Pacific Journal of Mathematics

Volume 172 No. 1 January 1996

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

A CLASS OF INCOMPLETE NON-POSITIVELY CURVED MANIFOLDS

B.H. BOWDITCH

In this paper, we describe a class of simply connected non-positively curved riemannian manifolds which satisfy some curvature constraints. Such manifolds have many of the properties of (complete) Hadamard manifolds, such as geodesic convexity and the existence of an ideal boundary.

1. Introduction.

The geometry of Hadamard (complete, simply-connected, non-positively curved riemannian) manifolds has been intensively studied for some time. A general account of the basic theory can be found in [**BaGS**]. However, there are interesting examples of non-positively curved manifolds which fail to be complete, while retaining many of the geometric properties of Hadamard manifolds. The best known is the Weil-Peterssen metric on Teichmüller space. This is negatively curved [**Ah**, **Tro**] and incomplete [**W1**], yet it admits an exhaustion by compact convex sets, and is thus geodesically convex [**W2**]. We describe some further examples in Chapter 2. Also, incomplete non-positively curved metrics have been used to construct interesting examples of complete non-positively curved manifolds by modifying the metric in a neighbourhood of the ends (see for example [**AbS**]).

These examples suggest that certain incomplete metrics may be of some interest in their own right. In this paper we restrict attention to metrics satisfying certain curvature constraints, and show that they behave, in many respects, like complete manifolds. We shall assume in particular that the curvature "blows up" along any path of finite length that leaves every compact set.

Let us first summarise a few properties of (complete) Hadamard manifolds. Firstly, the exponential map based at any point gives a diffeomorphism of $\mathbb{R}^n$ onto X. Moreover, there is a natural compactification, X_C, of X into a topological ball, formed by adjoining the *ideal sphere*, $X_I = X_C \backslash X$. A point of X_I may be thought of as an equivalence class of geodesic rays, where two rays are equivalent if they remain a bounded distance apart.

If, in addition, we assume that X has strictly negative curvature bounded away from 0, then it follows that X is a "visibility manifold", i.e. any two points of X_I may be joined by a bi-infinite geodesic [**EO**].

If we go further, and impose another curvature bound away from $-\infty$ (so that X has "pinched curvature"), then much more can be said about the geometry of X. For example, we have Anderson's result [**An**] that if $Q \subseteq X_C$ is any closed subset, and $\text{hull}(Q) \subseteq X_C$ is the closed convex hull of Q, then $X_I \cap \text{hull}(Q) = X_I \cap Q$. For further results about convex sets, see [**Bo**].

To generalise to incomplete (i.e. not necessarily complete) manifolds, let us assume that:

(A) X is a Riemannian manifold such that
(A1) X has non-positive curvature, and
(A2) X is simply connected.

We write d for the path-metric on X, and write $(\bar{X}, d)$ for the metric completion of (X, d). Given $x \in X$, write $\kappa(x)$ for the maximal sectional curvature of any tangent 2-plane at X.

Suppose we assume, in addition to (A), that:

(B) For all $a \in \bar{X} \backslash X$, there is some $K > 0$ and a neighbourhood U of a in $\bar{X}$ such that for all $x \in X \cap U$, we have $\kappa(x) \leq -1/K^2 d(x, a)^2$;

then, we claim that:

(1) X is geodesically convex. In fact, any two points $x, y \in \bar{X}$ may be joined by a geodesic segment $[x, y] \subseteq X \cup \{x, y\}$. Moreover, $[x, y]$ is, up to reparameterisation, uniquely length-minimising among all rectifiable paths in $\bar{X}$.

(2) The completion $\bar{X}$ is a CAT(0) space (as explained in Section 3.5).

(3) There is a natural compactification X_C of X so that X_C is homeomorphic to a closed ball, with X as its interior.

(4) There is a natural continuous injection $\iota : \bar{X} \longrightarrow X_C$ from $\bar{X}$ in the metric topology to X_C in its topology as a ball.

(5) Suppose $(x, y) \in (X_C \times X_C) \backslash (X_I^\infty \times X_I^\infty)$ where $X_I^\infty = X_C \backslash \iota(\bar{X})$. Then, x and y may be joined by a unique geodesic $[x, y] \subseteq X \cup \{x, y\}$, (where $[x, x] = \{x\}$). Moreover, $[x, y]$ is closed in X_C.

(6) The map $[(x, y) \mapsto [x, y]] : (X_C \times X_C) \backslash (X_I^\infty \times X_I^\infty) \longrightarrow \mathscr{C}(X_C)$ is continuous, where $\mathscr{C}(X_C)$ is the set of all closed subsets of X_C in the Hausdorff topology (Section 5.2).

Suppose, in addition to (A) and (B), that X satisfies:

(C) There exist $p_0 \in X$ and $L_0, R_0 > 0$, such that if $x \in X$ with $d(x, p_0) \geq R_0$, then $\kappa(x) \leq -1/L_0^2 d(x, p_0)^2$; then it follows also that:

(5′) If $(x, y) \in X_C \times X_C$, then x and y may be joined by a unique geodesic $[x, y] \subseteq X \cup \{x, y\}$.

(6′) The map $[(x, y) \mapsto [x, y]] : X_C \times X_C \longrightarrow \mathscr{C}(X_C)$ is continuous.

More precise statements of these results will be given later. They will all be proven in this paper: (1) Proposition 3.5.3. (2) Proposition 3.5.1, (3) Proposition 4.5.2, (4) Proposition 4.3.4, (5) Lemma 4.1.4, Lemma 5.3.1, (6) Proposition 5.3.4, (5′) Lemma 6.2.1, Proposition 6.2.3, (6′) Proposition 6.3.2.

If one adds additional hypotheses, such as pointwise pinching of curvature, then we have variations of Anderson's construction which enable us to construct convex sets in X. Thus, for example, with appropriate hypotheses, we can deduce that X has an exhaustion by compact convex sets. There is also the possibility of generalising some of the results of [**Bo**] to such spaces, though we shall not get involved with that here. Indeed we suspect that this programme could be carried further, and that, for example, many analytic results could be carried over to such spaces.

Note that in the complete case, pinched negative curvature is the same as pointwise negative curvature together with bounded geometry. "Bounded geometry" means that, for any fixed $r > 0$, the set of metric balls $\{N(x, r) \mid x \in X\}$ (defined up to isometry) all lie in a compact set in the C^2-topology. There is an analogous statement in the incomplete case. In this case, if X is negatively curved, properties (B) and (C) and pointwise pinching of curvature are all implied by a single hypothesis of "bounded geometry up to scale". To explain what we mean, let B be the closed unit ball in $\mathbb{R}^n$, with a standard orthonormal frame, F_0, at the origin, o. Let $\mathscr{S}$ be the space of smooth Riemannian metrics on B, with strictly negative curvature and with smooth boundary, ∂B, such that the frame F_0 is orthonormal in each metric, and such that ∂B is always the unit sphere about o. We give the space $\mathscr{S}$ the C^2 topology. Suppose that X satisfies (A). Suppose that $x \in X$, and $\lambda > 0$ is such that the ball $N(x, \lambda)$ is compact. Given any orthonormal frame, F, at x, let $e : B \longrightarrow N(x, \lambda)$ be the composition of a dilation by a factor of λ on $\mathbb{R}^n$ with the exponential map sending F_0 to F. Thus, e is a diffeomorphism, so we can pull back the metric on X to get a metric on B. This gives us a point of $\mathscr{S}$. We shall say that X has *bounded geometry up to scale* if there is a compact subset, $S \subseteq \mathscr{S}$, such that for all $x \in X$, we can choose $\lambda(x) > 0$ such that $N(x, \lambda(x))$ is compact, and such that for some frame at x, the the point of $\mathscr{S}$ constructed as above always lies in S. (Note that we are free to choose $\lambda(x)$ as small as we like. However, the sectional curvatures at the origin of metrics in S are all bounded away from 0. Thus, if $\lambda(x)$ is small, the scaling factor forces the curvature at x to be large. Similarly, if the curvature at x is small, then there must be a large compact metric ball centred on x.) We leave as an exercise the fact that this property implies properties (B) and (C).

As remarked earlier, one motive for studying incomplete manifold might

be to gain some further insight into the geometry of the Weil-Peterssen metric on the Teichmüller spaces. Wolpert [**W2**] shows that this is geodesically convex. As an example, he considers the case of once-punctured tori. In this case, the moduli space is a 2-dimensional Riemannian orbifold with two cone singularities (orbifold points), and a cusp singularity (with the cusp point removed), of the type obtained by spinning the graph of $f(x) = x^3, x > 0$ about the x-axis. It follows that the universal cover (i.e. Teichmüller space) in this case satisfies axioms (A) and (B) (see Chapter 2). For higher-dimensional spaces, the situation becomes more complicated. The asymptotics of the curvature tensor have been studied by Trapani [**Tra**]. It appears that in general property (B) fails. However, one might still hope for some modification of the hypothesis (B), for example, to take account to the directions of the tangent 2-planes along which the curvature blows up, sufficient to recover an ideal sphere analogous to Thurston's compactification.

In general, incomplete simply connected manifolds of negative curvature seem to have received little attention. Without some strong constraints on the curvature, they can behave in ways quite unlike Hadamard manifolds. For example, Hass [**Ha**] gives an example of a negatively curved metric on a 3-ball which contains a closed geodesic in its interior. This phenomenon is not possible in dimension 2, nor with constant curvature in any dimension. It might be interesting to explore further conditions under which this sort of behaviour would be prohibited.

2. Examples.

In this chapter we give some examples of the kind of incomplete manifolds we are considering. These particular examples have been chosen principally to illustrate the assertions made in the introduction. We begin with some manifolds satisfying properties (A) and (B).

Suppose $-\infty \leq a < b \leq \infty$, and that $f : (a, b) \longrightarrow (0, \infty)$ is a smooth function. Let t be an arc-length parameter along the graph of f, $\text{graph}(f) \subseteq (a, b) \times (0, \infty)$. Given $t \in \text{graph}(f)$ write $\rho(t) \in \mathbb{R} \cup \{\infty\}$ for the length of the tangent at this point to the intercept with the x-axis. (Figure 2.) We take the sign of $\rho(t)$ to be the same as that of df/dt. We may form a surface of revolution, S, by spinning $\text{graph}(f)$ about the x-axis. Now, S has two orthogonal foliations: one by generators of S which are intrinsically geodesic, and the other by circles of curvature $c(t) = 1/\rho(t)$. We see that S has Gaussian curvature equal to

$$k(t) = -\frac{dc}{dt} - c^2 = \frac{1}{\rho^2}\left(\frac{d\rho}{dt} - 1\right) = -\frac{1}{f}\frac{d^2 f}{dt^2}.$$

Thus, for S to non-positively curved, we need that f be convex. Such a surface, S, has two topological ends corresponding to the ends of the interval

(a, b). We see that the end corresponding to a will be complete if and only if $a = -\infty$, or else $a > -\infty$ and $f(x) \to \infty$ as $x \to a$. We call such an end a *tube*. If we have $a > -\infty$ and $f(x) \to 0$ and $\frac{df}{dx}(x) \to 0$ as $x \to a$, then we call the end a *cusp*. If S satisfies property (B), we see that it is necessary (but not sufficient) that either both the ends of S be tubes, or that one end be a tube, and the other be a cusp.

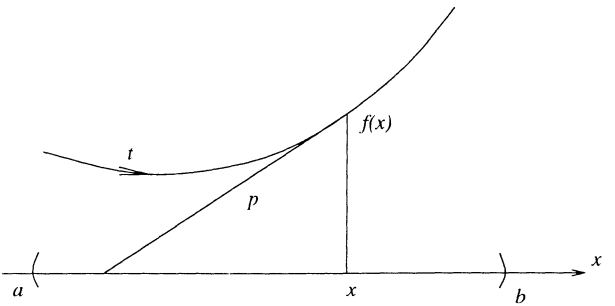


Figure 2.

As an explicit example, consider the graph of $f(x) = x^\beta$ for some $\beta > 1$, defined on the interval $(0, \infty)$. We have $-\frac{1}{f(x)}\frac{d^2 f}{dx^2}(x) = -\beta(\beta - 1)x^{-2}$. Now $x/t \to 1$ as $t \to 0$, and so the curvature of S blows up like $-1/t^2$ as we approach the cusp point at 0. We see that S satisfies (A1) and (B), and so its universal cover, $X = \tilde{S}$ satisfies (A) and (B). The metric completion $\bar{X}$ of X is obtained by adding a single point, p, at the origin 0. Thus, under the natural inclusion $\iota : \bar{X} \to X_C$, the point p maps to an ideal point $\iota(p) \in X_I$. The remaining ideal points can be thought of as the endpoints of the geodesic generators of X, as $t \to \infty$. Thus, the set X_I^∞ of these remaining ideal points has naturally the topology of an open interval. This is compactified into the circle, X_I, by adding the point $\iota(p)$.

