi_2 X & \xrightarrow{\bar{h}_L} & H_2 X_L & \twoheadrightarrow & H_2 L \\ \| & & \downarrow & & \downarrow \\ \mathbf{Z}\otimes_L \pi_2 X & \longrightarrow & \mathbf{Z}\otimes_L C_2 \tilde{X} & \twoheadrightarrow & \mathbf{Z}\otimes_L R. \end{array}$$

It is easily shown that both squares commute. Also

$$\pi_2 X \xrightarrow{h_L} H_2 X_L \quad \text{factors as} \quad \pi_2 X \longrightarrow \mathbf{Z}\otimes_L \pi_2 X \xrightarrow{\bar{h}_L} H_2 X_L.$$

Hence

$$\begin{aligned} \pi_2 X \xrightarrow{0} \pi_2 Y &\Leftrightarrow \pi_2 X \xrightarrow{0} H_2 X_L \\ &\Leftrightarrow \mathbf{Z}\otimes_L \pi_2 X \xrightarrow{0} H_2 X_L \\ &\Leftrightarrow \mathbf{Z}\otimes_L \pi_2 X \xrightarrow{0} \mathbf{Z}\otimes_L C_2 \tilde{X}. \quad \square \end{aligned}$$

Note 2.4. The proof of Theorem 2.3 shows that conditions (2), (3), and (4) are equivalent for any (not necessarily normal) subgroup $L \leq G$, provided we restate (4) as an isomorphism of (not necessarily free) G-modules. In fact, it is clear that (2)–(4) are *hereditary* in the sense that, if they are true for some subgroup $L \leq G$, then they hold for any subgroup $M \leq G$ containing L.

3. Subcomplexes of aspherical complexes

DEFINITION 3.1. Let G be a (finitely presented) group. G is *at most (finitely) two-dimensional* if and only if there exists an aspherical $[G, 2]_{(f)}$-complex.

For any group G and element $g \in G$, denote the image of g in G^{ab} by $\bar{g}$.

DEFINITION 3.2. Let L be a normal subgroup of G ($L \lhd G$) with quotient

H. L is said to be a *free crossed G-module of weight k* ($k \leq \infty$) if and only if there exists elements $\{g_1, \ldots, g_k\} \subset L$ such that L is the normal closure $\langle\langle g_1, \ldots, g_k \rangle\rangle_G$ of $\{g_i\}$ in G, $H_1 L$ is a free H-module with basis $\{\bar{g}_1, \ldots, \bar{g}_k\}$, and $H_2 L = 0$.

It is a very nice theorem of J. Ratcliffe [R, Theorem 2.2] that this is equivalent to the usual definition of a free crossed G-module (in this setting). The normal generators $\{g_1, \ldots, g_k\}$ of L are called a *basis* for the free crossed module L.

An interesting special case is when G is a free crossed *G-module* of weight k. Examples of weight 1 self free crossed modules are knot groups. By a theorem of M. Kervaire [K], any finitely presented self free crossed module of weight k is the fundamental group of a k-link of 3-spheres embedded in S^5.

Note that if L is a free crossed G-module of weight k, then L is a free crossed L-module of weight $k \cdot |G/L|$. Furthermore, if L is a free crossed L-module of weight k and H is any group, then, for $G = L * H$, the normal closure N of L (in G) is a free crossed G-module of weight k. To see this, notice that N is the free product $*_{h \in H} hLh^{-1}$ in G and that $1 \to N \to G \to H \to 1$ is a split extension. $H_1 L \cong \mathbf{Z}^k$, so $H_1 N \cong \mathbf{Z}H^k$; $H_2 L = 0$ implies $H_2 N = 0$. If the normal closure of $\{l_1, \ldots, l_k\}$ in L is equal to L, then $\langle\langle\{l_1, \ldots, l_k\}\rangle\rangle_G = N$.

As another example, one may show the following proposition.

PROPOSITION 3.3. *Let G be a 1-relator group with presentation $\{x_1, \ldots, x_n; Q^q\}$, where Q is not a proper power. Let $e_{x_i}(Q)$ denote the exponent sum of Q with respect to x_i. Then G is a free crossed G-module if and only if $[q = 1$ and $E = \gcd\{e_{x_i}(Q)\} = 1]$ if and only if $[H_1 G \cong Z^{n-1}]$.* □

Note. Let G be a finitely generated 1-relator group. Any two 1-relator presentations of G have the same number of generators. This follows because the models associated with both presentations are Cockcroft (Example 1.2) and are therefore minimal.

Let def G denote the *deficiency* of the finitely presented group G. The following theorem characterizes when one may add finitely many one-cells and two-cells to a $[G, 2]_f$-complex X to obtain an aspherical 2-complex. Let F_l denote a free group of rank l.

THEOREM 3.4. *Let X be a minimal $[G, 2]_f$-complex. One may add l 1-cells and k 2-cells to X to obtain an aspherical two-complex Y if and only if* (1) *there exists a normal subgroup $L < G * F_l$ which is a free crossed $G * F_l$-module of weight k having $H = (G * F_l)/L$ at most finitely two-dimensional and* (2) $\operatorname{def} G + l = \operatorname{def} H + k$.

Note that this theorem is true whether X is aspherical or not. For another equivalent condition see Theorem 3.6.

Proof. ($\Rightarrow$) Suppose $X < Y = (X \vee \bigvee_{i=1}^{l} S_i^1) \cup \{e_1^2, \ldots, e_k^2\}$ and Y is aspherical. Consider the homotopy sequence for the pair $(Y, \bar{X})$, where $\bar{X} = X \cup Y^{(1)}$:

$$\begin{array}{ccccccccccc} \pi_3(Y) & \to & \pi_3(Y, \bar{X}) & \to & \pi_2(\bar{X}) & \to & \pi_2(Y) & \to & \pi_2(Y, \bar{X}) & \xrightarrow{\partial} & \pi_1(\bar{X}) \to \pi_1(Y) \to 0. \\ \| & & & & \| & & \| & & & & \| \quad \| \\ 0 & & & & 0 & & 0 & & & & G * F_l \quad H \end{array}$$

The group $\pi_2(Y, \bar{X})$ is a free crossed $\pi_1(\bar{X})$-module on the characteristic maps for the k added two cells. We let $L = \operatorname{im} \partial$. Then $H = G/L$ is at most 2-dimensional. Because $\chi(X) = \chi_{\min}(G, 2)$, $\chi(Y) = \chi_{\min}(H, 2)$ and $\chi(X) + k - l = \chi(Y)$ we have (as $\chi(X) = 1 - \operatorname{def} G$) $k + \operatorname{def} H = \operatorname{def} G + l$.

($\Leftarrow$) Let X be a minimal $[G, 2]_f$-complex and identify $\pi_1 X$ with G. Let $\{g_1, \ldots, g_k\}$ be a basis for $L < G * F_l$ as a free crossed $G * F_l$-module. Attach $e_1^2, \ldots, e_k^2$ to $\bar{X} = X \vee \bigvee_{i=1}^{l} S_i^1$ using maps $\alpha_i : S_i^1 \to \bar{X}^{(1)}$ which represent $g_i \in G * F_l$ ($i = 1, \ldots, k$). Then $X < Y = \bar{X} \cup \{e_1^2, \ldots, e_k^2\}$ has $\pi_1 Y = H$ at most a finitely two-dimensional group. Thus there is an $[H, 2]_f$-complex W which is aspherical. Because $\operatorname{def} H = \operatorname{def} G - k + l = 1 - \chi(Y) = 1 - \chi(W)$, we have $\chi(Y) = \chi(W) = \chi_{\min}(H, 2)$. Therefore, Y is aspherical by Lemma 1.4. □

Note. One sees from the proof that really only the following was used:
$G * F_l$ has a normal subgroup L of weight k over $G * F_l$ such that
(a) $G * F_l / L$ is at most finitely 2-dimensional and
(b) $\operatorname{def}[(G * F_l)/L] + k = \operatorname{def} G + l$.
That L is a free crossed $G * F_l$-module of weight k is a *consequence* of the above statement.

COROLLARY 3.5. *Let X be a minimal $[G, 2]_f$-complex. One may add k two-cells to X to obtain a contractible space if and only if* $\det G = k =$ *weight G. This is true iff G is a free crossed G-module* (= *higher dimensional link group*) *of weight k which is Cockcroft.*

Proof. The first statement follows by specializing Theorem 3.1 to $l = 0$ and $H = \{1\}$. To see the second, we observe that a group G with $H_1 G \cong \mathbf{Z}^k$ and which is Cockcroft has $H_2 X \cong H_2 G \cong \mathbf{Z}^{k - \operatorname{def} G}$. So $H_2 G = 0$ implies $\operatorname{def} G = k$. A similar argument yields the converse. □

We would ask, more generally, *what does the fundamental group G of a subcomplex X of a finite contractible space Y look like*? By the above corollary, we see that $G * F_l$ is a higher dimensional link group with $\operatorname{def}(G * F_l) = \operatorname{def} G + l = \text{weight}(G * F_l)$. Does this imply that $\operatorname{def} G = \text{weight } G$? It is easy to see that such groups are E-groups (see [B], 123–130, for facts about E-groups).

We also have (using the same geometric techniques)

THEOREM 3.6. *Let X be a Cockcroft $[G, 2]$-complex. One may add $(k \leq \infty)$ 2-cells to X to obtain an aspherical two-complex Y if and only if there exists a free crossed G-module $L \lhd G$ (of weight k) which kills $\pi_2 X$.*

Proof. From the exactness of

$$\pi_2 X \to \pi_2 Y \to \pi_2(Y, X) \to \pi_1 X$$

we see that $\pi_2 Y = 0$ if and only if $\pi_2 X \to \pi_2 Y$ is zero and $\pi_2(Y, X) \to \pi_1 X$ is monic. But, by the theorem of J. H. C. Whitehead [W_2], $\pi_2(Y, X)$ is a free crossed $\pi_1 X = G$ module of weight k. □

Thus, a counter example to the Whitehead conjecture would arise if there is a group G with a "large" free crossed module L as a subgroup (in the sense that L kills $\pi_2 X$ for some Cockcroft $[G, 2]$-complex X).

4. Extending Adams' theorem

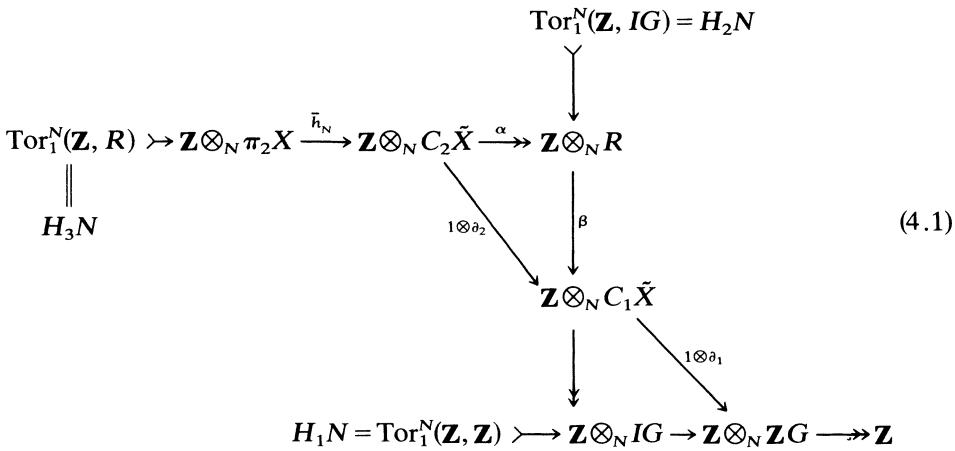
In this section we prove Theorem 1 of the introduction.

LEMMA 4.1. *Let X be a $[G, 2]$-complex and $\partial_2 : C_2\tilde{X} \to C_1\tilde{X}$ be the second boundary operator considered as a left G-module homomorphism of free G-modules $C_i\tilde{X}$. Let N be a subgroup of G. Then $1 \otimes \partial_2 : \mathbf{Z} \otimes_N C_2\tilde{X} \to \mathbf{Z} \otimes_N C_1\tilde{X}$ is a monomorphism if and only if $F(\pi_2 X) \subset K_N$ and $H_2 N = 0$. This happens iff $H_2(X_N) = 0$.*

Proof. Recall that R is the image of $\partial_2^X : C_2\tilde{X} \to C_1\tilde{X}$ and $IG = \operatorname{im} \partial_1^X$. From the exact sequences

$$\pi_2 X \rightarrowtail C_2\tilde{X} \twoheadrightarrow R, \qquad R \rightarrowtail C_1\tilde{X} \twoheadrightarrow IG, \quad \text{and} \quad IG \rightarrowtail \mathbf{Z}G \twoheadrightarrow \mathbf{Z},$$

we obtain the following sequences:



(4.1)

Note that the triangles commute and that the vertical and horizontal sequences are exact. Clearly $\ker(1\otimes\partial_2) = \alpha^{-1}H_2N$. So $H_2N = 0$ yields $\ker(1\otimes\partial_2) = \ker\alpha = \mathrm{im}\,\bar{h}_N = 0$, if $F(\pi_2X) \subset K_N$. Similarly $\ker(1\otimes\partial_2)$ is zero yields $\alpha^{-1}H_2N = 0$ which implies $\alpha^{-1}(0) = \mathrm{im}\,\bar{h}_N = 0$ and $H_2N = 0$. □

DEFINITION 4.2 [St_1]. A group G is called an *E-group* if H_1G is torsion free and the trivial G-module $\mathbf{Z}$ has a projective G-resolution

$$\cdots \rightarrow P_n \rightarrow \cdots \rightarrow P_2 \xrightarrow{\partial_2} P_1 \rightarrow P_0 \twoheadrightarrow \mathbf{Z}$$

such that the homomorphism $1\otimes_G \partial_2 : \mathbf{Z}\otimes_G P_2 \rightarrow \mathbf{Z}\otimes_G P_1$ is *injective.*

It follows that if G is an E-group, then $H_2(G) = 0$.

For a given group G, we define P_1G to be the *maximal perfect subgroup of the group G.* It is *uniquely defined* as the group generated by the family of all perfect subgroups of G. This subgroup is perfect because the group generated by any family of perfect subgroups is perfect. Because the normal closure of a perfect group is perfect, P_1G is a *normal subgroup* of G. P_1 is clearly a functor from the category of groups and homomorphisms to the category of perfect groups and homomorphisms.

There is another way to define P_1G. Let $\{G^\alpha \mid \alpha \text{ ordinal}\}$ denote the *derived series*: $G^\alpha = (G^{\alpha-1})'$ for α not a limit ordinal, $G^\alpha = \bigcap_{\beta<\alpha} G^\beta$ for α a limit ordinal. This sequence terminates [Dr, p. 20] at a perfect group, and since G^α contains any perfect subgroup of G, it terminates at P_1G.

The following theorem may have been known to J. F. Adams. It certainly follows from his techniques, when applied to the derived series of L. See [A] and [St_1].

THEOREM 4.3. *Suppose X is a non-aspherical $[G, 2]$-complex such that $Y = X \cup \{e_\alpha^2 \mid \alpha \in \mathcal{A}\}$ is aspherical. Let $L = \ker\{\pi_1 X \to \pi_1 Y = H\}$. Then the maximal perfect subgroup $P_1 L$ of L is superperfect, kills $\pi_2 X$, and is non-trivial.*

Proof. We first observe that L is an E-group because

$$\begin{array}{ccc} 1 \otimes_L \partial_2^X : \mathbf{Z} \otimes_L C_2 \tilde{X} & \to & \mathbf{Z} \otimes_L C_1 \tilde{X} \\ \| & & \| \\ C_2 X_L & & C_1 X_L \end{array}$$

is monic (Lemma 4.1) and $H_1 L \cong \mathbf{Z}H^{|\mathcal{A}|}$. By Theorem A(i) of [$\mathrm{St}_1$] each term of the derived series L^α of L is an E-group; hence $P_1 L$ is an E-group, therefore superperfect. In fact, the argument of the theorem cited above shows that

$$1 \otimes_{P_1 L} \partial_2^X : \mathbf{Z} \otimes_{P_1 L} C_2 \tilde{X} \to \mathbf{Z} \otimes_{P_1 L} C_1 \tilde{X}$$

is monic, hence $P_1 L$ kills $\pi_2 X$ by Lemma 4.1. $P_1 L$ is non-trivial because, if it were trivial then $P_1 L$ kills $\pi_2 X$ implies $F(\pi_2 X) \subset K_{P_1 L} = 0$. But $F(\pi_2 X) = 0$ iff $\pi_2 X = 0$, which contradicts the assumption that X was non-aspherical. □

A group G has cohomological dimension $\leq n$ ($\mathrm{cd}\, G \leq n$) if $H^i(G; M) = 0$ for all $i > n$ and all $\mathbf{Z}G$-modules M. Equivalently $\mathrm{cd}\, G \leq n$ if and only if the trivial G-module $\mathbf{Z}$ has an $\mathbf{Z}G$-protective resolution of length n:

$$0 \to P_n \to \cdots \to P_1 \to P_0 \to \mathbf{Z} \to 0. \qquad (*)$$

A group G has type *FP* (*FL*) if and only if G has a projective resolution $(*)$ of finite length with each P_i (free) of finite rank.

If G has type *FL*, we define the (naive) Euler characteristic $\chi(G) = \sum_{i=0}^n (-1)^i \operatorname{rank}_{ZG} P_i$. In this case, let $b_i G = \operatorname{rank}_Z H_i G$. Standard arguments show that $\chi(G) = \sum_{i=0}^n (-1)^i b_i G$ as well.

THEOREM 4.4. *Let X be a $[G, 2]$-complex and P be any superperfect normal subgroup of G such that P kills $\pi_2 X$. Then G/P has cohomological dimension ≤ 2 over Z. Furthermore, if X is a minimal $[G, 2]_f$-complex, then G/P has type FL and $\chi_{\min}(G, 2) = \chi(G/P)$.*

Proof. Consider the chain complex $C_*\tilde{X}$. In diagram 4.1 with $P = N$, we see that $H_1P = H_2P = 0$ together with $\bar{h}_P = 0$ (if and only if $K_P \supset F(\pi_2 X)$) shows that the sequence

$$0 \to \mathbf{Z}\otimes_P C_2\tilde{X} \to \mathbf{Z}\otimes_P C_1\tilde{X} \to \mathbf{Z}\otimes_P \mathbf{Z}G \to \mathbf{Z} \to 0$$

is an exact sequence of $\mathbf{Z}G$-modules. Because P is normal, $\mathbf{Z}\otimes_P \mathbf{Z}G \cong \mathbf{Z}(G/P)$ as (G/P)-modules. If X is a $[G, 2]_f$-complex, then clearly G/P has type FL and $\chi(G/P) = \chi(X) = \chi_{\min}(G, 2)$. □

After proving Theorem 4.4, we noticed that R. Strebel had proved a similar result for E-groups [St_1]. However, groups arising as the fundamental group of a subcomplex of an aspherical complex are *not* necessarily E-groups, as the second homology group is not necessarily zero (it is free abelian). The results do not imply one another, even though the basic trick is the same.

THEOREM 4.5. *Let X be any non-aspherical $[G, 2]$-complex and $X < Y$, an aspherical $[H, 2]$-complex. Then there exists a family of distinct non-trivial normal superperfect subgroups $P_i \lhd G$, $i \in I$, such that* cd $G/P_i \leq 2$ *for $i \in I$ and such that the smallest (normal) subgroup $P = \langle P_i \mid i \in I\rangle$ containing all P_i kills $\pi_2 X$. Hence, $P_1 G$ kills $\pi_2 X$.*

Proof. $X < X \cup Y^{(1)} = X \vee \bigvee S^1_\alpha = \bar{X} < Y = \bar{X} \cup \{e^2_\beta\}$. By theorem 4.3 there is a superperfect normal subgroup $\bar{P} \neq 1$ in $G * F$, where F is a free group isomorphic to $\pi_1(\bigvee S^1_\alpha)$. Also $\bar{P}$ kills $\pi_2\bar{X}$. Hence, by 4.4, cd $G * F/\bar{P} \leq 2$.

By Kuros' theorem, we have $\bar{P} = *_u (uGu^{-1} \cap \bar{P})$ for certain $u \in G * F$. The group $\bar{P}$ is superperfect implies that each $uGu^{-1} \cap \bar{P}$ is superperfect. Let $P_u = u^{-1}(uGu^{-1} \cap \bar{P})u$. Each P_u is a superperfect normal subgroup of G. The group $\bar{P} \neq 1$ implies that *some* of the $P_u \neq 1$. Choose the family $\{P_i\}$ to be those $P_u \neq 1$.

Consider the following diagram:

$$\begin{array}{ccc} G * F & \twoheadrightarrow & G * F/\bar{P} \\ \uparrow & & \uparrow \\ uGu^{-1} & \twoheadrightarrow & uGu^{-1}/\bar{P} \cap uGu^{-1} \end{array}$$

Thus $uGu^{-1}/\bar{P} \cap uGu^{-1} \approx G/P_u$ has cohomological dimension ≤ 2 for each u. Note that if *any* $\bar{P} \cap uGu^{-1} = 1$, then G itself has cohomological dimension ≤ 2.

Let $F_G(M)$ denote the 2-sided ideal *in* $\mathbf{Z}G$ generated by the coordinates of elements of the G-module $M \subset (\mathbf{Z}G)^\alpha$. We know that $F_{G*F}(\pi_2\bar{X}) \subset K_{\bar{P}} = \mathbf{Z}(G * F) \cdot I\bar{P}$, where $\pi_2\bar{X} \cong \mathbf{Z}(G * F) \otimes_G \pi_2 X$, by Theorem 4.3. It follows that

$F_{G*F}(\pi_2\bar{X}) = F_{G*F}(\pi_2 X)$, with $\pi_2 X$ considered as a $G*F$-module via the projection $\eta : G*F \to G$. Notice that $P = \eta(\bar{P})$. The surjection $\mathbf{Z}\eta : \mathbf{Z}(G*F) \to \mathbf{Z}(G)$ clearly carries $K_{\bar{P}} = \mathbf{Z}(G*F)\cdot I\bar{P}$ onto $K_P = \mathbf{Z}(G)\cdot IP$. Also $F_{G*F}(\pi_2\bar{X}) = F_{G*F}(\pi_2 X)$ is carried onto $F_G(\pi_2 X)$. Thus $F_{G*F}(\pi_2 X) \subset K_{\bar{P}}$ implies $F_G(\pi_2 X) \subset K_P$ and we are done. □

For the next corollary let X be a $[G, 2]$-complex which is not aspherical, but which is a subcomplex of an aspherical $[H, 2]$-complex. Thus, there must exist a non-trivial superperfect subgroup $P \lhd G$ such that cd $G/P \leq 2$. Because groups of finite cohomological dimension are torsion free, we have

COROLLARY 4.6. *Any element $g \in G$ such that $g^n \in P$ $(n \geq 1)$ must be in P. In particular, the torsion of G is contained in P.* □

In [B, p. 122], R. Bieri shows that the center of a non-abelian group of cohomology dimension ≤ 2 is cyclic. The exact sequence $P \rightarrowtail G \twoheadrightarrow \bar{G} = G/P$ induces a monomorphism

$$\mathfrak{Z}G/(P \cap \mathfrak{Z}G) \rightarrowtail \mathfrak{Z}(G/P)$$

($\mathfrak{Z}G$ is the center of G). If G/P is non-abelian, then $\mathfrak{Z}(G/P)$ is 0 or $\mathbf{Z}$; if G is finitely generated, $\mathfrak{Z}(G/P) = 0$, $\mathbf{Z}$, or $\mathbf{Z}\oplus\mathbf{Z}$ (this last occurs only if $G/P = \mathbf{Z}\oplus\mathbf{Z}$ is abelian).

COROLLARY 4.7. *Let G be a finitely presented group, X be a minimal $[G, 2]_f$-complex, and P be a superperfect normal subgroup of G with the cohomological dimension of $G/P \leq 2$. Then $\mathfrak{Z}G/(P \cap \mathfrak{Z}G) = 0$, $\mathbf{Z}$, or $\mathbf{Z}\oplus\mathbf{Z}$, with this last group occurring only if $\bar{G} = G/P$ is abelian. If* def $G \geq 1$ *and P doesn't kill $\pi_2 X$ or if P kills $\pi_2 X$ and* def $G \neq 1$, *then $\mathfrak{Z}G \subset P$.*

Proof. First, we assume that P kills $\pi_2 X$ and that def $G \neq 1$. Then, by Theorem 4.4, $\bar{G}$ has type FL and $\chi(\bar{G}) = \chi(X) = 1 - \text{def } G$. The deficiency of $G \neq 1$ implies that $\chi(\bar{G}) \neq 0$. Then, by corollary 3.6 of [S], we see that $\mathfrak{Z}(\bar{G})$ is trivial. Hence P contains $\mathfrak{Z}(\bar{G})$.

We assume that def $G \geq 1$ and that P does not kill $\pi_2 X$. Let $R_i = \ker\{\mathbf{Z}\otimes_P C_i\tilde{X} \to \mathbf{Z}\otimes_P C_{i-1}\tilde{X}\}$ $(i = 1, 2)$. The cohomological dimension of $\bar{G} \leq 2$ implies that R_1 is a projective $\bar{G}$-module. Because P is superperfect, we have an exact sequence

$$R_2 \rightarrowtail \mathbf{Z}\otimes_P C_2\tilde{X} \twoheadrightarrow R_1.$$

This shows that R_1 and R_2 are both finitely generated projective $\bar{G}$-modules. Now at this point in the proof, we must use the Euler characteristic of a group defined by J. Stallings in [S]. The rank of a finitely generated projective $\bar{G}$-module Q is a certain element rQ in the free abelian group T on the set of conjugacy classes of $\bar{G}$. Then ρQ is defined to be the coefficient of [1] in rQ. Accordingly, $\chi(\bar{G}) = \oplus_{i=0}^{2}(-1)^i \cdot \rho(\mathbf{Z}\otimes_P C_i\tilde{X}) - \rho R_2 = \chi_{\min}(G,2) - \rho R_2 = 1 - \operatorname{def} G - \rho R_2$. It follows from proposition 1 of [DV] that $\rho R_2 \geq 0$ and $\rho R_2 = 0$ iff $R_2 = 0$. Now P does *not* kill $\pi_2 X$ implies that $R_2 \neq 0$. Thus $\rho R_2 > 0$. Hence the deficiency of $G \geq 1$ implies that $\chi(\bar{G}) < 0$ and the result again follows from corollary 3.6 of [S]. □

We would like to thank the referee for simplifying the hypotheses of 4.7.

One may show that all the higher centers $\mathfrak{Z}^n G \subset P$ as well. To see that $\mathfrak{Z}^2 G \subset P$, notice that the hypotheses of 4.7 imply that $\mathfrak{Z}\bar{G} = 1$. Then the following diagram commutes:

$$\begin{array}{ccccc}
\mathfrak{Z}G & \rightarrowtail & P & \twoheadrightarrow & P/\mathfrak{Z}G \\
\| & & \downarrow & & \downarrow \\
\mathfrak{Z}G & \rightarrowtail & G & \overset{\eta}{\twoheadrightarrow} & G/\mathfrak{Z}G. \\
 & & \downarrow & & \downarrow \\
 & & \bar{G} & = & \bar{G}
\end{array}$$

Now $\mathfrak{Z}\bar{G} = 1$ implies that $\mathfrak{Z}(G/\mathfrak{Z}G) \subset P/\mathfrak{Z}G$ and hence that $\mathfrak{Z}^2 G = \eta^{-1}\mathfrak{Z}(G/\mathfrak{Z}G) \subset P$.

BIBLIOGRAPHY

[A] J. F. ADAMS, *A new proof of a theorem of W. H. Cockcroft*, J. London Math. Soc., *30* (1955), 482–488.

[B] R. BIERI, Homological dimension of discrete groups, Queen Mary College Notes, 1976, University of London, London E.1.

[Be] W. BECKMANN, Completely Aspherical 2-Complexes, Thesis, Cornell University, 1980.

[C] W. H. COCKCROFT, *On two-dimensional aspherical complexes*, Proc. London Math. Soc., *4* (1954), 375–384.

[CS] —— and R. SWAN, *On the homotopy type of certain two-dimensional complexes*, Proc. London Math. Soc., *11* (1961), 194–202.

[Co] J. COHEN, *Aspherical 2-complexes*, J. of Pure and Applied Algebra *12* (1978), 101–110.

[DV] E. DYER and A. VASQUEZ, *An invariant for finitely generated projectives over* $\mathbf{Z}[G]$, J. Pure and Applied Algebra, 7 (1976) 241–248.

[D] M. DYER, *On the second homotopy module of two-dimensional* CW*-complexes*, Proc. AMS, *55* (1976), 400–404.

[Dr] E. DROR, *A generalization of the Whitehead theorem*, Symposium on algebraic topology, vol. 249, Lecture Notes in Math., Springer-Verlag, 13–22.

[GR] M. A. GUTIERREZ and J. G. RATCLIFFE, *On the second homotopy group*, Quart, J. Math., *32* (1981), 45–55.

[H] J. HOWIE, *Aspherical and acyclic 2-complexes*, J. London Math. Soc., *20* (1979), 549–558.

[K] M. KERVAIRE, On higher dimensional knots, *Differential and Combinatorial Topology*, pp. 105–119, Princeton Univ. Press, Princeton, N.J., 1965.

[L] R. LYNDON, *Cohomology theory of groups with a single defining relation*, Ann. Math., *52* (1950), 650–665.

[MKS] W. MAGNUS, A. KARRASS, and D. SOLITAR, *Combinatorial Group Theory*, Interscience, New York, 1966.

[R] J. RATCLIFFE, *Free and projective crossed modules*, J. London Math. Soc., *22* (1980), 66–74.

[Se] J.-P. SERRE, *Cohomologie Des Groupes Discrets, Prospects in Mathematics*, Annals of Math. Studies, *70*, 1971.

[S] J. STALLINGS, *Centerless groups – an algebraic formulation of Gottlieb's theorem*, Topology, *4* (1965), 129–134.

[St_1] R. STREBEL, *Homological methods applied to the derived series of groups*, Comment. Math. Helv., *49* (1974), 302–332.

[St_2] ——, *Die Reihe der Derivierten von E-Gruppen*, Dissertation an der Eidgenössischen Technischen Hochschule Zürich (1973).

[W_1] J. H. C. WHITEHEAD, *On adding relations to homotopy groups*, Annals of Math. *42* (1941), 409–428.

[W_2] ——, *Combinatorial homotopy* II, Bull. American Math. Soc., *55* (1949), 453–496.

George Mason University
Fairfax, Virginia 22030

and

University of Oregon
Eugene, Oregon 97403

Received February 12/June 11 1981

Comment. Math. Helvetici **56** (1981) 447–464 0010-2571/81/003447-18$01.50 + 0.20/0
© 1981 Birkhäuser Verlag, Basel

Permutation modules and projective resolutions

JAMES HOWIE[1] and HANS RUDOLF SCHNEEBELI

1. Introduction

We adopt the convention of [2] that a *group-pair* $(G, \boldsymbol{S})$ consists of a group G and an indexed family $\boldsymbol{S} = \{S_i\}_I$ of subgroups of G (possibly with repetitions). For the most part no explicit indexing is referred to; we then write $\boldsymbol{S}$ as a shorthand for $\{S_i\}_I$. We let $\mathbf{Z}G/\boldsymbol{S}$ denote the left $\mathbf{Z}G$-module $\oplus_I \mathbf{Z}G/S_i$. Any module of the form $\mathbf{Z}G/\boldsymbol{S}$ is called a *permutation module.*

We study projective resolutions $\mathscr{P} \twoheadrightarrow \mathbf{Z}$ and we deal with special splittings of the kernels of the boundary map in some dimension n. We focus attention on such splittings where a $\mathbf{Z}G$-permutation module appears as a direct summand. The motivation for this set-up stems from our investigation [6], where a group is termed of finite quasi-projective dimension in case for some $\mathbf{Z}G$-projective resolution $\mathscr{P} \twoheadrightarrow \mathbf{Z}$ and some integer $n > 1$, the kernel of ∂_n splits into a $\mathbf{Z}G$-projective and a $\mathbf{Z}G$-permutation module. Some conclusions in [6] only depend on the existence of a direct permutation module summand. We justify the more general set-up taken up here by constructing examples of groups of infinite quasi-projective dimension to which our structure results still apply.

For finite groups, Gruenberg and Roggenkamp [4], [5] investigated non-projective decompositions of the augmentation ideal or of relation modules. Under a more special hypothesis on the decomposition, we avoid the restriction to finite groups and to low dimensions. Dealing with finite groups later on, similarity with results of [5] turns up.

We now introduce another piece of notation. Let $\mathscr{C} : \cdots \to C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} C_{n-2} \to \cdots$ be a chain complex. Then we denote the kernel of $\partial_{n-1} : C_{n-1} \to C_{n-2}$ by $K_n(\mathscr{C})$.

In *Section* 2 we assume that for some projective resolution $\mathscr{P} \twoheadrightarrow \mathbf{Z}$ and some n, there is a splitting $K_n(\mathscr{P}) \cong M \oplus \mathbf{Z}G/\boldsymbol{S}$. We show that each group in the family $\boldsymbol{S}$ is finite. The Tate cohomology of any non-trivial group in $\boldsymbol{S}$ has a divisor of n as its period. Our key result is a variation of a theorem attributed to Serre in [7].

---

[1] Partly supported by a William Gordon Seggie Brown Fellowship.

THEOREM 5. *Suppose* $(G, \boldsymbol{S})$ *is a group-pair and* q, r *are positive integers such that for every* $\mathbf{Z}G$*-module* M *the group* $H^q(G, M)$ *has a direct summand isomorphic to* $\prod_{i\in I} H^r(S_i, M)$. *Suppose also that* $i, j \in I$ *and* $g \in G$ *are such that* $S_i \cap gS_jg^{-1}$ *is not torsion-free. Then* $i = j$ *and* $g \in S_i$.

This result leads to a chain of corollaries some of which state facts proved earlier for groups of finite quasi-projective dimension in [6]. In particular, the hypotheses of Theorem 5 are satisfied for a pair $(G, \boldsymbol{S})$ if for some $\mathbf{Z}G$-projective resolution $\mathcal{P} \to \mathbf{Z}$ and some $n > 0$, the module $\mathbf{Z}G/\boldsymbol{S}$ is a direct summand of the kernel $K_n(\mathcal{P})$. The following statement already shows that our investigation extends beyond the class of groups of finite quasi-projective dimension.

We term a pair of groups (G, S) a *Frobenius pair*, if G is a Frobenius group and S a Frobenius complement of G.

COROLLARY 5.4. *Suppose* G *is a finite group and* $\mathbf{Z}G/S$ *is isomorphic to a direct summand of* $K_n(\mathcal{P})$ *for some* $\mathbf{Z}G$*-projective resolution* $\mathcal{P} \to \mathbf{Z}$ *and some* $n > 0$. *Then either* $S = 1$, $S = G$ *or* (G, S) *is a Frobenius-pair.*

The objective of *Section* 3 is a discussion of Frobenius groups from the point of view of homological algebra.

THEOREM 7. *Let* (G, S) *be a Frobenius pair. Then* $\mathbf{Z}G/S$ *is a* $\mathbf{Z}G$*-direct summand of* $\mathbf{Z} \oplus \mathbf{Z}G$.

The relationship between splitting of $\mathbf{Z} \oplus \mathbf{Z}G$ and of the augmentation ideal $IG = \ker(\mathbf{Z}G \to \mathbf{Z})$ is investigated in [4], [5].

THEOREM 9. *Let* (G, S) *be a Frobenius pair and let* n *denote the period of the Tate cohomology of* S. *Then there exists a finitely generated* $\mathbf{Z}G$*-free resolution* $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ *such that* $\mathbf{Z}G/S$ *is a* $\mathbf{Z}G$*-direct summand of* $K_n(\mathcal{P})$.

In a different direction, we use relative (co)-homology of a Frobenius pair (G, S) to compute the (co)-homology of G. We obtain results similar to [1, 55.1] in a rather elementary way from the long exact (co)-homology sequences of the pair (G, S). We refer to the Appendix for the discussion of some special properties of the restriction-corestriction maps in relative cohomology.

COROLLARY 7.2. *Let* (G, S) *be a Frobenius pair and denote by* N *the kernel of the projection* $G \twoheadrightarrow S$. *Then there exist natural isomorphisms of functors on*

$\mathbf{Z}G$-*modules for any* $r>1$ *as follows*:

(i) $H^r(G;-)\cong H^r(S;-)\oplus H^r(G,S;-)$.

(ii) $H^r(G,S;-)\cong H^r(N;-)^S$.

(iii) *For all* $r\in\mathbf{Z}$, *there are natural isomorphisms*

$$\hat{H}^r(S;(-)^N)\cong\hat{H}^r(S;\mathrm{res}_S^G(-))$$

of functors on $\mathbf{Z}G$-*modules.*

For any finite group H let π_H denote the set of primes dividing $|H|$. Then the decomposition (i) is a functorial splitting of $H^r(G,-)$ into π_S- and π_N-parts $(r>1)$.

In *Section* 4, we present some explicit computations related to examples. We first deal with Frobenius pairs of the type $(G,\mathbf{Z}/k\mathbf{Z})$. In this context, we observe decompositions of the relation modules occurring in the Lyndon resolution derived from a finite presentation of G. This is an opportunity to make explicit some connection with results of Gruenberg and Roggenkamp [5]. The easiest non-trivial example of this case is the symmetric group $S_3=\langle x,y\mid x^2=y^3=(xy)^2=1\rangle$.

In analogy with [6, Theorem 5] we use graph products to construct infinite groups G for which permutation modules occur as a direct summand of some $K_n(\mathscr{P})$. Of special interest is the group $S_3*_C S_3$, where $C\cong\mathbf{Z}/2\mathbf{Z}$, since this group has infinite quasi-projective dimension and it cannot be constructed with the help of graph products along the lines of [6, Theorem 5].

We finally remark that our working hypothesis is based on $\mathbf{Z}G$-projective resolutions. A natural source for such resolutions is free G-actions on acyclic spaces. It would be desirable to know the topological circumstances under which actions produce the phenomena discussed here since they lead from geometric considerations to structure results in group theory.

2. Projective resolutions

2.1. *Preliminaries and general facts*

Suppose $(G,\boldsymbol{S})$ is a group pair, where G is a subgroup of a group H. Then there is a $\mathbf{Z}H$-isomorphism $\mathbf{Z}H\otimes_G\mathbf{Z}G/\boldsymbol{S}\cong\mathbf{Z}H/\boldsymbol{S}$.

Suppose $(G,\boldsymbol{S})$ is a group pair, and U is a subgroup of G. If we restrict the $\mathbf{Z}G$-action on $\mathbf{Z}G/\boldsymbol{S}$ to a $\mathbf{Z}U$-action, we obtain a permutation module $\mathbf{Z}U/\boldsymbol{T}$, where the family $\boldsymbol{T}$ can be constructed as follows. Let I be the index set of the

family $\boldsymbol{S}=\{S_i\}_I$. For each $i \in I$, choose a set $\{t_\beta;\ \beta \in J_i\}$ of representatives of the double cosets UgS_i, and for each β in J_i define $T_{i\beta} = U \cap t_\beta S_i t_\beta^{-1}$. Then $\boldsymbol{T}$ is the family $\{T_{i\beta};\ i \in I, \beta \in J_i\}$. This construction also appears in [2, p. 305]. We call $(U, \boldsymbol{T})$ the pair *induced* from $(G, \boldsymbol{S})$ by U.

We now consider a $\mathbf{Z}G$-projective resolution $\mathscr{P} \twoheadrightarrow A$ of some $\mathbf{Z}G$-module A, and suppose that $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a submodule of $K_n(\mathscr{P})$ for some positive integer n. Then, by restricting the $\mathbf{Z}G$-action, we obtain a $\mathbf{Z}U$-projective resolution $\mathscr{P} \twoheadrightarrow A$, and a $\mathbf{Z}U$-submodule of $K_n(\mathscr{P})$ isomorphic to $\mathbf{Z}U/\boldsymbol{T}$, where $(U, \boldsymbol{T})$ is the induced pair.

In particular, if we take $U = S_i$ for some $i \in I$, then one subgroup in the family $\boldsymbol{T}$ is just $S_i \cap S_i = S_i$, so that the $\mathbf{Z}S_i$-projective P_{n-1} contains a $\mathbf{Z}S_i$-submodule isomorphic to $\mathbf{Z}S_i/S_i \cong \mathbf{Z}$. This is possible only if S_i is a finite group. We have thus proved the following.

PROPOSITION 1. *Suppose $\mathscr{P} \twoheadrightarrow A$ is a $\mathbf{Z}G$-projective resolution, and $\mathbf{Z}G/\boldsymbol{S}$ is a submodule of some $K_n(\mathscr{P})$. Then $\boldsymbol{S}$ is a family of finite subgroups.*

Note that any projective resolution $\mathscr{P}$ can be modified so that $K_n(\mathscr{P})$ contains a free direct summand. Hence only the non-trivial subgroups in $\boldsymbol{S}$ are of interest.

Next we observe that, if $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a direct summand of $K_n(\mathscr{P})$ for some projective resolution $\mathscr{P} \twoheadrightarrow A$, then, modulo a slight adjustment, the same is true for any other projective resolution of A. In particular, $\mathscr{P}$ may always be assumed to be free.

PROPOSITION 2. *Suppose A is a $\mathbf{Z}G$-module, $\mathscr{P} \twoheadrightarrow A$ and $\mathscr{Q} \twoheadrightarrow A$ are $\mathbf{Z}G$-projective resolutions, and $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a direct summand of $K_n(\mathscr{P})$. Then one can form a $\mathbf{Z}G$-projective resolution $\mathscr{Q}' \twoheadrightarrow A$ by adding a free $\mathbf{Z}G$-module F to $\mathscr{Q}$ in dimensions n and $n-1$, such that $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a direct summand of $K_n(\mathscr{Q}')$. Furthermore, if the index set I of the family $\boldsymbol{S}$ is finite, then F may be chosen of finite rank.*

Proof. By Schanuel's lemma, there are $\mathbf{Z}G$-projectives P and Q such that $K_n(\mathscr{P}) \oplus P \cong K_n(\mathscr{Q}) \oplus Q$. The first part of the proposition follows by choosing F large enough to contain Q as a direct summand, and noting that $K_n(\mathscr{Q}') \cong K_n(\mathscr{Q}) \oplus F$.

Now suppose I is finite. then $\mathbf{Z}G/\boldsymbol{S}$ is finitely generated. Hence, for any choice of basis for F, only finitely many basis elements are involved in the images of elements of $\mathbf{Z}G/\boldsymbol{S}$ in $K_n(\mathscr{Q}) \oplus F$. We may therefore replace F in the construction of $\mathscr{Q}'$ by a suitable free direct summand of finite rank.

2.2. *Resolutions of* $\mathbf{Z}$

From now on we restrict our attention to $\mathbf{Z}G$-projective resolutions $\mathcal{P} \twoheadrightarrow \mathbf{Z}$. Our next observation is that an integer n for which $\mathbf{Z}G/\boldsymbol{S}$ is a direct summand of $K_n(\mathcal{P})$ for some $\mathcal{P}$ is by no means unique. In fact, if there is one such integer, there are infinitely many.

PROPOSITION 3. *Suppose $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ is a $\mathbf{Z}G$-projective resolution such that $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a direct summand of $K_n(\mathcal{P})$. Then there exists, for each positive integer r, a $\mathbf{Z}G$-projective resolution $\mathcal{P}^{(r)} \twoheadrightarrow \mathbf{Z}$, such that $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a direct summand of $K_m(\mathcal{P}^{(r)})$.*

Proof. We define $\mathcal{P}^{(1)} = \mathcal{P}$, and $\mathcal{P}^{(r)}$ $(r \geqslant 2)$ inductively as follows. Suppose $\mathcal{P}^{(r-1)}$ has been defined, and $K_{(r-1)n}(\mathcal{P}^{(r-1)}) = \mathbf{Z}G/\boldsymbol{S} \oplus A$. Let $\mathcal{Q} \twoheadrightarrow A$ be any $\mathbf{Z}G$-projective resolution, and define $\mathcal{P}^{(r)} \twoheadrightarrow \mathbf{Z}$ to be

$$\begin{array}{ccccccccc} \left(\bigoplus_{i\in I} \mathbf{Z}G \otimes_{S_i} \mathcal{P}\right) \oplus \mathcal{Q} & & \longrightarrow & & P^{(r-1)}_{(r-1)n-1} & \to \cdots \to & P^{(r-1)}_0 & \to \mathbf{Z} \to & 0 \\ & \searrow & & \nearrow & & & & & \\ & \left(\bigoplus_{i\in I} \mathbf{Z}G \otimes_{S_i} \mathbf{Z}\right) \oplus A. & & & & & & & \end{array}$$

Then $K_{nr}(\mathcal{P}^{(r)}) \cong (\oplus_{i\in I} \mathbf{Z}G \otimes_{Si} K_n(\mathcal{P})) \oplus K_n(\mathcal{Q})$.

For each $i \in I$, $K_n(\mathcal{P})$ contains $\mathbf{Z}G/S_i$, and so $\mathbf{Z}$, as a $\mathbf{Z}S_i$-direct summand. Hence $K_m(\mathcal{P})$ contains $\oplus_I \mathbf{Z}G \otimes_{S_i} \mathbf{Z} \cong \mathbf{Z}G/\boldsymbol{S}$ as a $\mathbf{Z}G$-direct summand, as claimed.

The next proposition is a straightforward calculation of cohomology, and we omit the proof.

PROPOSITION 4. *Suppose $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ is a $\mathbf{Z}G$-projective resolution and $K_n(\mathcal{P}) = \mathbf{Z}G/\boldsymbol{S} \oplus A$. Then there are, for each $q > 0$, natural isomorphisms*

$$H^{n+q}(G; -) \simeq \left(\prod_{i\in I} H^q(S_i; -)\right) \oplus \mathrm{Ext}^q_{\mathbf{Z}G}(A; -)$$

$$H_{n+q}(G; -) \cong \left(\bigoplus_{i\in I} H_q(S_i; -)\right) \oplus \mathrm{Tor}^{\mathbf{Z}G}_q(A; -).$$

COROLLARY 4.1. *Suppose $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ is a $\mathbf{Z}G$-projective resolution such that $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a direct summand of $K_n(\mathcal{P})$. Then either n is even or $\boldsymbol{S}$ consists entirely of copies of $\{1\}$.*

Proof. Suppose n is odd and, for some $i \in I$, $S_i \neq \{1\}$.

By Proposition 1, S_i is finite. Choose a subgroup C of S_i of prime order. Then, regarded as a $\mathbf{Z}C$-module, $K_n(\mathscr{P})$ has a direct summand isomorphic to $\mathbf{Z}$, say $K_n(\mathscr{P}) \cong \mathbf{Z} \oplus A'$. Apply Proposition 4 to C with $q = 1$. Since n is odd, we have

$$0 = H_{n+1}(C; \mathbf{Z}) \cong H_1(C; \mathbf{Z}) \oplus \mathrm{Tor}_1^{\mathbf{Z}C}(A'; \mathbf{Z}) \neq 0.$$

This is a contradiction, so the proof is complete.

The importance of Proposition 4 is that it may be used together with our next result to deduce certain group-theoretic properties of the pair $(G, \boldsymbol{S})$. This result is a partial generalisation of a theorem of Serre [7].

THEOREM 5. *Suppose $(G, \boldsymbol{S})$ is a group-pair and q, r are positive integers such that, for every $\mathbf{Z}G$-module M, the group $H^q(G; M)$ has a direct summand isomorphic to $\prod_{i \in I} H^r(S_i; M)$. Suppose also that $i, J \in I$, $g \in G$ are such that $S_i \cap gS_jg^{-1}$ is not torsion-free. Then $i = j$ and $g \in S_i$.*

Proof. Choose a cyclic subgroup $C \neq 1$ of finite order in $S_i \cap gS_jg^{-1}$. Then we have, for any $\mathbf{Z}C$-module M, an isomorphism $H^q(C; M) \cong H^q(G; \mathrm{Hom}_{\mathbf{Z}C}(\mathbf{Z}G; M))$. By hypothesis, therefore, $H^q(C; M)$ has a direct summand isomorphic to

$$\prod_{i \in I} H^r(S_i; \mathrm{Hom}_{\mathbf{Z}C}(\mathbf{Z}G; M)) \cong \mathrm{Ext}^r_{\mathbf{Z}G}(\mathbf{Z}G/\boldsymbol{S}; \mathrm{Hom}_{\mathbf{Z}C}(\mathbf{Z}G; M)) \cong \mathrm{Ext}^r_{\mathbf{Z}C}(\mathbf{Z}G/\boldsymbol{S}; M).$$

Choose $M = \mathbf{Z}$ or $M = IC$ (the augmentation ideal of $\mathbf{Z}C$) according as r is even or odd. Then $H^q(C; M)$ is either cyclic of order $|C|$ or zero, according as $(q - r)$ is even or odd. In any case $H^q(C; M)$, and so also $\mathrm{Ext}^r_{\mathbf{Z}C}(\mathbf{Z}G/\boldsymbol{S}; M)$, is cyclic. But $\mathrm{Ext}^r_{\mathbf{Z}C}(\mathbf{Z}G/\boldsymbol{S}; M)$ contains a direct summand isomorphic to $\mathbf{Z}/|C|\mathbf{Z}$ for each point of the C-set $\bigcup_{i \in I} G/S_i$ fixed by C. Hence there is at most one such point, so the fixed points $1 \cdot S_i$ and $g \cdot S_j$ coincide. In other words, $i = j$ and $g \in S_i$.

COROLLARY 5.1. *Suppose $\mathbf{Z}G/\boldsymbol{S}$ is isomorphic to a direct summand of $K_n(\mathscr{P})$ for some $\mathbf{Z}G$-projective resolution $\mathscr{P} \twoheadrightarrow \mathbf{Z}$ and some positive integer n. Suppose also that $i, j \in I$, $g \in G$ are such that $S_i \cap gS_jg^{-1} \neq 1$. Then $i = j$ and $g \in S_i$.*

Proof. Follows from Propositions 1 and 4, and Theorem 5.

COROLLARY 5.2. *Suppose* $\mathbf{Z}G/S$ *is isomorphic to a direct summand of* $K_n(\mathcal{P})$ *for some* $\mathbf{Z}G$*-projective resolution* $\mathcal{P} \twoheadrightarrow \mathbf{Z}$*, and some positive integer* n*. Then either* $S = 1$ *or* S *is its own normaliser in* G*.*

COROLLARY 5.3. *Suppose* G *is an infinite group with non-trivial centre, and* $\mathbf{Z}G/S$ *is isomorphic to a direct summand of* $K_n(\mathcal{P})$ *for some* $\mathbf{Z}G$*-projective resolution* $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ *and some positive integer* n*. Then* $S = 1$*.*

COROLLARY 5.4. *Suppose* G *is a finite group and* $\mathbf{Z}G/S$ *is isomorphic to a direct summand of* $K_n(\mathcal{P})$ *for some* $\mathbf{Z}G$*-projective resolution* $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ *and some positive integer* n*. Then either* $S = 1$*,* $S = G$*, or* (G, S) *is a Frobenius pair.*

COROLLARY 5.5. *Suppose* $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ *and* $\mathcal{Q} \twoheadrightarrow \mathbf{Z}$ *are* $\mathbf{Z}G$*-projective resolutions and* m *and* n *are positive integers, such that* $\mathbf{Z}G/S$ *is a direct summand of* $K_m(\mathcal{P})$*, and* $\mathbf{Z}G/T$ *is a direct summand of* $K_n(\mathcal{Q})$*. Suppose also that* $g \in G$ *is such that* $U = S \cap gTg^{-1}$ *is non-trivial. Then either* $U = S$ *or* (S, U) *is a Frobenius pair.*

Proof. Follows from Corollary 5.4 by regarding $\mathcal{Q}$ as a $\mathbf{Z}S$-resolution.

Remarks. Serre's Theorem states that, under a hypothesis somewhat stronger than that of Theorem 5, any finite subgroup is contained in precisely one conjugate of precisely one of the subgroups S_i. Under this stronger hypothesis, the third possibility in Corollary 5.4, that (G, S) is a Frobenius pair, would be ruled out. Under our weaker hypothesis, however, we cannot rule the possibility out. Indeed, the permutation modules $\mathbf{Z}G/S$, where (G, S) is a Frobenius pair, all occur as direct summands of kernels in projective resolutions of $\mathbf{Z}$. We will prove this, and other facts about Frobenius groups, in the next section.

Of course, it is well known that any Frobenius complement has periodic Tate cohomology. Our next result states that, if $\mathbf{Z}G/S$ is isomorphic to a direct summand of some $K_n(\mathcal{P})$, then S has periodic Tate cohomology.

PROPOSITION 6. *Suppose* S *is a non-trivial subgroup of* G*, and* $\mathbf{Z}G/S$ *is isomorphic to a direct summand of* $K_n(\mathcal{P})$ *for some* $\mathbf{Z}G$*-projective resolution* $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ *and some positive integer* n*. Then* S *has periodic Tate cohomology, with period dividing* n*.*

Proof. By Proposition 1, S is finite. Also, regarded as a $\mathbf{Z}S$-module, $K_n(\mathcal{P})$ has

a direct summand isomorphic to $\mathbf{Z}$. Since $n \neq 1$ by Corollary 4.1, we may apply Proposition 2 to find a finitely-generated free $\mathbf{Z}S$-resolution $\mathscr{F} \twoheadrightarrow \mathbf{Z}$ with $F_0 \cong \mathbf{Z}S$ and $K_n(\mathscr{F}) = \mathbf{Z} \oplus A$ for some $\mathbf{Z}S$-module A.

Form a complete $\mathbf{Z}S$-resolution

$$\cdots \to F_1 \xrightarrow{\partial_0} \mathbf{Z}S \xrightarrow{\nu} \mathbf{Z}S \xrightarrow{\partial_0} F_1^* \to \cdots \qquad (*)$$

from $\mathscr{F}$ in the usual way by means of the dual $\mathscr{F}^*$ and the norm-map ν.

From the exact sequences

$$0 \to K_{n-1}(\mathscr{F})^* \to F_{n-1}^* \to K_n(\mathscr{F})^* \to \mathrm{Ext}^n_{\mathbf{Z}S}(\mathbf{Z}; \mathbf{Z}S) = 0$$

(see for example [3, p. 90, Theorem 6.1]), we deduce that $(*)$ is exact and Coker $\partial_{n-1}^* = K_n(\mathscr{F})^* \cong \mathbf{Z} \oplus A^*$. We use this to obtain natural isomorphisms.

$$H^q(S; -) \oplus \mathrm{Ext}^q_{\mathbf{Z}S}(A^*; -) \cong \hat{H}^{q-n}(S; -) \qquad (q \geqslant 1),$$

where $\hat{H}$ denotes Tate cohomology.

For any $\mathbf{Z}S$-module M and any $q \geqslant 1$, we have, from the above and Proposition 4, that $H^q(S; M)$ and $H^{q+n}(S; M)$ are direct summands of one another. It follows that $H^q(S; -)$ and $H^{q+n}(S; -)$ agree on finitely generated modules (and hence on all modules, since $H^q(S; -)$ commutes with direct limits).

Thus the Tate cohomology of S is periodic, of period dividing n, as claimed.

3. Frobenius groups

THEOREM 7. *Let (G, S) be a Frobenius pair. Then $\mathbf{Z}G/S$ is a direct summand of $\mathbf{Z} \oplus \mathbf{Z}G$.*

Remark. This result should be compared with [5, Lemma 4.1], which expresses the augmentation ideal of a Frobenius group as a direct sum of factors corresponding to the Frobenius complement and kernel respectively. As is shown in [5], there is a close correspondence between direct sum decompositions of $\mathbf{Z} \oplus \mathbf{Z}G$ and of the augmentation ideal of G.

Proof. By the definition of Frobenius pair, G acts transitively on a set X such that no element of $G \setminus 1$ fixes more than one element of X, and S is the stabiliser

of some element x_0 of X. By Frobenius' Theorem, G is a semi-direct product $N \lhd S$, where N acts freely and transitively on the G-orbit X of x_0. Hence X is the N-orbit of x_0, and there is an identification between elements of X and of N such that the action of S on X corresponds to the S-action by conjugation on N, and such that $x_0 \in X$ is identified with $1 \in N$. It follows that S acts freely by conjugation on $N\backslash 1$. Let $\Omega_1, \ldots, \Omega_t$ denote the orbits of the S-action on $N\backslash 1$, and choose orbit representatives $g_r \in \Omega_r$ for $1 \leqslant r \leqslant t$.

We define $\mathbf{Z}G$-morphisms $\phi : \mathbf{Z}G/S \to \mathbf{Z} \oplus \mathbf{Z}G$ and $\psi : \mathbf{Z} \oplus \mathbf{Z}G \to \mathbf{Z}G/S$ by

$$\phi : 1 \cdot S \mapsto \left(1, \sum_{r=1}^{t} \sum_{s \in S} sg_r\right)$$

$$\psi|_{\mathbf{Z}} : 1 \mapsto \sum_{g \in N} g \cdot S$$

$$\psi|_{\mathbf{Z}G} : 1 \mapsto -1 \cdot S.$$

Then we have

$$(\psi \circ \phi)(1 \cdot S) = \sum_{g \in N} g \cdot S - \sum_{r=1}^{t} \sum_{s \in S} sg_r \cdot S. \tag{1}$$

Since S acts freely on $N\backslash 1$ we have, for $1 \leqslant r \leqslant t$,

$$\sum_{s \in S} sg_r \cdot S = \sum_{s \in S} (sg_r s^{-1}) \cdot S = \sum_{g \in \Omega_r} g \cdot S.$$

Summing over all r and substituting in (1), we deduce that

$$(\psi \circ \phi)(1 \cdot S) = 1 \cdot S.$$

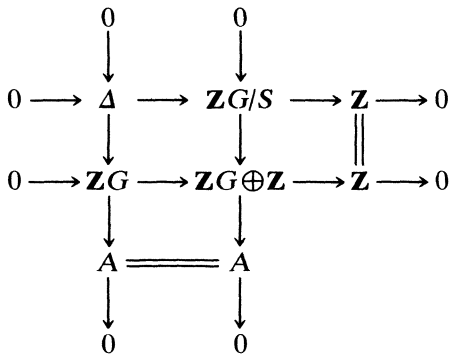
Hence $\psi \circ \phi = 1_{\mathbf{Z}G/S}$ and so $\mathbf{Z}G/S$ is isomorphic to $\mathbf{Z}G$-direct summand of $\mathbf{Z} \oplus \mathbf{Z}G$, as claimed.

COROLLARY 7.1. *Let (G, S) be a Frobenius pair and $r \geqslant 2$ an integer. Then there are natural isomorphisms*

$$H^r(G; -) \cong H^r(S; -) \oplus H^r(G, S; -)$$
$$H_r(G; -) \cong H_r(S; -) \oplus H_r(G, S; -).$$

Proof. We denote by A the complement of $\mathbf{Z}G/S$ in $\mathbf{Z} \oplus \mathbf{Z}G$. The following

commutative diagram has exact rows and columns



For $r \geqslant 1$ we read off the natural isomophism

$$H^r(G;-) \cong \mathrm{Ext}^r_{\mathbf{Z}G}(\mathbf{Z}G \oplus \mathbf{Z};-) \cong \mathrm{Ext}^r_{\mathbf{Z}G}(\mathbf{Z}G/S;-) \oplus \mathrm{Ext}^r_{\mathbf{Z}G}(A;-).$$

By Shapiro's Lemma,

$$\mathrm{Ext}^r_{\mathbf{Z}G}(\mathbf{Z}G/S;-) \cong H^r(S;-),$$

and from the long exact sequence induced by the first column we get, for $r \geqslant 2$,

$$\mathrm{Ext}^r_{\mathbf{Z}G}(A;-) \cong \mathrm{Ext}^{r-1}_{\mathbf{Z}G}(\Delta;-) \cong H^r(G, S;-).$$

This establishes the first of the stated natural isomorphisms. The second is proved in an analogous way.

COROLLARY 7.2. *Let (G, S) be a Frobenius pair and denote by N the kernel of the projection $G \twoheadrightarrow S$. For any finite group H let π_H denote the set of primes dividing $|H|$. Then the following hold:*

(i) *The decompositions in Corollary* 7.1 *are functorial splittings of $H^r(G;-)$, resp. $H_r(G;-)$, into π_S- and π_N-parts $(r \geqslant 2)$.*

(ii) *For all $r \geqslant 2$, there are natural isomorphisms of functors on $\mathbf{Z}G$-modules*

$$H^r(G, S;-) \cong H^r(N;-)^S.$$

(iii) *For all $r \in \mathbf{Z}$, there are natural isomorphisms of functors on $\mathbf{Z}G$-modules*

$$\hat{H}^r(S, (-)^N) \cong \hat{H}^r(S, \mathrm{res}^G_S(-)).$$

Proof. Statement (i) follows from the fact that $|S|$ annihilates $H^r(S;-)$ and $|N| = |G:S|$ annihilates $H^r(G,S;-)$ (because (G,S) is Frobenius – see Appendix). Statements (ii) and (iii) result from (i) when one compares the splitting of $H^r(G;A)$ in Corollary 7.1 with that given by the restriction-inflation sequence for the normal Hall subgroup N [1, p. 191, 55.1]. For (iii) one uses the periodicity of S to eliminate the lower bound on the dimension r.

Remark. For the particular coefficient module $\mathbf{Z}G/S$ and for $r=0$, statement (iii) of Corollary 7.2 takes the form

$$\hat{H}^0(S;\mathbf{Z}) \cong \hat{H}^0(S;\mathbf{Z}G/S). \tag{2}$$

Conversely, if (G,S) is a group-pair with G finite and $1 \neq S \neq G$, such that (2) holds, then (G,S) is a Frobenius pair. For the module $\mathbf{Z}G/S$ considered as a $\mathbf{Z}S$-module has the form

$$\mathbf{Z}G/S \cong \mathbf{Z} \oplus \left(\bigoplus_{i=1}^{n} \mathbf{Z}S/S_i \right)$$

and hence

$$\begin{aligned} \hat{H}^0(S;\mathbf{Z}) &\cong \hat{H}^0(S;\mathbf{Z}G/S) \quad \text{(by (2))} \\ &\cong \hat{H}^0(S;\mathbf{Z}) \oplus \left(\bigoplus_{i=1}^{n} \hat{H}^0(S_i;\mathbf{Z}) \right). \end{aligned}$$

Thus all the S_i are trivial and S acts freely on the set $(G/S)\backslash(1 \cdot S)$. Essentially the same computation gives the following converse of Theorem 7.

PROPOSITION 8. *Let (G,S) be a group pair, with G finite, such that $\mathbf{Z}G/S$ is a $\mathbf{Z}S$-direct summand of $\mathbf{Z} \oplus M$, where M is a $\mathbf{Z}S$-module satisfying $\hat{H}^0(S;M)=0$. Then (G,S) is a Frobenius pair.*

THEOREM 9. *Let (G,S) be a Frobenius pair, and let n denote the period of the Tate cohomology of S. Then there exists a finitely-generated $\mathbf{Z}G$-free resolution $\mathcal{P} \twoheadrightarrow \mathbf{Z}$ such that $\mathbf{Z}G/S$ is a $\mathbf{Z}G$-direct summand of $K_n(\mathcal{P})$.*

Proof. Since S has cohomological period n, one can construct, using the methods of [8, section 2], a finitely-generated $\mathbf{Z}S$-free resolution $\mathcal{F} \twoheadrightarrow \mathbf{Z}$ such that $K_n(\mathcal{F}) \cong \mathbf{Z} \oplus Q$ for some projective $\mathbf{Z}S$-module Q.

By Theorem 7, there exists a $\mathbf{Z}G$-module A such that $\mathbf{Z} \oplus \mathbf{Z}G \cong A \oplus \mathbf{Z}G/S$ as

$\mathbf{Z}G$-modules. Let $\mathscr{F}' \twoheadrightarrow A$ be any finitely-generated $\mathbf{Z}G$-free resolution. Then $\mathscr{P}' = \mathscr{F}' \oplus (\mathbf{Z}G \otimes_S \mathscr{F})$ is a finitely-generated $\mathbf{Z}G$-free resolution of $A \oplus \mathbf{Z}G/S \cong \mathbf{Z} \oplus \mathbf{Z}G$. Furthermore, $\mathbf{Z}G/S$ is a $\mathbf{Z}G$-direct summand of $K_n(\mathscr{P}')$.

We use the epimorphism $\varepsilon : P_0' \twoheadrightarrow \mathbf{Z} \oplus \mathbf{Z}G$ to produce a splitting $P_0' \cong Q' \oplus \mathbf{Z}G$, where Q' is the full preimage of $\mathbf{Z}$ under ε. Since Q' contains the kernel of ε, we may define a finitely-generated $\mathbf{Z}G$-free resolution

$$\begin{array}{ccccccccc} \mathscr{P}: \cdots & \longrightarrow & P_2' & \longrightarrow & P_1' & \longrightarrow & Q' & \longrightarrow & \mathbf{Z}' \longrightarrow 0 \\ & & & & \oplus & & \oplus & & \\ & & & & \mathbf{Z}G & \longrightarrow & \mathbf{Z}G & & \end{array}$$

such that $K_n(\mathscr{P}) \cong (K_n(\mathscr{P}')$ has a $\mathbf{Z}G$-direct summand isomorphic to $\mathbf{Z}G/S$.

4. Examples

4.1. *Frobenius pairs*

Suppose (G, S) is a Frobenius pair, and S has cohomological period n. Then Theorem 9 tells us that $\mathbf{Z}G/S$ is a direct summand of $K_n(\mathscr{F})$ for some $\mathbf{Z}G$-free resolution $\mathscr{F} \twoheadrightarrow \mathbf{Z}$. The proof of Theorem 9 relies on the decomposition $\mathbf{Z} \oplus \mathbf{Z}G \cong \mathbf{Z}G/S \oplus A$ to construct a suitable resolution $\mathscr{F}$. It does not help us to decide whether, for any given resolution $\mathscr{F} \twoheadrightarrow \mathbf{Z}$, the module $\mathbf{Z}G/S$ appears as a direct summand of $K_n(\mathscr{F})$.

We examine this situation more closely in the special case where $n = 2$ (that is, S is cyclic) and $\mathscr{F} \twoheadrightarrow \mathbf{Z}$ is the Lyndon resolution arising from a finite presentation of G. In this case, $K_2(\mathscr{F})$ is the relation module of the presentation.

Gruenberg and Roggenkamp [5, Propositions 5(i), 6] have shown that, in this situation, any relation module decomposes as the direct sum of two non-projective factors. In fact, following the argument in [5] reveals that one of the factors is isomorphic to $\mathbf{Z}G/S$.

Explicitly, there is an isomorphism [5, Lemma 4.1]

$$K_1(F) = IG \cong \mathbf{Z}G \otimes_S IS \oplus IN,$$

where N is the kernel of the projection $G \twoheadrightarrow S$; the modules IG, IS, IN are the augmentation ideals in $\mathbf{Z}G$, $\mathbf{Z}S$, $\mathbf{Z}N$ respectively, and the N-action on IN is extended to a G-action by letting S act by conjugation on elements of N.

Given a free presentation $\phi : \Phi \to G$ of finite rank, one chooses a basis $\{x_1, \ldots, x_n\}$ of Φ such that $g_1 = \phi(x_1) \in S$ and $g_i = \phi(x_i) \in N$ for $2 \leqslant i \leqslant n$. This

allows one to decompose the exact sequence

$$0 \longrightarrow K_2(\mathscr{F}) \longrightarrow (\mathbf{Z}G)^n \xrightarrow[\begin{pmatrix} 1-g_1 \\ \vdots \\ 1-g_n \end{pmatrix}]{} K_1(\mathscr{F}) = IG \longrightarrow 0$$

as the direct sum of two terms:

$$0 \longrightarrow \mathbf{Z}G/S \longrightarrow \mathbf{Z}G \xrightarrow[(1-g_1)]{} \mathbf{Z}G \otimes_S IS \longrightarrow 0$$

and

$$0 \longrightarrow B \longrightarrow (\mathbf{Z}G)^{n-1} \xrightarrow[\begin{pmatrix} 1-g_2 \\ \vdots \\ 1-g_n \end{pmatrix}]{} IN \longrightarrow 0$$

EXAMPLE 1. Let G be the symmetric group of degree 3, $\mathscr{F} \twoheadrightarrow \mathbf{Z}$ the Lyndon resolution associated to the presentation

$$\langle x, y \mid x^2 = (xy)^2 = y^3 = 1 \rangle,$$

and S the subgroup of order 2 generated by x.

Then, $\partial_1 : (\mathbf{Z}G)^2 \to \mathbf{Z}G$ is given by the matrix

$$\begin{pmatrix} 1-x \\ 1-y \end{pmatrix}$$

and $K_2(\mathscr{F}) = \operatorname{Ker} \partial_1$ has a direct summand isomorphic to $\mathbf{Z}G/S$. This is embedded in $F_1 = (\mathbf{Z}G)^2$ as the cyclic submodule generated by the element $(1+x, 0)$.

4.2. *Infinite groups*

Using Theorem 5, Corollary 5.4 and Theorem 9, we can completely classify those group pairs $(G, \boldsymbol{S})$, with G finite, for which $\mathbf{Z}G/\boldsymbol{S}$ occurs as a direct summand of some $K_n(\mathscr{P})$. They are pairs $(G, \{S, 1, 1, \ldots\})$ such that either

(a) (G, S) is a Frobenius pair;

(b) $S = G$, a group with periodic cohomology; or

(c) $S = 1$, in which case $\mathbf{Z}G/\boldsymbol{S}$ is free, and no further information about G can be deduced.

For group pairs $(G, \boldsymbol{S})$ with G infinite, the position is less clear, and a complete classification would appear to be impossible to find. A restricted insight into the complexity of the class of groups considered may be obtained from looking at constructions under which the class is closed. In this light, the following proposition and examples give some idea of how complicated the situation is.

PROPOSITION 10. *Suppose Γ is a graph of groups whose edge groups G_e satisfy* cd $G_e \leqslant n_0$ *for some fixed integer n_0. Suppose we are given, for each vertex v, a (possibly empty) family $\boldsymbol{S}_v$ of subgroups of the vertex group G_v, and a $\mathbf{Z}G$-projective resolution $\mathscr{P}_v \twoheadrightarrow \mathbf{Z}$ such that for some $n(v) \leqslant n_0$, $\mathbf{Z}G_v/\boldsymbol{S}_v$ is a direct summand of $K_{n(v)}(\mathscr{P}_v)$. Let $\boldsymbol{S}$ denote the union $\bigcup_v \boldsymbol{S}_v$, regarded as a family of subgroups of the fundamental group $G = \pi(\Gamma)$. Then there exists a $\mathbf{Z}G$-projective resolution $\mathscr{P} \twoheadrightarrow \mathbf{Z}$, and an integer r, such that $\mathbf{Z}G/\boldsymbol{S}$ is a direct summand of $K_r(\mathscr{P})$.*

Proposition 10 is a direct generalisation of [6, Theorem 5], and is proved in an analogous way, using mapping cones. We omit the details.

Proposition 10 allows us to construct a large class of examples, beginning with those we already know, such as Frobenius groups, and groups of finite quasi-projective dimension. Our next example does not arise in this way. That is, the group concerned is infinite, has infinite quasi-projective dimension, and cannot be properly expressed as the fundamental group of a graph of groups whose edge groups are torsion-free.

EXAMPLE 2. Define $G = A \underset{S}{*} B$, where A and B are isomorphic copies of the symmetric group of degree 3, and $S = \langle x \rangle$ is a common subgroup of order 2.

If we apply the exact functor $\mathbf{Z}G \otimes_A -$ to the $\mathbf{Z}A$-resolution in Example 1, we obtain an exact sequence

$$0 \longrightarrow \mathbf{Z}G/S \oplus M_A \longrightarrow (\mathbf{Z}G)^2 \longrightarrow \mathbf{Z}G \longrightarrow \mathbf{Z}G/A \longrightarrow 0$$

in which $\mathbf{Z}G/S$ is identified with the cyclic submodule of $(\mathbf{Z}G)^2$ generated by the element $(1+x, 0)$. Hence there is a commutative diagram with exact rows

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathbf{Z}G/S & \longrightarrow & \mathbf{Z}G & \xrightarrow{(1-x)} & \mathbf{Z}G & \longrightarrow & \mathbf{Z}G/S \longrightarrow 0 \\
 & & \downarrow{\scriptstyle \lambda_A} & & \downarrow{\scriptstyle (1,0)} & & \| & & \downarrow{\scriptstyle \varepsilon_A} \\
0 & \longrightarrow & \mathbf{Z}G/S \oplus M_A & \longrightarrow & (\mathbf{Z}G)^2 & \longrightarrow & \mathbf{Z}G & \longrightarrow & \mathbf{Z}G/A \longrightarrow 0
\end{array}$$

in which ε_A is the augmentation map, and λ_A is the canonical inclusion corresponding to the direct sum decomposition.

Replacing A by B in the above gives rise to another, similar diagram. From the two diagrams one can obtain a new commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbf{Z}G/S & \longrightarrow & \mathbf{Z}G & \longrightarrow & \mathbf{Z}G & \longrightarrow & \mathbf{Z}G/S & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle (\lambda_A, -\lambda_B)} & & \downarrow{\scriptstyle (1,0,-1,0)} & & \downarrow{\scriptstyle (1,-1)} & & \downarrow{\scriptstyle (\varepsilon_A, -\varepsilon_B)} & & \\ 0 & \longrightarrow & (\mathbf{Z}G/S \oplus M_A) \oplus (\mathbf{Z}G/S \oplus M_B) & \longrightarrow & (\mathbf{Z}G)^4 & \longrightarrow & (\mathbf{Z}G)^2 & \longrightarrow & \mathbf{Z}G/A \oplus \mathbf{Z}G/B & \longrightarrow & 0 \end{array}$$

in which the rows are exact and the vertical maps are all injective. It follows that the induced sequence of cokernels is also exact. This has the form

$$0 \longrightarrow \mathbf{Z}G/S \oplus (M_A \oplus M_B) \longrightarrow (\mathbf{Z}G)^3 \longrightarrow \mathbf{Z}G \longrightarrow \mathbf{Z} \longrightarrow 0.$$

Hence $\mathbf{Z}G/S$ is a direct summand of $K_2(\mathscr{F})$ for some $\mathbf{Z}G$-free resolution $\mathscr{F} \twoheadrightarrow \mathbf{Z}$.

Remark. The group G of Example 2 is expressed as an amalgamated free product with finite amalgamated subgroup, a construction which does not satisfy the hypotheses of Proposition 10. This suggests that it may be possible to weaken the restriction on edge groups in Proposition 10. However, it does not seem easy to find a general condition under which the conclusion of the proposition continues to hold. That some restriction on edge groups is necessary is shown by the following example.

EXAMPLE 3. $G = SL_2(\mathbf{Z}) \cong (\mathbf{Z}/4\mathbf{Z}) \underset{(\mathbf{Z}/2\mathbf{Z})}{*} (\mathbf{Z}/6\mathbf{Z})$

is an infinite group with non-trivial torsion and non-trivial centre. By Corollary 5.3, there is no non-trivial subgroup S such that $\mathbf{Z}G/S$ occurs as a direct summand in any $K_n(\mathscr{P})$.

Appendix on relative cohomology

Suppose G is a finite group of order r. Then the ordinary (co)-homology functors of G with arbitrary $\mathbf{Z}G$-module coefficients are annihilated by r.

Now let G be a cyclic group of order $2n$ and S the subgroup of order n. Then $G/S = C \cong \mathbf{Z}/2\mathbf{Z}$. We shall compute the group $H^2(G, S; \mathbf{Z}C)$. In the sequel, we follow the notation of [2].