Suppose, more generally, that $f : (0, b) \longrightarrow (0, \infty)$ is convex, and that $f(x) \to 0$ and $\frac{df}{dx}(x) \to 0$ as $x \to 0$. Then $\mu = \lim_{t\to\infty} \frac{df}{dt}(t) \in (0, 1]$ is well defined. (Thus $\mu = 1$ if $b \leq \infty$.) Let S be the surface of revolution, and $X = \tilde{S}$ the universal cover. We may coordinatise X using a radial coordinate $\theta \in \mathbb{R}$ and an arc length coordinate $t \in (0, \infty)$. In this way, S is the quotient of X by the map $[(t, \theta) \mapsto (t, \theta + 2\pi)]$. As before, X_C is formed by adjoining the arc $\{(\infty, \theta) \,|\, \theta \in \mathbb{R}\}$, and then taking the one-point compactification with the point 0 at the origin. Let l_θ be the geodesic generator $\{(t, \theta) \,|\, t \in (0, \infty)\}$ of X. The total Gauss curvature of the sector of X lying between l_{θ_1} and l_{θ_2} may be calculated as $C(\theta_0) = -\int_0^\infty (\theta_0 f)\left(\frac{1}{f}\frac{d^2 f}{dt^2}\right) dt = -\mu\theta_0$ where $\theta_0 = \theta_2 - \theta_1$. Applying Gauss-Bonnet, we find that the ideal points (∞, θ_1) and (∞, θ_2) can be joined by a bi-infinite geodesic in X if and only if $C(\theta_0) < -\pi$

i.e. if and only if $\theta_0 > \pi/\mu$. Now, $\mu \leq 1$, and so X cannot have the visibility property. Note that $\frac{d\rho}{dt} \to 1$ as $t \to \infty$, and so $k(t) = \frac{1}{\rho^2}\left(\frac{d\rho}{dt} - 1\right) = o(1/t^2)$. Thus Property (C) fails in this case.

By giving similar consideration to the case where both ends of S are tubes, we see that no surface constructed in this way can satisfy all of properties (A), (B) and (C).

The surfaces of revolution just described are a special case of the following more general construction.

Suppose M is a Riemannian manifold, and that $I \subseteq \mathbb{R}$ is an open interval. Let $f : I \longrightarrow (0, \infty)$ be a smooth function. We define a Riemannian metric on $X = M \times I$ by setting

$$ds^2 = dt^2 + f(t)^2 \sum_{i,j} g_{ij} dx^i dx^j,$$

where t is arc length in I, g_{ij} is the Riemannian metric on M with respect to the local coordinate system $(x^i)_i$, and ds is infinitesimal distance in X.

We remark that this is an example of a still more general construction of "warped products" described in the paper of Bishop and O'Neill [**BiO**]. In a warped product, the interval I may be replaced by any non-positively curved manifold. In the paper cited, there is a complete characterisation of when a warped product is non-positively curved.

In our special case, we can derive the relevant inequalities fairly simply as follows. Note that X has two orthogonal foliations, one by geodesics of the form $\{x\} \times I$ for $x \in M$, and the other by codimension-1 submanifolds of the form $M_t = M \times \{t\}$ for $t \in I$. Each M_t is totally umbilic, with principal curvatures equal to $c(t) = \frac{1}{f(t)} \frac{df}{dt}(t)$. In the intrinsic metric, M_t is isometric to M with the metric scaled by a factor of $f(t)$.

Write $A = \partial/\partial t$ for the vector field on X orthogonal to the M_t. Now suppose that Π is a tangent 2-plane at $(x,t) \in M \times I = X$. If Π is orthogonal to $A(x,t)$, then Π corresponds to a tangent 2-plane, Π_M at x in M. Write $S_M(\Pi_M)$ for the sectional curvature of M in Π_M. Thus, the sectional curvature, in Π, of the intrinsic metric of M_t is $S_M(\Pi_M)/f(t)^2$. Applying Gauss's Theorema Egregium [**S**], we see that the sectional curvature, $S(\Pi)$, of X in Π is given by

$$S(\Pi) = \frac{1}{f^2} S_M(\Pi_M) - c^2 = \frac{1}{f^2}\left(S_M(\Pi_M) - \left(\frac{df}{dt}\right)^2\right).$$

On the other hand, suppose that Π is a tangent 2-plane at (x,t) containing the vector $A(x,t)$. In this case the sectional curvature, $S(\Pi)$, of X in Π is

$$S(\Pi) = -\frac{dc}{dt} - c^2 = -\frac{1}{f}\frac{d^2 f}{dt^2}.$$

Now, if Y and Z are, locally, any two vector fields everywhere orthogonal to A, then a simple calculation shows that $R(A,Y,Y,Z) = 0$, where R is the Riemann curvature tensor. This symmetry implies that each sectional curvature of X at (x,t) lies between $-\frac{1}{f}\frac{d^2f}{dt^2}$ and $\frac{1}{f^2}S_M(\Pi_M) - c^2 = \frac{1}{f^2}\left(S_M(\Pi_M) - \left(\frac{df}{dt}\right)^2\right)$ for some tangent 2-plane Π_M at x in M. In particular, for X to be non-positively curved, it is sufficient that M be non-positively curved, and that f be convex. (For more detailed computations of this nature, see [**BiO**].)

Examples of this construction are the surfaces of revolution described above. In this case, we have $M = \mathbb{R}$ and f is thought of as a function of arc-length, t, along graph$(f) \equiv I$. In such a case, we must always have $\frac{df}{dt} < 1$.

With this last constraint removed, we can construct examples satisfying (A), (B) and (C). For example, with $M = I = \mathbb{R}$, and $f(t) = e^t$, we obtain the hyperbolic plane foliated by horospheres.

For another example, set $M = \mathbb{R}$, $I = (0,\infty)$ and $f(t) = t^\beta$ with $\beta > 1$. Now, the curvature $k(t)$ equals $-\frac{1}{f}\frac{d^2f}{dt^2} = -\beta(\beta-1)/t^2$. This case is qualitatively similar to the surface of revolution of $[x \mapsto x^\beta]$ described above, except that now, X satisfies (C), and has the visibility property.

As a third example, set $M = \mathbb{R}$, $I = (0,1)$ and $f(t) = t^2/(1-t)^2$. We see that $k(t) = -\frac{1}{f}\frac{d^2f}{dt^2} = -2(2t+1)/t^2(1-t)^2$. Thus $-k(t)$ grows like $1/t^2$ as $t \to 0$ and like $1/(1-t)^2$ as $t \to 1$. It follows that X satisfies (A) and (B). Since it is bounded (has finite diameter), it trivially satisfies (C). Both the completion, $\bar{X}$, and the compactification, X_C, of X may be identified set-theoretically as $(\mathbb{R} \times [0,1])/\sim$, where $(x,0) \sim (y,0)$ for all $x,y \in \mathbb{R}$. However, the topologies are different. Thus X_C may be thought of as the one point compactification of $X \times (0,1]$ by adding the point $0 \equiv \{(x,0)\}/\sim$, whereas $\bar{X}$ is noncompact—a base of neighbourhoods of 0 being given by $\{(\mathbb{R}\times[0,\epsilon))/\sim \,|\, \epsilon > 0\}$. Note that the natural map $\bar{X} \longrightarrow X_C$ is a continuous bijection.

One can construct higher dimensional examples, for example by taking M to be euclidean n-space $\mathbb{E}^n$, or hyperbolic n-space $\mathbb{H}^n$. Note that $M = \mathbb{E}^n$, $I = \mathbb{R}$ and $f(t) = e^t$ gives us $\mathbb{H}^{n+1}$. So does $M = \mathbb{H}^n$, $I = \mathbb{R}$ and $f(t) = \cosh t$.

There are many variations on this theme one can explore. One can also go on to construct further examples by gluing together examples of this type.

3. Geodesic convexity.

In this chapter, we aim at establishing properties (1) and (2) for manifolds satisfying (A) and (B). The following notation is used throughout.

Suppose X is a Riemannian manifold. We write T_xX for the tangent space to X at x, and TX for the total space of the tangent bundle. Given $\xi, \zeta \in T_xX$, we write $\langle \xi, \zeta \rangle$ and $|\xi| = \sqrt{\langle \xi, \xi \rangle}$ respectively for the Riemannian inner product and norm on T_xX. If $\xi, \zeta \neq 0$, set $\angle(\xi, \zeta) = \cos^{-1}(\langle \xi, \zeta \rangle / |\xi||\zeta|) \in [0, \pi]$ for the angle between ξ and ζ. We write d for the induced path-metric on X.

We shall use the term "geodesic" in the Riemannian sense of a curve whose first derivative is parallel. Thus, in terms of the metric d, a geodesic can be characterised as a constant-speed path, for which all sufficiently small subpaths are length-minimising.

3.1. Ruled maps. In this section we take X to be a Riemannian manifold of non-positive curvature (A1). For $x \in X$, we write $\kappa(x) \in [-\infty, 0]$ to be the maximal sectional curvature at x.

Suppose that $I = [t_0, t_1] \subseteq \mathbb{R}$ is a closed interval and $J \subseteq \mathbb{R}$ is any interval. We write $\operatorname{int} I$ and $\operatorname{int} J$ respectively for the interiors of I and J. Given a smooth map $\beta : I \times J \longrightarrow X$, we shall denote by β_u and β^t the maps

$$\beta_u = [t \mapsto \beta(t, u)] : I \longrightarrow X$$

and

$$\beta^t = [u \mapsto \beta(t, u)] : J \longrightarrow X$$

where $t \in I$ and $u \in J$. Thus $\beta_u(t) = \beta^t(u) = \beta(t, u)$. We refer to paths of the form β_u and β^t respectively as *longitudes* and *transversals*. We write $\partial\beta/\partial t$ and $\partial\beta/\partial u$ respectively for $\beta_*(\partial/\partial t)$ and $\beta_*(\partial/\partial u)$. We say that β is a *ruled map* if for all $u \in I$, β_u is a geodesic. Thus $\left|\frac{\partial\beta}{\partial t}(t, u)\right| = (\operatorname{length} \beta_u)/|t_1 - t_0|$.

Suppose that for $u \in J$, the geodesic β_u is non-constant. We see that the map $\left[t \mapsto \frac{\partial\beta}{\partial u}(t, u)\right]$ is the first variation of a geodesic along β_u. Thus, the component of $\frac{\partial\beta}{\partial u}(t, u)$ parallel to $\frac{\partial\beta}{\partial t}(t, u)$ is linear in t. Moreover, since X is non-positively curved, the Riemannian norm of the component orthogonal to $\frac{\partial\beta}{\partial t}(t, u)$ is convex (see the discussion of normalised ruled maps below). It follows that the map $\left[t \mapsto \left|\frac{\partial\beta}{\partial u}(t, u)\right|\right]$ is convex. This is also readily verified in the case where β_u is constant. Integrating, we find that the map $[t \mapsto \operatorname{length} \beta^t] : J \longrightarrow [0, \infty)$ is convex. In particular:

Lemma 3.1.1. *For all $t \in [t_0, t_1]$, we have*

$$\operatorname{length} \beta^t \leq \max(\operatorname{length} \beta^{t_0}, \operatorname{length} \beta^{t_1}).$$

We shall say that a ruled map $\beta : I \times J \longrightarrow X$ is *non-degenerate* if β_u is non-constant for all $u \in J$. In such a case, we say that $(t,u) \in I \times J$ is a *singular point* if β fails to be an immersion at that point, i.e. if $\frac{\partial \beta}{\partial u}(t,u)$ is some multiple of $\frac{\partial \beta}{\partial t}(t,u)$. We say that β is *non-singular* if there are no singular points in $\operatorname{int} I \times J$. In such a case, the pull back of the Riemannian metric to $\operatorname{int} I \times \operatorname{int} J$ is also a Riemannian metric of non-positive curvature. In fact, the curvature at (t,u) is at most $\kappa(\beta(t,u))$. This is Synge's Inequality (see [**S**]). In the particular context of ruled maps, it is discussed in a paper of Aleksandrov [**Alek**].

By a *ruled surface*, we shall mean the image, $P = \beta(I \times J) \subseteq X$, of a ruled map $\beta : I \times J \longrightarrow X$, where J is compact, and such that β is non-singular and injective on $\operatorname{int} I \times \operatorname{int} J$. We shall refer to the sets $\beta(I \times \{u\})$ for $u \in J$ as *generating geodesics*. We write $\kappa_P(x)$ for the intrinsic curvature of P at x. Thus $\kappa_P(x) \leq \kappa(x) \leq 0$. Of particular interest is the case where the boundary, ∂P, of P is a piecewise geodesic path. This motivates the following definition.

Definition. By a (*non-positively curved*) *n-gon* we mean a surface P, which is topologically a closed disc with boundary ∂P, together with a set $V \subseteq \partial P$ of n points, and a metric, ρ on P such that ρ restricted to the interior $\operatorname{int} P = P \backslash \partial P$ is a non-positively curved Riemannian metric, and such that each component of $\partial P \backslash V$ is geodesic.

We shall refer to the points of V as *vertices* and the components of $\partial P \backslash V$ as *edges*. At each vertex $v \in V$, the adjacent edges meet at some well-defined angle $\theta(v) \geq 0$. Since the metric is not assumed to be Riemannian at the point v itself, it may be possible to have $\theta(v) = 0$ (if the curvature grows sufficiently fast as we approach v). In such a case, we refer to v as a *cusp*. In all cases we consider, P will be convex, i.e. $\theta \leq \pi$ for all $v \in V$. Now, the Gauss-Bonnet formula tells us that

$$\sum_{v \in V} \theta(v) = (n-2)\pi + \int_P \kappa_P(x) d\omega(x),$$

where $\kappa_P(x)$ is the curvature at $x \in P$, and $d\omega$ is the area element. Note that we must always have $n \geq 3$.

By talking about ruled surfaces, we avoid having to worry about the technical complication of dealing with singular points; although intuitively we would expect such points to work in our favour since they concentrate negative curvature. The fact that singular points do not cause any real problems has been made precise by Aleksandrov [**Alek**].

Another another type of restriction we shall want to place on ruled maps is the following.

We say that a non-degenerate ruled map $\beta : I \times J \longrightarrow X$ is *normalised* if:

(R1) for all $u \in J$, the longitude $\beta_u = [t \mapsto \beta(t,u)]$ is a geodesic parameterised with respect to arc-length (i.e. $\left|\frac{\partial\beta}{\partial t}(t,u)\right| = 1$ for all (t,u)); and

(R2) for all $(t,u) \in I \times J$, we have

$$\left\langle \frac{\partial\beta}{\partial t}(t,u), \frac{\partial\beta}{\partial u}(t,u) \right\rangle = 0.$$

Thus, for a fixed u, the map $\left[t \mapsto \frac{\partial\beta}{\partial u}(t,u)\right]$ is a Jacobi field along the longitude β_u. We write $J(t) = \left|\frac{\partial\beta}{\partial u}(t,u)\right|$. From the Jacobi field equation [**S**], we know that, except where it vanishes, $J(t)$ is smooth in t, and that

$$\frac{d^2J}{dt^2}(t) \geq -\kappa(\beta(t,u))J(t).$$

Suppose that $\lambda : I \longrightarrow [0,\infty)$ satisfies $\lambda(t) \leq -\kappa(\beta(t,u))$ for all $t \in I = [t_0,t_1]$. The following is a simple consequence of the above differential inequality.

Proposition 3.1.2. *Suppose $f : I \longrightarrow [0,\infty)$ is smooth and satisfies $\frac{d^2f}{dt^2}(t) = \lambda(t)f(t)$ for all $t \in I$. If $f(t_0) = J(t_0)$ and $\frac{df}{dt}(t_0) \leq \frac{dJ}{dt}(t_0)$ then $f(t) \leq J(t)$ for all $t \in I$.*

Corollary 3.1.3. *Suppose $f : I \longrightarrow [0,\infty)$ is smooth and satisfies $\frac{d^2f}{dt^2}(t) = \lambda(t)f(t)$ for all $t \in I$. If $f(t_0) = J(t_0)$ and $f(t_1) = J(t_1)$, then $f(t) \leq J(t)$ for all $t \in I$.*

Of particular interest will be the case where λ has the form

$$\lambda(t) = 1/K^2(t+h)^2$$

for $t \geq 0$, and $K, h > 0$ fixed. The solutions of $\frac{d^2f}{dt^2}(t) = \lambda(t)f(t)$ have the form $(t+h)^{1+\mu}$ and $(t+h)^{-\mu}$ where $\mu = (\sqrt{1+4K^2}) - 1 > 0$. In particular, if $f(0) = 1$ and $\frac{df}{dt}(0) = 0$ we have the solution

$$f(t) = \frac{\mu}{2\mu+1}\left(\left(1+\frac{t}{h}\right)^{1+\mu} + \left(1+\frac{1}{\mu}\right)\left(1+\frac{t}{h}\right)^{-\mu}\right).$$

We shall refer to this later (Lemmas 3.4.1 and 6.1.1).

For the proof Lemma 3.4.1, we will need to describe a process of "normalising" ruled maps.

Suppose that $\alpha : I \times J \longrightarrow X$ is a non-degenerate ruled map, where now $I = [v_0, v_1]$. We are looking for a subset $S \subseteq \mathbb{R} \times J$ and a map $\rho : I \times J \longrightarrow S$ with the following properties:

(N1) ρ is a smooth diffeomorphism of $I \times J$ onto S.

(N2) For all $u \in J$, the set $S \cap (\mathbb{R} \times \{u\})$ is a closed interval of the form $[q_0(u), q_1(u)] \times \{u\}$.

(N3) For all $u \in J$, the map $\rho|(I \times \{u\})$ sends $I \times \{u\}$ linearly onto $[q_0(u), q_1(u)] \times \{u\}$.

(N4) The map $\beta = \alpha \circ \rho^{-1} : S \longrightarrow X$ is a normalised ruled map (i.e. it satisfies properties (R1) and (R2) above.)

We see that S has the form $S = \{(t, u) \in \mathbb{R} \times J \mid q_0(u) \leq t \leq q_1(u)\}$, where $q_0, q_1 : J \longrightarrow \mathbb{R}$ are smooth maps.

As before, we define longitudes, α_u, β_u, and transversals α^v, β^t, by $\alpha^v(u) = \alpha_u(v) = \alpha(v, u)$ and $\beta^t(u) = \beta_u(t) = \beta(t, u)$. For $i = 0, 1$, set $\gamma_i = \alpha^{v_i} : J \longrightarrow X$, and $\sigma_i = [u \mapsto (q_i(u), u)] : J \longrightarrow S$. Thus, $\gamma_i = \beta \circ \sigma_i$. We see that

$$\frac{d\gamma_i}{du}(u) = \frac{dq_i}{du}(u)\frac{\partial\beta}{\partial t}(\sigma_i(u)) + \frac{\partial\beta}{\partial u}(\sigma_i(u)),$$

and so

$$\frac{dq_i}{du}(u) = \left\langle \frac{d\gamma_i}{du}(u), \frac{\partial\beta}{\partial t}(\sigma_i(u)) \right\rangle.$$

Note that $\frac{\partial\beta}{\partial t}(\sigma_i(u))$ is the unit tangent vector $\xi_i(u) = \left(\frac{v_1 - v_0}{l(u)}\right)\frac{\partial\alpha}{\partial v}(v_i, u)$ to the geodesic α_u, where $l(u) = \operatorname{length}\alpha_u = \operatorname{length}\beta_u$.

Now, suppose that we are given α, and want to construct S and ρ, and hence β. We can obtain the functions q_i, up to an additive constant, by integrating the quantity $\left\langle \frac{d\gamma_i}{du}(u), \xi_i(u) \right\rangle$. Note that $\frac{d}{du}(q_1(u) - q_0(u)) = \frac{dl}{du}(u)$, and so we can arrange that $q_1(u) - q_0(u) = l(u)$ for all $u \in J$. This, then, defines the set $S \subseteq \mathbb{R} \times J$, and hence determines the map $\rho : I \times J \longrightarrow S$. One verifies that the map $\beta = \alpha \circ \rho^{-1}$ satisfies properties (R1) and (R2) as required.

3.2. The space of geodesics. For the moment, we can take X to be any Riemannian manifold. Let $(\bar{X}, d)$ be the metric completion of (X, d). Since (X, d) is a path-metric space it follows that $(\bar{X}, d)$ is a path-metric space. We claim that every point of $\bar{X} \setminus X$ is accessible by a smooth path of finite length:

Lemma 3.2.1. *Suppose $y \in \bar{X} \setminus X$; then there is a smooth path $\beta : [0, 1] \longrightarrow \bar{X}$ so that $\beta(0) = y$, $\beta((0, 1]) \subseteq X$ and* $\operatorname{length}\beta < \infty$.

Proof. Certainly, y is accessible by a rectifiable path of finite length in X, and we may use local convexity to approximate it by a smooth path. □

Now, write $\operatorname{path}(\bar{X})$ for the set of all paths from $[0, 1]$ to $\bar{X}$. Given $\alpha, \beta \in \operatorname{path}(\bar{X})$, write

$$d_{sup}(\alpha, \beta) = \max\{d(\alpha(t), \beta(t)) \mid t \in [0, 1]\}.$$

Thus d_{sup} is a metric on $\text{path}(\bar{X})$. We see easily that:

Proposition 3.2.2. $(\text{path}(\bar{X}), d_{sup})$ *is a complete metric space.*

We write $\text{path}(X) \subseteq \text{path}(\bar{X})$ for the subspace of paths lying entirely in X.

We define the *endpoint map*

$$\pi : \text{path}(\bar{X}) \longrightarrow \bar{X} \times \bar{X}$$

by $\pi(\beta) = (\beta(0), \beta(1))$. Clearly π is continuous.

Let $\text{geod}(\bar{X}) \subseteq \text{path}(\bar{X})$ be the subspace of those $\beta \in \text{path}(\bar{X})$ such that either β is constant, or else $\beta((0,1)) \subseteq X$ and $\beta|(0,1)$ is a constant-speed geodesic. Let

$$\text{geod}(X) = \text{geod}(\bar{X}) \cap \text{path}(X) = \text{geod}(\bar{X}) \cap \pi^{-1}(X \times X).$$

Now, let us suppose that X is non-positively curved (A1). In this case, the map $\pi : \text{geod}(X) \longrightarrow X \times X$ is a local homeomorphism:

Lemma 3.2.3. *Suppose* $b \in \text{geod}(X)$. *Let* $\pi(\beta) = (x, y)$. *Then, there are neighbourhoods* U *of* x *and* V *of* y *in* X, *and a neighbourhood* W *of* β *in* $\text{geod}(X)$ *such that* $\pi|W : W \longrightarrow U \times V$ *is a homeomorphism.*

Proof. This follows, exactly as in the complete case, using the Jacobi field equation, and the implicit function theorem. □

We see that, if X has dimension n, then $\text{geod}(X)$ is a $2n$-dimensional manifold, and inherits a smooth structure from $X \times X$.

Suppose that $\gamma : J \longrightarrow \text{geod}(X)$ is a smooth path. By definition, the paths $\gamma_i = [u \mapsto \gamma(u)(i)] : J \longrightarrow X$ for $i = 0, 1$ are smooth. We write $\hat{\gamma} : [0,1] \times J \longrightarrow X$ for the map given by $\hat{\gamma}(t, u) = \gamma(u)(t)$.

Lemma 3.2.4. *The map* $\hat{\gamma}$ *is smooth.*

Proof. From the implicit function theorem, exactly as in the complete case. □

Thus, $\hat{\gamma}$ is a ruled map. Note that $\gamma_i = \hat{\gamma}^i$ according to our previous notation. Applying Lemma 3.1.1, we see that γ is a rectifiable path in $(\text{path}(\bar{X}), d_{sup})$. In fact, if $J' \subseteq J$ is any subinterval, then

$$\text{length}\,(\gamma|J') \leq \max\left(\text{length}(\gamma_0|J'), \text{length}(\gamma_1|J')\right).$$

Since $(\text{path}(\bar{X}), d_{sup})$ is complete, we have the following:

Lemma 3.2.5. *Suppose* $\gamma : (0,1] \longrightarrow \mathrm{geod}(X)$ *is smooth, and* $\mathrm{length}\,\gamma_i < \infty$ *for* $i = 0, 1$. *Then,* γ *extends (uniquely) to a map* $\gamma : [0,1] \longrightarrow \mathrm{path}(\bar{X})$.

Suppose, in such a case, it happens that $\gamma(0)((0,1)) \subseteq X$, so that $\gamma(0)|(0,1)$ must be geodesic. Thus, by definition, $\gamma(0) \in \mathrm{geod}(\bar{X})$. Our aim in the next section is to show that this is always the case if X satisfies axiom (B), and $\gamma(0)$ is non-constant.

3.3. The path-lifting property. Suppose that X is non-positively curved (A1) and satisfies:

(B) For all $a \in \bar{X} \setminus X$, there is some $K > 0$ and a neighbourhood U of a in $\bar{X}$ such that for all $x \in X \cap U$ we have $\kappa(x) \leq -1/K^2 d(x,a)^2$.

We aim to show that $\pi : \mathrm{geod}(X) \longrightarrow X \times X$ is a covering map. A similar idea can be found in [**AlexB**]. This result will be based on the following path-lifting property.