Consider the exact sequence $\Delta \rightarrowtail \mathbf{Z}G/S \twoheadrightarrow \mathbf{Z}$, where $\mathbf{Z}G/S = \mathbf{Z}C$. Then $\Delta \cong \tilde{\mathbf{Z}}$, the $\mathbf{Z}G$-module $\mathbf{Z}$ with nontrivial G-action. In terms of a generator g for G we

can write down a free $\mathbf{Z}G$-resolution

$$\mathscr{F}: \cdots \longrightarrow \mathbf{Z}G \xrightarrow{1+g} \mathbf{Z}G \xrightarrow{\sum_{k=0}^{2n-1}(-1)^k g^k} \mathbf{Z}G \xrightarrow{1+g} \mathbf{Z}G \longrightarrow \tilde{\mathbf{Z}} \longrightarrow 0.$$

Now the complex $\mathrm{Hom}_{\mathbf{Z}G}(\mathscr{F}, \mathbf{Z}C)$ is of the form

$$\mathscr{F}' \cdots \longrightarrow \mathbf{Z}C \xrightarrow{1+c} \mathbf{Z}C \xrightarrow{n(1-c)} \mathbf{Z}C \xrightarrow{1+c} \mathbf{Z}C \longrightarrow \mathrm{Hom}_{\mathbf{Z}G}(\tilde{\mathbf{Z}}, \mathbf{Z}C) \longrightarrow 0$$

where $c = gS$ is a generator for C. In view of the definitions in [2], we may use $\mathscr{F}'$ to compute

$$H^2(G, S; \mathbf{Z}C) \cong \mathrm{Ext}^1_{\mathbf{Z}G}(\tilde{\mathbf{Z}}; \mathbf{Z}C) \cong \tilde{\mathbf{Z}}/n\tilde{\mathbf{Z}}.$$

Thus, in general, $H^2(G, S; -)$ is *not* annihilated by the index of S in G. However, our investigation will lead to the following result:

COROLLARY 2. *Suppose (G, S) is a Frobenius pair and S has index k in G. Then k annihilates the relative cohomology functors $H^q(G, S; -)$ for all $q \geq 1$.*

The proof of this result involves a slight digression.

If U is a subgroup of G, and M is a $\mathbf{Z}G$-module, let $\varepsilon : \mathbf{Z}G \otimes_U M \to M$ denote the "evaluation" map, $\varepsilon(g \otimes m) = g \cdot m$. Let σ denote the natural isomorphism

$$\mathrm{Ext}^*_{\mathbf{Z}G}(\mathbf{Z}G \otimes_U M; -) \xrightarrow{\cong} \mathrm{Ext}^*_{\mathbf{Z}U}(M; -).$$

Then we will refer to the natural transformation

$$\mathrm{res} = \sigma \circ \varepsilon^* : \mathrm{Ext}^*_{\mathbf{Z}G}(M; -) \longrightarrow \mathrm{Ext}_{\mathbf{Z}U}(M; -)$$

as *restriction*, since in the special case $M = \mathbf{Z}$, it is the usual restriction $H^*(G; -) \to H^*(U; -)$.

Similarly, if U has finite index k in G and $\{g_1, \ldots, g_k\}$ is a left transversal, we can define a map

$$\eta : M \longrightarrow \mathbf{Z}G \otimes_U M \quad \text{by} \quad \eta(m) = \sum_{r=1}^{k} g_r \otimes g_r^{-1} m.$$

Transfer or *corestriction* is then defined to be the natural transformation

$$\mathrm{cor} = \eta^* \circ \sigma^{-1} : \mathrm{Ext}^*_{\mathbf{Z}U}(M; -) \longrightarrow \mathrm{Ext}^*_{\mathbf{Z}G}(M; -).$$

Just as in the usual case $M = \mathbf{Z}$, we now have the following result.

PROPOSITION. *Let G be a group and U a subgroup of finite index k in G. Then for any pair of $\mathbf{Z}G$-modules M, A and any integer $q \geqslant 0$, the composite*

$$\mathrm{cor} \circ \mathrm{res} : \mathrm{Ext}^q_{\mathbf{Z}G}(M; A) \longrightarrow \mathrm{Ext}^q_{\mathbf{Z}U}(M, A) \longrightarrow \mathrm{Ext}^q_{\mathbf{Z}G}(M, A)$$

is just multiplication by the integer k.

Proof. One checks that $\varepsilon \circ \eta$ is multiplication by k on M.

Now let $(G, \boldsymbol{S})$ be a group pair and U a subgroup of G. As in Section 2, we can define a family $\boldsymbol{T}$ of subgroups of U such that there is a $\mathbf{Z}U$-module isomorphism $\mathbf{Z}G/\boldsymbol{S} \cong \mathbf{Z}U/\boldsymbol{T}$. (See also [2, pp. 305–306]). As above, we denote by Δ the kernel of the augmentation $\mathbf{Z}G/\boldsymbol{S} \to \mathbf{Z}$ and define $H^k(G, S; -) \cong \mathrm{Ext}^{k-1}(\Delta, -)$ following [2]. Specializing the above to $M = \Delta$, we obtain natural transformations

$$\mathrm{res} : H^*(G; \boldsymbol{S}; -) \longrightarrow H^*(U, \boldsymbol{T}; -)$$

and, provided U has finite index in G,

$$\mathrm{cor} : H^*(U, \boldsymbol{T}; -) \longrightarrow H^*(G, \boldsymbol{S}; -).$$

The Proposition now yields the following consequence:

COROLLARY 1. *Let $(G, \boldsymbol{S})$ and $(U, \boldsymbol{T})$ be pairs of groups as above and suppose U has finite index k in G. Then for any integer $q \geqslant 1$, the natural transformation*

$$\mathrm{cor} \circ \mathrm{res} : H^q(G, \boldsymbol{S}; -) \longrightarrow H^q(U, \boldsymbol{T}; -) \longrightarrow H^q(G, \boldsymbol{S}; -)$$

is just multiplication by k.

If (G, S) is a Frobenius pair, we set $U = S$ in Corollary 1. Then the family $\boldsymbol{T}$ consists of one copy of S together with $(k-1)$ copies of the trivial group. Since $H^q(S, \{S, 1, \ldots, 1\}; -) \cong H^q(S, S; -) = 0$ for $q \geqslant 1$, Corollary 2 follows because multiplication by k factors through 0.

REFERENCES

[1] BABAKHANIAN, A., Cohomological methods in group theory. Marcel Dekker Inc., New York 1972.
[2] BIERI, R. and ECKMANN, B., *Relative homology and Poincaré duality for group pairs.* J. Pure and Appl. Alg. *13* (1978) 277–319.
[3] CARTAN, H. and EILENBERG, S., Homological algebra. Princeton University Press 1956.
[4] GRUENBERG, K. W., Free abelianised extensions of finite groups. Homological Group Theory (C. T. C. Wall, ed.), 71–104; London Math. Soc. Lecture Note Series. No. 36, Cambridge Univ. Press 1979.
[5] GRUENBERG, K. W. and ROGGENKAMP, K. W., *Decomposition of the augmentation ideal and of the relation modules of a finite group.* Proc. London. Math. Soc. (3) *31* (1975) 149–166.
[6] HOWIE, J. and SCHNEEBELI, H. R., *Groups of finite quasi-projective dimension.* Comment. Math. Helvetici *54* (1979) 615–628.
[7] HUEBSCHMANN, J., *Cohomology theory of aspherical groups and of small cancellation groups.* J. Pure and Appl. Alg. *14* (1979) 137–143.
[8] SWAN, R. G., *Periodic resolutions for finite groups.* Annals of Math. *72* (1960) 267–291.

Department of Mathmetics
University of Edinburgh
Edinburgh EH9 3JZ
Scotland

Margelstrasse 14
CH-5430 *Wettingen*
Switzerland

Received October 3, 1980

Comment. Math. Helvetici **56** (1981) 465–473

0010-2571/81/003465-9$01.50 + 0.20/0
© 1981 Birkhäuser Verlag, Basel

Aspherical four-manifolds and the centres of two-knot groups

JONATHAN A. HILLMAN

§1. Introduction

In [6], Hausmann and Kervaire have shown that any finitely generated abelian group is the centre of the group of an n-knot, for each $n \geq 3$. This result is in sharp contrast to the classical case ($n = 1$) when the centre is either infinite cyclic or trivial. In this note we shall consider the remaining case $n = 2$, our principal result being that if the centre of the group of a 2-knot has rank greater than 1 then the closed 4-manifold obtained by surgery on the knot is aspherical, and the centre is $\mathbf{Z}^2$. We use equivariant Poincaré duality to show that the homology of the universal cover of such a 4-manifold is a stably free module concentrated in degree 2, followed by an Euler characteristic counting argument to show that this module has rank 0, and so must be trivial.

In §2 we shall invoke a result of Kaplansky to show that nonzero stably free modules over certain rings have well defined, strictly positive rank, thus preparing the way for our counting argument. This is first used in §3 to verify a conjecture of Murasugi, for groups with torsion free centre, and to sketch a new proof of a theorem of Gottlieb. In §4 we derive our main result. Twistspinning the trefoil knot in several ways gives fibred 2-knots whose groups have centre $\mathbf{Z}$, $\mathbf{Z} \oplus \mathbf{Z}/2\mathbf{Z}$ or $\mathbf{Z}$ and we conjecture that no other nontrivial group may be the centre of a 2-knot group. We show also that the centre of the group of a 2-link with more than one component must be a torsion group, and we determine the 2-knot groups which contain an abelian subgroup of finite index. Finally, in §5 we give a necessary and sufficient condition for a closed 4-manifold to be aspherical, which is applied to the results of surgery on certain 2-knots.

I would like to thank the referee for his comments on the first version of this paper. This work was done under a grant from the Science Research Council at the University of Durham.

§2. Hopfian rings

A ring R is *prehopfian* if whenever m and n are integers such that $0 < m < n$ there is no epimorphism of R^m onto R^n; it is *hopfian* if every onto endomorphism

of a finitely generated free R-module is an automorphism. If M is a stably free module over a prehopfian ring R, so that $M \oplus R^b = R^a$ for some a, b then R-rank $M = a - b$ is well defined and non-negative; if R is hopfian and $M \neq 0$ then R-rank $M > 0$.

LEMMA. *Let S be the multiplicative system in a ring R generated by a central element s which is not a divisor of zero. If M is a stably free R-module and n an integer such that the localization M_S maps onto R_S^n with nontrivial kernel, then M maps onto R^n with nontrivial kernel. Hence R (pre)hopfian implies that R_S is (pre)hopfian.*

Proof. We shall prove the first assertion by induction on n. It is trivially true if $n = 0$. Suppose that $\phi : M_S \to R_S^n$ is an epimorphism with kernel K. On multiplying ϕ by a suitable power of s, we may assume that $\phi = \psi_S$ for some map $\psi : M \to R^n$ and that the map $\psi' : M \to R$ obtained by composing ψ with projection onto the k^{th} factor is onto, for some $k \leqslant n$. Let $N = \ker \psi'$ and let $\theta : N \to R^{n-1}$ be the composition of $\psi \mid N$ with the projection onto the complementary free summand of R^n. Then $M \approx N \oplus R$, so N is stably free, and θ_S is onto with kernel $\ker \theta_S = (\ker \theta)_S = (\ker \psi)_S$ nontrivial. By the hypothesis of induction there is an epimorphism $\pi : N \to R^{n-1}$ with nontrivial kernel. The map $\pi \oplus \mathrm{id}_R : N \oplus R$ $(\approx M) \to R^n$ gives an epimorphism of M onto R^n with nontrivial kernel. The second assertion follows immediately.

If we restate the definition of "prehopfian" in terms of matrices, we see immediately that any ring which maps nontrivially to a field is prehopfian, and hence integral group rings are prehopfian. Similarly any subring of a (pre)hopfian ring is (pre)hopfian. As Kaplansky has shown that for any group G the complex group algebra $\mathbf{C}[G]$ is hopfian (for a proof see [10]), integral group rings are in fact hopfian. (For the remarks in this paragraph I am indebted to Professor P. M. Cohn. See also [4]).

§3. Centres and Euler characteristic

On the evidence of his work on 1-relator groups, Murasugi conjectured that the centre of a finitely presentable group other than $\mathbf{Z}^2$ of deficiency $\geqslant 1$ is infinite cyclic or trivial, and is trivial if the group has deficiency $\geqslant 2$ [11]. In that paper he showed that this is true for the groups of links in S^3, which all have deficiency $\geqslant 1$. (The knots and links in S^3 whose groups have nontrivial centre has been determined by Burde, Zieschang and Murasugi [2, 3]). Here we shall show that it

holds for a much larger class of (finitely presentable) groups, including all those with torsion free centre.

The deficiency of a finite presentation $P = \langle x_1, \ldots x_a \mid r_1, \ldots r_b \rangle$ of a group G is def $P = a - b$; the deficiency of a finitely presentable group G is def $G = \max\{\text{def } P \mid P \text{ presents } G\}$. Such a finite presentation P determines a 2-dimensional cell complex X_P with 1 0-cell, a 1-cells and b 2-cells, which has fundamental group G, and Euler characteristic $\chi(X_P) = 1 - a + b = 1 - \text{def } P$. We shall let $\zeta(G)$ and Γ denote the centre of the group G and the integral group ring $\mathbf{Z}[G]$ respectively.

THEOREM 1. *Let X be a finite 2-dimensional cell complex with fundamental group G such that $\zeta(G)$ contains an element of infinite order. Then $\chi(X) \geq 0$, and $\chi(X) = 0$ if and only if X is aspherical.*

Proof. Let $p : \tilde{X} \to X$ be the universal cover of X. The cell structure of X may be lifted to a cell structure for $\tilde{X}$ which is invariant under the action of G via covering transformations, and the cellular chain complex of $\tilde{X}$ may then be regarded as a finite chain complex of free left Γ-modules

$$C_* : 0 \to C_2 \to C_1 \to C_0 \to 0$$

where $c_i = \Gamma\text{-rank } C_i$ is the number of i-cells of X. Since $\tilde{X}$ is simply connected, $H_0(C_*) = \mathbf{Z}$ and $H_1(X_*) = 0$, while $H_2(C_*) = H_2(\tilde{X}) \approx \pi_2(X)$ is a submodule of C_2.

Let z be an element of $\zeta(G)$ of infinite order, and let S be the central multiplicative system generated by $z - 1$. Then S contains no divisors of zero, so Γ embeds in the localization Γ_S, which by the lemma of §2 is hopfian. Since the augmentation module $\mathbf{Z}$ is annihilated by $g - 1$ for all g in G, localizing the chain complex C_* leads to an exact sequence

$$0 \to H_2(C_*)_S \to C_{2S} \to C_{1S} \to C_{0S} \to 0,$$

from which it follows that $H_2(C_*)_S$ is a stably free Γ_S-module of rank $c_0 - c_1 + c_2 = \chi(X)$, which must therefore be nonnegative. As $H_2(C_*)$ is a submodule of C_2, which embeds in C_{2S}, it is 0 if and only if $H_2(C_*)_S = 0$.
The theorem follows readily.

Remark. A similar argument gives an easy proof of Gottlieb's theorem that an aspherical finite cell complex whose fundamental group has a nontrivial centre has Euler characteristic 0 [5]. For the group must be torsion free [1; page 63] and we may localize the cellular chain complex of the universal covering space to get an exact sequence of finitely generated free modules over a hopfian ring.

COROLLARY. *Let G be a finitely presentable group with a central element of infinite order. Then* def $G \leq 1$. *If* def $G = 1$ *and G is neither $\mathbf{Z}$ nor $\mathbf{Z}^2$, then G has cohomological dimension 2, $\zeta(G) = \mathbf{Z}$ and the commutator subgroup G' is free.*

Proof. The theorem together with the paragraph preceding it imply directly all but the last two assertions, which then follow from Theorem 8.8 of [1] which asserts that if G has finite cohomological dimension, then c.d. $\zeta(G) \leq$ c.d. $G - 1$, with equality only if G' is free.

As the groups of links in S^3 are all torsion free and have deficiency ≥ 1, this corollary implies immediately the above-mentioned results of Murasugi [11].

§4. Centres of two-knot groups

A 2-knot is a locally flat *PL* embedding of the (oriented) 2-sphere S^2 in an oriented homotopy 4-sphere Σ^4. The exterior of such a knot $K: S^2 \to \Sigma^4$ is the compact bounded 4-manifold $X_K = \Sigma^4 - N$, where N is an open regular neighbourhood of the image of K, and the group of K is $G = \pi_1(X_K)$. The group G is also the fundamental group of the closed orientable 4-manifold $Y_K = X_K \cup S^1 \times D^3$ obtained from Σ^4 by surgery on K.

THEOREM 2. *Let M be a closed PL 4-manifold with fundamental group G, and suppose*

(i) *$\zeta(G)$ contains an element z of infinite order;*

(ii) *$H^s(G; \Gamma) = 0$ for $s \leq 2$.*

Then M is aspherical if and only if $\chi(M) = 0$.

Proof. Let $p: \tilde{M} \to M$ be the universal cover of M, and let C_* be the cellular chain complex of $\tilde{M}$ with respect to a cell structure lifted through p, as in Theorem 1. Let C^* denote the dual complex $\mathrm{Hom}_\Gamma(C_*, \Gamma)$, a cochain complex of free right Γ-modules. Then there are Poincaré duality isomorphisms $\bar{H}^p(C^*) \to H_{4-p}(C_*)$ (where $\bar{A}$ denotes the left Γ-module deduced from a right Γ-module A via the anti-involution of the group ring: $g \cdot a = w_1(g) a \cdot g^{-1}$ for all a in A and g in G) [12; page 23].

Since $\tilde{M}$ is simply connected $H_0(C_*) = \mathbf{Z}$ and $H_1(C_*) = 0$, while since G is infinite $\tilde{M}$ is an open 4-manifold and so $H_4(C_*) = 0$. Since $\mathrm{Ext}^s_\Gamma(\mathbf{Z}, \Gamma) = H^s(G; \Gamma) = 0$ for $s \leq 2$, it follows from the universal coefficient spectral sequence

$$\mathrm{Ext}^q_\Gamma(H_p(C_*), \Gamma) \Rightarrow H^{p+q}(C^*)$$

that $H^1(C^*)=0$ and that $H^2(C^*)\approx \mathrm{Hom}_\Gamma(H_2(C_*),\Gamma)$. Poincaré duality then implies that $H=H_2(C_*)$ is isomorphic to $\overline{H^*}$, where $H^*=\mathrm{Hom}_\Gamma(H,\Gamma)$, and $H_3(C_*)=0$. Thus to prove that $\tilde{M}$ is contractible it shall suffice to show that $H^*=0$.

Let S be the central multiplicative system generated by $z-1$. Then as in Theorem 1, Γ embeds in Γ_S, while the only nonzero homology of C_{*S} is H_S in degree 2. If B is any Γ_S-module $H^3(\mathrm{Hom}_{\Gamma_S}(C_{*S},B))=H_1(C_{*S}\otimes B)=0$ by Poincaré duality and the Künneth theorem. Hence as in [12; page 26] H_S is stably free and we may split the boundary maps of the complex C_{*S} to obtain an isomorphism $H_S\oplus C_{1S}\oplus C_{3S}\approx C_{0S}\oplus C_{2S}\oplus C_{4S}$, so Γ_S-rank $H_S=\chi(M)$.

Since Γ_S is hopfian, by the lemma of §2, $\chi(M)=0$ implies that $H_S^*=H_S=0$, and since Γ embeds in Γ_S, H^* embeds in H_S^* and so must also be 0. Hence $H=0$ and $\tilde{M}$ is contractible, so M is aspherical. As the converse is a particular case of Gottlieb's theorem ([5]–see the Remark in §3) the theorem is proven.

COROLLARY 1. *If K is a 2-knot whose group G contains a central subgroup C isomorphic to $\mathbf{Z}^2$ then Y_K is aspherical and $\zeta(G)=\mathbf{Z}^2$.*

Proof. The quotient G/C is infinite, for otherwise the commutator subgroup G' would be finite [7; page 102] and so C would map monomorphically to $G/G'=\mathbf{Z}$. A spectral sequence argument (as in [1; page 158]) then shows that $H^s(G;\Gamma)=0$ for $s\leqslant 2$ and so by the theorem Y_K is aspherical and G has cohomological dimension 4. Therefore G is torsion free and c.d. $\zeta(G)\leqslant 3$, by Theorem 8.8 of [1]. The same theorem proves that c.d. $\zeta(G)=3$ only if G' is free, which cannot be the case, for otherwise c.d. $G\leqslant$ c.d. $G'+$c.d. $G/G'=2$. Since an abelian group of cohomological dimension $\leqslant 2$ must be either $\mathbf{Z}^2$ or a subgroup of the additive rational numbers [1; pages 101–110], the corollary follows.

The group of the 0-twist spun trefoil knot has centre $\mathbf{Z}$ but does not satisfy the second hypothesis of the theorem, and the result of surgery on this knot is not aspherical. The centres of the groups of the 1-, 3- and 6-twist spun trefoil knots are $\mathbf{Z}$, $\mathbf{Z}\oplus\mathbf{Z}/2\mathbf{Z}$ and $\mathbf{Z}^2$ respectively [14]. We conjecture that no other nontrivial group can be the centre of a 2-knot group.

We shall now consider briefly the centres of link groups. In the classical case the centre must be 1, $\mathbf{Z}$ or $\mathbf{Z}^2$, the latter being realised only by the (2-component) abelian link [2]. The argument of Hausmann and Kervaire may be readily modified to show that any finitely generated abelian group is the centre of the group of some μ-component boundary n-link, for each $\mu\geqslant 1$ and $n\geqslant 3$. If $n=2$ and $\mu\geqslant 2$ the centre must have rank 0 for otherwise an argument as in Theorem 2 leads to the contradictory conclusion that the localization H_S of the homology of

the universal cover of the manifold obtained by surgery on the link is a stably free Γ_S-module of Γ_S-rank $2-2\mu<0$. (To obtain the contradiction we need only know that Γ_S is prehopfian, and that $H^t(G;\Gamma)=0$ for $t\leqslant 1$).

We shall say that a group G just contains a subgroup A if the index of A in G finite.

COROLLARY 2. *Let M be a closed 4-manifold whose fundamental group Π just contains an abelian subgroup A. Then $\chi(M)\geqslant 0$, and if $\chi(M)=0$ the subgroup A has rank 1, 2 or 4.*

Proof. By passing to a finite covering space we may assume that M is orientable and that Π is a free abelian group. We may clearly assume that $\beta=\text{rank }\Pi$ is greater than 1. Let $\tilde{M}$ be the universal covering space of M. If S is as in the theorem the localized spectral sequence collapses (even if $\beta=1$ or 2) and Poincaré duality then implies that $H_2(\tilde{M})_S$ is the only nonzero localised homology module, and its $\mathbf{Z}[\Pi]$-rank is $\chi(M)$ which therefore must be nonnegative. If $\beta>2$ and $\chi(M)=0$ the theorem implies that M is aspherical and hence that $\beta=4$.

The manifolds $S^1\times S^3$, $S^1\times S^1\times S^2$, and $S^1\times S^1\times S^1\times S^1$ have Euler characteristic 0 and fundamental group $\mathbf{Z}$, $\mathbf{Z}^2$ and $\mathbf{Z}^4$ respectively. (It may be shown that they are determined up to homotopy type among orientable 4-manifolds by these properties).

COROLLARY 3. *Let K be a 2-knot whose group just contains an abelian subgroup A. Then either Y is aspherical (rank $A=4$) or the commutator subgroup G' is finite (rank $A=1$).*

Proof. It is readily verified that a group with abelianization $\mathbf{Z}$ just contains an infinite cyclic subgroup if and only if its commutator subgroup is finite, while no such group can just contain a subgroup isomorphic to $\mathbf{Z}^2$ (or $\mathbf{Z}^3$). The assertions now follow on applying Corollary 2 to the manifold Y.

Necessary conditions for a group with finite commutator subgroup to be a 2-knot group were given in [8], and these conditions were shown to be sufficient by Yoshikawa [14].

If G is a 2-knot group which just contains $\mathbf{Z}^4$, then the commutator subgroup G' just contains $\mathbf{Z}^3$ and is torsion free, by the corollary. Therefore it is the fundamental group of a flat Riemannian 3-manifold [13; page 103], which must be orientable [9]. Of the six possible such groups listed on page 117 of [13], only $\mathbf{Z}^3$ and the group G_6 admit meridional [8] automorphisms; moreover no such automorphism of $\mathbf{Z}^3$ can have finite order. The group G_6 is presented by

$\{\alpha, \gamma \mid \alpha\gamma^2\alpha^{-1}\gamma^2 = \gamma\alpha^2\gamma^{-1}\alpha^2 = 1\}$ and the automorphism $\Phi: G_6 \to G_6$ sending α to $\alpha\gamma$ and γ to α is meridional and of order 6. Since the group $G = G_6 *_\Phi$ presented by $\{G_6, t \mid tgt^{-1} = \Phi(g)$ for all g in $G_6\}$ is torsion free and just contains $\mathbf{Z}^4$, there is a flat 4-manifold M with fundamental group G [13; page 103]. As in [8] we may then obtain a 2-knot with group G by surgery on M. (In fact we may assume that M is the mapping torus of a diffeomorphism of a flat 3-manifold with fundamental group G_6, and hence that the knot is fibred.) Which high dimensional knot groups just contain abelian groups?

§5. Aspherical four-manifolds

If $f: M \to N$ is an $(n-1)$-connected degree 1 map between closed orientable $2n$-manifolds with fundamental group Π, the only obstruction to its being a homotopy equivalence is $H_n(f) = \ker(f_*: \pi_n(M) \to \pi_n(N))$, which is a stably free $\mathbf{Z}[\Pi]$-module, by lemma 2.3 of [12]. Arguing as in Theorem 2 we may show that $\mathbf{Z}[\Pi]$-rank $H_n(f) = (-1)^n(\chi(M) - \chi(N))$ and so f is a homotopy equivalence if it is an integral homology equivalence. In the nonorientable case an $(n-1)$-connected map f is a homotopy equivalence if it preserves the orientation character and induces an integral homology equivalence on the orientation covers.

In this section we shall adapt the argument outlined above to a case in which it is not known a priori that the map has degree 1.

THEOREM 3. *Let M be a closed 4-manifold with fundamental group F. Then the classifying map $f: M \to K(F, 1)$ is a homotopy equivalence if and only if F is a Poincaré duality group of formal dimension 4 and orientation character $w = w_1(M)$, and f induces an equivalence on homology with w-twisted rational coefficients.*

Proof. As these conditions are clearly necessary, we need only show that they are sufficient. By passing to the covering spaces associated with $\ker w$ we may suppose that M and F are orientable. Up to homotopy type we may suppose also that f is an inclusion and that $K(F, 1)$ is a finite cell complex containing M as a subcomplex. Let C_*, D_* and E_* be the cellular chain complexes of the universal covers $\tilde{M}$, $\tilde{K}(F, 1)$ and $(\tilde{K}(F, 1), \tilde{M})$ respectively, with respect to their natural F-invariant cell structures, and let $\Phi = \mathbf{Z}[F]$. Since F is a 4-dimensional duality group $H^s(F; \Phi) = 0$ for $s < 4$, and Poincaré duality together with the universal coefficient spectral sequence then give an isomorphism $H_2(C_*) \approx \overline{\mathrm{Hom}}_\Phi(H_2(C_*), \Phi)$ as in Theorem 2, while $H_s(C_*) = 0$ if $s \neq 0$ or 2. Since Φ maps monomorphically to $\mathbf{Q}[F]$, $H_2(C_*)$ embeds in $\mathbf{Q} \otimes H_2(C_*) = H_2(\tilde{M}; \mathbf{Q})$. As $\tilde{K}(F, 1)$ is contractible, the only possibly nontrivial homology module of $\mathbf{Q} \otimes E_*$ is

$\mathbf{Q}\otimes H_3(E_*) = \mathbf{Q}\otimes H(E_*) = H_2(\tilde{M}; \mathbf{Q})$, which is a stably free $\mathbf{Q}[F]$-module by Lemma 2.3 of [12]. Since f induces a rational homology equivalence the Euler characteristics of M and $K(F, 1)$ are equal. As these are also the Euler characteristics of C_* and D_*, and as the sequence

$$0 \to C_* \to D_* \to E_* \to 0$$

is exact, the Euler characteristic of E_* is 0. Therefore the stably free $\mathbf{Q}[F]$-module $H_2(\tilde{M}; \mathbf{Q}) = \mathbf{Q}\otimes H_3(E_*)$ has rank 0, and, by Kaplansky's theorem, must in fact be 0. Thus $H_2(C_*) = 0$ and f is a homotopy equivalence.

COROLLARY 4. *Let M be an orientable 4-manifold such that $H = \pi_1(M)$ is an orientable Poincaré duality group of formal dimension 4 and suppose that the cohomology ring $H^*(M; \mathbf{Q})$ is generated by $H^1(M; \mathbf{Q})$. Then M is aspherical.*

Proof. The classifying map from M to $K(H, 1)$ is clearly a rational (co)homology equivalence, and so we may apply the theorem.

COROLLARY 2. *Let K be a 2-knot whose group G is an orientable Poincaré duality group of formal dimension 4 such that $G' \neq G''$. Then the closed 4-manifold Y_K obtained by surgery on K is aspherical.*

Proof. Since $G' \neq G''$, there is a field E such that $E\otimes(G'/G'') \neq 0$. An argument as in Theorem 1 of [9] using the Wang sequence and Milnor duality with coefficients E instead of $\mathbf{Z}/2\mathbf{Z}$ shows that the classifying map $c : Y_K \to K(G, 1)$ has nonzero degree and hence induces a rational homology equivalence, so the above theorem applies.

Remark. If $(\mathbf{Z}/2\mathbf{Z})\otimes(G'/G'') \neq 0$ the orientability of G follows from the other conditions [9]. Is the corollary still true if $G' = G''$? In [9] we used an argument similar to that of Theorem 1 to show that if a 2-knot group G is a Poincaré duality group with one end (i.e. other than $\mathbf{Z}$) such that $\mathbf{Z}[G]$ embeds nicely in a division ring then the result of surgery on the knot is aspherical. An examination of that argument shows that the first condition can be weakened to $H^s(G; \Gamma) = 0$ for $s \leqslant 2$. Can the second condition be relaxed to G being torsion free, or dispensed with entirely?

REFERENCES

[1] BIERI, R., *Homological Dimensions of Discrete Groups*, Queen Mary College Lecture Notes, 1975.

[2] BURDE, G. and MURASUGI, K., *Links and Seifert fibre spaces*, Duke Math. J. 37 (1970), 89–93.

[3] —— and ZIESCHANG, H., *Eine Kennzeichnung der Torus-Knoten*, Math. Ann. *167* (1966), 169–175.
[4] COHN, P. M. *Some remarks on the invariant basis property*, Topology *4* (1965), 215–228.
[5] GOTTLIEB, D., *A certain subgroup of the fundamental group*, Amer. J. Math. *87* (1965), 840–856.
[6] HAUSMANN, J. C. and KERVAIRE, M. A., *Sur les centres des groupes de noeuds multidimensionelles*, C.R. Acad. Sc. Paris t. *278* (1978), Ser. A, 699–702.
[7] HEMPEL, J., *3-Manifolds*, Annals of Math. Studies 84, Princeton University Press, Princeton, N.J., 1976.
[8] HILLMAN, J. A., *High dimensional knot groups which are not two-knot groups*, Bull. Austral. Math. Soc. *17* (1977), 449–462.
[9] ——, *Orientability, asphericity and two-knot groups*, Houston J. Math. *6* (1980), 67–76.
[10] MONTGOMERY, M. S., *Left and right inverses in group algebras*, Bull. Amer. Math. Soc. *75* (1969), 539–540.
[11] MURASUGI, K., *On the centre of the group of a link*, Proc. Amer. Math. Soc. *14* (1965), 1052–1057. Erratum 18 (1966), 1142.
[12] WALL, C. T. C., *Surgery on Compact Manifolds*, Academic Press, New York – London, 1970.
[13] WOLF, J. A., *Spaces of Constant Curvature* (3rd edition), Publish or Perish Inc., Berkeley, California, 1977.
[14] YOSHIKAWA, K., *On 2-knots with the finite commutator subgroup*, Math. Sem. Notes Kobe Univ. *8* (1980), 321–330.

Department of Mathematics
University of Durham
Durham DH1 3LE
U.K.

Current address
Department of Mathematics
University of Texas
Austin TX 78712
U.S.A.

Received February 24, June 23 1981

Comment. Math. Helvetici **56** (1981) 474–486 0010-2571/81/003474-13$01.50 + 0.20/0
© 1981 Birkhäuser Verlag, Basel

Über Homöomorphismen n-dimensionaler Henkelkörper und endliche Erweiterungen von Schottky-Gruppen

BRUNO ZIMMERMANN

Es sei V_g^n ein n-dimensionaler Henkelkörper vom Geschlecht g (d.h. V_g^n entsteht aus dem n-Ball D^n durch Ankleben von g 1-Henkeln). In [12] haben wir gezeigt, daß sich jede endliche Gruppe von Isotopieklassen von Homöomorphismen eines 3-dimensionalen Henkelkörpers V_g^3 repräsentieren läßt durch eine isomorphe Gruppe von Homöomorphismen, und daß die analoge Aussage für Homotopieklassen von Homöomorphismen von V_g^3 im allgemeinen nicht richtig ist. Dies führte dazu, das analoge Problem in höheren Dimensionen zu betrachten.

Die Gruppe der Homotopieklassen von Homöomorphismen (bzw. Homotopieäquivalenzen) von V_g^n ist isomorph zur äußeren Automorphismengruppe Out F_g der freien Gruppe F_g vom Rang g, $F_g = \pi_1 V_g^n$. Es sei $H(V_g^n)$ die Gruppe aller Homöomorphismen von V_g^n und $p : H(V_g^n) \to \text{Out } F_g$ die kanonische Abbildung, die jedem Homöomorphismus von V_g^n den induzierten äußeren Automorphismus der Fundamentalgruppe zuordnet. Wir zeigen in dieser Arbeit, daß es eine Dimension $n = n(g)$ gibt, so daß sich jede endliche Untergruppe A von Our F_g realisieren läßt durch die Operation auf $\pi_1 V_g^n \cong F_g$ einer isomorphen Gruppe $\tilde{A}$ von Homöomorphismen von V_g^n (d.h. $p : H(V_g^n) \to \text{Out } F_g$ induziert einen Isomorphismus von $\tilde{A}$ auf A). Dabei können wir $\tilde{A}$ sogar als Gruppe von Isometrien wählen, falls wir V_g^n durch eine geeignete $(n-1)$-dimensionale Schottky-Gruppe uniformieren (in Abhängigkeit von der gegebenen endlichen Untergruppe A von Out F_g; im Inneren von V_g^n gibt es dann eine Riemannsche Metrik konstanter negativer Krümmung). Dies ist äquivalent zu folgender Aussage. Ist $1 \to F_g \hookrightarrow G \to A \to 1$ die durch A bestimmte Erweiterung von F_g, so ist G isomorph zu einer Gruppe konformer Abbildungen der $(n-1)$-Sphäre S^{n-1}, wobei $F_g \subset G$ einer Schottky-Gruppe entspricht.

Als Haupthilfsmittel dient uns die Charakterisierung endlicher Erweiterungen von freien Gruppen in [3].

Die Ordnung einer endlichen Gruppe A orientierungserhaltender Homöomorphismen eines 3-dimensionalen Henkelkörpers V_g^3 ist beschränkt

durch $12(g-1)$ [12]. Wir bestimmen im 4. Abschnitt alle Gruppen G, die als Erweiterung auftreten, falls diese Schranke angenommen wird.

1. Endliche Erweiterungen freier Gruppen

Eine Gruppe G ist genau dann endliche Erweiterung einer endlich erzeugten freien Gruppe, wenn G eine HNN-Erweiterung der folgenden Form ist [3]:

$$\langle t_1, \ldots, t_r, K \mid \text{Relationen von } K,\ t_1 L_1 t_1^{-1} = M_1, \ldots, t_r L_r t_r^{-1} = M_r \rangle, \tag{1.1}$$

wo K ein Baumprodukt endlich vieler endlicher Gruppen ist (den Eckengruppen von K, wobei über den Kantengruppen amalgamiert wird) und jede der Gruppen L_i, M_i Untergruppe einer Eckengruppe ist. Den K zugrunde liegenden Baum werden wir ebenfalls mit K bezeichnen, für die Ecken und Eckengruppen bzw. Kanten und Kantengruppen verwenden wir jeweils dieselben Symbole.

Ist F_g eine freie Untergruppe von G vom Rang g und Index j, so gilt folgende Formel ([3]; vgl. auch Bemerkung 3.3):

$$g = j\left(\frac{1}{f_1} + \cdots + \frac{1}{f_r} + \frac{1}{e_1} + \cdots + \frac{1}{e_s} - \frac{1}{v_1} - \cdots - \frac{1}{v_{s+1}}\right) + 1; \tag{1.2}$$

dabei ist f_i die Ordnung von L_i bzw. M_i, die e_j bzw. v_k sind die Ordnungen der Kanten- bzw. Eckengruppen von K.

Wir verwenden in §3 folgendes Lemma. Teil(a) wurde in [1] bewiesen. Wir geben einen kurzen Beweis unter Verwendung von 1.1 und 1.2 (die wir ohnehin später benötigen).

1.3. LEMMA. (a) *Sei α ein Automorphismus einer freien Gruppe F_g vom Rang g, so daß α^p ein innerer Automorphismus von F_g ist, p prim. Operiert α trivial auf $F_g/[F_g, F_g] \cong \mathbf{Z}^g$, so ist α ein innerer Automorphismus von F_g.*

(b) *Die Ordnungen endlicher Untergruppen der äußeren Automorphismengruppe* Out F_g *von F_g sind beschränkt für festes g.*

(c) *Es gibt nur endlich viele Isomorphietypen von Gruppen G, die endliche effektive Erweiterung einer freien Gruppe F_g von festem Rang g sind.*

Beweis. (a) Der Automorphismus α bestimmt eine Erweiterung $1 \to F_g \hookrightarrow G \xrightarrow{\pi} \mathbf{Z}_p \to 1$ zum abstrakten Kern $(\mathbf{Z}_p, F_g, \Omega : \mathbf{Z}_p \to \text{Out } F_g$ mit $\Omega(1) = [\alpha])$. Die Gruppe G ist von der Form 1.1, wobei jede Eckengruppe von K isomorph zu $\mathbf{Z}_p$ und jede Kantengruppe trivial ist. Da α trivial auf $F_g/[F_g, F_g]$ operiert, hat

$G/[G, G]$ den freien Rang g, und es gibt genau g Erzeugende $t_1, \ldots, t_g$ in der Darstellung 1.1 von G. Aus 1.2 folgt

$$g \geq p\left(g \cdot \frac{1}{p} + s - (s+1)\frac{1}{p}\right) + 1 = g + p \cdot s - s,$$

wobei Gleichheit genau dann gilt, wenn $L_i \cong \mathbf{Z}_p$ für alle i. Es folgt $s = 0$ und $L_i \cong \mathbf{Z}_p$ für alle i, d.h. $G = \langle x, t_1, \ldots, t_g \mid x^p = 1, t_i x t_i^{-1} = x \rangle$ mit $\pi(x) = 1 \in \mathbf{Z}_p$. Durch Konjugation operiert x trivial auf G bzw. F_g und repräsentiert die Automorphismenklasse von α.

(b) folgt aus (a) und der entsprechenden Aussage für endliche Untergruppen von $\mathrm{GL}(g, \mathbf{Z})$ (vgl. z.B. [10, Satz 191], wo gezeigt wird, daß der Kern der kanonischen Abbildung $\mathrm{GL}(g, \mathbf{Z}) \to \mathrm{GL}(g, \mathbf{Z}/p\mathbf{Z})$ torsionsfrei ist für $p \geq 3$ prim).

(c) Jede Zahl e_i in 1.2 ist echter Teiler zweier v_j, wobei jedes v_j vorkommt. Wegen

$$g \geq j\left(\left(\frac{1}{e_1} - \frac{1}{v_{\alpha_1}} - \frac{1}{v_{\alpha_2}}\right) + \left(\frac{1}{e_2} - \frac{1}{v_{\alpha_3}}\right) + \cdots + \left(\frac{1}{e_s} - \frac{1}{v_{\alpha_{s+1}}}\right)\right) + 1 \geq s$$

ist die Anzahl der Eckengruppen von K für festes g beschränkt. Für die Anzahl r der t_i gilt:

$$g \geq j\left(r \cdot \frac{1}{j} - 1\right) + 1 = r - j + 1.$$

Wegen (b) ist daher auch die Anzahl der t_i beschränkt. Daher gibt es nur endlich viele Möglichkeiten, eine Gruppe G mit Darstellung 1.1 und der Eigenschaft in (c) zu konstruieren. □

Wir notieren noch eine Folgerung aus [11, Theorem 3.2.9] und 1.3 (a):

1.4. KOROLLAR. *Sei* $1 \to F_g \hookrightarrow G \to A \to 1$ *eine endliche effektive Erweiterung einer freien Gruppe vom Rang* g. *Dann ist* $G/[F_g, F_g]$ *isomorph zu einer kristallographischen Gruppe des* $\mathbf{R}^g$.

Wir wissen nicht, ob man jede kristallographische Gruppe auf diese Art erhalten kann.

2. *n*-dimensionale Schottky-Gruppen

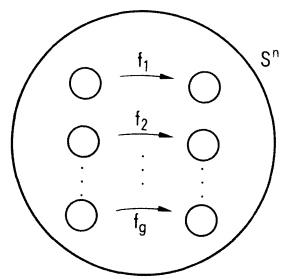
Es sei $M(n)$ die Gruppe der Kreisverwandschaften der n-Sphäre S^n. Dann ist $M(n)$ gleichzeitig die Gruppe aller konformen Abbildungen von S^n und läßt sich

kanonisch fortsetzen auf den $(n+1)$-Ball D^{n+1}, $S^n = \partial D^{n+1}$. Fassen wir das Innere $\mathring{D}^{n+1}$ von D^{n+1} als Modell des hyperbolischen $(n+1)$-Raumes auf ("Poincaré-Modell"), so ist $M(n)$ die Gruppe der Isometrien von $\mathring{D}^{n+1}$ und wird erzeugt von Spiegelungen an n-Sphären, die senkrecht auf dem Rand von D^{n+1} stehen (hierzu und zum folgenden vgl. [2] und [7]).

Die orthogonale Gruppe $O(n+1)$ läßt sich auffassen als die Untergruppe aller Abbildungen aus $M(n)$, die den Nullpunkt in D^{n+1} fixieren. Da jede endliche Untergruppe von $M(n)$ im Inneren von D^{n+1} einen Fixpunkt hat [2, Proposition 13], sind endliche Untergruppen von $M(n)$ konjugiert zu Untergruppen der orthogonalen Gruppe $O(n+1) \subset M(n)$.

Entsprechend dem klassischen Fall $M(2)$ lassen sich die Elemente von $M(n)$ einteilen in parabolische, elliptische und loxodromische Transformationen [2]. Eine loxodromische Transformation besitzt eine Verschiebungsachse im Inneren von D^{n+1} und genau zwei Fixpunkte auf dem Rand (die Endpunkte der Achse). Eine loxodromische Transformation heißt hyperbolisch, wenn sie eine Translation entlang ihrer Verschiebungsachse ist, d.h. keine zusätzliche Drehung um diese Achse induziert (hyperbolische Transformationen sind das Produkt zweier Spiegelungen an Spiegelkreisen senkrecht zur Achse).

Eine n-dimensionale Schottky-Gruppe S_g^n ist eine diskrete Untergruppe von $M(n)$, die von g loxodromischen Transformationen $f_1, \ldots, f_g$ frei erzeugt wird, die ein System von $2g$ $(n-1)$-Sphären in S^n folgendermaßen abbilden:



Dabei soll jeweils das Innere jeder Urbildsphäre (das keine der anderen $2g-1$ Sphären enthält) auf das Äußere der Bildsphäre abgebildet werden; die beiden Fixpunkte von f_i liegen im Inneren von Urbild- und Bildsphäre. Diese Definition ist in völliger Analogie zur Definition der klassischen Schottky-Gruppen (vgl. z.B. [6, S. 265] und [5, Chapter IV. 2]).

Die Menge der Limespunkte $\Omega(S_g^n)$ einer Schottky-Gruppe S_g^n ist eine nirgendwo dichte perfekte Teilmenge von S^n (es sei $g \geq 2$). Im Inneren von D^{n+1} liegen keine Limespunkte von S_g^n, und $(D^{n+1} - \Omega(S_g^n))/S_g^n$ ist ein n-dimensionaler Henkelkörper V_g^n. Wir sagen, daß S_g^n den Henkelkörper V_g^n uniformisiert. Dann

haben wir im Inneren von V_g^n eine Riemannsche Metrik konstanter negativer Krümmung, und wir können von Isometrien von V_g^n sprechen.

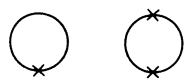
3. Geometrische Realisierung endlicher Untergruppen von Out F_g

Im nächsten Satz legen wir die PL-Kategorie zugrunde.

3.1. SATZ. (a) *Sei* Out F_g *die äußere Automorphismengruppe einer freien Gruppe* F_g *vom Rang* g. *Dann gibt es eine natürliche Zahl* $n = n(g)$, *so daß sich jede endliche Untergruppe* A *von Out* F_g *realisieren läßt durch eine isomorphe Gruppe von Homöomorphismen des* n-dimensionalen *Henkelkörpers* V_g^n *vom Geschlecht* g.

(b) *Es gilt* $n(g) \to \infty$ *für* $g \to \infty$, *d.h. es gibt keine natürliche Zahl* n, *die für alle* g *gleichzeitig die Folgerung von* (a) *erfüllt.*

Beweis. (a) Die Gruppe $A \subset \text{Out } F_g$ bestimmt eine Erweiterung $1 \to F_g \hookrightarrow G \xrightarrow{\pi} A \to 1$, wobei G eine Darstellung der Form 1.1 hat. Nach [8] operiert G auf einem Baum B, wobei $F_g \subset G$ frei operiert und die Stabilisatoren der Ecken bzw. Kanten von B durch die entsprechenden Gruppen in K gegeben sind (vgl. den Beweis von Satz 3.2). Dann ist $B_1 := B/F_g$ ein endlicher Graph mit $\pi_1 B \cong F_g$, und die Gruppe $A \cong G/F_g$ operiert auf B und induziert auf $F_g \cong \pi_1 B$ die vorgegebene Operation. Durch geeignetes Unterteilen können wir erreichen, daß B_1 keine Teilgraphen folgender Form enthält:



Es sei $n+2$ die Anzahl der Eckpunkte von B_1. Wir bilden die Eckpunkte von B_1 bijektiv auf die Eckpunkte des Standard-n-Simplexes $\Delta^{n+1} \subset \mathbf{R}^{n+1}$ ab. Dann gibt es eine eindeutig bestimmte kanonische Einbettung von B_1 in $\partial\Delta^{n+1}$, und die Operation von A auf B_1 läßt sich zu einer Operation von A auf $\partial\Delta^{n+1}$ fortsetzen. Die reguläre Umgebung $U(B_1)$ von B_1 in der zweiten Normalunterteilung von $\partial\Delta^{n+1}$ ist homöomorph zu einem n-dimensionalen Henkelkörper V_g^n, und A operiert auf $U(B_1)$ und realisiert die gewünschte Operation auf $F_g \cong \pi_1(U(B_1))$. Indem wir B_1 in höherdimensionale Simplices einbetten, sehen wir, daß sich A auch in jeder Dimension größer als n realisieren läßt. Wegen 1.3(c) gibt es daher die gewünschte Zahl $n = n(g)$.

(b) Wir uniformisieren V_g^n durch eine Schottky-Gruppe S_g^{n-1}, die auf D^n operiert. Für $n \geq 3$ ist $\partial D^n - \Omega(S_g^{n-1}) = S^{n-1} - \Omega(S_g^{n-1})$ universelle Überlagerung von

$\partial V_g^n = \#_g S^1 \times S^{n-2}$. Daher ist nach [4, Theorem 2.1] S^{n-1} die Endenkompaktifizierung der universellen Überlagerung von ∂V_g^n. Ist A eine endliche Untergruppe von Out F_g mit zugehöriger Erweiterung $1 \to F_g \hookrightarrow G \to A \to 1$, und operiert A auf V_g^n, so läßt sich $A \mid \partial V_g^n$ liften zu einer Operation von G auf der universellen Überlagerung von ∂V_g^n. Daher operiert G auch auf der Endenkompaktifizierung S^{n-1} und damit auch alle (endlichen) Eckengruppen von K. Wählen wir umgekehrt als Eckengruppen von K endliche Gruppen, die nicht auf S^{n-1} operieren (z.B. $(\mathbf{Z}_p)^\nu$ für großes ν), so kann $A \cong G/F_g$ nicht auf V_g^n operieren. □

Der folgende Satz zeigt, daß sich A sogar durch eine Gruppe von Isometrien realisieren läßt, wenn wir V_g^n durch eine geeignete Schottky-Gruppe S_g^{n-1} uniformieren.

3.1. SATZ. *Sei A eine endliche Untergruppe von* Out F_g *und* $1 \to F_g \hookrightarrow G \xrightarrow{\pi} A \to 1$ *die durch A bestimmte Erweiterung. Dann gibt es für geeignetes n einen Isomorphismus von G auf eine Untergruppe von $M(n-1)$, wobei $F_g \subset G$ auf eine Schottky-Gruppe abgebildet wird.*

Beweis. Wir legen das Standard-$(n+1)$-Simplex $\Delta^{n+1} \subset \mathbf{R}^{n+1}$ symmetrisch zum Nullpunkt und projezieren seinen Rand $\partial\Delta^{n+1}$ radial vom Nullpunkt aus auf die n-Sphäre $S^n \subset \mathbf{R}^{n+1}$. Wir erhalten eine Triangulation von S^n, wobei die Kanten der Triangulation auf Großkreisen (Geodätischen bzgl. der sphärischen Metrik) liegen. Analog zum Beweis von 3.1 betten wir den Graphen B_1 in S^n ein und setzen A zu einer Gruppe orthogonaler Transformationen von S^n fort.

Die Gruppe G operiert auf B mit einem Fundamentalbereich der Form

$$F = K \cup \{r \text{ Kanten } T_1, \ldots, T_r\},$$

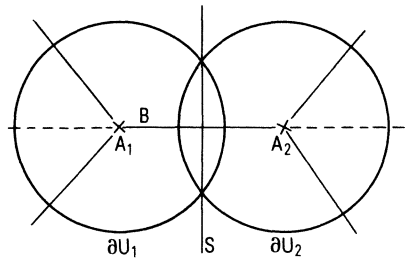
wobei der Anfangspunkt jeder Kante T_i in K liegt und $t_i \in G$ den Endpunkt von T_i nach K abbildet (vgl. [8, §4]). Die Stabilisatoren in G der Ecken bzw. Kanten von K sind die entsprechenden Ecken- bzw. Kantengruppen, der Stabilisator der Kante T_i ist die Gruppe L_i, der Kante $t_i(T_i)$ (die nicht zu F gehört) die Gruppe M_i.

Wir projezieren F nach $B_1 = B/F_g$ und bezeichen das Bild mit F_1. Dabei wird K isomorph auf einen Baum in F_1 abgebildet, den wir wieder mit K bezeichnen. Ist $t_i \in F_g$, so liegt der Endpunkt der Projektion $\tilde{T}_i$ von T_i in K. Alle Stabilisatoren in G werden isomorph auf die entsprechenden Stabilisatoren in A abgebildet, wobei der Isomorphismus durch π gegeben ist. Kein Element $\neq 1$ aus A operiert trivial auf B_1. Wir unterscheiden 3 Fälle.

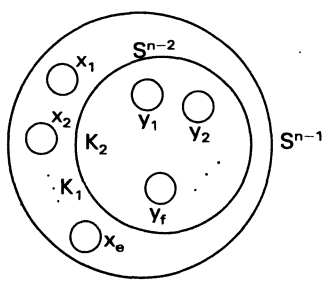
1. *Fall.* $G = A_1 *_B A_2$, d.h. es kommen keine t_i vor und K hat die Form

$A_1 \overset{*}{-} \overset{*}{B} \overset{*}{-} A_2$.

Nach obigem ist K in $B_1 \subset S^n$ eingebettet. Es sei δ der Abstand der Punkte A_1 und A_2. Wir wählen Kugelumgebungen U_1 und U_2 mit Radius $\frac{2}{3}\delta$ (bezüglich der sphärischen Metrik) der Punkte A_1 und A_2. Dann bildet der Stabilisator in A von A_i (den wir ebenfalls mit A_i bezeichen) die Umgebung U_i auf sich ab ($i = 1, 2$), und A_i operiert konform auf $\partial U_i \cong S^{n-1}$.



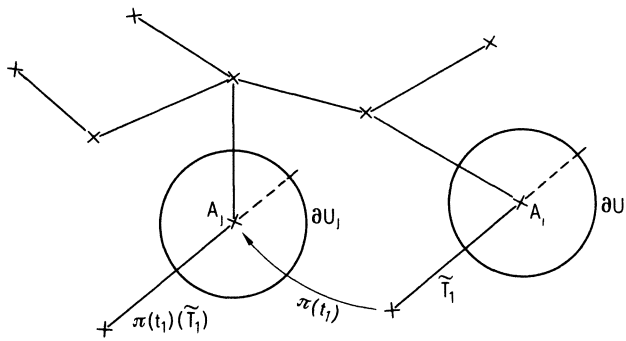
Die Operationen des Stabilisators $B \subset A_1, A_2$ auf ∂U_1 und ∂U_2 sind konform äquivalent (durch die Spiegelung an einer $(n-1)$-Sphäre S, die senkrecht auf dem Mittelpunkt der Kante B steht und die $(n-2)$-Sphäre $\partial U_1 \cap \partial U_2$ enthält). Kein Element $\neq 1$ aus B operiert trivial auf ∂U_i, da dasselbe für die Operation von A auf B_1 gilt. Wir können daher annehmen, daß A_1 und A_2 konform auf der $(n-1)$-Sphäre S^{n-1} operieren mit gemeinsamer Untergruppe B. Die Gruppe B hat 2 Fixpunkte R_1 und R_2 auf S^{n-1}, die den Schnittpunkten des Großkreises, auf dem die Kante B liegt, mit ∂U_1 bzw. ∂U_2 entsprechen. Wir können ohne Einschränkung annehmen, daß A_2 orthogonal auf S^{n-1} operiert, und daß R_1 und R_2 Diametralpunkte sind. Wir wählen eine hinreichend kleine ε-Umgebung V von R_1, so daß $a(V) \cap V = \varnothing$ für alle Elemente a aus $A_1 - B$ bzw. $A_2 - B$. Es gilt $B(V) = V$. Es sei τ die Spiegelung an der $(n-2)$-Sphäre $S^{n-2} = \partial V$, mit $\tau(R_1) = R_2$ und $\tau b \tau^{-1} = b$ für alle Elemente b aus B. Es sei G' die von A_1 und $\tau A_2 \tau^{-1}$ erzeugte Untergruppe von $M(n-1)$. Es seien $S^{n-2} = \partial V$, $x_1, \ldots, x_e$ die Bilder von S^{n-2} unter A_1, und K_1 das Ringgebiet in S^{n-1}, das von S^{n-2}, $x_1, \ldots, x_e$ begrenzt wird. Entsprechend seien S^{n-2}, $y_1, \ldots, y_f$ die Bilder von S^{n-2} unter $\tau A_2 \tau^{-1}$ und K_2 das Ringgebiet begrenzt von S^{n-2}, $y_1, \ldots, y_f$.



Aus der Operation von G' auf der Menge $G'(S^{n-2})$ von Sphären in S^{n-1} folgt $G' = A_1 *_B \tau A_2 \tau^{-1} \underset{\alpha}{\cong} G$ (keine Normalform des freien Produkts mit Amalgam operiert trivial auf S^{n-2}; für einen analogen Schluß im Falle von Schottky-Gruppen vgl. [5, Chapter IV. 2]). Wir betrachten die Menge von Ringgebieten $M := G'(K_1) \cup G'(K_2)$. Dann erhalten wir einen Fundamentalbereich von $\alpha(F_g) \subset G'$, indem wir mit $K_1 \cup K_2$ starten und sukzessive $|G/F_g|$ an $K_1 \cup K_2$ angrenzende Ringgebiete in M hinzufügen. Es folgt, daß $\alpha(F_g)$ eine Schottky-Gruppe ist. Der Graph B läßt sich aus M folgendermaßen zurückgewinnen: Es sei $\tilde{B}$ der Graph, dessen Eckpunkte die Ringgebiete aus M sind. Zwei Eckpunkte seien durch eine Kante verbunden, falls die entsprechenden Ringgebiete benachbart sind. Dann operiert G' auf $\tilde{B}$ und es gilt $(\tilde{B}, G') \cong (B, G)$ (äquivarianter Isomorphismus). Betrachten wir die Operation von G' auf $D^n (S^{n-1} = \partial D^n)$, so haben A_1 bzw. $\tau A_2 \tau^{-1}$ Fixpunkte P_1 bzw. P_2 im Inneren von D^n, $P_1 \neq P_2$, die den Mittelpunkten von U_1 bzw. U_2 entsprechen. Ist e die geodätische (nichteuklidische) Strecke zwischen P_1 und P_2, so ist $(G'(e), G')$ ein Baum, der ebenfalls äquivariant isomorph ist zu (B, G).

2. *Fall.* $G = K$, d.h. es kommen keine t_i vor. Wir gehen induktiv vor. Wir starten mit einer Kante von K und konstruieren die entsprechende Gruppe wie im 1. Fall. Danach nehmen wir sukzessive je eine weitere angrenzende Kante hinzu und konstruieren die entsprechende Gruppe wieder analog zum 1. Fall; der einzige Unterschied ist, daß wir an Stelle von A_1 im 1. Fall jetzt schon ein Baumprodukt haben, daß bereits konform auf S^{n-1} operiert. Einen Fundamentalbereich von F_g konstruiert man analog zum 1. Fall aus Ringgebieten, die den Eckpunkten des Baumes B entsprechen.

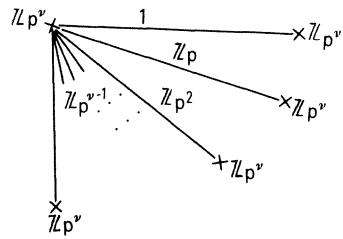
3. *Fall.* Es kommen t_i vor.
Wir konstruieren K wie im 2. Fall und führen den Beweis für t_1 (für die restlichen t_i geht er vollkommen analog). Wir betrachten die Teilmenge $F_1 \subset B_1 \subset S^n$.



Es gilt $L_1 \subset A_i$, $M_1 \subset A_j$, und $\pi(t_1) \in A$ ist eine orthogonale Abbildung mit $\pi(t_1)L_1\pi(t_1^{-1}) = M_1$, wobei L_1 die Kante $\tilde{T}_1$ fixiert und M_1 die Kante $\pi(t_1)(\tilde{T}_1)$ (eventuell $\pi(t_1) = 1$). Die Gruppe K operiert bereits auf S^{n-1}. Die Gruppe $L_1 \subset K$ hat 2 Fixpunkte S_1, S_2 auf S^{n-1}, die den Schnittpunkten von $\tilde{T}_1$ mit ∂U_i entsprechen. Analog hat M_1 zwei Fixpunkte T_1, T_2 auf S^{n-1}, die den Schnittpunkten von $\pi(t_1)(\tilde{T}_1)$ mit ∂U_j entsprechen. Es gibt eine konforme Abbildung k_1: $S^{n-1} \to S^{n-1}$ mit $k_1 L_1 k_1^{-1} = M_1$ und $k_1(S_1) = T_1$, $k_1(S_2) = T_2$ (diese Abbildung erhalten wir aus $\pi(t_1)$). Es sei f_1 eine hyperbolische Transformation von S^{n-1} mit Fixpunkten T_1 und T_2. Dann gilt $f_1 M_1 f_1^{-1} = M_1$, da M_1 die Achse von f im Inneren von $D^n (S^{n-1} = \partial D^n)$ punktweise festläßt. Es folgt $f_1^n k_1 L_1 k_1^{-1} f_1^{-n} = M_1$, und für großes n bildet $f_1^n k_1$ eine hinreichend kleine Umgebung V_1 von S_1 auf das komplement einer hinreichend kleinen Umgebung V_2 von T_2 ab. Dann können wir für t_1 konforme Abbildung $f_1^n k_1$ wählen (dies folgt wiederum aus der Normalform für Elemente in HNN-Gruppen). Einen Fundamentalbereich von F_g erhält man, indem man aus der Vereinigung der Ringgebiete, die den Ecken von K entsprechen, das Innere $\mathring{V}_1$ und $\mathring{V}_2$ von V_1 und V_2 wegnimmt, und endlich viele angrenzende unter G äquivalente Gebiete hinzunimmt. □

Nicht jede endliche Gruppe A von Homöomorphismen von V_g^n ist äquivalent zu einer Gruppe von Isometrien, da es endliche Gruppen von Homöomorphismen von Sphären gibt, die zu keiner orthogonalen Gruppe äquivalent sind.

In 3.1(b) haben wir gezeigt, daß die Dimension, in der sich eine endliche Untergruppe A von Out F_g realisieren läßt, mit g beliebig groß werden kann. Beschränken wir uns auf Realisierungen durch Gruppen von Isometrien, so tritt der Fall schon für endliche zyklische Gruppen auf. Wählen wir $\nu = \nu(n)$ hinreichend groß, so operiert folgendes Baumprodukt nicht auf S^n (man betrachte die Fixpunktmenge von $\mathbf{Z}_p$ in D^{n+1}).



3.3. *Bemerkung.* Aus der Operation von A auf dem Graphen B_1 mit $\pi_1 B_1 \cong F_g$ ergibt sich leicht ein Beweis der Formel 1.2. Es gilt $g = 1 - b + a$, wobei b die Anzahl der Ecken und a die Anzahl der Kanten von B_1 ist. Die Anzahl der Ecken von B_1, die unter A äquivalent sind zu einer Ecke A_i von $K \subset B_1$, ist gleich

$|A/A_i|$. Die Anzahl der Kanten äquivalent zu einer Kante B_j von K ist gleich $|A/B_j|$. Die Anzahl der Kanten äquivalent zu $\tilde{T}_k$ ist gleich $|A/L_k|$. Hieraus ergibt sich sofort 1.2 (falls F_g normal in G liegt; andernfalls gehen wir zu einem Normalteiler über und wenden das Verfahren zweimal an).

Satz 3.2 läßt sich verallgemeinern für endliche effektive Erweiterungen der freien Gruppe F_∞ von abzählbarem Rang. Eine Schottky-Gruppe S^n_∞ mit abzählbar vielen Erzeugenden sei analog zum endlichen Fall definiert durch abzählbar viele Paare von Sphären (die sich natürlich in S^n häufen). Für jede endliche effektive Erweiterung $1 \to F_\infty \hookrightarrow G \to A \to 1$ läßt sich nach [9] die Gruppe G einbetten in eine Gruppe $\bar{G}$, die endliche Erweiterung einer freien Gruppe F_g für geeignetes g ist. Wir können ohne Einschränkung annehmen, daß diese Erweiterung ebenfalls effektiv ist. Aus 3.2 folgt dann:

3.4. KOROLLAR. *Sei* $1 \to F_\infty \hookrightarrow G \to A \to 1$ *eine endliche effektive Erweiterung der freien Gruppe* F_∞ *von abzählbar unendlichem Rang. Dann läßt sich* G *realisieren als Untergruppe der Gruppe* $M(n)$ *konformer Abbildungen von* S^n *für geeignetes* n, *wobei* $F_\infty \subset G$ *einer Schottky-Gruppe entspricht.*

Die Sätze 3.2 bzw. 3.4 bleiben richtig für beliebige endliche Erweiterungen freier Gruppen (d.h. nicht notwendigerweise effektive). Die Operation von A auf B_1 braucht in diesem Fall nicht effektiv zu sein, wir können B_1 aber in eine hinreichend hochdimensionale Sphäre S^n einbetten und die Operation von A auf B_1 zu einer effektiven Operation von A auf S^n fortsetzen.

4. Homöomorphismen 3-dimensionaler Henkelkörper

Die Ordnung einer endlichen Gruppe A orientierungserhaltender Homöomorphismen eines 3-dim. Henkelkörpers V^3_g ist beschränkt durch $12(g-1)$ (es sei $g>1$); dies ist eine Folgerung aus der Formel von Riemann-Hurwitz ([12, Abschnitt 2]). Wir können V^3_g durch eine Schottky-Gruppe S^2_g uniformisieren, so daß A als Gruppe konformer Abbildungen von ∂V^3_g operiert (wobei wir jetzt auch Schottky-Gruppen zulassen, die durch $2g$ Jordankurven an Stelle der $2g$ 1-Sphären definiert sind, sog. nichtklassiche Schottky-Gruppen). Liften wir A in die Überlagerung $S^2 - \Omega(S^2_g)$ von ∂V^3_g, so erhalten wir eine Erweiterung

$$1 \to S^2_g \hookrightarrow G \to A \to 1,$$

wobei G eine Gruppe orientierungserhaltender konformer Abbildungen von S^2

ist, d.h. eine Untergruppe von $\mathrm{PSL}_2(\mathbf{C})$. Die universelle Überlagerung von ∂V_g^3 bzw. $S^2 - \Omega(S_g^2)$ ist die hyperbolische Ebene $D^2 = \{z \in \mathbf{C} \mid |z| < 1\}$ (Poincaré-Modell). Liften wir G nach D^2, so erhalten wir eine Fuchssche Gruppe F und eine kanonische Projektion $F \to G$. Genau dann hat A maximale Ordnung $12(g-1)$, wenn F isomorph ist zur Viereckgruppe

$$V := \langle x, y, z, w \mid x^2 = y^2 = z^2 = w^3 = xyzw = 1 \rangle$$

(vgl. [12]).

4.1. SATZ. *Genau dann hat A maximale Ordnung $12(g-1)$, wenn G isomorph ist zu einer Gruppe der folgenden Form:*

$$D_2 *_{\mathbf{Z}_2} D_3, \qquad D_3 *_{\mathbf{Z}_3} A_4, \qquad D_4 *_{\mathbf{Z}_4} S_4, \qquad D_5 *_{\mathbf{Z}_5} A_5;$$

dabei bezeichnet D_i die Diedergruppe mit $2i$ Elementen, A_4 die Tetraedergruppe, S_4 die Oktaedergruppe und A_5 die Dodekaedergruppe. Jede der vier Gruppen operiert auf S^2 als Untergruppe von $\mathrm{PSL}_2(\mathbf{C})$ mit einer Schottky-Gruppe als Normalteiler von endlichem Index.

Beweis. Hat A maximale Ordnung $12(g-1)$, so haben wir eine surjektive Abbildung $V \to G$ mit torsionsfreiem Kern. Die Gruppe G hat eine Darstellung der Form 1.1. Da $V/[V, V]$ endlich ist, treten keine Erzeugenden t_i auf. Da V von 3 Elementen der Ordnung 2 erzeugt wird, hat G eine Darstellung der Form $A_1 *_B A_2$, $A_1 A_2$, B endlich. Die Gruppen A_1 bzw. A_2 haben Fixpunkte P_1 bzw. P_2 im Inneren von D^3 ($S^2 = \partial D^3$), mit $P_1 \neq P_2$. Die nichteuklidische Gerade durch P_1 und P_2 wird von B punktweise festgelassen. Daher hat B mindestens 2 Fixpunkte auf S^2. Es folgt, daß B eine zyklische Gruppe ist (die anderen Möglichkeiten sind D_i, A_4, S_4 und A_5, die alle keinen gemeinsamen Fixpunkt besitzen).

Aus 1.2 folgt

$$\frac{1}{|B|} - \frac{1}{|A_1|} - \frac{1}{|A_2|} = \frac{1}{12}.$$

Einfaches Nachrechnen ergibt, daß es für $\{|A_1|, |A_2|\}$ nur die folgenden geometrisch realisierbaren Möglichkeiten gibt: $\{4, 6\}$ falls $B \cong \mathbf{Z}_2$, $\{6, 12\}$ falls $B \cong \mathbf{Z}_3$, $\{8, 24\}$ falls $B \cong \mathbf{Z}_4$ und $\{10, 60\}$ falls $B \cong \mathbf{Z}_5$. Dies entspricht den 4 Fällen des Satzes.

Wir konstruieren eine Operation von $D_2 *_{\mathbf{Z}_2} D_3$ auf S^2. Es sei $D^3 \subset \mathbf{R}^3$ der 3-Ball und α bzw. β die Drehung um 120 bzw. 180 Grad um die x-Achse und γ

die Drehung um 180 Grad um die z-Achse. Es sei A_1 die von α und γ und $\tilde{A}_2$ die von β und γ erzeugte Gruppe orthogonaler Abbildungen von D^3 bzw. S^2. Es sei τ die Spiegelung an einem kleinen Kreis um einen der beiden Fixpunkte von γ auf S^2. Dann gilt $\tau\gamma\tau^{-1}=\gamma$, und die von A_1 und $A_2:=\tau\tilde{A}_2\tau^{-1}$ erzeugte Gruppe G konformer Abbildungen von S^2 ist isomorph zu $D_2 *_{\mathbf{Z}_2} D_3$. In allen anderen Fällen geht die Konstruktion analog.

Schottky-Untergruppen finden wir als Kerne z.B. der Abbildungen

$$D_2 *_{\mathbf{Z}_2} D_3 \to D_3 \oplus \mathbf{Z}_2 \ (g=2), \qquad D_2 *_{\mathbf{Z}_2} D_3 \to S_4 \ (g=3),$$
$$D_3 *_{\mathbf{Z}_3} A_4 \to S_4 \ (g=3), \qquad D_4 *_{\mathbf{Z}_4} S_4 \to S_4 \ (g=3),$$
$$D_5 *_{\mathbf{Z}_5} A_5 \to S_5 \ (g=6).$$

Durch Übergang zu charakteristischen Untergruppen der Kerne finden wir Schottky-Normalteiler von beliebig hohem Rang g. □

4.2. KOROLLAR. *Es gibt jeweils unendlich viele Werte von* g, *für die die obere Schranke* $12(g-1)$ *einer endlichen Gruppe orientierungserhaltender Homöomorphismen von* V_g^3 *angenommen bzw. nicht angenommen wird.*

Beweis. Wir haben noch zu zeigen, daß für unendlich viele Werte von g die obere Schranke $12(g-1)$ nicht angenommen wird. Es sei $g=p+1$, wobei $p>12$ eine Primzahl sei, so daß 3 kein Teiler von $p-1$ ist. Wir nehmen an, daß es eine Gruppe von Homöomorphismen A der maximalen Ordnung $12p$ für $g=p+1$ gibt. Es sei

$$1 \to S_g^2 \hookrightarrow G \xrightarrow{\pi} A \to 1$$

die entsprechende Erweiterung. Nach einem Sylow-Satz besitzt A einen Normalteiler $\mathbf{Z}_p$ der Ordnung p. Wir haben eine Abbildung $\tilde{\pi}: G \to A \to A/\mathbf{Z}_p$ mit torsionsfreiem Kern, und $A/\mathbf{Z}_p$, hat die Ordnung 12. Es folgt $G \cong D_2 *_{\mathbf{Z}_2} D_3$ und $A/\mathbf{Z}_p \cong D_3 \oplus \mathbf{Z}_2$ (durch Nachprüfen aller Möglichkeiten). Wegen $(p, 12)=1$ ist A nach einem Satz von Schur das semidirekte Produkt von $\mathbf{Z}_p$ mit $D_3 \oplus \mathbf{Z}_2$, dessen Elemente wir in der Form (a, b) schreiben. Es operiert $\mathbf{Z}_3 \subset D_3$ trivial auf $\mathbf{Z}_p$, da $\mathbf{Z}_p$ wegen $(p-1, 3)=1$ keinen nicht-trivialen Automorphismus der Ordnung 3 besitzt. Es ist $\pi \mid D_2$ und $\pi \mid D_3$ jeweils ein Isomorphismus aufs Bild. Außerdem gilt $\pi((123))=(0,(123)) \in A$ $((123) \in D_3 \cong S_3)$. Es sei $x \in D_2 \cong \mathbf{Z}_2 \oplus \mathbf{Z}_2$, so daß $\tilde{\pi}(x)$ den $\mathbf{Z}_2$-Faktor von $D_3 \oplus \mathbf{Z}_2$ erzeugt. Dann kommutiert $\pi(x)$ mit $\pi((123))$ und daher mit $\pi(D_3)$. Da A von $\pi(D_3)$ und $\pi(x)$ erzeugt wird, folgt $A \cong D_3 \oplus \mathbf{Z}_2$, was ein Widerspruch ist. □

LITERATUR

[1] BAUMSLAG, G. und TAYLOR, T., The center of groups with one defining relator. Math. Ann. *175* (1968), 315–319.
[2] GREENBERG, L., Discrete subgroups of the Lorentz group. Math. Scand. *10* (1962), 85–107.
[3] KARRAS, A., PIETROWSKI, A. und SOLITAR, D., Finite and infinite extensions of free groups. J. Austral. Math. Soc. *16* (1972), 458–466.
[4] KULKARNI, R. S., Some topological aspects of Kleinian groups. Amer. J. Math. *100* (1978), 897–911.
[5] LEHNER, J., Discontinuous groups and automorphic functions. Mathem. Surveys, Amer. Math. Soc. 1964.
[6] MARDEN, A., Geometrically finite Kleinian groups, in: Discrete groups and automorphic functions. Academic Press 1977.
[7] MOSTOW, G. D., Quasi-conformal mappings in n-space and the rigidity of hyperbolic space forms. Publ. Inst. Haute Etudes Sci. *34* (1968), 55–104.
[8] SERRE, J. P., Arbres, amalgames et SL_2, Astérisque *46*, 1977.
[9] SCOTT, G. P., An embedding theorem for groups with a free subgroup of finite index. Bull. London Math. Soc. *6* (1974), 304–306.
[10] SPEISER, A., Die Theorie der Gruppen endlicher Ordnung. Springer, Berlin 1937.
[11] WOLF, J. A., Spaces of constant curvature. McGraw-Hill 1967.
[12] ZIMMERMANN, B., Über Abbildungsklassen von Henkelkörpern. Arch. Math. *33* (1979), 379–382.

Ruhr Universität Institut f. Mathematik
Postfach 102148
D-4630 *Bochum* 1

Eingegangen den 24. August 1980

Comment. Math. Helvetici **56** (1981) 487–523 0010–2571/81/004487–37$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Affine manifolds with nilpotent holonomy

DAVID FRIED, WILLIAM GOLDMAN, MORRIS W. HIRSCH

Introduction

An affine manifold is a differentiable manifold together with an atlas of coordinate charts whose coordinate changes extend to affine automorphisms of Euclidean space. These charts are called affine coordinates. A map between affine manifolds is called affine it its expression in affine coordinates is the restriction of an affine map between vector spaces. Thus we form the category of affine manifolds and affine maps.

Let M be a connected affine manifold of dimension $n \geq 1$, locally isomorphic to the vector space E. Its universal covering $\tilde{M}$ inherits a unique affine structure for which the covering projection $\tilde{M} \to M$ is an affine immersion. The group π of deck transformations acts on $\tilde{M}$ by affine automorphisms.

It is well known that there is an affine immersion $D : \tilde{M} \to E$, called the developing map. This follows, for example, from Chevalley's Monodromy Theorem; a proof is outlined in Section 2. Such an immersion is unique up to composition with an affine automorphism of E. Thus for every $g \in \pi$ there is a unique affine automorphism $\alpha(g)$ of E such that $D \circ g = \alpha(g) \circ D$. The resulting homomorphism $\alpha : \pi \to \mathrm{Aff}(E)$ from π into the group of affine automorphisms of E is called the affine holonomy representation. It is unique up to inner automorphisms of $\mathrm{Aff}(E)$. The composition $\lambda : \pi \to \mathrm{GL}(E)$ is called the linear holonomy representation. The affine structure on M is completely determined by the pair (D, α).

M is called complete when D is a homeomorphism. This is equivalent to geodesic completeness of the connection on M (in which parallel transport is locally defined by affine charts as ordinary parallel transport in E). It is notorious that compactness does not imply completeness.

The main results of this paper are about affine manifolds whose affine holonomy groups $\alpha(\pi)$ are nilpotent. An important class of such manifolds are the affine nilmanifolds $\pi \backslash G$. Here π is a discrete subgroup of a simply connected nilpoint Lie group G. It is assumed that G has a left-invariant affine structure; the space of right cosets of π then inherits an affine structure.

A tensor field on an affine manifold M is called polynomial if its components in affine cordinates are polynomial functions. In particular there is an exterior algebra of polynomial exterior differential forms on M, closed under exterior differentiation.

Our main results are summarized as follows.

THEOREM A. *Let M be a compact affine manifold whose affine holonomy group is nilpotent. Then the following conditions are equivalent:*

(a) *M is complete*
(b) *the developing map is surjective*
(c) *the linear holonomy is unipotent*
(d) *the linear holonomy preserves volume*
(e) *the affine holonomy is irreducible*
(f) *the affine holonomy is indecomposable*
(g) *M is a complete affine nilmanifold*
(h) *M has a polynomial volume form*
(i) *M is orientable and the de Rham cohomology of M is the cohomology of the complex of polynomial exterior forms.*

In Section 1 we collect several algebraic facts about affine representations and cohomology which will be used throughout the rest of the paper. In particular we introduce the notations of irreducibility and indecomposability of an affine representation. Section 2 is geometric and discusses affine structures and their holonomy and development. The parallelism on an affine manifold is used to describe several classes of tensors which have special descriptions in affine coordinates. Left-invariant structures on Lie groups, which provide many examples of affine manifolds, are introduced.