Lemma 3.3.1. *Suppose* $\gamma : [0,1] \longrightarrow \mathrm{path}(\bar{X})$ *with* $\gamma((0,1]) \subseteq \mathrm{geod}(X)$, *and* $\gamma|(0,1]$ *smooth. For* $i = 0, 1$, *write* γ_i *for the path* $[u \mapsto \gamma(u)(i)] : [0,1] \longrightarrow \bar{X}$. *Suppose that for* $i = 0, 1$, *we have* $\mathrm{length}\,\gamma_i < \infty$. *Then* $\gamma(0) \in \mathrm{geod}(\bar{X})$.

Proof. By definition, any constant path lies in $\mathrm{geod}(\bar{X})$, so we can suppose that $\gamma(0)$ is non-constant. As remarked at the end of the last section, it suffices to show that $\gamma(0)((0,1)) \subseteq X$. Without loss of generality, we can suppose that $\gamma(u)$ is non-constant for all $u \in [0,1]$. Define $\alpha : [0,1]^2 \longrightarrow \bar{X}$ by $\alpha(v,u) = \gamma(u)(v)$. Thus, $\alpha : [0,1] \times (0,1]$ is a non-degenerate ruled map. Now, the normalising procedure of Section 3.1 gives us a map $\rho : [0,1] \times (0,1] \longrightarrow \mathbb{R} \times (0,1]$ so that $\beta = \alpha \circ \rho^{-1} : S_0 \longrightarrow X$ is a normalised ruled map, where $S_0 = \rho([0,1] \times (0,1]) = \{(t,u) \in \mathbb{R} \times (0,1] \mid q_0(u) \leq t \leq q_1(u)\}$. We have $\gamma_i|(0,1] = \beta \circ \sigma_i$ where $\sigma_i(u) = (q_i(u), u)$. Thus, for all $u \in (0,1]$,

$$\frac{d\gamma_i}{du}(u) = \frac{dq_i}{dt}(u)\frac{\partial\beta}{\partial t}(\sigma_i(u)) + \frac{\partial\beta}{\partial u}(\sigma_i(u))$$

and so

$$\left|\frac{dq_i}{dt}(u)\right| \leq \left|\frac{d\gamma_i}{du}(u)\right|.$$

We see that $\int_0^1 \left|\frac{dq_i}{dt}(u)\right| du \leq \mathrm{length}\,\gamma_i < \infty$, and so $q_i(u)$ tends to a limit, $q_i(0)$, as u tends to 0. Also, since $l(u) = \mathrm{length}\,\alpha_u = q_1(u) - q_0(u)$ for all $u \in (0,1]$, and since $\alpha_0 = \gamma(0)$ is non-constant, we see that $q_0(0) < q_1(0)$. Let $S = \{(t,u) \in \mathbb{R} \times [0,1] \mid q_0(u) \leq t \leq q_1(u)\}$. We may extend ρ to a homeomorphism $\rho : [0,1]^2 \longrightarrow S$ mapping $[0,1] \times \{0\}$ linearly to $[q_0(0), q_1(0)] \times \{0\}$. Thus, β extends to a map $\beta = \alpha \circ \rho^{-1} : S \longrightarrow \bar{X}$. As

before, we define longitudes, β_u, and transversals, β^t, by $\beta_u(t) = \beta^t(u) = \beta(t,u)$. We want to show that $\beta_0((q_0,q_1)) = \gamma(0)((0,1)) \subseteq X$.

Suppose, for contradiction, that there is some $t \in (q_0,q_1)$ with $\beta(t,0) \in \bar{X} \setminus X$. For notational convenience, we shall assume that $t = 0$, i.e. that $\beta(0,0) \in \bar{X} \setminus X$. Let $a = \beta(0,0)$.

Let U be the neighbourhood of a in $\bar{X}$ given by the hypothesis (B) above. We can find $t_0 > 0$ and $u_0 > 0$ such that $[-t_0,t_0]\times[0,u_0] \subseteq S$ and $\beta([-t_0,t_0]\times[0,u_0]) \subseteq U$.

Now, for all $(t,u) \in S$, we have that

$$\left|\frac{\partial\beta}{\partial u}(t,u)\right| \leq \max_{i=0,1}\left|\frac{\partial\beta}{\partial u}(\sigma_1(u))\right| \leq \max_{i=0,1}\left|\frac{d\gamma_i}{du}(u)\right|.$$

The first inequality follows from Corollary 3.1.3 (with $\lambda \equiv 0$) and the second comes from the formula for $\frac{d\gamma_i}{du}(u)$ given above. In particular, we see that for all $t \in [-t_0,t_0]$,

$$\int_0^{u_0}\left|\frac{\partial\beta}{\partial u}(t,u)\right| du \leq \max_{i=0,1} \text{length}(\gamma_i|[0,u_0]) < \infty.$$

Given $u \in [0,u_0]$, set

$$h(u) = \int_0^u \left|\frac{\partial\beta}{\partial u}(0,w)\right| dw.$$

Thus, $h(u) = \text{length}(\beta^0|[0,u]) \geq d(a,\beta(0,u))$. Since the longitude β_u is a geodesic parameterised by arc-length, we have, for all $t \in [-t_0,t_0]$

$$d(\beta(0,u),\beta(t,u)) = |t|$$

and so

$$d(a,\beta(t,u)) \leq |t| + h(u).$$

Thus, by hypothesis (B), we have

$$\kappa(\beta(t,u)) \leq -1/K^2(|t| + h(u))^2.$$

Fix, for the moment, some $u \in (0,u_0]$. For $t \in [-t_0,t_1]$ set $J(t) = \left|\frac{\partial\beta}{\partial u}(t,u)\right|$. If $J(0) \neq 0$, then J is differentiable at 0. Suppose $\frac{dJ}{dt}(0) \geq 0$. Then, applying Proposition 3.1.2 on the interval $[0,t_0]$ and using the formula given after the Proposition, we find that

$$J(t_0) \geq \frac{\mu}{2\mu+1}\left(1 + \frac{t_0}{h(u)}\right)^{1+\mu} J(0).$$

If, on the other hand, $\frac{dJ}{dt}(0) \leq 0$, then, by symmetry, we get the same lower bound for $J(-t_0)$. Thus, in all cases, we get that

$$J(-t_0) + J(t_0) \geq \frac{\mu}{2\mu+1}\left(1 + \frac{t_0}{h(u)}\right)^{1+\mu} J(0).$$

Thus,

$$\begin{aligned}
\infty &> \int_0^{u_0} \left|\frac{\partial\beta}{\partial u}(-t_0, u)\right| du + \int_0^{u_0} \left|\frac{\partial\beta}{\partial u}(t_0, u)\right| du \\
&\geq \frac{\mu}{2\mu+1} \int_0^{u_0} \left(1 + \frac{t_0}{h(u)}\right)^{1+\mu} \left|\frac{\partial\beta}{\partial u}(0, u)\right| du \\
&= \frac{\mu}{2\mu+1} \int_0^{u_0} \left(1 + \frac{t_0}{h(u)}\right)^{1+\mu} \frac{dh}{du}(u) du \\
&= \frac{\mu}{2\mu+1} \int_0^{h(u_0)} \left(1 + \frac{t_0}{w}\right)^{1+\mu} dw \\
&= \infty.
\end{aligned}$$

This contradicts the existence of $a \in \gamma(0)((0,1)) \cap (\bar{X} \backslash X)$. Thus $\gamma(0)((0,1)) \subseteq X$, and so $\gamma(0) \in \mathrm{geod}(\bar{X})$ as required. □

Corollary 3.3.2. *The map $\pi : \mathrm{geod}(X) \longrightarrow X \times X$ is a covering map.*

Proof. By Lemma 3.2.3, we know that π is a local homeomorphism. Lemmas 3.2.5 and 3.3.1 together tell us that π has the path-lifting property for smooth paths. The result follows by standard arguments. □

3.4. Properties of geodesics. In this section we shall add the assumption (A2) that X is simply connected, i.e., altogether we are assuming that X satisfies hypotheses (A) and (B).

Now, $X \times X$ is simply connected, and so by Corollary 3.3.2, we see that each component of $\mathrm{geod}(X)$ maps homeomorphically to $X \times X$ under π. Choose any point $x_0 \in X$, and let $\mathrm{geod}_0(X)$ be the component of $\mathrm{geod}(X)$ containing the constant path at x_0. Let π_0 be the restriction of π to $\mathrm{geod}_0(X)$ so that $\pi_0 : \mathrm{geod}_0(X) \longrightarrow X \times X$ is a homeomorphism. Given $x, y \in X$, write $[x \to y] = \pi_0^{-1}(x, y)$. We see easily that for all $x \in X$, $[x \to x]$ is the constant path at x.

Lemma 3.4.1. $\mathrm{geod}(X) = \mathrm{geod}_0(X)$.

Proof. Suppose, for contradiction, that $\mathrm{geod}(X) \neq \mathrm{geod}_0(X)$. Choose any $x \in X$. Since $\pi : \mathrm{geod}(X) \longrightarrow X \times X$ is a covering map, there is some

$\alpha \in \text{geod}(X) \backslash \text{geod}_0(X)$ with $\pi(\alpha) = (x, x)$. Thus $\alpha \neq [x \to x]$. Without loss of generality can suppose that $x \notin \alpha((0,1))$. (Otherwise choose a smaller segment of α and reparameterise.) For each $t \in (0,1)$, the path α meets the path $[x \to \alpha(t)]$ in $\alpha(t)$, at an angle different from 0 or π. Thus, as t ranges through $[0,1]$, the geodesics $[x \to \alpha(t)]$ span a non-positively curved 1-gon, which is impossible by Gauss-Bonnet (Section 3.1). □

In summary, we have shown:

Proposition 3.4.2. *Any two points of X are joined by a unique geodesic (defined on the domain $[0,1]$). Moreover, this geodesic varies smoothly in its endpoints.*

Given $x, y \in X$, write $[x, y] \subseteq X$ for the image of $[x \to y]$. Thus $[x, x] = \{x\}$ and $[x, y] = [y, x]$.

For a fixed $x \in X$, the function ρ defined by $\rho(z) = \text{length}\,[x \to z]$ is smooth on $X \backslash \{x\}$. Moreover, any geodesic $[x \to y]$ is orthogonal to the level sets of ρ, and so a standard argument of Riemannian geometry shows that:

Proposition 3.4.3. *For all $x, y \in X$, the geodesic $[x \to y]$ is, up to reparameterisation, the unique length-minimising rectifiable path from x to y. (In particular, $d(x,y) = \text{length}\,[x \to y]$.)*

Now, given $x \in X$ and $y \in X \setminus \{x\}$, we write $\overrightarrow{xy} = \frac{1}{d(x,y)} \frac{d\alpha}{dt}(0)$, where $\alpha = [x \to y]$. In other words, $\overrightarrow{xy}$ is the unit tangent vector at x along $[x, y]$. If $z \in X \setminus \{x\}$, write $y\hat{x}z = \angle(\overrightarrow{xy}, \overrightarrow{xz})$ for the angle between $\overrightarrow{xy}$ and $\overrightarrow{xz}$.

Given the existence and uniqueness of geodesics, the following comparison theorems follow exactly as in the complete case. Let $(\mathbb{E}^2, d')$ be the euclidean plane,

Proposition 3.4.4. (Angle Comparison Theorem of Aleksandrov). *Suppose $x, y, z \in X$ are distinct points. Choose $x', y', z' \in \mathbb{E}^2$, so that $d'(x', y') = d(x, y)$, $d'(y', z') = d(y, z)$ and $d'(z', x') = d(z, x)$. Then $x\hat{y}z \leq x'\hat{y}'z'$, $y\hat{z}x \leq y'\hat{z}'x'$ and $z\hat{x}y \leq z'\hat{x}'y'$.*

We refer to $x'y'z'$ as a *comparison triangle* for xyz.

Proposition 3.4.5. (CAT(0) inequality). *Suppose $x, y, z \in X$ are distinct points. Suppose $u \in [x, y]$ and $v \in [x, z]$. Choose a comparison triangle $x'y'z'$ in $\mathbb{E}^2$ for xyz. Let $u' \in [x', y']$ and $v' \in [x', z']$ be the points with $d'(x', u') = d(x, u)$ and $d'(x', v') = d(x, v)$. Then $d'(u', v') \leq d(u, v)$.*

We thus say that (X, d) is a "CAT(0)-space". More precisely, a CAT(0)-space is a path-metric space in which every pair of points may be joined

by a "geodesic", in the sense of a length-minimising path, and where the conclusion of Proposition 3.4.5 is satisfied where $[x, y]$ may be interpreted as any choice of geodesic from x to y. In fact, it follows, in retrospect, that in a CAT(0)-space, there is a unique geodesic joining any pair of points, and so $[x, y]$ is uniquely defined. For further discussion of such spaces, see Ballmann's article in Chapter 10 of [**GH**], or the book by Bridson and Haefliger [**BrH**].

As a corollary of Proposition 3.4.5, we have the convexity of the distance function:

Proposition 3.4.6. *Suppose $I, J \subseteq \mathbb{R}$ are intervals, and that $\alpha : I \longrightarrow X$ and $\beta : J \longrightarrow X$ are geodesics parameterised proportionately to arc-length. Then the function $[(t, u) \mapsto d(\alpha(t), \beta(u))] : I \times J \longrightarrow [0, \infty)$ is convex.*

3.5. The completion. Finally in this chapter, we describe the geometry of the completion $(\bar{X}, d)$ of (X, d). We are again assuming that X satisfies hypotheses (A) and (B).

Now, the metric completion of any CAT(0)-space is a CAT(0)-space, so we see immediately that:

Proposition 3.5.1. *$(\bar{X}, d)$ is a CAT(0)-space.*

In particular, every pair of points are joined by a unique geodesic. Recall, however, that the term "geodesic" is here being used in the metric space sense of a constant-speed globally length-minimising path. We should therefore check that this agrees with the notion of "geodesic" already defined in Section 3.2. As before, we write $\mathrm{geod}(\bar{X})$ for the space of such geodesics.

Note that it's easy to see that a path $\alpha \in \mathrm{geod}(\bar{X})$ is globally length-minimising, in other words, that $\mathrm{length}\, \alpha = d(x, y)$ where $(x, y) = \pi(\alpha)$. To do this, choose $t \in (0, \frac{1}{2}]$. Since geodesics in X are globally length-minimising (Proposition 3.4.3), we have that $\mathrm{length}\, (\alpha|(t, 1-t)) = d(\alpha(t), \alpha(1-t))$. The observation follows by letting $t \to 0$. Now, since $(\bar{X}, d)$ is CAT(0), it now follows that if $\alpha, \beta \in \mathrm{geod}(\bar{X})$ with $\pi(\alpha) = \pi(\beta)$, then $\alpha = \beta$. (This can also be verified directly, by a similar limiting argument.) It remains to show that such paths always exist:

Lemma 3.5.2. *Any two points of $\bar{X}$ can be joined by a path in* $\mathrm{geod}(\bar{X})$.

Proof. Suppose $x, y \in \bar{X}$. Since every constant path lies in $\mathrm{geod}(X)$, we can suppose that $x \neq y$. By Lemma 3.2.1, both x and y are accessible by smooth paths of finite length in X. From the geodesic convexity of X (Proposition 3.4.2) and Lemma 3.3.1, we see that x and y can be joined by a path in $\mathrm{geod}(\bar{X})$. □

In summary, we have shown:

Proposition 3.5.3. *For all $x, y \in \bar{X}$, there is a unique $\alpha \in \mathrm{geod}(\bar{X})$ with $\pi(\alpha) = (x, y)$. Moreover* $\mathrm{length}\,\alpha = d(x, y)$. *In fact, α is the unique constant-speed globally length-minimising path in $\bar{X}$ from x to y.*

We can now use the term "geodesic" without ambiguity. As with X, we write $[x \to y]$ for the unique path in $\mathrm{geod}(\bar{X})$ joining x to y. We write $[x, y] \subseteq \bar{X}$ for the image of $[x \to y]$. As before, $[x, y] = [y, x]$ and $[x, x] = \{x\}$. Note that for all $x, y \in \bar{X}$, we have $[x, y] = \{z \in \bar{X} | d(x, z) + d(z, y) = d(x, y)\}$.

Note that since $(\bar{X}, d)$ is CAT(0), it follows that the distance function is convex (cf. Lemma 3.4.6). In particular, geodesics vary continuously on their endpoints, and so:

Proposition 3.5.4. *The map $\pi : \mathrm{geod}(\bar{X}) \longrightarrow \bar{X} \times \bar{X}$ is a homeomorphism.*

We remark that if we fix one endpoint, then geodesics vary in a C^1 fashion:

Proposition 3.5.5. *Given $a \in \bar{X}$, define $f_a : X \times (0, 1) \longrightarrow X$ by $f_a(x, t) = [a \to x](t)$. Then f_a is C^1.*

Proof. Clearly, if $a \in X$, then f_a is smooth. If $a \in \bar{X} \setminus X$, we choose a sequence of points $a_n \in X$ with $a_n \to a$, and check that the derivatives of the functions f_{a_n} converge. This can be done by considering Jacobi fields along $[x, a_n]$ (c.f. the case of horofunctions [**Hel**]). □

4. The compactification.

In this chapter, we assume that X satisfies axioms (A) and (B). We shall describe the compactification $X_C = X \cup X_I$, where X_I is the "ideal sphere". Thus, X_I may be thought of, set theoretically, as the union of $X_I^0 \equiv \bar{X} \setminus X$ and a set, X_I^∞ of asymptote classes of geodesic rays. We shall show that X_C is homeomorphic to a closed ball (Proposition 4.5.2.)

4.1. Geodesic rays. A *geodesic ray* based at $x \in \bar{X}$ is a path $\alpha : [0, \infty) \longrightarrow \bar{X}$ such that $\alpha(0) = x$, and $\alpha((0, \infty)) \subseteq X$, and such that $\alpha | (0, \infty)$ is a geodesic parameterised by arc length.

We know (Proposition 3.5.3) that geodesics are length-minimising in $\bar{X}$. In particular, α must be a proper map.

Suppose α, β are geodesic rays. By Lemma 3.4.6, the map

$$[t \mapsto d(\alpha(t), \beta(t))]$$

is convex. Thus, if $d(\alpha(t), \beta(t))$ is bounded above, then

$$d(\alpha(t), \beta(t)) \leq d(\alpha(0), \beta(0)) \quad \text{for all } t.$$

Definition. We say that the rays α and β are *asymptotic* if $d(\alpha(t), \beta(t))$ is bounded as $t \longrightarrow \infty$.

Clearly this is an equivalence relation on the set of geodesic rays. Note that:

Lemma 4.1.1. *If α and β are asymptotic rays, then the map*

$$[t \mapsto d(\alpha(t), \beta(t))]$$

is monotonically non-increasing.

Corollary 4.1.2. *Two asymptotic rays based at the same point are equal.*

Proposition 4.1.3. *Suppose that β is a geodesic ray, and $x \in \bar{X}$. Then there is a (unique) geodesic ray based at x asymptotic to β.*

Proof. For this, we need only the convexity of the distance function (Lemma 3.4.6), and the completeness of $\bar{X}$.

For $n \in \mathbb{N}$, set $l_n = d(x, \beta(n))$. Let $\alpha_n : [0, l_n] \longrightarrow \bar{X}$ be the geodesic from x to $\beta(n)$ parameterised by arc-length. Note that $n - l_0 \leq l_n \leq n + l_0$. From Lemma 3.4.6 applied to β and α_n, we see that $d(\alpha(t), \beta(t)) \leq l_0$ provided $t \leq n - l_0$. Thus, if $m \geq n \geq l_0$, then $d(\alpha_n(n - l_0), \alpha_m(n - l_0)) \leq 2l_0$. Now, by Lemma 3.4.6 applied to α_n and α_m, we see that for all $t \in [0, n - l_0]$, we have $d(\alpha_n(t), \alpha_m(t)) \leq \frac{2l_0 t}{n - l_0}$. Thus, for a fixed t, the sequence $(\alpha_n(t))$ is a Cauchy sequence, and so tends to a limit $\alpha(t) \in \bar{X}$. Now each α_n is length-minimising, and so $d(\alpha(t), \alpha(u)) = |t - u|$ for all $t, u \in [0, \infty)$. Thus by Proposition 3.5.3, we see that $\alpha((0, \infty)) \subseteq X$ and $\alpha|(0, \infty)$ is geodesic. For all $n \geq t + l_0$, we have $d(\beta(t), \alpha_n(t)) \leq l_0$, and so $d(\alpha(t), \beta(t)) \leq l_0$. Thus α and β are asymptotic. □

Now, let X_I^∞ be the set of asymptote classes of geodesic rays. We write X_I^0 for the set $\bar{X} \setminus X$, and define the *ideal sphere*, X_I, as a disjoint union $X_I = X_I^0 \sqcup X_I^\infty$. We write $X_C = X \sqcup X_I$ for the *compactification* of X, and $\iota : \bar{X} \longrightarrow X_C$ for the natural inclusion. We shall describe the topology on these spaces in Section 4.3.

Suppose that $x \in \bar{X} \equiv X \cup X_I^0$ and that $y \in X_I^\infty$. Lemma 4.1.3 tells us that there is a unique geodesic ray β based at x and in the class y. We say that β *tends to* the point y. Write $[x, y] = \beta([0, \infty)) \cup \{y\} \subseteq X_C$, and refer to $[x, y]$ as the *geodesic joining* x to y. Given the existence and uniqueness of geodesics in $\bar{X}$, we have established that:

Lemma 4.1.4. *Given $(x, y) \in X_C \times X_C \setminus (X_I^\infty \times X_I^\infty)$, then there is a unique geodesic $[x, y]$ joining x to y.*

We may extend the notations $\overrightarrow{xy}$ and $y\hat{x}z$ to the case where $x \in X$ and $y, z \in X_C \setminus \{x\}$.

Note that from the proof of Proposition 4.1.3, we see that if $z, x \in X$, $y \in X_I^\infty$, and $y_n \in [z,y] \cap X$ is a sequence of points tending to y, then the vectors $\overrightarrow{xy_n}$ tend to $\overrightarrow{xy}$ in the unit tangent space at x.

If we fix $y \in X_I^\infty$, then the vector field $[x \mapsto \overrightarrow{xy}] : X \longrightarrow TX$, is C^1, where TX is the total tangent bundle to X. This may be proven using the convergence of Jacobi fields just as in the complete case. We may also define a positive-time flow $\phi : X \times [0,\infty) \longrightarrow X$ along this field. Thus, $\phi(x,t) = \beta(t)$, where β is the geodesic ray based at x tending to y. As in the complete case, we have:

Proposition 4.1.5. *The flow $\phi : X \times [0,\infty) \longrightarrow X$ is C^2.*

4.2. Horofunctions. In this section, we describe the "horofunctions" (or "Busemann functions") about a point $y \in X_I^\infty$. The results will be used again in Chapter 6, though, for the moment, it is something of a digression.

Fix $y \in X_I^\infty$. Suppose $a \in \bar{X}$. Let β be the geodesic ray based at a tending to y. Given any $x \in \bar{X}$, the function $[t \mapsto t - d(x, \beta(t))]$ is monotonically increasing in t. Moreover it is bounded above (by $d(x,a)$). It thus tends to a well-defined limit $h_a(x) = \lim_{t\to\infty}(t - d(x,\beta(t)))$. We see easily that $|h_a(x) - h_a(x')| \leq d(x,x')$ for all $x, x' \in \bar{X}$. Thus, $h_a : \bar{X} \longrightarrow \mathbb{R}$ is continuous. Also, one can show that h_a is C^2. This follows as in the complete case (see [**HeI**]). We refer to h_a as a *horofunction* about y.

To see that h_a is at least C^1 on X is elementary. For a fixed t, write $f_t(x) = t - d(x,\beta(t))$. Thus $f : X \longrightarrow \mathbb{R}$ is smooth on X, and its gradient, $\operatorname{grad} f_t$ at x equals $\overrightarrow{xy_t}$ where $y_t = \beta(t)$. From the Angle Comparison Theorem (Proposition 3.4.4) we can verify that $\overrightarrow{xy_t}$ tends to $\overrightarrow{xy}$ as $t \to \infty$. Moreover, this convergence is uniform on compact subsets of X. Thus f is C^1, and $\operatorname{grad} f(x) = \overrightarrow{xy}$.

As a consequence, we may deduce that any two horofunctions about y differ by a constant.

Lemma 4.2.1. *If $a, b, x \in \bar{X}$, then $h_b(x) = h_b(a) + h_a(x)$.*

Proof. From the previous paragraph, we know that for all $x \in X$, we have $\operatorname{grad}(h_b - h_a)(x) = 0$, and so $h_b - h_a$ is constant on X. By continuity, it is constant on all of $\bar{X}$. Since $h_a(a) = 0$, we must have $h_b(x) - h_a(x) = h_b(a)$ as required. □

We remark that we do not really need the differentiable structure on X in order to deduce Lemma 4.2.1. In fact, it follows from the CAT(0) inequality. The important observation is that if we have a "long" rectangle in a CAT(0)-space, then the sum of the two diagonals is approximately equal to the sum

of the two long edges. More specifically, suppose $x, y, z, w \in \bar{X}$, are any four points, then $|d(x,y) + d(z,w) - d(y,z) - d(x,w)| \leq \frac{1}{R}(d(x,z)^2 + d(y,w)^2)$, where $R = \min(d(x,y), d(z,w), d(y,z), d(x,w))$. Here xz and yw are the "short" sides. The exact form of the right-hand term of the inequality is unimportant. We just need to note that if the rectangle is sufficiently long, while the lengths of the short sides remain bounded, then the first term can be made arbitrarily small. We leave the reader to work out the details of this, and relate it to the definition of horofunctions.