In Section 3 we define another large class of affine manifolds: the radiant manifolds. These are characterized by the property that the affine holonomy has a stationary point. By choosing the origin to be such a point we identify the affine holonomy with the linear holonomy. More generally, for any affine manifold M we define a cohomology class c_M which vanishes when precisely M is radiant. This radiance obstruction is one of our principal tools.

Section 4 uses the cohomology theory developed in Section 1 to obtain several results about nonradiant affine manifolds having nilpotent affine holonomy group.

In Section 5 we prove a technical result about nilpotent linear groups.

In Section 6, we prove the equivalence of (a)–(f) in Theorem A, using techniques of Section 4 and the technical lemma from Section 5. (See 6.11). Deformation of affine structures is discussed, and some solvable examples are given.

In Section 7 we prove that every compact complete affine manifold with nilpotent fundamental group is an affine nilmanifold, establishing (g) of Theorem A. Finally, in Section 8 we discuss polynomial tensors on affine manifolds with nilpotent holonomy, completing the proof of Theorem A.

We have greatly profited from conversations with John Smillie. Many of the results here were first proved in his thesis under the assumption of abelian affine holonomy. We are indebted to him for many valuable ideas.

1. Affine representations and cohomology of groups

Throughout this paper E denotes the real n-dimensional vector space $\mathbf{R}^n$, $n \geq 1$. The group of linear automorphisms of E is $\mathrm{GL}(E)$. An *affine map* $E \to E$ is the composition of a linear map and a translation. The group of affine automorphisms of E is denoted by $\mathrm{Aff}(E)$.

Let $g \in \mathrm{Aff}(E)$. There are unique elements $L \in \mathrm{GL}(E)$ and $b \in E$ such that $g(x) = Lx + b$. We call L the *linear part* of g and b *the translational part* of g. Notice that $L = dg(x)$, the derivative of g at x, for all $x \in E$; and $b = g(0)$.

An *affine representation* of a group G is a homomorphism $\alpha : G \to \mathrm{Aff}(E)$. For each $g \in G$ the affine automorphism $\alpha(g)$ decomposes into a linear part $\lambda(g)$ and a translational part $u(g)$. In this way E becomes a G-module via λ and $u : G \to E$ is a *crossed homomorphism*, or *cocycle*, for λ: for all $x \in E$, $g, h \in G$,

$$\alpha(g)(x) = \lambda(g)x + u(g) \tag{1}$$

$$u(gh) = u(g) + \lambda(g)u(h) \tag{2}$$

Conversely, given any (linear) representation $\lambda : G \to \mathrm{GL}(E)$, and $u : G \to E$ satisfying (2), formula (1) defines an affine representation of G having linear part λ.

It is readily verified that $y \in E$ is a stationary point for α – that is, $\alpha(g)(y) = y$ for all $g \in G$ – if and only if

$$u(g) = y - \lambda(g)y \tag{3}$$

for all $g \in G$. In this case u is called a *principal crossed homomorphism*, or a *coboundary*, and we write $u = \delta y$. It is easily proved that $u = \delta y$ if and only if α is conjugate to the linear action λ by the translation $T_y : E \to E$, $T_y(x) = x - y$ (that is, $\lambda(g) = T_y \circ \alpha(g) \circ T_y^{-1}$ for all g).

We assume familiarity with the language of G-modules and their cohomology

groups: whenever a linear representation $\lambda : G \to \mathrm{GL}(E)$ is given, E is called a G-module, and vector spaces $H^i(G : E_\lambda)$ are defined for $i = 0, 1, \ldots$. Often the subscript λ will be omitted. The only G-modules we consider are defined by real finite-dimensional representations.

Of primary interest are H^0 and H^1. By definition $H^0(G : E_\lambda)$ is the space of stationary points of λ, while $H^1(G; E_\lambda)$ is the quotient space of crossed homomorphism for λ modulo principal crossed homomorphisms. The element of $H^1(G; E_\lambda)$ represented by u is the cohomology class of u, denoted by $[u]$. One may interpret $H^1(G; E_\lambda)$ as the set of translational conjugacy classes of affine representations of G having linear part λ.

The *radiance obstruction* of α is the cohomology class $c_\alpha = [u] \in H^1(G; E)$ where u is the translational part of α. The following simple observation will be very useful: $c_\alpha = 0$ *if and only if* α *is conjugate by a translation to its linear part*; *and also* $c_\alpha = 0$ *if and only if* α *has a stationary point.*

Also of interest are the homology groups $H_i(G; E_\lambda)$, of which we need only H_0. For a general definition the reader is referred to Atiyah–Wall [3]. The homology group $H_0(G; E_\lambda)$ is by definition E/K where K is the linear subspace spanned by

$$\{x - \lambda(g)x : x \in E \text{ and } g \in G\}.$$

Let $E^* = \mathrm{Hom}(E, \mathbf{R})$ denote the dual space of E. The *contragredient representation* of λ is

$$\lambda^t : G \to \mathrm{GL}(E^*)$$
$$\lambda^t(g) : w \mapsto w \circ \lambda(g^{-1}), \qquad w \in E^*.$$

It is readily proved that $H^0(G; E^*)$ and $H_0(G; E)$ are dual vector spaces under the paring

$$H^0(G, E^*) \times H_0(G, E) \to \mathbf{R}$$
$$(w, x + K) \mapsto w(x)$$

The following result, proved in Hirsch [20] (see also Dwyer [46]), will be used repeatedly:

1.1. LEMMA. *Let* G *be a nilpotent group and* E *a* G*-module. Then* $H^0(G; E) = 0$ *implies* $H^i(G; E) = 0$ *for all* $i \geq 0$.

From this we deduce:

1.2. THEOREM. *Let G be a nilpotent group, E a G-module. Then the following are equivalent*:

(a) $H^0(G; E) = 0$, (b) $H^0(G; E^*) = 0$,
(c) $H_0(G; E) = 0$, (d) $H_0(G; E^*) = 0$

Proof. Suppose $H^0(G; E) = 0$. Let $H_0(G; E) = E/K$ as before. The exact sequence $0 \to K \to E \to E/K \to 0$ determines an exact sequence

$$0 \to H^0(G; K) \to H^0(G; E) \to H^0(G; E/K) \to H^1(G; K).$$

Since $H^0(G; E) = 0$, by exactness, $H^0(G; K) = 0$. By Lemma 1.1, $H^1(G; K) = 0$, and now by exactness $H^0(G; E/K) = 0$. But G acts trivially on E/K, so $H^0(G; E/K) = 0$ implies $E/K = 0$. Thus $H_0(G; E) = 0$.

This proves (a)$\Rightarrow$(c). The other implications follow from this and the dualities

$$H^0(G; E^*) = H_0(G; E)^*,$$
$$H_0(G; E^*) = H^0(G; E)^*.$$ QED.

1.3. COROLLARY. *Let G be a nilpotent group and E a G-module. If $H^1(G; E) \neq 0$, then G leaves fixed a nonzero vector in E and a nonzero linear functional in E^*.*

A G-module E is *unipotent* if for every $g \in G$ the operator $g - I$ is a nilpotent linear endomorphism of E (i.e. $(g - I)^n = 0$ where $n = \dim E$). We call an affine representation unipotent if its linear part defines a unipotent module.

It is well known that E is unipotent if and only if E has a vector space basis in terms of which G is represented by upper triangular matrices all of whose diagonal entries are 1's (see e.g. Humphreys [23]). Even if E is not unipotent there is a unique maximal submodule $E_U \subset E$ upon which G acts unipotently, called the *Fitting submodule.*

1.4. LEMMA. $H^0(G; E/E_U) = 0$.

Proof. If $x + E_U \in H^0(G; E/E_U)$, then the submodule spanned by x and E_U is unipotent. Hence $x \in E_U$ by maximality. QED.

An affine representation $\alpha : G \to \mathrm{Aff}(E)$ is *reducible* if $\alpha(G)$ leaves invariant a proper affine subspace $F_0 \subset E$. Let $F \subset E$ be the linear subspace of which F_0 is a coset. Then $\alpha(G)$ permutes the cosets of F, so there is induced an affine representation $G \to \mathrm{Aff}(E/F)$, having a stationary point at the coset F_0. If α is not reducible it is called *irreducible.*

An affine representation with a stationary point is called *radiant.* This is a special case of reducible (and also of decomposable, defined below). Every linear representation is a radiant affine representation, and every radiant representation is conjugate to a linear one by a translation.

A reducible affine representation α may have the stronger property of being *decomposable* (compare Zassenhaus [45]). This means that there is a splitting $E = E_1 \oplus F$, $E_1 \neq E$, invariant under the linear part of α, with some coset of E_1 invariant under α. We call $E_1 \oplus F$ a *decomposition* of α. The radiant case $E_1 = 0$, $F = E$ is allowed. If α does not have a decomposition, α is *indecomposable.* Evidently irreducibility implies indecomposability.

1.5. THEOREM. *Let $\alpha : G \to \mathrm{Aff}(E)$ be an indecomposable affine representation of a nilpotent group G. Then α is unipotent.*

The proof of this theorem will be broken up into the following two results, 1.6 and 1.7:

1.6. LEMMA. *Let $E_U \subset E$ be the Fitting submodule of the linear part of an affine representation α of G. If G is nilpotent then some coset of E_U is invariant under $\alpha(G)$.*

Proof. Let $v : G \to E$ be the translational part of α. It is easy to verify that the composition $v' : G \xrightarrow{v} E \to E/E_U$ is a crossed homomorphism for the induced linear action λ' of G on E/E_U. Thus v' is the translational part of an affine representation $\alpha' : G \to \mathrm{Aff}(E/E_U)$ whose linear part is λ'. The natural projection $E \to E/E_U$ is equivariant respecting (α, α').

By Lemma 1.4, $H^0(G; E/E_U) = 0$, so by Lemma 1.1, $H^1(G; E/E_U) = 0$ because G is nilpotent. In particular the radiance obstruction $c_{\alpha'} \in H^1(G; E/E_U)$ vanishes. Therefore $\alpha'(G)$ fixes some $x + E_U \in E/E_U$; this means that $\alpha(G)$ leaves invariant the coset $x + E_U$. QED.

Lemma 1.6 shows that the conclusion of Theorem 1.5 holds under the stronger assumption that α is irreducible.

The proof of 1.5 is completed by the following *splitting theorem*, 1.7. The analogue for Lie algebras is well known.

1.7. THEOREM. *If G is a nilpotent group and E a G-module with Fitting submodule E_U, there exists a unique submodule F such that $E = E_U \oplus F$.*

Proof. Put $E/E_U = V$. Let $\beta : G \to \mathrm{GL}(V)$ be the induced representation. By 1.4, $H^0(G; V) = 0$.

Let $P : E \to V$ be the canonical projection, which is equivariant. Let $S : V \to E$ be a linear map such that $P \circ S = I_V$, the identity map of V.

A submodule $F \subset E$ is complementary to E_U if and only if $F = T(V)$ where $T : V \to E$ is an equivariant linear map with $P \circ T = I_V$. Thus we must prove there is a unique such T.

Any linear $T : V \to E$ with $P \circ T = I_V$ can be uniquely expressed as $T = R + S$ where $R : V \to E_U$, and T is equivariant if and only if

$$R + S = g \circ (R + S) \circ \beta(g)^{-1}$$

or

$$R = g \circ R \circ \beta(g)^{-1} + g \circ S \circ \beta(g)^{-1} - S. \tag{4}$$

We must prove there is a unique R satisfying (4).

Define a linear action γ of G on $\mathrm{Hom}(V, E_U)$ by

$$\gamma : G \to \mathrm{GL}(\mathrm{Hom}(V, E_U)), \qquad \gamma(g)(R) = g \circ R \circ \beta(g)^{-1}$$

and a cocycle u for γ by

$$u : G \to \mathrm{Hom}(V, E_U), \qquad u(g) = g \circ S \circ \beta(g)^{-1} - S.$$

To see that $u(g)$ maps V into E_U notice that $P \circ u(g) = 0$ because P is equivariant and $P \circ S = I_V$.

Now (4) says that R is a stationary point of that affine action ζ of G on $\mathrm{Hom}(V, E_U)$ defined by γ and u. Thus we must prove that ζ has a unique stationary point. This will follow from 1.1 and 1.2 if we prove that $H^0(G; \mathrm{Hom}(V, E_U)) = 0$.

To this end suppose $R : V \to E_U$ is fixed under all $\gamma(g)$, i.e., suppose

$$g \circ R = R \circ \beta(g), \quad \text{all} \quad g \in G. \tag{5}$$

Let $d = \dim E_U$. Since $g \mid E_U$ is unipotent, for all $g_1, \ldots, g_d \in G$ we have

$$(I - g_d) \circ \cdots \circ (I - g_1) \mid E_U = 0.$$

From this and (5), R vanishes on the linear span in V of all vectors of the form

$$[I-\beta(g_d)]\cdots[I-\beta(g_1)]x, \qquad x\in V. \tag{6}$$

Hence it suffices to prove every vector in V is a linear combination of vectors of the form (6). This will be evident if we show that V is spanned by vectors of the form $[I-\beta(g)]x$, i.e., that $H_0(G, V)=0$. But this follows from $H^0(G, V)=0$ and 1.2. QED.

An alternative proof of 1.7 can be based on the analogous result about Lie algebras, by passing to the Lie algebra of the identity component of the algebraic closure of G in $\mathrm{GL}(E)$.

2. Development, holonomy and parallelism

Let M denote an n-dimensional manifold, $n\geq 1$. We shall always assume M is connected and without boundary.

An *affine atlas* Φ on M is a covering of M by coordinate charts such that each coordinate change between overlapping charts in Φ extends to an affine automorphism of $E=\mathbf{R}^n$. A maximal affine atlas is an *affine structure*, on M, and M together with an affine structure is an *affine manifold*. Each chart in the affine structure defines *affine coordinates*.

Let M be an affine manifold. Let $\mathcal{S}$ denote the sheaf of germs of affine coordinate systems on M. The germ at $x\in U$ of the affine chart $f: V\to E$ is denoted by $[f]_x$.

The group $\mathrm{Aff}(E)$ acts stalkwise on $\mathcal{S}$ by composition of germs: $g\in\mathrm{Aff}(E)$ sends $[f]_x$ to $[g\circ f]_x$. It is easy to see that $\mathrm{Aff}(E)$ acts simply transitively on stalks. In fact $\mathcal{S}$ is a principal $\mathrm{Aff}(E)$-bundle over M when $\mathrm{Aff}(E)$ is given the discrete topology.

The following result, whose proof is left to the reader, gives another description of $\mathcal{S}$.

2.1. THEOREM. *$\mathcal{S}$ is a disconnected covering space of M. If $[f_1]_x, [f_2]_x\in\mathcal{S}$ are affine coordinate germs at $x\in M$, there is a deck transformation taking f_1 to f_2, induced by composition with a unique affine automorphism of E. Thus the group of deck transformations acts transitively on each fibre of $\mathcal{S}$, and the components of $\mathcal{S}$ are isomorphic as covering spaces of M. These covering spaces are regular.*

We fix one component of $\mathcal{S}$ and call it the *holonomy covering space* $\hat{M}$ of M. There is a canonical immersion $\hat{D} : \hat{M} \to E$ which assigns $f(x)$ to the germ $[f]_x$. This immersion is affine in the sense that in affine coordinates, it appears as an affine map.

Let Γ denote the group of deck transformations of $\hat{M} \to M$. An element $g \in \Gamma$ sends $[f]_x$ to $[\hat{\alpha}(g) \circ f]_x$, where $\hat{\alpha}(g)$ is the affine automorphism of E determined by g. The resulting map $\hat{\alpha} : \Gamma \to \mathrm{Aff}(E)$ is a faithful affine representation of Γ on E. It is easy to see that $\hat{D}$ is equivariant respecting α, i.e., for each $g \in \Gamma$ the following diagram commutes:

$$\begin{array}{ccc} \hat{M} & \xrightarrow{\hat{D}} & E \\ {\scriptstyle g}\downarrow & & \downarrow{\scriptstyle \hat{\alpha}(g)} \\ \hat{M} & \xrightarrow{\hat{D}} & E \end{array}$$

Now let $p : \tilde{M} \to M$ be a universal covering, and fix a lift of p to a covering $\hat{p} : \tilde{M} \to \hat{M}$. Denote by $D : \tilde{M} \to E$ the composition $\hat{D} \circ \hat{p}$; we call D the *developing map* of the affine manifold M. We give $\tilde{M}$ the affine structure inherited from M; then $D : \tilde{M} \to E$ is an affine immersion.

Let π denote the group of deck transformations of $\tilde{M} \to M$. Choice of a base point in $\tilde{M}$ identifies π with the fundamental group of M. Since $\tilde{M} \to \hat{M}$ is a regular covering, there is a natural surjective homomorphism $\pi \to \Gamma$ with respect to which $\tilde{M} \to \hat{M}$ is equivariant.

Denote by $\alpha : \pi \to \mathrm{Aff}(E)$ the composite homomorphism $\pi \to \Gamma \xrightarrow{\hat{\alpha}} \mathrm{Aff}(E)$. We call α the *affine holonomy* of M. The image $\alpha(\pi) \subset \mathrm{Aff}(E)$ is isomorphic to Γ; it is called the *affine holonomy group* of M. It is clear that $D : \tilde{M} \to E$ is equivariant respecting α.

The definitions of D and α depend on various choices of covering spaces. But once the universal covering $\tilde{M} \to M$ is fixed, changing $\hat{M}$ or the lift $\tilde{M} \to \hat{M}$ changes D only by composition with an element of $\mathrm{Aff}(E)$; and it changes α by conjugation with that element.

Let $\lambda : \pi \to \mathrm{GL}(E)$ be the homomorphism assigning to each $g \in \pi$ the linear part of $\alpha(g)$. We call λ the *linear holonomy*; it is well defined up to conjugation by an element of $\mathrm{GL}(E)$.

We shall assume when discussing an affine manifold M that a developing map $D : \tilde{M} \to E$ has been selected. The affine and linear holonomy representations are then determined.

This somewhat abstract description of holonomy can be brought down to earth by working through the definitions. In terms of loops on M based at $x_0 \in M$, the

homomorphism

$$\alpha : \pi_1(M, x_0) \to \mathrm{Aff}(E)$$

can be described as follows. Let $s : [0, 1] \to M$ be a loop at x_0, representing $[s] \in \pi_1(M, x_0)$. Subdivide $[0, 1]$ by $0 = t_0 < \cdots < t_m = 1$, so that for each $i = 1, \ldots, m$ there is an affine chart (f_i, U_i) with $s : [t_{i-1}, t_i] \to U_i$. (The choice of (f_0, U_0) determines a component of $\mathcal{S}$.) Let V_j be the component of $f_j(U_j \cap U_{j+1})$ that contains $f_j(s(t_j))$, $j = 1, \ldots, m-1$. There is a unique $A_j \in \mathrm{Aff}(E)$ such that

$$f_{j+1} \circ f_j^{-1} = A_j \text{ on } V_j, \qquad j = 1, \ldots, m.$$

Then

$$\alpha([s]) = \begin{cases} A_m \circ \cdots \circ A_1 \text{ if } m \geq 2 \\ \text{the identity map } I \text{ of } E \text{ if } m = 1. \end{cases}$$

We emphasize that there is some freedom of choice in the developing map D. In practice we shall alter D only by composing it with a translation of E.

The affine manifold M is *complete* if D maps $\tilde{M}$ homeomorphically onto E. In this case we can take $\tilde{M} = E$. The affine holonomy is then an inclusion $\pi \hookrightarrow \mathrm{Aff}(E)$ of a discrete subgroup which acts freely and properly discontinuously on E. Thus M is identified with the orbit space E/π.

M can be compact without being complete, e.g. the 1-dimensional affine manifold $\mathbf{R}_+/\{2^n\}$. Examples have been constructed of compact affine manifolds whose developing maps fail to be covering-space projections into their images (see also Section 3).

The following result illustrates the interplay between the topology of an affine manifold, its developing map, and its affine holonomy representation:

2.2. THEOREM. *Let M be a compact complete affine manifold. Then the affine holonomy representation is irreducible.*

Proof. Let $M = E/\pi$. Suppose $F \subset E$ is an affine subspace invariant under π. Then F is the universal covering of F/π. Both F/π and E/π are Eilenberg-MacLane spaces of type $K(\pi; 1)$. The inclusion $F/\pi \to E/\pi$ induces an isomorphism of fundamental groups and is therefore a homotopy equivalence. Since F/π and E/π are compact manifolds it follows that $\dim F/\pi = \dim E/\pi$. Hence $F = E$.

QED.

On an affine manifold M there is a natural linear connection ∇ having zero curvature and torsion: in affine coordinates it appears as the standard connection on E. Covariant differentiation in M corresponds to ordinary differentiation in E.

It is proved in Auslander and Markus [9] that an affine manifold M is complete if and only if the connection ∇ is geodesically complete.

The operation of parallel transport of a tensor along a path in M depends only on the homotopy class of the path (fixing end points). This corresponds to ordinary parallel transport in E in successive affine charts.

A tensor (field) T on M is *parallel* if its covariant derivative vanishes. This means that in every connected affine chart the components of T are constant. Equivalently, the induced tensor field on $\tilde{M}$ is the pullback by the developing map of a constant tensor field T_0 on E. Thus T_0 is an element of the tensor algebra $T(E)$ which is fixed under the induced action of the linear holononomy group on $T(E)$. In every affine chart T appears as T_0.

We shall be particularly concerned with parallel vector fields, parallel 1-forms, and parallel volume elements. A parallel vector field on M corresponds to a vector in E which is stationary under the linear holonomy. The set of such vectors is $H^0(\pi; E)$. A *parallel 1-form* corresponds to a linear map $E \to \mathbf{R}$ stationary under the contragredient action, i.e. an element of $H^0(\pi; E^*)$. A *parallel volume form* is a nonzero exterior n-form on E stationary under the induced action of the linear holonomy group on $\Lambda^n E^*$. Since this action is precisely the determinant of the linear holonomy, M has a parallel volume form if and only if the linear holonomy λ has determinant identically one, i.e. when λ is a representation in the special linear group $\mathrm{SL}(E)$.

The condition that an affine manifold M have a parallel volume form is an extremely useful one. We refer to it by saying that *M has parallel volume.*

A basic conjecture, going back to Markus [28], is that a compact orientable affine manifold has parallel volume if and only if it is complete. This is trivially true for flat Riemannian manifolds. Furness and Fedida [15] prove it for flat pseudo-Riemannian tori in dimensions ≤ 3. In his thesis [38] Smillie proves the conjecture in the case of abelian affine holonomy; for nilpotent affine holonomy it follows from 6.6 and 6.8 below.

More general than parallel tensors are *polynomial tensor fields.* By these we mean tensor fields on affine manifolds whose coefficients in affine coordinates are polynomials. A tensor field T is polynomial of degree $< p$ if and only if the iterated covariant derivative $\nabla_{v_p} \circ \cdots \circ \nabla_{v_1}(T)$ vanishes for all $v_p, \ldots, v_1 \in M_x$, all $x \in M$. (Here ∇ is the covariant differentiation associated to the affine structure.)

Many examples of affine manifolds come from left-invariant affine structures on Lie groups. If G is a Lie group, an affine structure on G is *left-invariant* if for each $g \in G$ the operation $L_g : G \to G$ of left multiplication by g is an automorph-

ism of the affine structure. In other words, in affine coordinates L_g is expressed by an affine map.

The development map of such a structure blends together with left multiplication in a remarkable way. Suppose G is simply connected; let $D : G \to E$ be the developing map. For each $g \in G$ there is a unique affine automorphism $\alpha(g)$ of E such that the following diagram commutes:

$$\begin{array}{ccc} g & \xrightarrow{D} & E \\ {\scriptstyle L_g}\downarrow & & \downarrow{\scriptstyle \alpha(g)} \\ G & \xrightarrow{D} & E \end{array}$$

Clearly $\alpha : G \to \mathrm{Aff}(E)$ is an affine representation. In particular $\alpha(G)$ preserves the connected open set $D(G) \subset E$, and acts transitively on it. Conversely if $\dim G = \dim E$ and $\alpha : G \to \mathrm{Aff}(E)$ is an affine representation having an open orbit $U = D(G)x_0$ for some $x_0 \in E$, then there is a unique left-invariant affine structure on G whose developing map is given by

$$D : G \to U \subset E, \qquad D(g) = \alpha(g)x_0.$$

Notice that when $x_0 = 0 \in E$, D is just the translational part of the affine representation.

The left-invariant affine structure on G is complete precisely when α is a simply transitive affine action of G on E. In this case α is evidently irreducible.

Let $D : G \to E$ be the developing map of a left-invariant affine structure. Then $D : G \to D(G)$ is a covering space, so the structure is complete if and only if the action of G on E is transitive. It is known that in this case G must be solvable (Auslander [6], Milnor [31]). It is conjectured that every solvable Lie group has such a structure.

For examples of left-invariant affine structures see Auslander [6], [7]; Scheuneman [35], [36]; and Section 6.

Suppose now that Γ is a discrete subgroup of G, where G has a left-invariant affine structure. Then the homogeneous space Γ/G of right cosets inherits an affine structure. When G is a simply-connected nilpotent Lie group, $\Gamma\backslash G$ is called an *affine nilmanifold.*

An affine representation $\alpha : G \to \mathrm{Aff}(E)$ is conveniently presented by the linear representation

$$\alpha' : G \to \mathrm{GL}(E \times \mathbf{R}),$$
$$\alpha' : (g) : (x, t) \mapsto (\lambda(g)x + tu(g), t)$$

where λ is the linear part and u the translational part of α.

As an example take $G=\mathbf{R}^2$ (the group) and $E=\mathbf{R}^2$ (the vector space). An affine representation $\phi : G \to \mathrm{Aff}(E)$ is defined by

$$\phi(s,t) : z \to \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} z + \begin{bmatrix} s^2/2+t \\ s \end{bmatrix},$$

for $(s,t)\in G$, $z=\begin{bmatrix} x \\ y \end{bmatrix}\in E$.

The associated linear representation on $\mathbf{R}^3$ assigns to $(s,t)\in G$ the matrix

$$\begin{bmatrix} 1 & s & \frac{s^2}{2}+t \\ \cdot & 1 & s \\ \cdot & \cdot & 1 \end{bmatrix} = \exp\left\{ \begin{bmatrix} \cdot & s & \cdot \\ \cdot & \cdot & s \\ \cdot & \cdot & \cdot \end{bmatrix} + \begin{bmatrix} \cdot & \cdot & t \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \right\}$$

It is easy to verify that $\phi(G)$ acts simply transitively on E; thus ϕ defines a left-invariant affine structure on G. The developing map $D : G \to E$ defined by evaluation at 0 is the translational part of ϕ:

$$D(s,t)=\begin{bmatrix} (s^2/2)+t \\ s \end{bmatrix}.$$

The left-invariant vector fields on G are spanned by $\partial/\partial t$ and $\partial/\partial s$; in the affine coordinates on G defined by D these correspond respectively to $\partial/\partial x$ and $y(\partial/\partial x+\partial/\partial y)$. The integral curves of $\partial/\partial x$ are horizontal lines; those of $y(\partial/\partial x+\partial/\partial y)$ are the parabolas $y^2-2x=\text{constant}$.

The functions y and y^2-2x on E transform under G by addition of constants. Therefore the 1-forms dy and $y\,dy-dx$ are the expression in affine coordinates of left-invariant 1-forms on G.

The vector field $\partial/\partial x$ and the form dy (on E) are translation invariant; hence they correspond respectively to a parallel vector field and a parallel 1-form on G.

Since the linear part of α preserves the area 2-form $dx\wedge dy$ on E, this form defines a parallel 2-form on G.

For every uniform discrete subgroup $\Gamma\subset G$ we get a complete affine 2-manifold $\Gamma\backslash G$ diffeomorphic to the 2-torus T^2. By Nagano and Yagi [32] and Kuiper [26] these are the only complete structures on T^2 other than the flat Riemannian structure $\mathbf{Z}^2\backslash\mathbf{R}^2$.

The choice of Γ, however, is crucial. For example it is easy to see that $\Gamma\backslash G$ has a closed geodesic (= 1-dimensional affine submanifold) if and only if Γ contains a pure translation $\varphi(0,t_0)$, $t_0\neq 0$.

There are many incomplete affine nilmanifold structures on T^2. For example, for each $a \geq 0$, $b > 0$ there is a radiant affine representation

$$\psi_{a,b} : G \to \mathrm{GL}(\mathbf{R}^2) \subset \mathrm{Aff}(\mathbf{R}^2),$$

$$(s, t) \to \exp \begin{bmatrix} s & as+bt \\ 0 & s \end{bmatrix}$$

$$= e^s \begin{bmatrix} 1 & as+bt \\ 0 & 1 \end{bmatrix}.$$

Each of these actions is simply transitive on the upper half-plane.

For each $\mu > 0$, $\nu > 0$ radiant actions are defined by

$$\theta_{\mu,\nu} : G \to \mathrm{GL}(\mathbf{R}^2) \subset \mathrm{Aff}(\mathbf{R}^2),$$

$$\theta_{\mu,\nu}(s, t) = \begin{bmatrix} \mu^s & 0 \\ 0 & \nu^t \end{bmatrix}$$

These are all simply transitive in the first quadrant.

Some other examples are given below after the proof of Theorem 6.4, and in 6.7.

By assuming that an affine structure is compatible with a complex structure on a manifold one defines the notion of an affine complex manifold. (An affine structure is compatible with a complex structure if and only if the associated flat affine connection is holomorphic with respect to the complex structure; alternatively a complex structure is compatible with an affine structure if and only if the associated almost complex structure is a parallel tensor of type (1, 1). It is easy to see that these two notions are equivalent.) All the results in this paper apply to such manifolds.

It is interesting to note that Fillmore and Scheuneman [11], Suwa [41], and Sakane [34] have proved, independently, that every complex surface having a complete affine complex structure has a finite cyclic covering which is a complex affine nilmanifold. For references and more information on these structures in the complex case the reader is referred to Gunning [19] and Inoue–Kobayashi–Ochiai [24].

Smillie [38], Auslander [6] and Milnor [31] relate affine structures to matrix algebras. See Matsushima [29] for a connection between affine structures on homogeneous complex manifolds and algebras.

3. Radiant manifolds and parallel tensors

In this section we study affine manifolds whose affine holonomy has a stationary point; we call such manifolds radiant.

There are many examples of radiant manifolds: certain affine nilmanifolds; the Hopf manifolds; the cartesian product of any compact surface and the circle. On the other hand the affine structures of radiant manifolds have rather special properties. There are no complete radiant manifolds (except for vector spaces), and they tend to have few parallel tensors. In a sense they are at the opposite extreme from complete affine manifolds.

We begin by defining the radiance obstruction for any affine manifold. This will play a key role in the analysis of nonradiant affine manifolds in Section 4.

M always denotes a connected n-dimensional affine manifold, $n \geq 1$, with local affine coordinates in the vector space $E = \mathbf{R}^n$. We fix a universal cover $p : \tilde{M} \to M$, with group of deck transformations π. Let $D : \tilde{M} \to E$ be a developing map for M and $\alpha : \pi \to \mathrm{Aff}(E)$ the corresponding affine holonomy representation (defined in Section 2). The linear part of α is the linear holonomy $\lambda : \pi \to \mathrm{GL}(E)$, and the translational part α is the cocycle $u : \pi \to E$ for λ.

The *radiance obstruction* of M is the cohomology class

$$c_M = c_\alpha \in H^1(\pi, E_\lambda),$$

where $c_\alpha = [u]$ is the radiance obstruction of the affine representation α (defined in Section 1). *This cohomology class depends only on the affine holonomy of* M (assuming the universal covering space $\tilde{M} \to M$ has been fixed), and not on the choice of the developing map. To see this let $\iota : \mathrm{Aff}(E) \to \mathrm{Aff}(E)$ be the identity affine representation. Let

$$\alpha^* : H^1(\mathrm{Aff}(E); E) \to H^1(\pi; E)$$

be the homomorphism induced by α. Let $c_\iota \in H^1(\mathrm{Aff}(E) : E)$ be the radiance obstruction of ι; then $\alpha^* c_\iota = c_M$. Let c'_M be the radiance obstruction corresponding to a different developing map $D' : \tilde{M} \to E$. There exists $g \in \mathrm{Aff}(E)$ such that $D' = g \circ D$, and one verifies that $c'_M = g_\# c_\iota$, where $g_\#$ is the automorphism of $H^1(\mathrm{Aff}(E); E)$ induced by conjugation by g. But it is well known that $g_\#$ is the identity, hence $c'_M = c_M$.

In a forthcoming paper [14] we shall discuss the radiance obstruction in more detail.

It follows from the results of Section 1 that $c_M = 0$ precisely when $\alpha(\pi)$ has a stationary point in E. In this case we call M a *radiant manifold*. We can compose

the developing map with a translation of E so that the stationary point is the origin, i.e., α takes values in $\mathrm{GL}(E)$. In this case $\alpha = \lambda$.

Since a nontrivial group of deck transformations cannot have a stationary point[1] we see that a complete affine manifold is not radiant unless it is affinely isomorphic to a vector space.

When M is radiant we shall tacitly assume that the origin in E is stationary under $\alpha(\pi)$. In this case $\alpha(\pi)$ fixes the radiant vector field $R = \Sigma x_i \partial/\partial x_i$ on E. There is a unique vector field $\tilde{X}$ on $\tilde{M}$ which is D-related to R. Clearly $\tilde{X}$ is fixed under π; therefore $\tilde{X}$ is p-related to a vector field X on M, called the *radiant vector field* of M.

It is easy to find examples of such manifolds. The simplest examples are Hopf manifolds: these are the orbit spaces $(E-\{0\})/\pi$ where $\pi \subset \mathrm{GL}(E)$ is the cyclic subgroup generated by an *expansion* $A \in \mathrm{GL}(E)$, i.e., the eigenvalues of A all have absolute value greater than 1. For another example, the Cartesian product of a compact orientable surface Σ of genus ≥ 2 and the circle S^1 can be given a radiant affine structure as follows. Use one of the well known faithful representations of $\pi_1(\Sigma)$ as a discrete subgroup of $\mathrm{SO}(1, 2)$ to obtain a properly discontinuous free action of $\pi_1(\Sigma)$ on one sheet of a two-sheeted hyperboloid $H \subset \mathbf{R}^3 - 0$. Let $\mathbf{Z} \subset \mathrm{GL}(\mathbf{R}^3)$ be generated by kI where $0 < k < 1$. Then $\pi_1(\Sigma) \times \mathbf{Z}$ acts freely as deck transformations on a component of the interior of the light cone; the orbit space is $\Sigma \times S^1$.

There are many radiant affine structures on the product of a closed surface of genus ≥ 1 and a circle, for which the developing map is not a covering-space projection of $\tilde{M}$ onto its image. This phenomenon was discovered by Thurston (see [40]) and independently by Smillie [38]. The developing image of such a 3-torus is the complement of the three coordinate axes in $\mathbf{R}^3$; for the product of a surface of genus ≥ 2 and a circle, the developing map can be onto the complement of the origin. On the other hand, we know of no example of a compact affine manifold whose developing map is surjective but which is not complete (compare Theorem 6.9).

3.1. THEOREM. *A compact radiant manifold does not have a parallel volume form.*

Proof. Suppose M is radiant with a parallel volume form. A trivial computation in affine coordinates shows that the flow of the radiant vector field increases volume. Hence M cannot be compact. QED.

---

[1] In fact no element of the group (except the identity) can have a fixed point. It follows easily that every element of the linear holonomy of a complete affine manifold has 1 as an eigenvalue. This might be called "Hirach's principle", following Sullivan [39].

3.2. THEOREM. *A compact radiant manifold M does not have a nonzero parallel 1-form.*

Proof. Let $\omega : E \to R$ be a linear map invariant under the linear holonomy. Then ω is also invariant under the affine holonomy. Therefore the composite map

$$\tilde{M} \xrightarrow{D} E \xrightarrow{\omega} \mathbf{R}$$

covers a map $f : M \to \mathbf{R}$. Let $x \in \tilde{M}$ cover a local minimum point of f. Then $D(x) \in E$ is a local minimum point of ω, so ω, being linear, must be identically zero. QED.

There are numerous compact radiant manifolds having parallel *line* fields, for example a Hopf manifold $(E-\{0\})/\pi$, where $\pi \subset \mathrm{GL}(E)$ is a cyclic subgroup generated by an expansion having a real eigenvalue.

With later applications in mind we prove the following result about decomposable affine manifolds. Note that in the case of radiant manifolds ($E_1 = 0$) it shows that any point fixed by the affine holonomy action lies outside the developing image (this was first stated by Nagano–Yagi [32] but their proof is incorrect). Thus the *radiant vector field on a compact radiant manifold is nonsingular.*

3.3. THEOREM. *Suppose M is a compact affine manifold with decomposable holonomy, say $E = E_1 \oplus F$ with E_1 α-invariant and F λ-invariant, $F \neq 0$. Then each component of $M_1 = p(D^{-1}E_1)$ is incomplete.*

Proof. Clearly M_1 is an affine submanifold of M. Assume a component N of M_1 is complete.

Consider the α-invariant vector field $S(x, y) = (0, y)$ on $E = E_1 \oplus F$. Then S is D-related to a π-invariant vector field $\tilde{Q}$ on $\tilde{M}$ which induces a vector field Q on M. By the compactness of M, the vector fields Q and $\tilde{Q}$ are integrable for all time and determine flows ϕ_t and $\tilde{\phi}_t$ on M and $\tilde{M}$ respectively. If the linear flow determined by S is denoted ψ_t, we have $\psi_t \circ D = D \circ \tilde{\phi}_t$.

Suppose $N \subset M$ is a component of $p^{-1}E_1$. Then $D \mid \tilde{N}$ is a homeomorphism onto E_1 by completeness.

The submanifold $E_1 \subset E$ is a repellor for the flow ψ_t. It follows that $\tilde{N}$ is a repellor for the flow $\tilde{\phi}_t$. Let $B = \{x \in M : \phi_t x \to \tilde{N} \text{ as } t \to -\infty\}$ denote the repelling basin of $\tilde{N}$.

Now $D : \tilde{N} \to E_1$ is a homeomorphism and $D : \tilde{M} \to E$ is a local homeomorphism. It follows that D maps a neighborhood $B_0 \subset B$ of $\tilde{N}$

homeomorphically onto a neighborhood $W \subset C$ of E_1. Therefore D maps $\tilde{\phi}_t(B_0)$ homeomorphically onto $\psi_t(W)$ for all $t \geq 0$. Since $\bigcup_{t>0} \psi_t(w) = E$ it follows that D maps $\bigcup_{t>0} \tilde{\phi}_t(B_0)$ homeomorphically onto E. This implies that $D : \tilde{M} \to E$ is a homeomorphism. Therefore M is complete, in contradiction to Theorem 2.2.

QED.

4. Affine manifolds with nilpotent holonomy

In this section we begin to exploit the assumption of nilpotent affine holonomy group. The same notation and conventions as in Section 3 are used.

4.1. THEOREM. *Let M be a compact affine manifold with nilpotent affine holonomy group. Then the following conditions are equivalent:*

(a) *M is not radiant;*

(b) *M has both a nonzero parallel vector field and a nonzero parallel 1-form.*

(c) $H^1(\pi; E) \neq 0$.

Proof. (a)$\Rightarrow$(b): Nonradiant means $c_M \neq 0$, so $H^1(\pi; E) \neq 0$. By 1.3, $H^0(\pi; E) \neq 0$ and $H^0(\pi; E^*) \neq 0$, which is (b).

(b)$\Rightarrow$(a): follows from 3.2.

(c)$\Leftrightarrow$(b): follows from 1.3. QED.

Notice that the implication (a)$\Rightarrow$(b) in 4.1 is valid even without compactness.

4.2. THEOREM. *Let M be a compact nonradiant affine manifold having nilpotent affine holonomy group. Then M fibres over the circle S^1.*

Proof. A nonzero parallel 1-form on M is a nowhere-vanishing closed 1-form. It is well known that the existence of such form implies that M fibres over S^1 (Tischler [42]). QED.

We conjecture that 4.2 holds even if M is radiant.

As a corollary of Theorem 4.1 and the nonsingularity of radiant vector fields, we see that a compact affine manifold with nilpotent affine holonomy group has Euler characteristic zero. This also follows from Hirsch–Thurston [22]; to use this result, it suffices that the linear holonomy group be solvable. Kostant and Sullivan [25] have shown that every complete compact affine manifold has zero Euler characteristic. It has long been conjectured that every affine manifold (except $\mathbf{R}^n$) which is complete *or* compact has zero Euler characteristic. (In the noncompact

case the Euler characteristic is defined as the alternating sum of the Betti numbers, which are conjectured to be finite.)

We now concentrate attention more directly on the holonomy representation.

An immediate consequence of Theorem 1.5 is:

4.3. THEOREM. *Let M be an affine manifold whose affine holonomy representation is nilpotent and indecomposable. Then the linear holonomy representation is unipotent.*

Little is known about the structure of noncompact complete affine manifolds, even those with nilpotent holonomy (but see Milnor [31]). The following result shows the importance of unipotent in this case.

4.4. THEOREM. *Let M be a noncompact complete affine manifold with nilpotent affine holonomy group. Then M is a flat affine vector bundle over a complete affine manifold with unipotent linear holonomy.*

Proof. Let $E_U \subset E$ be the Fitting component of the linear holonomy. We may assume E_U invariant under $\alpha(\pi)$. By 4.1 M has a nonzero parallel vector field, so $E_U \neq 0$.

By the splitting Theorem 1.7 there is a decomposition $E = E_U \oplus F$ of α. The vector bundle is

$$q : M = E/\alpha(\pi) \to E_U/\alpha(\pi),$$

with total space M, base space $E_U/\alpha(\pi)$, and fibre F. It is easy to see that there are local trivializations

$$f_i : q^{-1}(W_i) \to W_i \times F$$

over an open cover $\{W_i\}$ of $E_U/\alpha(\pi)$, which are affine isomorphisms, and whose transition functions $g_{ij} : W_i \to \mathrm{GL}(F)$ are constants. Thus the bundle is flat.

QED.

5. Nilpotent G-modules without expansions

The purpose of this section is to prove the following technical result. It will be used in Section 6 to prove the existence of expansions in the linear holonomy of certain compact affine manifolds.

Recall that $E = \mathbf{R}^n$, $n \geq 1$.

5.1. LEMMA. *Let G be a nilpotent group and E a G-module. Suppose that G does not contain an expansion of E. Then for every integer $r \geq 0$ there exists a C^r map $\Psi : E \to \mathbf{R}$ with the following properties*:

(a) $\Psi > 0$ *almost everywhere*;

(b) Ψ *is G-invariant*;

(c) *there exists $a > 0$ such that*

$$\Psi(e^t x) = e^{ta}\Psi(x)$$

for all $t \in \mathbf{R}$, $x \in E$.

Proof. If the lemma is true for a normal subgroup $G_0 \subset G$ of finite index, it is true for G. For let $\Psi_0 : E \to \mathbf{R}$ be a C^r G_0-invariant map satisfying (a) and (c). Let the left cosets of G_0 be $g_1 G_0, \ldots, g_\nu G_0$. Then the map

$$\Psi : E \to \mathbf{R}, \qquad \Psi(x) = \sum_{i=1}^{\nu} \Psi_0(g_i x)$$

satisfies the lemma.

Let $F = \mathbf{C} \oplus E^*$ be the complexification of the dual space E^* of E. The contragredient representation of G on E^* extends to a (complex) representation $\rho : G \to \mathrm{GL}(F)$. Let $H \subset \mathrm{GL}(F)$ be the identity component of the algebraic closure of $\rho(G)$. Then H is a connected nilpotent Lie group. Set $G_0 = \rho^{-1}(H)$. Then $G_0 \subset G$ is a normal subgroup of finite index.

From the primary decomposition of the representation of the Lie algebra of H in F induced by ρ we get a ρ-invariant splitting $F = \oplus F_k$. Each F_k has a basis $\mathcal{B}_k$ representing the operators $\rho(h) \mid F_k$, $h \in H$ as complex matrices

$$\rho_k(h) = \lambda_k(h) I + N_k(h)$$

where $N_k(h)$ is an upper-triangular nilpotent matrix. In particular the set of eigenvalues of $\rho(h)$ is $\{\lambda_k(h)\}$.

Let $f_k \in F_k$ be the first basis vector in $\mathcal{B}_k$, $k = 1, \ldots, m$. Then

$$\rho_k(h) f_k = \lambda_k(h) f_k \tag{1}$$

for all $h \in H$.

Define group homomorphisms

$$\varphi_k : G_0 \to \mathbf{R}, \qquad \varphi_k(g) = \log |\lambda_k(g)|; \qquad k = 1, \ldots, m.$$

Set

$$\varphi = (\varphi_1, \ldots, \varphi_m) : G_0 \to \mathbf{R}^m.$$

Suppose G_0 contains no expansion of E. Then $\varphi(G_0)$ is disjoint from the positive orthant $P \subset \mathbf{R}^m$,

$$P = \{y \in \mathbf{R}^m : y_k > 0,\ k = 1, \ldots, m\}.$$

Since $\varphi(G_0)$ is a subgroup of $\mathbf{R}^m$ it follows that the linear span $L \subset \mathbf{R}^m$ of $\varphi(G_0)$ is also disjoint from P. This implies that the orthogonal complement of L contains a nonzero vector $v \in \bar{P}$. This vector $v = (v_1, \ldots, v_m)$ thus has the following properties:

$$\text{each} \quad v_k \geq 0 \tag{2}$$

$$\sum_{k=1}^{m} v_k = a > 0 \tag{3}$$

$$\sum_{k=1}^{m} v_k \varphi_k(g) = 0 \quad \text{for all} \quad g \in G_0. \tag{4}$$

Evidently the vector cv, $c > 0$, has the same properties. Therefore we can also choose v to satisfy:

$$\text{For each } k, \text{ either } v_k = 0 \text{ or } v_k > 2r, \text{ where } r \text{ comes from Lemma 5.1.} \tag{5}$$

For each k we have chosen a vector $f_k \in \mathrm{Hom}_{\mathbf{C}}(\mathbf{C} \otimes E, \mathbf{C})$ satisfying (1). We embed E in $\mathbf{C} \otimes E$ in the natural way and define

$$\Psi : E \to \mathbf{R},$$

$$\Psi(x) = \prod_{k=1}^{m} |f_k(x)|^{v_k}.$$

By (5) Ψ is differentiable of class C^r. By (2) and (3) Ψ satisfies (a), (c) of 5.1.

To show that Ψ is G_0-invariant let $g \in G_0$, $x \in E$. Then

$$\Psi(g^{-1}x) = \prod_{k=1}^{m} |f_k(g^{-1}x)|^{v_k} = \prod_{k=1}^{m} |(\lambda(g) f_k)x|^{v_k}$$

$$= \left(\prod_{k=1}^{m} |\lambda_k(g)|^{v_k}\right)\left(\prod_{k=1}^{m} |f_k(x)|^{v_k}\right) = \left(\exp \sum_{k=1}^{m} v_k \varphi_k(g)\right)\Psi(x) = \Psi(x)$$

by (4). This completes the proof of 5.1. QED.

In general Ψ in 5.1 cannot be chosen to be C^∞. For example, let

$$G = \left\{ \begin{bmatrix} \lambda^{\sqrt{2}} & 0 \\ 0 & \lambda^{-1} \end{bmatrix} ; \lambda > 0 \right\} \text{ on } E = \mathbf{R}^2.$$

Any such ψ must be of the form $C(xy^{\sqrt{2}})^p$ where $C, p > 0$, and will not be C^∞.

6. Compact affine manifolds with nilpotent holonomy

In this section we exploit compactness to prove three main results: 6.1, 6.8, 6.9. The proofs rely on the integration of vector fields. The first two were proved by Smillie [38] for the abelian case.

The first, Theorem 6.1, is used to find expansions of E/E_U in the linear holonomy. This turns out to be a powerful geometric tool. The second, Theorem 6.8, shows that unipotent holonomy implies completeness. This is also true for noncompact affine nilmanifolds.

As is Section 3, $D : \tilde{M} \to E$ is the developing map, π is the group of deck transformations of the universal cover $\tilde{M} \to M$, etc. Let $\Lambda \subset \mathrm{GL}(E)$ denote the linear holonomy group and $\Gamma \subset \mathrm{Aff}(E)$ the affine holonomy group.

6.1. THEOREM. *Let M be a compact affine manifold. Let $E_0 \subset E$ be a proper linear subspace invariant under the affine holonomy. If the image $\Lambda_1 \subset \mathrm{GL}(E/E_0)$ of Λ is nilpotent then some element of Λ_1 expands E/E_0.*

Proof. Let $q : E \to E/E_0$ denote the canonical projection. Let R be the radiant vector field on E/E_0. Let $\{(\varphi_i, U_i)\}$ be an affine atlas on M. There is a C^∞ vector field X on M which in every affine chart (φ_i, U_i) is represented by a vector field on $\varphi_i(U_i)$ which is q-related to R. For there is clearly such a vector field X_i on U_i; set $X = \sum_{\mu i} X_i$ where $\{\mu_i\}$ is a C^∞ partition of unity subordinate to the open cover $\{U_i\}$.

Suppose Λ_1 does not contain any expansion of E/E_0. Then by 5.1 there is a Λ_1-invariant C^1 map $\Psi : E/E_0 \to \mathbf{R}$, $\Psi > 0$ almost everywhere and

$$d\Psi_z R(z) = a\Psi(z) \tag{1}$$

for some $a > 0$ and all $z \in E/E_0$ (this is the differential equivalent of 5.1(c)).

The composite map

$$\tilde{f} : \tilde{M} \xrightarrow{D} E \xrightarrow{q} E/E_0 \xrightarrow{\Psi} \mathbf{R}$$

is π-invariant. Thus $\tilde{f}$ covers a C^1 map $f: M \to \mathbf{R}$. In affine coordinates f appears as $\Psi \circ q$. It follows that

$$df_y X(y) = af(y)$$

for all $y \in M$. This implies that if $\xi : \mathbf{R} \to M$ is an integral curve of X then

$$f(\xi(t)) = e^{ta} f(\xi(c)).$$

Now there exists $x_0 \in M$ with $f(x_0) > 0$, and the integral curve ξ_0 through x_0 is defined for all t because M is compact. Then $\lim_{t\to\infty} f(\xi_0(t)) = \infty$. But f, being continuous, is bounded. This contradiction proves 6.1. QED.

We derive several consequences from 6.1.

6.2. THEOREM. *Let M be a compact radiant manifold with nilpotent linear holonomy group Λ. Then Λ contains an expansion.*

Proof. Take $E_0 = 0$ in 6.1.

6.3. COROLLARY. *M has no nonzero parallel covariant or contravariant tensors.*

Proof. Let ω be a parallel covariant (resp. contravariant) tensor on M; let $\tilde{\omega}$ be the corresponding (constant) tensor on E invariant under Λ. By 6.2, Λ contains an expansion g. Let the eigenvalues of g be $\lambda_1, \ldots, \lambda_n$ (counted with multiplicity), $|\lambda_i| > 1$. Now $\tilde{\omega}$ is an element of the tensor algebra $\oplus^l E^*$ (resp. $\oplus^l E$). The map on this algebra induced by g has eigenvalues $\lambda_{i_1} \cdots \lambda_{i_l}$ (resp. $\lambda_{i_1}^{-1} \cdots \lambda_{i_l}^{-1}$), which makes it an expansion (resp. a contraction). Consequently the g-invariant tensor $\tilde{\omega}$ must be zero. QED.

Conversely, suppose M is a nonradiant compact affine manifold with nilpotent affine holonomy. Then by 4.1 M has nonzero parallel vector fields and 1-forms.

Another characterization of radiance is the following:

6.4. THEOREM. *Let M be a compact affine manifold with nilpotent affine holonomy group. Then the following conditions are equivalent:*

(a) *M is radiant.*

(b) *The linear holonomy group contains an expansion.*

(c) *The linear holonomy group contains g such that $I - g$ is invertible.*

Proof. (a)⇒(b) by 6.2 and clearly (b)⇒(c). Suppose (c). By Hirsch [21] it follows that $H^1(\Gamma; E) = 0$. Therefore the radiance obstruction of M vanishes, implying (a). QED.

We digress to consider deformations of affine structures.

Conditions (c) and (b) of Theorem 6.4 evidently persists under sufficiently small deformations of the linear holonomy representations. It follows that if M is a *compact radiant affine manifold with nilpotent holonomy group, then all sufficiently small deformations of the affine structure are also radiant.* In other words, for this class of manifolds radiance is an open condition. (See Goldman [16] for the definition of deformation of affine structures.)

Consider also the radiant affine structure on $M^2 \times S^1$ discussed in Section 3, where M^2 is a compact surface of genus ≥ 2. Let $g \in \pi_1(M \times S^1)$ be the image of a generator of S^1. Then g is central and $I - \lambda(g)$ is invertible. This suffices to prove that any sufficiently small deformation of the affine holonomy has a stationary point (Hirsch [21]). Thus the radiance of these affine structures is also persistent under small deformations.

In this direction D. Fried has proved [13] that there are no complete affine structures on these 3-manifolds. It is conjectured that all affine structures on these manifolds are radiant.

Corollary 6.3, and the persistence of radiance discussed above, cannot be extended to the case of solvable holonomy. To construct a counterexample we start from the subgroup $G \subset \mathrm{GL}(3, \mathbf{R})$ consisting of all the matrices of the form

$$e^{kt}\begin{bmatrix} e^t & \cdot & u \\ \cdot & e^{-t} & v \\ \cdot & \cdot & 1 \end{bmatrix} = \begin{bmatrix} e^{(k+1)t} & \cdot & u \\ \cdot & e^{(k-1)t} & v \\ \cdot & \cdot & e^{kt} \end{bmatrix}$$

where $t, u, v \in \mathbf{R}$, and $k > 0$ is a real constant to be determined later. Clearly G preserves the upper half-space

$$W = \{(x, y, z) \in \mathbf{R}^3 : z > 0\}$$

and acts simply transitively. Thus G inherits a left-invariant radiant (and hence incomplete) affine structure from W.

It is easily seen that G is the semidirect product of the two-dimensional vector subgroup N defined by $t = 0$ with the subgroup $H \cong \mathbf{R}$ defined by $u = v = 0$. The representation $\mathbf{R} \to \mathrm{Aut}(\mathbf{R}^2)$ which defines the semidirect product is given by

$$\varphi : t \to \begin{bmatrix} e^t & \cdot \\ \cdot & e^{-t} \end{bmatrix}; \qquad t \in \mathbf{R}$$

It is well-known that $G = N \rtimes H$ possesses discrete cocompact subgroups Γ. To see this take any integral matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbf{Z})$ with $a + d > 2$. Then A is conjugate (in $\mathrm{SL}_2(\mathbf{R})$) to some diagonal matrix $\begin{bmatrix} e^{\lambda} & 0 \\ 0 & e^{-\lambda} \end{bmatrix}$, $\lambda \neq 0$, and therefore one may form the semidirect product $\mathbf{R}^2 \rtimes \mathbf{Z}$ defined by this representation, with a generator of $\mathbf{Z}$ corresponding to A. Since A was originally defined as an integral matrix, it preserves a lattice $\mathbf{Z}^2 \subset \mathbf{R}^2$. (It must be emphasized, however, that this lattice is *not* generated by a basis which diagonalizes A). Then we may form the semidirect product $\Gamma = \mathbf{Z}^2 \rtimes \mathbf{Z}$ which is a discrete subgroup of $G = \mathbf{R}^2 \rtimes \mathbf{R} = N \rtimes H$, with $\Gamma \backslash G$ compact. Using the left-invariant structure on G, we thus obtain affine structures on the compact 3-manifold $M^3 = \Gamma \backslash G$; indeed, as k varies continuously a whole one-parameter family of affine structures is defined in this way.

Let us now examine the case $k = 1$. Then G is the group of matrices

$$A(u, v, t) = \begin{bmatrix} e^{2t} & \cdot & u \\ \cdot & 1 & v \\ \cdot & \cdot & e^{t} \end{bmatrix}; \qquad t, u, v \in \mathbf{R}.$$

It is apparent that the vector field $\partial/\partial y$ parallel to the second basis vector is fixed by G, and hence defines a parallel vector field on G, and also on $\Gamma \backslash G$. Thus we obtain a compact affine manifold with both a radiant vector field and a parallel vector field. These two vector fields generate an affine action on the group $\mathrm{Aff}(\mathbf{R})$ on the affine 3-manifold.

The existence of a parallel vector field on a radiant affine manifold is easily seen to be equivalent to the condition that the holonomy have more than one stationary point. (The stationary points of an affine action form an affine subspace, parallel to the space of parallel vector fields.)

We now describe a deformation of this affine structure. For each $s \in \mathbf{R}$, define an affine representation $A_s : G \to \mathrm{Aff}(E)$ as follows. Let $A_s(u, v, t)$ be the affine map of $E = \mathbf{R}^3$ with linear part $A(u, v, t)$ and translational part $(0, st, 0)$. For each s, A_s defines an action of G which is simply transitive on the upper half space. By passing to a quotient $\Gamma \backslash G$ we obtain new affine structures on $M^3 = \Gamma \backslash G$. These new affine manifolds are no longer radiant for $s \neq 0$, as is easily verified. Thus radiance is not an open condition on holonomy of affine structures. A glance at the circle shows that neither is radiance a closed condition.

The case $k = \pm\frac{1}{2}$ of (2) is also interesting. Taking $k = \frac{1}{2}$ we represent G by matrices

$$\begin{bmatrix} e^{3t/2} & \cdot & u \\ \cdot & e^{-t/2} & v \\ \cdot & \cdot & e^{t/2} \end{bmatrix}.$$

then G preserves the parallel 2-form $\omega = dy \wedge dz$ as well as the radiant vector field R. (Compare 6.3.) In addition to the affine vector field R the polynomial vector field $z^2(\partial/\partial x)$ and rational vector field $z^{-1}(\partial/\partial y)$ are left-invariant. The affine 1-form $\iota_R\omega = z\,dy - y\,dz$ is also left-invariant. Thus a compact radiant manifold may admit a 1-form which is affine (polynomial of degree 1) but not parallel (polynomial of degree 0). Compare 3.2.

In a forthcoming paper [14] we will show that on a compact radiant manifold all polynomial closed 1-forms and volume forms are zero.

It is interesting to note that G also carries the *complete* left-invariant affine structure corresponding to the simply transitive affine action $\alpha : G \to \mathrm{Aff}(E)$ defined as follows. Represent G as $\mathbf{R} \ltimes \mathbf{R}^2$ as above. Let $(t, \xi) \in \mathbf{R} \ltimes \mathbf{R}^2$ act on $\mathbf{R} \times \mathbf{R}^2$ by

$$(t, \xi) : (q, \eta) \to \left(q + t, \begin{bmatrix} e^t & \cdot \\ \cdot & e^{-t} \end{bmatrix} \eta + \xi\right).$$

See also Example 6.7 below. Auslander [6] gives a different complete left-invariant affine structure on G.

We return to the general theory.

Let $E_U \subset E$ be the Fitting component (= maximal unipotent submodule) of the linear holonomy. Recall that when the affine holonomy group is nilpotent we may assume (by 1.6) that E_U is invariant under the affine holonomy; and there is a unique splitting $E = E_U \oplus F$ invariant under the linear holonomy.

With this notation we have the following generalization of 6.2.

6.5. THEOREM. *Let M be a compact affine manifold with nilpotent affine holonomy group. If $F \neq 0$ then some element of the linear holonomy group expands F.*

Proof. Apply 6.1 with $E_0 = E_U$; then $E/E_0 \approx F$ as a π-module. QED.

For the special case of abelian holonomy, 6.5 and 6.2 were first proved by John Smillie.

An important consequence of 6.5 is:

6.6. THEOREM. *Let M be a compact affine manifold with nilpotent affine holonomy group. Assume M has a parallel volume form. Then the linear holonomy is unipotent.*

Proof. Let g be any element of the linear holonomy group Λ. Then $1 = \mathrm{Det}\, g = \mathrm{Det}\,(g \mid E_U) \cdot \mathrm{Det}\,(g \mid F)$. But $g \mid E_U$ is a unipotent operator, so

$\mathrm{Det}(g \mid E_U) = 1$. Therefore $\mathrm{Deg}(g \mid F) = 1$ and so $g \mid F$ cannot be an expansion. Therefore $F = 0$ by 6.5. This means $E = E_U$. QED.

6.7. EXAMPLE. Theorem 6.6 fails for solvable holonomy. For example let $T^n = \mathbf{R}^n/\mathbf{Z}^n$ be the usual Euclidean affine torus, $\mathbf{Z}^n \subset \mathbf{R}^n$ being the integer lattice. Let $f : T^n \to T^n$ be the affine automorphism represented by a matrix $A \in \mathrm{SL}(n, \mathbf{Z})$. Let $\mathbf{Z}$ act on $T^n \times \mathbf{R}$, the generator sending (x, t) to $(f(x), t+1)$. Then $M = (T^n \times \mathbf{R})/\mathbf{Z}$ has a complete affine structure covered by the product affine structure on $T^n \times \mathbf{R}$. If A is not unipotent then the linear holonomy of M contains the nonunipotent operator $A \times I$ on $\mathbf{R}^n \times \mathbf{R}$. The affine holonomy group Γ is solvable (even polycyclic) since it embeds in the exact sequence

$$1 \to \mathbf{Z}^n \to \Gamma \to \mathbf{Z} \to 1.$$

We now turn to the problem of characterizing complete affine manifolds. The following theorem, one of the few general methods of proving completeness, shows the geometrical importance of unipotence.

6.8. THEOREM. *Let M be an affine manifold. Suppose either*

(a) *M is compact and has unipotent holonomy,*

or

(b) *M is an affine nilmanifold $\Gamma\backslash G$ and the corresponding linear representation $G \to \mathrm{GL}(E)$ is unipotent.*

Then M is complete.

Proof. First assume M is compact. From unipotence it follows that there is a flag of linear subspaces $0 = F_0 \subset F_1 \subset \cdots \subset F_n = E$, $\dim F_i = i$, invariant under the linear holonomy, such that the induced action on each F_i/F_{i-1} is trivial. Therefore there are linear maps $l_i : F_i \to \mathbf{R}$ which are invariant under the linear holonomy, and which have kernel F_{i-1}. (See e.g. Humphreys [23].)

The invariant flag on E determines a family of fields $\mathscr{F}_i$ of parallel i-planes on M, $i = 0, \ldots, n$. Each field $\mathscr{F}_i$ is integrable, as is clear in affine coordinates. The invariant linear map l_i determines a parallel 1-form ω_i defined on $\mathscr{F}_i$ and vanishing on $\mathscr{F}_{i-1}$. Using partitions of unity, one readily constructs vector fields X_i on M tangent to $\mathscr{F}_i$ with $l_i(X_i) = 1$. The X_i are covered by integrable vector fields $\tilde{X}_i$ on $\tilde{M}$, since M is compact.

We now show that $\tilde{M}$ develops *onto* E. Suppose one fixes a base point m_0 in $\tilde{M}$ which develops to the origin in E. Let $v \in E$. We construct a path in $\tilde{M}$ beginning at m_0 whose development ends at v as follows. Starting at m_0, flow along X_n for

time $l_n(v)$, ending at the point $m_1 \in \tilde{M}$ with $D(m_1) = v_1 \in E$ where $v - v_1 \in \operatorname{Ker} l_n = F_{n-1}$. Then flow along X_{n-1} for time $l_{n-1}(v - v_1)$ ending at $m_2 \in \tilde{M}$ with $D(m_2) = v_2 \in E$ where $v - v_2 \in \operatorname{Ker} l_{n-1} = F_{n-1}$. Continuing this way, we reach $v = v_n$ after n steps.

We verify that the developing map is injective. Assume $\gamma_0 : [a, b] \to \tilde{M}$ is a path which develops onto a closed loop $\delta : [a, b] \to E$ with $\delta(a) = \delta(b) = 0$. We may deform γ_0 with end points fixed so that the new path γ_1 develops in F_{n-1}: we use $\gamma_t : [a, b] \to \tilde{M}$, where $\gamma_t(s)$ is the image at time $-tl_n(\delta(s))$ of $\gamma_0(s)$ under the flow of $\tilde{X}_n$, $0 < t \leq 1$, $a \leq s \leq b$. By further such deformations, we eventually obtain a path from $\gamma_0(a)$ to $\gamma_0(b)$ which develops to a point. Hence $\gamma_0(a) = \gamma_0(b)$, so the developing map is injective and M is complete.

When M is an affine nilmanifold, not assumed compact, the proof is similar. We may as well assume $M = G$. Fix a flag at the identity $e \in G$ invariant under the linear part of the affine action of G. Let $X_1, \ldots, X_n$ be tangent vectors at e forming a basis for the flag, and extend them to vector fields on G by left-multiplication. Being left-invariant, each of these vector fields is integrable for all $t \in \mathbf{R}$. The rest of the proof is analogous to the compact case. QED.

The same proof shows that in the incomplete case, the projection $E_U \oplus F \to E_U$ maps the developing image onto E_U by a fibration.

The following application of 6.8 might be true even without nilpotence.

6.9. THEOREM. *Let M be a compact affine manifold with nilpotent affine holonomy group. Suppose the developing map $D : \tilde{M} \to E$ is surjective. Then M is complete.*

Proof. Let $E = E_U \oplus F$ be the Fitting splitting of the linear holonomy. If $F \neq 0$ then Theorem 3.3 holds that each component N of $p(D^{-1}E_U)$ is incomplete. Nevertheless N has unipotent holonomy and so, by Theorem 6.8, must be complete. These facts are only compatible if $F = 0$. It follows that $M = N$ is complete. QED.

At this point it may be useful to summarize the implications proved thus far between various properties of affine manifolds:

6.10. SUMMARY. For compact affine manifolds:

(a) Surjective developing map and nilpotent affine holonomy $\Rightarrow$ indecomposable affine holonomy (6.9).

(b) Complete $\Rightarrow$ irreducible affine holonomy (2.2).

(c) Unipotent linear holonomy $\Rightarrow$ complete (6.8).

(d) Nilpotent affine holonomy group and parallel volume $\Rightarrow$ unipotent linear holonomy (6.6).

(e) Nilpotent linear holonomy group and radiant $\Rightarrow$ expansion in linear holonomy (6.2).

For possibly noncompact affine manifolds:

(f) Indecomposable, nilpotent affine holonomy $\Rightarrow$ unipotent linear holonomy (4.3).

(g) A left-invariant affine structure on a nilpotent Lie group G is complete if and only if the corresponding affine representation of G is unipotent. (6.8, 1.5).

In a subsequent paper we will show that for compact affine manifolds, parallel volume $\Rightarrow$ irreducible affine holonomy.

The following case is especially neat, as all the conditions coincide:

6.11. THEOREM. *Let M be a compact affine manifold with nilpotent affine holonomy group. Then the following are equivalent*:

(a) *M is complete*

(b) *The linear holonomy is unipotent.*

(c) *M has parallel volume.*

(d) *The affine holonomy is irreducible.*

(e) *The affine holonomy is indecomposable.*

(f) *The developing map is surjective.*

Proof. (f)$\Rightarrow$(e) by 6.9. (e)$\Rightarrow$(b) by 4.3. (b)$\Rightarrow$(a) by 6.8. Clearly (a)$\Rightarrow$(f). Also (b)$\Rightarrow$(c), and (c)$\Rightarrow$(b) by 6.6. Finally, (a)$\Rightarrow$(d) by 2.2, and (d)$\Rightarrow$(e). QED.

In the next section we adjoin another equivalent condition: that M be a complete affine nilmanifold.

In a forthcoming paper [17] it will be shown that (d)$\Rightarrow$(a) fails for certain compact affine 3-manifolds with *solvable* fundamental group.

7. Complete affine nilmanifolds

The following theorem follows from a more general theorem announced by Auslander [8]. Unfortunately the proof is wrong, as Auslander shows in [6]. Many of the ideas in this paper, and in particular the proof of our Theorem 7.1, come from Auslander's work. (One can, however, deduce Theorem 7.1 from the correct results in [8] (see [8], page 811).)

7.1. THEOREM. *Let M be a compact complete affine manifold with nilpotent fundamental group. Then M is an affine nilmanifold.*

Proof. We take the universal covering space of M to be vector space E. The fundamental group is the group of deck transformation $\pi \subset \mathrm{Aff}(E)$; thus $M = E/\pi$.

To prove the theorem we must find a subgroup $G \subset \mathrm{Aff}(E)$, containing π, which acts simply transitively on E.

The linear action of π on E is unipotent (Corollary 3.4). Hence for every $g \in G$ there is a unique element $\log(g)$ in the Lie algebra $\mathrm{Aff}(E)$ of $\mathrm{Aff}(E)$ whose linear part is a nilpotent transformation of E and whose exponential is g.

By a celebrated theorem of Malcev [27] there is a simply connected nilpotent Lie group H containing π as a discrete uniform subgroup. Recall that the Lie algebra $\mathfrak{h}$ of H is generated by elements $L(g) \in \mathfrak{h}$, $g \in \pi$ subject only to the relations

$$C(L(g_1), L(g_2)) = L(g_1 g_2) \tag{1}$$

where $C : \mathfrak{h} \times \mathfrak{h} \to \mathfrak{h}$ satisfies the Baker–Campbell–Hausdorff formula (Varadarajan [43]):

$$\exp C(X, Y) = \exp X \exp Y.$$

Clearly (1) holds with $L(g) = \log(g)$, so there is a homomorphism of Lie algebras $\mathfrak{h} \to \mathrm{Aff}(E)$ and an associated homomorphism of Lie groups $H \to \mathrm{Aff}(E)$. The image of H is a connected subgroup $G \subset \mathrm{Aff}(E)$; and since for each $g \in G$ the linear part of $\log g$ is a nilpotent transformation, the image of G in $\mathrm{GL}(E)$ is unipotent. It follows that E has a basis putting this image in upper triangular unipotent form. The group of all affine automorphisms of E with such linear parts is simply connected. Thus G is simply connected. By construction, $\pi \subset G$.

Define $\tilde{f} : G \to E$ by $\tilde{f}(g) = g(0)$. Then $\tilde{f}$ is equivariant respecting the inclusion $G \subset \mathrm{Aff}(E)$. There is an induced map $f : \pi\backslash G \to M$. Now $\pi\backslash G$ and M are Eilenberg–MacLane spaces of type $K(\pi, 1)$, and f is a homotopy equivalence. Since they are compact manifolds, they have the same dimension and f must be surjective.

It follows that $\tilde{f} : G \to E$ is surjective; therefore G acts transitively on E. Since $\dim G = \dim E$, $\tilde{f}$ is a covering space projection. Since E is simply connected, $\tilde{f}$ is a homeomorphism. Therefore G acts simply transitively on E. Hence $f : \pi\backslash G \to M$ is an affine isomorphism, as claimed. QED.

In a forthcoming paper [13] we shall prove that a compact complete affine manifold with solvable holonomy has a finite covering which is an affine solvmanifold.

8. Polynomial tensors on affine manifolds with nilpotent holonomy

A tensor field on an affine manifold M is called *polynomial of degree* $\leq r$ if in affine coordinates its components are polynomial functions of (total) degree $\leq r$. It is of interest to know which fields are of this type, and which real cohomology classes are represented by polynomial exterior differential forms. (One can ask similar questions about other classes of tensors. For example: which tensors on M correspond in affine charts to real analytic tensors defined on all of E?).

The following theorem is the main result of this section:

8.1. THEOREM. *Let M be a compact affine manifold with nilpotent holonomy. If M is a complete affine nilmanifold the inclusion of the complex of polynomial exterior forms into the de Rham complex induces an isomorphism on cohomology.*

In Theorem 8.4 below a converse result is proved.

The proof of 8.1 relies on the theorem of Nomizu [33] (see also Raghunathan [47]) which identifies the real cohomology of a compact nilmanifold $\Gamma\backslash G$ with the cohomology of left-invariant differential forms on G. Therefore it suffices to prove that when G is a simply connected nilpotent Lie group with left-invariant complete affine structure, then every left-invariant tensor field on G is polynomial in affine coordinates. Now such a structure on G is defined by a simply transitive affine action of G on the vector space E; and left-invariant tensor fields on G correspond $\alpha(G)$-invariant tensor fields on E. Therefore it is enough to prove:

8.2. THEOREM. *Let G be a nilpotent Lie group and $\alpha : G \to \mathrm{Aff}(E)$ a simply transitive affine action. Then every $\alpha(G)$-invariant tensor field on E is polynomial.*

Let α be as in 8.2; then by Theorem 1.5 $\alpha(G)$ is a unipotent subgroup of $\mathrm{Aff}(E) \subset \mathrm{GL}(E \times \mathbf{R})$. Thus $\alpha(G)$ is a connected unipotent group of matrices. It is well known that such a group is a unipotent algebraic group, and in a unique way. On algebraic groups there is a natural notion of an algebraic tensor field.

The algebraic structure on G can be made explicit as follows. The exponential map $\exp : \mathfrak{G} \to G$ is a diffeomorphism, as is its inverse $\log : G \to \mathfrak{G}$. The Lie

algebra $\mathfrak{G}$, being a vector space, has a natural structure as an algebraic variety. The maps exp and log are isomorphisms of algebraic varieties.

It is easy to see that every left-invariant tensor field on G is algebraic, that is, it is polynomial *in the coordinates defined by exp*. What we have to prove is that it is *also* polynomial in the coordinates defined by the developing map

$$D : G \to E, \qquad D(g) = \alpha(g)(0)$$

defined by evaluation at the origin. To this end we prove:

8.3. PROPOSITION. *The developing map $D : G \to E$ is an isomorphism of algebraic varieties. In particular the composite maps*

$$\mathfrak{G} \xrightarrow{\exp} G \xrightarrow{D} E$$

and

$$\mathfrak{G} \xleftarrow{\log} G \xleftarrow{D^{-1}} E$$

are polynomial.

Proof. As remarked above, $\alpha : G \to \alpha(G)$ is an isomorphism of algebraic groups. Now D is the composition of α with the map

$$\mathrm{Aff}(E) \to E, \qquad g \mapsto g(0)$$

which is evidently algebraic. This proves that $D : G \to E$ is algebraic.

To prove that $D^{-1} : E \to G$ is algebraic we induct on $\dim G = \dim E$. The case $\dim G = 0$ is trivial.

Let $\lambda : G \to \mathrm{GL}(E)$ be the linear part of α; clearly λ is algebraic.

Since λ is unipotent there exists a nonzero linear functional $\psi : E \to \mathbf{R}$ which is λ-invariant. Thus if $g \in G$ then $\psi \circ \alpha(g)(x) - \psi(x) = \psi \circ D(g)$ for all $x \in E$. It follows easily that $\psi \circ D$ defines a homomorphism $G \to \mathbf{R}$. Let G_1 denote the kernel of this homomorphism and let $E_1 = \mathrm{Ker}\, \psi$. Clearly $\alpha(G_1)$ acts simply transitively on E_1.

Suppose, inductively, that the inverse map $f_1 : E_1 \to G_1$ to $D_1 = D|_{G_1}$ is algebraic. The homomorphism $G \to R$ has a left-inverse so we may write G as a semidirect product $G = G_1 \rtimes \{g_t\}_{t \in \mathbf{R}}$ where $t \mapsto g_t$ is a one-parameter subgroup of G which acts on E/E_1 by translation by t. Thus in a basis of E containing a basis

of E_1, $\alpha(g_t)$ is represented by

$$\left[\begin{array}{c|c} Q(t) & r(t) \\ \hline 0 \cdots 0 & 1 \end{array}\right]\left[\begin{array}{c} p(t) \\ \hline t \end{array}\right]$$

where $Q : \mathbf{R} \to \mathrm{GL}(E_1)$ and $r, p : \mathbf{R} \to E_1$ are polynomial maps. Define $f : E_1 \oplus \mathbf{R} = E \to G$ by

$$f(x, t) = g_t \cdot [f_1 \circ Q(-t)(x - p(t))].$$

Clearly f is a composition of algebraic maps and hence algebraic. Furthermore for $t \in \mathbf{R}$ and $h \in G_1$,

$$\begin{aligned} f \circ D(g_t h) &= f(\alpha(g_t)D(h)) = f((Q(t)D(h) + p(t), t)) \\ &= g_t \cdot f_1(Q(-t)((Q(t)D(h) + p(t)) - p(t))) = g_t \cdot f_1(D(h)) = g_t h \end{aligned}$$

so that f is inverse to D. Thus D^{-1} is algebraic, and the proof of 8.3 is complete. QED.

Now that 8.3 has been established, 8.2 and 8.1 are consequences because of Nomizu's theorem, as explained above.