Suppose that h is a horofunction about y. We have seen that $|\operatorname{grad} h| = 1$ everywhere, and so the level sets of h give us a codimension-1 foliation of X by C^2 submanifolds. Given $t \in \mathbb{R}$, write $S(t) = X \cap h^{-1}(t)$. We refer to $S(t)$ as a *horosphere* about y. Let $B(t) = X \setminus h^{-1}([t, \infty))$. Thus $B(t)$ is a closed convex subset of X with boundary $S(t)$. We call $B(t)$ a horoball about t.

Given a horoball B about y, we may define the *nearest point retraction* ρ of X onto B. Thus, for all $x \in X$, $\rho(x)$ is the nearest point on $[x,y] \cap B$ to x. We see that $\rho(x) = x$ for all $x \in B$, and $\rho(X \setminus B) = S = \partial B$. We have observed that S is a C^2-submanifold. We have

Lemma 4.2.2. *The nearest point retraction $\rho|(X \setminus B) : X \setminus B \longrightarrow S$ is C^2.*

Proof. Let h be the horofunction with $h(S) = \{0\}$. Apply Proposition 4.1.5, noting that $\rho(x) = \phi(x, -h(x))$ for all $x \in X \setminus B$. □

4.3. The compactified topology. Choose any basepoint $p \in X$, and let $T^1_p(X)$ be the unit tangent space at p. Now each vector in $T^1_p(X)$ determines the germ of a geodesic emanating from p. We may continue this geodesic until either we arrive at some point of X^0_I, or until we form a geodesic ray tending to some point of X^∞_I. Lemma 4.1.4 thus gives an identification of $X_I = X^0_I \cup X^\infty_I$ with $T^1_p(X)$. Thus, X_I is given the topology of an $(n-1)$-sphere. This topology turns out to be independent of the choice of basepoint $p \in X$. Moreover, it may be extended to give X_C the topology of a closed n-ball. In this, and the next two sections we give an account of this.

The identification $\bar{X} \equiv X \cup X^0_I \subseteq X_C$ gives us a metric d on $X \cup X^0_I$. We may extend this to a map $d : X_C \times X_C \longrightarrow [0, \infty]$ by setting $d(x,x) = 0$ and $d(x,y) = \infty$ when $x \in X^\infty_I$ and $y \in X_C \setminus \{x\}$. Given $x \in X \cup X^0_I$, and $r \geq 0$, we write

$$N(x,r) = \{y \in X_C \mid d(x,y) \leq r\}.$$

If $p \in X$, write

$$C(p,x,r) = \{y \in X_C \mid d(x,[p,y]) \leq r\}.$$

In other words, $y \in C(p,x,r)$, if and only if $[p,y]$ meets $N(x,r)$. Clearly $N(x,r) \subseteq C(p,x,r)$. The following is a simple consequence of the CAT(0) inequality.

Lemma 4.3.1. *Suppose that $p \in X$ and $y \in X_I^\infty$. Given $z \in [p,y] \cap X$, and $r, r' > 0$, then there is some $w \in [p,y] \cap X$ such that $C(p,w,r') \subseteq C(p,z,r)$.*

We may now define a topology, $\tau(X_C, p)$, on X_C, relative to the point $p \in X$. We describe neighbourhood bases for points $y \in X_C$ as follows. If $y \in X$, we take as neighbourhood base the collection $\{N(y,\epsilon) \mid \epsilon > 0\}$. If $y \in X_I^0$, we take as neighbourhood base $\{C(p,y,\epsilon) \mid \epsilon > 0\}$. If $y \in X_I^\infty$, we take as neighbourhood base $\{C(p,x,\epsilon) \mid x \in [p,y] \cap X, \epsilon > 0\}$. Note that, in the last case, by Lemma 4.3.1, we could equally well take as neighbourhood base $\{C(p,x,r) \mid x \in [p,y] \cap X\}$ for any fixed $r > 0$. It is easily verified that these sets form the basis for a topology $\tau(X_C, p)$ on X_C. Clearly, its restriction to X agrees with the metric topology. However, its restriction to $X \cup X_I^0 \equiv \bar{X}$ is, in general, coarser than the metric topology. We aim to show that $\tau(X_C, p)$ is independent of $p \in X$. The following lemma will be used in several places in the rest of this paper.

Lemma 4.3.2. *Given $a \in X_I^0$, and $h, \eta > 0$, we can find $r > 0$ with the following property. Suppose $(y,z) \in (X_C \times X_C) \setminus (X_I^\infty \times X_I^\infty)$ and $x \in N(a,r) \cap X$. If $d(a,[y,z]) \geq h$, then $y\hat{x}z \leq \eta$.*

Proof. By hypothesis (B), we can find $K, h_0 > 0$ such that if $d(x,a) \leq h_0$, then $\kappa(x) \leq -1/K^2 d(x,a)$. Suppose $h, \eta > 0$. Let $r > 0$, depending on h and η, be as determined below. We can assume that $r < h' = \min(h, h_0)$. Let $R = h' - r$.

Now let x, y, z be as in the statement of the lemma. For the moment, we assume that $y, z \in X \cup X_I^0$. The general case will follow by continuity. We want that $y\hat{x}z \leq \eta$.

Since $d(a,[y,z]) \geq h$, we have that $x \notin [y,z]$. Let $\theta = y\hat{x}z$. We can suppose that $\theta > 0$. Now, x, y, z are the vertices of a ruled surface obtained by joining x to each point $w \in [y,z]$ by a geodesic $[x,w]$. Thus $P = \bigcup\{[x,w] \mid w \in [y,z]\}$ is a non-positively curved 3-gon. In fact, if q lies in $\operatorname{int} P = P \setminus \partial P$, then the intrinsic curvature $\kappa_P(q)$ is at most $\kappa(q)$. By Gauss-Bonnet, we find that

$$-\int_P \kappa(q) d\omega(q) \leq -\int_P \kappa_P(q) d\omega(q) \leq \pi,$$

where $d\omega$ is the area element of P.

Suppose $t \in (0,R)$, and $w \in [y,z] \setminus \{y,z\}$. Let $q(w,t)$ be the point of $[x,w]$ with $d(x,q(w,t)) = t$. (Figure 4a.) Now $d(a,q(w,t)) \leq d(x,a) +$

$d(x, q(w,t)) \leq r + t$. Thus $-\kappa(q(w,t)) \geq 1/K^2 d(a, q(w,t))^2 \geq 1/K^2(r+t)^2$. By the Angle Comparison Theorem (Proposition 3.4.4), we see that the path traced out by $q(w,t)$ as w moves on $[y,z]$ has length at least θt. Thus,

$$\begin{aligned}\pi &\geq -\int_P \kappa(p) d\omega(q) \\ &\geq \int_0^R \frac{\theta t}{K^2(r+t)^2} dt \\ &= \frac{\theta}{K^2}(-\log(r/h') + (r/h') - 1) \\ &= \frac{\theta}{K^2} f(r/h')\end{aligned}$$

where $f(s) = s - \log s - 1$. Now $f(s) \to \infty$ as $s \to 0$, and so if r/h' is sufficiently small, we have $\theta \leq \pi K^2 / f(r/h') \leq \eta$ as required.

To deal with the case where $y \in X_I^\infty$ and $z \in X \cup X_I^0$, choose a sequence of points $y_n \in [y,z] \cap X$ with $y_n \to y$. As observed in Section 4.1, we have $\overrightarrow{xy_n} \to \overrightarrow{xy}$, and so the general case follows by continuity. □

Proposition 4.3.3. *The topology $\tau(X_C, p)$ is independent of $p \in X$.*

Proof. Suppose $p, p' \in X$. Certainly $\tau(X_C, p)$ and $\tau(X_C, p')$ agree on X. We thus want to show that for all $y \in X_I$, the neighbourhood bases with respect to p and p', as described above, are equivalent.

Suppose, first, that $y \in X_I^\infty$. Let $l = d(p,p')$ and suppose $r > 0$. Given $x \in [p,y] \cap X$, we want to find $x' \in [p',y] \cap X$ with $C(p',x',r) \subseteq C(p,x,r)$. By Lemma 4.3.1, we have $z \in [p,y]$ so that $C(p,z,r+2l) \subseteq C(p,x,r)$. By Lemma 4.1.1, we can find $x' \in [p',y]$ with $d(z,x') \leq l$. If $w \in C(p',x',r)$ so that $d(x',[p',w]) \leq r$, then the CAT(0) inequality, applied to the triangle wpp', tells us that $d(x',[p,w]) \leq d(x',[p',w]) + d(p,p') \leq r + l$. Thus $d(z,[p,w]) \leq (r+l) + d(z,x') \leq r + 2l$, and so $w \in C(p,z,r+2l)$. We have shown that $C(p',x',r') \subseteq C(p,x,r)$ as required.

Now suppose that $y \in X_I^0$. Given $\epsilon > 0$, we want to find $\epsilon' > 0$ so that $C(p',y,\epsilon') \subseteq C(p,y,\epsilon)$. We can assume that $\epsilon < d(y,p')$. Let $h_0 = d(y,[p,p'])$ and $h = \min(h_0, \epsilon)$. Lemma 4.3.2 gives us some $\epsilon' > 0$ such that if $x \in N(y,\epsilon') \cap X$ and $(a,b) \in (X_C \times X_C) \backslash (X_I^\infty \times X_I^\infty)$, then $d(y,[a,b]) \leq h$ or $a\hat{x}b \leq \pi/3$. Now suppose that $z \in C(p',y,\epsilon')$, so that there some $x \in [p',z] \cap N(y,\epsilon') \cap X$. Since $d(y,[p,p']) \geq h$, we have $p\hat{x}p' \leq \pi/3$. Thus $p\hat{x}z \geq \pi - \pi/3 = 2\pi/3$ and so $d(y,[p,z]) \leq h \leq \epsilon$. Thus $z \in C(p,y,\epsilon)$. We have shown that $C(p',y,\epsilon') \subseteq C(p,y,\epsilon)$. □

We shall write $\tau(X_C)$ for the topology thus defined on X_C. The following is easily verified.

Proposition 4.3.4. *The natural inclusion $\iota : \bar{X} \longrightarrow X_C$ is continuous.*

Here, and in the rest of this paper, we adopt the convention that $\bar{X}$ has the metric topology, whereas $X \cup X_I^0$ has the subspace topology induced from $\tau(X_C)$.

It is not very hard to see that $(X_C, \tau(X_C))$ is compact hausdorff. We shall not give a direct proof here, since we show, in the next two sections, that it is homeomorphic to a closed n-dimensional ball.

4.4. Starlike sets. Let $\mathbb{E}^n$ be n-dimensional euclidean space, and let $\underline{0} \in \mathbb{E}^n$ be any point. We identify the unit tangent space $T^1_{\underline{0}}\mathbb{E}^n$ with the unit sphere S^{n-1}. We may identify $\mathbb{E}^n$ with $(S^{n-1} \times [0,\infty))/\!\sim$, where $(\xi, 0) \sim (\zeta, 0)$ for all $\xi, \zeta \in S^{n-1}$, otherwise equivalence classes are single points. We may identify the compactified space $\mathbb{E}^n_C$ with $(S^{n-1} \times [0,\infty])/\!\sim$. We write $\langle \xi, t \rangle$ for the $\sim$-class of (ξ, t). Thus, $\underline{0} = \langle \xi, 0 \rangle$ for all $\xi \in S^{n-1}$.

Note that a subset $\Sigma \subseteq \mathbb{E}^n$ is open and starlike about $\underline{0}$ if and only if it has the form $\{\langle \xi, t \rangle \mid 0 \leq t < f(\xi)\}$, where $f : S^{n-1} \longrightarrow (0,\infty]$ is lower-semicontinuous. We write

$$\Sigma_C = \{\langle \xi, t \rangle \in \mathbb{E}^n_C \mid 0 \leq t \leq f(\xi)\}\,.$$

Thus, Σ_C is also starlike about $\underline{0}$, and a subset of the closure of Σ in $\mathbb{E}^n_C$. Write $\Sigma_I = \Sigma_C \setminus \Sigma = \{\langle \xi, t \rangle \in \mathbb{E}^n_C \mid t = f(\xi)\}$. (Note that this notation is consistent with that previously defined if $\Sigma = \mathbb{E}^n = X$.) We put a topology $\tau(\Sigma_C)$ on Σ_C as follows. We demand that the subspace topology on Σ induced from $\tau(\Sigma_C)$ agrees with that induced from $\mathbb{E}^n$. If $\langle \xi, t \rangle \in \Sigma_I$, we take as a base of neighbourhoods the collection $\{D(U,u)\}$ where

$$D(U,u) = \{\langle \zeta, v \rangle \in \Sigma_C \mid \zeta \in U, v \geq u\}$$

and U ranges over all neighbourhoods of ζ in ζ, and u ranges over the interval $(0,t)$. Thus, in general, the topology $\tau(\Sigma_C)$ on Σ_C is coarser than the subspace topology induced from $(\mathbb{E}^n_C, \tau(\mathbb{E}^n_C))$. (Note that $\tau(\mathbb{E}^n_C)$ agrees with our previous definition with $\Sigma = \mathbb{E}^n = X$.)