Remark. In general we do not know a sharp bound on the degree of a left-invariant tensor field. A (probably crude) bound can be obtained by estimating the degrees of polynomials in the proof of 8.3. In the notation of 8.3 we have algebraic maps

$$f : E \to G, \qquad Q : \mathbf{R} \to \mathrm{GL}(E_1), \qquad P : \mathbf{R} \to E_1.$$

We give $\mathbf{R}$ and E_1 their natural vector space structure; thus $\deg p$ is well-defined. From the definition of p it is easy to compute that

$$\deg p \leq n = \dim E.$$

We give $\mathrm{GL}(E_1)$ matrix coordinates coming from linear coordinates on E_1. Then $\deg Q$ is well-defined and one sees easily that $\deg Q \leq n - 1$. Now G is a subgroup of $\mathrm{Aff}(E)$. In a natural way $\mathrm{Aff}(E) \subset \mathrm{GL}(E \times \mathbf{R})$, and we give $\mathrm{GL}(E \times \mathbf{R})$ matrix coordinates. Thus $\deg f$ is defined; and $\deg f_1$ is defined similarly. From the

formula

$$f(x, t) = g_t \cdot [f_1 \circ Q(-t)(x - p(t))],$$

taken from the proof of 8.3, we find

$$\deg f \leq (\deg f_1)(\deg Q + \deg p) \leq (\deg f_1)(2n - 1).$$

By induction on n we get

$$\deg f \leq 1 \cdot 3 \cdot 5 \cdots (2n-1) = (2n-1)!/2^{n-1}(n-1)!.$$

Now let X be a left-invariant vector field on G, considered as an $\alpha(G)$-invariant vector field $X : E \to E$. Let $x \in E$; set $f(x) = g \in G$. Then

$$X(x) = X(D \circ f(x)) = X(D(g)) = X(\alpha(g)(0)) = \lambda(g)X(0).$$

Thus $X(x) = \lambda(f(x))X(0)$. This expresses $X : E \to E$ as the composition

$$X : E \xrightarrow{f} G \xrightarrow{\lambda} \mathrm{GL}(E) \xrightarrow{h} E$$

where h is evaluation at $X(0)$. It is clear that λ and h both have degree 1. Thus $\deg X \leq \deg f$. Similarly the degree of a left-invariant 1-form is bounded by $\deg f$.

It follows that if T is a left-invariant (p, q)-tensor field on G then as a tensor field on E

$$\deg T \leq (p+q) \deg f \leq (p+q)(2n-2)!/2^{n-1}(n-1)!$$

It would be interesting to have a sharper bound. In Fried [4] there is an example where a 4-dimensional G has a left-invariant vector field of degree 5. If $\dim G < 4$, however, every left-invariant tensor of type (p, q) has degree $\leq 2(p+q)$ (see [13]).

We note that for any Lie group with left-invariant affine structure, all *right*-invariant vector fields are polynomial of degree ≤ 1 since in affine coordinates they generate 1-parameter subgroups of $\mathrm{Aff}(E)$, so they are affine vector fields on E.

The following theorems complement 6.11 and 7.1 by characterizing compact complete affine nilmanifolds in terms of differential forms.

8.4. THEOREM. *Let M be a compact affine manifold with nilpotent affine holonomy group. Then the following conditions are equivalent:*

(a) *M is a complete affine nilmanifold.*

(b) *M is orientable and every de Rham cohomology class is represented by a polynomial differential form.*

(c) *M has a polynomial volume form.*

(d) *M has a parallel volume form.*

Proof. Theorem 8.1 shows that (a)$\Rightarrow$(b). Clearly (b)$\Rightarrow$(c); and 6.11, 7.1 show that (d)$\Rightarrow$(a).

We must show that (c)$\Rightarrow$(d). If the linear holonomy is unipotent (d) is obvious. We shall prove that if the linear holonomy is not unipotent there is no polynomial volume form.

Denote the affine holonomy by $\alpha : \pi \to \mathrm{Aff}(E)$ and the linear holonomy by $\lambda : \pi \to \mathrm{GL}(E)$. Let $E = E_U \oplus F$ be the Fitting splitting of λ. We assume λ is not unipotent, i.e., $F \neq 0$. Since M has a volume form, M is orientable. By 6.5 there exists $g_1 \in \pi$ such that $\lambda(g_1) \mid F$ is a contraction. Since $\lambda(\pi) \mid E_U$ is unipotent, and $\mathrm{Det}\,\lambda(g_1) > 0$ because M is orientable, we have

$$\mathrm{Det}\,\lambda(g_1) = \delta, \qquad 0 < \delta < 1.$$

Fix $x_0 \in E$. It follows from the unipotence of $\lambda(g_1)$ on E_U and the contracting of $\lambda(g_1)$ on F that there is a polynomial p in one variable such that for all $m \geq 0$, $|\alpha(g_1)^m x_0| \leq p(m)$.

Suppose that θ is a polynomial n-form on E invariant under the affine holonomy $\alpha(\pi)$. Let μ be a nonzero parallel n-form on E and write $\theta(x) = f(x)\mu$ where $f : E \to \mathbf{R}$ is a polynomial. Then $\theta(x) = (\mathrm{Det}\,\lambda(g) \cdot \theta(\alpha(g)x)$ for all $g \in \pi$, so $f(x) = (\mathrm{Det}\,\lambda(g))f(\alpha(g)x)$ particular, for all $m \geq 0$, $f(x) = (\mathrm{Det}\,\lambda(g_1))^m f(\alpha(g_1)^m x)$, or

$$f(x) = \delta^m f(\alpha(g_1)^m x).$$

Since f is a polynomial, $|f(\alpha(g_1)^m x_0)$ is bounded in absolute value by a polynomial $q(m)$ (which depends on x). Hence

$$|f(x_0)| \leq \delta^m q(m) \quad \text{for all} \quad m \geq 0.$$

Since $0 < \delta < 1$, it follows that $f(x_0)$ is zero. As x_0 was arbitrary, this completes the proof of Theorem 8.4.

It is to be noted that Auslander [4] and Auslander–Markus [10] have constructed many examples of compact complete affine-nilmanifolds having parallel Lorentz metrics. See also Milnor [31] for additional examples.

BIBLIOGRAPHY

[1] ARROWSMITH, D. and FURNESS, P., *Locally symmetric spaces*, J. London Math. Soc. (2) *10* (1975), 487–499.

[2] ——. *Flat affine Klein bottles*, Geometriae Dedicata *5* (1976), 109–115.

[3] ATIYAH, M. and WALL, C., *Cohomology of groups*, in *Algebraic Number Theory*, Cassels and Frohlich (Eds), 94–115, Thompson Book Co. 1967.

[4] AUSLANDER, L., *Examples of locally affine spaces*, Annals of Math. *64* (1956), 255–259.

[5] ——, *An exposition of the theory of solvmanifolds*, Bull. Amer. Math. Soc. *79* (1973), 227–285.

[6] ——, *Simply transitive groups of affine matrices*, Amer. J. Math. *99* (1977), 809–821.

[7] ——, *Some compact solvmanifolds and locally affine spaces*, J. Math. and Mech. *7* (1958), 963–975.

[8] ——, *The structure of complete locally affine manifolds*, Topology *3* (1964), 131–139.

[9] AUSLANDER, L. and MARKUS, L., *Holonomy of flat affinely connected manifolds*, Annals of Math. *62* (1955), 139–151.

[10] ——, *Flat Lorentz 3-manifolds*, Memoirs Amer. Math. Soc. 1959.

[11] FILLMORE, J. and SCHEUNEMAN, J., *Fundamental groups of compact complete locally affine surfaces*, Pacific J. Math. *44* (1973), 487–496.

[12] FRIED, D., *Polynomials on affine manifolds*. To appear.

[13] —— and GOLDMANN, W., *3-dimensional affine crystallographic groups*. To appear.

[14] GOLDMAN, W. and HIRSCH, M., *Polynomial forms on affine manifolds*, Pacific J. Math. (in press).

[15] FURNESS, P. and FEDIDA, E., *Sur les structures pseudo-riemanniennes plates des variétés compactes*, C.R. Acad. Sci, Paris, *286* (1978), 1969–171.

[16] GOLDMAN, W., *Obstructions to deforming geometric structures*, to appear.

[17] ——, *Two examples of affine manifolds*, to appear.

[18] GOLDMANN, W. and HIRSCH, M., *A generalization of Bieberbach's theorem*, Invent. Math. (in press).

[19] GUNNING, R., *Uniformization of complex manifolds: the role of connections*, Princeton U. Press, 1978.

[20] HIRSCH, M., *Flat manifolds and the cohomology of groups*, in *Algebraic and geometric topology*, Lecture Notes in Mathematics 664, Springer-Verlag, 1977.

[21] ——, *Stability of stationary points and cohomology of groups*, Proc. Amer. Math. Soc. *79* (1980), 191–196.

[22] HIRSCH, M. and THURSTON, W., *Foliated bundles, flat manifolds, and invariant measures*, Annals of Math. *101* (1975), 369–390.

[23] HUMPHREYS, J., *Linear algebraic groups*, Springer-Verlag, 1975.

[24] INOUE, M., KOBAYASHI, S. and OCHIAI, T., *Holomorphic affine connections on compact complex surfaces*, J. Fac. Sci. U. Tokyo *27* (1980), 247–264.

[25] KOSTANT, B. and SULLIVAN, D., *The Euler characteristic of an affine space form is zero*, Bull. Amer. Math. Soc. *81* (1975), 937–938.

[26] KUIPER, N., *Sur les surfaces localement affines*, Colloques de géométrie différentielle, Strasbourg (1953), 79–87.

[27] MALCEV, A. *On a class of homogeneous spaces*, Izvestija Akademia Nauk SSR Ser. Math. *13* (1949); Amer. Math. Soc. Translations *9* (1962), 276–307.

[28] MARKUS, L., *Cosmological models in differential geometry*, U. Minnesota 1962 (mimeographed).

[29] MATSUSHIMA, Y., *Affine structures on complex manifolds*, Osaka J. Math. *5* (1968), 215–222.
[30] MILNOR, J., *Curvatures of left-invariant metrics on Lie groups*, Adv. in Math. *21* (1976), 293–329.
[31] ——, *On fundamental groups of complete affinely flat manifolds*, Adv. in Math. 25 (1977), 178–187.
[32] NAGANO, T. and YAGI, K., *The affine structures on the real two-torus, I*, Asapa J. Math. *11* (1974), 181–210.
[33] NOMIZU, K., *On the cohomology ring of compact homogeneous spaces of nilpotent Lie groups*, Annals of Maths. *59* (1954), 531–538.
[34] SAKANE, Y., *On compact complex affine manifolds*, J. Math. Soc. Japan *29* (1977), 135–145.
[35] SCHEUNEMAN, J., *Examples of compact locally affine spaces*, Bull. Amer. Math. Soc. *77* (1974), 589–592.
[36] ——, *Affine structures on three-step nilpotent Lie algebras*, Proc. Amer. Math. Soc. *46* (1974), 451–454.
[37] ——, *Translations in certain groups of affine motions*, Proc. Amer. Math. Soc. *47* (1975), 223–228.
[38] SMILLIE, J., *Affinely flat manifolds*, Doctoral thesis, U. Chicago (1977).
[39] SULLIVAN, D., *A generalization of Milnor's inequality concerning affine foliations and affine manifolds*, Comment. Math. Helv. *51* (1976), 183–189.
[40] SULLIVAN, D. and THURSTON, W., *Manifolds with canonical coordinates: some examples*. Institut des Hautes Etudes, 1979, mimeographed.
[41] SUWA, T., *Compact quotient spaces of C^2 by affine transformation groups*, J. Diff. Geom. *10* (1975), 239–252.
[42] TISCHLER, D., *On fibering certain foliated manifolds over S^1*, Topology *9* (1970), 153–154.
[43] VARADARAJAN, V., *Lie groups, Lie algebras, and their representations*, Printice Hall, 1974.
[44] VITTER, A., *Affine structures on compact complex manifolds*, Invent. Math. *17* (1972), 231–244.
[45] ZASSENHAUS, H., *Beweis eines Satzes über diskrete Gruppen*, Abhandl. Math. Sem. Univ. Hamburg *12* (1938), 289–312.
[46] DWYER, W. G., *Vanishing homology over nilpotent groups*, Proc. Amer. Math. Soc. *49* (1975), 8–12.
[47] RAGHUNATHAN, M. S. *Discrete subgroups of Lie groups*, Springer-Verlag 1972.

University of California, Santa Cruz
University of California, Berkeley
University of California, Berkeley

Received March 24, 1981

Comment. Math. Helvetici **56** (1981) 524–557 0010–2571/81/004524–34$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Double collisions for a classical particle system with nongravitational interactions

RICHARD MCGEHEE

1. Introduction

In 1907 K. F. Sundman proved that a solution of the Newtonian gravitational three body problem for which the angular momentum is nonzero can be written in terms of a power series convergent for all time [12]. The major obstacle overcome by Sundman was presented by collisions between the particles. The condition that the angular momentum be nonzero allowed him to eliminate the possibility of a collision between all three particles. Since he could not eliminate the possibility of a collision between two of the particles, he was forced to investigate the nature of solutions with double collisions.

Sundman discovered that a solution which has a double collision as $t \to t^*$ can be written as a convergent power series in $(t^*-t)^{1/3}$. This solution cannot be extended holomorphically to t^*. However, using analytic continuation in the complex plane around the branch point t^*, Sundman found a real analytic extension for $t > t^*$. Introducing a new variable, which locally has the form $(t^*-t)^{1/3}$ at each collision, he was able to extend the solution through each double collision and to write it as a convergent power series in the new variable. A complete description of Sundman's construction can be found in the book of Siegel and Moser [10].

From the vantage point provided us by the current theory of dynamical systems, we can question the significance of Sundman's power series expansion. Today an ordinary differential equation is viewed as a vector field on a manifold, and its solutions are viewed as determining a flow on that manifold. One attaches as much importance to the smoothness of the flow with respect to initial data as to its smoothness with respect to time. Sundman's extension of an orbit through a double collision would be considered of questionable significance if it were unrelated to nearby orbits. Since Sundman's extension accounts only for the dependence on time and not for the dependence on initial data, we do not know *a priori* whether his extension is related to nearby orbits.

Levi-Civita has given us a different method for extending orbits through

double collisions [6]. He eliminated the singularities in the vector field by transforming the equations to ones without singularities. The extension through double collisions then is given by the transformed equations and is automatically a smooth function of initial data.

It happens that Sundman's extension and Levi-Civita's extension coincide. One is tempted to believe that somehow the power series technique carries with it the dependence on initial data. Since holomorphic functions are completely determined by their behavior on small sets, and since the solutions of the three body problem are holomorphic, one can easily imagine that an analytic extension in time will automatically give analytic dependence on initial data. It does so for the gravitational potential, but we shall see below that it fails to do so for other potentials.

To contrast these two methods of extension we study a particle system with nongravitational interactions. Instead of assuming that the pairwise potential energy is r^{-1}, where r is the distance between the two particles, we assume that the potential energy is $r^{-\alpha}$, where α is some positive real number. We then study Sundman's technique versus Levi-Civita's. We shall see that solutions can be extended as a power series in time whenever α is in a certain dense subset of the rationals. However, only a few of these values of α produce double collisions which can be extended according to Levi-Civita. For the other values, Sundman's technique gives extensions which are not continuous with respect to initial data.

Thus we see that it is only a peculiar property of the gravitational interaction which gives us the same orbit extensions for the two different techniques. If one were to carry out Sundman's entire program for a three body problem with a nongravitational potential, one might obtain a power series solution, convergent for all time, but for which small changes in initial data could produce drastic changes in the orbits.

2. Equations of motion

We study a system of two particles. Since the two particles will always move in some fixed plane, we take R^2 for the position space. We fix the center of mass at the origin, so the system reduces to that of a single particle of unit mass in a central force field. We take the potential energy function to be

$$U(x) = -|x|^{-\alpha}, \qquad \alpha > 0,$$

where $x \in R^2$ is the position of the single particle. The motion is described by the

differential equation

$$\ddot{x} = -\text{grad } U(x) = -\alpha |x|^{-\alpha-2} x, \tag{2.1}$$

where the double dot represents the second time derivative.

A double collision occurs if the two particles coincide and corresponds to $x = 0$ in the reduced problem. The origin is a "singularity" of equation (2.1), i.e. it is a point where the equation is undefined. Indeed, the potential energy approaches $-\infty$ as $x \to 0$.

A "singularity" of a solution of (2.1) is a time t^* beyond which the solution cannot be extended as a smooth function of real t. A precise definition of this notion is given in Section 3 below. We shall show that, for equation (2.1), singularities of the solution correspond to singularities of the differential equation, i.e. $x \to 0$ as $t \to t^*$. The only singularities are therefore double collisions.

In Section 4 we develop some properties of the solutions of equation (2.1). We shall see that $\alpha = 2$ is a dividing line between two very different types of behavior. For example, equation (2.1) has many circular periodic orbits. For $\alpha < 2$, these orbits are all stable; for $\alpha > 2$, they are all unstable. As another example, the angular momentum must be zero for a collision orbit if $\alpha < 2$. However, if $\alpha > 2$, there are collision orbits with arbitrary angular momentum.

It is interesting to note that the case $\alpha = 2$ has another property not discussed here. A system of n particles moving along a line with pairwise interaction given by an inverse square potential is completely integrable [9].

In Section 5 we discuss Sundman's technique of extending singular solutions. We call his technique "branch regularization." Roughly speaking, a singularity t^* of a solution can be considered as a branch point, where time t is regarded as complex. We then ask whether we can find a real analytic branch which extends the solution. We are able to find such a branch whenever α has the form $2((q/p)-1)$, where q and p are relatively prime integers with q odd. If p is even, as it is for the gravitational potential, then the extension can be described as a "reflection," i.e. the velocity vector reverses direction at collision, and the particles effectively bounce off each other. If p is odd, then the extension can be described as a "transmission," i.e. the direction of the velocity vector is preserved at collision, and the particles effectively pass through each other.

In Section 6 we turn to Levi-Civita's idea of focusing attention on the singularities of the differential equation rather than on the singularities of the solution. It is difficult to describe his technique in general terms, since it involves an apparently *ad hoc* change of variable. However, Easton has given a general definition of regularization in the spirit of Levi-Civita [4], Easton's idea is to use an isolating block to examine whether orbits passing close to collision determine

an extension for an orbit ending in collision. Easton calls his technique "regularization by surgery." However, this author feels that the use of surgery plays a secondary role to the use of isolating blocks and therefore prefers to call the technique "block regularization."

In Sections 6 and 7 we apply Easton's definition to equation (2.1). We find that the equation is block regularizable if and only if $\alpha = 2(1-n^{-1})$, where n is a positive integer. We then discuss the similarities and differences between branch regularization and block regularization.

Finally in Section 8 we exhibit a transformation analogous to that given by Levi-Civita. Our transformation works for those values of α for which equation (2.1) is block regularizable. It reduces to Levi-Civita's transformation when $\alpha = 1$.

3. Singularities of solutions

In this section we prove that the only singularities of solutions of equation (2.1) are collisions. We begin with the general definition of a singularity of a solution of the differential equation

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}). \tag{3.1}$$

Here $\mathbf{F}$ is a real analytic vector field on some open set $U \subset R^n$, and the dot denotes differentiation with respect to time t. If $\mathbf{x}(0) \in U$, then the standard existence and uniqueness theorems give us a unique solution $\mathbf{x}(t)$ defined for $t^- < t < t^+$, where $-\infty \leqslant t^- < 0 < t^+ \leqslant +\infty$ and (t^-, t^+) is the maximal interval over which the solution can be extended.

DEFINITION. If $t^+ < \infty$, then the solution $\mathbf{x}(t)$ is said to *end in a singularity at* t^+. If $t^- > -\infty$, then $\mathbf{x}(t)$ is said to *begin in a singularity* at t^-. In either case, $t^* = t^+$ or t^- is said to be a *singularity* of the solution $\mathbf{x}(t)$.

We can write equation (2.1) as a first order system by introducing the momentum vector $y = \dot{x}$. The equation then becomes

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -\alpha\, |x|^{-\alpha-2} x, \end{aligned} \tag{3.2}$$

which is a Hamiltonian system with Hamiltonian function

$$H(x, y) = \tfrac{1}{2}|y|^2 - |x|^{-\alpha}. \tag{3.3}$$

The function H represents the total energy of the particle and is a conserved quantity; i.e. H is constant along solutions of equations (3.2).

We now state the main result of this section. The proof is given after the proof of Lemma 3.2.

THEOREM 3.1. *Suppose that* $(x(t), y(t))$ *is a solution of* (3.2) *with a singularity at* t^*. *Then* $x(t) \to 0$ *as* $t \to t^*$.

LEMMA 3.2. *Suppose that* $(x(t), y(t))$ *is a solution of* (3.2) *with* $H(x, y) = h$, *and suppose that* $|x(t_0)| > 2\varepsilon > 0$. *Then there exist constants* δ *and* M, *depending on* h *and* ε, *but independent of* $(x(t_0), y(t_0))$, *such that*

$$|y(t)| < M \quad \text{and} \quad |x(t)| > \varepsilon, \quad \text{whenever} \quad |t - t_0| < \delta.$$

Proof. We take

$$M = (2h + 2\varepsilon^{-\alpha})^{1/2} \quad \text{and} \quad \delta = \varepsilon / M.$$

It suffices to show that

$$|x(t)| > \varepsilon \quad \text{whenever} \quad |t - t_0| < \delta, \tag{3.4}$$

since $H(x(t), y(t)) = h$ implies that

$$|y(t)| = (2h + 2|x(t)|^{-\alpha})^{1/2},$$

and hence that

$$|y(t)| < M \quad \text{whenever} \quad |x(t)| > \varepsilon. \tag{3.5}$$

Suppose that (3.4) is false. Then there exists a t_1 such that $|t_1 - t_0| < \delta$ and $|x(t_1)| \leqslant \varepsilon$. By continuity one then can find a t_2 such that

$$\begin{aligned} &|t_2 - t_0| < \delta, \\ &|x(t_2)| = \varepsilon, \end{aligned} \tag{3.6}$$

and

$$|x(t)| > \varepsilon \quad \text{whenever} \quad |t - t_0| < |t_2 - t_0|.$$

Since $y = \dot{x}$, (3.5) now implies that

$$|\dot{x}(t)| < M \quad \text{for} \quad |t - t_0| < |t_2 - t_0|.$$

Hence $|x(t_2)| > |x(t_0)| - M|t_2 - t_0|$, i.e. $\varepsilon > 2\varepsilon - M|t_2 - t_0|$, or $|t_2 - t_0| > \varepsilon/M = \delta$. But this contradicts (3.6), establishing (3.4) and proving the lemma.

Proof of Theorem 3.1. Let $H(x, y) = h$. Assume that $x(t)$ does not approach 0 as $t \to t^*$. Then there exist an $\varepsilon > 0$ and a sequence $\{t_n\}$, with $t_n \to t^*$, such that

$$|x(t_n)| > 2\varepsilon \quad \text{for all} \quad n. \tag{3.7}$$

Let δ and M be given by Lemma 3.2. Choose n so that $|t^* - t_n| < \delta$. Then

$$|y(t)| < M \quad \text{and} \quad |x(t)| > \varepsilon \quad \text{whenever} \quad |t - t_n| < \delta.$$

Equations (3.2) then imply that

$$|\dot{x}(t)| < M \quad \text{and} \quad |\dot{y}(t)| < \alpha\varepsilon^{-\alpha-1} \quad \text{whenever} \quad |t - t_n| < \delta.$$

Hence $(x(t), y(t))$ actually approaches some limiting point (x^*, y^*) as $t \to t^*$. By (3.7), $x^* \geq 2\varepsilon$. Therefore the solution can be extended beyond t^*, which contradicts the hypothesis that t^* is a singularity. Hence $x(t) \to 0$ as $t \to t^*$ and the theorem is proved.

4. A geometric description of the flow

We now present a description of the orbit structure of system (3.2), with special emphasis on the orbits near collision. The transformations used here are essentially the same as those used by this author and others to study collisions in Newtonian gravitational systems. (See [8] for specific references.) Similar transformations have been used also by Devaney in his study of the anisotropic Kepler problem [3].

It is convenient to introduce the constants

$$\beta = \frac{\alpha}{2} \quad \text{and} \quad \gamma = \frac{1}{\beta + 1}.$$

We identify the real plane R^2 with the complex plane C^1 and consider x as a complex number or as a vector in the Euclidean plane, depending on context. As

usual, a bar denotes the complex conjugate. We shall use the angular momentum integral

$$\Omega(x, y) = \text{Im}\,(\bar{x}y).$$

We introduce the new real coordinates $r>0$, θ, w, and v by letting

$$\begin{aligned} x &= r^{\gamma}e^{i\theta}, \\ y &= r^{-\beta\gamma}(v+iw)e^{i\theta}. \end{aligned} \tag{4.1}$$

System (3.2) then becomes

$$\begin{aligned} \dot{r} &= (\beta+1)v, \\ \dot{\theta} &= r^{-1}w, \\ \dot{w} &= r^{-1}(\beta-1)wv, \\ \dot{v} &= r^{-1}(w^2+\beta(v^2-2)). \end{aligned} \tag{4.2}$$

If we take the energy integral $H(x, y)$ given by (3.3) to have the constant value h and the angular momentum integral $\Omega(x, y)$ to have the constant value c, then we can write

$$w^2+v^2-2 = 2hr^{\alpha\gamma}, \tag{4.3}$$

$$r^{(1-\beta)\gamma}w = c. \tag{4.4}$$

We now define the constant energy manifold

$$\mathbf{M}(h) = \{(r, \theta, w, v) \in R^4 : r \geqslant 0 \text{ and (4.3) holds}\}. \tag{4.5}$$

System (4.2) determines a vector field on $\mathbf{M}(h)$ which is undefined when $r=0$. We let

$$\mathbf{N} = \{(r, \theta, w, v) \in \mathbf{M}(h) : r = 0\}, \tag{4.6}$$

which is the manifold of states corresponding to collision for equation (2.1). From the definition of $\mathbf{M}(h)$, we see that

$$\mathbf{N} = \{(r, \theta, w, v) \in R^4 : r = 0 \quad \text{and} \quad w^2+v^2 = 2\}$$

and hence that **N** is independent of h. Since θ is considered modulo 2π, **N** is a two dimensional torus.

As mentioned above, the vector field given by (4.2) is not defined on **N**. A collision corresponds to an orbit which approaches **N** in finite time. However, we can scale the vector field in such a way that the new vector field can be extended to **N**. We accomplish this scaling by introducing a new time parameter τ given by

$$dt = r\,d\tau. \tag{4.7}$$

Equations (4.2) then become

$$\begin{aligned} r' &= (\beta+1)rv, \\ \theta' &= w, \\ w' &= (\beta-1)wv, \\ v' &= w^2+\beta(v^2-2), \end{aligned} \tag{4.8}$$

where the prime denotes differentiation with respect to τ. For this new vector field, **N** is an invariant set. As we shall see in Lemma 4.1 below, collision orbits now approach **N** asymptotically as $\tau \to \pm\infty$.

The solutions of the entire system (4.8) are determined by the last two of those equations in the following sense. The equations for w' and v' do not involve r and θ and therefore may be solved independently. Then r and θ can be determined from the first two of equations (4.8). If $h \neq 0$, then r could also be determined from equation (4.3).

We therefore consider the last two of equations (4.8) as a separate system:

$$\begin{aligned} w' &= (\beta-1)wv, \\ v' &= w^2+\beta(v^2-2). \end{aligned} \tag{4.9}$$

These equations admit the integral

$$\Lambda(w, v) = |w|^{\alpha}\,|v^2+w^2-2|^{1-\beta}. \tag{4.10}$$

This integral is derived from the angular momentum and energy integrals as follows.

$$\Lambda = |\Omega(x, y)|^{\alpha}\,|2H(x, y)|^{1-\beta}$$

is a function of integrals for system (3.2) and is therefore itself an integral. Written in terms of r, θ, w, and v, this expression becomes equation (4.10).

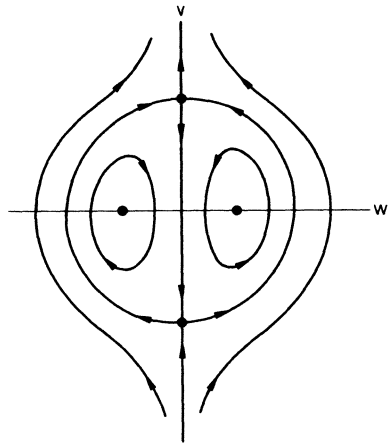


Figure 4.1.

The orbit structure of system (4.9) is completely determined by the integral Λ. There are three distinct cases depending on whether $\beta < 1$, $\beta = 1$, or $\beta > 1$. These cases are shown in Figures 4.1, 4.2, and 4.3 respectively. In each case, the circle $\{w^2 + v^2 = 2\}$ is invariant and corresponds to the collision manifold $\mathbf{N}$. The constant energy manifold $\mathbf{M}(h)$ projects to $\{w^2 + v^2 \leqslant 2\}$ when $h < 0$, to $\{w^2 + v^2 \geqslant 2\}$ when $h > 0$, and to $\{w^2 + v^2 = 2\}$ when $h = 0$.

In the remainder of this section we describe the relevant aspects of the solutions of system (4.8). Most of the statements can be derived from an examination of the integral Λ. Those that will be used in later sections are labeled as Lemmas and proved.

When $\beta < 1$, the flow is similar to that of the Kepler problem, which is given by $\beta = \frac{1}{2}$. If we take $h \geqslant 0$, which corresponds to hyperbolic or parabolic orbits in the Kepler problem, and if we take $w \neq 0$, which means that the angular momentum is nonzero, then all solutions are unbounded in both directions. The

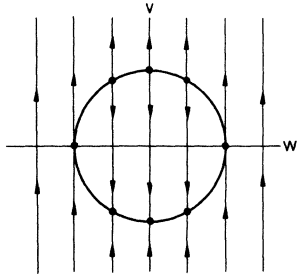


Figure 4.2.

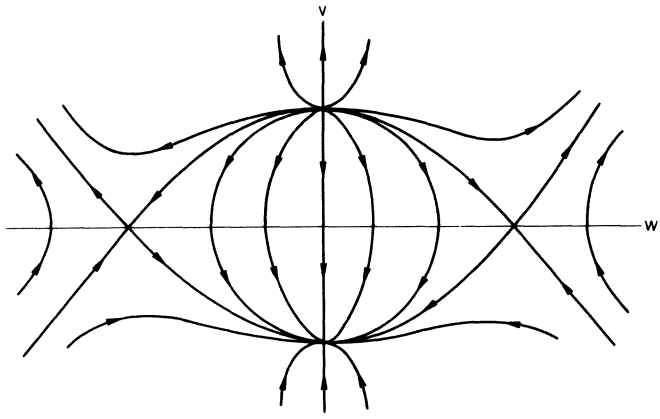


Figure 4.3.

orbits with zero angular momentum either begin or end in collision. If we take $h<0$, which corresponds to elliptic or circular orbits in the Kepler problem, then all solutions are bounded. There are two circular orbits, corresponding to the critical points $(w, v)=(\pm\sqrt{\alpha}, 0)$ in Figure 4.1. These circular orbits are stable. Taking the angular momentum to be zero distinguishes a one parameter family of orbits beginning and ending in collision. All other solutions move along invariant tori. One feature that distinguishes the Kepler problem is that each torus is foliated by periodic orbits. One does not expect this foliation for arbitrary $\alpha<2$. Instead one expects most of the tori to be filled with quasi-periodic orbits.

For $\beta=1$, the variable w becomes the angular momentum integral, as can be seen from equation (4.4). In this case, the invariant circle $\{w^2+v^2=2\}$ becomes a circle of critical points for system (4.9). When $h<0$, every solution begins and ends in collision. When $h>0$ and $|w|\leqslant\sqrt{2}$, every orbit either begins or ends in collision and becomes unbounded in the other direction. When $h>0$ and $|w|>\sqrt{2}$, every orbit is unbounded in both time directions. When $h=0$, the variable v also becomes an integral, as can be seen from equation (4.3). If $v\neq 0$, then the first of equations (4.8) immediately implies that $r=r_0e^{(\beta+1)v\tau}$ and hence that every solution begins or ends in collision and becomes unbounded in the other direction. Setting $v=0$ distinguishes a one parameter family of circular periodic orbits $r=r_0$, where r_0 is constant.

When $\beta>1$, the flow is drastically different from the Kepler problem. The circular periodic orbits, corresponding to the critical points $(w, v)=(\pm\sqrt{\alpha}, 0)$ in Figure 4.3, now occur for positive energy and are unstable. Among the orbits asysmptotic to one of the circular orbits, some are collision orbits while the others are unbounded. When $h>0$ there are unbounded collision orbits as well as orbits

beginning and ending in collision and orbits unbounded in both directions. When $h \leqslant 0$ all solutions begin and end in collision except those with $h = 0$ and $c = 0$. These begin or end in collision and are unbounded in the other direction.

In the following lemmas, we shall consider r as a function of t or of τ, depending on whether we are considering a solution of system (4.2) or of system (4.8). The orbits of both systems are the same; only the rate at which solutions move along the orbits is different. Using equation (4.7), we can write τ as a function of t along a particular orbit:

$$\tau(t) = \tau_0 + \int_{t_0}^{t} \frac{ds}{r(s)}. \tag{4.11}$$

Here s is a dummy variable for t, so the $r(s)$ under the integral is a solution of (4.2). Note that $\tau_0 = \tau(t_0)$ and that $r(\tau_0) = r(t_0)$, where the first r now denotes a solution of (4.8). Note also that any solution of (4.2) must have $r(t) > 0$ for all t.

We now show that orbits with singularities in the original time variable t become orbits for which r approaches 0 asymptotically in the new time variable τ.

LEMMA 4.1. *Suppose that* $(r, \theta, w, v)(t)$ *satisfies* (4.2) *and that* $r \to 0$ *as* $t \to t^* \pm$. *Then* $\tau(t) \to \mp\infty$.

Proof. We prove that $t \to t^*-$ implies that $\tau(t) \to +\infty$. The other case is similar. Equation (4.11) implies that $\tau(t)$ is defined for all $t_0 < t < t^*$ and is increasing. Hence $\tau(t) \to \tau^*$ as $t \to t^*-$, where τ^* may be infinite. Then $r(\tau) \to 0$ as $\tau \to \tau^*$. Since equations (4.8) are defined and smooth on $\mathbf{M}(h)$, and since $\mathbf{N}$ is a compact invariant set, it is impossible for a solution to approach $\mathbf{N}$ in finite time. Therefore $\tau^* = \infty$, and the lemma is proved.

For the following two lemmas, we use the notation

$$\mathbf{S}^{\pm} = \{(r, \theta, w, v) \in \mathbf{N} : v = \pm\sqrt{2}\}. \tag{4.12}$$

By the definition of $\mathbf{N}$,

$$\mathbf{S}^{\pm} = \{(r, \theta, w, v) \in R^4 : r = 0,\ w = 0,\ v = \pm\sqrt{2}\},$$

and hence $\mathbf{S}^+$ and $\mathbf{S}^-$ are both circles. We denote by $\omega(\mathbf{p})$ the omega limit set of the point $\mathbf{p} = (r, \theta, w, v)$ under the flow on $\mathbf{M}(h)$ defined by equations (4.8). Similarly, $\alpha(\mathbf{p})$ denotes the alpha limit set.

LEMMA 4.2. *Assume* $\beta \neq 1$. *Let* $\mathbf{p}(\tau) = (r, \theta, w, v)(\tau)$ *be a solution of* (4.8),

and write $\mathbf{p}_0 = \mathbf{p}(0)$. *Then*

(a) $r \to 0$ *as* $\tau \to +\infty$ *if and only if* $\omega(\mathbf{p}_0) \subset \mathbf{S}^-$, *and*

(b) $r \to 0$ *as* $\tau \to -\infty$ *if and only if* $\alpha(\mathbf{p}_0) \subset \mathbf{S}^+$.

Proof. As above, we prove only (a), since (b) is similar. Since $\mathbf{N}$ is exactly the set on $\mathbf{M}(h)$ where $r = 0$, $r \to 0$ as $\tau \to +\infty$ if and only if $\omega(\mathbf{p}_0) \subset \mathbf{N}$. Since $\mathbf{S}^- \subset \mathbf{N}$, it therefore suffices to show that $\omega(\mathbf{p}_0) \subset \mathbf{N}$ implies that $\omega(\mathbf{p}_0) \subset \mathbf{S}^-$. Using the energy relation (4.3) with $r = 0$, we can rewrite the equation for v' in system (4.8) restricted to $\mathbf{N}$ as

$$v' = (1-\beta)w^2.$$

Therefore v defines a Liapunov function on $\mathbf{N}$. Since a Liapunov function must be constant on an omega limit set [1], we must have $v' = 0$ on $\omega(\mathbf{p}_0)$. But $v' = 0$ on $\mathbf{N}$ exactly on $\mathbf{S}^+$ and $\mathbf{S}^-$. We can rule out $\mathbf{S}^+$ since v is positive in a neighborhood of $\mathbf{S}^+$ and the first of equations (4.8) implies that r is increasing whenever r is positive. Therefore $\omega(\mathbf{p}_0) \subset \mathbf{S}^-$ and the proof is complete.

LEMMA 4.3. *Assume* $\beta \neq 1$. *Let* $(r, \theta, w, v)(t)$ *be a solution of* (4.2) *such that* $r \to 0$ *as* $t \to t^* \pm$. *Then, as* $t \to t^* \pm$,

$$w(t) \to 0, \qquad v(t) \to \pm\sqrt{2}, \quad \textit{and} \quad \theta(t) \to \theta^*,$$

for some constant θ^*. *Furthermore, if* $\beta < 1$, *then* $w(t) \equiv 0$ *and* $\theta(t) \equiv \theta^*$.

Proof. As before, we prove only the case $t \to t^* -$. By Lemma 4.1 it suffices to prove the same limits for system (4.8) as $\tau \to +\infty$. By Lemma 4.2 it suffices to prove these limits for all points in $\mathbf{M}(h)$ on the stable manifold of $\mathbf{S}^-$. On $\mathbf{S}^-$, $\theta' = 0$, so $\mathbf{S}^-$ consists entirely of rest points. A computation shows that the eigenvalues at each of these rest points are 0, $-\sqrt{2}(\beta - 1)$, and $-2\sqrt{2}\beta$. The zero eigenvalue corresponds to the tangential direction along $\mathbf{S}^-$, while the two others correspond to the normal directions. For $\beta \neq 1$, $\mathbf{S}^-$ has hyperbolic normal structure, and hence the stable manifold of $\mathbf{S}^-$ is the union of the stable manifolds of each of the points on $\mathbf{S}^-$ [5]. Thus the solution $(r, \theta, w, v)(t)$ approaches some point on $\mathbf{S}^-$, i.e. $\theta \to \theta^*$, $w \to 0$, and $v \to -\sqrt{2}$.

When $\beta < 1$, one normal eigenvalue is positive and the other is negative. In this case the stable manifold is two dimensional and, in fact, consists only of points with $w = 0$. (See Figure 4.1) Since $\theta' = w = 0$, θ must be constant. Therefore $w \equiv 0$ and $\theta \equiv \theta^*$, and the proof is complete.

Equation (4.4) tells us that the angular momentum is zero if w is zero and r is

positive. Therefore the final statement of Lemma 4.3 implies that, for $\beta < 1$, the angular momentum must be zero on a collision orbit. Furthermore, all the motion takes place along a fixed line. These properties are not true for $\beta \geq 1$, as is seen in Figures 4.2 and 4.3.

5. Branch regularization of solutions

In this section we define the classical notion of regularization of solutions. We then determine for which α equation (2.1) is regularizable in this sense. We begin with a general definition concerning the solutions of the equation

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}). \tag{5.1}$$

DEFINITION. Let $\phi_1(t)$ and $\phi_2(t)$ be solutions of (5.1). Suppose that ϕ_1 ends in a singularity at time t^* and that ϕ_2 begins in a singularity at the same time. Suppose that there is a multivalued analytic function having a branch at t^* and extending both ϕ_1 and ϕ_2, Then ϕ_1 is said to be a *branch extension* of ϕ_2 at t^*, and ϕ_2 is said to be a *branch extension* of ϕ_1 at t^*,

DEFINITION. A solution ϕ of equation (5.1) which either begins or ends in a singularity at t^* is said to be *branch regularizable* at t^* if it has a unique branch extension at t^*. Equation (5.1) is said to be *branch regularizable* if every solution is branch regularizable at every singularity.

It is important to note that time t is considered to be complex in the above definitions. The solutions ϕ_1 and ϕ_2 are real analytic, i.e. they are real for real values of t. However, they are extensions of one another through complex values of t.

It is also important not to be confused by our use of complex numbers to represent points in the Euclidean plane. A solution $x(t)$ of equation (2.1) is not a a real analytic function of t. Instead, $x_1(t)$ and $x_2(t)$ are real analytic functions of t, where $x = x_1 + ix_2$.

We are interested in branch extensions for solutions of equation (2.1) or, equivalently, of system (3.2). Since $y = \dot{x}$, $y(t)$ will automatically have a branch extension whenever $x(t)$ has one. It is therefore sufficient to consider only $x(t)$.

We denote the rational numbers by Q. Whenever we write $p/q \in Q$, we mean that p and q are relatively prime integers. A special subset of Q is important here:

$$Q_0 = \{p/q \in Q : q \text{ is odd and } 0 < p < q\}.$$

The following theorem tells us for which values of α equation (2.1) is branch regularizable. Recall that $\alpha = 2\beta$ and that $\gamma = (1+\beta)^{-1}$. Since $\alpha > 0$, we must have $0 < \gamma < 1$.

THEOREM 5.1. *Equation* (2.1) *is branch regularizable if and only if* $\gamma \in Q_0$.

We break this theorem into two parts, one for each direction of the implication. Theorem 5.1 is an immediate consequence of the following two theorems. Note that Theorem 5.2 is actually stronger than the implication stated in Theorem 5.1.

THEOREM 5.2. *If* $\gamma \notin Q_0$, *then no singular solution of equation* (2.1) *is branch regularizable.*

THEOREM 5.3. *If* $\gamma \in Q_0$, *then every singular solution of equation* (2.1) *is branch regularizable.*

To prove these theorems we must know the exact analytical dependence on time of the solutions near a singularity. This information is provided by the following three lemmas, which will be proved after they are first used to establish Theorems 5.2 and 5.3.

LEMMA 5.4. *Suppose* $\beta < 1$. *Let* $x(t)$ *be a solution of equation* (2.1) *such that* $x(t) \to 0$ *as* $t \to 0+$. *Then*

$$x(t) = \kappa t^{\gamma} \Phi(t^{2\beta\gamma}),$$

where κ *is a complex constant with* $|\kappa| = 1$ *and where* Φ *is a real analytic function defined on a neighborhood of* 0, *with* $\Phi(0) = (\sqrt{2(\beta+1)})^{\gamma}$.

LEMMA 5.5. *Suppose* $\beta > 1$. *Let* $x(t)$ *be a solution of equation* (2.1) *such that* $x(t) \to 0$ *as* $t \to 0+$. *Then*

$$x(t) = \kappa[t^{\gamma}\Psi_1(t^{2\beta\gamma}, t^{2(\beta-1)\gamma}) + it^{\beta\gamma}\Psi_2(t^{2\beta\gamma}, t^{2(\beta-1)\gamma})],$$

where κ *is a complex constant with* $|\kappa| = 1$ *and where* Ψ_1 *and* Ψ_2 *are real analytic functions defined on a neighborhood of* $(0, 0)$, *with* $\Psi_1(0, 0) = (\sqrt{2(\beta+1)})^{\gamma}$.

LEMMA 5.6. *Suppose* $\beta = 1$. *Let* $x(t)$ *be a solution of equation* (2.1) *such that* $x(t) \to 0$ *as* $t \to 0+$. *Let* c *be the constant value of the angular momentum* Ω *and* h

be the constant value of the energy H along the solution. Then $c^2 \leq 2$, and $h > 0$ if $c^2 = 2$. Also,

$$x(t) = \kappa\sqrt{2}t^{1/2}t^{ic/2a}(a+ht)^{1/2}(a+ht)^{-ic/2a} \quad \text{if} \quad c^2 < 2, \text{ and}$$

$$x(t) = \kappa\sqrt{2ht}e^{-ic/2ht} \quad \text{if} \quad c^2 = 2,$$

where $a = (2-c^2)^{1/2}$ and κ is a complex constant with $|\kappa| = 1$.

Proof of Theorem 5.2. First note that, since the equations of motion are autonomous, we need consider only solutions $x(t)$ with a singularity at $t^* = 0$. Theorem 3.1 implies that $x(t) \to 0$ as $t \to 0$. Now note that, if $x = \phi(t)$ is a solution of equation (2.1), then $x = \phi(-t)$ is also a solution. Therefore it suffices to show that no solution satisfying $x(t) \to 0$ as $t \to 0+$ has a real analytic branch for $t < 0$.

If γ is irrational, or if $\gamma = p/q$, with q even, then t^γ has no real branch for $t < 0$. The formulas of Lemmas 5.4 and 5.5 then immediately imply that $x(t)$ has no real branch for $t < 0$, if $\beta \neq 1$. If $\beta = 1$, which corresponds to $\gamma = p/q = \frac{1}{2}$, then the formulas of Lemma 5.6 immediately imply that $x(t)$ has no real branch for $t < 0$. In any case, no singular solution is branch regularizable when $\gamma \notin Q_0$.

Proof of Theorem 5.3. By the considerations at the beginning of the proof of Theorem 5.2, it suffices to show that every solution $x(t)$ satisfying $x(t) \to 0$ as $t \to 0+$ has a unique branch extension for $t < 0$. We first show that every such solution has a unique real analytic continuation to $t < 0$. We then show that this continuation is a solution.

Since $\gamma = p/q$, we have that $\beta\gamma = (q-p)/q$ and $(\beta-1)\gamma = (q-2p)/q$. Using Lemmas 5.4 and 5.5, we can then write

$$x = \phi(t) = \kappa[u^p X_1(u^2) + iu^{q-2p}X_2(u^2)], \quad t > 0,$$

where

$$u = t^{1/q},$$

$$X_1(z) = \begin{cases} \Phi(z^{q-p}), & \text{when} \quad \beta < 1, \\ \Psi_1(z^{q-p}, z^{q-2p}), & \text{when} \quad \beta > 1, \end{cases}$$

and

$$X_2(z) = \begin{cases} 0, & \text{when} \quad \beta < 1, \\ \Psi_2(z^{q-p}, z^{q-2p}), & \text{when} \quad \beta > 1. \end{cases}$$

Since q is odd, $t^{1/q}$ has a unique real analytic continuation to $t<0$, namely $-(-t)^{1/q}$. Therefore $\phi(t)$ has a unique real analytic continuation to $t<0$, which can be written

$$\phi_b(t)=\kappa[(-v)^p X_1(v^2)+i(-v)^{q-2p}X_2(v^2)], \quad t<0,$$

where

$$v=(-t)^{1/q}.$$

We have only left to show that $x=\phi_b(t)$ is also a solution of equation (2.1). For $t<0$ we can write

$$\phi_b(t)=\begin{cases}-\phi(-t), & \text{if } p \text{ is odd,}\\ \overline{\phi(-t)}, & \text{if } p \text{ is even.}\end{cases}$$

One can check easily that $x=-\phi(-t)$ and $x=\overline{\phi(-t)}$ are both solutions of equation (2.1) for $t<0$, since $x=\phi(t)$ is a solution for $t>0$. Therefore every singular solution has a unique branch extension, and the proof is complete.

Before proceeding to the proofs of the lemmas, we digress briefly to discuss the regularization provided by the proof of Theorem 5.3. If $\beta<1$, then X_2 is identically zero, and hence we can write

$$\phi_b(t)=\begin{cases}-\phi(-t), & \text{if } p \text{ is odd,}\\ \phi(-t), & \text{if } p \text{ is even.}\end{cases}$$

Since the motion of a singular orbit takes place along a fixed line when $\beta<1$, we can interpret this extension as a transmission if p is odd and a reflection if p is even. Note that the Newtonian gravitational potential corresponds to $\alpha=1$, $\beta=\frac{1}{2}$, and $\gamma=p/q=\frac{2}{3}\in Q_0$. Hence collisions are branch regularizable and the regularization is a reflection.

Four other examples are shown in Figure 5.1. In each case $q=5$, so the real analytic continuation is given by a rotation through an angle of 5π in the complex t-plane. When $\beta=\frac{1}{4}$, p is even, so the extension is a reflection. When $\beta=\frac{2}{3}$, p is odd, so the extension is a transmission. When $\beta=\frac{3}{2}$, we have $\beta\geq 1$, so there are collision orbits with nonzero angular momentum. Such an orbit and its branch extension are shown in the figure.

We note in passing that equation (2.1) is explicitly solvable in terms of circular or elliptic functions for the following values of α: $\frac{1}{2}$, $\frac{2}{3}$, 1, $\frac{4}{3}$, $\frac{3}{2}$, 2, 3, 4, and 6 [13].

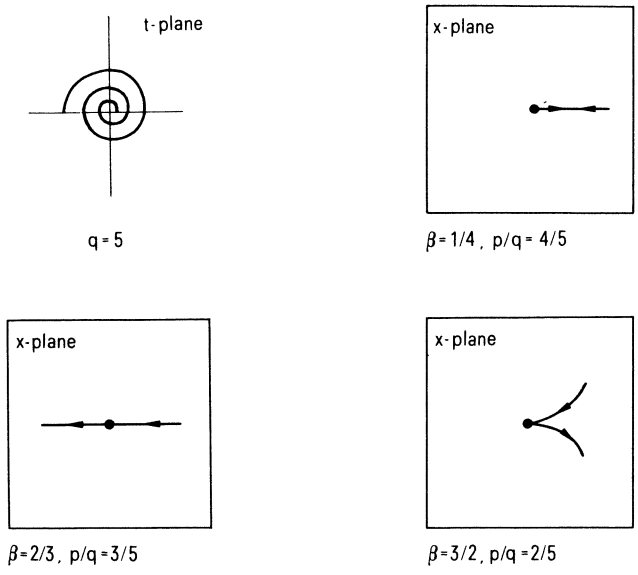


Figure 5.1.

The equation is branch regularizable for all of these values except $\frac{2}{3}$, 2, and 6. Thus there appears to be no relationship between regularizability and explicit solvability.

The remainder of this section is devoted to the proofs of Lemmas 5.4, 5.5, and 5.6. We start by stating and proving two more lemmas. The first could be proved using formal power series expansions and convergence arguments. However, in keeping with the geometric spirit of this paper, we prove it using the stable manifold theorem.

LEMMA 5.7. *Let k, α_1, and α_2 be positive, and let $f: C^2 \to C^1$ be real analytic near $(0, 0)$, with $f(0, 0) = 0$. Assume that the function ϕ satisfies $\phi(x) > 0$ for $x > 0$ and $\phi(x) \to 0$ as $x \to 0+$, and suppose that $y = \phi(x)$ is a solution of*

$$\frac{dy}{dx} = k + f(y^{\alpha_1}, y^{\alpha_2}). \tag{5.2}$$

Then there exists a function $g: C^2 \to C^1$, real analytic near $(0, 0)$, with $g(0, 0) = 0$, and such that

$$\phi(x) = x(k + g(x^{\alpha_1}, x^{\alpha_2})).$$

Proof. We consider the differential equation (5.2) as a vector field in the plane by introducing a time variable t which satisfies $x\,dt = dx$ and writing

$$\begin{aligned} \dot{x} &= x, \\ \dot{y} &= x(k + f(y^{\alpha_1}, y^{\alpha_2})), \end{aligned} \tag{5.3}$$

where the dot denotes differentiation with respect to t. This system is a real analytic vector field on

$$\mathbf{Q} \equiv \{(x, y) \in R^2 : x > 0 \quad \text{and} \quad y > 0\}.$$

We now introduce the new variables

$$\xi_1 = x^{\alpha_1}, \qquad \xi_2 = x^{\alpha_2}, \quad \text{and} \quad \eta = y/x. \tag{5.4}$$

Writing system (5.3) in these variables we obtain

$$\begin{aligned} \dot{\xi}_1 &= \alpha_1 \xi_1, \\ \dot{\xi}_2 &= \alpha_2 \xi_2, \\ \dot{\eta} &= -\eta + k + f(\xi_1 \eta^{\alpha_1}, \xi_2 \eta^{\alpha_2}). \end{aligned} \tag{5.5}$$

Note that the point $(\xi_1, \xi_2, \eta) = (0, 0, k) \equiv \mathbf{p}$ is a rest point for this system. Since $k > 0$, system (5.5) is real analytic in a neighborhood of $\mathbf{p}$. Applying the real analytic version of the stable manifold theorem [11] to $\mathbf{p}$, we find that the local unstable manifold $\mathbf{W}$ of $\mathbf{p}$ can be written

$$\mathbf{W} = \{(\xi_1, \xi_2, \eta) \in R^3 : \eta = k + g(\xi_1, \xi_2) \text{ for } (\xi_1, \xi_2) \text{ near } (0, 0)\},$$

where $g : C^2 \to C^1$ is real analytic near $(0, 0)$ with $g(0, 0) = 0$.

Now observe that system (5.5) has a two dimensional invariant manifold

$$\mathbf{M} \equiv \{(\xi_1, \xi_2, \eta) \in R^3 : \xi_1 > 0, \xi_2 > 0, \eta > 0, \quad \text{and} \quad \xi_1^{\alpha_2} = \xi_2^{\alpha_1}\},$$

and that equations (5.4) define a real analytic diffeomorphism from $\mathbf{Q}$ to $\mathbf{M}$ which carries the vector field (5.3) to the vector field (5.5). Furthermore, since the function ϕ can be extended to $x = 0$ as a C^1 function with $\phi(0) = 0$ and $\phi'(0) = k$, the graph of ϕ is mapped to the unique orbit in $\mathbf{W} \cap \mathbf{M}$. Hence ϕ can be written

$$\phi(x) = y = x\eta = x(k + g(x^{\alpha_1}, x^{\alpha_2})),$$

which completes the proof.

LEMMA 5.8. *Let* λ, α_1, *and* α_2 *be positive, and let* $f: C^2 \to C^1$ *be real analytic near* $(0, 0)$. *Then the function*

$$\psi(x) \equiv \int_0^x \zeta^{\lambda-1} f(\zeta^{\alpha_1}, \zeta^{\alpha_2})\, d\zeta$$

can be written

$$\psi(x) = x^{\lambda} g(x^{\alpha_1}, x^{\alpha_2}),$$

where $g: C^2 \to C^1$ *is real analytic near* $(0, 0)$.

Proof. Substituting $\zeta = xt$ into the above integral gives us

$$\psi(x) = x^{\lambda} \int_0^1 t^{\lambda-1} f((xt)^{\alpha_1}, (xt)^{\alpha_2})\, dt.$$

The result follows immediately upon writing

$$g(\xi_1, \xi_2) \equiv \int_0^1 t^{\lambda-1} f(\xi_1 t^{\alpha_1}, \xi_2 t^{\alpha_2})\, dt.$$

Proof of Lemma 5.4. Transformation (4.1) gives us a solution $(r, \theta, w, v)(t)$ of system (4.2) such that $r \to 0$ as $t \to 0+$. Lemma 4.3 implies that $v(t) \to +\sqrt{2}$ as $t \to 0+$ and that $w(t) \equiv 0$ and $\theta(t) \equiv \theta^*$. Using equation (4.3) we find that

$$v(t) = (2 + 2hr(t)^{2\beta\gamma})^{1/2}$$

for small positive t. The first of equations (4.2) then implies that

$$\frac{dr}{dt} = (\beta + 1)(2 + 2hr^{2\beta\gamma})^{1/2}$$

for small positive t. Applying Lemma 5.7, we obtain a function R which is real analytic in a neighborhood of 0 with $R(0) = \sqrt{2}(\beta + 1)$ such that

$$r(t) = tR(t^{2\beta\gamma}).$$

Transformation (4.1) then gives us

$$x(t) = \kappa t^{\gamma} \Phi(t^{2\beta\gamma}),$$

where $\kappa = e^{i\theta^*}$ and $\Phi(z) = R(z)^{\gamma}$.

Proof of Lemma 5.5. Transformation (4.1) gives us a solution $(r, \theta, w, v)(t)$ of system (4.2) such that $r \to 0$ as $t \to 0+$. Lemma 4.3 implies that $v(t) \to +\sqrt{2}$, $w(t) \to 0$, and $\theta(t) \to \theta^*$ as $t \to 0+$. Using equations (4.4) and (4.3) we compute that

$$w(t) = cr^{(\beta-1)\gamma}$$

and that

$$v(t) = (2 + 2hr(t)^{2\beta\gamma} - c^2 r(t)^{2(\beta-1)\gamma})^{1/2}$$

for small positive t. The first of equations (4.2) then implies that

$$\frac{dr}{dt} = (\beta+1)(2 + 2hr^{2\beta\gamma} - c^2 r^{2(\beta-1)\gamma})^{1/2}$$

for small positive t. Applying Lemma 5.7, we obtain a function R which is real analytic in a neighborhood of $(0, 0)$ with $R(0, 0) = \sqrt{2}(\beta+1)$ such that

$$r(t) = tR(t^{2\beta\gamma}, t^{2(\beta-1)\gamma}).$$

The second of equations (4.2) combined with equation (4.4) now implies that

$$\frac{d\theta}{dt} = ct^{(\beta-1)\gamma-1} R(t^{2\beta\gamma}, t^{2(\beta-1)\gamma})^{(\beta-1)\gamma-1}.$$

Applying Lemma 5.8 we find that

$$\theta(t) = \theta^* + t^{(\beta-1)\gamma} \Theta(t^{2\beta\gamma}, t^{2(\beta-1)\gamma}),$$

where Θ is real analytic near $(0, 0)$. Transformation (4.1) then gives us that

$$\begin{aligned} x &= r^\gamma e^{i\theta^*}[\cos(\theta - \theta^*) + i \sin(\theta - \theta^*)] \\ &= \kappa t^\gamma [\Psi_1(\xi_1, \xi_2) + i\sqrt{\xi_2}\Psi_2(\xi_1, \xi_2)], \end{aligned} \tag{5.6}$$

where $\xi_1 = t^{2\beta\gamma}$, $\xi_2 = t^{2(\beta-1)\gamma}$,

$$\Psi_1(\xi_1, \xi_2) = R(\xi_1, \xi_2)^\gamma \cos(\sqrt{\xi_2}\Theta(\xi_1, \xi_2)),$$

and

$$\Psi_2(\xi_1, \xi_2) = R(\xi_1, \xi_2)^\gamma \frac{\sin(\sqrt{\xi_2}\Theta(\xi_1, \xi_2))}{\sqrt{\xi_2}}.$$

Since R and Θ are real analytic in a neighborhood of $(0, 0)$, with $R(0, 0) > 0$, and since cos is an even real analytic function, Ψ_1 is real analytic in a neighborhood of $(0, 0)$. Note that $\Psi_1(0, 0) = R(0, 0)^\gamma$. Since sin is an odd real analytic function, Ψ_2 is real analytic in a neighborhood of $(0, 0)$. Writing (5.6) in terms of t finishes the proof.

Proof of Lemma 5.6. Since $\beta = 1$, we have that $\alpha = 2$ and $\gamma = \frac{1}{2}$. Equation (4.4) then becomes

$$w = c,$$

while equation (4.3) becomes

$$v^2 + c^2 - 2 = 2hr, \tag{5.7}$$

from which follows the inequality

$$c^2 - 2 \leqslant 2hr.$$

Since $r \to 0$ as $t \to 0$, we must have that $c^2 \leqslant 2$.

We now wish to find the explicit solutions of system (3.2) or, equivalently, of system (4.2). Rewriting (4.2) with $\beta = 1$, we have

$$\begin{aligned} \dot{r} &= 2v, \\ \dot{\theta} &= c/r, \\ \dot{w} &= 0, \\ \dot{v} &= 2h. \end{aligned} \tag{5.8}$$

Using equation (5.7), we see that

$$v^2 \to 2 - c^2 = a^2 \quad \text{as} \quad t \to 0+.$$

The possibility that v would approach a negative value is eliminated by the first of

equations (5.8). Hence we have that

$$v \to a \quad \text{as} \quad t \to 0+.$$

We can now solve system (5.8) explicitly to obtain

$$v(t) = a + 2ht,$$

and

$$r(t) = 2at + 2ht^2, \tag{5.9}$$

Substituting the transformation (4.1) into system (5.8), we obtain the following differential equation for x:

$$\dot{x} = \frac{v + ic}{r} x = \frac{a + ic + 2ht}{2t(a + ht)} x. \tag{5.10}$$

First consider the case $c^2 < 2$, so that $a > 0$. The general solution of equation (5.10) can be written

$$x(t) = Kt^{1/2} t^{ic/2a} (a + ht)^{1/2} (a + ht)^{-ic/2a},$$

where K is a complex constant. Equation (5.9) implies that $|K|^2 = 2$, which gives the desired formula for the case $c^2 < 2$.

Now consider the case $c^2 = 2$, so that $a = 0$. Equation (5.7) becomes

$$v^2 = 2hr.$$

If h were zero, then v would also be zero and r would be constant, which would contradict the hypothesis that $r \to 0$. Thus we must have $h > 0$. Equation (5.10) now becomes

$$\dot{x} = \frac{ic + 2ht}{2ht^2} x,$$

and the general solution can be written

$$x(t) = Kte^{-ic/2ht}.$$

Equation (5.9) now implies that $|K|^2 = 2h$, which gives the desired formula for the case $c^2 = 2$ and completes the proof.

6. An isolating block about collision

Our next goal is to describe the geometric notion of regularization defined by Easton. We proceed by first developing the notation of isolating blocks, following Conley and Easton [2]. We next find an isolating block for our central force problem. We then complete the business of the present section by using equations (4.8) to establish some properties of the block. In Section 7 we give Easton's definition of regularization [4] and use the results of the present section to apply his definition to the central force problem. Finally, in Section 8, we finish our discussion of regularization by describing a transformation similar to that of Levi-Civita [6].

We begin with isolating blocks. Let $\mathbf{M}$ be a smooth manifold and let $\psi: \mathbf{M} \times R^1 \to \mathbf{M}$ be a flow on $\mathbf{M}$. A subset $\mathbf{N} \subset \mathbf{M}$ is called *invariant* if $\psi(\mathbf{N}, R^1) = \mathbf{N}$.

DEFINITION. A compact invariant set $\mathbf{N} \subset \mathbf{M}$ is called *isolated* if there exists an open set $\mathbf{U}$ containing $\mathbf{N}$ such that $\Psi(\mathbf{x}, R^1) \subset \mathbf{U}$ implies $\mathbf{x} \in \mathbf{N}$. The set $\mathbf{U}$ is called an isolating neighborhood for $\mathbf{N}$.

Now let $\mathbf{B}$ be a compact subset of $\mathbf{M}$ with non-empty interior and suppose that $\mathbf{b} = \partial\mathbf{B}$ is a smooth submanifold of $\mathbf{M}$. Define

$$\begin{aligned} \mathbf{b}^+ &\equiv \{\mathbf{x} \in \mathbf{b} : \psi(\mathbf{x}, (-\varepsilon, 0)) \cap \mathbf{B} = \phi \text{ for some } \varepsilon > 0\}, \\ \mathbf{b}^- &\equiv \{\mathbf{x} \in \mathbf{b} : \psi(\mathbf{x}, (0, \varepsilon)) \cap \mathbf{B} = \phi \text{ for some } \varepsilon > 0\}, \\ \mathbf{t} &\equiv \{\mathbf{x} \in \mathbf{b} : \dot{\psi}(\mathbf{x}, 0) \text{ is tangent to } \mathbf{b}\}. \end{aligned} \tag{6.1}$$

DEFINITION. $\mathbf{B}$ is called an *isolating block* if $\mathbf{t} = \mathbf{b}^+ \cap \mathbf{b}^-$.

DEFINITION. Let $\mathbf{N}$ be an isolated invariant set, and let $\mathbf{B}$ be an isolating block. Then $\mathbf{B}$ is said to *isolate* $\mathbf{N}$ if $\text{int}(\mathbf{B})$ is an isolating neighborhood for $\mathbf{N}$.

The following theorem was proved by Conley and Easton [2].

THEOREM 6.1. *If $\mathbf{N}$ is an isolated invariant set, then there exists an isolating block which isolates $\mathbf{N}$. If $\mathbf{B}$ is an isolating block, then there exists an isolated invariant set (possibly empty) which is isolated by $\mathbf{B}$.*

Since it is convenient in our application to define $\mathbf{B}$ in terms of a real-valued function on $\mathbf{M}$, we introduce some more notation. Let $I: \mathbf{M} \to R^1$ be smooth. Write

$$I^*(\mathbf{x}, t) \equiv I(\psi(\mathbf{x}, t)),$$

and define

$$\dot{I}(x) \equiv \dot{I}^*(\mathbf{x}, 0) \quad \text{and} \quad \ddot{I}(x) \equiv \ddot{I}^*(\mathbf{x}, 0),$$

where $\dot{I}^*$ and $\ddot{I}^*$ as usual denote derivatives with respect to time t. The following lemma is proved by Wilson and Yorke [14]. The symbol "D" denotes derivative.

LEMMA 6.2. *Let* $I: \mathbf{M} \to [0, \infty)$, *and let* $\delta_0 > 0$. *Suppose that* $DI(\mathbf{x}) \neq 0$ *whenever* $0 < I(\mathbf{x}) \leq \delta_0$. *Suppose also that* $\ddot{I}(\mathbf{x}) > 0$ *whenever* $0 < I(\mathbf{x}) \leq \delta_0$ *and* $\dot{I}(\mathbf{x}) = 0$. *Then* $\mathbf{N} \equiv I^{-1}(0)$ *is an isolated invariant set and* $I^{-1}([0, \delta])$ *is an isolating block for* $\mathbf{N}$ *for each* $\delta \in (0, \delta_0]$.

We now define the subsets of $\mathbf{b}$ which are asymptotic to $\mathbf{N}$.

$$\mathbf{a}^+ = \{\mathbf{x} \in \mathbf{b}^+ : \psi(\mathbf{x}, [0, \infty)) \subset \mathbf{B}\},$$

$$\mathbf{a}^- = \{\mathbf{x} \in \mathbf{b}^- : \psi(\mathbf{x}, (-\infty, 0]) \subset \mathbf{B}\}.$$

By definition, if $\mathbf{x} \in \mathbf{b}^+ - \mathbf{a}^+$, then there exists a $t > 0$ such that $\psi(\mathbf{x}, t) \notin \mathbf{B}$. Thus we may define the time spent in the block for a point $\mathbf{x} \in \mathbf{b}^+ - \mathbf{a}^+$ by

$$T(\mathbf{x}) \equiv \inf \{t > 0 : \psi(\mathbf{x}, t) \notin \mathbf{B}\}.$$

Note that $\psi(\mathbf{x}, [0, T(\mathbf{x})]) \in \mathbf{B}$ and that $\psi(\mathbf{x}, T(\mathbf{x})) \in \mathbf{b}^-$. Now define the map across the block

$$\Psi: \mathbf{b}^+ - \mathbf{a}^+ \to \mathbf{b}^- : \mathbf{x} \to \psi(\mathbf{x}, T(\mathbf{x})).$$

The following theorem was also proved by Conley and Easton [2].

THEOREM 6.3. *If* $\mathbf{B}$ *is an isolating block, then* $\Psi: \mathbf{b}^+ - \mathbf{a}^+ \to \mathbf{b}^- - \mathbf{a}^-$ *is a diffeomorphism.*

DEFINITION. An isolating block $\mathbf{B}$ is said to be *trivializable* if Ψ extends uniquely to a differmorphism $\mathbf{b}^+ \to \mathbf{b}^-$.

The following lemma shows that trivializability is actually a property of an isolated invariant set. The proof follows from techniques in Conley's notes [1] and will be omitted here.

LEMMA 6.4. *Suppose that* $\mathbf{N}$ *is an isolated invariant set and that* $\mathbf{B}_1$ *and* $\mathbf{B}_2$ *isolate* $\mathbf{N}$. *Then* $\mathbf{B}_1$ *is trivializable if and only if* $\mathbf{B}_2$ *is trivializable.*

DEFINITION. Let $\mathbf{B}$ isolate $\mathbf{N}$. Then $\mathbf{N}$ is said to be *trivializable* if $\mathbf{B}$ is trivializable.

We now return to the central force problem. For the remainder of this section we take ψ to be the flow on $\mathbf{M}(h)$ determined by equations (4.8). The manifold $\mathbf{M}(h)$ is given by (4.5), while the invariant set $\mathbf{N}$ is given by (4.6). Dots now denote differentiation with respect to τ. Define

$$I:\mathbf{M}(h)\to R^1:(r,\theta,w,v)\to r,$$

$$\mathbf{B}(h,\delta)\equiv\{x\in\mathbf{M}(h):I(\mathbf{x})\leqslant\delta\}.$$

LEMMA 6.5. *Let* $\beta<1$. *Given any* h, *there exists a* $\delta_0>0$ *such that* $\mathbf{B}(h,\delta)$ *is an isolating block for* $\mathbf{N}$ *whenever* $0<\delta\leqslant\delta_0$.

Proof. We use Lemma 6.2. The tangent space to $\mathbf{M}(h)$ at the point $\mathbf{x}=(r,\theta,w,v)$ is given by

$$\{(\dot r,\dot\theta,\dot w,\dot v):-h\alpha\gamma r^{\alpha\gamma-1}\dot r+w\dot w+v\dot v=0\}.$$

Since $DI(\mathbf{x})(\dot r,\dot\theta,\dot w,\dot v)=\dot r$, we have that $DI(\mathbf{x})\neq 0$ if $(w,v)\neq(0,0)$. Equation (4.3) implies that

$$|w^2+v^2-2|=2\,|h|\,\delta^{\alpha\gamma},$$

where $I(\mathbf{x})=\delta$. Therefore, if δ_0 is small enough, then $(w,v)\neq(0,0)$ whenever $\delta\leqslant\delta_0$. Hence $DI(\mathbf{x})\neq 0$ when $0<I(\mathbf{x})\leqslant\delta_0$. Now, using equations (4.8), we see that

$$\dot I=(\beta+1)rv\quad\text{and}\quad\ddot I=(\beta+1)r(w^2+(2\beta+1)v^2-2\beta).$$

If $I(\mathbf{x})=\delta$ and if $\dot I(\mathbf{x})=0$, then $r=\delta$ and $v=0$. Using (4.3), we then have

$$\ddot I=2(\beta+1)\delta(h\delta^{\alpha\gamma}+1-\beta).$$

Since $1-\beta>0$, we may choose δ_0 small enough so that $\ddot I>0$ whenever $0<\delta\leqslant\delta_0$. Finally, note that $\mathbf{N}=I^{-1}(0)$. Hence, by Lemma 6.2, $\mathbf{B}(h,\delta)$ is an isolating block for $\mathbf{N}$ and the proof is complete.

By examining Figures 4.2 and 4.3, one can observe that $\mathbf{N}$ is not isolated if $\beta \geqslant 1$. Theorem 6.1 then tells us that it is impossible to construct an isolating block about $\mathbf{N}$ when $\beta \geqslant 1$.

We now exhibit for the block $\mathbf{B}(h, \delta)$ the various subsets defined above. We fix h and assume that $0 < \delta \leqslant \delta_0$, where δ_0 is given by Lemma 6.5. We abbreviate $\mathbf{B} = \mathbf{B}(h, \delta)$ and $\mathbf{x} = (r, \theta, w, v)$.

$$\mathbf{b} = \{\mathbf{x} \in \mathbf{M}(h) : r = \delta\},$$

$$\mathbf{b}^+ = \{\mathbf{x} \in \mathbf{b} : v \leqslant 0\},$$

$$\mathbf{b}^- = \{\mathbf{x} \in \mathbf{b} : v \geqslant 0\},$$

$$\mathbf{t} = \{\mathbf{x} \in \mathbf{b} : v = 0\},$$

$$\mathbf{a}^+ = \{\mathbf{x} \in \mathbf{b}^+ : w = 0\},$$

$$\mathbf{a}^- = \{\mathbf{x} \in \mathbf{b}^- : w = 0\}.$$

The projections of some of these sets to the (w, v)-plane are shown in Figure 6.1 for the case $h > 0$.

The following theorem gives us the values of β for which $\mathbf{N}$ is trivializable. The proof will be broken into two parts. The first part is proved in Lemma 6.7 below, while the second part is postponed until Section 8. The relevance of this theorem to the question of whether collisions can be regularized will be discussed in the next section.

THEOREM 6.6. *The set* $\mathbf{N}$ *given by equation* (4.6) *is a trivializable isolated invariant set for equations* (4.8) *if and only if* $\beta = 1 - n^{-1}$, *where* n *is a positive integer.*

LEMMA 6.7. *Suppose that* $\mathbf{N}$ *is a trivializable isolated invariant set for equations* (4.8). *Then* $\beta = 1 - n^{-1}$, *where* n *is a positive integer.*

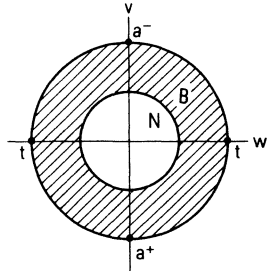


Figure 6.1.

Proof. Since $\mathbf{N}$ is not an isolated invariant set when $\beta \geq 1$, we must have $\beta < 1$. Using the definition of $\mathbf{B}$ and equation (4.3), we write

$$\mathbf{b} = \{(r, \theta, w, v) \in R^4 : r = \delta \text{ and } w^2 + v^2 = 2 + 2h\delta^{\alpha\gamma}\}.$$

Using the integral Λ (4.10) for system (4.9), we can write the map across the block as

$$\Psi : \mathbf{b}^+ - \mathbf{a}^+ \to \mathbf{b}^- - \mathbf{a}^- : (r, \theta, w, v) \to (r, \Psi_\theta(\theta, w), w, -v),$$

where Ψ_θ is the second component of Ψ. Here we are using (θ, w) as coordinates on $\mathbf{b}^+$, so $r = \delta$ and $v = -(2 + 2h\delta^{\alpha\gamma} - w^2)^{1/2}$. Since equations (4.8) are independent of θ,

$$\Psi_\theta(\theta, w) = \theta + \Gamma(w).$$

The function Γ is defined for all w such that $0 < w^2 \leq 2 + 2h\delta^{\alpha\gamma}$. By symmetry,

$$\Gamma(-w) = -\Gamma(w). \tag{6.2}$$

By hypothesis, $\mathbf{B}$ is trivializable. Therefore Ψ extends to a continuous map $\mathbf{b}^+ \to \mathbf{b}^-$. Thus

$$\theta + \Gamma(0+) = \theta + \Gamma(0-) + 2\pi n,$$

where n is an integer. By (6.2), $\Gamma(0-) = -\Gamma(0+)$. Hence we must have

$$\Gamma(0+) = \pi n. \tag{6.3}$$

The number $\Gamma(0+)$ can be computed using geometric methods. Consider a point $\mathbf{p}_0 \in \mathbf{a}^+$. Recall the definition (4.12) of $\mathbf{S}^\pm$. By Lemma 4.3, $\omega(\mathbf{p}_0)$ is a point $\mathbf{s}^-$ in $\mathbf{S}^-$. The orbit through $\mathbf{p}_0$ is the stable manifold of $\mathbf{s}^-$. Now let $\mathbf{p} \in \mathbf{b}^+$ be close to $\mathbf{p}_0$. The orbit through $\mathbf{p}$ follows closely the stable manifold of $\mathbf{s}^-$, passes close to $\mathbf{s}^-$, and then follows closely the unstable manifold of $\mathbf{s}^-$. We therefore must determine the unstable manifold of $\mathbf{s}^-$.