Now, if $a, b \in (0,\infty]$ and $h : [0,a] \longrightarrow [0,b]$ is a homeomorphism, with $h(0) = 0$, then the map $\hat{h} = [\langle \xi, t \rangle \mapsto \langle \xi, h(t) \rangle]$ gives a homeomorphism of the ball $N(\underline{0}, a)$ onto $N(\underline{0}, b)$ (where $N(\underline{0}, \infty) = \mathbb{E}^n_C$). Moreover, if $\Sigma \subseteq N(\underline{0}, a) \subseteq \mathbb{E}^n$ is open and starlike about $\underline{0}$, then so is $\Sigma' = \hat{h}(\Sigma)$. Also, $\Sigma'_C = \hat{h}(\Sigma_C)$, and $\hat{h}|\Sigma_C : (\Sigma_C, \tau(\Sigma_C)) \longrightarrow (\Sigma'_C, \tau(\Sigma'_C))$ is a homeomorphism.

Write B^n for the closed unit n-ball (as a manifold), and write $\operatorname{int} B^n = B^n \setminus \partial B^n$ for its interior.

Lemma 4.4.1. *Suppose $\Sigma \subseteq \mathbb{E}^n$ is open and starlike. Then, the pair*

(Σ_C, Σ), with the topology given by $\tau(\Sigma_C)$, is homeomorphic to the pair $(B^n, \operatorname{int} B^n)$.

Proof. Let $\mathbb{E}^n_\infty = \mathbb{E}^n \cup \{\underline{\infty}\}$ be the one point compactification of $\mathbb{E}^n$. Let $B_0 = N(\underline{0}, 1) \subseteq \mathbb{E}^n \subseteq \mathbb{E}^n_\infty$, be the unit ball about $\underline{0}$. From the discussion prior to the statement of the lemma, we see that we can assume that $\Sigma \subseteq B_0$.

Let $g : \mathbb{E}^n \longrightarrow \mathbb{E}^n_\infty \setminus \{\underline{0}\}$ be the inversion given by $g(\langle \xi, t\rangle) = \langle \xi, 1/t\rangle$ for $t > 0$ and $g(\underline{0}) = \underline{\infty}$. Restricted to B_0, the map g gives a homeomorphism of B_0 onto $B_\infty = \mathbb{E}^n_\infty \setminus \operatorname{int} B_0$. Let $\Omega = g(\Sigma)$, and $\Omega_C = g(\Sigma_C)$. Let $\partial\Omega$ be the topological boundary of Ω in $\mathbb{E}^n_\infty$, so that $\partial\Omega \subseteq B_\infty \setminus \{\underline{\infty}\}$ and $\Omega_C = \Omega \cup \partial\Omega$. We define the map $\rho : \Omega_C \setminus \{\underline{\infty}\} \longrightarrow [0, \infty)$ by $\rho(x) = d_{euc}(x, \partial\Omega)$, where d_{euc} is the euclidean distance.

Certainly, ρ is continuous on Ω, and $\rho(x) = 0$ if and only if $x \in \Omega_C \setminus \Omega$. Moreover, if $\langle \xi, t\rangle, \langle \xi, u\rangle \in \Omega$ with $t < u$, then $\rho(\langle \xi, t\rangle) < \rho(\langle \xi, u\rangle)$. We now define $h : \Omega_C \longrightarrow \mathbb{E}^n_C$ by $h(\langle \xi, t\rangle) = \langle \xi, 1 + \rho(\langle \xi, t\rangle)\rangle$ and $h(\underline{\infty}) = \underline{\infty}$. Clearly, h maps Ω_C bijectively onto B_∞, and $h|\Omega$ is a homeomorphism onto $\operatorname{int} B_\infty$. It follows that $j = g^{-1}hg$ maps Σ_C bijectively onto B_0, and that $j|\Sigma$ is a homeomorphism onto $\operatorname{int} B_0$. Moreover, a simple exercise shows that j is, in fact, a homeomorphism from $(\Sigma_C, \tau(\Sigma_C))$ to B_0. □

With a bit more work, one can make a stronger statement, namely:

Lemma 4.4.2. *Suppose that $\Sigma \subseteq \mathbb{E}^n$ is open and starlike. Then, there is a homeomorphism of (Σ_C, Σ) to $(B^n, \operatorname{int} B^n)$ whose restriction to Σ is a smooth diffeomorphism onto $\operatorname{int} B^n$.*

Proof (Sketch). One way to do this is to approximate the map ρ, from the proof of Lemma 4.4.1, by a smooth map, ρ', with $\partial\rho'/\partial t > 0$ everywhere on $\Omega \setminus \{\underline{\infty}\}$. Define $\sigma : B_\infty \setminus \{\underline{\infty}\} \longrightarrow (0, \infty)$ by $\rho(\langle \xi, \sigma(\langle \xi, t\rangle)\rangle) = t$. We want to smooth out σ on $\operatorname{int} B_0 \setminus \{\underline{\infty}\}$ to get a smooth map σ' with $\partial\sigma'/\partial t > 0$. Given any positive integer n, define $\sigma_n : S^{n-1} \longrightarrow (0, \infty)$ by $\sigma_n(\xi) = \sigma(\langle \xi, 1 + 1/n\rangle)$. We approximate each σ_n by a smooth map $\sigma'_n : S^{n-1} \longrightarrow (0, \infty)$ so that $|\sigma'_n(\xi) - \sigma_n(\xi)| \leq 1/2n(n+1)$ for all $\xi \in S^{n-1}$. In this way, we arrange that $\sigma'_{n+1}(\xi) < \sigma_n(\xi)$ for all $\xi \in S^{n-1}$. By interpolation, we get a smooth function $\sigma' : B(\underline{0}, 2) \setminus B_0 \longrightarrow (0, \infty)$ so that $\sigma'(\langle \xi, 1 + 1/n\rangle) = \sigma'_n(\xi)$ and $\partial\sigma'/\partial t > 0$. We now extend to a smooth function $\sigma' : \operatorname{int} B_\infty \setminus \{\underline{\infty}\} \longrightarrow (0, \infty)$ so that $\partial\sigma'/\partial t > 0$ everywhere, and $\sigma'(\langle \xi, t\rangle) = t$ for all sufficiently large t. The identity $\rho'(\langle \xi, \sigma'(\langle \xi, t\rangle)\rangle) = t$ allows us to define a smooth map $\rho' : \Omega \setminus \{\underline{\infty}\} \longrightarrow (0, \infty)$, with $\partial\rho'/\partial t > 0$. We extend ρ' to a map $\Omega_C \setminus \{\underline{\infty}\} \longrightarrow [0, \infty)$ by setting $\rho'(\Omega_C \setminus \Omega) = \{0\}$. We now proceed as in Lemma 4.4.1. It may be verified that the map $j' : \Sigma_C \longrightarrow B_0$ thus defined is a diffeomorphism on Σ. □

4.5. The logarithm map. In this section, we relate the discussion of starlike sets to our compactified manifold X_C.

Choose any point $p \in X$ and then identify the unit tangent space $T^1_p(X)$ with S^{n-1} via an isometry $\phi : T^1_p(X) \longrightarrow S^{n-1}$. Recall the description of $\mathbb{E}^n_C$ as a quotient of $S^{n-1} \times [0, \infty]$, given in the previous section. We define a map $\log : X_C \longrightarrow \mathbb{E}^n_C$ as follows. Set $\log(p) = \underline{0}$, and for $x \in X_C \setminus \{p\}$, set $\log(x) = \langle \phi(\overrightarrow{px}), d(p,x) \rangle$, were $d(p,x) = \infty$ for $x \in X^\infty_I$. By Lemma 4.1.4, we see that log is a bijection onto its image $\Sigma_C(X) = \log(X_C) \subseteq \mathbb{E}^n_C$. Moreover $\log | X$ gives a diffeomorphism of X onto $\Sigma(X) = \log(X) \subseteq X$. This follows as in the complete case. Thus, Σ is open and starlike about $\underline{0}$. Also, we have that, set theoretically, $(\Sigma(X))_C = \Sigma_C(X)$.

Lemma 4.5.1. *The map* $\log : (X_C, \tau(X_C)) \longrightarrow (\Sigma_C(X), \tau(\Sigma_C(X)))$ *is a homeomorphism.*

Proof. The fact that log is continuous is a simple consequence of the Angle Comparison Theorem (Proposition 3.4.4). We have also noted that $\log | X$ is a diffeomorphism. It remains therefore to show that $\exp = \log^{-1} : \Sigma_C(X) \longrightarrow X_C$ is continuous at all points of $\Sigma_C(X) \setminus \Sigma(X)$.

Suppose that $y = \exp(\langle \xi, t \rangle) \in X^0_I$. Given $r \in (0, t)$, let $x = \exp(\langle \xi, t - r/2 \rangle)$. Thus, $x \in [p, y]$ with $d(x, y) = r/2$. By the continuity of $\exp | X$, we can find $U \subseteq S^{n-1}$ which is a neighbourhood of ξ, such that if $\xi' \in U$, then $\langle \xi, t - r/2 \rangle \in \Sigma$ and $d(x, \exp(\langle \xi, t - r/2 \rangle)) \leq r/2$. It follows that $d(y, \exp(\langle \xi', t - r/2 \rangle)) \leq r/2$, and so $\exp(\langle \xi', t' \rangle) \in C(p, y, r)$ whenever $t' \geq t - r/2$ and $\langle \xi', t' \rangle \in \Sigma_C(X)$. This shows that $\exp(D(U, t - r/2)) \subseteq C(p, y, r)$, and so exp is continuous at $\langle \xi, t \rangle$.

The case where $\exp(\langle \xi, t \rangle) \in X^\infty_I$ is similar. □

Putting Lemma 4.5.1 together with Lemma 4.4.2, we have:

Proposition 4.5.2. *The pair* (X_C, X), *in the topology* $\tau(X_C)$, *is homeomorphic to the pair* $(B^n, \operatorname{int} B^n)$ *where* B^n *is the unit* n*-dimensional ball, and* $\operatorname{int} B^n$ *is its interior. Moreover, we can arrange that the homeomorphism restricted to* X *gives a smooth diffeomorphism onto* $\operatorname{int} B^n$.

In particular, we see that X_C is compact metrisable.

5. Continuity properties.

As in the previous chapter, we are assuming that X satisfies axioms (A) and (B). Our aim here is to investigate how geodesics move as we vary the endpoints.

5.1. Lower semicontinuity of the distance function. We extend the metric d on $\bar{X} \equiv X \cup X_I^0$ to a map $d : X_C \times X_C \longrightarrow [0, \infty]$ by setting $d(x,x) = 0$ and $d(x,y) = \infty$ if $x \in X_I^\infty$ and $y \in X_C \backslash \{x\}$. We claim that this map is lower-semicontinuous on $X_C \times X_C$ given the product topology $\tau(X_C) \times \tau(X_C)$.

Lemma 5.1.1. *Suppose that $x, y \in X \cup X_I^0$ and $x \neq y$. Given and $h > 0$, there exist neighbourhoods U of x and V of y in $\tau(X_C)$ such that if $u \in U$ and $v \in V$, and $(u,v) \notin X_I^\infty \times X_I^\infty$, then $d(x,[u,v]) \leq h$ and $d(y,[u,v]) \leq h$. In fact, we can find $u', v' \in [u,v]$ with $d(x,u') \leq h$, $d(y,v') \leq h$ and $u' \in [u,v']$.*

Proof. We shall deal with the case where x and y both lie in X_I^0. The remaining cases are simpler. We can assume that $h < \frac{1}{4}d(x,y)$. By Lemma 4.3.2, there is some $\epsilon_1 > 0$ such that if $a \in N(x,\epsilon_1) \cap X$ and $(z,w) \in (X_C \times X_C) \backslash (X_I^\infty \times X_I^\infty)$, then either $d(x,[z,w]) \leq h$, or else $z\hat{a}w \leq \pi/2$. There is a similar constant ϵ_2 corresponding to b. Let $\epsilon = \min(\epsilon_1, \epsilon_2, h)$. Let $U = C(y,x,\epsilon)$ and $V = C(x,y,\epsilon)$. From the definition $\tau(X_C) = \tau(X_C, x) = \tau(X_C, y)$ we see that U, V are neighbourhoods of x, y respectively in $\tau(X_C)$. Suppose that $u \in U$ and $v \in V$, so that $d(x,[u,y]) \leq \epsilon$ and $d(y,[v,x]) \leq \epsilon$. Choose $a \in [u,y] \cap N(x,\epsilon) \cap X$ and $b \in [v,x] \cap N(x,\epsilon) \cap X$. By the Angle Comparison Theorem (Proposition 3.4.4), we see that $a\hat{b}x \leq \pi/3$, and so $a\hat{b}v \geq 2\pi/3 > \pi/2$. Thus $d(y,[a,v]) \leq h$. So, again by the Angle Comparison Theorem, we have $v\hat{a}y \leq \pi/3$ and so $u\hat{a}v \geq 2\pi/3 > \pi/2$. Thus $d(x,[u,v]) \leq h$. Similarly, $d(y,[u,v]) \leq h$.