The unstable manifold of $\mathbf{s}^-$ is a subset of $\mathbf{N}$, so we study the flow on $\mathbf{N}$ determined by equations (4.8). On $\mathbf{N}$, $w^2 + v^2 = 2$, so we introduce the angular

variable χ by

$$w + iv = \sqrt{2}e^{i\chi}.$$

Equations (4.8), restricted to $\mathbf{N}$, then become

$$\theta' = w,$$

$$\chi' = (1-\beta)w.$$

Hence, in the (θ, χ) variables, the unstable manifolds of points on $\mathbf{S}^-$ are just straight lines with slope $1-\beta$. (See Figure 6.2.) We are interested in the branch of the unstable manifold of $\mathbf{s}^-$ for which $w \geqslant 0$. Therefore we take $-\pi/2 \leqslant \chi \leqslant \pi/2$. Write

$$\mathbf{s}^- = (\delta, \theta_0, 0, -\sqrt{2}).$$

Then the unstable manifold of the point $\mathbf{s}^-$ is exactly the stable manifold of the point

$$\mathbf{s}^+ \equiv (\delta, \theta_0 + \pi(1-\beta)^{-1}, 0, \sqrt{2}).$$

We now can determine $\Gamma(0+)$. The orbit through $\mathbf{p}$ first follows the stable manifold of $\mathbf{s}^-$, then follows the unstable manifold of $\mathbf{s}^-$, which coincides with the stable manifold of $\mathbf{s}^+$, and finally follows the unstable manifold of $\mathbf{s}^+$. Note that θ does not change along the stable manifold of $\mathbf{s}^-$ or along the unstable manifold of $\mathbf{s}^+$. Therefore, as $\mathbf{p} \to \mathbf{p}_0$, the change in θ along the orbit approaches the difference in θ between $\mathbf{s}^-$ and $\mathbf{s}^+$. This difference is $\pi(1-\beta)^{-1}$. Hence $\Gamma(0+) = \pi(1-\beta)^{-1}$. Combining this result with (6.3), we have $\beta = 1 - n^{-1}$, where n is an integer. Since $0 \leqslant \beta < 1$, n is positive and the proof is complete.

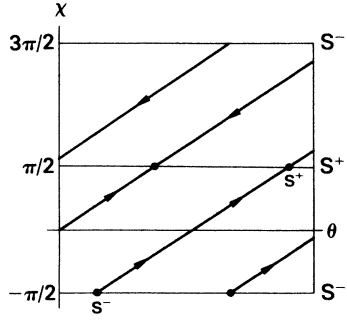


Figure 6.2.

7. Block regularization of the vector field

We now turn to Easton's definition of regularization. We are interested in whether equations (3.2) can be regularized. We consider the equivalent system (4.2). The set $\mathbf{N}$ given by (4.6) is the set of singularities of equations (4.2), i.e. it is the set where the vector field fails to be defined. In the previous section we worked with system (4.8), for which $\mathbf{N}$ is an invariant set. On $\mathbf{M}(h)-\mathbf{N}$ the orbits for the two systems are identical; only the parameterization is different. Later in this section we shall make use of this relationship between the two systems, but first we describe Easton's definition.

Let $\mathbf{M}$ be a smooth manifold, let $\mathbf{N}$ be a compact subset of $\mathbf{M}$, and let $\mathbf{F}$ be a vector field on $\mathbf{M}-\mathbf{N}$. In the previous section $\mathbf{N}$ was an invariant set. In this section, $\mathbf{N}$ is the set of singularities of the vector field $\mathbf{F}$. Let ϕ be the flow on $\mathbf{M}-\mathbf{N}$ given by $\mathbf{F}$. We use the term "flow" loosely here, since we do not require that $\phi(\mathbf{x}, t)$ be defined for all t.

Again let $\mathbf{B}$ be a compact subset of $\mathbf{M}$ with non-empty interior, and suppose that $\mathbf{b}=\partial\mathbf{B}$ is a smooth manifold which does not intersect $\mathbf{N}$. As before, $\mathbf{b}^+$, $\mathbf{b}^-$, and $\mathbf{t}$ are defined by (6.1). The definition of isolating block is also the same. Let $\mathcal{O}(\mathbf{x})$ denote the orbit through $\mathbf{x}$, i.e.

$$\mathcal{O}(\mathbf{x})=\{\phi(\mathbf{x}, t): \phi(\mathbf{x}, t) \text{ is defined}\}.$$

DEFINITION. An isolating block $\mathbf{B}$ is said to isolate the singularity set $\mathbf{N}$ if $\mathbf{N}\subset \operatorname{int}(\mathbf{B})$ and if $\mathcal{O}(\mathbf{x})\not\subset\mathbf{B}$ for all $\mathbf{x}\in\mathbf{B}-\mathbf{N}$.

The subsets $\mathbf{a}^+$ and $\mathbf{a}^-$ are the same as before, except that now we must allow for solutions which are not defined for all t. Thus

$$\mathbf{a}^+=\{\mathbf{x}\in\mathbf{b}^+: \phi(\mathbf{x}, t)\in\mathbf{B} \text{ for all } t\geqslant 0 \text{ for which } \phi(\mathbf{x}, t) \text{ is defined}\},$$

$$\mathbf{a}^-=\{\mathbf{x}\in\mathbf{b}^-: \phi(\mathbf{x}, t)\in\mathbf{B} \text{ for all } t\leqslant 0 \text{ for which } \phi(\mathbf{x}, t) \text{ is defined}\}.$$

The map $\Phi:\mathbf{b}^+-\mathbf{a}^+\to\mathbf{b}^-$ is defined in exactly the same way as the map Ψ in Section 6. Again Theorem 6.3 holds [4], and again we have the same definition of a trivializable block $\mathbf{B}$.

DEFINITION. The singularity set $\mathbf{N}$ is said to be *block regularizable* if there exists a trivializable block $\mathbf{B}$ which isolates $\mathbf{N}$.

Easton gives a general procedure, which he calls "regularization by surgery," whereby one can replace the given vector field by a vector field without singularities [4]. For our purposes we give the following interpretation of the above

definition. Suppose that one can isolate the singularities of a given system with an isolating block, and suppose that the map across the block can be extended. Solutions passing close to the singularities then will determine uniquely an extension for a solution ending in a singularity. Thus one can construct an extended flow with the property of differentiability with respect to initial data. If, on the other hand, the map across the block does not extend, then such an extended flow does not exist.

As a side remark we note that it may be useful in some applications to relax the definition of regularization and to require only that the map across the block extends to a homeomorphism. If the differentiability of the flow were of little interest, then such a relaxation would probably be more appropriate.

We now return again to our central force problem. Note that whether a certain set is or is not an isolating block is independent of the parameterization of the flow. The map across the block is also independent of the parameterization. Therefore, $\mathbf{B}(h, \delta)$ is an isolating block for system (4.2) if and only if it is an isolating block for system (4.8), and $\Phi = \Psi$. Hence $\mathbf{B}(h, \delta)$ is trivializable for (4.2) if and only if it is trivializable for (4.8). We then have, as an immediate consequence of Theorem 6.6, our main result about block regularization of system (3.2).

THEOREM 7.1. *The singularity set* $\mathbf{N}$ *for system* (4.3) *is block regularizable if and only if* $\beta = 1 - n^{-1}$, *where* n *is a positive integer.*

The extension provided by this theorem was computed in the proof of Lemma 6.7. There we saw that orbits passing close to collision are deflected through an angle of $\Gamma(0+) = \pi(1-\beta)^{-1}$ in the x-plane. If this angle is a multiple of π, then the particle emerges in the same direction regardless of whether it passes the singularity on the right or on the left. Hence, if $\beta = 1 - n^{-1}$, then the extension of a singular orbit is a rotation through $n\pi$. This extension is a reflection for even n and a transmission for odd n. The proof of Lemma 6.7 also shows that the map across the block extends to a homeomorphism when $\beta = 1 - n^{-1}$. To complete the proof of Theorem 6.6, we have only to prove that this homeomorphism is in fact a diffeomorphism. This proof is given in the next section.

In Figure 7.1 we illustrate four examples of the regularization provided by Theorem 7.1. In case $\beta = \frac{1}{4}$, $\Gamma(0+) = 4\pi/3$, so no unique extension is determined. In case $\beta = \frac{1}{2}$, $\Gamma(0+) = 2\pi$, so the unique extension is a reflection. In case $\beta = \frac{2}{3}$, $\Gamma(0+) = 3\pi$, so the unique extension is a transmission. In case $\beta = \frac{3}{2}$, $\beta \geqslant 1$, so there are no nearby orbits from which to determine an extension.

It is interesting to compare Figure 7.1 with Figure 5.1. In the cases $\beta = \frac{1}{4}$ and $\beta = \frac{3}{2}$, branch regularization determines an extension while block regularization

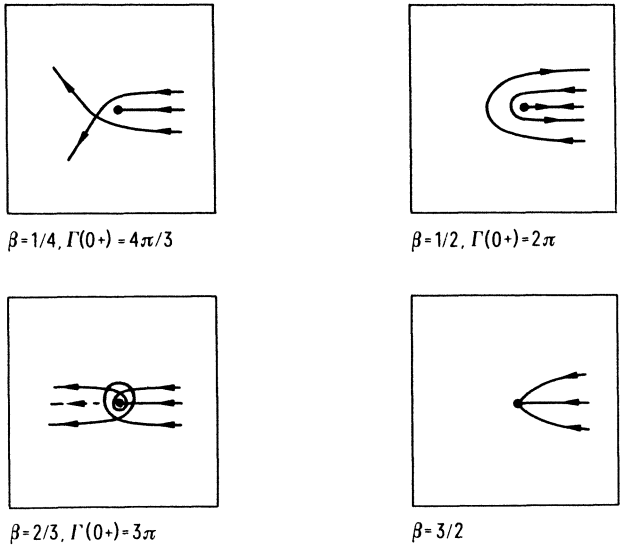


Figure 7.1.

does not. In the cases $\beta = \frac{1}{2}$ and $\beta = \frac{2}{3}$, both regularizations provide the same extensions.

One easily sees that system (3.2) is branch regularizable whenever it is block regularizable and that the two techniques give the same extension. If it is block regularizable, then $\beta = 1 - n^{-1}$, so $p/q = (1+\beta)^{-1} = n/(2n-1)$. Therefore $q = 2n-1$ is odd, and the system is branch regularizable. The extension is a reflection when $p = n$ is even and a transmission when $p = n$ is odd.

In general, block regularization does not imply branch regularization. In a previous paper [7] this author gave an example of a block regularizable system which is not branch regularizable. It seems reasonable to suppose that there are systems which are both branch and block regularizable, but for which the two extensions are different. However, this author does not know such an example.

8. Levi-Civita regularization

We now return to the original Hamiltonian system (3.2) and introduce a change of variables analogous to the Levi-Civita transformation of the Kepler problem [6]. Consider the Hamiltonian function $H(x, y)$ given by equation (3.3) and assume that

$$\alpha = 2\beta = 2(1-n^{-1}),$$

where n is a positive integer. As usual, we fix the total energy $H(x, y) = h$. Define new complex variables z and w by

$$\begin{aligned} x &= z^n \\ y &= w\bar{z}^{1-n}. \end{aligned} \tag{8.1}$$

Note that these equations define a canonical transformation with multiplier n, i.e.

$$\operatorname{Re}(d\bar{x}\, dy) = n \operatorname{Re}(d\bar{z}\, dw).$$

Note also that the transformation is an n-to-1 mapping. Now consider the new Hamiltonian function

$$\begin{aligned} K(z, w) &\equiv |z|^{2(n-1)}(H(z^n, w\bar{z}^{1-n}) - h) \\ &= \tfrac{1}{2}|w|^2 - h\,|z|^{2(n-1)} - 1. \end{aligned}$$

We consider (8.1) as an isoenergetic transformation from the manifold $\{H = h\}$ to the manifold $\{K = 0\}$. The vector field (3.2) on $\mathbf{M}(h)$ transforms to

$$\begin{aligned} \frac{dz}{d\sigma} &= w, \\ \frac{dw}{d\sigma} &= 2(n-1)h\,|z|^{2(n-2)}z, \end{aligned} \tag{8.2}$$

on the manifold $\{K(z, w) = 0\}$. The new time variable σ is given by

$$dt = n\,|z|^{2(n-1)}\, d\sigma.$$

Equations (8.2) extend to $z = 0$ and hence the singularity at collision has been "regularized" in the sense of Levi-Civita. Since $|w|^2 = 2$ when $z = 0$, solutions of (8.2) pass right through $z = 0$. To recover the behavior of system (3.2) we apply transformation (8.1). The case when $n = 3$ ($\alpha = \frac{4}{3}$) is shown in Figure 8.1.

We now state and prove our final lemma. This lemma, together with Lemma 6.7, establishes Theorem 6.6 and hence Theorem 7.1.

LEMMA 8.1. *Let $\beta = 1 - n^{-1}$, where n is a positive integer. Then the invariant set $\mathbf{N}$ for equations* (4.8) *is trivializable.*

Proof. We use transformation (8.1). The block $\mathbf{B}(h, \delta)$ written in the (z, w)

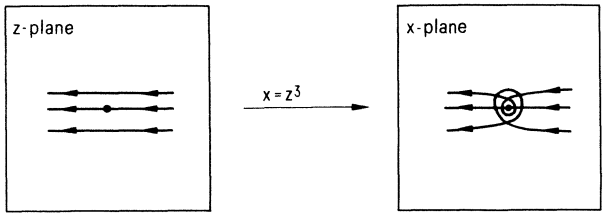


Figure 8.1.

coordinates becomes

$$\mathbf{B}' = \{(z, w) \in C^2 : K(z, w) = 0, |z|^n \leqslant \delta\}.$$

Vector field (8.2) has no invariant set inside $\mathbf{B}'$. Hence the asymptotic sets $\mathbf{a}^+$ and $\mathbf{a}^-$ for $\mathbf{B}'$ are empty. By Theorem 6.3, the map across the block is a diffeomorphism from the entire incoming set to the entire outgoing set. Transforming back to the coordinates of system (4.8), we see that Ψ extends to a diffeomorphism.

Acknowledgements

The author is indebted to Jürgen Moser for many helpful comments and criticisms during the course of this research. Much of the work on this paper was accomplished while the author was supported by the Forschungsinstitut für Mathematik, ETH Zürich. The work also was supported partially by NSF Grant MSC 79–01998.

REFERENCES

[1] CONLEY, C., *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conference Series in Mathematics, 38, AMS, Providence, 1978.
[2] CONLEY, C. and EASTON, R., *Isolated invariant sets and isolating blocks*, *Trans. Amer. Math. Soc. 158* (1971), 35–61.
[3] DEVANEY, R., *Collision orbits in the anisotropic Kepler problem*, Inventiones Math. *45* (1978), 221–251.
[4] EASTON, R., *Regularization of vector fields by surgery*, J. Differential Equations *10* (1971), 92–99.
[5] FENICHEL, N., *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J. *21* (1971), 193–226.
[6] LEVI-CIVITA, T., *Sur la régularisation du problème des trois corps*, Acta Math. *42* (1920), 99–144.
[7] MCGEHEE, R., *Triple collision in Newtonian gravitational systems*, Dynamical Systems, Theory and Applications, J. Moser, ed., Lecture Notes in Physics, Springer-Verlag, 1975, 550–572.
[8] ——, *Singularities in classical celestial mechanics*, Proceedings of the International Congress of Mathematicians, Helsinki, 1978, 827–834.

[9] MOSER, J., *Three integrable Hamiltonian systems connected with isospectral deformation*, Adv. in Math. *16* (1975), 197–220.
[10] SIEGEL, C. L. and MOSER, J., *Lectures on Celestial Mechanics*, Springer-Verlag, 1971.
[11] SIEGEL, C. L., *Der Dreierstoss*, Ann. Math. *42* (1941), 127–168.
[12] SUNDMAN, K., *Recherches sur le problème des trois corps*, Acta Societatis Scientierum Fennicae *34* (1907).
[13] WHITTAKER, E. T., *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Dover, N.Y., 1944.
[14] WILSON, F. W., Jr. and YORKE, J., *Lyapunov functions and isolating blocks*, J. Differential Equations *13* (1973), 106–123.

School of Mathematics
University of Minnesota
Minneapolis, MN 55455 USA

Received June 30, 1981

Comment. Math. Helvetici **56** (1981) 558–580 0010–2571/81/004558–34$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Ueber extremale quasikonforme Abbildungen

RICHARD FEHLMANN

1. Einleitung

Wir betrachten quasikonforme Abbildungen f des Einheitskreises $D = \{z \,|\, |z|<1\}$ auf sich. Diese induzieren durch stetige Fortsetzung einen orientierungserhaltenden Homöomorphismus μ der Kreislinie ∂D auf sich. Betrachten wir die Klasse Q_f (gelegentlich bezeichnen wir sie auch mit Q_μ) aller quasikonformen Abbildungen mit denselben Randwerten wie f, so gibt es eine (oder mehrere) extremale quasikonforme Abbildung f_0 in Q_f, d.h. eine mit kleinstmöglicher maximaler Dilatation. Dies folgt aus der Normalität der Menge aller K-quasikonformen Abbildungen in Q_f für festes K. Im Folgenden bezeichne K_f die maximale Dilatation von f und $K_0 = \inf_{f \in Q_{f_0}} K_f$ diejenige einer Extremalen f_0.

Wir wissen: $f \in Q_{f_0}$ ist genau dann für seine Randwerte extremal, wenn für dessen komplexe Dilatation κ gilt

$$\sup_{\phi \in v} \left| \iint_D \kappa(z)\phi(z)\, dx\, dy \right| = \|\kappa\|_\infty. \tag{1}$$

Dabei ist v die Menge aller holomorphen quadratischen Differentiale ϕ in D mit L_1-Norm

$$\|\phi\| = \iint_D |\phi(z)|\, dx\, dy = 1.$$

Die Notwendigkeit dieser Bedingung wurde von Hamilton [3] gezeigt und die Hinlänglichkeit von Reich und Strebel [10]. Sie hat zur Konsequenz: ist κ_0 komplexe Dilatation einer extremalen Abbildung f_0, so gibt es eine Folge in D holomorpher quadratischer Differentiale ϕ_n mit $\|\phi_n\| = 1$, sodass

$$\left| \iint_D \kappa_0 \phi_n\, dx\, dy \right| \to \|\kappa_0\|_\infty =: k_0, \qquad n \to \infty.$$

Eine solche Folge heisst Hamiltonfolge. Da v normal ist, können wir o.B.d.A. annehmen, eine solche Hamiltonfolge konvergiere lokal gleichmässig in D gegen

eine holomorphe Funktion ϕ_0. Ist $\|\phi_0\| = 1$, so folgt

$$\left| \iint_D \kappa_0 \phi_0 \, dx \, dy \right| = k_0.$$

Ist $0 < k_0 < 1$ und berücksichtigen wir, dass wir ϕ_0 durch $\phi_0 e^{i\theta}$ ersetzen können, so folgt

$$\kappa_0 = k_0 \frac{\bar{\phi}_0}{|\phi_0|} \quad \text{f.ü.}$$

und damit ist f_0 eine Teichmüllersche Abbildung sowie ϕ_0 ihr zugehöriges holomorphes quadratisches Differential. Dieses hat endliche Norm, also ist f_0 sogar eindeutig extremal, d.h. für $f \in Q_{f_0}$, $f \neq f_0$, gilt $K_f > K_{f_0} = K_0$ (vgl. [12]).

Ist $0 \leqq \|\phi_0\| < 1$, so können wir annehmen, dass $\phi_n \to 0$ für $n \to \infty$ lokal gleichmässig in D [14]. Wir sprechen von einer degenerierenden Hamiltonfolge. Es ist also von Interesse, diesen Fall zu untersuchen. Wir werden in Abschnitt 4 zeigen, dass es dann einen wesentlichen Randpunkt gibt, d.h. einen Punkt $\zeta \in \partial D$, sodass die maximale Dilatation K_0 der Extremalen eine untere Schranke für die maximale Dilatation jeder Erweiterung der Randabbildung in nur eine Umgebung von ζ ist. Hieraus ergibt sich die folgende Alternative für extremale quasikonforme Abbildungen. Es handelt sich um eine Teichmüllersche Abbildung mit holomorphem quadratischem Differential endlicher Norm (sie ist also eindeutig extremal) oder eine mit mindestens einem wesentlichen Randpunkt. Diese Aussage ist eine Verschärfung des Rahmenabbildungskriteriums [14].

An dieser Stelle wollen wir das Problem auch allgemeiner stellen. Sei Γ eine abgeschlossene Menge auf ∂D und $f \in Q_\mu$. Es gibt dann auch extremale Abbildungen f_Γ in der grösseren Klasse Q_μ^Γ aller quasikonformen Abbildungen des Einheitskreises auf sich, die auf Γ mit μ übereinstimmen. Dann gilt, sofern Γ mindestens 4 Punkte hat: f_Γ ist in Q_μ^Γ genau dann extremal, wenn für deren komplexe Dilatation κ_Γ

$$\sup_{\phi \in v_\Gamma} \left| \iint_D \kappa_\Gamma \phi \, dx \, dy \right| = \|\kappa_\Gamma\|_\infty \tag{2}$$

ist, wobei v_Γ die Menge der $\phi \in v$ ist, die noch reell längs $\partial D \setminus \Gamma$ sind, d.h. wo $\phi(z)\, dz^2 \in \mathbb{R}$ für $z \in \partial D \setminus \Gamma$ [9].

Sei $K_\Gamma = \inf_{f \in Q_\mu^\Gamma} K_f$. Für eine Extremale f_Γ gilt dann: Sie ist Teichmüllersch mit holomorphem quadratischem Differential ϕ_Γ in v_Γ oder es gibt einen bezüglich Γ wesentlichen Randpunkt ζ, d.h. einen Punkt $\zeta \in \Gamma$, sodass K_Γ eine untere Schranke für die maximale Dilatation jeder quasikonformen Abbildung ist,

welche eine Umgebung U von ζ (bezüglich $\bar{D}$) abbildet und auf $U \cap \Gamma$ mit μ übereinstimmt.

2. Die Dilatation der Randabbildung und die Hauptungleichung

Ist A eine offene Menge in $\mathbb{C}$ und $B \subset \mathbb{C}$, so verstehen wir unter einer K-quasikonformen (K-q.k.) Abbildung $f: A \to B$ eine Injektion, die K-quasikonform in jeder Zusammenhangskomponente von A mit Bildgebieten in B ist. Eine solche Abbildung heisst quasikonform (q.k.), wenn sie K-quasikonform für ein gewisses K ist.

Ist μ die Randabbildung einer q.k. Abbildung des Einheitskreises D auf sich, so verstehen wir unter der Dilatation H von μ die Zahl

$$H := \inf \{K_f \mid f: U(\partial D) \to D, \text{ q.k., } f|_{\partial D} = \mu\}$$

und unter der (lokalen) Dilatation H_ζ von μ in ζ ($\zeta \in \partial D$) die Zahl

$$H_\zeta := \inf \{K_f \mid f: U(\zeta) \to D, \text{ q.k., } f|_{U(\zeta)\cap\partial D} = \mu|_{U(\zeta)\cap\partial D}\}.$$

Dabei werden die Infima über alle bezüglich $\bar{D}$ offenen Umgebungen $U(\partial D)$ von ∂D bezw. $U(\zeta)$ von ζ und über alle derartigen Funktionen f genommen.

Wir bemerken sogleich, dass die Abbildung $\zeta \mapsto H_\zeta$ auf ∂D nach oben halbstetig ist. Ist nämlich $a > H_\zeta$, so gibt es eine a-quasikonforme Erweiterung von μ in eine Umgebung U von ζ in $\bar{D}$. Also ist $H_{\zeta'} \leqq a$ für $\zeta' \in U \cap \partial D$. Deshalb wird das Supremum von $\{H_\zeta \mid \zeta \in \partial D\}$ in einem Punkt $\zeta_0 \in \partial D$ angenommen, und es ist $H_{\zeta_0} = \operatorname{Max}_{\zeta\in\partial D} H_\zeta$.

Die Dilatation H und insbesondere die lokalen Dilatationen H_ζ bilden natürlich untere Schranken für die zur Extremalen gehörige maximale Dilatation K_0. Ein wichtiges Werkzeug, un diese Grössen zu vergleichen, ist die Hauptungleichung ([10], p. 380). Sie besagt das folgende:

Sind f und f_1 zwei quasikonforme Selbstabbildungen des Einheitskreises D in der Klasse Q_μ mit komplexen Dilatationen κ und κ_1, so gilt für alle $\phi \in v$

$$1 \leqq \iint_D |\phi(z)| \frac{\left|1 - \kappa(z)\dfrac{\phi(z)}{|\phi(z)|}\right|^2}{1-|\kappa(z)|^2} \cdot \frac{\left|1 + \mu_1(w)\dfrac{\kappa(z)}{\mu(w)}\dfrac{\phi(z)}{|\phi(z)|}\left[\dfrac{1-\bar{\kappa}(z)\bar{\phi}(z)/|\phi(z)|}{1-\kappa(z)\phi(z)/|\phi(z)|}\right]\right|^2}{1-|\mu_1(w)|^2} dx\,dy. \tag{1}$$

Dabei ist $w=f(z)$ und μ und μ_1 sind die komplexen Dilatationen von f^{-1} und f_1^{-1}.

Daraus ergibt sich die Hinlänglichkeit der Bedingung (1.1) [10] und auch das Rahmenabbildungskriterium [14]: Ist $H<K_0$, so gibt es keine degenerierende Hamiltonfolge. Die Extremale f_0 ist also eindeutig extremal, denn sie ist Teichmüllersch mit quadratischem Differential endlicher Norm.

3. Eine Abschätzung der Dilatation der Randabbildung und ein daraus resultierender Fortsetzungssatz

Sei μ eine Abbildung der Kreislinie ∂D auf sich, die sich quasikonform in den ganzen Einheitskreis fortsetzen lässt. Wir sagen dann kurz, μ sei eine quasisymmetrische (q.s.) Abbildung von ∂D auf sich. Wenden wir nämlich auf Bild und Urbild die Abbildung $-i\cdot\log z$ an und bezeichnen die induzierte Abbildung der reellen Achse auf sich wieder mit μ, so ist μ eine quasisymmetrische Abbildung gemäss der Definition in [2], denn sie lässt sich ja quasikonform in die obere Halbebene erweitern. Ferner gilt $\mu(x+2\pi)=\mu(x)+2\pi$ für alle $x\in\mathbf{R}$.

Es gibt also ein $\rho\geqq 1$ mit

$$Q(\mu, x, t):=\frac{\mu(x+t)-\mu(x)}{\mu(x)-\mu(x-t)}\in\left[\frac{1}{\rho}, \rho\right]$$

für alle $x\in\mathbf{R}$ und $t>0$. (μ heisst dann ρ-quasisymmetrisch.)

Für $E\subset\mathbf{R}$ definieren wir

$$\rho(E):=\inf\left\{\rho\geqq 1 \mid Q(\mu, x, t)\in\left[\frac{1}{\rho}, \rho\right]; x-t, x, x+t\in E\right\},$$

und verstehen unter der lokalen Quasisymmetrie von μ in x die Zahl

$$\rho_x:=\inf\{\rho(I) \mid x\in I,\ I \text{ ein offenes Intervall}\}.$$

Wie die lokale Dilatation ist auch $x\mapsto\rho_x$ nach oben halbstetig. Weil $\rho_x<\infty$ in $x\in\mathbf{R}$, gibt es ein x_0, $0\leqq x_0\leqq 2\pi$, mit $\rho_{x_0}=\mathrm{Max}_{0\leqq x\leqq 2\pi}\rho_x$. Wegen $\mu(x+2\pi)=\mu(x)+2\pi$ ist also $\mathrm{Max}_{x\in\mathbf{R}}\rho_x<\infty$.

Für jeden Homöomorphismus μ der Kreislinie auf sich nennen wir die so erhaltene Abbildung μ kurz die induzierte reelle Abbildung.

LEMMA 3.1. *μ sei ein orientierungserhaltender Homöomorphismus der Kreislinie ∂D auf sich. Für die induzierte reelle Abbildung gelte*

$$\rho_x < \infty \quad \text{für jedes reelle } x.$$

Dann lässt sich μ quasikonform in den Einheitskreis erweitern, und für die Dilatation H von μ gilt:

$$H \leqq \left(\max_{x \in \mathbf{R}} \rho_x\right)^2.$$

Für die lokale Dilatation H_ζ von μ in $\zeta \in \partial D$ gilt:

$$H_\zeta \leqq \rho_x^2.$$

Beweis. Wir haben schon gesehen, dass $\operatorname{Max}_{x \in \mathbf{R}} \rho_x$ endlich ist. Sei nun $A > \operatorname{Max}_{x \in \mathbf{R}} \rho_x$. Dann liegt jedes $x \in \mathbf{R}$ in einem offenen Intervall J, in welchem μ A-quasisymmetrisch ist, d.h. wo $\rho(J) \leqq A$ ist. Die Funktion

$$l_x := \sup \{l > 0 \mid \mu \quad \text{ist } A\text{-q.s. in } (x-l, x+l)\}$$

ist stetig, denn man sieht sofort, dass

$$l_x - |\Delta x| \leqq l_{x+\Delta x} \leqq l_x + |\Delta x|$$

gilt, und sie ist positiv oder konstant unendlich. Wegen $l_{x+2\pi} = l_x$, gibt es ein $l_0 > 0$, sodass μ A-quasisymmetrisch ist in $[x - l_0, x + l_0]$ für jedes $x \in \mathbf{R}$. Wir konstruieren nun die Beurling–Ahlfors-Fortsetzung [2] in der oberen Halbebene. Im horizontalen Streifen $\{x + iy \mid 0 \leqq y \leqq l_0\}$ wird diese A^2-quasikonform, und da sie auch die horizontale Periode 2π hat, erhalten wir eine A^2-q.k. Erweiterung von μ in einen Kreisring $r < |z| \leqq 1$. Da $A > \operatorname{Max}_{x \in \mathbf{R}} \rho_x$ beliebig war, gilt also $H \leqq (\operatorname{Max}_{x \in \mathbf{R}} \rho_x)^2$.

Nach einem Erweiterungssatz von Lehto und Virtanen ([7], p. 100) gibt es nun eine q.k. Abbildung des Einheitskreises, die in einem kleineren Kreisring $\tilde{r} \leqq |z| \leqq 1$, $\tilde{r} > r$, mit der konstruierten Erweiterung zusammenfällt.

Die Abschätzung für die lokale Dilatation erhalten wir ebenso, indem wir $A > \rho_x$ wählen.

Seien $x_1 < x_2 < x_3 < x_4$ reelle Zahlen oder ∞, dann bezeichne $M(x_1, x_2, x_3, x_4)$ den Modul des Vierecks, bestehend aus der oberen Halbebene und den Ecken

x_1, x_2, x_3, x_4. Wir definieren

$$\mathcal{P}(r) := M(-1, 0, r, \infty).$$

Die a- und b-Seiten seien so gewählt, dass $\mathcal{P}$ monoton wächst. $\mathcal{P}$ ist stetig in $[0, \infty]$, wenn wir $\mathcal{P}(0) = 0$ und $\mathcal{P}(\infty) = \infty$ setzen. Es gilt

$$\mathcal{P}(r) = 1 + \Theta(r) \log r \quad \text{für} \quad r \geqq 1,$$

Wobei Θ monoton von $0{,}2284\cdots$ bis $1/\pi$ wächst (vgl. [2]).

Für eine reellwertige Funktion μ in einem Intervall der reellen Achse, welche sich quasikonform in eine Umgebung dieses Intervalls fortsetzen lässt, sind lokale Dilatation H_x und lokale Quasisymmetrie ρ_x in x definiert, und wir können ρ_x durch H_x abschätzen.

LEMMA 3.2. *Sei I ein offenes Intervall und $\mu : I \to \mathbf{R}$ stetig monoton wachsend. Für ein $x_0 \in I$ existiere eine quasikonforme Fortsetzung von μ, die in einer Umgebung von x_0 (bezügl. der oberen Halbebene) mit μ übereinstimmt. Also $H_{x_0} < \infty$. Dann gilt*

$$\rho_{x_0} \leqq \mathcal{P}^{-1}(H_{x_0}).$$

Beweis. Sei $A > H_{x_0}$. Es gibt also eine A-quasikonforme Abbildung f eines Halbkreises $\{z = x + iy \mid |z - x_0| \leqq r, y \geqq 0\}$ in die obere Halbebene mit $f(x) = \mu(x)$ für reelle x.

Wir betrachten Folgen x_n, t_n mit $x_n \to x_0$, $t_n \downarrow 0$ für $n \to \infty$. Auf das Urbild wenden wir die Abbildung $z \mapsto (z - x_n)/t_n$ an und auf das Bild $w \mapsto (w - \mu(x_n))/(\mu(x_n) - \mu(x_n - t_n))$. Weil $x_n - t_n$, x_n, $x_n + t_n$ dabei in -1, 0, 1 übergehen und $\mu(x_n - t_n)$, $\mu(x_n)$ in -1, 0, so gilt für die neue induzierte Abbildung $f_n : f_n(-1) = -1$, $f_n(0) = 0$, $f_n(1) = Q(\mu, x_n, t_n)$, und f_n ist in einem Gebiet A-q.k., das mit $n \to \infty$ gegen die obere Halbebene geht. $\{f_n\}$ ist normal ([7], p. 76), also konvergiert eine Teilfolge gegen eine A-q.k. Abbildung g der oberen Halbebene auf sich mit $g(-1) = -1$, $g(0) = 0$ und $g(\infty) = \infty$. Also ist

$$\mathcal{P}(g(1)) = M(-1, 0, g(1), \infty) \in \left[\frac{1}{A}, A\right].$$

Somit ist $\mathcal{P}^{-1}(1/A) \leqq g(1) \leqq \mathcal{P}^{-1}(A)$, und weil $g(1)$ der Limes einer Teilfolge von $Q(\mu, x_n, t_n)$ ist, kann es keine Folgen $x_n \to x_0$, $t_n \downarrow 0$ geben mit

$Q(\mu, x_n, t_n) \notin [1/B, B]$ für jedes n, wenn $B > \mathcal{P}^{-1}(A)$ ist. Also ist

$$\rho_{x_0} \leqq \mathcal{P}^{-1}(H_{x_0}),$$

denn $A > H_{x_0}$ war beliebig.

Weil bei der konformen Abbildung $-i \log z$ natürlich $H_\zeta = H_x$ gilt für $x = -i \log \zeta$, und weil auf Grund der Monotonie von $\mathcal{P}$ eben $\mathcal{P}^{-1}(\mathrm{Max}_{\zeta \in \partial D} H_\zeta) = \mathrm{Max}_{\zeta \in \partial D} \mathcal{P}^{-1}(H_\zeta)$ ist, erhalten wir aus Lemma 3.1. und 3.2. den Fortsetzungssatz:

SATZ 3.1. *Ist μ ein orientierungserhaltender Homöomorphismus der Kreislinie ∂D auf sich und gilt in jedem Punkt $\zeta \in \partial D$ für die lokale Dilatation von μ in ζ*

$$H_\zeta < \infty,$$

dann lässt sich μ quasikonform in den Einheitskreis erweitern (μ ist also quasisymmetrisch) und für die Dilatation H von μ gilt

$$H \leqq \left(\mathcal{P}^{-1}\left(\operatorname*{Max}_{\zeta \in \partial D} H_\zeta\right)\right)^2,$$

d.h. H ist durch eine Konstante nach oben beschränkt, die nur von $\mathrm{Max}_{\zeta \in \partial D} H_\zeta$ abhängt.

Bemerkung 3.1. Zusammen mit dem schon erwähnten Erweiterungssatz ([7], p. 100) können wir schliessen: Unter den Voraussetzungen des Satzes 3.1. gibt es für jede Zahl $\hat{H} > (\mathcal{P}^{-1}(\mathrm{Max}_{\zeta \in \partial D} H_\zeta))^2$ eine quasikonforme Fortsetzung von μ in den ganzen Einheitskreis, die in einer Umgebung der Kreislinie noch $\hat{H}$-quasikonform ist.

4. Bestimmung der maximalen Dilatation einer extremalen quasikonformen Abbildung im Falle degenerierender Hamiltonfolgen

Die Funktionenklasse v ist enthalten in der grösseren Klasse $L_1(D)$ der komplexwertigen messbaren Funktionen ϕ in D mit

$$\|\phi\| = \iint_D |\phi| \, dx \, dy < \infty.$$

Wir betrachten zunächst degenerierende Folgen (ϕ_m) in $L_1(D)$, also $\phi_m \in L_1(D)$

und $\phi_m \to 0$ lokal gleichmässig in D für $m \to \infty$. Ist $I \subseteqq \partial D$ ein abgeschlossenes oder offenes Intervall mit Endpunkten z_1, z_2, so sei das Argument $\arg z$ in einer Umgebung von I stetig festgelegt, die z_i seien so numeriert, dass $\arg z_1 \leqq \arg z \leqq \arg z_2$ ist für $z \in I$, und S_I bezeichne den Sektor

$$S_I := \{z \in D \mid \arg z_1 \leqslant \arg z \leqq \arg z_2\}.$$

$|I|$ steht für die Bogenlänge von I. Wir definieren

$$\theta(I) := \overline{\lim_{m \to \infty}} \iint_{S_I} |\phi_m| \, dx \, dy.$$

Für festes $\zeta \in \partial D$ definieren wir

$$\theta(\zeta) := \inf \{\theta(I) \mid \zeta \in I,\ I \text{ ein offenes Intervall auf } \partial D\}.$$

Für Intervalle I, J mit $J \subset I$ gilt natürlich $\theta(J) \leqq \theta(I)$. Verlangen wir ferner $\|\phi_m\| \leqq 1$, so gilt $\theta(I) \leqq 1$. Ist I_n das abgeschlossene oder offene Intervall auf ∂D mit Endpunkten $\zeta e^{-i\,1/n}$, $\zeta e^{i\,1/n}$, so überlegt man sich leicht, dass gilt

$$\theta(\zeta) = \lim_{n \to \infty} \theta(I_n).$$

$\zeta \mapsto \theta(\zeta)$ ist somit auf ∂D definiert und hängt von der degenerierenden Folge (ϕ_m) ab. Geht man zu einer Teilfolge (ϕ_{m_k}) über, so wird die entsprechende Funktion θ höchstens kleiner. Wir beweisen nun das

LEMMA 4.1. *Sei (ϕ_m) eine Folge in $L_1(D)$ mit $\|\phi_m\| \leqq 1$, die lokal gleichmässig gegen Null konvergiert. Weiter seien $l > 0$, $\varepsilon > 0$. Dann gibt es eine Unterteilung $\{z_1, \ldots, z_N\}$ von ∂D von einer Feinheit l, sodass für eine geeignet gewählte Teilfolge (ϕ_{m_k}) gilt:*

$$\sum_{i=1}^{N} \theta(z_i) < \varepsilon.$$

Dabei verstehen wir unter einer Feinheit l, dass für die Intervalle I_i, $i = 1, \ldots, N$ mit Endpunkten z_i, z_{i+1} $(z_{N+1} := z_1)$ $|I_i| \leqq l$ für $i \leqq N$ gilt.

Beweis. Wir nehmen an, es existiere ein $\zeta_1 \in \partial D$ mit $\theta(\zeta_1) \geqq \varepsilon / 8\pi \cdot l$. I_n sei das Intervall auf ∂D mit Endpunkten $\zeta_1 e^{-i/n}$, $\zeta_1 e^{i/n}$. Wegen $\theta(I_n) \downarrow \theta(\zeta_1)$ gibt es zu

jedem n ein m_n mit

$$\iint\limits_{S_{I_n}} |\phi_{m_n}|\, dx\, dy \geqq \frac{\varepsilon}{8\pi} \cdot l - \frac{1}{n},$$

also ist (ϕ_{m_n}) eine Teilfolge mit

$$\varliminf_{n\to\infty} \iint\limits_{S_I} |\phi_{m_n}|\, dx\, dy \geqq \frac{\varepsilon}{8\pi} \cdot l. \tag{1}$$

Bei Wahl einer Teilfolge verkleinern sich die $\theta(\zeta)$ höchstens, während der obige Limes inferior höchstens grösser wird, also Ungleichung (1) erhalten bleibt. Gibt es für die neue Folge noch ein ζ_2 mit $\theta(\zeta_2) \geqq \varepsilon/8\pi \cdot l$, wählen wir analog eine weitere Teilfolge. Nach endlich vielen Schritten sind wir fertig, da $\|\phi_m\| \leqq 1$. Wir haben also durch Uebergang zu einer Teilfolge, die wir wieder (ϕ_m) nennen wollen, erreicht, dass $\theta(\zeta) \leqq \varepsilon/8\pi \cdot l$ für $\zeta \in \partial D \setminus \{\zeta_1, \ldots, \zeta_k\}$ (vgl. [8]). Wir wählen nun eine Unterteilung $z_1, \ldots, z_N$ von ∂D, sodass für die Intervalle I_i, $i = 1, \ldots, N$ mit Endpunkten z_i, z_{i+1} $(z_{N+1} := z_1)$ gilt $|I_i| < l$. Wir können $l < \pi$ voraussetzen und für die Anzahl N der Intervalle $2\pi/l \leqq N \leqq 4\pi/l$ erreichen durch $l/2 \leqq |I_i| < l$. Ebenfalls erreichen wir, dass $\{z_i \mid i \leqq N\} \cap \{\zeta_1, \ldots, \zeta_k\} = \emptyset$ ist. Also gilt $\theta(z_i) \leqq \varepsilon/8\pi \cdot l$, $i \leqq N$, und damit

$$\sum_{i=1}^{N} \theta(z_i) \leqq N \frac{\varepsilon}{8\pi} \cdot l \leqq \frac{\varepsilon}{2} < \varepsilon.$$

Wir bemerken, dass wir zu jedem z_i ein offenes Intervall J_i mit $z_i \in J_i$ auf ∂D wählen können, sodass gilt

$$\sum_{i=1}^{N} \theta(J_i) < \varepsilon.$$

Wegen $\sum_{i=1}^{N} \theta(z_i) \leqq \varepsilon/2$ brauchen wir bloss $\theta(J_i) \leqq \theta(z_i) + \varepsilon/2N$ zu verlangen.

KOROLLAR 4.1. *Unter den Voraussetzungen von Lemma* 4.1. (*die Teilfolge, für welche* $\sum_{i=1}^{N} \theta(z_i) < \varepsilon$ *gilt, bezeichnen wir wieder mit* (ϕ_m)), *seien die offenen Intervalle* J_i *o.B.d.A. disjunkt gewählt gemäss obiger Bemerkung. Ist* γ_i *ein Jordanbogen, der in* D *einen Punkt von* J_i *mit einem von* J_{i+1} *verbindet* $(J_{N+1} := J_1)$ *und dabei eine Umgebung* G_i *vom abgeschlossenen Intervall zwischen* J_i *und* J_{i+1} *von* D

abtrennt, und ist $D_\varepsilon := D \setminus \bigcup_{i=1}^{N} \bar{G}_i$, *so gilt*

$$\overline{\lim_{m\to\infty}} \iint\limits_{D_\varepsilon} |\phi_m|\, dx\, dy \leqq \varepsilon.$$

Dies ergibt sich unmittelbar, da wegen $\phi_m \to 0$ lokal gleichmässig in D für $m \to \infty$

$$\iint\limits_{D_\varepsilon \setminus \bigcup_{i=1}^{N} S_{i_1}} |\phi_m|\, dx\, dy \to 0, \qquad m \to \infty$$

gilt, und damit ist

$$\overline{\lim_{m\to\infty}} \iint\limits_{D_\varepsilon} |\phi_m|\, dx\, dy \leqq \sum_{i=1}^{N} \overline{\lim_{m\to\infty}} \iint\limits_{S_{i_1}} |\phi_m|\, dx\, dy = \sum_{i=1}^{N} \theta(J_i) < \varepsilon.$$

Wir betrachten nun wieder eine quasisymmetrische Abbildung μ der Kreislinie ∂D auf sich. Wir nehmen an, es existiere eine degenerierende Hamiltonfolge (ϕ_m), $\phi_m \in v \subset L_1(D)$. Es gilt also $\phi_m \to 0$ lokal gleichmässig in D und

$$\iint\limits_D \phi_m \kappa_0\, dx\, dy \to \|\kappa_0\|_\infty, \qquad m \to \infty.$$

Dabei ist κ_0 die komplexe Dilatation einer extremalen Erweiterung f_0 von μ. Dann gilt [10]

$$\iint\limits_D |\phi_m| \frac{\left|1 - \kappa_0 \dfrac{\phi_m}{|\phi_m|}\right|^2}{1 - |\kappa_0|^2}\, dx\, dy \to \frac{1}{K_0}, \qquad m \to \infty.$$

Wir beweisen nun den

SATZ 4.1. *Sei* μ *ein orientierungserhaltender Homöomorphismus der Kreislinie* ∂D *auf sich, der sich quasikonform in den Einheitskreis fortsetzen lässt. Zu der komplexen Dilatation* κ_0 *einer extremalen Fortsetzung* f_0 *existiere eine degenerierende Hamiltonfolge* (ϕ_m) *in* v. *Also*

$$\iint\limits_D \kappa_0 \phi_m\, dx\, dy \to \|\kappa_0\|_\infty, \qquad m \to \infty.$$

Dann gilt für die maximale Dilatation K_0 dieser Erweiterung

$$K_0 = \operatorname*{Max}_{\zeta \in \partial D} H_\zeta.$$

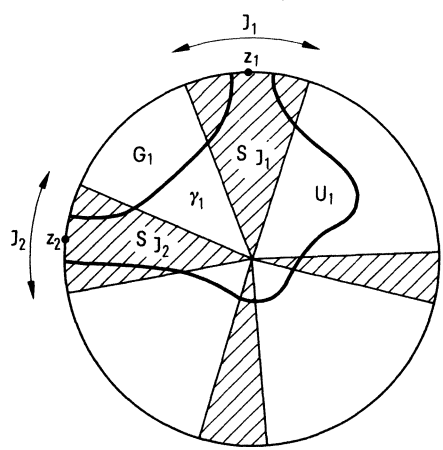
Es gibt also einen wesentlichen Randpunkt $\zeta_0 \in \partial D$ mit $H_{\zeta_0} = K_0$.

Beweis. Sei $H' > \operatorname{Max}_{\zeta \in \partial D} H_\zeta$. Für $\zeta \in \partial D$ definieren wir

$$l_\zeta := \sup \left\{ |I| \;\middle|\; \begin{array}{l} \text{I ein offenes Intervall auf } \partial D \text{ mit Mittelpunkt } \zeta, \text{ und } \mu|_I \text{ lasse} \\ \text{sich } H'\text{-q.k. in eine Umgebung von } I \text{ fortsetzen} \end{array} \right\}$$

Wie früher sehen wir, $\zeta \mapsto l_\zeta$ ist stetig auf ∂D; es gibt also ein $l_0 > 0$ mit $l_0 = \operatorname{Min}_{\zeta \in \partial D} l_\zeta$. Ein beliebiges Intervall auf ∂D der Länge l_0 lässt also eine H'-q.k. Erweiterung von μ in eine Umgebung zu.

Zu (ϕ_m), $l = l_0$ und $\varepsilon > 0$ wählen wir nun nach Lemma 4.1. eine Teilfolge (ϕ_{m_k}) und eine Unterteilung $\{z_1, \ldots, z_N\}$ von ∂D mit $|I_i| \leqq l$, $i = 1, \ldots, N$. Dabei bezeichnen I_i wie vorher die offenen Intervalle mit Endpunkten z_i, z_{i+1}, und, entsprechend unserer Bemerkung am Schluss von Lemma 4.1, seien die offenen Intervalle J_i $(z_i \in J_i)$ gewählt mit $\sum_{i=1}^{N} \theta(J_i) < \varepsilon$. Weil die Länge von I_1 höchstens l ist, gibt es eine H'-q.k. Erweiterung h_1 von μ in eine einfach zusammenhängende Umgebung U_1 von I_1. Darin wählen wir einen analytischen Bogen γ_1, der einen Punkt in $J_1 \cap I_1$ mit einem Punkt in $J_2 \cap I_1$ verbindet und dabei eine Umgebung G_1 vom abgeschlossenen Intervall zwischen J_1 und J_2 von D abtrennt. Wir können zum Beispiel die Umgebung U_1 konform auf einen Halbkreis abbilden,



sodass $\bar{I}_1$ ins Innere des Durchmessers zu liegen kommt und dort einen orthogonalen Halbkreis wählen, der durch die entsprechenden Punkte des Durchmessers läuft. Das durch γ_1 von D abgetrennte Gebiet nennen wir G_1. Durch h_1 wird es auf ein Gebiet im Bildkreis H'-q.k. abgebildet. Dieses Gebiet wiederum wird durch die inverse Abbildung f_0^{-1} der extremalen q.k. Erweiterung f_0 von μ auf ein Gebiet G_1' abgebildet. $f_0^{-1} \circ h_1$ ist auf dem Randintervall $\partial D \cap \partial G_1$ die identische Abbildung. G_1' wird also durch $\partial D \cap \partial G_1$ und den quasikonformen Bogen $f_0^{-1} \circ h_1(\gamma_1)$ berandet, welcher dieselben Endpunkte hat wie γ_1.

Nun betrachten wir eine H'-q.k. Erweiterung h_2 von $\mu|_{I_2}$ in eine Umgebung U_2 von I_2. Das Bild von I_2 unter μ liegt ausserhalb der abgeschlossenen Menge $h_1(\bar{G}_1)$. Wir dürfen also annehmen, $h_2(U_2) \cap h_1(\bar{G}_1) = \emptyset$ und $U_2 \cap \bar{G}_1 = \emptyset$. Dann wählen wir einen analytischen Bogen γ_2, der in U_2 einen Punkt von $J_2 \cap I_2$ mit einem Punkt in $J_3 \cap I_2$ verbindet und das Gebiet G_2 von D abtrennt. Es ist also $\bar{G}_1 \cap \bar{G}_2 = \emptyset$ und $h_1(\bar{G}_1) \cap h_2(\bar{G}_2) = \emptyset$. Wieder setzen wir $G_2' = f_0^{-1} \circ h_2(G_2)$. Diese Konstruktion setzen wir fort, sodass schliesslich sowohl die $\bar{G}_i$ als auch die $h_i(\bar{G}_i)$ paarweise disjunkt sind. Wir erhalten (in Abhängigkeit von $\varepsilon > 0$) Gebiete $G_1, \ldots, G_N$ in D und setzen

$$D_\varepsilon := D \setminus \bigcup_{i=1}^{N} \bar{G}_i, \qquad D_\varepsilon' := D \setminus \bigcup_{i=1}^{N} \bar{G}_i'.$$

Dies sind zwei einfach zusammenhängende Gebiete. Wir wollen nun die durch μ und die h_i auf ∂D_ε induzierte Abbildung quasikonform in D_ε fortsetzen. Die erweiterte Abbildung ist dann quasikonform in D (Hebbarkeit analytischer Bögen ([7], p. 47)) und hat die Randwerte von μ. Sei

$$D_\varepsilon^{\sim} = D \setminus \bigcup_{i=1}^{N} h_i(\bar{G}_i).$$

Die durch μ und die h_i von ∂D_ε auf $\partial D_\varepsilon^{\sim}$ induzierte Abbildung nennen wir μ_ε. μ_ε hat die Eigenschaft, dass sie in eine Umgebung jeden Punktes $\zeta \in \partial D_\varepsilon$ H'-q.k. erweitert werden kann. Entweder ist nämlich $\zeta \in \partial D \setminus \bigcup_{i=1}^{N} \gamma_i$, wegen $H' > \operatorname{Max}_{\zeta \in \partial D} H_\zeta$ ist die Erweiterung also möglich, oder dann ist $\zeta \in \gamma_i$ für ein gewisses i, μ_ε also sowieso die Einschränkung einer H'-q.k. Abbildung.

Nennen wir die lokale Dilatation von μ_ε in ζ nun H_ζ^ε, so gilt demnach

$$H_\zeta^\varepsilon \leqq H' \qquad \forall\, \zeta \in \partial D_\varepsilon.$$

Aus Satz 3.1. folgt: μ_ε lässt sich quasikonform in D_ε erweitern, und die Dilatation H^ε von μ_ε ist also kleiner oder gleich $(\mathcal{P}^{-1}(H'))^2$. Für das folgende sei $\hat{H}$ eine

feste Zahl mit $H>(\mathscr{P}^{-1}(H'))^2$. Wir können μ_ε also $\hat{H}$-q.k. in eine Umgebung von ∂D_ε (bezüglich $\bar{D}_\varepsilon$) erweitern. Gemäss unserer Bemerkung 3.1. gibt es sogar eine q.k. Abbildung f_ε von D_ε in $D_\varepsilon^{\sim}$, welche in einer kompakten Umgebung U von ∂D_ε bezüglich $\bar{D}_\varepsilon$ $\hat{H}$-q.k. ist. Wir setzen f_ε durch die h_i in die G_i fort und nennen die erweiterte Abbildung wieder f_ε. Wir haben also eine q.k. Erweiterung f_ε von μ. Ihre maximale Dilatation in den G_i ist kleiner oder gleich H', in U höchstens $\hat{H}$, und in ganz D nennen wir sie K_ε. $D_\varepsilon \setminus U$ ist in einer kompakten Menge in D enthalten, also gilt für $U'=f_0^{-1}\circ f_\varepsilon(U)$

$$\iint\limits_{D_\varepsilon'\setminus U'} |\phi_m|\,dx\,dy \to 0 \quad \text{für} \quad m\to\infty. \tag{2}$$

Nach der Hauptungleichung gilt, wenn κ_ε die komplexe Dilatation von f_ε^{-1} ist, $w=f_0(z)$ sowie

$$K_\varepsilon(w)=\frac{1+|\kappa_\varepsilon(w)|}{1-|\kappa_\varepsilon(w)|}$$

ist, für jedes m

$$1\leqq\iint\limits_D |\phi_m(z)|\frac{|1-\kappa_0\phi_m/|\phi_m||^2}{1-|\kappa_0|^2}K_\varepsilon(w)\,dx\,dy.$$

Wir schätzen $K_\varepsilon(f_0(z))$ in G_i' und U' ab. Weil die Dilatation von f_ε in $f_\varepsilon^{-1}(w)$ dieselbe ist, wie die von f_ε^{-1} in w und G_i' und U' durch $f_\varepsilon^{-1}\circ f_0$ auf G_i und U abgebildet werden, ist $K_\varepsilon(f_0(z))$ in G_i' kleiner oder gleich H' und in U' kleiner oder gleich $\hat{H}$. Aus der Hauptungleichung folgt also sicher

$$1\leqq\iint\limits_D |\phi_m|\frac{|1-\kappa_0\phi_m/|\phi_m||^2}{1-|\kappa_0|^2}H'\,dx\,dy + \iint\limits_{D_\varepsilon'} |\phi_m|\,K_0\hat{H}\,dx\,dy + \iint\limits_{D_\varepsilon'\setminus U'} |\phi_m|\,K_0K_\varepsilon\,dx\,dy$$

weil

$$\frac{|1-\kappa_0\phi_m/|\phi_m||^2}{1-|\kappa_0|^2}\leqq\frac{(1+|\kappa_0|)^2}{1-|\kappa_0|^2}=\frac{1+|\kappa_0|}{1-|\kappa_0|}\leqq K_0.$$

Lassen wir $m \to \infty$ gehen, so gilt, weil (ϕ_m) eine Hamiltonfolge ist,

$$\iint_D |\phi_m| \frac{|1-\kappa_0 \phi_m/|\phi_m||^2}{1-|\kappa_0|^2}\, dx\, dy \to \frac{1}{K_0},$$

nach Korollar 4.1., angewandt auf G_i' und D_ε', und mit (2) ist

$$1 \leqq \frac{H'}{K_0} + \hat{H} K_0 \varepsilon.$$

Weil $\varepsilon > 0$ beliebig war, ist $K_0 \leqq H'$, und da $H' > \mathrm{Max}_{\zeta \in \partial D} H_\zeta$ beliebig war, also $K_0 \leqq \mathrm{Max}_{\zeta \in \partial D} H_\zeta$.

Damit ist die Behauptung bewiesen.

5. Das analoge Resultat im Fall der Randwertvorgabe auf einer abgeschlossenen Teilmenge von ∂D

Wir geben nun eine Uebersicht, wie man den Satz 4.1. auf den Fall verallgemeinert, wo μ eine quasisymmetrische Abbildung der Kreislinie auf sich ist und die Randwerte durch μ nur auf einer abgeschlossenen Teilmenge Γ von ∂D festgelegt sind. Wir definieren deshalb die Dilatation H_Γ von μ bezüglich Γ durch

$$H_\Gamma := \inf \{K_f \mid f : U(\Gamma) \to D, \text{ q.k., } f|_\Gamma = \mu|_\Gamma\}$$

und die (lokale) Dilatation H_ζ^Γ von μ bezüglich Γ in $\zeta \in \Gamma$ durch

$$H_\zeta^\Gamma := \inf \{K_f \mid f : U(\zeta) \to D, \text{ q.k., } f|_{U(\zeta)\cap\Gamma} = \mu|_{U(\zeta)\cap\Gamma}\},$$

wobei die Infima über alle bezüglich $\bar{D}$ offenen Umgebungen $U(\Gamma)$ von Γ bezw. $U(\zeta)$ von ζ und über alle derartigen Funktionen f genommen werden.

Die Hauptungleichung (2.1) gilt dann für Funktionen f und f_1 in der Klasse Q_μ^Γ für alle $\phi \in v_\Gamma$. Eine Hamiltonfolge für die Klasse Q_μ^Γ ist also eine Folge (ϕ_m) mit $\phi_m \in v_\Gamma$ und

$$\left| \iint_D \kappa_\Gamma \phi_m\, dx\, dy \right| \to \|\kappa_\Gamma\|_\infty \quad \text{für} \quad m \to \infty.$$

Dabei ist κ_Γ die komplexe Dilatation einer extremalen Abbildung f_Γ in Q_μ^Γ. Da

diese ϕ_m durch Spiegelung über $\partial D \setminus \Gamma$ analytisch fortsetzbar sind, gilt für eine degenerierende Folge $\phi_m \to 0$ lokal gleichmässig in $\bar{D} \setminus \Gamma$ für $m \to \infty$ (vgl. [9]).

Der Fortsetzungssatz 3.1 gilt dann in der folgenden Form:

SATZ 5.1. *Sei μ ein orientierungserhaltender Homöomorphismus der Kreislinie ∂D auf sich und Γ eine abgeschlossene Teilmenge von ∂D. In jedem Punkt $\zeta \in \Gamma$ gelte für die lokale Dilatation von μ bezüglich Γ in ζ*

$$H_\zeta^\Gamma < \infty.$$

Dann lässt sich $\mu|_\Gamma$ quasikonform in den Einheitskreis erweitern, und die Dilatation H_Γ der dadurch induzierten Randabbildung bezüglich Γ ist durch eine Konstante nach oben beschränkt, die nur von $\operatorname{Max}_{\zeta \in \Gamma} H_\zeta^\Gamma$ abhängt.

Beweis. Sei $H_1 > \operatorname{Max}_{\zeta \in \Gamma} H_\zeta^\Gamma$. Wir wollen $\mu|_\Gamma$ zunächst auf endlich viele Intervalle, die Γ enthalten, erweitern und dabei die Dilatation durch H_1 beschränkt halten. Γ ist kompakt, also gibt es endlich viele offene Intervalle, die Γ überdecken und die Eigenschaft haben, dass sich $\mu|_\Gamma$ in jedem dieser Intervalle H_1-q.k. in eine Umgebung fortsetzen lassen. Durch Trennung überlappender Intervalle[1] erhalten wir endlich viele disjunkte offene Intervalle I_k, $k = 1, \ldots, n$, die Γ überdecken, sowie eine Erweiterung μ_1 von $\mu|_\Gamma$ auf $\bigcup_{k=1}^n I_k$, deren lokale Dilatation H_ζ in jedem Punkt dieser Vereinigung höchstens H_1 ist. Weil Γ eine abgeschlossene Teilmenge von $\bigcup_{k=1}^n I_k$ ist, finden wir abgeschlossene Intervalle $\bar{J}_k \subset I_k$ mit $\Gamma \subset \bigcup_{k=1}^n \bar{J}_k$. Für jedes $k = 1, \ldots, n$ erweitern wir $\mu_1|_{\bar{J}_k}$ durch Spiegelung über die beiden Enden hinaus. In einem Randpunkt ζ von $\bar{J}_k$ wird dabei die lokale Quasisymmetrie nicht grösser als $1 + \rho_\zeta + \rho_\zeta^2$, wobei ρ_ζ die lokale Quasisymmetrie von $\mu_1|_{I_k}$ in ζ ist ([4], Korollar 2). Nach Lemma 3.2. bleibt somit die lokale Quasisymmetrie in ζ kleiner oder gleich $1 + \mathcal{P}^{-1}(H_\zeta) + \mathcal{P}^{-1}(H_\zeta)^2$. Wegen $H_\zeta \leqq \mathcal{P}^{-1}(H_\zeta)^2$ (Lemma 3.1. und 3.2.) und $H_\zeta \leqq H_1$ ist $(1 + \mathcal{P}^{-1}(H_1) + \mathcal{P}^{-1}(H_1)^2)^2$ eine obere Schranke der lokalen Dilatationen der durch Spiegelung in gewisse offene Intervalle J_k', $\bar{J}_k \subset J_k' \subset I_k$, erweiterten Abbildung μ_1. Entsprechend Lemma 1 in [4] ändern wir diese Erweiterung in $J_k' \setminus \bar{J}_k$ stetig differenzierbar ab. Dabei wird die lokale Quasisymmetrie in den Endpunkten von $\bar{J}_k$ nicht vergrössert, und die Ableitung in $J_k' \setminus \bar{J}_k$ ist überall positiv. Wir erweitern diese Abbildung stetig differenzierbar, streng monoton, ohne die Ableitung Null werden zu lassen zwischen den Intervallen J_k'. In diesen Punkten ist die lokale Dilatation also eins.

Damit haben wir $\mu|_\Gamma$ zu einer Abbildung μ_1 der Kreislinie auf sich erweitert, die die Voraussetzungen von Satz 3.1. erfüllt, und der Satz ist bewiesen.

---

[1] Gehört ein Punkt des Durchschnitts nicht zu Γ, so liegt mit ihm ein ganzes Intervall ausserhalb Γ.

Wir bemerken noch, dass wir, weil $H_1 > \operatorname{Max}_{\zeta\in\Gamma} H_\zeta^\Gamma$ beliebig war, für die Dilatation H von μ_1 die Abschätzung

$$H \leqq \left(1+\mathscr{P}^{-1}\left(\operatorname*{Max}_{\zeta\in\Gamma} H_\zeta^\Gamma\right)+\left(\mathscr{P}^{-1}\left(\operatorname*{Max}_{\zeta\in\Gamma} H_\zeta^\Gamma\right)\right)^2\right)^2$$

erhalten.

Wir verallgemeinern nun den Satz 4.1.

SATZ 5.2. *Sei μ ein orientierungserhaltender Homöomorphismus der Kreislinie ∂D auf sich und Γ eine abgeschlossene Menge in ∂D. Ferner lasse sich $\mu|_\Gamma$ quasikonform ins Innere fortsetzen, d.h. Q_μ^Γ sei nicht leer. Wenn es dann eine degenerierende Hamiltonfolge (ϕ_m) für die Klasse Q_μ^Γ gibt, so gilt für die maximale Dilatation K_Γ einer extremalen Abbildung f_Γ in Q_μ^Γ*

$$K_\Gamma = \operatorname*{Max}_{\zeta\in\Gamma} H_\zeta^\Gamma.$$

Mit Hilfe von Satz 5.1. lässt sich nun der Beweis von Satz 3.1. leicht auf diesen Fall übertragen. Zunächst bemerken wir, dass wir für die Unterteilungspunkte $\{z_1, \ldots, z_N\}$ in Lemma 4.1. verlangen können, dass je ein offenes Intervall $J_k^\sim \ni z_k$ entweder in Γ enthalten ist oder ganz ausserhalb Γ liegt. Dies ist möglich, denn falls kein Intervall $J_k^\sim \ni z_k \in \Gamma$ in Γ enthalten ist, gibt es in beliebiger Nähe von z_k einen Punkt $z_k^\sim$ und mit ihm ein ganzes Intervall aussen an Γ. Dort wählen wir einen neuen Punkt, wir nennen diesen nun z_k, und behalten die Eigenschaften $|I_i| < l$ und $\theta(z_k) \leqq \varepsilon/8\pi \cdot l$ bei. Die offenen Intervalle $J_k \ni z_k$ mit

$$\sum_{i=1}^{N} \theta(J_i) < \varepsilon$$

können wir dann innerhalb Γ oder ganz ausserhalb Γ wählen. Wir wählen $H' > \operatorname{Max}_{\zeta\in\Gamma} H_\zeta^\Gamma$ und definieren

$$l_\zeta := \sup\left\{|I| \;\middle|\; \begin{array}{l} I \text{ ein offenes Intervall auf } \partial D \text{ mit Mittelpunkt } \zeta, \text{ und } \mu|_{I\cap\Gamma} \\ \text{lasse sich } H'\text{-q.k. in eine Umgebung von } I \text{ fortsetzen} \end{array}\right\}$$

Es genügt dann, diejenigen Kreisbögen $z_k z_{k+1}$ durch analytische Bögen γ_k von D abzutrennen, die Punkte von Γ enthalten. Die neu induzierte Randabbildung μ_ε von ∂D_ε auf $\partial D_\varepsilon^\sim$ ist durch die H'-q.k. Erweiterungen auf abgeschlossenen Intervallen auf ∂D_ε vorgegeben, welche die γ_k enthalten. Diese Abbildung μ_ε erfüllt die Voraussetzungen von Satz 5.1, wobei der Menge Γ nun die Vereinigung dieser endlich vielen abgeschlossenen Intervalle entspricht. Wählen wir nun

die Konstante $\hat{H}$ (unabhängig von ε!) hinreichend gross, zeigt dieselbe Berechnung wie früher

$$1 \leqq \frac{H'}{K_\Gamma} + \hat{H} K_\Gamma \varepsilon$$

und damit

$$K_\Gamma = \operatorname*{Max}_{\zeta \in \Gamma} H_\zeta^\Gamma.$$

Wir erwähnen noch, dass damit für das Extremalproblem mit Randwertvorgabe auf einer abgeschlossenen Menge Γ gilt: Es gibt eine eindeutige extremale Abbildung, und sie ist Teichmüllersch mit quadratischem Differential $\phi \in v_\Gamma$, oder es existiert mindestens ein bezüglich Γ wesentlicher Randpunkt $\zeta_0 \in \Gamma$, d.h. ein Punkt ζ_0 mit $K_\Gamma = H_{\zeta_0}^\Gamma$.

6. Anwendungen

a. Konstruktion einer Randabbildung mit überall wesentlichen Randpunkten

Wir konstruieren eine quasisymmetrische Randabbildung μ, für welche die lokale Dilatation H_ζ konstant gleich K_0 ist. M.a.W., jeder Randpunkt ist wesentlich. Dabei soll $K_0 > 1$ sein, da die Problemstellung sonst trivial ist.

Sei f die extremale q.k. Abbildung in einer Homotopieklasse von Abbildungen zwischen zwei kompakten Riemannschen Flächen. Ihre maximale Dilatation sie $K > 1$. f ist eine Teichmüllersche Abbildung mit quadratischem Differential endlicher Norm und komplexer Dilatation

$$\kappa = k \frac{\bar{\phi}}{|\phi|}.$$

Die Riemannschen Flächen R und $R' = f(R)$ haben die Gestalt $R = D|_G$, $R' = D|_{G'}$, wo G und G' diskrete Gruppen von Decktransformationen in $D = \{z \mid |z| < 1\}$ sind. f induziert eine K-q.k. Abbildung $\hat{f}$ des Einheitskreises auf sich. ϕ induziert das quadratische Differential $\hat{\phi}$ auf D, und es ist $\|\hat{\phi}\| = \infty$, weil in jedem der unendlich vielen kompakten Fundamentalgebiete die Norm von $\hat{\phi}$ gleich $\|\phi\|_R$ ist. $\hat{f}$ ist also eine Teichmüllersche Abbildung mit quadratischem Differential unendlicher Norm. $\hat{f}$ induziert eine Randabbildung μ der Kreislinie auf sich, und diese hat eine extremale K_0-q.k. Fortsetzung f_0. Wäre f_0 eine

Teichmüllersche Abbildung mit holomorphem quadratischem Differential endlicher Norm, so wäre $K_0 < K$ wegen der eindeutigen Extremalität. Aus demselben Grunde liesse sich f_0 nach unten durchdrücken [6], was einen Widerspruch zur Extremalität von f zur Folge hätte. Dies sieht man in folgender Weise: $\hat{f}$ induziert einen Isomorphismus θ der Gruppen G und G' durch

$$\theta(g) = \hat{f} \circ g \circ \hat{f}^{-1}, \qquad g \in G.$$

Also gilt

$$\hat{f} = \theta(g) \circ \hat{f} \circ g^{-1}, \qquad \text{in } D, \text{ insbesondere}$$
$$\mu = \theta(g) \circ \mu \circ g^{-1} \qquad \text{für } g \in G.$$

Damit ist aber auch $\theta(g) \circ f_0 \circ g^{-1}$ für jedes $g \in G$ eine Erweiterung von μ, wegen der eindeutigen Extremalität also

$$f_0 = \theta(g) \circ f_0 \circ g^{-1}, \qquad g \in G.$$

Also lässt sich f_0 nach unten durchdrücken.

Wir haben gesehen, μ besitzt als extremale Erweiterung f_0 keine Teichmüllersche Abbildung mit holomorphem quadratischem Differential endlicher Norm, also gibt es einen wesentlichen Randpunkt $\zeta_0 \in \partial D$, d.h.

$$H_{\zeta_0} = K_0.$$

($K_0 > 1$, denn sonst hätten wir wieder eindeutige Extremalität und obigen Widerspruch.)

Nun zeigen wir, dass die Bahn $G(\zeta_0) := \{g(\zeta_0) \mid g \in G\}$ dicht liegt auf ∂D (d.h. G und G' sind Gruppen erster Art). Weil R kompakt ist, ist jedes $g \in G \setminus \{id\}$ hyperbolisch ([1], p. 97). Wäre $G(\zeta_0) = \{\zeta_0\}$, so hätten alle $g \in G$ dieselben zwei Fixpunkte, und das Fundamentalgebiet wäre nicht kompakt. Somit gibt es ein $\zeta_1 \neq \zeta_0$ mit $\zeta_1 \in G(\zeta_0)$. Weil bei kompaktem Fundamentalgebiet die "limit-set" L schon ganz ∂D ist, gilt nach Lehner ([5], p. 18):

Zu jedem $z \in \partial D \setminus \{\zeta_0, \zeta_1\}$ gibt es ein $z' \in \partial D$ mit $z \in \overline{G(w)}$ für jedes $w \in \partial D \setminus \{z'\}$. In unserem Fall also ist $z \in \overline{G(\zeta_0)}$ oder $z \in \overline{G(\zeta_1)}$. Wegen $G(\zeta_0) = G(\zeta_1)$ gilt $z \in \overline{G(\zeta_0)}$ für jedes $z \in \partial D$. Also $\overline{G(\zeta_0)} = \partial D$.

Sei nun $z \in G(\zeta_0)$. Es gibt somit ein $g \in G$ mit $g(\zeta_0) = z$. Wegen $\mu = \theta(g) \circ \mu \circ g^{-1}$ ist $H_z = H_{\zeta_0}$, denn g und $\theta(g)$ sind ja konform. Also ist

$$H_z = K_0 \quad \text{für alle} \quad z \in G(\zeta_0).$$

$z \mapsto H_z$ ist aber halbstetig nach oben und $\overline{G(\zeta_0)} = \partial D$, somit gilt

$$H_z = K_0 > 1 \quad \text{für jedes} \quad z \in \partial D.$$

b. Vergleich der Grössen H_ζ, K_ζ sowie H_Γ, K_Γ, H_Γ^0, K_Γ^0

Wir betrachten zu H, H_ζ, H_Γ noch einige weitere Grössen. Wir definieren für $\zeta \in \partial D$,

$$K_\zeta := \inf \{K_f \mid f : D \to D, \text{ q.k., } f|_I = \mu|_I\},$$

wobei das Infimum über alle offenen Intervalle I auf ∂D genommen wird, die ζ enthalten, sowie über alle derartigen Funktionen f. Offensichtlich ist $H_\zeta \leqq K_\zeta$. Wir werden in diesem Abschnitt die Gleichheit beweisen. Zunächst verallgemeinern wir diese Definition. Sei Γ eine abgeschlossene Teilmenge von ∂D. H_Γ und K_Γ sind also definiert. Wir setzen

$$H_\Gamma^0 := \inf_{\{\Gamma'\}} H_{\Gamma'}, \qquad K_\Gamma^0 := \inf_{\{\Gamma'\}} K_{\Gamma'},$$

wobei das Infimum über alle abgeschlossenen Mengen Γ' in ∂D genommen wird, die Γ im Innern enthalten. Ist zum Beispiel $\Gamma = \{\zeta\}$, so gilt $H_\zeta = H_\Gamma^0$ und $K_\zeta = K_\Gamma^0$; oder für $\Gamma = \partial D$ ist $H_\Gamma^0 = H_\Gamma = H$, $K_\Gamma^0 = K_\Gamma = K_0$. Wir beweisen, dass gilt

$$K_\Gamma^0 = \operatorname{Max} \{H_\Gamma^0, K_\Gamma\}.$$

Daraus erhalten wir für $\Gamma = \{\zeta\}$ wegen $K_{\{\zeta\}} = 1$ sofort

$$K_\zeta = H_\zeta.$$

Im Falle $\Gamma = \{\zeta_1, \zeta_2, \ldots, \zeta_n\}$ ist $H_\Gamma^0 = \operatorname{Max}_{i \leqq n} H_{\zeta_i}$ sowie $K_\Gamma = K_n$ die maximale Dilatation der Teichmüllerschen Extremalen für dieses n-Eck [13]. Also gilt

$$K_{\{\zeta_1, \ldots, \zeta_n\}}^0 = \operatorname{Max} \{K_n, H_{\zeta_1}, H_{\zeta_2}, \ldots, H_{\zeta_n}\}.$$

Für $n \leqq 3$ ist $K_n = 1$, also

$$K_{\{\zeta_1, \zeta_2, \zeta_3\}}^0 = \operatorname{Max} \{H_{\zeta_1}, H_{\zeta_2}, H_{\zeta_3}\}.$$

Mit Hilfe von Satz 5.1. beweisen wir jetzt den

SATZ 6.1. *Sei μ eine quasisymmetrische Abbildung der Kreislinie ∂D auf sich und Γ eine abgeschlossene Teilmenge von ∂D. Dann gilt*

$$K_\Gamma^0 = \operatorname{Max}\{H_\Gamma^0, K_\Gamma\}.$$

Beweis. Weil $K_\Gamma^0 \geqq H_\Gamma^0$ und $K_\Gamma^0 \geqq K_\Gamma$ ist, folgt die Behauptung zum Beispiel aus $K_\Gamma^0 = H_\Gamma^0$. Wir haben also den Fall $K_\Gamma^0 > H_\Gamma^0$ zu untersuchen. Falls $\Gamma = \partial D$ ist, haben wir $K_\Gamma^0 = K_\Gamma$, also dürfen wir annehmen, Γ sei nicht die ganze Kreislinie ∂D. Ferner sei Γ nicht-leer, da sonst sowieso $K_\Gamma^0 = K_\Gamma = H_\Gamma^0 = 1$ gilt. Sei (ε_n) eine Folge mit $\varepsilon_n \downarrow 0$. Γ_n seien endliche Vereinigungen disjunkter abgeschlossener Intervalle so, dass Γ im Innern von Γ_n und Γ_n in der ε_n-Umgebung von Γ bezüglich ∂D enthalten ist. Es gilt

$$\bigcap_{n \in \mathbf{N}} \Gamma_n = \Gamma.$$

Die Folge Γ_n ist monoton fallend. Zunächst sehen wir, dass $H_{\Gamma_n} \downarrow H_\Gamma^0$ und $K_{\Gamma_n} \downarrow K_\Gamma^0$. Aus $H_\Gamma^0 < K_\Gamma^0$ schliessen wir $H_{\Gamma_n} < K_{\Gamma_n}$ für $n \geqq n_0$. Die Extremalen f_n für die Randwerte μ auf Γ_n sind also Teichmüllersche Abbildungen mit quadratischen Differentialen $\phi_n \in v_{\Gamma_n}$, denn kein Punkt von Γ_n kann bezüglich Γ_n wesentlich sein. Aus der Normalität von $\{\phi_n\}$ schliessen wir auf eine in $\bar{D} \setminus \Gamma$ lokal gleichmässig konvergente Teilfolge, die wir wieder ϕ_n nennen wollen. Also $\phi_n \to \phi_\Gamma$ in $\bar{D} \setminus \Gamma$ für $n \to \infty$. Wir behaupten zunächst, $\|\phi_\Gamma\| \neq 0$. Ist nämlich $\|\phi_\Gamma\| = 0$, d.h. konvergieren die ϕ_n lokal gleichmässig in $\bar{D} \setminus \Gamma$ gegen Null, so schliessen wir wieder, dass $H_\Gamma^0 = K_\Gamma^0$ gilt in folgender Weise:

Sei $H' > H_\Gamma^0$. Es gibt also eine H'-q.k. Fortsetzung h, welche in einer Umgebung $U(\Gamma)$ von Γ definiert ist und in Γ_n die Randwerte von μ hat für $n \geqq n_0$. Wie früher wählen wir wieder analytische Kurven γ_k in $U(\Gamma)$, welche die Endpunkte der Intervalle von Γ_{n_0} verbinden und Gebiete G_k von D abtrennen. Wir setzen $\tilde{D} = D \setminus \bigcup \bar{G}_k$ und haben in Intervallen auf $\partial \tilde{D}$, welche die γ_k enthalten, die Randwerte durch h vorzugeben. Nach Satz 5.1. können wir diese Abbildung $\tilde{K}$-q.k. in $\tilde{D}$ erweitern (für ein gewisses $\tilde{K}$). Wir nennen die erweiterte Addildung $\tilde{f}$, und weil $\tilde{f}$ und f_n auf Γ_n übereinstimmen für alle $n \geqq n_0$, gilt die Hauptungleichung

$$1 \leqq \iint_D |\phi_n| \frac{|1 - \kappa_n \phi_n / |\phi_n||^2}{1 - |\kappa_n|^2} \cdot \frac{1 + |\tilde{\kappa}|}{1 - |\tilde{\kappa}|} \, dx \, dy, \qquad n \geqq n_0.$$

Die Abbildung f_n hat aber komplexe Dilatation

$$\kappa_n = k_n \frac{\bar{\phi}_n}{|\phi_n|}, \qquad K_{\Gamma_n} = \frac{1 + k_n}{1 - k_n},$$

also

$$\frac{|1-\kappa_n\phi_n/|\phi_n||^2}{1-|\kappa_n|^2}=\frac{1}{K_{\Gamma_n}},$$

und damit

$$K_{\Gamma_n}\leqq\iint\limits_D|\phi_n|\frac{1+|\tilde{\kappa}|}{1-|\tilde{\kappa}|}\,dx\,dy,\qquad n\geqq n_0.$$

Lassen wir $n\to\infty$ gehen, so folgt, weil $\phi_n\to 0$ lokal gleichmässig in $\bar{D}\setminus\Gamma$ und $(1+|\tilde{\kappa}|)/(1-|\tilde{\kappa}|)\leqq H'$ in $U(\Gamma)$

$$K_\Gamma^0=\lim_{n\to\infty}K_{\Gamma_n}\leqq H'.$$

$H'>H_\Gamma^0$ war beliebig, also $K_\Gamma^0=H_\Gamma^0$ im Widerspruch zur Voraussetzung.

Wie früher haben wir also $0<\|\phi_\Gamma\|\leqq 1$. Mit $K_\Gamma^0=:(1+k_\Gamma^0)/(1-k_\Gamma^0)$ haben wir

$$\kappa_n=k_n\frac{\bar{\phi}_n}{|\phi_n|}\xrightarrow[n\to\infty]{}k_\Gamma^0\frac{\bar{\phi}_\Gamma}{|\phi_\Gamma|}\quad\text{f.ü.}$$

Die Abbildungen f_n konvergieren also gegen eine Teichmüllersche Abbildung f_Γ mit quadratischem Differential in v_Γ und komplexer Dilatation $k_\Gamma^0\bar{\phi}_\Gamma/|\phi_\Gamma|$. Ausserdem stimmt f_Γ auf Γ mit μ überein, ist also die eindeutig bestimmte Extremale für μ bezüglich Γ. Damit ist $K_\Gamma=K_\Gamma^0$ und die Behauptung damit bewiesen.

Wir bemerken noch, dass keine analoge Beziehung zwischen H_Γ, H_Γ^0 und K_Γ gilt und führen das folgende Beispiel an. Wir strecken ein achsenparalleles Quadrat mit den Ecken ζ_1, ζ_2, ζ_3, ζ_4 durch $F_{K_0}(x+iy)=K_0x+iy$, $K_0>1$. Wir wählen $\Gamma=\{\zeta_1,\zeta_2,\zeta_3,\zeta_4\}$. Also ist $H_\Gamma=1$, da Γ nur aus isolierten Punkten besteht, ferner $1<H_\Gamma^0<K_0=K_\Gamma=K_\Gamma^0$.

c. Beweis einer Vermutung von Sethares

Als weitere Anwendung wollen wir eine in der Dissertation von Sethares [11] aufgestellte Vermutung beweisen. Wir betrachten eine Teichmüllersche Abbildung f, welche in einem Gebiet definiert sei, das den abgeschlossenen Einheitskreis $\bar{D}$ enthält. f hat also lokal die Darstellung $f=\Psi^{-1}\circ F_k\circ\Phi$, wo Φ, Ψ konform sind und $F_K(x+iy)=Kx+iy$ ist, $K>1$, mit Ausnahme in isolierten

Punkten. $\phi := \Phi'^2$ und $\psi := \Psi'^2$ sind eindeutige Funktionen mit Nullstellen und Polen in diesen isolierten Punkten bezw. deren Bildpunkten (vgl. [12]). Ist z_0 eine Nullstelle oder ein Pol von ϕ der Ordnung n, so ist $f(z_0)$ eine Nullstelle bezw. ein Pol von ψ derselben Ordnung. Wir fragen nun, ob f für den induzierten Randhomöomorphismus μ von ∂D auf $f(\partial D)$ extremal ist. Ist ϕ holomorph in D und hat auf ∂D nur Pole der Ordnung $n \leqq 2$. so ist f, wie Sethares gezeigt hat, eindeutig extremal. Existieren auf dem Rand Pole höherer Ordnung, so lässt sich f abändern, ohne die maximale Dilatation zu vergrössern und ohne die Randwerte zu verändern. Auch ist ersichtlich, dass die Extremalität von f aus der Existenz von Polen zweiter Ordnung von ϕ auf ∂D folgt. Die Vermutung lautet dann: Hat ϕ auf ∂D keine Pole zweiter Ordnung, aber mindestens einen Pol höherer Ordnung, so ist f nicht extremal. Wir beweisen nun den

SATZ 6.2. *Sei f eine Teichmüllersche Abbildung in einem Gebiet, das den abgeschlossenen Einheitskreis $\bar{D}$ enthält, ϕ das zugehörige meromorphe quadratische Differential sowie K die maximale Dilatation von f. Dann sind für die quasikonforme Erweiterung von $f|_{\partial D}$ ins Innere die folgenden Fälle möglich:*

(a) *ϕ hat auf ∂D mindestens einen Pol zweiter Ordnung. Dann ist f extremal, und die Pole zweiter Ordnung sind wesentliche Randpunkte.*

In diesem Fall ist f eindeutig extremal genau dann, wenn ϕ in D holomorph ist und auf ∂D höchstens Pole der Ordnung $n \leqq 2$ hat.

(b) *ϕ hat auf ∂D keinen Pol zweiter Ordnung. Dann ist das Maximum der lokalen Dilatationen $\operatorname{Max}_{\zeta \in \partial D} H_\zeta$ echt kleiner als K.*

f ist d.u.n.d. eindeutig extremal, wenn ϕ in D holomorph ist und auf ∂D höchstens Pole erster Ordnung hat. Andernfalls ist f nicht extremal.

Beweis. Man berechnet die lokale Dilatation H_ζ der Randabbildung μ in jedem Punkt $\zeta \in \partial D$. Ist ζ ein Pol zweiter Ordnung von ϕ, so ist $H_\zeta = K$. Diese Stellen sind also wesentliche Randpunkte, und mithin ist f extremal. In regulären Punkten $\zeta \in \partial D$ von ϕ (für die Definitionen im Zusammenhang mit quadratischen Differentialen siehe [15]) berechnet man leicht, dass $H_\zeta = 1$ gilt. Für Nullstellen oder Pole der Ordnung $n \neq 2$ erhält man $H_\zeta < K$. Im Falle (b) gibt es unter der Annahme, dass f extremal ist, keinen wesentlichen Randpunkt, und nach Satz 4.1 muss f damit eine Teichmüllersche Abbildung mit in D holomorphem quadratischem Differential ϕ von endlicher Norm sein. ϕ hat also auf ∂D höchstens Pole erster Ordnung, und f ist mithin eindeutig extremal. Damit ist (b) bewiesen, und (a) ergibt sich in folgender Weise:

Hat ϕ auf ∂D mindestens einen Pol zweiter Ordnung, so ist f nach dem soeben Gesagten extremal. Ist ϕ zudem holomorph in D und hat auf ∂D höchstens Pole zweiter Ordnung, so ist f nach Sethares eindeutig extremal. Besitzt ϕ auf ∂D einen Pol der Ordnung $n \geqq 3$, so können wir, wie Sethares gezeigt hat, f

abändern, indem wir F_K in einem Gebiet der Form $\{z \mid \theta_1 < \arg z < \theta_2\}$ durch die durch die Randwerte induzierte Extremale $\tilde{F} \neq F_K$ ersetzen ($f = \Psi^{-1} \circ F_K \circ \Phi$). Hat ϕ in $z_0 \in D$ einen Pol, so betrachten wir eine Kreislinie γ von z_0 in D, die durch keinen Pol und keine Nullstelle von ϕ geht. Weil das Bild einer analytischen Kurve unter F_K wieder analytisch ist, ist $f(\gamma)$ analytisch. Die Dilatation der von f zwischen γ und $f(\gamma)$ induzierten Randabbildung ist also eins. Im Innern von γ können wir f also durch die dazugehörige Extremale $\tilde{f}$ abändern. Diese ist konform oder eine Teichmüllersche Abbildung mit holomorphem quadratischem Differential, also ist $f \neq \tilde{f}$. f kann also nie eindeutig extremal sein, wenn ϕ einen Pol in D besitzt.

Auf die explizite Berechnung der lokalen Dilatationen H_ζ wollen wir an dieser Stelle verzichten.

LITERATUR

[1] BERS, L., *Quasiconformal mappings and Teichmüller's theorem. Analytic Functions.* Princeton University Press, Princeton, N.J. (1960), 89–119.

[2] BEURLING, A. and AHLFORS, L. V., *The boundary correspondence under quasiconformal mappings.* Acta Math. *96* (1956), 125–142.

[3] HAMILTON, R., *Extremal quasiconformal mappings with prescribed boundary values.* Trans. Amer. Math. Soc. *138* (1969), 399–406.

[4] KELINGOS, J. A., *Boundary correspondence under quasiconformal mappings.* Michigan Math. Journal, *13* (1966), 235–249.

[5] LEHNER, J., *A short course in automorphic functions.* Holt, Rinehart and Winston, New York (1966).

[6] LEHTO, O., *Group isomorphisms induced by quasiconformal mappings. Contributions to Analysis* (1974).

[7] LEHTO, O., und VIRTANEN, K. I., *Quasikonforme Abbildungen.* Springer Verlag, Berlin–Heidelberg–New York (1965).

[8] ORTEL, M., *Quasiconformal mappings, extremal for their boundary values.* Preprint.

[9] REICH, E., *On the relation between local and global properties of boundary values for extremal quasiconformal mappings.* Annals of Math. Studies *79* (1974), 391–407.

[10] REICH, E. and STREBEL, K., *Extremal quasiconformal mappings with given boundary values. Contributions to Analysis.* A Collection of Papers Dedicated to Lipman Bers. Academic Press, New York and London (1974).

[11] SETHARES, G. C., *The extremal property of certain Teichmüller mappings.* Comment. Math. Helv. *45* (1970), 353–362.

[12] STREBEL, K., *Zur Frage der Eindeutigkeit extremaler quasikonformer Abbildungen des Einheitskreises II.* Comment. Math. Helv. *39* (1964), 77–89.

[13] STREBEL, K., *Ueber quadratische Differentiale mit geschlossenen Trajektorien und extremale quasikonforme Abbildungen.* Festband zum 70. Geburtstag von Prof. Rolf Nevanlinna. Springer Verlag, Berlin–Heidelberg–New York (1966), 105–127.

[14] STREBEL, K., *On the existence of extremal Teichmüller mappings.* Journal d'analyse Math. *30* (1976), 464–480.

[15] STREBEL, K., *Quadratic Differentials.* Erscheint demnächst in: Ergebnisse der Mathematik.

Math. Inst. Freiestr. 36 8032 Zurich

Eingegangen den 15. Oktober 1980

Comment. Math. Helvetici **56** (1981) 581–598 0010-2571/81/004581-18$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

The fundamental group and the spectrum of the Laplacian

ROBERT BROOKS*

It has been known for many years that there is a close relationship between the fundamental group of a closed manifold M and the Riemannian geometry of its universal cover $\tilde{M}$. Roughly speaking, a closed manifold picks out a family of metrics on its universal cover, namely, those metrics which are invariant under deck transformations. These metrics are all quasi-isometrically equivalent, because the original manifold is compact. Then "asymptotic properties" of the fundamental group are reflected in asymptotic geometric properties of the universal cover in this quasi-isometry class of metrics.

A well-known example of this is the theorem of Milnor–Švarč [14] which states that the fundamental group $\pi_1(M)$ has exponential growth in the sense of groups if and only if the volume $V(r)$ of a ball of radius r in $\tilde{M}$ grows exponentially in r.

The object of this paper is to add another example of this type. If N is any (possibly non-compact) Riemannian manifold, and Δ denotes the Laplace-Beltrami operator, with sign chosen so that it is a positive operator, we set

$$\lambda_0(N) = \inf \frac{\int_N \Delta f \cdot f}{\int_N f^2},$$

where f runs over all smooth functions on N with compact support. λ_0 has the following interpretation: it is well-known that the action of Δ on smooth functions with compact support extends to an unbounded positive, self-adjoint operator on $L^2(N)$. Then $\lambda_0(N)$ is the greatest lower bound of the spectrum of Δ (see [6]).

From the expression given for $\lambda_0(N)$, it seems that one derivative of the metric appears, since that is true of Δ. However, writing $\int_N \Delta f \cdot f = \int_N \|\text{grad } f\|^2$ shows that λ_0 does not involve any derivatives in the metric, and one sees that the property $\lambda_0(N) = 0$ is unchanged when the metric on N is changed to a quasi-isometrically equivalent one.

* Partially supported by NSF Grant #MCS 7802679.

With this understood, our main result is that the condition $\lambda_0(\tilde{M}) = 0$ depends only on the fundamental group $\pi_1(M)$.

THEOREM 1. *$\lambda_0(\tilde{M}) = 0$ if and only if $\pi_1(M)$ is an amenable group.*

See [10] and §1 below for results concerning amenable groups.

An immediate consequence of this is the following:

THEOREM 2. *Let N be a simply connected Riemannian manifold, and Γ a group of isometries of N such that N/Γ is a compact manifold. Then Γ is amenable if and only if any other group of isometries satisfying the same condition is amenable.*

Recalling that a group containing a free group on two generators as a subgroup is not an amenable group, one might regard Theorem 2 as a generalization of a theorem of Tits [15], valid when N is a linear group.

The question of the existence of complete manifolds N such that $\lambda_0(N) > 0$ has been studied by McKean ([13], see also [19]) who proved that if N is a complete, simply-connected manifold of negative sectional curvature bounded away from 0, then $\lambda_0(N) > 0$. He also produced a lower bound for $\lambda_0(N)$ in terms of the bound on the sectional curvature, something that our topological methods cannot provide. It follows from McKean's theorem and Theorem 1 that the fundamental group of a compact manifold of negative sectional curvature is not amenable – this result is known from several points of view (see [3], [8], [9]).

The novelty of Theorem 1 is, however, that there are no curvature assumptions at all – in particular we find manifolds N which have the property that $\lambda_0(\tilde{N}) > 0$, which are not $K(\pi, 1)$'s or obtained from $K(\pi, 1)$'s in any geometric way. To see this, we observe that any finitely-presented group arises as the fundamental group of a compact manifold of dimension 4. In particular, this is true of the free group on two generators. The resulting manifold cannot be a $K(\mathbf{Z} * \mathbf{Z}, 1)$, since it has non-trivial H^4, and the figure-eight, which is a $K(\mathbf{Z} * \mathbf{Z}, 1)$, does not. Choosing any metric on this manifold N gives a metric on the universal cover $\tilde{N}$ with $\lambda_0(\tilde{N}) > 0$.

Theorem 1 also compares with the theorem of Milnor [14] by observing that a group with subexponential growth is necessarily amenable. This is well-known, and we provide a proof in §1. A consequence of Theorem 1 is then that if $\tilde{M}$ has subexponential growth, then $\lambda_0(\tilde{M}) = 0$. One may ask if this theorem is valid without the assumption that $\tilde{M}$ arises as the universal cover of a compact manifold. In fact, this result is valid for complete manifolds (see [2].) It is reasonable to expect that other theorems from group theory can be translated into Riemannian geometry with the aid of Theorem 1. We do not pursue these questions here.

The outline of this paper is as follows: In §1 we review the notion of an amenable group. In §2 we prove one direction of Theorem 1: under the assumption that $\pi_1(M)$ is amenable, we exhibit functions f_i with

$$\frac{\int f_i \cdot \Delta f_i}{\int f_i^2} \to 0 \quad \text{as} \quad i \to \infty.$$

Section 4 then gives a proof of the other direction of Theorem 1, after a brief discussion of integral currents in §3. In §5 we present some remarks on the heat kernel, connecting Theorem 1 with Kesten's Theorem [12].

This paper came about from several sources. Most immediately it arose from reading the inspiring paper of McKean [13], while searching for a geometric version of Kesten's Theorem [12]. Kesten's theorem was first brought to my attention by the Ph.D. thesis of F. Eisenberg [9], and I benefitted also by many conversations with Joel Cohen. The philosophy suggesting the possibility of Theorem 1 comes directly from [14].

The author would like to acknowledge conversations with William Allard, Jeff Cheeger, Blaine Lawson, and Scott Wolpert, all of whom were very helpful in the development of this work.

§1. Følner's theorem

Let G be a countably generated discrete group. G is said to be amenable if there is a finitely additive left-invariant mean μ on G, i.e., a bounded linear functional

$$\mu : L^\infty(G) \to \mathbf{R}$$

having the properties:

(i) $\inf_{g \in G} (f(g)) \leq \mu(f) \leq \sup_{g \in G} (f(g))$

(ii) for all $g \in G$, $\mu(g \cdot f) = \mu(f)$, where $g \cdot f(x) = f(g^{-1}x)$.

It is clear from the definition that any finite group is amenable: let

$$\mu(f) = \frac{1}{|G|} \sum_{g \in G} f(g).$$

However, in general, μ's constructed for infinite groups tend to be rather transcendental, their existence relying on some variant of the Axiom of Choice.

For applications, it is desirable to have alternative characterizations of this notion. We give one such characterization below, and one in §4.

The first of these is a combinatorial characterization due to Følner [17]. To state it, for any finite subset E of G, we let $\#(E)$ denote the cardinality of E, and $g \cdot E = \{gh : h \in E\}$ for $g \in G$.

THEOREM (Følner). *G is amenable if and only if, for every k in the interval $0 < k < 1$, and arbitrary, finitely many elements $a_1, \ldots, a_n$ of G, there exists a finite subset E of G such that*

$$\#(E \cap a_i \cdot E) \geq k \#(E), \quad \text{for} \quad i = 1, \ldots, n.$$

As a simple application of this theorem, we prove:

PROPOSITION 1. *If G has subexponential growth, then G is amenable.*

Proof. Let $0 < k < 1$ be given, and $a_1, \ldots, a_n$ a finite subset of G. Let E_r be the set of all words of length $\leq r$ in the elements $a_1, \ldots, a_n$ and their inverses.

CLAIM. There exists an r such that for all i, $\#(E_r \cap a_i \cdot E_r) \geq k \cdot \#(E_r)$.

Proof. Indeed, we see that $E_{r-1} \subseteq E_r \cap a_i \cdot E_r$ for all i. If the claim is false, then $\#(E_{r-1}) < k \#(E_r)$, for all r; this implies that $\#(E_r) \geq (\text{const})(1/k)r$, and then G has exponential growth.

The proposition follows by setting $E = E_r$ for this r.

Remarks. This proposition is well-known. A proof along different lines may be found in [11], and a proof similar to the above for $G = \mathbf{Z}$ is given in [10].

From this theorem one can easily establish a large class of amenable groups. It follows that all abelian groups are amenable; one sees directly from the definition that subgroups and quotients of amenable groups are amenable, and the extension of an amenable group by an amenable group is amenable.

The best example of a non-amenable group is the free group on two generators – the fact that this is non-amenable follows from the Banach–Tarski–Hausdorff paradox. In particular any group containing a free subgroup on two generators is non-amenable, and this is how non-amenable groups arise geometrically.

We now give a geometric interpretation of Følner's theorem. To state it, let M be a compact manifold, and $\tilde{M}$ its universal cover. We denote by F a fundamental domain in M for the action of $\pi_1(M)$ which arises in the following way: we pick a

smooth triangulation of M, and for each n-simplex in this triangulation, $n = \dim(M)$, we pick one simplex in $\tilde{M}$ covering this simplex. F is the union of all of these simplices – in particular, F is a union of finitely many smooth n-simplices, but F is not necessarily connected.

PROPOSITION 2. *Let M be a compact manifold, and $\tilde{M}$ its universal cover. Then $\pi_1(M)$ is amenable if and only if, for every fundamental domain F as above, and for every ε, there is a finite subset E of $\pi_1(M)$ such that if*

$$H = \bigcup_{g \in E} g\, F$$

then H satisfies the isoperimetric inequality

$$\frac{\text{area}\,(\partial H)}{\text{vol}\,(H)} < \varepsilon.$$

Proof. First we assume that $\pi_1(M)$ is amenable, and pick F and ε as above. Let $a_1, \ldots, a_n$ be the finite set of elements of $\pi_1(M)$ such that $a_i(F) \cap F \neq \emptyset$. Note that $a_1, \ldots, a_n$ generate $\pi_1(M)$. Let $A = \text{area}\,(\partial F)$, $V = \text{vol}\,(F)$. Given k, we choose E a finite set as in Følner's theorem, and consider H as above. Then

$$\text{vol}\,(H) = V \cdot \#(E),$$

and

$$\begin{aligned}\text{area}\,(\partial H) &\leq A \cdot \#\{g \in E : a_i \cdot g \notin E \quad \text{for some} \quad i\} \\ &\leq A \cdot n \cdot (1-k)\#(E),\end{aligned}$$

where the last follows from the choice of E. Then

$$\frac{\text{area}\,(\partial H)}{\text{vol}\,(H)} \leq A \cdot n(1-k)$$

and choosing $k \geq 1 - \varepsilon/(A \cdot n)$ gives the desired inequality

Now assume that $\pi_1(M)$ satisfies the conditions of the proposition, and let $a_1, \ldots, a_n$ and k be given as in Følner's theorem.

We choose a fundamental domain F so that $a_i \cdot F \cap F \neq \emptyset$, $i = 1, \ldots, n$, which is particularly easy in view of the fact that we do not demand that F is connected. We may extend the set $\{a_1, \ldots, a_n\}$ so that it includes all elements of G such that $g \cdot F \cap F \neq \emptyset$.

Now, given ε, we choose H and E as in the conclusion of the proposition. We want to estimate the set $\{g \in E : g \cdot F \cap \partial H \neq \emptyset\}$.

On the one hand, there is a constant l, depending only on F, such that if $g \cdot F \cap \partial H \neq \emptyset$, then area $(g \cdot F \cap \partial H) \geq l$. Thus

$$\#\{g : g \cdot F \cap \partial H \neq \emptyset\} \leq \frac{\text{area } (\partial H)}{l}.$$

On the other hand, this set contains the set $E - a_i \cdot E$ for each i.

Finally, if V denotes the volume of F, then

$$\#(E) = \frac{\text{vol } (H)}{V}.$$

Then

$$\frac{\#(E \cap a_i \cdot E)}{\#(E)} = 1 - \frac{\#\{g : g \cdot F \cap \partial G \neq \emptyset\}}{\#(E)} \geq 1 - \frac{\text{area } (\partial H)}{\text{vol } (H)} \cdot \frac{V}{l}$$

$$\geq 1 - \varepsilon \cdot \frac{V}{l}.$$

Choosing ε small gives us a number $\geq k$, and so $\pi_1(M)$ satisfies the criterion of Følner's Theorem.

§2. Some test functions

In this section, we prove that if $\pi_1(M)$ is amenable, then $\lambda_0(\tilde{M}) = 0$. The idea of the proof is to find test functions f whose support lies in a union H of fundamental domains chosen according to Følner's Theorem, such that Δf is concentrated near ∂H.

We now fix for the discussion a fundamental domain F, two numbers $0 < \varepsilon_1 < \varepsilon_2 < 1$, which we will assume fixed but sufficiently small (depending on the shape of F), and $\psi : [0, 1] \to [0, 1]$ a smooth function such that $\psi(x) = 0$ for $x < \varepsilon_1$, $\psi(x) = 1$ for $x > \varepsilon_2$.

Now let $\varepsilon > 0$, and E a finite subset of $\pi_1(M)$ be given as in the conclusion of Proposition 2. Then let χ_H be the characteristic function of $H = \bigcup_{g \in E} g \cdot F$, and let

$$f_E(x) = \psi(\text{dist } (x, H)) \cdot \chi_H.$$

f_E is smooth with support contained in H.

We now estimate

$$\frac{\int_{\tilde{M}} f_E \cdot \Delta f_E}{\int_{\tilde{M}} f_E^2}.$$

On the other hand, there is a C_1, depending only on ψ and F, such that

$$\int_{\tilde{M}} f_E \cdot \Delta f_E \leq C_1 \cdot \text{area}\,(\partial H)$$

since Δf_E has support only in an ε_2-neighborhood of ∂H, and here its values are determined in each fundamental domain $g \cdot F$ by ψ and the finitely many ways the fundamental domains adjoining $g \cdot F$ either occur or do not occur in H.

On the other hand

$$\int_{\tilde{M}} f_E^2 \geq C_2 \cdot \text{vol}\,(H)$$

since in each fundamental domain $g \cdot F$ occurring in H, $f_E^2 \equiv 1$ except within an ε_2-neighborhood of $g \cdot \partial F$. Combining these two gives

$$\frac{\int_{\tilde{M}} f_E \cdot \Delta f_E}{\int_{\tilde{M}} f_E^2} \leq \frac{C_1}{C_2} \cdot \frac{\text{area}\,(\partial H)}{\text{vol}\,(H)},$$

and this last can be made arbitrarily small by Proposition 2. Thus,

$$\lambda_0(\tilde{M}) = \inf \frac{\int_{\tilde{M}} f \cdot \Delta f}{\int_{\tilde{M}} f^2} = 0.$$

§3. A variational problem

In this section, let M denote a bounded, open subset of an n-dimensional Riemannian manifold M', such that M has smooth boundary B. We consider the

problem of minimizing the isoperimetric constant

$$h(M) = \inf_S \frac{\text{area } (\partial S)}{\text{vol } (S)}$$

where S runs over open submanifolds of M with smooth boundary. Our main result is:

THEOREM 3. *In the above situation, there is a non-zero integral current T realizing this minimum, such that the generalized mean curvature of ∂T is bounded in terms of $h(M)$ and the mean curvature of B.*

It is remarked in [4] that $h(M) > 0$.

Let S_i be a sequence of such S such that (area (∂S_i)/vol (S_i)) decreases to $h(M)$. According to [1], there exists positive constants a and b such that if vol $(S_i) \leq a$, then

$$\text{vol } (S_i) \leq b \text{ area } (\partial S_i)^{n/(n-1)},$$

so that

$$\left(\frac{\text{area } (\partial S_i)}{\text{vol } (S_i)}\right)^{n/(n-1)} (\text{vol } (S_i))^{1/(n-1)} \geq \frac{1}{b}.$$

It follows that there is a constant c such that vol $(S_i) > c$.

Since we also have that vol $(S_i) < \text{vol } (M) < \infty$, it follows that there exists d such that $N(S_i) = \text{vol } (S_i) + \text{vol } (\partial S_i) < d$, and we may apply [16, 4.2.17] to conclude the existence of an integral current T such that

$$\liminf_{i \to \infty} \text{vol } (T - S_i) = 0.$$

In particular, vol $(T) \geq c$, and

$$\frac{\text{area } (\partial T)}{\text{vol } (T)} \leq \liminf_{i \to \infty} \frac{\text{area } (\partial S_i)}{\text{vol } (T)} = \liminf_{i \to \infty} \frac{\text{area } (\partial S_i)}{\text{vol } (\partial S_i)} = h(M).$$

Now let ξ be a compactly supported vector-field in M', and let h_t be the flow generated by ξ. We consider

$$\frac{d}{dt} \frac{\text{area } (h_t(\partial T))}{\text{vol } (h_t(T))};$$

it will be convenient to first handle the case where $\xi \mid B$ is tangent to B. In this case, we have that $h_t(T)$ is again an integral current supported in M, so that by the minimizing property of T,

$$\frac{d}{dt}\frac{\text{area}\,(h_t(\partial T))}{\text{vol}\,(h_t(T))} = 0 \qquad (*)$$

Denote by ν the $\|\partial T\|$-measurable unit vector field satisfying

$$\frac{d}{dt}\,\text{vol}\,(h_t(S))\,|_{t=0} = \int \xi \cdot \nu \, d\,\|\partial T\|;$$

ν is the unit "outward normal" to ∂T. If $\mathscr{m}$ denotes the generalized mean curvature of ∂T, then it is standard that we also have

$$\frac{d}{dt}\,\text{area}\,(h_t(\partial T)) = \int \xi \cdot \mathscr{m}\nu \, d\,\|\partial T\|$$

so that $(*)$ becomes

$$0 = \frac{d}{dt}\frac{\text{area}\,(h_t(\partial T))}{\text{vol}\,(h_t(T))} = \frac{1}{\text{vol}\,(h_t(T))}\int \xi \cdot \mathscr{m}\nu \, d\,\|\partial T\| - h(M)\int \xi \cdot \nu \, d\,\|\partial T|$$

$$= \frac{1}{\text{vol}\,(h_t(T))}\int \xi \cdot (\mathscr{m} - h(M))\nu \, d\,\|\partial T\| \qquad (**)$$

and we see that $\mathscr{m} \equiv h(M)$ on $\text{supp}\,(\partial T) - B$.

Now let $x \in \text{supp}\,(\partial T) \cap B$, and let U be a neighborhood about x. Let ρ be the signed distance function from B

$$\rho(y) = \text{dist}\,(y, B) \quad \text{if} \quad t \in U - M$$
$$\rho(y) = -\text{dist}\,(y, B) \quad \text{if} \quad y \in U \cap M.$$

Then $\text{grad}\,(\rho)$ is a unit vector field pointing outward from B. If f is a smooth function with compact support in U, let $\mu(f) = (d/dt)(\text{area}\,(g_t(\partial T))$ where g_t is the flow generated by $f \cdot \text{grad}\,(\rho)$. We wish to estimate $\mu(f)$ in terms of $|f|$ and $|H|$, where H is the mean curvature vector of B, since this will give us the desired bound for the mean curvature of ∂T at x.

To state our main result, we will need the following related quantities:

$$\mu_1(f) = \int_{M-B} f \operatorname{grad}(\rho) \cdot \nu \, d\|\partial T\|$$

$$\mu_2(f) = \int_B f \, d\|\partial T\|$$

$$\mu_3(f) = \int_B fH \cdot \nu \, d\|\partial T\|$$

$$\mu_4(f) = \lim_{h \downarrow 0} \int_{-h < \rho(x) \leq 0} f[1 - |\nu \cdot \operatorname{grad}(\rho)|^2] \, d\|\partial T\|.$$

Our main observation is:

LEMMA. $\mu(f) = h(M) \cdot \mu_1(f) + \mu_3(f) + \mu_4(f)$.

In particular, this shows that the limit defining μ_4 exist.

Proof. Let φ_h be the Lipshitz function given by

$$\varphi_h(x) = 1 \qquad \text{if} \quad x \geq 0$$

$$= 1 + \frac{1}{h} x \quad \text{if} \quad -h \leq x \leq 0$$

$$= 0 \qquad \text{if} \quad x \leq -h.$$

Then

$$\mu(f) = \mu((1 - \psi_h)f) + \mu(\psi_h f) \quad \text{where} \quad \psi_h = \varphi_h \circ \rho.$$

Then $\mu((1-\psi_n)f) = h(M) \int_M (1-\psi_f) \cdot f \operatorname{grad} \rho \cdot \nu \, d\|dT\|$ since $(1-\psi_h)f$ is supported away from B. As $h \to 0$, this becomes $h(M) \cdot \mu_1(f)$. Now $\mu((\psi_h f)) = \int_{M \cup B} \nabla(\psi_h \cdot f \operatorname{grad} \rho) \cdot \pi_T \, d\|\partial T\|$ where π_T denotes orthogonal projection onto the tangent bundle of ∂T, see [18]. On the other hand,

$$\nabla(\psi_h \cdot f \operatorname{grad} \rho) = (\varphi_h' \circ \rho) f \, d\rho \cdot \operatorname{grad} \rho + (\varphi_h \circ \rho) \, df \cdot \operatorname{grad}(\rho) + (\varphi_h \circ \rho) f \nabla \operatorname{grad}(\rho).$$

The first term then gives

$$\int_{M \cup B} (\varphi_h' \circ \rho) f \, d\rho \cdot \operatorname{grad} \rho = \frac{1}{h} \int_{1-h < x < 0} f \cdot [1 - |\nu \cdot \operatorname{grad} \rho|^2] \, d\|\partial T\|$$

The second term tends to 0 as $h \to 0$, since because $\text{grad}(\rho) = \nu \, \|\partial S\|$-almost everywhere on B, the contribution of the boundary is 0, while as $h \to 0$ the contribution of $M - B$ vanishes.

The third term is then easily seen to be $\int_B fH \cdot \text{grad}\, \rho \, d\, \|\partial T\|$ as $h \to 0$, and the lemma is established.

Of these terms, only μ_4 is not estimated in terms of our desired quantities. To handle this, we observe that

$$\frac{d}{dt} \text{vol}\,(g_t(T)) = \mu_1(f) + \mu_2(f).$$

If $f \geq 0$ everywhere, then $-f \cdot \text{grad}\,(\rho)$ points everywhere interior to M, so that by the minimizing property of T,

$$\frac{d}{dt}\left(\frac{\text{area}\,(g_t(\partial T))}{\text{vol}\,(g_t(T))}\right) = \frac{1}{\text{vol}\,(T)}\left([\mu(-f) - h(M)[\mu_1(-f) + \mu_2(-f)]\right) \geq 0$$

and we get

$$\mu_3(-f) + \mu_4(-f) \geq h(M)\mu_2(-f)$$

and hence

$$0 \leq \mu_4(f) \leq h(M)\mu_2(f) - \mu_3(f)$$

when $f \geq 0$. Now we have, for general f,

$$\begin{aligned} |\mu(f)| &= |h(M)\mu_1(f) + \mu_3(f) + \mu_4(f)| \\ &\leq |h(M)\mu_1(f)| + |\mu_3(f)| + \mu_4(f^+) + \mu_4(f^-) \\ &\leq |h(M)\mu_1(f)| + |h(M)\mu_2(|f|)| + 2\,|\mu_3(|f|)| \\ &\leq h(M) \int_{M \cup B} |f|\, \rho \cdot \nu \,|\, d\, \|\partial T\| + 2 \int_B |f| \cdot |H|\, d\, \|\partial T\| \end{aligned}$$

from which we see that the mean curvature of ∂T is bounded in absolute value by $1 + 2\,|H|$, and the theorem is established.

We remark that it seems likely that this bound can be improved. Perhaps a more careful argument would give the mean curvature of ∂T at $\text{supp}\,(\partial T) \cap B$ to be bounded by the mean curvature of B.

§4. Proof of Theorem 1

In this section we complete the proof of Theorem 1 by showing that if $\lambda_0(\tilde{M}) = 0$, then $\pi_1(M)$ is amenable. The idea of the proof is to use the assumption that $\lambda_0(\tilde{M}) = 0$ to find a sequence of functions f_i with compact support in $\tilde{M}$ such that

$$\frac{\int f_i \cdot \Delta f_i}{\int f_i^2} = \varepsilon_i \to 0 \quad \text{as} \quad i \to \infty.$$

From this we conclude from an argument of Cheeger ([4], and see below) the existence of hypersurfaces S_i dividing $\tilde{M}$ into a bounded and an unbounded component, such that the isoperimetric ratio

$$\frac{\text{area}\,(S_i)}{\text{vol}\,(\text{int}\,(S_i))}$$

(where $\text{int}\,(S_i)$ denotes the bounded component of M) goes to 0 as $i \to \infty$.

Then we would like to replace $\text{int}\,(S_i)$ with the fundamental domains which intersect $\text{int}\,(S_i)$ in hopes of applying Proposition 2, and concluding that $\pi_1(M)$ is amenable.

A moment's thought, however, will convince one that this approach does not quite work. Given a hypersurface S_i, one may attach to it a long, thin spike which will increase both the surface area and the enclosed volume by an arbitrarily small amount, but which will wander through a large number of fundamental domains.

To overcome this difficulty, we apply the theorem of §3 to replace the hypersurfaces S_i by integral currents T_i, whose isoperimetric ratio is not greater than that of S_i, but whose mean curvature is bounded independent of i. It will be seen below that this is sufficient to imply that the isoperimetric ratio of the union of fundamental domains meeting $\text{supp}\,(T_i)$ tends to 0 as $i \to \infty$, and so Proposition 2 applies to conclude that $\pi_1(M)$ is amenable.

We begin first with a review of Cheeger's inequality [4] from our present viewpoint.

Let N be a compact manifold with non-empty boundary, and let f be a function satisfying Dirichlet boundary conditions (for us it will suffice to assume that $f \equiv 0$ in a neighborhood of ∂N). For almost all $t > 0$, the level surface S_t is a smooth hypersurface in N, by Sard's Theorem; S_t divides N into two components corresponding to $f(x) > t$ and $f(x) < t$. Let $\text{int}\,(S_t)$ denote the first of these components.

Then Cheeger's inequality says that, for some t, we have the inequality

$$\frac{\int f \cdot \Delta f}{\int f^2} \geq \frac{1}{4}\left(\frac{\text{area}\,(S_t)}{\text{vol}\,(\text{int}\,S_t)}\right)^2.$$

In particular, taking the infimum of both sides, and allowing the hypersurface S to range over all possible hypersurfaces, gives the lower bound $\lambda_0(N) \geq \frac{1}{4}h^2$ for the spectrum of Δ on N with Dirichlet boundary conditions, in terms of the isoperimetric constant

$$h = \inf_S \frac{\text{area}\,(S)}{\text{vol}\,(\text{int}\,(S))}.$$

In the situation at hand, we are given that $\lambda_0(\tilde{M}) = 0$; it follows from the definitions that there are functions f_i with compact support with the property that

$$\frac{\int_{\tilde{M}} f_i \cdot \Delta f_i}{\int_{\tilde{M}} (f_i)^2} = \varepsilon_i \to 0 \quad \text{as} \quad i \to \infty.$$

We now pick a fundamental domain F for the action of $\pi_1(M)$, and let H_i be a finite union of translates of F so that $\text{supp}\,(f_i) \subset H_i$. It follows from Cheeger's inequality that there exist hypersurfaces S_i contained in H_i such that

$$\frac{\text{area}\,(S_i)}{\text{vol}\,(\text{int}\,(S_i))} \to 0 \quad \text{as} \quad i \to \infty.$$

In order to apply the results of §3, we need a uniform way of "smoothing the boundary" of H_i, so that the mean curvature of the boundary is uniformly bounded. This is easily done – cover a neighborhood of the boundary of F by open sets U_i, and for each U_i choose smooth functions g_i which are $\equiv 1$ for $x \in F$, and which are $\equiv 0$ whenever $\text{dist}\,(x, F) \geq \varepsilon_1$, where ε_1 is an arbitrarily small constant. If U_i contains a corner of F, so that U_i meets more than two translates of F, we choose one such g_i for all the possible ways each of these translates may either be included or not included in H_i.

We now cover H_i by translates of the U_i's, choose g_i's as above, and glue them together by a partition of unity to get a smooth function g. For some ε_2, then, $g^{-1}(1-\varepsilon_2)$ will be a smooth hypersurface of $\tilde{M}$, and we set $\tilde{H}_i = H_i \cup g^{-1}((1-\varepsilon_2, 1))$, $\partial\tilde{H}_i = g^{-1}(1-\varepsilon_2)$. The result of §3 now applies to give:

LEMMA. *There are integral currents T_i in $\tilde{M}$ such that*:

(i) $\dfrac{\text{area}\,(\partial T_i)}{\text{vol}\,(T_i)} \to 0 \quad \text{as} \quad i \to \infty$

(ii) *The mean curvature of ∂T_i is bounded in absolute value independent of i.*

We are now ready to apply Proposition 2 to conclude the amenability of $\pi_1(M)$. To that end, let $\varepsilon > 0$ and fundamental domain F be given.

For each T_i, let K_i be the union of translates of F meeting T_i

$$K_i = \bigcup_{g\cdot F \cap T_i \neq \varnothing} g \cdot F$$

PROPOSITION 3. *There is a constant κ, depending only on F, such that*

$$\frac{\text{area}\,(\partial K_i)}{\text{vol}\,(K_i)} \leq \kappa \cdot \frac{\text{area}\,(\partial T_i)}{\text{vol}\,(T_i)}.$$

Given Prop. 3, Proposition 2 implies the theorem, since we may choose i sufficiently large so that

$$\kappa \cdot \frac{\text{area}\,(T_i)}{\text{vol}\,(\text{int}\,(T_i))} < \varepsilon.$$

Proof of Proposition 3. It is clear that $\text{vol}\,(K_i) \geq \text{vol}\,(T_i)$, since $T_i \subseteq K_i$, so it suffices to show that

$$\text{area}\,(\partial K_i) \leq \kappa \cdot \text{area}\,(\partial T_i).$$

To see this, we cover M with finitely many geodesic balls B_δ of radius δ, where 3δ is less than the injectivity radius of M. We lift these balls to a covering B_δ^k of $\tilde{M}$.

For each ball B_δ^k such that $B_\delta^k \cap T_i \neq \varnothing$, let x_k be a point in this intersection. Then the balls $B_{2\delta}(x_k)$ cover T_i.

CLAIM. There is a constant K such that

$$\text{area}\,(B_{2\delta}(x_k) \cap T_i) \geq K.$$

Proof. It follows from the Monotonicity Theorem ([18], Remark 3.15, or [1]), that there is a constant M such that

$$\frac{\text{area}\,(B_r(x))e^{Mr}}{r^{n-1}}$$

is monotone increasing in r, where area $(B_r(x))$ is the volume of a geodesic ball of radius r about x in T_i, and M can be estimated from the mean curvature of T_i and the second fundamental form of some isometric embedding of $\tilde{M}$ into $\mathbf{R}^N$. Furthermore, as $r \to 0$ the limit of this expression is $\geq \alpha(n-1)$, the volume of a unit ball in $\mathbf{R}^{n-1}$, whenever $x \in T_i$.

It follows that area $(B_{2\delta}(x_k)) \cap T^i \geq e^{-M(2\delta)}(2\delta)^{n-1}\alpha(n-1)$, establishing the claim.

It follows from the compactness of M that there is a number L so that at most L disks $B_{2\delta}(x_k)$ meet at any given point of $\tilde{M}$. It similarly follows that there is a number m such that any $B_{2\delta}(x_k)$ lies in at most m fundamental domains.

Setting $A = \text{area}\,(\partial F)$, and letting N denote the number of balls $B_{2\delta}(x_k)$ which meet T_i, we have

$$\text{area}\,(T_i) \geq \frac{N \cdot K}{L} \geq \frac{K}{L}\, m\, \frac{\text{area}\,(\partial K_i)}{A}$$

and setting $\text{const} = \dfrac{AL}{Km}$ establishes the theorem.

§5. The heat equation

We now give an alternate description of amenable groups, due to Kesten [12].

For G a countable discrete group, a probability distribution P on G is a function $P: G \to [0, 1]$ such that $\sum_{g \in G} P(g) = 1$. We will assume that P is symmetric – i.e., $P(g) = P(g^{-1})$ – and we will also assume that $P(e) > 0$.

In a natural way, P determines a left-invariant random walk on G, by the following construction: given an element $x \in G$, the probability, when taking a step from x, of landing on xg, is precisely $P(g)$.

If $\{g_i\}$ denotes some indexing of the group G, let M denote the infinite matrix

$$M = (m_{i,j}) \quad \text{where} \quad m_{i,j} = P(g_i^{-1} g_j)$$

M is clearly a symmetric matrix, all of whose entries are positive, and whose rows and columns sum to 1. Let also $P^{(k)}$ denote the function

$$P^{(k)}(g) = \sum_{g_1 \cdot g_2 \cdots g_k = g} P(g_1) \cdot P(g_2) \cdots P(g_k).$$

Then $P^{(k)}$ denotes the probability distribution associated to the random walk corresponding to taking k steps in the random walk determined by P; the associated matrix $M^{(k)}$ is obtained from M by k-fold matrix multiplication.

We define the Kesten number $\kappa(P)$ by the formula

$$\kappa(P) = \lim_{k \to \infty} (m_{g,g}^k)^{1/k}$$

where g is any element of G, for instance e.

$\kappa(P)$ has a number of interpretations. One sees readily that if $\lambda_1 < \kappa(P) < \lambda_2$, then for sufficiently large k we have $(\lambda_1)^k < P^{(k)}(e) < (\lambda_2)^k$ – thus $\kappa(P)$ is an asymptotic estimate on the probability that after k steps of the random walk P, one has returned to the starting point.

Another interpretation arises by allowing P to act on $l^2(G)$ according to the convolution law

$$(P * f)(g) = \sum_{g'} f(gg') \cdot P(g').$$

Then $\kappa(P)$ is precisely the spectral radius of this operator P.

With this understood, we state the following:

THEOREM (Kesten) [12]. *Assuming that the support of P generates the group G, then $\kappa(P) = 1$ if and only if G is amenable.*

In particular, whether or not $\kappa(P) = 1$ does not depend on P. The value of $\kappa(P)$, however, doesn't have an intrinsic group-theoretic meaning, and depends on P.

Now we introduce a probability distribution on $\pi_1(M)$ in the following geometric way:

Let $H_t(x, y)$ denote the fundamental solution for the heat equation in $\tilde{M}$ – the existence and uniqueness of $H_t(x, y)$ is established in [7], see also [5]. We also fix for the discussion a fundamental domain F in $\tilde{M}$.

For each t, let $P_t(g)$ be the probability distribution defined by the formula

$$P_t(g) = \frac{1}{\text{vol}(F)} \int_{Fxg \cdot F} H_t(x, y)\, dx\, dy.$$

It is evident that $P_t(g)$ is a symmetric probability distribution on $\pi_1(M)$ whose support is all of $\pi_1(M)$. Let κ_t denote the Kesten number associated to P_t.

Intuitively, P_t describes the following random walk on $\tilde{M}$: at time 0, a unit amount of heat is uniformly distributed throughout F. The heat then distributes itself about $\tilde{M}$, according to the heat equation, until at time t. At this point, the heat is redistributed evenly within each fundamental domain. Then the heat flows according to the heat equation until time $2t$, at which time the heat is again redistributed evenly in each fundamental domain, and the process continues. Then κ_t is an asymptotic measure of the heat remaining in F.

We compare this with the random walk on $\tilde{M}$ which is given by the undisrupted flow of heat for all time. We define a number κ_∞, the analogue of the Kesten number in this case, by the formula

$$\kappa_\infty = \lim_{t \to \infty} \left(\int_{F \times F} H_t(x, y)\, dx\, dy \right)^{1/t} = e^{-\lambda_0(\tilde{M})}.$$

where the last equality follows from the spectral representation of $H_t(x, y)$. Thus Theorem 1 and Kesten's Theorem give

THEOREM 4. *For any t, $\kappa_t = 1$ if and only if $\lambda_0(\tilde{M}) = 0$.*

More precisely, one sees easily that

$$(P_t(e))^{1/t} = \left(\frac{1}{\text{vol}(F)} \int_{F \times F} H_t(x, y)\, dx\, dy \right)^{1/t} \to e^{-\lambda_0(\tilde{M})} \quad \text{as} \quad t \to \infty,$$

and that $\kappa_t \geq P_t(e)$, so that $\lim_{t \to \infty} (\kappa_t)^{1/t} = e^{-\lambda_0(M)}$. Hence it is reasonable to ask:

Question. Do we have $\lim_{t \to \infty} (\kappa_t)^{1/t} = e^{-\lambda_0(M)}$?

An affirmative answer would give a new proof of the result of §2. If one could also show that the right-hand side cannot be 1 without each term in the limit on the left-hand side being 1, then one would have a new proof of Theorem 1.

REFERENCES

[1] W. K. ALLARD, *On the First Variation of a Varifold*, Ann. Math. *95* (1972), 417–491.
[2] R. BROOKS, *Exponential Growth and the Spectrum of the Laplacian*, Proc. AMS *82* (1981), 473–477.
[3] ——, *Some Remarks on Bounded Cohomology*, Ann. Math. Studies *97* (1980), 53–63.
[4] J. CHEEGER, *A Lower Bound for the Smallest Eigenvalue of the Laplacian*, in Gunning, *Problems in Analysis* (1970), 195–199.
[5] J. CHEEGER and S. T. YAU, *A Lower Bound for the Heat Kernel*, preprint.
[6] S. Y. CHENG and S. T. YAU, *Differential Equations on Riemannian Manifolds and their Geometric Applications*, Comm. Pure and Appl. Math., *28* (1975), 333–354.
[7] H. DONNELLY, *Asymptotic Expansions for the Compact Quotients of Properly Discontinuous Group Actions*, Ill. J. of Math. *23* (1979), 485–496.
[8] P. EBERLEIN, *Some Properties of the Fundamental Group of a Fuchsian Manifold*, Inventiones Math. *19* (1973), 5–13.
[9] F. EISENBERG, Ph.D. thesis, Stony Brook (1979).
[10] F. P. GREENLEAF, *Invariants Means on Topological Groups and their Applications*, von Nostrand, Reinhold & Co., 1969.
[11] M. HIRSCH and W. THURSTON, *Foliated Bundles, Invariant Measures, and Flat Manifolds*, Ann. Math. *101* (1975), 369–390.
[12] H. KESTEN, *Full Banach Mean Values on Countable Groups*, Math. Scand. *7* (1959), 146–156.
[13] H. P. MCKEAN, *An Upper Bound to the Spectrum of Δ on a Manifold of Negative Curvature*, J. Diff. Geometry *4* (1970), 359–366.
[14] J. MILNOR, *A Note on Curvature and Fundamental Group*, J. Diff. Geom. *2* (1968), 1–7.
[15] J. TITS, *Free Subgroups in Linear Groups*, J. Algebra *20* (1972), 250–270.
[16] H. FEDERER, *Geometric Measure Theory*, Springer Verlag (1969).
[17] E. FØLNER, *On groups with Full Banach Mean Value*, Math. Scand. *3* (1955), 243–254.
[18] H. B. LAWSON, *Minimal Varieties in Real and Complex Geometry*, Séminaire de Mathématiques Superieures, vol. 57, Université de Montréal, 1974.
[19] S. T. YAU, *Isoperimetric Constants and the First Eigenvalue of a Compact Riemannian Manifold*, Ann. Sci. Éc. Norm. Sup. (1975) 487–507.

Dept. of Mathematics and Institute for Physical Science and Technology
University of Maryland
College Park, MD 20742 USA.

Received August 13, 1980/July 8, 1981

Comment. Math. Helvetici **56** (1981) 599–614 0010–2571/81/004599–16$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Self homotopy equivalences of virtually nilpotent spaces*

E. DROR, W. G. DWYER and D. M. KAN

§ 1. Introduction

The aim of this paper is to prove Theorem 1.1 below, a generalization to *virtually nilpotent spaces* of a result of Wilkerson [13] and Sullivan [12]. We recall that a CW complex Y is *virtually nilpotent* if

(i) Y is connected,

(ii) $\pi_1 Y$ is *virtually nilpotent* (i.e. has a nilpotent subgroup of finite index) and

(iii) for every integer $n > 1$, $\pi_1 Y$ has a subgroup of finite index which acts nilpotently on $\pi_n Y$. The class of virtually nilpotent spaces is much larger than the class of nilpotent spaces. For instance such non-nilpotent spaces as the Klein bottle and the real projective spaces are virtually nilpotent, and so is, of course, any connected space with a finite fundamental group.

1.1. THEOREM. *Let Y be a virtually nilpotent finite* CW *complex. Then the classifying space of the topological monoid of the self homotopy equivalences of Y is of finite type* (i.e. has the homotopy type of a CW complex with a finite number of cells in each dimension). *In fact it has the somewhat stronger property that each of its homotopy groups is of finite type* (i.e. has a classifying space of finite type).

1.2. *Remark.* For abelian groups, being of finite type is the same as being finitely generated, but for non-abelian groups, being of finite type is stronger than being finitely generated or even being finitely presented.

1.3. *Remark.* As it is easy to verify that, in Theorem 1.1, the higher homotopy groups in question are finitely generated, the main content of Theorem 1.1 is that *the group of homotopy classes of self homotopy equivalences of Y is of finite type.*

1.4. ORGANIZATION OF THE PAPER. The paper consists essentially of three parts:

(i) After a brief discussion (in § 2) of the notion of *finite type* for groups and simplicial sets, we reduce Theorem 1.1 (in § 3) to a similar (and in fact equivalent)

* This research was in part supported by the National Science Foundation.

statement (3.6) about the *homotopy automorphism complex* (i.e. complex of self loop homotopy equivalences) *of a simplicial group* and then (in §4) to a similar statement (4.1) about the *homotopy automorphism complex of a simplicial virtually nilpotent group.* The arguments are standard, except in 3.5, where we make the transition from pointed simplicial sets to simplicial groups and have to show that the usual notion of a function complex of maps between two simplicial groups is indeed the "correct" one.

(ii) In the next two sections we make the crucial transition from *homotopy automorphisms* to *automorphisms*, i.e. we reduce Theorem 4.1 to a similar statement (6.1) about the *automorphism complex of a simplicial virtually nilpotent group.* A key step in the argument is a curious lemma (5.1) which states that, *under suitable circumstances, the homotopy groups of the automorphism complex of a simplicial module differ by only a finite amount from the homotopy groups of the homotopy automorphism complex.*

(iii) The last two sections are devoted to a proof of Theorem 6.1. It turns out that it suffices to show that the automorphism complexes involved are dimension wise of finite type and this we then do by combining variations on arguments of Baumslags proof that the automorphism group of a finitely generated virtually nilpotent group is finitely presented [1, Ch. 4] with the result of Borel and Serre that arithmetic subgroups of algebraic groups are of finite type [2, §11]. Of course it would be nice if one could do this without resorting to such non-homotopical notions as algebraic groups and their arithmetic subgroups.

§2. Finite type

We start with a brief review of the notions of finite type for simplicial sets and for groups, and note in particular (2.9) that *a connected simplicial set with finitely generated higher homotopy groups is of finite type if and only if its fundamental group is of finite type.*

2.1. SIMPLICAL SETS OF FINITE TYPE. A simplicial set X is said to be of *finite type* if, for every integer $n \geq 0$, there exists a map $f_n : F_n \to X$ such that

(i) F_n is finite (i.e. has only a finite number of non-degenerate simplices) and

(ii) f_n induces, for every vertex $v \in F_n$ and every integer $0 \leq i \leq n$, an isomorphism $\pi_i(F_n; v) \approx \pi_i(X; f_n v)$.

This definition readily implies ([8], [11, Ch. III and Ch. VI]):

2.2. PROPOSITION. *A simplicial set X is of finite type if and only if its realization $|X|$ has the homotopy type of a* CW *complex with a finite number of cells in each dimension.*

2.3. PROPOSITION. *A reduced* (i.e. only one vertex) *simplicial set X is of finite type if and only if its simplicial loop group GX has the loop homotopy type of a free simplicial group which is finitely generated in each dimension.*

The next two propositions are very useful ones.

2.4. PROPOSITION. *Let U be a bisimplicial set such that, for every integer $k \geq 0$, the simplicial set $U_{k,*}$ is of finite type. Then the diagonal diag U is also of finite type.*

2.5. PROPOSITION. *Let $p : E \to B$ be a fibration onto and assume that all its fibres are of finite type. Then E is of finite type if and only if B is so.*

Proofs. The proof of 2.4 is easy once the diagonal of the bisimplicial set U has been identified with the "realization" of U [3, Ch. XII, 3.4]. The "if" part of 2.5 is straightforward. To prove the "only if" part of 2.5, let U and V be the bisimplicial sets such that, for every integer $k \geq 0$,

$$V_{k,*} = B \quad \text{and} \quad U_{k,*} = E \times_B \cdots \times_B E \ (k+1 \text{ factors})$$

and let $U \to V$ be the obvious map. Then it is not hard to verify that, for every integer $n \geq 0$, the induced map $U_{*,n} \to V_{*,n} = B_n$ is a weak homotopy equivalence and so is therefore [3, p. 335] the induced map diag $U \to$ diag $V = B$. The desired result now follows from 2.4, the "if" part of 2.5 and the fact that $U_{0,*} = E$ and that, for every integer $k \geq 0$, the face maps $d_i : U_{k+1,*} \to U_{k,*}$ are fibrations with the fibres of p as fibres.

Next we consider

2.6. GROUPS OF FINITE TYPE. A group G is said to be of *finite type* if the simplicial set $K(G, 1)$ is of finite type.

2.7. EXAMPLES. Using 2.5 one readily verifies that the following groups are of finite type:

(i) all *finitely generated free groups,*
(ii) all *finite groups*
(iii) all *finitely generated abelian groups,*
(iv) all *finitely generated nilpotent groups,*
(v) all *finitely generated virtually nilpotent* (see § 1) *groups,* and
(vi) all *homotopy groups of a simplicial set which is virtually nilpotent* (see § 1) *and of finite type.*

Less obvious are

2.8. EXAMPLES. (i) Every *arithmetic subgroup of an algebraic group* is of finite type. This is a result of Borel–Serre [2, § 11].

(ii) The *group of automorphisms of a finitely generated virtually nilpotent group* is of finite type. To prove this one combines Baumslag's proof of [1, th. 4.7] with 2.5 and 2.8 (i).

We end with several propositions which will be needed later.

2.9. PROPOSITION. *Let X be a connected simplicial set and assume that $\pi_n X$ is of finite type for $n > 1$. Then $\pi_1 X$ is of finite type if and only if X is of finite type.*

2.10. PROPOSITION. *Let C be a simplicial group such that C_n is of finite type for all $n \geq 0$. Then its classifying complex $\bar{W}C$ [11, Ch. IV] is of finite type.*

2.11. PROPOSITION. *Let $G \to \{G_i\}$ be a pro-isomorphism of groups* [2, Ch. III] *in which each G_i is of finite type. Then G is also of finite type.*

Proofs. Propositions 2.9 and 2.10 follow readily from Propositions 2.5 and 2.4 respectively, while Proposition 2.11 is an immediate consequence of the fact that any retract of a simplicial set of finite type is also of finite type.

§ 3. Reduction to simplicial groups

In this section we reduce Theorem 1.1 to similar and equivalent results for *simplicial sets* (3.2), *reduced* (i.e. only one vertex) *simplicial sets* (3.4) and *simplicial groups* (3.6). Most of the arguments are routine. However, in the last reduction one runs into the problem that the loop group functor G is not a simplicial functor with respect to the usual simplicial structures on the categories of reduced simplicial sets and simplicial groups. To get around this difficulty we introduce on the category of reduced simplicial sets a new simplicial structure which is better behaved with respect to the functor G and which gives rise to function complexes homotopically equivalent to the usual ones. Of course one could instead have appealed to the rather general Proposition 5.4 of [5].

3.1. REDUCTION TO SIMPLICIAL SETS. For a CW complex Y let *haut* Y denote its *simplicial monoid of self homotopy equivalences*, i.e. the simplicial monoid which has as its n-simplices the homotopy equivalences $|\Delta[n]| \times Y \to Y$,

and for a fibrant (i.e. satisfying the extension condition [11, § 1]) simplicial set X, let haut X denote its simplicial monoid of homotopy automorphisms, i.e. the simplicial monoid which has as its n-simplices the weak homotopy equivalences $\Delta[n]\times X \to X$. Using

(i) the adjointness of the realization functor $|\ |$ and the singular functor Sin,

(ii) the fact that, for every CW complex Y and fibrant simplicial set X, the adjunction maps $|\mathrm{Sin}\ Y| \to Y$ and $X \to \mathrm{Sin}|X|$ are homotopy equivalences, and

(iii) the fact that, for every simplicial set X, the obvious maps $|\Delta[n]\times X| \to |\Delta[n]|\times|X|$ are homeomorphisms, one readily verifies that *the induced maps* π_n haut $X \to \pi_n$ haut $|X|$ *are isomorphisms for all* $n \geq 0$. As haut Y is clearly isomorphic to the singular complex of the topological monoid of self homotopy equivalences of Y, if follows that Theorem 1.1 is equivalent to

3.2. THEOREM. *Let X be a virtually nilpotent fibrant simplicial set which has the (weak) homotopy type of a finite simplicial set. Then* π_n haut X *is of finite type for all* $n \geq 0$.

3.3. REDUCTION TO REDUCED SIMPLICIAL SETS. For a reduced (i.e. only one vertex) fibrant simplicial set K, denote by $\mathrm{haut}_* K$ the submonoid of haut K which "keeps the vertex fixed" and note that there is an obvious fibration haut $K \to K$ with $\mathrm{haut}_* K$ as fibre. Using 2.5 and 2.7 (vi) one then readily sees that Theorem 3.2 is equivalent to

3.4. THEOREM. *Let K be a virtually nilpotent fibrant simplicial set which is reduced and has the (weak) homotopy type of a finite simplicial set. Then* π_n $\mathrm{haut}_* K$ *is of finite type for all* $n \geq 0$.

3.5. REDUCTION TO SIMPLICIAL GROUPS. We start with constructing a new simplicial structure on the category of reduced simplicial sets along the lines of [7, § 12], i.e. for a simplicial set X and a reduced simplicial set K, we denote by $X \cdot K$ the reduced simplicial set which is the quotient of $X \times K$ by the equivalence relation: $(x_1, k_1) \sim (x_2, k_2)$ *if and only if* $k_1 = k_2 = s_0^i k$ *for some non-degenerate* $k \in K$ *and* $d_0^{i+1} x_1 = d_0^{i+1} x_2$, and note that this definition readily implies the existence of a *natural isomorphism* $(X' \times X) \cdot K \approx X' \cdot (X \cdot K)$.

Next, for a fibrant reduced simplicial set K, denote by $\mathrm{haut}_0 K \subset \mathrm{haut}_* K$ the subcomplex consisting of the maps $\Delta[n]\times K \to K$ which factor through $\Delta[n] \cdot K$. Then $\mathrm{haut}_0 K$ *is clearly a submonoid of* $\mathrm{haut}_* K$. Moreover the usual retraction $\Delta[1]\times\Delta[n] \to \Delta[n]$ of $\Delta[n]$ onto its first (or last) vertex induces a retraction of $\Delta[n] \cdot K$ onto K and hence [5, § 6] *the induced maps* π_n $\mathrm{haut}_0 K \to \pi_n$ $\mathrm{haut}_* K$ *are isomorphisms for all* $n \geq 0$.

Finally, for a simplicial group C, denote by haut C its *simplicial monoid of homotopy automorphisms*, i.e. the simplicial monoid which has as its n-simplices the homomorphisms $\Delta[n]\otimes C \to C$ which are weak (loop homotopy) equivalences [11, Ch. VI] (i.e. induce isomorphisms on π_n for all $n \geq 0$). Using [11, Ch. VI],

(i) the adjointness of the loop group functor G and the classifying complex functor $\bar{W}$,

(ii) the fact that, for every fibrant reduced simplicial set K and every free simplicial group C, the adjunction maps $K \to \bar{W}GK$ and $G\bar{W}C \to C$ are respectively a homotopy equivalence and a loop homotopy equivalence, and

(iii) the fact that, for every reduced simplicial set K, the homomorphisms $\Delta[n]\otimes GK \to G(\Delta[n]\cdot K)$, given by [7, p. 118] $(x, \tau k) \to \tau(s_0 x, k)$, are actually isomorphisms, one verifies that *the induced maps* $\pi_n \operatorname{haut}_0 K \to \pi_n \operatorname{haut} GK$ *are isomorphisms for all* $n \geq 0$. It now follows that Theorem 3.4 is equivalent to

3.6. THEOREM. *Let C be a free simplicial group which is finitely generated* (i.e. has a finite number of non-degenerate generators) *and has a virtually nilpotent classifying complex* $\bar{W}C$ [11, Ch. IV]. *Then* π_n haut C *is of finite type for all* $n \geq 0$.

§ 4. Reduction to simplicial virtually nilpotent groups

Now we reduce Theorem 3.6 to a similar result for simplicial virtually nilpotent groups (4.1). To state this result denote, for a homomorphism of simplicial groups $C \to \pi$, by $\operatorname{haut}_\pi C \subset \operatorname{haut} C$ the *simplicial monoid of homotopy automorphisms of C over π*, i.e. the simplicial monoid which has as its n-simplices the commutative diagrams

$$\begin{array}{ccc} \Delta[n]\otimes C & \to & C \\ & \searrow \quad \swarrow & \\ & \pi & \end{array}$$

in which the top map is in haut C and the other maps are the obvious ones. Furthermore, for a (simplicial) group B, let $\Gamma_i B$ be the i-th term of its *lower central series* (i.e. $\Gamma_1 B = B$ and $\Gamma_i B = [\Gamma_{i-1} B, B]$ for $i > 1$. Then one has

4.1. THEOREM. *Let* $1 \to B \to C \to \pi \to 1$ *be an exact sequence of simplicial groups such that*

(i) π *is discrete and finite,*

(ii) C *is free and finitely generated* (see 3.6), *and*

(iii) *the classifying complex* $\bar{W}B$ [11, Ch. IV] *is nilpotent.*
Then the groups $\pi_n \operatorname{haut}_\pi C/\Gamma_i B$ $(n \geq 0, i \geq 1)$ *are of finite type.*

That indeed this Theorem 4.1 implies Theorem 3.6 is an immediate consequence of 2.11 and the following three propositions.

4.2. PROPOSITION. *Let* $1 \to B \to C \to \pi \to 1$ *be as in* 4.1. *Then the obvious maps*

$$\pi_n \operatorname{haut}_\pi C \to \{\pi_n \operatorname{haut}_\pi C/\Gamma_i B\} \qquad n \geq 0$$

are pro-isomorphisms of groups [3, Ch. III].

Proof. In view of [3, Ch. III] the obvious map $C \to \{C/\Gamma_i B\}$ is a weak pro-homotopy equivalence and it is not difficult to show, using induction on the number of non-degenerate generators of C, that so is the induced map of function complexes over π

$$\hom_\pi (C, C) \to \{\hom_\pi (C, C/\Gamma_i B)\}$$

and the desired result now follows readily from the obvious isomorphisms

$$\hom_\pi (C, C/\Gamma_i B) \approx \hom_\pi (C/\Gamma_i B, C/\Gamma_i B)$$

4.3. PROPOSITION. *Let* $C \to \pi$ *be a homomorphism of simplicial groups such that* $\pi_0 C$ *is finitely generated and* π *is discrete and finite. Then* $\pi_n \operatorname{haut}_\pi C = \pi_n \operatorname{haut} C$ *for* $n \geq 1$ *and* $\pi_0 \operatorname{haut}_\pi C$ *is a subgroup of finite index of* $\pi_0 \operatorname{haut} C$.

Proof. This follows readily from the fact that a finitely generated group (such as $\pi_0 C$) has only a finite number of subgroups of a given finite index.

4.4. PROPOSITION. *Let* C *be a finitely generated free simplicial group with a virtually nilpotent classifying complex* $\bar{W}C$. *Then there exists an exact sequence* $1 \to B \to C \to \pi \to 1$ *of simplicial groups such that*

(i) π *is discrete and finite, and*
(ii) $\bar{W}B$ *is nilpotent.*

Proof. In view of [8] if suffices to show that *every virtually nilpotent finite CW complex* Y *has a nilpotent finite cover.* To prove this let $\varphi \subset \pi_1 Y$ be a nilpotent subgroup of finite index which acts nilpotently on $\pi_n Y$ for $2 \leq n \leq \dim Y$. Then φ acts on the universal cover $\tilde{Y}$ of Y. As φ acts nilpotently on $\pi_n \tilde{Y}$ for $2 \leq n \leq \dim Y$ and $\dim \tilde{Y} = \dim Y$, it follows that φ acts nilpotently on $H_n \tilde{Y}$ for all $n \geq 0$ and therefore on $\pi_n \tilde{Y}$ for all $n \geq 2$. The desired result is now immediate.

§ 5. Automorphisms of simplicial modules

In preparation for the next step in our reduction (in § 6) we prove here a lemma for simplicial modules (5.1) which seems to be of interest in its own right.

For a simplicial π-module M, let $\mathrm{haut}_\pi M$ be its *simplicial monoid of homotopy automorphisms* (an n-simplex of which is a π-module homomorphism $\Delta[n]\otimes M \to M$ which is a homotopy equivalence) and let $\mathrm{aut}_\pi M \subset \mathrm{haut}_\pi M$ be its maximal *simplicial subgroup of automorphisms.* Then one has:

5.1. LEMMA. *Let π be a finite group and let M be a finitely generated* (3.6) *simplicial π-module which, in each dimension, is torsion free as an abelian group. Then the obvious maps $\pi_n \mathrm{aut}_\pi M \to \pi_n \mathrm{haut}_\pi M$ $(n \geq 0)$ have finite kernels and cokernels.*

5.2. *Remark.* Lemma 5.1 remains true if M is *not* required to be torsion free in each dimension, but we don't need this extra generality.

Proof. The proof consists of three parts and will often, explicitly or implicitly, use the fact that [11, Ch. V] *there exists an isomorphism of categories N between the category of simplicial π-modules and the category of differential graded π-modules which are trivial in negative dimensions.* First we note that 5.1 holds if $NM_i = 0$ for $i \neq n, n+1$ and $\pi_i M = 0$ for $i \neq n$. Next we consider a finite direct sum of such simplicial π-modules and finally we treat the general case.

I. Assume that M is as in 5.1 and that in addition $NM_i = 0$ for $i \neq n, n+1$ and $\pi_i M = 0$ for $i \neq n$. Then the boundary map $\partial : NM_{n+1} \to NM_n$ is a monomorphism and we can therefore consider NM_{n+1} as a submodule of NM_n. Furthermore $N(\Delta[K]\otimes M)_n$ is a direct sum of copies of NM_n indexed by the n-simplices of $\Delta[k]$ and if, for every n-simplex $p \in \Delta[k]$ and element $x \in NM_n$, we denote by $x_p \in N(\Delta[k]\otimes M)_n$ the copy of x that lies in the summand indexed by p, then a straightforward calculation yields that the image of the boundary map $\partial : N(\Delta[k]\otimes M)_{n+1} \to N(\Delta[k]\otimes M)_n$ is generated by the elements

$$x_p - x_q \quad \text{where} \quad x \in NM_n \quad \text{and} \quad p, q \in \Delta[k]_n$$

$$x_p \quad \text{where} \quad x \in NM_{n+1} \quad \text{and} \quad p \in \Delta[k]_n$$

Next one notes that a k-simplex $f \in \hom_\pi (M, M)$ is completely determined by a collection of π-modules maps $f_p : NM_n \to NM_n$ indexed by the n-simplices of $\Delta[k]$ and it is not difficult to verify that conversely such a collection $\{f_p\}$

(i) comes from a k-simplex of $\hom_\pi (M, M)$ iff each f_p maps NM_{n+1} into itself and all f_p induce the same endomorphism of NM_n/NM_{n+1},

(ii) comes from a k-simplex of $\mathrm{haut}_\pi M$ iff, in addition to the conditions of (i), the f_p induce an automorphism of NM_n/NM_{n+1} and

(iii) comes from a k-simplex of $\mathrm{aut}_\pi M$ iff, is addition to the conditions of (i) and (ii), each f_p is an automorphism of NM_n.

From this, together with the usual combinatorial formulas for the homotopy groups of a complex satisfying the extension condition [10, p. 5], it is not hard to deduce that *the homotopy groups of* $\mathrm{aut}_\pi M$ *and* $\mathrm{haut}_\pi M$ *vanish in dimension* >0 *and that the map* $\pi_0 \mathrm{aut}_\pi M \to \pi_0 \mathrm{haut}_\pi M$ *can be identified with the inclusion, into the group of automorphisms of* NM_n/NM_{n+1} *that lift to endomorphisms of* NM_n, *of those automorphisms of* NM_n/NM_{n+1} *that lift to automorphisms of* NM_n. To see that this inclusion is of finite index, one notes that NM_n determines an element in the finite group $\mathrm{Ext}_\pi(NM_n/NM_{n+1}, NM_{n+1})$ and that the automorphisms of NM_n/NM_{n+1} that stabilize this element are contained in the image of the inclusion in question.

II. Assume that $M \approx M^0 \oplus \cdots \oplus M^r$ where each M^n $(0 \le n \le r)$ is as in I; in particular $NM_i^n = 0$ if $i \neq n, n+1$ and $\pi_i M^n = 0$ if $i \neq n$. Then one readily verifies, using the functor N, that $\hom_\pi(M^i, M^j) = 0$ for $i > j$. Hence *the* k*-simplices of* $\mathrm{aut}_\pi M$ (*resp.* $\mathrm{haut}_\pi M$) *are in* 1–1 *correspondence with* $(n \times n)$*-matrices* $\{s_{i,j}\}$ (*resp.* $\{t_{i,j}\}$) *with*

$$s_{i,j},\ t_{i,j} \in \hom_\pi(M^i, M^j)_k \quad \text{for} \quad i < j.$$

$$s_{ij} = t_{i,j} = 0 \quad \text{for} \quad i > j$$

$$s_{i,i} \in (\mathrm{aut}_\pi M^i)_k \qquad t_{i,i} \in (\mathrm{haut}_\pi M^i)_k$$

and the desired result is now immediate.

III. Finally assume that M is merely as in 5.1. For every integer $k \ge -1$, let $E_k M \subset M$ be the maximal simplicial submodule which is trivial in dimensions $\le k$. As M is finitely generated there is an integer r such that $E_r M = 0$ and the finiteness of π now readily implies the existence of an isomorphism of simplicial π-modules

$$M \otimes Q \approx (E_{-1}M/E_0 M \otimes Q) \oplus \cdots \oplus (E_{r-1}M/E_r M \otimes Q)$$

Using the functor N it is not hard to show that there is a sequence of positive integers $t = (t_0, \ldots, t_r)$ with t_{k-1} dividing t_k $(1 \le k \le r)$ such that, for the simplicial submodule $t^{-1}M \subset M \otimes Q$ given by

$$x \in t^{-1}M \quad \textit{iff} \quad t_k x \in E_k M + E_{k-1}(M \otimes Q) \quad \textit{for all} \quad 0 \le k \le r$$

there is an isomorphism of simplicial modules

$$t^{-1}M \approx E_{-1}t^{-1}M/E_0t^{-1}M \oplus \cdots \oplus E_{r-1}t^{-1}M/E_rt^{-1}M$$

i.e. $t^{-1}M$ *satisfies the conditions of* II. Moreover the naturality of the construction t^{-1} implies the existence of a commutative diagram

$$\begin{array}{ccc} \mathrm{aut}_\pi M & \longrightarrow & \mathrm{aut}_\pi t^{-1}M \\ \downarrow{\scriptstyle \mathrm{incl.}} & & \downarrow{\scriptstyle \mathrm{incl.}} \\ \mathrm{haut}_\pi M & \longrightarrow & \mathrm{haut}_\pi t^{-1}M \end{array}$$

in which, because M was assumed to be dimension wise torsion free as an abelian group, the horizontal maps are 1–1 and it remains to show that the maps they induce on the homotopy groups have finite kernels and cokernels in all dimensions ≥ 0.

To do this for the top map we note that a k-simplex $f \in \mathrm{aut}_\pi M$ (resp. $\mathrm{aut}_\pi t^{-1}M$) is completely determined by a collection of automorphisms $f_p : M_{\dim p} \to M_{\dim p}$ (resp. $t^{-1}M_{\dim p} \to t^{-1}M_{\dim p}$) indexed by the simplices $p \in \Delta[k]$ of dimension $\leq r+k$. From this and the fact that, in each dimension, M has index in $t^{-1}M$, it is not difficult to deduce that the image of $\mathrm{aut}_\pi M$ in $\mathrm{aut}_\pi t^{-1}M$ also has finite index in each dimension and the desired result follows.

The bottom map is the restriction to certain components of a homomorphism of simplicial abelian groups $\hom_\pi (M, M) \to \hom_\pi (t^{-1}M, t^{-1}M)$ which is 1–1 and has a finite cokernel in each dimension. Hence the induced map $\pi_n \mathrm{haut}_\pi M \to \pi_n \mathrm{haut}_\pi t^{-1}M$ has a finite kernel for $n \geq 0$ and a finite cokernel for $n > 0$. That this map also has a finite cokernel for $n = 0$ follows from the fact that (see above) the composition $\pi_0 \mathrm{aut}_\pi M \to \pi_0 \mathrm{haut}_\pi M \to \pi_0 \mathrm{haut}_\pi t^{-1}M$ does.

§ 6. Reduction to automorphisms

The next reduction step is to show that Theorem 4.1 is equivalent to a similar result for the maximal *simplicial* sub-*groups of automorphisms* $\mathrm{aut}_\pi C/\Gamma_i B \subset \mathrm{haut}_\pi C/\Gamma_i B$, i.e.

6.1. THEOREM. $1 \to B \to C \to \pi \to 1$ *be as in* 4.1. *Then the groups* $\pi_n \mathrm{aut}_\pi C/\Gamma_i B$ $(n \geq 0, i \geq 1)$ *are of finite type.*

This equivalence follows immediately from

6.2. LEMMA. *Let* $1 \to B \to C \to \pi \to 1$ *be as in* 4.1 (i) *and* (ii). *Then the obvious maps* $\pi_n \operatorname{aut}_\pi C/\Gamma_i B \to \pi_n \operatorname{haut}_\pi C/\Gamma_i B$ $(n \geq 0, i \geq 1)$ *have finite kernels and cokernels.*

Proof. Note that there is a pull back diagram

$$\begin{array}{ccc} \operatorname{aut}_\pi C/\Gamma_i B & \longrightarrow & \operatorname{haut}_\pi C/\Gamma_i B \\ \downarrow & & \downarrow \\ \operatorname{aut}_\pi C/\Gamma_2 B & \longrightarrow & \operatorname{haut}_\pi C/\Gamma_2 B \end{array}$$

in which, since C is free, the map on the right is a fibration. Hence it suffices to prove the lemma for $i = 2$ only, which we will do by reducing this case to Lemma 5.1.

Consider the commutative diagram

$$\begin{array}{ccccccc} Z^1(\pi; B/\Gamma_2 B) & \xrightarrow{a} & \operatorname{aut}_\pi C/\Gamma_2 B & \xrightarrow{b} & \operatorname{aut}_\pi B/\Gamma_2 B & \xrightarrow{c} & H^2(\pi; B/\Gamma_2 B) \\ \downarrow \simeq & & \downarrow & & \downarrow & & \downarrow \simeq \\ Z^1(\pi; B/\Gamma_2 B) & \xrightarrow{a} & \operatorname{haut}_\pi C/\Gamma_2 B & \xrightarrow{b} & \operatorname{haut}_\pi B/\Gamma_2 B & \to \operatorname{end}_\pi B/\Gamma_2 B \xrightarrow{c} & H^2(\pi; B/\Gamma_2 B) \end{array}$$

constructed as follows:

(i) The maps b are induced by the functor which, to every epimorphism $H \to \pi$ with abelian kernel, assigns this kernel (as a π-module).

(ii) For every 1-cocycle $z \in Z^1(\pi; B_n/\Gamma_2 B_n)$ (i.e. function $z : \pi \to B_n/\Gamma_2 B_n$ such that $z(xy) = xz(y) + z(x)$ for all $x, y \in \pi$), the map $az : \Delta[n] \otimes C/\Gamma_2 B \to C/\Gamma_2 B$ assigns to k-simplices $p \in \Delta[n]$ and $q \in C/\Gamma_2 B$, the k-simplex $p'(zq') \cdot q \in C/\Gamma_2 B$, where q' denotes the image of q in π and p' is the simplicial operator such that $p = p' i_n$, where $i_n \in \Delta[n]$ is the non-degenerate n-simplex.

(iii) For an n-simplex $r \in \operatorname{end}_\pi B/\Gamma_2 B = \hom_\pi (B/\Gamma_2 B, B/\Gamma_2 B)$ we put $cr = k_n - r'_* k_n$, where $k_n \in H^2(\pi; B_n/\Gamma_2 B_n)$ is the extension class [8, Ch. IV] of $B_n/\Gamma_2 B_n \to C_n/\Gamma_2 B_n \to \pi$ and $r' : B_n/\Gamma_2 B_n \to B_n/\Gamma_2 B_n$ is the restriction of r to the non-degenerate n-simplex $i_n \in \Delta[n]$.

Then it is not difficult to verify that the maps a are 1–1, that $Z^1(\pi; B/\Gamma_2 B)$ acts principally on $\operatorname{aut}_\pi C/\Gamma_2 B$ and $\operatorname{haut}_\pi C/\Gamma_2 B$ and that the maps b map the resulting quotients isomorphically onto the subcomplexes of $\operatorname{aut}_\pi B/\Gamma_2 B$ and $\operatorname{haut}_\pi B/\Gamma_2 B$ which go to 0 under c. Moreover the first of these quotients is a

simplicial group which acts principally on $\mathrm{aut}_\pi B/\Gamma_2 B$ and this readily implies that the image of $\mathrm{aut}_\pi B/\Gamma_2 B$ in $H^2(\pi; B/\Gamma_2 B)$ is fibrant; as $H^2(\pi; B/\Gamma_2 B)$ is dimension wise finite, so is this image and its homotopy groups are thus finite in all dimensions ≥ 0. To obtain a similar result for the image of $\mathrm{haut}_\pi B/\Gamma_2 B$ in $H^2(\pi; B/\Gamma_2 B)$, one notes that the map $c: \mathrm{end}_\pi B/\Gamma_2 B \to H^2(\pi; B/\Gamma_2 B)$ is a translation by $k \in H^2(\pi; B/\Gamma_2 B)$ of a homomorphism and that therefore the image of $\mathrm{end}_\pi B/\Gamma_2 B$ under c is fibrant. As $\mathrm{haut}_\pi B/\Gamma_2 B$ is a union of components of $\mathrm{end}_\pi B/\Gamma_2 B$, the same holds for the image of $\mathrm{haut}_\pi B/\Gamma_2 B$ under c and the desired result now readily follows.

§ 7. Automorphisms of diagrams of π-kernels

In this section we obtain a lemma (7.3) on automorphisms of diagrams of π-kernels in the sense of Eilenberg–Maclane, which will be used in § 8 to prove Theorem 6.1.

We start with a brief discussion of

7.1. π-KERNELS AND CENTRAL MAPS BETWEEN THEM. For a group G, let ζG denote its *center*, *aut* G its group of *automorphisms*, *in* $G \approx G/\zeta G$ its group of *inner automorphisms* and $\mathrm{out}\, G = (\mathrm{aut}\, G)/(\mathrm{in}\, G)$ its group of *outer automorphisms*. Given a group π, a *π-kernel* then is [9, Ch. IV] a pair (G, ψ) where G is a group and $\psi: \pi \to \mathrm{out}\, G$ a homomorphism. Similarly we define a *central map* $(G, \psi) \to (G', \psi')$ between two π-kernels as a pair (g, p) consisting of a homomorphism $g: G \to G'$ which sends ζG into $\zeta G'$ and a homomorphism

$$\pi \times_{\mathrm{out} G} \mathrm{aut}\, G \xrightarrow{p} \pi \times_{\mathrm{out} G'} \mathrm{aut}\, G'$$

over π which, over the identity of π, agrees with the homomorphism *in* $G \approx G/\zeta G \to G'/\zeta G' \approx$ *in* G' induced by g.

7.2. EXAMPLE. If $1 \to B \to C \to \pi \to 1$ is as in 4.1 (i) and (ii), then each $B_n/\Gamma_i B_n$ $(n \geq 0, i > 1)$ is a π-kernel in an obvious manner and all face operators between them become central maps [10, p. 347]. Moreover the same holds for the degeneracy operators if each $B_n/\Gamma_i B_n$ is nilpotent of class exactly $i-1$; otherwise they need not be "center preserving". This is automatic if C_0 and hence B_0 is free on more than one generator.

7.3. LEMMA. *Let π be a finite group, let D be a finite category* (i.e. its nerve is a finite simplicial set) *and let F be a functor from D to the category of π-kernels and*

central maps between them such that, for every object $d \in D$, Fd is finitely generated, nilpotent and torsion free as a group. Then the group $\mathrm{aut}_\pi F$ *of self natural equivalences of F is of finite type.*

To prove this we will freely use some elementary algebraic group theory as can be found, for instance, in [6, §21] and [4, IV, 2.2].

Proof. First we consider the case that D has only one object and its identity map and show, essentially following Baumslag [1, Ch. 4]: if F is a π-kernel which is finitely generated, nilpotent and torsion free as a group, then $\mathrm{aut}_\pi F$ is of finite type.

We start with proving that the group $\mathrm{aut}\, F$ of *group automorphisms* of F is of finite type. Let MF denote the (uniquely divisible nilpotent) Malcev completion of F [1, p. 50] and let LF be the finite dimensional nilpotent Lie algebra over the field Q of the rationals associated with F [1, p. 48]. The Baker–Campbell–Hausdorff formula gives rise to a natural set isomorphism $\log: MF \approx LF$, the group F admits a natural embedding $F \subset MF$, there exists [1, p. 51] at least one *lattice subgroup* of MF containing F (i.e. a group $F \subset F' \subset MF$ such that $\log F' \subset LF$ is a free abelian group which spans LF as a vector space) and the intersection $\bar{F}$ of all such lattice subgroups is itself a lattice subgroup, which is natural in F and contains F as a subgroup of finite index [13]. Moreover, as F and $\bar{F}$ are nilpotent, $\mathrm{aut}\, F$ is of finite index in $\mathrm{aut}\, \bar{F}$ [1, p. 61]. Furthermore $\mathrm{aut}\, \bar{F}$ is isomorphic to the subgroup of $\mathrm{aut}\, LF$ consisting of those Lie algebra automorphisms that carry $\log \bar{F}$ isomorphically onto itself (called the *stabilizer* in $\mathrm{aut}\, LF$ of the lattice $\log \bar{F}$). By construction $\mathrm{aut}\, LF$ is the group of rational points of a linear algebraic group **aut** LF over Q (which operates on the vector space LF). By definition the stabilizer in $\mathrm{aut}\, LF$ of the lattice $\log \bar{F}$ is an arithmetic subgroup of $\mathrm{aut}\, LF$. As [2, § 11] *every arithmetic subgroup of an algebraic group is of finite type*, so is $\mathrm{aut}\, \bar{F}$ and therefore $\mathrm{aut}\, F$.

Next we note that, using the Baker–Campbell–Hausdorff formula, it is not difficult to verify that $M(F/\zeta F)$ is in a natural manner the group of rational points of a unipotent algebraic group $\mathbf{M}(F/\zeta F)$ over Q. Conjugation produces a monomorphism $M(F/\zeta F) \to \mathrm{aut}\, LF$ which is easily seen to lift to an algebraic group map $\mathbf{M}(F/\zeta F) \to$ **aut** LF. The image of this map is contained in the unipotent radical of **aut** LF and is therefore a closed algebraic subgroup of **aut** LF, denoted by **in** LF. It follows that the image of $M(F/\zeta F)$ in $\mathrm{aut}\, LF$ (which we denote by $\mathrm{in}\, LF$) is the group of rational points of **in** LF.

Finally lift the structure homomorphism $\psi: \pi \to \mathrm{out}\, F$ to a function $\varphi: \pi \to \mathrm{aut}\, F$ and note that, for every element $y \in \pi$, the function $\mathrm{com}\, \varphi y: \mathrm{aut}\, LF \to \mathrm{aut}\, LF$ given by $t \to [t, L\varphi y]$ gives rise to a subset $(\mathrm{com}\, \varphi y)^{-1}(\mathrm{in}\, LF) \subset \mathrm{aut}\, LF$

which is readily verified to be a subgroup and not to depend on the choice of φ. The functions com φy clearly lift to algebraic maps **com** φy : **aut** $LF \to$ **aut** LF and, using the fact that **in** LF is a closed algebraic subgroup of **aut** LF, it is not difficult to see that each $(\mathbf{com}\ \varphi\psi)^{-1}(\mathbf{in}\ LF)$ is a closed algebraic subgroup of **aut** LF. If

$$\mathrm{aut}_\pi LF = \bigcap_{y \in \pi} (\mathrm{com}\ \varphi y)^{-1}(\mathrm{in}\ LF) \quad \text{and} \quad \mathbf{aut}_\pi LF = \bigcap_{y \in \pi} (\mathbf{com}\ \varphi y)^{-1}(\mathbf{in}\ LF)$$

then it follows that $\mathbf{aut}_\pi LF$ is a closed algebraic subgroup of **aut** LF which has $\mathrm{aut}_\pi LF$ as its group of rational points. Moreover a straightforward calculation yields that an element of aut F is in $\mathrm{aut}_\pi F$ if and only if its image in aut LF is contained in $\mathrm{aut}_\pi LF$ and from this it readily follows that $\mathrm{aut}_\pi F$ is of finite type.

To prove Lemma 7.3 in general, let F now be a functor as in 7.3. Then LF is a functor from the category D to the category of Lie algebras and it is easy to check that its group of self natural equivalences aut LF is the group of rational points of a linear algebraic group **aut** LF which operates on the vector space $\oplus_{d \in D} LFd$, that aut F (the group of self natural group automorphisms) is isomorphic to a subgroup of finite index of the stabilizer in aut LF of the lattice $\oplus_{d \in D} \log \overline{Fd}$ and that the subgroup $\mathrm{aut}'_\pi LF \subset \mathrm{aut}\ LF$ given by the pull back diagram

$$\begin{array}{ccc} \mathrm{aut}'_\pi LF & \longrightarrow & \mathrm{aut}\ LF \\ \downarrow & & \downarrow \\ \prod_{d \in D} \mathrm{aut}_\pi\ LFd & \longrightarrow & \prod_{d \in D} \mathrm{aut}\ LFd \end{array}$$

is the group of rational points of a closed algebraic subgroup $\mathbf{aut}'_\pi LF \subset \mathbf{aut}\ LF$.

For every object $d \in D$, now identify $M(Fd/\zeta Fd)$ with in LFd under the obvious isomorphism, lift the structure homomorphism $\psi_d : \pi \to \mathrm{out}\ Fd$ to a function $\varphi_d : \pi \to \mathrm{aut}\ Fd$ and let $p_d : \mathrm{aut}'_\pi LF \to \mathrm{aut}_\pi LFd$ denote the projection, and, for every map $f : d \to d' \in D$, denote by f_* both the induced map $M(Fd/\zeta Fd) \to M(Fd'/\zeta Fd')$ and the structure map

$$\pi \times_{\mathrm{out}\,Fd} \mathrm{aut}\ Fd \to \pi \times_{\mathrm{out}\,Fd'} \mathrm{aut}\ Fd'$$

Then we denote by $\mathrm{aut}_\pi LF \subset \mathrm{aut}'_\pi LF$ the intersection of the equalizers of the diagrams

$$\mathrm{aut}'_\pi LF \underset{(\mathrm{com}\, f_* \varphi_d y) \cdot p_{d'}}{\overset{f_* \cdot (\mathrm{com}\, \varphi_d y) \cdot p_d}{\rightrightarrows}} M(Fd'/\zeta Fd')$$

where f runs through the maps of D and y runs through the elements of π. As the obvious maps $\mathbf{M}(Fd/\zeta Fd) \to \mathbf{in}\, LFd$ are algebraic maps of unipotent groups over Q which induce isomorphisms $M(Fd/\zeta Fd) \approx \text{in}\, LFd$ on the groups of rational points, these maps are isomorphisms themselves. Using this it is not difficult to verify that $\text{aut}_\pi LF$ is the group of rational points of a closed algebraic subgroup $\mathbf{aut}_\pi LF \subset \mathbf{aut}'_\pi LF$. Once again a straightforward calculation yields that an element of $\text{aut}\, F$ is in $\text{aut}_\pi F$ of and only if its image in $\text{aut}\, LF$ is contained in $\text{aut}_\pi LF$, readily implying the desired result.

§ 8. Final reduction

We now complete the proof of Theorem 1.1 by reducing Theorem 6.1 to Lemma 7.3.

First we note that it is not difficult to verify (by obstruction theory) *that the groups* $\pi_n \text{haut}_\pi C/\Gamma_i B$ $(n \geq 1, i \geq 1)$ *are finitely generated abelian and hence of finite type.* This fact, together with 2.9, 2.10 and 6.2 readily implies that if suffices to show that *the groups* $(\text{aut}_\pi C/\Gamma_i B)_n$ *are of finite type for all* $n \geq 0$.

To do this make the harmless assumption that C_0 is free on more than one generator. Then (7.2) $B/\Gamma_i B$ is a simplicial object over the category of π-kernels and central maps between them and one can define in the obvious manner its simplicial group of automorphisms $\text{aut}_\pi B/\Gamma_i B$. Now construct a sequence of maps.

$$1 \to Z^1(\pi; \Gamma_{i-1}B_n/\Gamma_i B_n) \xrightarrow{a} (\text{aut}_\pi C/\Gamma_i B)_n \xrightarrow{b} (\text{aut}_\pi B/\Gamma_i B)_n \xrightarrow{c} H^2(\pi; \Gamma_{i-1}B_n/\Gamma_i B_n) \to 1$$

as follows (cf. §6 and [9, Ch. IV]).

(i) The map b is the homomorphism induced by the functor which (see 7.2) assigns to every epimorphism $H \to \pi$, its kernel as a π-kernel and to every "center of the kernel preserving" map between such epimorphisms, the induced central map.

(ii) For every 1-cocycle $z \in Z^1(\pi, \Gamma_{i-1}B_n/\Gamma_i B_n)$, the map $az : \Delta[n] \otimes C/\Gamma_i B \to C/\Gamma_i B$ assigns to k-simplices $p \in \Delta[n]$ and $q \in C/\Gamma_i B$, the k-simplex $p'(zq') \cdot q \in C/\Gamma_i B$, where q' denotes the image of q in π and p' is the simplicial operator such that $p = p' i_n$, where i_n denotes the non-degenerate n-simplex of $\Delta[n]$. *Clearly a is a homomorphism.*

(iii) Restriction of a simplex $r : \Delta[n] \otimes B/\Gamma_i B \to B/\Gamma_i B \in (\text{aut}_\pi B/\Gamma_i B)_n$ to the simplex $i_n \in \Delta[n]$ yields an automorphism $r' : B_n/\Gamma_i B_n \to B_n/\Gamma_i B_n$. If $r'_* k$ denotes the extension of $B_n/\Gamma_i B_n$ by π induced by r' from the given extension

$k: B_n/\Gamma_i B_n \to C_n/\Gamma_i B_n \to \pi$, then $cr \in H^2(\pi; \Gamma_{i-1}B_n/\Gamma_i B_n)$ will be the element which (see [9, Ch. IV]) sends the equivalence class of $r'_* k$ to that of k. It is not difficult to verify that *the map c so defined is a crossed homomorphism.*

A simplified version of the argument that appears in the proof of 6.2 now yields that *the above sequence is exact* and that thus our problem is reduced to showing that *the groups* $(\mathrm{aut}_\pi B/\Gamma_i B)_n$ *are of finite type for all* $n \geq 0$.

Let $c\Delta[n]$ denote the category of $\Delta[n]$ (i.e. $c\Delta[n]$ has as objects the simplices of $\Delta[n]$ and has a map $p \to q$ for every simplicial operator f such that $fp = q$) and let E be the functor from $c\Delta[n]$ to the category of π-kernels and central maps which, to each k-simplex of $\Delta[n]$, assigns $B_k/\Gamma_i B_k$. Then the group $\mathrm{aut}_\pi E$ of self natural equivalences of E is clearly isomorphic to $(\mathrm{aut}_\pi B/\Gamma_i B)_n$. Furthermore, let r be an integer such that all non-degenerate generators of C (and hence of B) are in dimensions $\leq r$, let $c^{n+r}\Delta[n] \subset C\Delta[n]$ denote the full subcategory generated by the simplices in dimensions $\leq n+r$ and let F be the restriction of E to $c^{n+r}\Delta[n]$. Then one readily verifies that the restriction $\mathrm{aut}_\pi E \to \mathrm{aut}_\pi F$ is an isomorphism and the desired result thus follows from 7.3.

BIBLIOGRAPHY

[1] G. BAUMSLAG, *Lecture notes on nilpotent groups*, Reg. Conf. Ser. Math. *2* (1970), AMS.
[2] A. BOREL and J. P. SERRE, *Corners and arithmetic groups*, Comment Math. Helv. *48* (1973), 436–491.
[3] A. K. BOUSFIELD and D. M. KAN, *Homotopy limits, completions and localizations*, Lect. Notes in Math. *304* (1972), Springer.
[4] M. DEMAZURE and P. GABRIEL, *Groupes algebriques*, North-Holland (1970).
[5] W. G. DWYER and D. M. KAN, *Function complexes in homotopical algebra*, Topology *19* (1980), 427–440.
[6] J. E. HUMPHREYS, *Linear algebraic groups*, Grad. Texts in Math. *21* (1975), Springer.
[7] D. M. KAN, *On homotopy theory and c.s.s. groups*, Ann. of Math. *68* (1958), 38–53.
[8] D. M. KAN, *A relation between CW complexes and free c.s.s. groups*, Am. J. of Math. *81* (1959), 512–520.
[9] S. MACLANE, *Homology*, Springer (1963).
[10] W. MAGNUS, A. KARRASS and D. SOLITAR, *Combinatorial group theory*, Wiley (1966).
[11] J. P. MAY, *Simplicial objects in algebraic topology*, Van Nostrand (1967).
[12] D. SULLIVAN, *Infinitesimal computations in topology*, Publ. Math. I.H.E.S. *47* (1978), 269–331.
[13] C. W. WILKERSON, *Applications of minimal simplicial groups*, Topology *15* (1976), 111–130.

Hebrew University, Jerusalem, Israel

Univ. of Notre Dame, Notre Dame, Ind.

Mass. Inst. of Techn., Cambridge, Mass.

Received July 29, 1981

Comment. Math. Helvetici **56** (1981) 615–626 0010-2571/81/004615-12$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Cyclic group actions on odd-dimensional spheres

C. KEARTON* AND S. M. J. WILSON

Abstract. We show that for any simple $(2q-1)$-knot k, $q>1$, and any positive integer n, the knot $\#_1^n k$ is the fixed-point set of a $\mathbf{Z}_n$-action on S^{2q+1}. Further, we show that for many values of n there are examples of $(2q-1)$-knots, $q\geq 2$, which are the fixed-point sets of inequivalent $\mathbf{Z}_n$-actions.

0. Introduction

An n-knot is a locally-flat PL pair (S^{n+2}, S^n), where S^n denotes the n-sphere. A $(2q-1)$-knot is *simple* if the complement of S^{2q-1} has the homotopy type of a circle up to but not including dimension q. For $q>1$ such knots have been classified in [L] in terms of the S-equivalence classes of their Seifert matrices, and in [K, T1, T2] in terms of their Blanchfield pairings. Using these classification results, for any simple $(2q-1)$-knot k, with $q>1$, and for any positive integer n, we construct a simple $(2q-1)$-knot k_n such that the n-fold cyclic cover of S^{2q+1} branched over k_n is again S^{2q+1}, and such that k_n lifts to $\#_1^n k$, the sum of n copies of k. An immediate corollary is that for any such k and n, there is a $\mathbf{Z}_n$-action on S^{2q+1} with fixed point set $\#_1^n k$.

The construction in this paper is purely algebraic, and may be contrasted with the geometric construction in [G], where for any m-knot k ($m\geq 2$) Gordon constructs an m-knot which is the fixed-point set of a $\mathbf{Z}_n$-action and whose fundamental group is isomorphic to that of $\#_1^n k$.

As an application of our construction we are able for many values of n to find examples of $(2q-1)$-knots which are the fixed-point sets of inequivalent $\mathbf{Z}_n$-actions. The technique is to pick simple $(2q-1)$-knots k and l such that $\#_1^n k = \#_1^n l$, and such that $k_n \neq l_n$.

1. The main construction

Let k be a simple $(2q-1)$-knot, $q>1$, and $n>1$ an integer. Let A be a non-singular Seifert matrix of k, and set $\varepsilon=(-1)^q$. Following Trotter [T1], we set $S=(A+\varepsilon A')^{-1}$, $T=-\varepsilon A' A^{-1}$.

* This paper was written whilst the first author was in receipt of a Research Grant from the Science Research Council of Great Britain.

PROPOSITION 1.1. *The pair* (S, T) *has the following properties.*

(i) S *is integral, unimodular,* ε*-symmetric.*
(ii) $(I-T)^{-1}$ *exists and is integral.*
(iii) $T'ST = S$.
(iv) $A = (I-T)^{-1}S^{-1}$.

Moreover, any pair of rational matrices (S, T) satisfying (i)–(iii) yields a Seifert matrix A by the formula (iv).

Proof. It is well known (see [L], [T1]) that $A+\varepsilon A'$ is unimodular, and so S is integral and unimodular. Clearly S is ε-symmetric.

Now $I-T = I+\varepsilon A'A^{-1} = (A+\varepsilon A')A^{-1} = S^{-1}A^{-1}$, from which (ii) and (iv) follow at once. Property (iii) is easily checked.

Now suppose that we are given a pair of rational matrices (S, T) satisfying (i)–(iii); then we can define the matrix $A = (I-T)^{-1}S^{-1}$, which by (i) and (ii) is a non-singular matrix over the integers. We have

$$\begin{aligned} A+\varepsilon A' &= (I-T)^{-1}S^{-1} + \varepsilon(S')^{-1}(I-T')^{-1} \\ &= (I-T)^{-1}S^{-1} + S^{-1}(I-ST^{-1}S^{-1})^{-1} \qquad \text{by (i), (iii)} \\ &= (I-T)^{-1}S^{-1} + S^{-1}S(I-T^{-1})^{-1}S^{-1} \\ &= [(I-T)^{-1} - (I-T)^{-1}T]S^{-1} = (I-T)^{-1}(I-T)S^{-1} = S^{-1} \end{aligned}$$

which is unimodular. It follows that A is a Seifert matrix. □

Now we define matrices U, V by

$$U = \begin{bmatrix} 0 & \cdots & 0 & T \\ I & & & 0 \\ & \ddots & & \vdots \\ 0 & & I & 0 \end{bmatrix}, \quad V = \begin{bmatrix} S & & 0 \\ & \ddots & \\ 0 & & S \end{bmatrix},$$

there being $n \times n$ blocks in each case.

THEOREM 1.2. *The pair* (V, U) *determines a simple* $(2q-1)$*-knot* k_n. *The* n*-fold branched cyclic cover of* k_n *is the knot* $\#_1^n k = k + \cdots + k$ (n *times*).

Proof. We have to check that the pair (V, U) satisfies conditions (i)–(iii) of

Proposition 1.1. Clearly V satisfies (i), and it is easy to check that

$$(I-U)^{-1}=\begin{pmatrix}(I-T)^{-1} & T(I-T)^{-1} & \cdots & T(I-T)^{-1}\\ \vdots & \ddots & \ddots & \vdots\\ & & & T(I-T)^{-1}\\ (I-T)^{-1} & \cdots & \cdots & (I-T)^{-1}\end{pmatrix}.$$

But $T(I-T)^{-1}=-\varepsilon A'A^{-1}(I+\varepsilon A'A^{-1})^{-1}=-\varepsilon A'(A+\varepsilon A')^{-1}$, which is an integer matrix. Hence (ii) is satisfied. To check (iii) is a simple matrix multiplication.

Hence (V, U) determine a unique simple $(2q-1)$-knot k_n. A routine computation shows that

$$U^n=\begin{pmatrix}T & & 0\\ & \ddots & \\ 0 & & T\end{pmatrix},$$

and hence the pair (V, U^n) satisfies (i)–(iii), and in fact represents the knot $\#_1^n k$.

Let K_n denote the complement of k_n, and $\tilde{K}_n$ the infinite cyclic cover of K_n. If u is a generator of the group of covering translations, then K_n is obtained from $\tilde{K}_n$ by quotienting out by the action of u. Similarly the n-fold cyclic cover of K_n is obtained from $\tilde{K}_n$ by quotienting out by the action of u^n.

Algebraically this can be described as follows, using Trotter's description of $H_q(\tilde{K}_n)$ in [T1]. Let $\mathbf{B}$ be a basis of $\mathbf{Q}^m$ corresponding to (V, U) where T is an $r\times r$ matrix; then $H_q(\tilde{K}_n)$ is the $\mathbf{Z}[u, u^{-1}]$-module generated by $\mathbf{B}$, the action of u being given by U. The fact that $(1-u): H_q(\tilde{K}_n)\to H_q(\tilde{K}_n)$ is an isomorphism means that when we quotient out by the action of u we get a homology circle. But the form of U^n means that $(1-u^n): H_q(\tilde{K}_n)\to H_q(\tilde{K}_n)$ is also an isomorphism, and hence the n-fold cyclic cover of K_n is a homology circle. Therefore the n-fold branched cyclic cover of k_n is a homotopy sphere, and hence a sphere. □

COROLLARY 1.3. *If k is a simple $(2q-1)$-knot, $q>1$, then $\#_1^n k$ is the fixed point set of a $\mathbf{Z}_n$-action on S^{2q+1}.*

PROPOSITION 1.4. *Let B be the Seifert matrix of k_n corresponding to (V, U). Then*

$$B=\begin{pmatrix}A & -\varepsilon A' & \cdots & -\varepsilon A'\\ \vdots & \ddots & \ddots & \vdots\\ & & & -\varepsilon A'\\ A & \cdots & \cdots & A\end{pmatrix}.$$

Proof:

$$B = (I-U)^{-1}V^{-1} = \begin{pmatrix} (I-T)^{-1} & T(I-T)^{-1} & \cdots & T(I-T)^{-1} \\ \vdots & & & \\ & & & T(I-T)^{-1} \\ (I-T)^{-1} & \cdots & & (I-T)^{-1} \end{pmatrix} \begin{pmatrix} S^{-1} & & \\ & & 0 \\ & & \\ 0 & & S^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} (I-T)^{-1}S^{-1} & T(I-T)^{-1}S^{-1} & \cdots & T(I-T)^{-1}S^{-1} \\ \vdots & & & \\ & & & T(I-T)^{-1}S^{-1} \\ (I-T)^{-1}S^{-1} & \cdots & & (I-T)^{-1}S^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} A & TA & \cdots & TA \\ \vdots & & & \\ & & & TA \\ A & \cdots & & A \end{pmatrix}$$

$$= \begin{pmatrix} A & -\varepsilon A' & \cdots & -\varepsilon A' \\ \vdots & & & \\ & & & -\varepsilon A' \\ A & \cdots & & A \end{pmatrix} \square$$

Next we prove a result which relates an Alexander matrix of k to one for k_n. Recall that an Alexander matrix $M(t)$ of k is a matrix over $\mathbf{Z}[t, t^{-1}]$ which presents $H_q(\tilde{K})$ as a $\mathbf{Z}[t, t^{-1}]$-module; that is, there is an exact sequence of $\mathbf{Z}[t, t^{-1}]$-modules

$$F \xrightarrow{M(t)} G \longrightarrow\!\!\!\!\!\rightarrow H_q(\tilde{K})$$

where F and G are free $\mathbf{Z}[t, t^{-1}]$-modules.

PROPOSITION 1.5. *Let $M(t)$ be a square Alexander matrix for the knot k; then $M(t^n)$ is an Alexander matrix for k_n.*

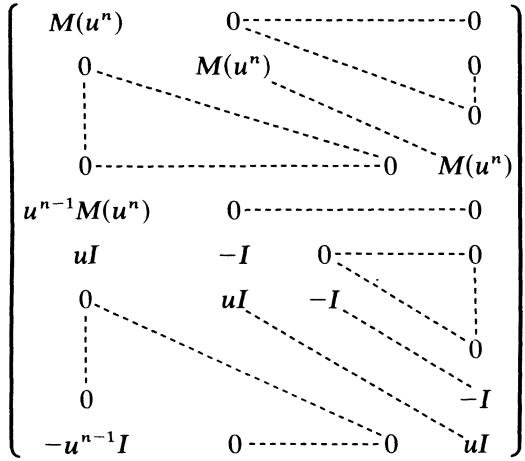
Proof. We can describe the $\mathbf{Z}[u, u^{-1}]$-module structure of $H_q(\tilde{K}_n)$ in the following way. Let $L_1, \ldots, L_n$ be n copies of the $\mathbf{Z}[t, t^{-1}]$-module $H_q(\tilde{K})$. Then for $1 \leq i < n$, $u : L_i \to L_{i+1}$ is a $\mathbf{Z}[t, t^{-1}]$-isomorphism, and $u : L_n \to L_1$ is defined so that $u^n : L_1 \to L_2 \to \cdots \to L_n \to L_1$ coincides with $t : L_1 \to L_1$. Thus a presentation matrix for $H_q(\tilde{K}_n)$ as a $\mathbf{Z}[u, u^{-1}]$-module is

$$\begin{bmatrix}
M(u^n) & 0 & \cdots & 0 \\
0 & & & 0 \\
\vdots & & & \vdots \\
0 & \cdots & 0 & M(u^n) \\
uI & -I & 0 & \cdots \\
0 & uI & -I & 0 \\
0 & & & 0 \\
\vdots & & & \vdots \\
0 & \cdots & 0 & -I \\
-u^n I & 0 & \cdots & uI
\end{bmatrix}.$$

Two elementary row operations give

$$\begin{bmatrix}
M(u^n) & 0 & \cdots & 0 \\
0 & & & \vdots \\
\vdots & & & M(u^n) \\
0 & \cdots & 0 & 0 \\
u^{n-1}M(u^n) & 0 & \cdots & 0 \\
uI & -I & 0 & \cdots \\
0 & uI & -I & 0 \\
\vdots & & & \\
-u^{n-1}I & 0 & \cdots & uI \\
-u^n I & 0 & \cdots & 0
\end{bmatrix}
\begin{matrix} 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \\ 0 \\ uI \end{matrix}.$$

We eliminate the final row and column to give



Now subtract u^{n-1} times the first row from the n^{th} to obtain a row of zeros, which may be eliminated. Continuing in this way we eventually arrive at the matrix $M(u^n)$. □

THEOREM 1.6. *The knot k_n depends only on k and n, and not upon the choice of Seifert matrix A.*

Proof. Let A be an $r \times r$ matrix, and let $\Lambda = \mathbf{Z}[t, t^{-1}, (1-t)^{-1}]$, a subring of the field $\mathbf{Q}(t)$, the field of rational functions in one variable over the rationals. According to Trotter's viewpoint [T1], k gives rise to an ε-symmetric bilinear form [,] on $\mathbf{Q}^r$ represented by the matrix S, and a Λ-module M contained in $\mathbf{Q}^r$ where the action of t is represented by T. A choice of Seifert matrix corresponds to an admissible lattice contained in M (see [T1] for definitions). Although our construction is given in terms of the matrices S and T, it is clear that it could be phrased in terms of M and [,], and hence that it does not depend upon the choice of A.

Alternatively, one can use the formula of Proposition 1.4 to show that if A is S-equivalent to A_1, then B is S-equivalent to B_1. □

2. Knots having distinct $\mathbf{Z}_n$-actions, n odd

In this section we shall show that for many odd integers n, there exist simple $(4q+1)$-knots $(q \geq 1)$, k and l, such that $\#_1^n k = \#_1^n l$ but $k_n \neq l_n$.

Let $\lambda_m(t)$ denote the m^{th} cyclotomic polynomial, where m is not a prime

power. Let ζ be a primitive m^{th} root of unity, $K = \mathbf{Q}(\zeta)$ and $F = \mathbf{Q}(\zeta + \zeta^{-1})$, the fixed field of K under complex conjugation. Let h_K denote the class number of K, h_F that of F, and $h_- = h_K/h_F$. According to the work of Bayer [Ba: Example 6.2], the number of distinct simple $(4q+1)$-knots $(q \geqslant 1)$ with Alexander polynomial $\lambda_m(t)$ is $h_- 2^d$ if $m = 2p^\alpha$ and $h_- 2^{d-1}$ otherwise, where $2d = [K:\mathbf{Q}]$. The factor h_- represents the number of isomorphism classes of $\mathbf{Z}[t, t^{-1}]$-modules supporting a Blanchfield pairing [Ba: Corollary 1.3], and the factor $2^d(2^{d-1})$ represents the number of non-isometric pairings which a given module supports. Note that Bayer's work is couched in terms of pairings on $\mathbf{Z}[\zeta]$-modules which are hermitian with respect to complex conjugation ($t \to t^{-1}$ becomes $\zeta \to \zeta^{-1} = \bar{\zeta}$), and we shall adopt this viewpoint.

Let U be the group of units of (the ring of integers of) K, U_0 the group of units of F, and $N: K \to F$ the norm. If I is a principal ideal, then let $\langle u \rangle$ denote the hermitian form h on I given by $h(a, b) = ua\bar{b}$. As in [Ba: Prop. 2.1], the set of isometry classes of unimodular hermitian forms on a given ideal (not necessarily principal) is in one-one correspondence with $U_0/N(U)$.

Now suppose that h_- has a factor $n > 1$, where n is odd and $(m, n) = 1$. Let a be an ideal of $\mathbf{Q}(\zeta)$ admitting a non-singular hermitian form h, with a being of order n in $\ker N: C_K \to C_F$. Then $\perp_1^n(a, h)$ has determinant $\langle u \rangle$ for some $u \in U_0/N(U)$; see [Ba: Definition 1.9] for the definition of determinant. Since the order of $U_0/N(U)$ is 2^d or 2^{d-1}, and n is odd, there exists $v \in U_0/N(U)$ such that $v^n = u$. Then $\perp_1^n \langle v \rangle$ has determinant $\langle v^n \rangle = \langle u \rangle$.

Set $K = (a, h) \perp (a, -h)$, $L = \langle v \rangle \perp \langle -v \rangle$. Then $\perp_1^n K$, $\perp_1^n L$ are indefinite and have the same rank, signatures and determinant. Hence by [Ba: Corollary 4.10] they are isometric. But K is not isometric to L, for the determinant of K is (a^2, α), and a^2 is non-zero in $\ker N: C_K \to C_F$ since n is odd.

In fact, if k, l are the simple $(4q+1)$-knots corresponding to K, L respectively, we can show that $k_n \neq l_n$. For let $M(t)$ be an Alexander matrix of k, so that by Proposition 1.5 $M(t^n)$ is an Alexander matrix of k_n. The work of Fox and Smythe [F–S] enables us to obtain a row ideal class from the matrix $M(\zeta)$, and the work of Hillman [H: Chap. III, Theorem 12] identifies this with the ideal a^2 in the determinant of K. But the Alexander polynomial of k_n is $\lambda_m(t^n)$, which has $\lambda_m(t)$ as one of its factors since $(m, n) = 1$. Let τ be a primitive m^{th} root of unity such that $\tau^n = \zeta$. Setting $t = \tau$ in the Alexander matrix $M(t^n)$, we obtain $M(\tau^n) = M(\zeta)$, and hence obtain a Fox-Smythe invariant a^2 again. In the case of l_n, these ideal invariants are all trivial, hence $k_n \neq l_n$.

Taking the n-fold branched cyclic covers of k_n, l_n we obtain respectively the knots $\#_1^n k$, $\#_1^n l$. Since $\perp_1^n K$ is isometric to $\perp_1^n L$, we have $\#_1^n k = \#_1^n l$.

Many examples may be obtained from the tables in [Sch].

For the case of $(4q-1)$-knots, $q \geq 1$, and $m \neq 2p^r, p^r$, where p is a prime, then as in [Ba: §5], $\zeta - \zeta^{-1}$ is a unit and so we can multiply all the pairings above by $\zeta - \zeta^{-1}$ to obtain skew-hermitian pairings. The argument then goes through as before. We are grateful to Dr. Bayer for pointing out this extension to the case of $(4q-1)$-knots.

3. Number theory

This section deals with some results from algebraic number theory, which will be used in the next section to deal with the case $n = 2$.

Let K be an algebraic number field, $R = \text{int}(K)$ its ring of integers, $\mathbf{Z}_{(p)}$ the p-adic integers, $R_p = R \otimes \mathbf{Z}_{(p)}$, $K_p = K \otimes \mathbf{Z}_{(p)}$, $U(R) = \prod_P R_p^\times$ and $J(K) = U(R) \cdot \coprod_p K_p^\times$, where $\coprod$ denotes the direct sum. $K^\times$ is considered as a subgroup of $J(K)$ under the "diagonal" map. If $C(K)$ denotes the ideal class group of K, then we have $C(K) \cong J(K)/U(R) \cdot K^\times$ an isomorphism which is natural with respect to ring extensions.

Now suppose that L is an algebraic number field, Γ a group of automorphisms of L, $S = \text{int}(L)$, $K = L^\Gamma$ the subfield of L fixed under Γ, and $R = \text{int}(K) = S^\Gamma$.

LEMMA 3.1. $\ker[C(R) \to C(S)] \cong \ker[H^1(\Gamma, S^\times) \to H^1(\Gamma, U(S))]$, *where the first map is induced by ring extension, the second by the "diagonal" map* $S^\times \to U(S)$.

Proof. Consider the exact sequence

$$0 \to U(S) \cdot L^\times \to J(L) \to C(S) \to 0,$$

Since $J(L)^\Gamma = J(K)$, we obtain a commutative diagram

$$\begin{array}{ccccccc} 0 \to & U(R) \cdot K^\times & \to & J(K) & \to & C(R) & \to 0 \\ & \downarrow & & \| & & \downarrow & \\ 0 \to & (U(S) \cdot L^\times)^\Gamma & \to & J(L)^\Gamma & \to & C(S)^\Gamma & \end{array}$$

Applying the Snake Lemma [Bass: p. 26] we find that

$$\begin{aligned} \ker[C(R) \to C(S)] &= \ker[C(R) \to C(S)^\Gamma] \\ &\cong \text{coker}[U(R) \cdot K^\times \to (U(S) \cdot L^\times)^\Gamma]. \end{aligned}$$

Now consider the exact sequence

$$0 \to S^\times \to U(S) \oplus L^\times \to U(S) \cdot L^\times \to 0$$

where the first map is $s \mapsto (s, s^{-1})$. From cohomology theory we obtain the exact sequence

$$0 \to R^\times \to U(R) \oplus K^\times \to (U(S) \cdot L^\times)^\Gamma \to H^1(\Gamma, S^\times) \to H^1(\Gamma, U(S)) \oplus H^1(\Gamma, L^\times).$$

Since by Hilbert 90, $H^1(\Gamma, L^\times) = 0$, we have

$$\begin{aligned} \operatorname{coker}[U(R) \oplus K^\times \to (U(S) \cdot L^\times)^\Gamma] &= \operatorname{coker}[U(R) \cdot K^\times \to (U(S) \cdot L^\times)^\Gamma] \\ &\cong \ker[H^1(\Gamma, S^\times) \to H^1(\Gamma, U(S))], \end{aligned}$$

and the result follows. □

Now let

$$L = \mathbf{Q}(\sqrt{-123}, \sqrt{-31}) \qquad S = \text{int}(L)$$

$$K = \mathbf{Q}(\sqrt{-123}) \qquad \Gamma = \text{Gal}(L/K), \quad R = \text{int}(K)$$

$$K' = \mathbf{Q}(\sqrt{3813}) \qquad \Gamma' = \text{Gal}(L/K') \quad R' = \text{int}(K').$$

The action of the non-trivial elements of Γ, Γ' will be denoted respectively by $\bar{}$, $\bar{}$. Our immediate purpose is to show that $C(R) \to C(S)$ is injective.

LEMMA 3.2. *The fundamental unit of* R' *is* $v = 247 + 4\sqrt{3813}$.

Proof. Certainly $247^2 - 16.3813 = 1$, so v is a unit of R'. If v is not the fundamental unit, then there exist positive integers a, b, c, d such that $(a + b\sqrt{N})(c + d\sqrt{N}) = 4v$, where $N = 3813$. Thus

$$ac + bdN + (ad + bc)\sqrt{N} = 4(247 + 4\sqrt{N}).$$

But $ac + bdN \geq N > 4.247$, so this is impossible. □

By the Dirichlet Unit Theorem, rank $(S^\times) = 1$ and $S^\times = \langle \pm 1 \rangle \times \langle u \rangle$ for some u (a fundamental unit).

LEMMA 3.3. *S has $u=\sqrt{-123}+2\sqrt{-31}$ as a fundamental unit.*

Proof. Note that $u\bar{u}=v$, so $u\in S^{\times}$. If $u=\pm w^n$ for some $w\in S$ (± 1 are the only units of finite order) then $w\bar{w}\in R'^{\times}$ and so $w\bar{w}=\pm v^m$ for some $m\in\mathbf{Z}$. But then $(v^m)^n=\pm v$ whence $m=n=\pm 1$. Hence the result. □

LEMMA 3.4. *$H^1(\Gamma, S^{\times})\to H^1(\Gamma, S_{31}^{\times})$ is injective, and hence so is $H^1(\Gamma, S^{\times})\to H^1(\Gamma, U(S))$.*

Proof. For an abelian Γ-group A we use the representation

$$H^1(\Gamma, A)\cong\frac{\{a\in A : a\tilde{a}=1\}}{\{a/\tilde{a} : a\in A\}}.$$

This representation is natural with respect to extension of A. Now

$$u\tilde{u}=(\sqrt{-123}+2\sqrt{-31})(\sqrt{-123}-2\sqrt{-31})=-123+124=1,$$

and $u/\tilde{u}=u^2/u\tilde{u}=u^2$ so

$$H^1(\Gamma, S^{\times})=S^{\times}/\langle u^2\rangle=\{(1),(-1),(u),(-u)\}.$$

We must show that none of -1, u, $-u$ is of the form $s/\tilde{s}$ for some $s\in S_{31}^{\times}$.

If for some $s\in S_{31}^{\times}$, $s/\tilde{s}=-1$, then $s=-\tilde{s}$ and so $s=r\sqrt{-31}$ for some $r\in R_{31}$. Hence s is not a unit.

If for some $s\in S_{31}^{\times}$, $s/\tilde{s}=u$, then, as $S_{31}=R_{31}[u]$, $s=a+bu$ with $a, b\in R_{31}$, and so $a+bu=(a+b\tilde{u})u=au+b$. Hence $a=b$ and $s=a(1+u)$. As s is a unit, $N_{L/\mathbf{Q}}(s)$ is also a unit. But $N_{L/\mathbf{Q}}(s)=(a\bar{a})^2(1+u)(1+\tilde{u})(1+\bar{u})(1+\bar{\tilde{u}})=(a\bar{a})^2\cdot 16.31$ and this is not a unit in $\mathbf{Z}_{31}$.

A similar argument disposes of the possibility $s/\tilde{s}=-u$ and the result is proved. □

Remark 3.5. We can now see that the primes of S above 31 are principal. By our calculation above, $N_{L/\mathbf{Q}}((1+u)/2)=31$ and so, since $(1+u)/2$ is integral ($N_{L/K}((1+u)/2=(1+\sqrt{-123})/2\in R$), $((1+u)/2)_S$ is a prime of S above 31. But $L/\mathbf{Q}$ is galois so all the primes of S above 31 are conjugate and hence principal. In fact, $(31)_S=((1+u)/2)_S^2((1-u)/2)_S^2$.

PROPOSITION 3.6. *$(3, \sqrt{-123})_{S[1/31]}$ is not principal.*

Proof. $(3, \sqrt{-123})_R$ is clearly not principal, for the equation $a^2+123b^2=12$ has no solutions over $\mathbf{Z}$. Since $C(R)\to C(S)$ is injective by Lemmas 3.1 and 3.4, $(3, \sqrt{-123})_S$ is not principal.

In passing from $C(S)$ to $C(S[1/31])$ we kill off the ideal classes represented by ideals dividing $(31)_S$; since by Remark 3.5 these ideals are all principal, the map $C(S) \to C(S[1/31])$ is injective. The result follows. □

4. The case $n = 2$

In this section we construct two simple $(4q+1)$-knots k and l such that $k + k = l + l$ but $k_2 \neq l_2$.

LEMMA 4.1. *Let* $\Delta(t) = 31t^2 - 61t + 31$. *Then* $\mathbf{Q}(\sqrt{-123}, \sqrt{-31})$ *is a splitting field for* $\Delta(t^2)$.

Proof. Let τ be a root of $\Delta(t^2)$; then we can take $\tau^2 = (61 + \sqrt{-123})/62$, so that $\Delta(t)$ splits in $\mathbf{Q}(\sqrt{-123})$. Now $31\tau^2 = (61 + \sqrt{-123})/2 = -[(1 - \sqrt{-123})/2]^2$, and so $\tau = (1 - \sqrt{-123})/2\sqrt{-31}$. Hence $\tau \in \mathbf{Q}(\sqrt{-123}, \sqrt{-31})$. But the conjugates of τ are τ, $-\tau$, $\bar{\tau} = 1/\tau$ and $-1/\tau$, so $\Delta(t^2)$ splits in $\mathbf{Q}(\sqrt{-123}, \sqrt{-31})$. □

Let J denote the ideal $(3, \sqrt{-123})$ over the ring $\mathbf{Z}[\tau^2, \tau^{-2}] = R[1/31]$ in the notation of Section 3. Note that $J = \bar{J}$ and $J\bar{J} = (3)$, where $\bar{}$ here denotes complex conjugation. Hence we can define a non-singular hermitian form $b : J \times J \to R[1/31]$ by $b(\alpha, \beta) = \alpha\bar{\beta}/3$. Let $(J \oplus J, B)$ denote the orthogonal direct sum $(J, b) \perp (J, b)$, and set

$$e = ((6 + \sqrt{-123})/31, (51 + \sqrt{-123})/31)$$

$$f = ((51 - \sqrt{-123})/31, (-6 + \sqrt{-123})/31).$$

It is easily checked that $B(e, e) = B(f, f) = 1$ and that $B(e, f) = 0$. Hence $(J, b) \perp (J, b) \cong \langle 1 \rangle \perp \langle 1 \rangle$.

Let k be the simple $(4q+1)$-knot $(q \geq 1)$ represented by (J, b) and l the corresponding knot represented by $\langle 1 \rangle$. Then $k + k = l + l$, but since J is a non-principal ideal by Proposition 3.6, $k \neq l$. Let $M(t)$ be a square Alexander matrix for k; then by Proposition 1.5, $M(t^2)$ is an Alexander matrix for k_2. The Fox–Smythe row ideal class of k_2 is obtained from the matrix $M(\tau^2)$ over the ring $\mathbf{Z}[\tau, \tau^{-1}] = S[1/31]$, and by [H: Chap. III, Theorem 12] this is the ideal $J_{S[1/31]}$. By Proposition 3.6, this ideal is non-principal. Since the corresponding invariant for l_2 is trivial, we have $k_2 \neq l_2$.

REFERENCES

[Bass] H. BASS. *Algebraic K-theory.* W. A. Benjamin, Inc. (1968).

[Ba] E. BAYER. *Unimodular hermitian and skewhermitian forms.* Jour. Algebra (to appear).

[B] R. C. BLANCHFIELD. *Intersection theory of manifolds with operators with applications to knot theory.* Annals of Math. *65* (1957), 340–356.

[F–S] R. H. FOX and N. SMYTHE. *An ideal class invariant of knots.* Proc. Amer. Math. Soc. *15* (1964), 707–709.

[G] C. MCA. GORDON. *On the higher-dimensional Smith conjecture.* Proc. London Math. Soc. (3) *29* (1974), 98–110.

[H] J. A. HILLMAN. *Alexander ideals.* Springer Lecture Notes 895.

[K] C. KEARTON. *Classification of simple knots by Blanchfield duality.* Bull. Amer. Math. Soc. *79* (1973), 952–955.

[L] J. LEVINE. *An algebraic classification of some knots of codimension two.* Comment. Math. Helv. *45* (1970), 185–198.

[Sch] G. SCHRUTKA V. RECHTENSTAMM. *Tabelle der (relativ-) Klassenzahlen von Kreiskörpern.* Abh. Deutsch Akad. Wiss. Berlin 1964, Math. Nat. Kl. Nr 2.

[T1] H. F. TROTTER. *On S-equivalence classes of Seifert matrices.* Invent. math *20* (1973), 173–207.

[T2] ——. *Knot modules and Seifert matrices.* Knot Theory, ed. J.-C. Hausmann, Lecture Notes in Mathematics *685* (1978), Springer-Verlag.

Department of Mathematics
University of Durham
Durham, DH1 3LE, England.

Received September 23, 1981

Note added in proof: The second author has recently shown that for any integer n there is an integer m, prime to n and not a prime power, such that, if ζ is an mth root of 1, there is an ideal class in $C_{\mathbb{Q}[\zeta]}$ of order n with norm 1 in $C_{\mathbb{Q}[\zeta+\zeta^{-1}]}$. Thus the results of section 2 are valid for any odd n.

Comment. Math. Helvetici **56** (1981) 627–637 0010–2571/81/004627–11$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Depth and equivariant cohomology

J. DUFLOT

0. Introduction

Let G be a finite group and let X be a space on which G acts continuously. Choose a classifying bundle $PG \to BG$ for principal G-bundles. The group G acts freely on the contractible space PG, and there is a diagonal action of G on $PG \times X$. Let $PG \times^G X$ denote the orbit space of this diagonal action.

Let p be a prime integer. The mod-p equivariant cohomology ring of the G-space X is defined by the formula

$$H_G^*(X, Z/pZ) = H^*(PG \times^G X, Z/pZ).$$

The main result of this paper is:

THEOREM 1 (see Section 2). *The depth of $H_G^*(X, Z/pZ)$ is greater than or equal to the maximum rank of a central p-torus acting trivially on X.*

Here a *p-torus* is a direct product of cyclic groups of order p, and the *rank* of a p-torus A is the number of cyclic factors of A.

In a series of papers (Q1, Q2) D. Quillen investigated the algebraic structure of this ring. For example, suppose X has finite-dimensional mod-p cohomology. In this case Quillen proves the following

THEOREM (Theorem 7.7 of (Q1)). *The Krull dimension of the commutative ring*

$$H_G(X, Z/pZ) = \begin{cases} H_G^{ev}(X, Z/pZ) & p \text{ odd} \\ H_G^*(X, Z/pZ) & p = 2 \end{cases}$$

is equal to the maximum rank of a p-torus A of G such than $X^A \neq \emptyset$.

In Section 4 of this paper, we use Theorem 1 and this dimension theorem of

Quillen's to show that certain cohomology modules are Cohen-Macaulay (Theorem 2).

Most of this work is contained in my thesis, written at M.I.T. under the guidance of Daniel Quillen. He suggested that I study these cohomology rings from a commutative-algebraic point of view; he also gave me most of my general insights into the subject and most of my specific methods of attack. I am very grateful to him for this guidance.

1. Preliminaries

The purpose of this preliminary section is to state some basic definitions and results, and to set notation.

Let G be a finite group. There is a classifying bundle $PG \to BG$ for principal G-bundles with paracompact base. The spaces PG and BG may be assumed to be (paracompact) CW-complexes.

Suppose that G acts continuously on a topological space X. Following Borel (B) we define

$$X_G = PG \times^G X$$

to be the orbit space of the diagonal action of G on $PG \times X$. We assume that the space X is such that X_G is a paracompact, locally contractible, Hausdorff space. For example, take X to be locally compact, paracompact, locally contractible and Hausdorff.

If R is a commutative ring, define the equivariant cohomology ring of the G-space X with coefficient in R to be

$$H_G^*(X, R) = H^*(X_G, R),$$

where the right hand side of this equation is ordinary singular cohomology with coefficients in R. The restrictions on X enable us to say that this definition of equivariant cohomology agrees with that of Borel (B) and Quillen (Q1) (these authors use sheaf cohomology); so we may use some results of their work.

We will make use of the following properties:

(a) (Q1, Section 1) $H_G^*(X, R)$ is independent of the choice of classifying bundle for G.

(b) (Q1, (1.5)) Functoriality: If $u: G \to G'$ is a homomorphism of finite groups and $f: X \to X'$ is u-equivariant, then there is a homomorphism $(u, f)^*: H_{G'}^*(X', R) \to H_G^*(X, R)$. If f and u are inclusions, this homomorphism will be denoted "res".

(c) If $X = \mathrm{pt}$ is a point, then $H_G^*(\mathrm{pt}, R) = H^*(PG \times^G X, R) = H^*(BG, R)$. In other words, $H_G^* = H_G^*(\mathrm{pt}, R)$ is classical group cohomology with coefficients in the trivial G-module R.

2. A regular sequence in $H_G^*(X, Z/pZ)$

In this section, the cohomology groups have coefficients in Z/pZ, where p is a fixed prime, unless otherwise indicated.

Let

$$H = \begin{cases} H_G^{\mathrm{ev}} = \bigoplus_{i \geq 0} H_G^{2i}(\mathrm{pt}) & p \text{ odd} \\ H_G^* = \bigoplus_{i \geq 0} H_G^i(\mathrm{pt}) & p = 2. \end{cases}$$

The ring H is a commutative ring. The graded group $H_G^*(X)$ may be considered as an H-module via the map $X \to \mathrm{pt}$. An *H-sequence* on $M = H_G^*(X)$ (or on any H-module M) may be defined in the following way (K):

A sequence of elements $x_1, x_2, \ldots, x_N, \ldots$ of positive degree in H is said to be an *H-sequence* on M (or a *regular* sequence on M) if x_1 is not a zero divisor on M, and if for each $i > 1$, x_i is not a zero divisor on $M/(x_1, \ldots, x_{i-1})M$.

Let $n > 0$. For each i such that $1 \leq i < n$, the sequence of elements $x_1, \ldots, x_n$ of H is an H-sequence on $H_G^*(X)$ if and only if $x_1, \ldots, x_i$ is an H-sequence on $H_G^*(X)$ and $x_{i+1}, \ldots, x_n$ is an H-sequence on $H_G^*(X)/(x_1, \ldots, x_i)H_G^*(X)$.

Theorems of Evens (E) and Venkov (V, see also (Q1)) show that H is Noetherian and that $H_G^*(X)$ is a finitely generated H-module if $H^*(X)$ is finite-dimensional over Z/pZ. In this case, any two maximal H-sequences have the same (finite) length (e.g. see (K); in Theorem 121 of (K), take the ideal I to be the positive degree elements of H). This common length we call the *depth* of $H_G^*(X)$.

Here is the main theorem.

THEOREM 1. *Let A be a p-torus that contained in the center of G. Suppose also that A acts trivially on X. Then there is a regular sequence on $H_G^*(X)$ of length equal to rank (A). Thus, if $H^*(X)$ is finite dimensional over Z/pZ, then*

$$\textit{depth } H_G^*(X) \geq \textit{rank } (A).$$

This theorem will be proved by induction on the rank of A. Before we give the proof of Theorem 1, we need some preliminary results.

Let A be a cyclic group of order p contained in the center of G, such that $X^A = X$. $(X^A = \{x \in X \mid ax = x \ \forall a \in A\})$. Consider the representation $\rho : A \to \mathbf{C}^*$ given by $\rho(a) = \exp(2\pi i/p)$, where a is a fixed generator for A. The representation ρ of A gives an A-action on $\mathbf{C}$. Using this action we have a line bundle, also called ρ, over the classifying space BA for A (as in (A)):

$$PA \times^A \mathbf{C} \xrightarrow[\rho]{} BA.$$

The first Chern class for this bundle is $c_1(\rho) \in H^2_A(\mathrm{pt}, Z) = H^2(BA, Z)$; via the homomorphism $Z \to Z/pZ$, we have a mod-p Chern class for ρ, $c_1(\rho) \in H^2(BA) = H^2_A$.

Since A acts trivially on X, there is an "equivariant" Lyndon–Hochschild–Serre spectral sequence (see Section 3):

$$E_2^{**} = H^*(X_{G/A}, \{H^*_A\}) \Rightarrow H^*(X_G).$$

Since A is central in G, and coefficients are in a field,

$$E_2^{p,q} = H^p(X_{G/A}, H^q_A) \simeq H^p(X_{G/A}) \otimes_{Z/pZ} H^q_A \quad \text{(see Section 3)}.$$

It is well known that

$$H^*_A = \begin{cases} Z/pZ(c_1(\rho)) \otimes_{Z/pZ} \Lambda(x) & p \text{ odd} \\ & \text{(a polynomial algebra on } c_1(\rho) \text{ tensored with} \\ & \text{an exterior algebra on } x \text{ where the Bockstein} \\ & \text{of } x \text{ equals } c_1(\rho).) \\ Z/2Z(y) & p = 2 \\ & \text{(a polynomial algebra on } y, \text{ where } y \cdot y = c_1(\rho).) \end{cases}$$

Following Evens (E) we have:

LEMMA 1. *Let* $\alpha \in H^{2M}_G(X)$ *be any cohomology class such that* $H^{2M}_G(X) \xrightarrow[\text{res}]{} H^{2M}_A = E_2^{0,2M}$ *takes* α *to* $c_1(\rho)^M$, *where* $M > 0$. *Then*

(a) $c_1(\rho) \in E_r^{0,2M}$ *for every* $r \geq 2$,

(b) *Multiplication by* $c_1(\rho)$ *induces an isomorphism* $E_r^{*,j} \xrightarrow{\simeq} E_r^{*,j+2M}$ *for every* $r \geq 2$ *and* $j \geq 0$ $(E_r^{*,j} = \bigoplus_{i \geq 0} E_r^{i,j})$, *and*

(c) $E^{**}_{2M+2} = E^{**}_\infty$

Proof.

(a) Referring to the fibration (see Section 3) giving rise to the spectral

sequence

$$BA \to PG \times^G X \to P(G/A) \times^{G/A} X$$

it is enough to note that $c_1(\rho)$ is the restriction to the cohomology of the fibre BA of the class α in the cohomology of the total space $PG \times^G X$. Thus, for every $r \geq 2$, $d_r(c_1(\rho)^M) = 0$, when d_r is the appropriate r-th differential of the spectral sequence.

(b) Multiplication by $c_1(\rho)$ is an isomorphism

$$H^*(X_{G/A}) \otimes_{Z/pZ} H^j_A \xrightarrow{\simeq} H^*(X_{G/A}) \otimes_{Z/pZ} H^{j+2M}_A$$

for every $j \geq 0$ since $c_1(\rho)$ is a polynomial generator of H^*_A. So, for $r = 2$, (b) is true.

Suppose (b) has been proven for $r \geq 2$. Suppose $j \geq 0$. Consider the following diagram:

$$\begin{array}{ccccc} E_r^{i-r,j+r-1+2M} & \xrightarrow[d_r]{} & E_r^{i,j+2M} & \xrightarrow[d_r]{} & E_r^{i+r,j-r+1+2M} \\ \uparrow{\scriptstyle \cdot c_1(\rho)^M} & & \uparrow{\scriptstyle \cdot c_1(\rho)^M} & & \uparrow{\scriptstyle \cdot c_1(\rho)^M} \\ E_r^{i-r,j+r-1} & \xrightarrow[d_r]{} & E_r^{i,j} & \xrightarrow[d_r]{} & E_r^{i+r,j-r+1} \end{array}$$

d_r is the differential of the spectral sequence. The diagram is commutative since

$$d_r(c_1(\rho)^M \cdot x) = c_1(\rho)^M \cdot d_r(x) \pm d_r(c_1(\rho)^M) \cdot x$$

by the multiplicative property of d_r, and since $d_r(c_1(\rho)^M) = 0$.

If $j - r + 1 \geq 0$, then by induction all the vertical maps are isomorphisms, so the induced map on homology,

$$E_{r+1}^{i,j} \xrightarrow[\cdot c_1(\rho)^M]{} E_{r+1}^{i,j+2M},$$

is an isomorphism.

If $j - r + 1 < 0$, $E_r^{i+r,j-r+1} = 0$, and a diagram case shows that

$$E_{r+1}^{i,j} \xrightarrow[\cdot c_1(\rho)^M]{} E_{r+1}^{i,j+2M}$$

is an isomorphism in this case also.

(c) $d_{2m+2}: E_{2M+2}^{i,j} \to E_{2M+2}^{i+2M+2,j-2M+1}$. Suppose that $j \leq 2M$; then

$d_{2M+2}(E^{*,j}_{2M+2}) = 0$ since the spectral sequence is first quadrant. If $j > 2M$, there is an integer $m > 0$ such that $j = 2mM + k$, where $k \leq 2M$. Let $y \in E^{i,j}_{2M+2} = E^{i,2mM+k}_{2M+2}$; then (b) and induction on m show that $y = (c_1(\rho)^M)^m \cdot x$ for some $x \in E^{i,k}_{2M+2}$. So

$$d_{2M+2}(y) = c_1(\rho)^{mM} \cdot d_{2M+2}(x) \pm d_{2M+2}(c_1(\rho)^{mM} \cdot x = 0 \pm 0 = 0.$$

So $d_{2M+2} \equiv 0$, and $E^{**}_{2M+2} = E^{**}_{\infty}$. QED

Thus, we have

COROLLARY 1. *The cohomology class α of Lemma 1 is not a zero divisor on $H^*_G(X)$.*

Proof. Lemma 1 shows that multiplication by $c_1(\rho)^M$ is injective on E^{**}_{∞}. So multiplication by α, which restricts to $c_1(\rho)^M$, must be injective on $H^*_G(X)$. QED

Let $l = N/p$ be the index of A in G, where N is the order of G. Corresponding to the one-dimensional representation ρ of A, there is the l-dimensional induced representation $\text{ind}(\rho)$ of G. There is an l-dimensional vector bundle, also called $\text{ind}(\rho)$, over the classifying space BG for G:

$$PG \times^G \mathbf{C}^l \to BG.$$

The mod-p Chern classes for this vector bundle are $c_i(\text{ind}(\rho)) \in H^{2i}(BG) = H^{2i}_G$ for $0 \leq i \leq l$.

LEMMA 2. *The cohomology class $e = c_l(\text{ind}(\rho))$ is a non-zero divisor on $H^*_G(X)$.*

Proof. Since A is central, the Mackey induction formula (e.g., see (Se)) shows that

$$\text{res ind}(\rho)) = l\rho = \rho \underbrace{\oplus \cdots \oplus}_{l \text{ times}} \rho.$$

Thus $\text{res}(c_l(\text{ind}(\rho))) = c_l(\text{res}(\text{ind}(\rho)) = c_l(l\rho) = c_1(\rho)^l$ by various properties of Chern classes. (Here "res" is ambiguously used to denote either the restriction of the representation $\text{ind}(\rho)$ of G to the subgroup A, or the restriction map in cohomology from H^*_G to H^*_A.)

We may now proceed with the proof of Theorem 1.

Proof of Theorem 1. We begin the induction by noting that Lemma 2 proves the Theorem in case $\text{rank}(A) = 1$.

Now, suppose rank $(A) = n > 1$. Thus, A is central, and $X^A = X$. Let $l = N/p^n$ be the index of A in G. Let A_1 be a subgroup of rank 1 in A, and write $A = A_1 \times B$ where B is a p-torus of rank $n-1$.

There is a one-dimensional representation $\hat{\rho}: A_1 \times B \to \mathbf{C}$ of A given by ρ (the same ρ as in Lemma 1) on A_1 and the trivial representation on B. Let $e = c_l(\mathrm{ind}_{A\to G}(\hat{\rho})) \in H^{2l}_G \leq H$ be the top Chern class of the l-dimensional representation $\mathrm{ind}_{A\to G}(\hat{\rho})$ of G. If $\mathrm{res}_{A_1\to G}: H^*_G \to H^*_{A_1}$, then we have $\mathrm{res}_{A_1\to G}(e) = c_1(\rho)^l$. This follows again from the Mackey induction formula, which implies that $\mathrm{res}_{A_1\to G}(\mathrm{ind}_{A\to G}(\hat{\rho})) = \mathrm{res}_{A_1\to A}(\mathrm{res}_{A\to G}(\mathrm{ind}_{A\to G}(\hat{\rho}))) = \mathrm{res}_{A_1\to A}(l\hat{\rho}) = l\rho$; and standard properties of Chern classes.

Thus, Lemma 2 says that e is not a zerodivisor on $H^*_G(X)$.

The finite group G acts on $\mathbf{C}^{2l}$ via $\mathrm{ind}_{A\to G}(\hat{\rho})$ and therefore on $\mathbf{C}^{2l} \times X$ diagonally. So there is a vector bundle

$$\begin{array}{ccc} PG \times^G (\mathbf{C}^{2l} \times X) & \xrightarrow[\xi]{} & PG \times^G X, \\ \| & & \\ (PG \times X) \times^G \mathbf{C}^{2l} & & \end{array}$$

and the associated (orientable) sphere bundle is

$$PG \times^G (S^{2l-1} \times X) \xrightarrow[\xi']{} PG \times^G X;$$

recall that $\mathrm{ind}_{A\to G}(\hat{\rho})$ is unitary, since $\hat{\rho}$ is unitary. Associated to this sphere bundle ξ' is a mod-p Euler class; it is the top Chern class of the vector bundle ξ. Therefore this Euler class is equal to e.

There is the exact Gysin sequence for ξ' (S):

$$\cdots \to H^{i-2l}(X_G) \xrightarrow[\cdot e]{} H^i(X_G) \xrightarrow{\theta} H^i((S^{2l-1} \times X)_G) \to \cdots.$$

The map $H^{i-2l}(X_G) \to H^i(X_G)$ is multiplication by e as indicated, and since e is a non-zero-divisor, this map in injective.

So, there is a short exact sequence of H-modules

$$0 \to H^*(X_G) \xrightarrow[\cdot e]{} H^*(X_G) \xrightarrow[\theta]{} H^*((S^{2l-1} \times X)_G) \to 0,$$

where $H^*(X_G) = \oplus_{i \geq 0} H^i(X_G)$. Multiplication by e is an H-module map since e has even degree.

This short exact sequence shows that there is an H-module isomorphism

$$H^*((S^{2l-1}\times X)_G) \xleftarrow[\tilde{\tilde{\theta}}]{} H^*(X_G)/(e)H^*(X_G)$$

induced by θ.

The isomorphism $\tilde{\tilde{\theta}}$ provides the inductive step. For, how does B act on $S^{2l-1}\times X$? In fact, B acts trivially. To see this it is enough to note that

(1) $\mathrm{res}_{B\to G}(\mathrm{ind}_{A\to G}(\hat{\rho})) = l$ (the l-dimensional trivial representation).

(*Proof*: B is central since A is. So $\mathrm{res}_{B\to G}(\mathrm{ind}_{A\to G}(\hat{\rho})) = \mathrm{res}_{B\to A}(\mathrm{res}_{A\to G}(\mathrm{ind}_{A\to G}(\hat{\rho}))) = \mathrm{res}_{B\to A}(l\hat{\rho}) = l\cdot \mathrm{res}_{B\to A}(\hat{\rho}) = l.$) and

(2) $X^B = X$; this follows because $X^A = X$ and $B \leq A$.

Since $\mathrm{rank}(B) < \mathrm{rank}(A)$, B is central and $(S^{2l-1}\times x)^B = S^{2l-1}\times X$, we may use induction to obtain an H-sequence $e_1, e_2, \ldots, e_{n-1}$ of length $n-1$ on $H^*((S^{2l-1}\times X)_G)$. Using the isomorphism $\tilde{\tilde{\theta}}$; $e, e_1, e_2, \ldots, e_{n-1}$ is an H-sequence of length n on $H^*(X_G)$. QED

We notice that it is possible to explicitly write down an H-sequence on $H^*(X_G)$ in the following manner. Write $A = A_1\times A_2\times\cdots\times A_n$ as a direct product of cyclic groups of order p. For $1\leq i\leq n$, let $\rho_i : A\to \mathbf{C}^*$ be the one dimensional representation of A given by the trivial representation of A on all but the i-th factor of A and by (our usual) ρ on A_i. If $e_i = c_l(\mathrm{ind}_{A\to G}(\rho_i))$ is the l-th Chern class of ρ_i ($l = N/p^n$) the proof of Theorem 1 shows that $e_1, e_2, \ldots, e_n$ is an H-sequence on $H^*(X_G)$.

Also,

$$H^*_G(X)/(e_1,\ldots,e_i)H^*_G(X) \cong H^*\left(\left(\left(\underbrace{S^{2l-1}\times\cdots\times S^{2l-1}}_{i \text{ factors}}\right)\times X\right)_G\right),$$

where G acts on $(S^{2l-1})^i$ via $\mathrm{ind}(\rho_j)$ on the j-th factor (for $1\leq j\leq i$) and on $(S^{2l-1})^i\times X$ diagonally.

3. A spectral sequence

If G and X are as in Section 2, let N be a normal subgroup of G acting *trivially* on X.

In this section we point out that there is a fibration

(*) $BN \to X_G \to X_{G/N}$

giving rise to a Serre spectral sequence

$$E_2^{**} = H^*(X_{G/N}, \{H_N^*\}) \Rightarrow H^*(X_G).$$

($\{\cdot\}$ denotes local coefficients.)

To get the fibration (*) we start with a classifying bundle $P(G/N) \to B(G/N)$ for principal G/N bundles. Then, there is a classifying bundle $PG \to BG$ for principal G-bundles and a commutative diagram

$$\begin{array}{ccc} PG & \to & P(G/N) \\ \downarrow & & \downarrow \xi \\ BG & \underset{f}{\to} & B(G/N) \end{array}$$

with f a *fibration.*

Since N acts trivially on $P(G/N)$ there is a commutative diagram

$$\begin{array}{ccc} PG & & \\ & \searrow^{\tilde{f}} & \\ \downarrow & PG/N & \to P(G/N) \\ & \downarrow \xi' & \square \quad \downarrow \xi \\ & BG & \underset{f}{\to} B(G/N) \end{array}$$

and the big square $\square$ is cartesian.

Replacing the fibres of the principal G/N bundles ξ and ξ' by the G/N-space X and noting that $PG/N \times^{G/N} X$ is homeomorphic to $PG \times^G X$, we have a commutative diagram of fibrations with the indicated square cartesian.

$$\begin{array}{ccccc} BN & \to & X_G & \to & X_{G/N} \\ \| & & \downarrow & \square & \downarrow \\ BN & \to & BG & \to & B(G/N) \end{array}$$

Since the fibration $X_G \to X_{G/N}$ is induced from the map $BG \to B(G/N)$, if the local coefficient system is trivial for the latter fibration, it is trivial for the former fibration (S).

Now, if N is central in G, it is a fact that the local coefficient system $\{H_N^*\}$ is

trivial for the fibration $BG \to B(G/N)$. Therefore, in this case, the local coefficient system $\{H_N^*\}$ is trivial for the fibration $X_G \to X_{G/N}$. Since we are using field coefficients we see that

$$E_2^{**} = H^*(X_{G/N}, H_N^*) \cong H^*(X_{G/N}) \otimes_{Z/pZ} H_N^*$$

if N is central.

(Note: The results in this section are of course also true if G is a compact Lie group, and N is a closed normal subgroup.)

4. Certain H-modules are Cohen-Macaulay

Suppose H is a (graded) commutative, Noetherian ring and M is a (graded) commutative Noetherian H-algebra that is also a finitely generated (graded) H-module. We may define (as is usual) the *H-depth* of M as the maximal rank of an H-sequence (defined as in Section 2) on M and the *H-dimension* of M as the Krull dimension of $H/\mathrm{ann}_H(M)$. Standard results in commutative algebra ensure that the H-depth of M equals the M-depth of M; and the H-dimension of M equals the M-dimension of M (M, Se2). Thus we may refer without ambiguity to the *depth* of M and the *dimension* of M without the "H" or "M" prefixes. We also have the standard result that depth $M \geq \dim M$. We say that M is *Cohen-Macaulay* if depth $M = \dim M$.

(The reader who is used to seeing these results for *local* rings is reminded of the accurate (at least for these results) analogy between local and graded rings. The ideal of positive degree elements plays the part that the maximal ideal does in local algebra.)

Fix a prime p. Let G be a finite group and let X be a G-space with finite-dimensional (over Z/pZ) mod-p cohomology. Define H as in Section 2. Let

$$M = H_G(X, Z/pZ) = \begin{cases} H_G^{\mathrm{ev}}(X, Z/pZ) & p \text{ odd} \\ H_G^*(X, Z/pZ) & p \text{ even.} \end{cases}$$

As in Section 2, M is a finitely generated H-module via the map $X \to \mathrm{pt}$.

In this section we show that H and M are Cohen-Macaulay for certain groups G and certain G-spaces X. Namely, we have:

THEOREM 2. *Let G be a finite group with a unique p-torus A of maximal rank. Assume A is also central. Let X be a G-space such that $X^A = X$. Then M and H are Cohen-Macaulay.*

Proof. According to Quillen (Q1: also, see Section 0), $\dim M = \operatorname{rank}(A) = \dim H$. Theorem 1 shows that depth $H \geq \operatorname{rank}(A)$, and that depth $M \geq \operatorname{rank}(A)$. Now, since dimension always dominates depth, we obtain the theorem. QED

If G is a finite group let $C_G(A) = \{g \in G \mid ga = ag \ \forall a \in A\}$ be the centralizer of A in G.

COROLLARY 2. *If G is a finite group and X is a G-space with finite dimensional (mod-p) cohomology, then for any maximal p-torus A in G, $H_{C_G(A)}(X^A)$ is Cohen-Macaulay.*

Proof. If A is maximal, then A is the unique maximal p-torus of $C_G(A)$. (For, if B is another p-torus in $C_G(A)$, then for every a in A and b in B, $ab = ba$. Thus, the subgroup generated by A and B, $\langle A, B\rangle$, is a p-torus in G containing A. Since A is maximal, $\langle A, B\rangle = A$ and $B \leq A$.) It is clear that A is central in $C_G(A)$ and acts trivially on X^A. QED

REFERENCES

(A) M. F. ATIYAH, *Characters and Cohomology of Finite Groups*, Publ. I.H.E.S. *9* (1960), 23–64.

(B) A. BOREL, et al., *Seminar on Compact Transformation Groups*, Ann. of Math. Study *46*, 1960.

(E) L. EVENS, *The Cohomology Ring of a Finite Group*, Trans. A.M.S., *101* (1961) 224–239.

(K) I. KAPLANSKY, *Commutative Rings*, The University of Chicago Press, Chicago, 1974.

(M) H. MATSUMURA, *Commutative Algebra*, Benjamin Press, New York, 1970.

(Q1, Q2) D. G. QUILLEN, *The Spectrum of an Equivariant Cohomology Ring I, II*, Ann. of Math. *94* (1971) 549–602.

(Se1) J. P. SERRE, *Linear Representation of Finite Groups*, Springer-Verlag, New York.

(Se2) J. P. SERRE, *Algebre Locale Multiplicites*, Springer-Verlag, New York, 1975.

(S) E. SPANIER, *Algebraic Topology*, McGraw-Hill, New York, 1966.

(V) B. B. VENKOV, *Cohomology Algebras for Some Classifying Spaces*, Dokl. Akad. Nauk. SSSR *127* (1959) 943–944, Math. Rev. *21* #7500.

Dept. of Mathematics,
Brandeis University,
Waltham, Massachusetts 02554, U.S.A.

Received December 24, 1980/October 27, 1981

Comment. Math. Helvetici **56** (1981) 638–660 0010–2571/81/004638–23$01.50+0.20/0
© 1981 Birkhäuser Verlag, Basel

Transverse foliations of Seifert bundles and self homeomorphism of the circle

DAVID EISENBUD, ULRICH HIRSCH and WALTER NEUMANN*

1. Introduction

In this paper we give criteria for a Seifert circle bundle over a compact surface to admit foliations whose leaves are all transverse to the fibers, and we discuss which foliations may be deformed to foliations of this type.

Our criteria for transverse foliations, presented in Section 3, are simple numerical inequalities involving the Seifert pairs of the fibration and the euler number of the base (Theorems 3.1 to 3.4). They generalize criteria of Milnor [Mi] and Wood [W], who treat the case of locally trivial circle bundles with orientable total space (see also Sullivan [Su] for a higher dimensional generalization of Milnor and Wood). They are complete except for the case that the base is S^2, in which case we only have partial results (Theorems 3.3, 5.3). Our criteria are valid both for the case of C^0 foliations and, as we show in Section 4, for analytic foliations, hence also for any intermediate degree of smoothness.

We reduce the geometric question of the existence of foliations to algebra in a way similar to that of Milnor and Wood. We let $\mathcal{D}$ be the group of self-homeomorphisms $f: \mathbf{R} \to \mathbf{R}$ which are lifts of self-homeomorphisms of the circle. $\mathcal{D}$ contains the group $\mathcal{D}^+ = \{f: \mathbf{R} \to \mathbf{R} \mid f$ monotonically increasing and $f(r+1) = f(r)+1$ for all $r \in \mathbf{R}\}$ as a subgroup of index 2 (the "flip" $t(r) = -r$ is a coset representative for the non-trivial coset of $\mathcal{D}^+$ in $\mathcal{D}$). For each real number γ we write $\mathrm{sh}(\gamma) \in \mathcal{D}^+$ for the "shift" by γ, that is $\mathrm{sh}(\gamma): r \mapsto r+\gamma$ for $r \in \mathbf{R}$.

It turns out that the problem of transversely foliating a Seifert fibered manifold M is equivalent to the problem of finding a homomorphism $\pi_1(M) \to \mathcal{D}$ which takes the class of a nonsingular fiber of M to the element $\mathrm{sh}(1) \in \mathcal{D}^+$ (Theorem 3.5). This in turn is equivalent to a problem of representating a certain product of conjugates of shifts as a product of a certain number of commutators in

*The authors wish to express their gratitude to the SFB 40 and the mathematical institute of the University of Bonn, under whose roof most of the work of this paper was done. The first and last authors are also grateful for the partial support of the NSF.

$\mathscr{D}^+$ or $\mathscr{D}$, or as a product of a certain number of squares of elements of $\mathscr{D}-\mathscr{D}^+$.

For this reason we begin, in Section 2, by characterizing the elements of $\mathscr{D}^+$ that can be written as products of g commutators, or of g squares of elements of $\mathscr{D}-\mathscr{D}^+$, etc., and by partially characterizing the elements that can be written as a product of conjugates of a given collection of shifts. Perhaps the most surprising result is that these classes of elements can be simply characterized in terms of the invariants

$$\underline{m}f=\min_{r\in\mathbf{R}}(f(r)-r), \qquad \bar{m}f=\max_{r\in\mathbf{R}}(f(r)-r),$$

of an element $f\in\mathscr{D}^+$ (this min and max exist, since $f(r)-r$ is a continuous perodic function on $\mathbf{R}$). In Section 4 we prove the same characterizations with $\mathscr{D}^+$ and $\mathscr{D}$ replaced by their subgroups $\widetilde{\mathrm{PSL}}(2,\mathbf{R})$ and $\widetilde{\mathrm{PGL}}(2,\mathbf{R})$ respectively, obtained by lifting the subgroups $\mathrm{PSL}(2,\mathbf{R})$ and $\mathrm{PGL}(2,\mathbf{R})$ of $\mathrm{Homeo}(S^1)$. A different characterization for products of commutators in $\widetilde{\mathrm{PSL}}(2,\mathbf{R})$ was given by Wood.

In Section 5 we describe a natural analytic family of examples due to Maria Carmen del Gazolas [dG]. We are grateful to the referee for bringing them to our attention. They are more general than the examples we originally had here and yield better results for base S^2.

The last section (Section 6) of our paper contains results on deforming foliations of Seifert fibered manifolds to make them transverse. Here we only consider transversely oriented C^2 foliations on Seifert manifolds whose base is not S^2 or P^2. Roughly speaking the theorem is that this can be done if the foliation has no compact leaves. The techniques are the same as for the case of locally trivial bundles, done by Thurston [T] and Levitt [L], so, at the referee's suggestion we omit proofs.

We fix the notations $\mathscr{D}$, $\mathscr{D}^+$, $\underline{m}$, $\bar{m}$ and, for $\gamma\in\mathbf{R}$, $\mathrm{sh}(\gamma)$ for use throughout this paper. In addition, we use Knuth's "floor and ceiling" notations $\lfloor\gamma\rfloor$ and $\lceil\gamma\rceil$ for $\gamma\in\mathbf{R}$ to denote respectively, the greatest integer $\leq\gamma$ and least integer $\geq\gamma$.

2. Self-homeomorphisms of the circle

We will actually need information on the "lifted" homeomorphisms in $\mathscr{D}$ and $\mathscr{D}^+$. To make the connection explicit, note that the center of $\mathscr{D}^+$ is

$$Z(\mathscr{D}^+)=\{\mathrm{sh}(n)\mid n\in\mathbf{Z}\}\cong\mathbf{Z},$$

and that $\mathscr{D}^+/Z(\mathscr{D}^+)$ is the group of orientation preserving self-homeomorphisms of the circle $S^1=\mathbf{R}/\mathbf{Z}$.

We write $a^b = b^{-1}ab$ for the *conjugate* of a by b. A *commutator* is an element of the form $a^{-1}b^{-1}ab$.

Before we describe our main results characterizing certain products of special elements of $\mathscr{D}^+$, we describe some elementary properties of $\mathscr{D}^+$ and of the functions $\underline{m}$ and $\bar{m}$.

LEMMA 2.1. *Let f, f_i, g be elements of $\mathscr{D}^+$. Then*

(1) *If $x, y \in \mathbf{R}$, then $x - y \in \mathbf{Z}$ implies $fx - fy = x - y$. If, on the contrary; $n < x - y < n+1$ for some $n \in \mathbf{Z}$, then $n < fx - fy < n+1$.*

(2) $0 \le \bar{m}f - \underline{m}f < 1$.

(3) $\underline{m}f^{-1} = -\bar{m}f$, *and dually,* $\bar{m}f^{-1} = -\underline{m}f$.

(4) $\sum_{i=1}^n \underline{m}f_i \le \underline{m}(\prod_{i=1}^n f_i) \le \underline{m}f_n + \sum_{i=1}^{n-1} \bar{m}f_i < (\sum_{i=1}^n \underline{m}f_i) + n - 1$, *and dually,*

$$\sum_{i=1}^{n} \bar{m}f_i \ge \bar{m}\left(\prod_{i=1}^{n} f_i\right) \ge \bar{m}f_n + \sum_{i=1}^{n-1} \underline{m}f_i > \left(\sum_{i=1}^{n} \bar{m}f_i\right) - n + 1.$$

(5) $\lfloor \underline{m}f^g \rfloor = \lfloor \underline{m}f \rfloor$ *and* $\lceil \underline{m}f^g \rceil = \lceil \underline{m}f \rceil$ *(so, in particular, if $\underline{m}f \in \mathbf{Z}$, then $\underline{m}f^g = \underline{m}f$) and the same for $\bar{m}$.*

Proof. (1) Apply f to the equality $y + n = x$ or to the inequality $y + n < x < y + n + 1$, for $n \in \mathbf{Z}$.

(2) Choose $x, y \in \mathbf{R}$ with $fx - x = \underline{m}f$ and $fy - y = \bar{m}f$ and observe that if $\bar{m}f - \underline{m}f \ge 1$ then x and y contravene (1).

(3) is clear, as is the first inequality in (4). For the second inequality choose $x \in \mathbf{R}$ with $f_n x - x = \underline{m}f_n$ and observe that $(\prod f_i)x - x \le \underline{m}f_n + \sum_{i=1}^{n-1} \bar{m}f_i$. The third inequality now follows by (2), and the dual inequalities follow by (3).

(5) By (1) we have $\lfloor x - y \rfloor = \lfloor g^{-1}x - g^{-1}y \rfloor$ for $x, y \in \mathbf{R}$. Choosing $y = gz$ and $x = fgz$ gives $\lfloor fgz - gz \rfloor = \lfloor f^g z - z \rfloor$, and minimizing over $z \in \mathbf{R}$ gives $\lfloor \underline{m}f \rfloor = \lfloor \underline{m}f^g \rfloor$. Replacing $\lfloor\ \rfloor$ by $\lceil\ \rceil$ and/or min by max in this argument proves the rest of (5).

We shall also need a special subset of $\mathscr{D}^+$. We say $f \in \mathscr{D}^+$ has a *stable fixed point* at $r \in \mathbf{R}$ if for any $s \in \mathbf{R}$ sufficiently close to r, the iterates $f^n(s)$ converge to r as n goes to infinity. Equivalently, $f(r-\varepsilon) > r - \varepsilon$ and $f(r+\varepsilon) < r + \varepsilon$ for all ε sufficiently small. We say $f \in \mathscr{D}^+$ has an *unstable fixed point at* $r \in \mathbf{R}$ if r is a stable fixed point of f^{-1}. For $k > 0$ we write

$$f \in \mathrm{SUF}(k)$$

(SUF stands for: Stable or Unstable Fixpoints) if f has exactly $2k$ fixed points on some (or equivalently every) half open unit interval $[r, r+1)$, and exactly k of them are stable and the other k are unstable. It is clear that the fixed points of f then alternate type (stable or unstable) along the real line. The significance of this definition is given by the following well known lemma.

LEMMA 2.2. *Any two elements of* SUF(k) *are conjugate in* $\mathscr{D}^+$.

Proof. We prove this for $k=1$, the only case we ever use. The general proof will then be clear. Thus let $f, g \in \mathrm{SUF}(1)$. Let x be an unstable fixed point of f and let y be the unique fixed point of f in the interval $(x, x+1)$, so y is necessarily stable. Let x' and y' be chosen similarly for g. We shall construct an $h \in \mathscr{D}^+$ with $hx = x'$, $hy = y'$ and $hgh^{-1} = f$.

It is enough to construct h on each of the intervals $[x, y]$ and $[y, x+1]$ so that the equation $hgh^{-1} = f$ holds where it is defined. We do this for $[x, y]$; the construction for $[y, x+1]$ is similar. Choose z and z' with $x < z < y$ and $x' < z' < y'$ and let $h_0 : [z, f(z)] \to [z', g(z')]$ be any monotonic homeomorphism. For $n \in \mathbf{Z}$ define $h_n = g^n h_0 f^{-n} : f^n[z, f(z)] \to g^n[z', g(z')]$. Since the intervals $f^n[z, f(z)]$ partition the interval (x, y) and the intervals $g^n[z', g(z')]$ partition (x', y'), these maps h_n fit together to give the desired map $h : [x, y] \to [x', y']$.

We now come to the main results of this section. The first result needed for our geometric application is:

THEOREM 2.3. *Let f be an element of $\mathscr{D}^+$.*

(1) *f can be written as a product of $g \geq 1$ commutators of elements of $\mathscr{D}^+$ if and only if $\underline{m}f < 2g-1$ and $\bar{m}f > 1-2g$.*

(2) *f can be written as a product of $g \geq 2$ squares of elements of $\mathscr{D}-\mathscr{D}^+$ if and only if $\underline{m}f < g-1$ and $\bar{m}f > 1-g$.*

Remark. (2) remains valid if "squares of elements of $\mathscr{D}-\mathscr{D}^+$" is replaced either by "elements of SUF(1)" or "elements of $\mathscr{D}^+$ having fixed points".

The situation for commutators in $\mathscr{D}$, which we need to handle the non-orientable case of our geometric problem, is much simpler:

THEOREM 2.4. *Every element of $\mathscr{D}^+$ is a commutator of an element of $\mathscr{D}-\mathscr{D}^+$ with an element of $\mathscr{D}^+$.*

We shall also need a partial characterization of certain products of conjugates in $\mathscr{D}^+$:

THEOREM 2.5. *Let $f_1, \ldots, f_n \in \mathscr{D}^+$ and $r \in \mathbf{R}$ be given with*

$$\sum \lfloor \underline{m} f_i \rfloor < r < \sum \lceil \bar{m} f_i \rceil.$$

Then there exist elements $e_i \in \mathscr{D}^+$ such that $d = \prod (f_i^{e_i})$ satisfies

$$\underline{m} d \leq r \leq \bar{m} d.$$

To deal with Seifert fibrations over the sphere we need to know when d can be the identity in the above theorem, when the f_i are shifts. We only have a partial answer to this question.

THEOREM 2.6. *Let $\gamma_1, \ldots, \gamma_k$ be real numbers.*

(1) *There exist $e_i \in \mathcal{D}^+$ such that $\prod \mathrm{sh}(\gamma_i)^{e_i} \in \mathrm{SUF}(1)$ if and only if $\sum \lfloor \gamma_i \rfloor \leq -1$ and $\sum \lceil \gamma_i \rceil \geq 1$.*

(2) *If $\sum \gamma_i = 0$ or $\sum \lfloor \gamma_i \rfloor \leq -2$ and $\sum \lceil \gamma_i \rceil \geq 2$, then there exist $e_i \in \mathcal{D}^+$ such that $\prod \mathrm{sh}(\gamma_i)^{e_i} = 1$.*

(3) *Conversely, if e_i exist as in (2), the either $\sum \gamma_i = 0$, or $\sum \lfloor \gamma_i \rfloor \leq -1$ and $\sum \lceil \gamma_i \rceil \geq 1$ with at least one of these inequalities strict.*

The rest of this section gives the proofs of Theorems 2.3 to 2.6. We start with the proof of 2.3, which is an induction, starting at the case $g = 1$ in part (1) and $g = 2$ in part (2). Since we shall often need this special case, we state it as a lemma, for easy reference.

LEMMA 2.7. *The following conditions on an element $f \in \mathcal{D}^+$ are equivalent:*

(1) *$\underline{m} f < 1$ and $\bar{m} f > -1$.*

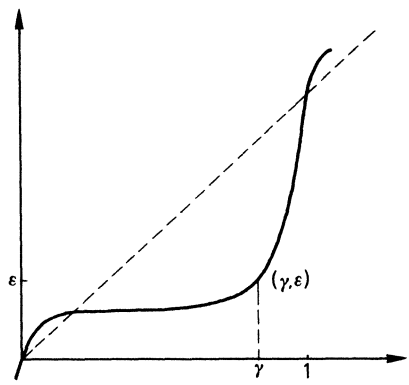
(2) *f is the product of two elements of* SUF(1).

(3) *f is a commutator in $\mathcal{D}^+$.*

(4) *f is the product of two squares of elements of $\mathcal{D} - \mathcal{D}^+$.*

Proof. We begin with the most delicate point, the constructive statement (1)$\Rightarrow$(2).

Let $\varepsilon, \gamma \in \mathbf{R}$ with $0 < \varepsilon < \gamma < 1$, to be chosen later, and let $c \in \mathcal{D}^+$ be a function whose graph looks like



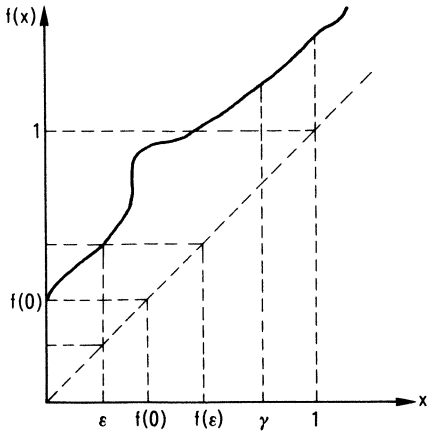
That is, $c(0)=0$, $c(\gamma)=\varepsilon$, and $c \in \mathrm{SUF}(1)$, with an unstable fixed point at 0 and a stable one in the interval $(0, \varepsilon)$.

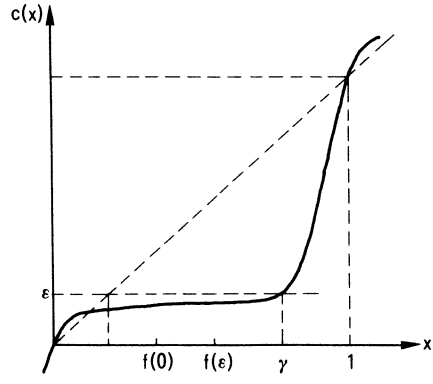
Given that f satisfies (1) and is not the identity, we may suppose, by inverting f if necessary, that $\bar{m}f>0$. Conjugating f by a shift if necessary, we can then further suppose that $0<f(0)<1$. Chosen ε close enough to zero and γ close enough to 1 that

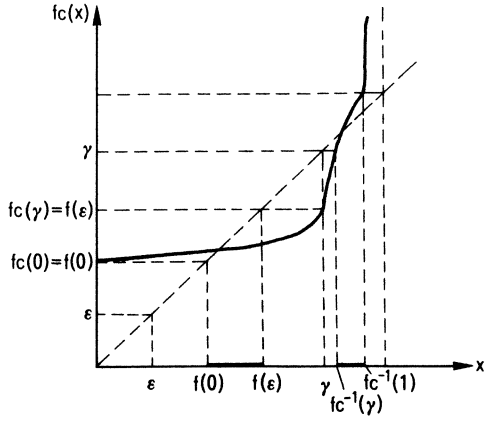
$$0<\varepsilon<f(0)<f(\varepsilon)<\gamma<1<f(\gamma).$$

We may write $f=(fc)c^{-1}$; we will see that this is nearly the required factorization.

First, since $fc(0)>0$, $fc(\gamma)<\gamma$, and $fc(1)>1$, the function fc must have at least one fixed point on each of $(0, \gamma)$ and $(\gamma, 1)$. On $(0, \gamma)$, the fixed points must of course occur on $fc(0, \gamma)=f(0, \varepsilon)$, while on $(\gamma, 1)$ they must occur on $(fc)^{-1}(\gamma, 1) \subset (\gamma, 1)$.







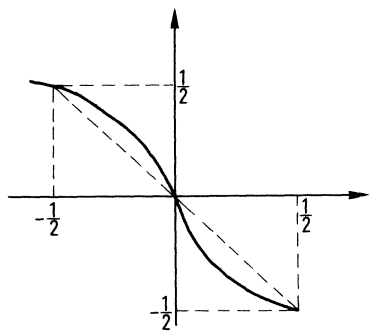
If f is nicely behaved, say f and f^{-1} Lipschitz, then by making c sufficiently flat on $f(0, \varepsilon)$ and sufficiently steep on $(\gamma, 1)$, we can assure that fc has exactly one (stable respectively unstable) fixed point on each of these intervals, so the factorization $f = (fc)c^{-1}$ is the desired one.

In general, we proceed as follows. Let g be the unique element of $\mathscr{D}^+$ which agrees with fc on $[0, 1] - f(0, \varepsilon) - (fc)^{-1}(\gamma, 1)$, and which is linear on $f[0, \varepsilon]$ and on $(fc)^{-1}[\gamma, 1]$. Clearly g is in SUF(1) (with a stable fixed point on $f(0, \varepsilon)$ and an unstable one on $(fc)^{-1}(\gamma, 1)$). We shall replace the factorization $f = (fc)c^{-1}$ by a factorization $f = g(c')^{-1}$, so it suffices to show that $c' = f^{-1}g$ is in SUF (1).

Now $c' = c(fc)^{-1}g$, so $c' = c$ except on $f(0, \varepsilon)$ and $(fc)^{-1}(\gamma, 1)$. Since $c'f(\varepsilon) = cf(\varepsilon) < c(\gamma) = \varepsilon < f(0)$, we see that $c'f(0, \varepsilon)$ is disjoint from $f(0, \varepsilon)$, so c' has no fixed point on $f(0, \varepsilon)$. Similarly $c'(fc)^{-1}(1) = c(fc)^{-1}(1) = f^{-1}(1) < \gamma < (fc)^{-1}(\gamma)$, so $c'((fc)^{-1}(\gamma, 1))$ is disjoint from $(fc)^{-1}(\gamma, 1)$, and so c' has no fixed points there either. Thus $c' \in$ SUF (1), and the proof of (1)$\Rightarrow$(2) is concluded.

(2)$\Rightarrow$(3). Write $f = dc$, with $c, d \in$ SUF (1). By Lemma 2.2, d^{-1} is a conjugate of c in $\mathscr{D}^+$, say $d = b^{-1}c^{-1}b$, so $f = b^{-1}c^{-1}bc$, as required.

(2)$\Rightarrow$(4). It suffices to show that any $c \in$ SUF (1) is a square of an element of $\mathscr{D} - \mathscr{D}^+$. By Lemma 2.2 it suffices to show this for just one such c. Let $a \in \mathscr{D} - \mathscr{D}^+$ have the following graph on $[-\frac{1}{2}, \frac{1}{2}]$ that is $a(0) = 0$, $a(\frac{1}{2}) = -\frac{1}{2}$, $a(-\frac{1}{2}) = \frac{1}{2}$, and



$a(x) < -x$ for $x \in (0, \frac{1}{2})$ and $a(x) > -x$ for $x \in (-\frac{1}{2}, 0)$. Clearly $c = a^2$ has a stable fixed point at $\frac{1}{2}$, an unstable one at 0, and no others in the interval $[0, 1)$.

(4)$\Rightarrow$(1). If $c \in \mathscr{D}^+$ has a fixed point then $-1 < \underline{m}c \leq 0$ and $0 \leq \bar{m}c < 1$. Thus if both c and d have fixed points, then $\underline{m}(cd) \leq \underline{m}c + \bar{m}d < 0 + 1 = 1$ and similarly $\bar{m}(cd) > -1$.

(3)$\Rightarrow$(1). Say $f = c^{-1}d^{-1}cd$. Using Lemma 2.1 we see

$$\underline{m}f = \underline{m}(c^{-1}d^{-1}cd) \leq \underline{m}(c^{-1}) + \bar{m}(d^{-1}cd) \leq -\bar{m}c + \lceil \bar{m}(d^{-1}cd \rceil$$
$$= -\bar{m}c + \lceil \bar{m}c \rceil < 1.$$

Similarly $\bar{m}f > -1$.

The inductive step in the proof of Theorem 2.3 will use the following consequence of Lemma 2.7.

PROPOSITION 2.8. *Let $f \in \mathscr{D}^+$ and $0 < \varepsilon < 1$ be given. Then*
(1) *$f = kh$ for some $h, k \in \mathscr{D}^+$ with*

$$0 < \underline{m}h < 1,$$
$$\underline{m}k < \underline{m}f - 2 + \varepsilon < \bar{m}k.$$

(2) *$f = kh$ for some $h, k \in \mathscr{D}^+$ with $h \in \mathrm{SUF}(1)$ and $\underline{m}k < \underline{m}f - 1 + \varepsilon < \bar{m}k$.*

Proof. By Lemma 2.7, $\mathrm{sh}(-\underline{m}f + 1 - \varepsilon)f = k'h'$, where each of k' and h' is in SUF(1). The decomposition $f = (\mathrm{sh}(\underline{m}f - 1 + \varepsilon)k')h'$ satisfies (2), while $f = (\mathrm{sh}(\underline{m}f - 2 + \varepsilon)k')(\mathrm{sh}(1)h')$ satisfies (1).

Proof of Theorem 2.3. The "only if" statements follow at once from Lemma 2.7 and Lemma 2.1(4). We prove the "if" statements by induction on g, the cases $g = 1$ of (1) and $g = 2$ of (2) being Lemma 2.7. Inverting f if necessary, we may assume $\bar{m}f > 0$.

(1) If $\underline{m}f < 2g - 1$, then we use Proposition 2.8(1) to write $f = kh$ with $0 < \underline{m}h < 1$ and $\underline{m}k < \underline{m}f - 2 + \varepsilon < \bar{m}k$. By Lemma 2.7, h can be written as a commutator, and if ε is small enough, then k will be a product of $g - 1$ commutators by induction hypothesis.

(2) If $\underline{m}f < g - 1$, then we apply Proposition 2.8(2). The element h will, as noted in the proof of Lemma 2.7, be the square of an element of $\mathscr{D} - \mathscr{D}^+$, while k will be a product of $g - 1$ such squares by induction hypothesis.

Proof of Theorem 2.4. We first note two simple equations. If $t \in \mathscr{D}$ is given by $t(x) = -x$, and if $a \in \mathscr{D} - \mathscr{D}^+$ is arbitrary, then

$$t^{-1} \cdot \mathrm{sh}(\gamma) \cdot t = \mathrm{sh}(-\gamma) \quad \text{for all} \quad \gamma \in \mathbf{R},$$
$$a^{-1} \cdot \mathrm{sh}(n) \cdot a = \mathrm{sh}(-n) \quad \text{for all} \quad n \in \mathbf{Z}.$$

Given $f \in \mathscr{D}^+$, we wish to write it as $f = a^{-1}b^{-1}ab$ with $a \in \mathscr{D} - \mathscr{D}^+$ and $b \in \mathscr{D}^+$. If f is a shift we use the equation $t^{-1} \cdot (\mathrm{sh}(\gamma))^{-1} \cdot t \cdot \mathrm{sh}(\gamma) = \mathrm{sh}(2\gamma)$, so we may assume f is not a shift. We can then find $n \in \mathbf{Z}$ such that $\underline{m}f - 2n < 1$ and $\bar{m}f - 2n > -1$ (namely let $2n$ be the one of $\lfloor \underline{m}f \rfloor$ and $\lfloor \underline{m}f \rfloor + 1$ which is even). Put $f' = \mathrm{sh}(-2n)f$.

Then f' satisfies condition (1) of Lemma 2.7, so $f' = cd$ with $c, d \in \mathrm{SUF}(1)$. Now $tc^{-1}t^{-1}$ is still in SUF(1), so by Lemma 2.2, $tc^{-1}t^{-1} = e^{-1}de$ for some $e \in \mathscr{D}^+$. Putting $a = et$, we get $c = a^{-1}d^{-1}a$, so $f' = a^{-1}d^{-1}ad$. Finally $f = \mathrm{sh}(2n)f' =$

$a^{-1} \cdot \mathrm{sh}(-n) \cdot a \cdot \mathrm{sh}(n) \cdot a^{-1}d^{-1}ad = a^{-1}(\mathrm{sh}(n)d)^{-1}a(\mathrm{sh}(n)d)$ is in the desired form.

Proof of Theorem 2.5. First note that given $f \in \mathscr{D}^+$ and $0 < \varepsilon < 1$, we can find a conjugate f' of f with $\underline{m}f' \leq \lfloor \underline{m}f \rfloor + \varepsilon$. Indeed, if $\underline{m}f = \lfloor \underline{m}f \rfloor$ this is trivial; otherwise choose $x \in \mathbf{R}$ with $fx = x + \lfloor \underline{m}f \rfloor + t$ and $0 < t < 1$, and choose $a \in \mathscr{D}^+$ with $a(x) = x$ and $a(x+\varepsilon) = x + t$, and then $a^{-1}fa(x) = x + \lfloor \underline{m}f \rfloor + \varepsilon$, so $\underline{m}(a^{-1}fa) \leq \lfloor \underline{m}f \rfloor + \varepsilon$.

We can thus replace each f_i in Theorem 2.5 by a conjugate f_i' with $\underline{m}f_i' \leq \lfloor \underline{m}f_i \rfloor + \varepsilon$ for some small ε. We next observe that by conjugating each f_i' by a suitable shift we can move the points at which they attain their $\underline{m}$'s so as to achieve also: $\underline{m}(\prod f_i') = \sum \underline{m}f_i'$.

Thus if $r \in \mathbf{R}$ is as in the theorem, we can find $b_i \in \mathscr{D}^+$ such that $\underline{m}(\prod f_i^{b_i}) < r$. Dually, one finds $c_i \in \mathscr{D}^+$ such that $r < \bar{m}(\prod f_i^{c_i})$.

Now let $e_i : [0, 1] \to \mathscr{D}^+$ be a continuous path with $e_i(0) = b_i$ and $e_i(1) = c_i$ for each i ($\mathscr{D}^+$ is a convex subset of $\mathbf{R}^{\mathbf{R}}$, hence pathwise connected). If $d_t = \prod f_i^{e_i(t)}$ for $0 \leq t \leq 1$ then, since $\underline{m}$ and $\bar{m}$ are continuous, the intermediate value theorem implies $\underline{m}(d_t) \leq r \leq \bar{m}(d_t)$ for some t, proving the theorem.

Proof of Theorem 2.6. The proof uses the following lemma.

LEMMA 2.9. (1) *If* $-1 < \gamma_1 < 0 < \gamma_2 < 1$, *then there exists* $e \in \mathscr{D}^+$ *such that* $\mathrm{sh}(\gamma_1)(\mathrm{sh}(\gamma_2))^e \in \mathrm{SUF}(1)$.

(2) *If* $-1 < \gamma < 1$ *and* $f \in \mathrm{SUF}(1)$, *then there exists* $e \in \mathscr{D}^+$ *such that* $f(\mathrm{sh}(\gamma))^e \in \mathrm{SUF}(1)$.

This lemma is easy to prove directly, but since it follows immediately from a stronger result, Lemma 4.2, to be proved later, we postpone its proof for now.

Returning to the proof of Theorem 2.6(1), note that the "only if" is immediate from Lemma 2.1, so we must prove the "if". Assume therefore $\sum \lfloor \gamma_i \rfloor \leq -1$ and $\sum \lceil \gamma_i \rceil \geq 1$. Our first step is to "normalize" the γ_i's.

Observe that inserting or deleting a $\gamma_i = 0$ does not change the problem. Also, since $\mathrm{sh}(n)$ is in the center of $\mathscr{D}^+$ for $n \in \mathbf{Z}$, if we replace each γ_i by $\gamma_i + n_i$ with $n_i \in \mathbf{Z}$ and $\sum n_i = 0$, then we also do not change the problem. We can thus normalize and reindex γ_i so that they become:

$$\gamma_0, \gamma_1, \ldots, \gamma_l; \qquad \gamma_0 \in \mathbf{Z}, \quad 0 < \gamma_i < 1 \quad \text{for} \quad i = 1, \ldots, l.$$

The conditions $\sum \lfloor \gamma_i \rfloor \leq -1$ and $\sum \lceil \gamma_i \rceil \geq 1$ then become: $\gamma_0 = -n$ with $1 \leq n \leq l-1$. By eliminating γ_0 by subtracting 1 from n of the γ_i, we can renormalize once more to obtain:

$$\gamma_1, \ldots, \gamma_l; \qquad -1 < \gamma_i < 1 \quad \text{for} \quad i = 1, \ldots, l. \tag{*}$$

Since exactly n of these γ_i are negative, and $1 \le n \le l-1$, we can assume $-1 < \gamma_1 < 0$ and $0 < \gamma_2 < 1$.

We now apply part (1) of Lemma 2.9 to find $e \in \mathscr{D}^+$ with $\mathrm{sh}(\gamma_1)\,\mathrm{sh}(\gamma_2)^e \in \mathrm{SUF}(1)$. Applying part (2) of the lemma iteratively, with $\gamma = \gamma_3, \ldots, \gamma_l$, then completes the proof.

To prove 2.6(2), note that if $\sum \gamma_i = 0$ we can take all $e_i = 1$. If $\sum \lfloor \gamma_i \rfloor \le -2$ and $\sum \lceil \gamma_i \rceil \ge 2$, then we can renormalize as above, to get the γ_i in the form $(*)$ with $-1 < \gamma_1, \gamma_3 < 0$ and $0 < \gamma_2, \gamma_4 < 1$. Then $\{\gamma_1, \gamma_2\}$ and $\{\gamma_3, \ldots, \gamma_l\}$ each satisfy the inequalities of part (1) of the theorem, so (2) follows from (1) and Lemma 2.2.

(3) Assume some product $\prod \mathrm{sh}(\gamma_i)^{e_i}$ equals 1. If all the γ_i are integers then we must have $\sum \gamma_i = 0$. Otherwise, for each i we have $\lfloor \gamma_i \rfloor \le \underline{m}(\mathrm{sh}(\gamma_i)^{e_i}) \le \bar{m}(\mathrm{sh}(\gamma_i)^{e_i}) \le \lceil \gamma_i \rceil$, with both the first and third inequality strict for some i, so the inequalities $\sum \lfloor \gamma_i \rfloor < 0$ and $\sum \lceil \gamma_i \rceil > 0$ follow from Lemma 2.1(4).

Now suppose $\sum \lfloor \gamma_i \rfloor = -1$ and $\sum \lceil \gamma_i \rceil = +1$. Then the normalization procedure used in the proof of part (1) implies that we may assume $k = 2$ and $-1 < \gamma_1 < 0 < \gamma_2 < 1$. The equation $\mathrm{sh}(\gamma_1)^{e_1}\,\mathrm{sh}(\gamma_2)^{e_2} = 1$ implies $\mathrm{sh}(\gamma_2)$ is conjugate to $\mathrm{sh}(-\gamma_1)$. The following lemma thus shows $\gamma_1 + \gamma_2 = 0$, completing the proof of (3).

LEMMA 2.10. *If $a \in \mathscr{D}^+$ is conjugate to $\mathrm{sh}(\gamma)$ then $\underline{m}a \le \gamma \le \bar{m}a$.*

Proof. By Lemma 2.1(5), $\lim_{n\to\infty} \underline{m}(a^n)/n = \gamma$. But by Lemma 2.1(4), $\lim \underline{m}(a^n)/n \ge \underline{m}(a)$. Thus $\underline{m}a \le \gamma$, and similarly $\bar{m}a \ge \gamma$.

Remark. For any $a \in \mathscr{D}^+$ the number $s(a) = \lim \underline{m}(a^n)/n = \lim \bar{m}(a^n)/n$ is a well defined (and well known) conjugacy invariant, and satisfies $\underline{m}a \le s(a) \le \bar{m}(a)$.

3. Seifert manifolds with transverse foliations

We consider a Seifert bundle $p : M \to F$ over a closed surface F, the fiber being S^1. The total space M may be orientable or not. Such a bundle may be described as follows: There exists some finite non-empty collection $D_1, \ldots, D_k$ of disjoint closed discs in F so that $p^{-1}(F - U \operatorname{int} D_i) \to F - U \operatorname{int} D_i$ is a locally trivial fibration admitting a section $s : F - U \operatorname{int} D_i \to M$, while $p^{-1}D_i \cong D^2 \times S^1$ is a solid torus. With each D_i is associated a coprime "Seifert" pair (α_i, β_i) of integers, with $\alpha_i \ge 1$, so that the class of $s(\partial D_i)$ in $\pi_1(p^{-1}D_i) = \mathbf{Z}$ is $-\beta_i$, and over D_i the map p is given in suitable coordinates by

$$p^{-1}D_i = D^2 \times S^1 \ni (re^{i\theta}, e^{i\psi}) \mapsto re^{i(\alpha_i\theta - \nu_i\psi)} \in D_i.$$

Here ν_i is an inverse of β_i modulo α_i and we are identifying D^2 and D_i with the

unit disc in $\mathbf{C}$. The integer α_i is thus the class in $\pi_1(p^{-1}D_i) = \mathbf{Z}$ of the "general fiber" $p^{-1}(u)$, $u \in D_i - \{0\}$.

We shall write the Seifert pairs as rational numbers β_i/α_i. The collection $\{\beta_1/\alpha_1, \ldots, \beta_k/\alpha_k\}$ of Seifert pairs associated to $p : M \to F$ is not unique, but depends on the choice of the D_i's and of s. By changing these choices, $\{\beta_i/\alpha_i\}$ can be changed in the following ways:

(a) permute the indices;

(b) add or delete a Seifert pair 0/1;

(c) alter each β_i/α_i by an integer, but keeping $\sum \beta_i/\alpha_i$ fixed;

(d) (only if M is non-orientable:) replace any β_i/α_i by $-\beta_i/\alpha_i$:

(e) (only if M is non-orientable:) replace any β_i/α_i by $(\beta_i \pm 2\alpha_i)/\alpha_i = (\beta_i/\alpha_i) \pm 2$.

If M is orientable, the Seifert fibration is completely classified (up to orientation preserving homeomorphisms) by the *Seifert invariant*

$$(g; \beta_1/\alpha_1, \ldots, \beta_k/\alpha_k),$$

where $g = g(F)$ is the genus of the base surface F (with the convention that g is negative for F non-orientable). The fact that the Seifert invariant is well determined up to (a), (b), and (c) above can also be formulated: M is classified by g, by the unordered set of β_i/α_i modulo 1 (omitting those that are zero modulo 1), and by $e(M \to F) = -\sum \beta_i/\alpha_i$. This number $e(M \to F)$ is called the *euler number* of the fibration, see [N–R].

Staying with the case M orientable, note that the Seifert invariant can always be put in the form

$$(g; \beta_0/1, \beta_1/\alpha_1, \ldots, \beta_k/\alpha_k),$$
$$0 < \beta_i/\alpha_i < 1 \quad \text{for} \quad i = 1, \ldots, k,$$

and this form, called *normal form*, is unique up to permutation of the indices $i = 1, \ldots, k$.

If M is non-orientable then one must include in the classifying data the information as to whether the fibers can be given mutually consistent orientations or not, and then the resulting "Seifert invariant" (well defined up to (a), (b), (c), (d), and (e)) classifies. We will not need a notation for this, so we do not introduce one. The euler number of the fibration is of course not defined in this case.

The above discussion is a modified presentation of Seifert's original classification [S], where the invariants are given in normalized form. The unnormalized version was introduced (in the oriented case) in [N] and [N–R].

We will say that M admits a *transverse foliation* if M has a codimension 1

foliation whose leaves are transverse to the fibers of $p : M \to F$. Foliations here are always C^0 foliations, however, when we give conditions below for existence of foliations, these foliations can actually be chosen to be analytic, by the results of the next section.

We can now state the main geometric results of this paper. Let $p : M \to F$ be a Seifert fibration over the closed surface F with Seifert pairs $\beta_1/\alpha_1, \ldots, \beta_k/\alpha_k$. Let χ be the euler characteristic of F, so $\chi = 2-2g$ if F is orientable (that is, $g \geq 0$) and $\chi = 2+g$ if F is non-orientable (that is $g < 0$).

THEOREM 3.1. *If M is non-orientable, then it admits a transverse foliation. This foliation can be chosen with all leaves compact.*

THEOREM 3.2. *If M is orientable and F is not the sphere (that is $\chi \neq 2$), then M admits a transverse foliation if and only if*

$$\sum \lfloor \beta_i/\alpha_i \rfloor \leq -\chi \quad \text{and} \quad \sum \lceil \beta_i/\alpha_i \rceil \geq \chi$$

or $M \to F$ has normal form Seifert invariant $(-1; 0/1)$. For a normal form invariant $(g; \beta_0/1, \beta_1/\alpha_1, \ldots, \beta_k/\alpha_k)$ these inequalities may be rewritten $\chi - k \leq \beta_0 \leq -\chi$.

THEOREM 3.3. *If $F = S^2$ (that is $\chi = 2$) and $\sum \beta_i/\alpha_i = 0$ or $\sum \lfloor \beta_i/\alpha_i \rfloor \leq -\chi = -2$ and $\sum \lceil \beta_i/\alpha_i \rceil \geq \chi = 2$, then M admits a tranverse foliation.*

Conversely, if $M \to S^2$ admits a transverse foliation then either $\sum \beta_i/\alpha_i = 0$ or $\sum \lfloor \beta_i/\alpha_i \rfloor \leq -1$ and $\sum \lceil \beta_i/\alpha_i \rceil \geq 1$ and at least one of these inequalities is strict.

Before proving 3.1 to 3.3, we pause to note the corresponding results for foliations with only compact leaves in case M is orientable. Recall that in this case the euler number of the fibration, $e(M \to F) = -\sum \beta_i/\alpha_i$, is defined.

THEOREM 3.4. *Suppose that M is orientable. The following statements are equivalent*

(i) $e(M \to F) = 0$;

(ii) *M has a transverse foliation with a compact leaf*;

(iii) *M has a smooth transverse foliation with all leaves compact.*

Proof. The euler number $e(M \to F)$ is the obstruction to finding a "rational section" to the Seifert fibration, that is, a closed compact surface $\bar{F} \looparrowright M$ immersed transverse to all fibers. This is implicit in [N–R], but can be seen explicitly as follows using the naturality properties of e of [N–R, Theorem 1.2] (which contained a misprint: "homeomorphism" should have been "continuous map"). By taking a double cover if necessary one can reduce to the case $g \geq 0$.

The Seifert fibration is then given by an S^1-action, and by factoring a $(\mathbf{Z}/a) \subset S^1$ one can reduce to the case of a genuine S^1-bundle. Finally, for a genuine S^1-bundle the existence of a rational section is equivalent, by standard obstruction theory, to the rational euler class in $H^2(F : \mathbf{Q})$ being trivial.

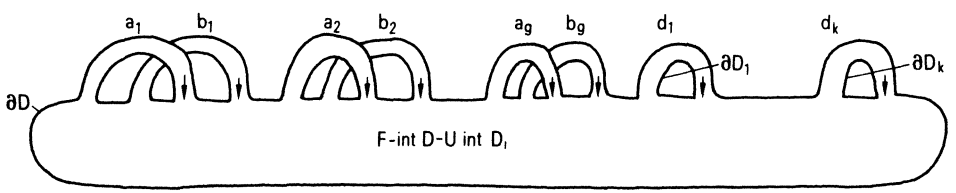
Thus (ii)$\Rightarrow$(i). Clearly (iii)$\Rightarrow$(ii).

Finally (i)$\Rightarrow$(iii) by Conner and Raymond [C–R]. Various approaches to this are also discussed in [N–R]. In fact, a complete set of models for the foliations given by (iii) is as follows. Those with orientable base are given by fiber products $M = Z \times_{\mathbf{Z}/a} S^1$, where $(X, \mathbf{Z}/a)$ is an oriented $(\mathbf{Z}/a)$-action on a closed surface and $\mathbf{Z}/a$ acts standardly on S^1. The foliation is by fibers of $M \to S^1/(\mathbf{Z}/a) \cong S^1$. Those with non-orientable base are given by fiber products $M = X \times_{D_{2a}} S^1$, where D_{2a} is the dihedral group $\langle t, g \mid t^2 = g^a = 1, t^{-1}gt = g^{-1} \rangle$ acting standardly on S^1 and acting on the closed oriented surface X in such a way that g is orientation preserving and t is free and orientation reversing on $X/\langle g \rangle$. The foliation is by fibers of $M \to S^1/D_{2a} \cong [0, 1]$.

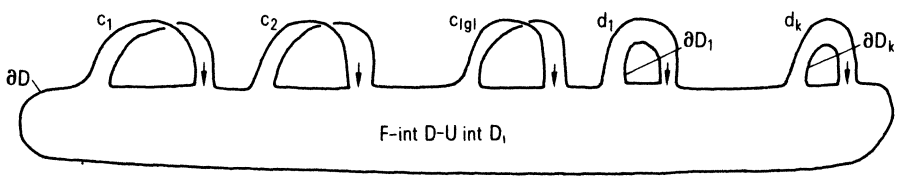
Generalizing the ideas of Milnor and Wood, we reduce 3.1–3 to algebraic theorems as follows.

Suppose $M \to F$ has a transverse foliation. We will utilize the notions of the introduction to this section, and in particular consider disks $D_1, \ldots, D_k$ in F so that $M \to F$ admits a section s over $F - U \operatorname{int} D_i$.

Cut F open along some additional disk, D, and open out $F - \operatorname{int} D - U \operatorname{int} D_i$ to represent it as a disk with oriented handles a_j, b_j, d_i:



in case F is orientable; or as a disc with handles c_i, d_i:



in case F is non-orientable.

Using the section s, we may trivialize the fibration over $F - \operatorname{int} D - \bigcup \operatorname{int} D_i$. Moreover we may assume this trivialization of the fibration agrees with the foliation over the disk $F - \operatorname{int} D - \bigcup \operatorname{int} D_i - \{\text{Handles}\}$. Then by following the

leaves of the foliation around a handle we may associate to each handle an element of $\mathscr{D}^+$, if the fiber orientation is preserved over a path going around that handle, or of $\mathscr{D}-\mathscr{D}^+$ if the fiber orientation is reversed over a path around the handle. The reason why the element corresponding to a handle is in $\mathscr{D}$ rather than Homeo (S^1) is that our section s on the handle tells us how to lift to $\mathscr{D}$.

Of course the fiber orientation is preserved around each handle d_i, and we see that the element of $\mathscr{D}^+$ associated to d_i is conjugate to the shift sh $(-\beta_i/\alpha_i)$. Indeed, in the coordinates in $p^{-1}D_i = D^2\times S^1$ introduced at the beginning of this section, let m denote a meridean $S^1\times\{1\}\subset\partial(D^2\times S^1)$, l denote a longitude $\{1\}\times S^1$, h denote a typical fiber $p^{-1}(u)$ with $u\in\partial D_i$, and q denote $s(\partial D_i)$. Then our explicit description of p implies the following homology relations in $H_1(\partial(D^2\times S^1))$: $h=\nu_i m+\alpha_i l$, $q=-(\mu_i m+\beta_i l)$, where μ_i is defined by $\nu_i\beta_i-\mu_i\alpha_i=1$. This implies the homology relation $m=\beta_i h+\alpha_i q$. But m represents the homology class of the intersection of a leaf of our foliation with $\partial(D^2\times S^1)$, so this homology relation says that as the leaf winds α_i times around the handle it also winds β_i times around the fiber in our given trivialization of the bundle structure on $\partial(D^2\times S^1)$. Thus in one circuit of the handle in the q-direction we get a conjugate of sh (β_i/α_i), so d_i is conjugate to sh $(-\beta_i/\alpha_i)$, as claimed.

We write a_j, b_j, c_j, and d_i again for the element of $\mathscr{D}$ associated to the corresponding handle. Since ∂D is the boundary of a disk, over which the fibration and foliation will be trivial, the element of $\mathscr{D}$ induced over ∂D must be 1. Thus, in the case F orientable, we have

$$\prod_{j=1}^{g} a_j b_j^{-1} a_j^{-1} b_j \prod_{i=1}^{k} d_i = 1, \tag{$*$}$$

while if F is non-orientable we get

$$\prod_{j=1}^{|g|} c_j^2 \prod_{i=1}^{k} d_i = 1. \tag{$**$}$$

Conversely, given elements a_j, b_j, d_i and $\mathscr{D}$ satisfying $(*)$ or c_j, d_i satisfying $(**)$, with d_i conjugate to sh $(-\beta_i/\alpha_i)$, the above discussion yields a construction of a transverse foliation for the corresponding Seifert fibration. Thus the problem of the existence of transverse foliations becomes the problem of finding appropriate factorizations $(*)$ or $(**)$.

Proof of Theorem 3.1. In this case, either F is orientable, and some a_j and/or b_j is in $\mathscr{D}-\mathscr{D}^+$, or F is non-orientable and some c_j lies in $\mathscr{D}^+$.

In the former case suppose it is a_l which is required to lie in $\mathscr{D}-\mathscr{D}^+$. We use Theorem 2.4 to find $a_l\in\mathscr{D}-\mathscr{D}^+$ and $b_l'\in\mathscr{D}^+$ such that $a_l(b_l')^{-1}a_l^{-1}b_l'$ is the shift by

$\sum \beta_i/\alpha_i$. If b_l is required to lie in $\mathscr{D}^+$ take $b_l = b_l'$; if b_l is required to lie in $\mathscr{D}-\mathscr{D}^+$ take $b_l = a_l b_l'$. If we choose the remaining a_j's and b_j's so their commutators are trivial we have satisfied equation $(*)$, with $d_i = \mathrm{sh}(-\beta_i/\alpha_i)$.

To see the statement about compact leaves, observe that we do not really need Theorem 2.4, we can work more explicitly. Namely choose a_l in the above argument to be $a_l = t$ with $tx = -x$, and choose $b_l' = \mathrm{sh}(\gamma)$ with $\gamma = (\sum \beta_i/\alpha_i)/2$. Since $t\,\mathrm{sh}(-\gamma)t^{-1}\,\mathrm{sh}(\gamma) = \mathrm{sh}(2\gamma)$, this does what is required, and clearly leads to a smooth foliation with compact leaves.

In the case that F is non-orientable it is enough to note that any shift has a square root in $\mathscr{D}^+$ to see that $(**)$ is solvable, and to choose this square root as a rational shift to get a smooth foliation with compact leaves.

Proof of Theorem 3.2. We first consider the case that F is orientable. Then our discussion shows that a transverse foliation exists if and only if we can represent some product of conjugates of the shifts $\mathrm{sh}(\beta_i/\alpha_i)$,

$$d = \prod_{i=1}^{k} \mathrm{sh}(\beta_i/\alpha_i)^{e_i},$$

as a product of g commutators. By Theorem 2.3 we can do this if and only if we can find d as above with

$$\underline{m}d < 2g-1 = 1-\chi$$
$$\bar{m}d > 1-2g = \chi-1.$$

But by Lemma 2.1, any d as above satisfies $\sum \lfloor \beta_i/\alpha_i \rfloor \leq \underline{m}d$ and $\sum \lceil \beta_i/\alpha_i \rceil \geq \bar{m}d$, so $\sum \lfloor \beta_i/\alpha_i \rfloor < 2g-1$ and $\sum \lceil \beta_i/\alpha_i \rceil > 1-2g$. Since these are inequalities between integers, the necessity of the condition in Theorem 3.2 is shown. Conversely, suppose the inequalities of Theorem 3.2 are satisfied. If the β_i/α_i are all integral then $d = \mathrm{sh}(\sum \beta_i/\alpha_i)$ does what is required. Otherwise $\sum \lfloor \beta_i/\alpha_i \rfloor < \sum \lceil \beta_i/\alpha_i \rceil$ and the inequalities $\sum \lfloor \beta_i/\alpha_i \rfloor \leq -\chi$ and $\sum \lceil \beta_i/\alpha_i \rceil \geq \chi$ imply that we can find $\sum \lfloor \beta_i/\alpha_i \rfloor < r < \sum \lceil \beta_i/\alpha_i \rceil$ with $\chi - 1 < r < 1-\chi$. Theorem 2.5 then implies the existence of the desired element d.

If F is non-orientable of genus $g \leq -2$, then the proof is just the same, using $(**)$ in place of $(*)$. We may thus assume that $g = -1$, so $\chi = 1$ and $F = \mathbf{R}P^2$. In this case condition $(**)$ becomes that the element d above can be chosen to equal the square of an element of $\mathscr{D}-\mathscr{D}^+$.

If we eliminate the trivial case that the β_i/α_i are all integral and $\sum \beta_i/\alpha_i = 0$, which is the exceptional case of Theorem 3.2, then the sufficiency of the conditions $\sum \lfloor \beta_i/\alpha_i \rfloor \leq -\chi = -1$ and $\sum \lceil \beta_i/\alpha_i \rceil \geq \chi = 1$ is just part (1) of Theorem 2.6

together with the observation, made in the proof of 2.7, that any element of SUF(1) is a square of an element of $\mathscr{D}-\mathscr{D}^+$.

To see the converse, observe first that if $c \in \mathscr{D}-\mathscr{D}^+$ with $c^2 \neq 1$, then $\underline{m}c^2 < 0 < \bar{m}c^2$ (since $c^2x < x \Leftrightarrow c^2x > cx$, since c reverses orientation). Thus if $d = c^2 \neq 1$ the inequalities $\sum \lfloor \beta_i/\alpha_i \rfloor \leq -1$ and $\sum \lceil \beta_i/\alpha_i \rceil \geq 1$ follow from Lemma 2.1, while if $d = 1$ then these inequalities follow from Theorem 2.6(3), unless $\sum \beta_i/\alpha_i = 0$. But in the latter case the inequalities still hold, unless the β_i/α_i are all integral, which is the exceptional case of Theorem 3.2.

Proof of 3.3. Here $(*)$ becomes $d = 1$, with d as in the previous proof, so Theorem 3.3 becomes parts (2) and (3) of Theorem 2.6.

There is an interesting corollary of the above proofs. Namely if F is orientable, then the fundamental group $\pi_1(M)$ can be presented as follows: it is generated by elements a_i, b_i, d_i, and z subject to the relation $(*)$ above and the additional relations

$$d_i^{\alpha_i} z^{\beta_i} = 1, \qquad d_i z = z d_i,$$
$$a_i z = z a_i^{\pm 1}, \qquad b_i z = z b_i^{\pm 1},$$

where the exponents are $+1$ or -1 according as the corresponding element a_j or b_j preserves or reverses fiber orientation in M. For F non-orientable the corresponding statement holds using elements c_j, d_i, z, and replacing $(*)$ by $(**)$. These are by an easy Van Kampen argument, see Seifert [5]. The element z is represented by a generic fiber of M; The other elements have their obvious geometric meanings. Our proof thus showed the following.

THEOREM 3.5. *The Seifert manifold M admits a transverse foliation if and only if there exists a homomorphism $\varphi : \pi_1(M) \to \mathscr{D}$ with $\varphi(z) = \mathrm{sh}(1)$, where $z \in \pi_1(M)$ is the class of a generic fiber of M.*

Theorems 3.1 to 3.3 thus give numerical conditions for existence of such a φ.

Given φ as in the above theorem, one can reconstruct M with its transverse foliation as follows. $G = \pi_1(M)/\langle z \rangle$ can be represented as a group of isometries of a geometry X, where X is S^2, euclidean space $\mathbf{R}^2$, or the hyperbolic plane $\mathbf{H}$. Thus $\pi_1(M)$ acts (non-effectively) on X. Via φ it also acts on $\mathbf{R}$, so it acts diagonally on $X \times \mathbf{R}$ and we can form the quotient space $X \times_{\pi_1(M)} \mathbf{R} = M$. The Seifert fibration is given by $X \times_{\pi_1(M)} \mathbf{R} \to X/\pi_1(M) = X/G = F$, and the foliation is induced from the foliation of $X \times \mathbf{R}$ by fibers of $X \times \mathbf{R} \to \mathbf{R}$.

Remark. The foliation will have all leaves compact if and only if the image of φ acts discretely on $\mathbf{R}$. One can check that φ can always be so chosen if $X = S^2$ or $\mathbf{R}^2$.

To close this section we discuss briefly the case of Seifert manifolds with boundary. In this case M always admits transverse foliations, but a specified transverse foliation on ∂M may not extend to one on M. The condition for such an extension to exist can be derived just as in the closed case. For simplicity we just discuss the case of orientable M.

Let the boundary components of M be denoted $T_1, \ldots, T_r$. We make the following simplifying assumption: $r \geq 1$ *and the foliation restricted to each* T_i *has non-compact leaves.* This is no loss of generality, since, if the foliation had only compact leaves on some T_i, then we could eliminate this boundary component by pasting in a solid torus over which the foliation extends.

Choose a section to the Seifert fibration on each boundary component T_j. Then a transverse foliation on T_j determines an element $h_j \in \mathscr{D}^+$, by taking holonomy in Homeo$^+$ (S^1) and using the section to lift to $\mathscr{D}^+$. This element is well defined up to conjugacy, so $\lfloor \underline{m} h_j \rfloor$ and $\lceil \bar{m} h_j \rceil$ are well defined. Moreover, our choice of sections on ∂M lets us define the Seifert invariant of M,

$$(g, r; \beta_1/\alpha_1, \ldots, \beta_k/\alpha_k)$$

(we now include the number of boundary components in our notation), well defined up to the usual indeterminacy, namely any $\beta_i/\alpha_i = 0/1$ can be added or deleted and each β_i/α_i can be altered by an integer so long as $\sum \beta_i/\alpha_i$ remains constant.

We put $\chi = 2-2g$ if $g \geq 0$ and $\chi = 2+g$ if $g < 0$, so χ is the euler characteristic of F with its boundary components capped by discs.

THEOREM 3.6. *If* $g \neq 0$ *or* -1 *then the inequalities* $\sum \lfloor \beta_i/\alpha_i \rfloor + \sum \lfloor \underline{m} h_j \rfloor \leq -\chi$ *and* $\sum \lceil \beta_i/\alpha_i \rceil + \sum \lceil \bar{m} h_j \rceil \geq \chi$ *are necessary and sufficient for the given foliation on* ∂M *to extend to a transverse foliation on* M. *This is true also for* $g = -1$ *unless the* β_i/α_i *are all integral and* $r = 1$. *If* $g = 0$ *these inequalities, with* χ *replaced by* 1, *are necessary, but not sufficient in general.*

Note that the inequalities are independent of the choice of sections on ∂M, since altering the section on T_j, say, alters $\lfloor \underline{m} h_j \rfloor$ and $\lceil \bar{m} h_j \rceil$ by an integer, and alters $\sum \lfloor \beta_i/\alpha_i \rfloor$ and $\sum \lceil \beta_i/\alpha_i \rceil$ by the negative of this integer.

The proof of this theorem is just like the closed case, so we omit it. We do not have complete results for $g = 0$. However, in the exceptional case with $g = -1$, omitted in the above theorem, we can assume, by changing the section on ∂M if

necessary, that the Seifert invariant is $(-1, 1; 0/1)$, and a necessary and sufficient condition for extension is that h_1 be the square of an element of $\mathscr{D}-\mathscr{D}^+$.

4. The projective linear group

In this section we show that the results of Section 2 also hold in the universal cover $\widetilde{\mathrm{PGL}}(2, \mathbf{R})$ of the projective linear group $\mathrm{PGL}(2, \mathbf{R})$. In particular this gives a rather pleasanter characterization of products of g commutators in $\widetilde{\mathrm{PSL}}(2, \mathbf{R})$ than that given by Wood [W]. It also shows that the transverse foliations whose existence is given by the theorems of Section 3 can always be chosen with "structure group" $\mathrm{PGL}(2, \mathbf{R})$, and in particular they can be chosen as analytic foliations.

We denote $\widetilde{\mathrm{PGL}}(2, \mathbf{R})$ by G. since $\mathrm{PGL}(2, \mathbf{R})$ acts on $\mathbf{R}P^1 = S^1$, we have an inclusion $\mathrm{PGL}(2, \mathbf{R}) \subset \mathrm{Homeo}(S^1)$, which induces an inclusion $G \subset \mathscr{D}$. Denote $G^+ = G \cap \mathscr{D}^+ = \widetilde{\mathrm{PSL}}(2, \mathbf{R})$. G^+ has index 2 in G with the nontrivial coset represented by the "flip" $tx = -x$.

Let $A = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \in \mathrm{PSL}(2, \mathbf{R})$. Then A has two fixed points on $\mathbf{R}P^1 = S^1$, so it has a unique lift $\tilde{A} \in G^+$ with fixed points. Let $K \subset G^+$ be the conjugacy class of $\tilde{A}$ (this is a slight modification of Wood's notation).

THEOREM 4.1. *Theorems* 2.3, 2.4, 2.5, 2.6 *and the remark after* 2.3 *are all valid with* $\mathscr{D}, \mathscr{D}^+$ *and* SUF(1) *replaced by* G, G^+ *and* K.

This can be deduced without too much trouble from a comparison of Wood's results with ours, but we give a self-contained version, based on the following improvement of Wood's Proposition 5.1.

LEMMA 4.2. *Given* $f, g \in G^+$ *with* $-1 < \bar{m}f$, $\underline{m}f < 0$, $0 < \bar{m}g$, $\underline{m}g < 1$, *there exist conjugates* f', g', *of* f, g *in* G^+ *with* $f'g' = \tilde{A}$. *In particular, by replacing either* f *or* g *by a conjugate one can achieve* $fg \in K$ *and* $gf \in K$ (*note* $gf = g(fg)g^{-1} \in K \Leftrightarrow fg \in K$).

Proof. Since $\mathrm{SL}(2, \mathbf{R})$ is a connected 2-fold cover of $\mathrm{PSL}(2, \mathbf{R})$, the group G^+ is also the universal cover of $\mathrm{SL}(2, \mathbf{R})$. Let $F, G \in \mathrm{SL}(2, \mathbf{R})$ be the images of f and g. We first show we can solve the relevant conjugacy problem in $\mathrm{SL}(2, \mathbf{R})$.

Elements $H \in \mathrm{SL}(2, \mathbf{R}) - \{\pm 1\}$ are classified up to conjugacy as follows. If $|\mathrm{tr}\, H| > 2$ then $\mathrm{tr}\, H$ classifies H up to conjugacy. For each value of $\mathrm{tr}\, H$ with $|\mathrm{tr}\, H| \leq 2$ there are exactly two conjugacy classes, distinguished as follows: there

exists $v \in \mathbf{R}^2$ such that v, Hv forms a basis of $\mathbf{R}^2$, and this basis will be an oriented basis for one conjugacy class and non-oriented for the other. The corresponding element $[H] \in \mathrm{PSL}(2, \mathbf{R})$ has either 2 fixed points on $\mathbf{R}P^1 = S^1$, one fixed point, or is conjugate to a rotation, according as $|\mathrm{tr}\, H| > 2$, $|\mathrm{tr}\, H| = 2$, or $|\mathrm{tr}\, H| < 2$.

Note also that, since $\mathrm{SL}(2, \mathbf{R})$ is the 2-fold cover of $\mathrm{PSL}(2, \mathbf{R})$, the element $1 \in \mathrm{SL}(2, \mathbf{R})$ lifts to $\mathrm{sh}(2n) \in G^+$ with $n \in \mathbf{Z}$ and $-1 \in \mathrm{SL}(2, \mathbf{R})$ lifts to $\mathrm{sh}(2n+1) \in G^+$ with $n \in \mathbf{Z}$.

We claim that our element F above is conjugate to a unique $F_0 \in \mathrm{SL}(2, \mathbf{R})$ of the form

$$F_0 = \begin{pmatrix} 0 & 1 \\ -1 & a \end{pmatrix}, \qquad a \in \mathbf{R},$$

and every such F_0 occurs. Of course, a is given by $a = \mathrm{tr}\, F$. Indeed, our condition on f is equivalent to saying $-1 < fr - r < 0$ for some $r \in \mathbf{R}$, in other words, there exists $v \in \mathbf{R}^2$ such that v, Fv is an unoriented basis of $\mathbf{R}^2$. Thus $F \in \mathrm{SL}(2, \mathbf{R}) - \{\pm 1\}$, and if $|\mathrm{tr}\, F| \leq 2$ then just the "negative" conjugacy class is permitted. Since F_0 is in this conjugacy class if $|a| \leq 2$, our claim follows.

Similarly G is conjugate to a unique G_0 of the form

$$G_0 = \begin{pmatrix} b & -\frac{1}{2} \\ 2 & 0 \end{pmatrix}.$$

Since $F_0 G_0 = \begin{pmatrix} 2 & 0 \\ * & \frac{1}{2} \end{pmatrix}$ is conjugate to $A = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$, by a further conjugation if necessary we can find F', G' conjugate to F, G such that $F'G' = A$.

Now let $f', g' \in G^+$ be the lifts of F', G' which satisfy $-1 < \bar{m} f'$, $\underline{m} f' < 0$, $0 < \bar{m} g'$, $\underline{m} g < 1$. They are unique, since lifts of elements of $\mathrm{SL}(2, \mathbf{R})$ are determined up to even integral shifts. Then f', g' are conjugate to f, g, and $f'g' = \mathrm{sh}(2n)\tilde{A}$ for some $n \in \mathbf{Z}$, since $f'g'$ is a lift of A. If F', G' are each conjugate to rotations, that is $|a| < 2$ and $|b| < 2$, then $-1 < \underline{m} f' \leq \bar{m} f' < 0$ and $0 < \underline{m} g' \leq \bar{m} g' < 1$, so $-1 < \underline{m}(f'g') \leq \bar{m}(f'g') < 1$. Thus in this case $n = 0$ and $f'g' = \tilde{A}$. Thus by continuity this holds for any value of a and b, completing the proof.

We now return to the proof of 4.1. We must first prove the analog of Lemma 2.7, with $\mathscr{D}, \mathscr{D}^+$, $\mathrm{SUF}(1)$ replaced by G, G^+, K. For the implication $(1) \Rightarrow (2)$, observe that if $\underline{m} f < 1$ and $\bar{m} f > -1$ then, by replacing f by f^{-1} if necessary we can assume f is as in Lemma 4.2. Then choosing $g \in K$, so $g^{-1} \in K$, Lemma 4.2 implies that f is the product of two elements of K. For the proof of $(2) \Rightarrow (4)$ we observe that $\tilde{A}$ is the square of a lift to G of the element $\begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix} \in \mathrm{PGL}(2, \mathbf{R})$, so any element of K is the square of an element of $G - G^+$. The rest of the proof of the

analogues of 2.7, 2.8, 2.3, 2.4 now are exactly as before. For the proof of the analogue of 2.5, note that the conjugating element a used in the proof of 2.5 can be chosen in G^+, so that proof also goes through with no change. Finally Lemma 2.9 is a special case of Lemma 4.2, so the proof of the analogue of 2.6 goes through with no change. This completes the proof of 4.1. Using Theorem 3.5 and the discussion after it, we get:

COROLLARY 4.3. *The necessary and/or sufficient conditions of Section* 3 *for the existence of a transverse foliation on a Seifert manifold* M *are also necessary and/or sufficient for the existence of an analytic such foliation, and for the existence of a homomorphism* $\varphi : \pi_1(M) \to G = \widetilde{\mathrm{PGL}}(2, \mathbf{R})$ *taking the class* h *of a non-singular fiber to sh* $(1) \in G$.

Remark. The homomorphism $\varphi : \pi_1(M) \to G$ of this corollary can practically never be found injective, with discrete image in G. Indeed this will be so if and only if the foliation is an Anosov foliation, discussed in the next section, which considerably restricts the possibilities for M.

Problem. Are the results of Section 2 valid also for groups between G and $\mathscr{D}$? This is particularly interesting for $C^r\mathscr{D} = \{f \in \mathscr{D} \mid f \text{ is } C^r\text{-smooth}\}$. All one needs is a suitable substitute for (1)$\Rightarrow$(2) in Lemma 2.7, the rest of the proof then runs itself.

5. Examples

We describe briefly examples of Maria Carmen del Gazolas [dG]. They generalize the well known Anosov foliation of a quotient $M = \Gamma\backslash\mathrm{PSL}(2, \mathbf{R})$ of $\mathrm{PSL}(2, \mathbf{R})$ by a discrete subgroup, induced by the foliation of $\mathrm{PSL}(2, \mathbf{R})$ by fibers of $\mathrm{PSL}(2, \mathbf{R}) \to \mathrm{PSL}(2, \mathbf{R})/U$, where $U \subset \mathrm{PSL}(2, \mathbf{R})$ is the subgroup of upper triangular matrices.

Let S be a compact surface with a riemannian metric with a finite number k of conical metric singularities with cone angles $2\pi\delta_i$, $i = 1, \ldots, k$. A neighborhood of such a point is obtained from a solid angle of measure $2\pi\delta_i$ by identifying its sides. Define the Euler characteristic of S by

$$\chi(S) = \chi(S_0) - \sum_{i=1}^{k} (1 - \delta_i)$$

where $\chi(S_0)$ is Euler characteristic of the underlying topological surface. The Gauss–Bonnet formula is $2\pi\chi(S) = \int_S K\, dv$, so if we assume S has constant curvature K, we see that K has the same sign as $\chi(S)$.

Now suppose each δ_i is a rational number γ_i/α_i. Then the unit tangent bundle $M = T^1S$ of S is well defined and is a Seifert bundle with an α_i-fold fiber over the i-th cone point. It is not hard to compute the Seifert invariant explicitly using the definition at the start of Section 3. Namely choose the section s to be a vector field which is radially outward at each of the cone points of S and has a unique singularity of index $\chi(S_0) - k$ otherwise. A simple calculation shows this leads to Seifert invariant

$$(g; (k - \chi(S_0))/1, -\gamma_1/\alpha_1, \ldots, -\gamma_k/\alpha_k)$$

which has normal form

$$(g; \beta_0/1, \beta_1/\alpha_1, \ldots, \beta_k/\alpha_k)$$

with

$$\beta_i = b_i\alpha_i - \gamma_i \qquad b_i = \lceil \gamma_i/\alpha_i \rceil > 0$$

$$\beta_0 = -b + k - \chi(S_0), \qquad b = \sum_1^k b_i.$$

Notice that $\beta_0 \leq -\chi(S_0)$. Moreover $e(M \to S) = \chi(S)$, so if S has non-positive curvature we get $e(M \to S) \leq 0$, or equivalently $-\sum_{i=1}^k (\beta_i/\alpha_i) \leq \beta_0$.

Now suppose S is as above with constant curvature ≤ 0, that is, it is hyperbolic or euclidean. Then the usual parallelism on $M = T^1S$ gives a foliation on M transverse to the fibers. Precisely, two vectors in euclidean or hyperbolic space are parallel if their directed geodesics stay a bounded distance apart in positive time. The unit tangent bundle of a euclidean or hyperbolic surface element is thus canonically foliated by families of parallel vectors. This foliates M away from its singular fibers, and this foliation is easily seen to extend across the singular fibers.

We have already observed that a necessary condition for an M with normal form invariant $(g; \beta_0/1, \beta_1/\alpha_1, \ldots, \beta_k/\alpha_k)$ to arise as one of these examples is

$$-\sum_{i=1}^k (\beta_i/\alpha_i) \leq \beta_0 \leq -\chi(S_0). \tag{5.1}$$

Since $-M$ has normal form invariant $(g; -k - \beta_0/1, (\alpha_1 - \beta_1)/\alpha_1, \ldots, (\alpha_k - \beta_k)/\alpha_k)$, the same condition for $-M$ becomes, after trivial simplification,

$$\chi(S_0) - k \leq \beta_0 \leq -\sum_{i=1}^k (\beta_i/\alpha_i). \tag{5.2}$$

If one shows that these necessary conditions are also sufficient, one has an alternate proof of sufficiency in Theorems 3.2 and 3.3. This presumably can be done in general; in particular, if the desired cone angles $2\pi\delta_i$ of S are all at most 2π, it is not hard to show the existence of S. This already gives sufficient examples to show that the sufficient condition of Theorem 3.3 for the $g=0$ case can be weakened as follows:

PROPOSITION 5.3. *If M is orientable and Seifert fibered over S^2, then a sufficient condition for M to admit a transverse foliation is that either* (5.1) *or* (5.2) *is valid. Here $\chi(S_0)=2$.*

It seems a reasonable conjecture that this condition is also necessary. The results of Greenberg [G] prove this for transverse foliations with structure group $\mathrm{PSL}(2,\mathbf{R})$ when $k=3$. We have some slight improvements of the necessary condition of Theorem 3.3 in general, but they do not come close to this conjecture, so we omit them.

The Anosov foliation mentioned at the beginning of this section are the special case of Maria Carmen del Gazolas' examples for cone angles of the form $2\pi/\alpha_i$.

6. Making foliations tranverse

Very many foliations of Seifert manifolds are isotopic to transverse foliations. We will state two results in this direction which generalize the results of Thurston [T] and Levitt [L] on S^1-bundles. The proofs rely on the results and techniques of [L] and [T], and we omit them. The possibility of making such a generalization was also observed by Johannson in discussions with one of the authors in 1976.

We consider a closed 3-manifold M with Seifert fibration $p: M \to F$ and a foliation $\mathcal{F}$ of M which is transversely orientable and C^2. We will suppose that no leaf of $\mathcal{F}$ is a torus or a Klein-bottle.

THEOREM 6.1. *If $\chi(F)<-1$ then $\mathcal{F}$ is homotopic to a foliation transverse to every fiber of p. The same is true for $\chi(F)\geq -1$ provided that*

(1) *$F\neq S^2$ or $\mathbf{RP}^2$,*

(2) *If $F=T^2$ then p has exceptional fibers, and*

(3) *either (a) The orientation of each fiber of p is preserved along each curve in F, or* (b) *The leaves of $\mathcal{F}$ are orientable.*

In proving Theorem 6.1, it is convenient to isotope $\mathcal{F}$ to make one leaf

transverse and then apply the following somewhat more general result:

THEOREM 6.2. *If $\mathcal{F}$ is transverse to one fiber of p then $\mathcal{F}$ is isotopic to a foliation transverse to every fiber of p.*

REFERENCES

[C-R] P. E. CONNER and F. RAYMOND, *Injective actions of toral groups*, Topology *10* (1971), 283–296.

[dG] MARIA CARMEN DEL GAZOLAS, *Foliations on Seifert bundles*, to appear.

[G] L. GREENBERG, *Homomorphisms of triangle groups into* PSL(2, **C**), *Riemann surfaces and Related Topics*, Proceedings of the 1978 Stony Brook conference (Princeton Univ. Press 1980), 167–181.

[L] G. LEVITT, *Feuilletages des variétés de dimension* 3 *qui sont des fibrés en circles.* Comment Math. Helv. *53* (1978), 572–594.

[Mi] J. MILNOR, *On the existence of a connection with curvature zero*, Comment. Math. Helv. *32* (1957–58), 215–223.

[Ne] W. D. NEUMANN, S^1*-actions and the* α*-invariant of their involutions*, Bonner Math. Schriften *44* (Bonn 1970).

[N-R] W. D. NEUMANN and F. RAYMOND, Seifert manifolds, plumbing, μ-invariant and orientation reversing maps, Alg. and Geom. Topology, Springer Lecture Notes in Math. *664* (1978), 163–196.

[No] S. P. Novikov, Topology of foliations, Trans. of the Moscow Math. Soc. for the year 1965, *14* (1967), 268–304.

[O] P. ORLIK, Seifert fibre spaces, Springer Lecture Notes in Math. *291* (1972).

[R] R. ROUSSARIE, *Plongements dans les variétés feuilletées et classification de feuilletages sans holonomie*, Publ. math. I.H.E.S. *43* (1974), 101–141.

[Se] H. SEIFERT, *Topologie dreidimensionaler gefaserter Raüme*, Acta. Math. *60* (1933), 147–238.

[Sch] A. J. SCHWARTZ, *A generalization of a Poincaré–Bendixson theorem to closed two dimensional manifolds*, Amer. J. Math. *85* (1963), 453–458.

[Su] D. SULLIVAN, *A generalization of Milnor's inequalities concerning affine foliations and affine manifolds*, Comment. Math. Helv. *51* (1976), 183–184.

[T] W. P. THURSTON, Foliations of 3-manifolds which are circle bundles, thesis, Berkeley 1972.

[W] J. W. WOOD, *Bundles with totally disconnected structure group*, Comment. Math. Helv. *46* (1971), 257–273.

Department of Mathematics
University of Maryland
College Park, MD 20742 USA

Received May 19, 1980, November 1981