Note that $d(x,y) \geq d(u,x) + d(x,y) - 2\epsilon$. Thus if $u', v' \in [u,v]$ with $d(x,u') \leq h$ and $d(y,v') \leq h$ then $d(u,u') \geq d(u,v') + d(x,y) - 2\epsilon - 2h \geq d(u,v')$ since $\epsilon \leq h$ and $4h \leq d(x,y)$. If $u \notin X_I^\infty$, this shows that $u' \in [u,v']$. If $u \in X_I^\infty$, choose $u_0 \in [u,u'] \cap [u,v'] \cap X$, and apply the same argument with u_0 replacing u. □

Lemma 5.1.2. *Suppose $x \in X_I^0$, $y \in X_I^\infty$ and $z \in [x,y] \cap X$. Given $h > 0$, then there is some $\epsilon > 0$ and a neighbourhood V about y in $\tau(X_C)$ such that if $u \in N(x,\epsilon)$ and $v \in V$, then $d(z,[u,v]) \leq h$.*

Proof. Take $\epsilon = h/2$. By the definition of $\tau(X_C)$, the set $V = C(x,z,\epsilon)$ is a neighbourhood of y. Suppose $v \in V$ and $u \in N(x,\epsilon)$. Then $d(z,[x,v]) \leq \epsilon$, and so by CAT(0) applied to xuv, we find that $d(z,[u,v]) \leq 2\epsilon \leq h$. □

Lemma 5.1.3. *Suppose $x \in X \cup X_I^0$ and $y \in X_I^\infty$ and $z \in [x,y] \cap X$. Given any $h > 0$, there are neighbourhoods U of x and V of y in $\tau(X_C)$ such that if $u \in U$ and $v \in V$ and $(u,v) \notin X_I^\infty \times X_I^\infty$, then $d(x,[u,v]) \leq h$ and $d(z,[u,v]) \leq h$. Moreover, we can find $u', v' \in [u,v]$ with $d(x,u') \leq h$ and*

$d(z, v') \leq h$ *and* $u' \in [u, v']$.

Proof. As with Lemma 5.1.1. Use Lemma 5.1.2. □

Proposition 5.1.4. *The map* $d : X_C \times X_C \longrightarrow [0, \infty]$ *is lower semicontinuous, where* $X_C \times X_C$ *is given the product topology* $\tau(X_C) \times \tau(X_C)$.

Proof. Suppose $x, y \in X_C$. If $x = y$, then $d(x, y) = 0$ and there is nothing to prove. If $x \in X \cup X_I^0$ and $y \in X_C \backslash \{x\}$, the result follows from Lemmas 5.1.1 and 5.1.3. The only remaining case is where $x, y \in X_I^\infty$ and $x \neq y$, so that $d(x, y) = \infty$. Choose any $p \in X$. Let $\theta = x\hat{p}y > 0$. Given any $r > 0$, let $R = r \operatorname{cosec}(\theta/4)$. Since $\tau(X_C) = \tau(X_C, p)$, by applying the Angle Comparison Theorem (Proposition 3.4.4), we can find neighbourhoods U of x and V of y such that if $u \in U$ and $v \in V$, then $d(p, u) \geq R$, $d(p, v) \geq R$, $x\hat{p}v \leq \theta/4$ and $y\hat{p}v \leq \theta/4$. Thus $u\hat{p}v \geq \theta/2$ and so, again by angle comparison, $d(u, v) \geq r$. □

5.2. The Hausdorff topology. We have seen that X_C is homeomorphic to a ball and hence metrisable. A metric on X_C induces a Hausdorff distance on the set, $\mathscr{C}(X_C)$, of all closed subsets of X_C and hence a topology on $\mathscr{C}(X_C)$. Since X_C is compact, it's not hard to see that the topology on on $\mathscr{C}(X_C)$ is independent of the choice of metric on X_C. We call this topology the *Hausdorff topology* on $\mathscr{C}(X_C)$.

A more natural description of the Hausdorff topology is in terms of uniformities (see [**K**]). Here we shall deal only with bases of uniformities. Given a set Y, write $\Delta = \Delta(Y) \subseteq Y \times Y$ for the diagonal $\{(x, x) \mid x \in Y\}$. Given a subset $W \subseteq Y \times Y$, write $W^2 = \{(x, y) \in Y \times Y \mid (\exists z \in Y)((x, z) \in W, (z, y) \in W)\}$. We say that a subset $W \subseteq Y \times Y$ is *symmetric* if $(x, y) \in W$ whenever $(y, x) \in W$. A collection $\mathscr{W}$ of symmetric subsets of $Y \times Y$ form a *uniform basis* for Y if the following hold:

(1) $\Delta \subseteq W$ for all $W \in \mathscr{W}$.

(2) For all $W_1, W_2 \in \mathscr{W}$, there is some $W_3 \in \mathscr{W}$ with $W_3 \subseteq W_1 \cap W_2$.

(3) For all $W \in \mathscr{W}$, there is some $V \in \mathscr{W}$ with $V^2 \subseteq W$.

Two such bases $\mathscr{W}_1$ and $\mathscr{W}_2$ are *equivalent* if for all $W_1 \in \mathscr{W}_1$ there is some $W_2 \in \mathscr{W}_2$ with $W_2 \subseteq W_1$, and for all $W_2' \in \mathscr{W}_2$ there is some $W_1' \in \mathscr{W}_1$ with $W_1' \subseteq W_2'$. Thus, two bases give rise to the same uniformity if and only if they are equivalent. (For our purposes, we can define a uniformity as an equivalence class of bases.)

Given a subset $W \subseteq Y \times Y$ and a subset $A \subseteq Y$, write $WA = \{x \in Y \mid (\exists y \in A)((x, y) \in W)\}$. Thus if $\Delta \subseteq W$, then $A \subseteq WA$.

A uniform basis $\mathscr{W}$ on Y induces a topology on Y, where a neighbourhood of the point $x \in Y$ is given by $\mathscr{W}\{x\} = \{W\{x\} \mid W \in \mathscr{W}\}$. This topology depends only on the uniformity. It is hausdorff if and only if $\bigcap \mathscr{W} = \Delta$.

Note that a metric d on Y induces a uniformity with basis $\{\{(x,y) \in Y \times Y \mid d(x,y) \leq \epsilon\} \mid \epsilon > 0\}$. This uniformity, in turn, induces the metric topology. If Y is compact, then this is the unique uniformity of Y inducing the metric topology.

Suppose that $\mathscr{W}$ is a uniform basis on Y. Write $\mathscr{C}(Y)$ for the set of subsets that are closed in the induced topology. Given $W \in \mathscr{W}$, write $P(W) = \{(A,B) \in \mathscr{C}(Y) \times \mathscr{C}(Y) \mid A \subseteq WB, B \subseteq WA\}$, and set $P(\mathscr{W}) = \{P(W) \mid W \in \mathscr{W}\}$. One checks that $P(\mathscr{W})$ is a uniform basis on $\mathscr{C}(Y)$. If $(Y, \mathscr{W})$ is hausdorff (respectively metrisable) then $(\mathscr{C}(Y), P(\mathscr{W}))$ is hausdorff (metrisable). We refer to the topology induced on $\mathscr{C}(Y)$ by $P(\mathscr{W})$ as the *Hausdorff topology*.

Since X_C is compact metrisable, it admits a unique uniformity, and so $\mathscr{C}(X_C)$ has a well-defined Hausdorff topology. In the next section shall show that geodesics vary continuously in this topology. We spend the rest of this section giving an explicit description of the uniformity on X_C.

Fix $p \in X$, and suppose that $A \subseteq X \cup X_I^0$. Given $\epsilon > 0$, define $\Omega(p, A, \epsilon) \subseteq X_C \times X_C$ as follows. The pair (x,y) lies in $\Omega(p,A,\epsilon)$ if either there is some $a \in A$ with $d(a,[p,x]) \leq \epsilon$ and $d(a,[p,y]) \leq \epsilon$, or else if $x, y \in X \cup X_I^0$ and $d(x,y) \leq 2\epsilon$.

Clearly $\Omega(p,A,\epsilon)$ is symmetric and if $B \subseteq A$ and $\delta \leq \epsilon$, then $\Omega(p,B,\delta) \subseteq \Omega(p,A,\epsilon)$.

Lemma 5.2.1. *For all $A \subseteq X \cup X_I^0$ and $\epsilon > 0$, we have $\Omega(p,A,\epsilon)^2 \subseteq \Omega(p,A,3\epsilon)$.*

Proof. Suppose $(x,y),(y,z) \in \Omega(p,A,\epsilon)$. There are three cases.

(1) There are points $a, b \in A$, $a_0 \in [p,x]$, $a_1, b_1 \in [p,y]$ and $b_0 \in [p,z]$ with $d(a,a_i) \leq \epsilon$ and $d(b,b_i) \leq \epsilon$ for $i = 0,1$. Without loss of generality, we have $d(p,b_1) \geq d(p,a_1)$. (Figure 5a).

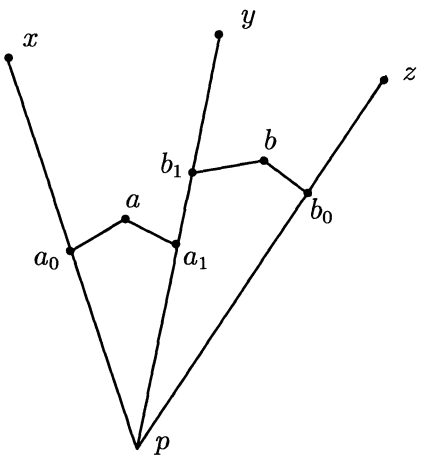


Figure 5a.

Applying CAT(0) to pb_0b_1, we have $d(a_1, [p,z]) \leq 3\epsilon$. Thus $(x,z) \in \Omega(p, A, 3\epsilon)$.

(2) $d(y,z) \leq 2\epsilon$ and there is some $a \in A$ with $d(a,[p,x]) \leq \epsilon$ and $d(a,[p,y]) \leq \epsilon$. Applying CAT(0) to pyz, we find that $d(a,[p,z]) \leq 3\epsilon$, and so $(x,z) \in \Omega(p,A,\epsilon)$.

(3) If $d(x,y) \leq 2\epsilon$ and $d(y,z) \leq 2\epsilon$, then $d(x,z) \leq 4\epsilon$ and so $(x,z) \in \Omega(p,A,2\epsilon)$.

□

Given $r > 0$, write $A(p,r) = X_I^0 \cup (X \setminus \operatorname{int} N(p,r))$, and set $W(p,r,\epsilon) = \Omega(p, A(p,r), \epsilon)$. Clearly $\Delta \subseteq W(p,r,\epsilon)$ for all $r > 0$ and $\epsilon > 0$. Let $\mathscr{W} = \mathscr{W}_p = \{W(p,r,\epsilon) \mid r > 0, \epsilon > 0\}$. Applying Lemma 5.2.1, we see that $\mathscr{W}$ is a uniform base on X_C.

Lemma 5.2.2. *The uniform base $\mathscr{W}_p$ induces the topology $\tau(X_C)$ on X_C.*

Proof. We need to check that if $x \in X_C$, then $\mathscr{W}\{x\}$ gives a neighbourhood base for x in $\tau(X_C) = \tau(X_C, p)$.

Case (1): $x \in X$.

If $\epsilon < d(X_I^0, [x,p])$ and $r > d(x,p) + \epsilon$, then $W(p,r,\epsilon)\{x\} = N(x,\epsilon)$.

Case (2): $x \in X_I^0$.

Clearly $C(p,x,\epsilon) \subseteq W(p,r,\epsilon)\{x\}$ for all $r > 0$ and $\epsilon > 0$. Now, $[p,x] \cap X_I^0 = \{x\}$. Given any $\epsilon \in (0, d(p,x))$, let $y \in [p,x]$ be the point with $d(x,y) = \epsilon/3$. Let $\delta = \delta(\epsilon) = \frac{1}{2} d(X_I^0, [p,y]) > 0$, so $\delta \leq \epsilon/6$. Now, suppose $r > d(p,x) + \delta$. If $z \in W(p,r,\delta)\{x\}$, then either $d(z,x) \leq 2\delta \leq \epsilon$, and so $z \in C(p,x,\epsilon)$, or else there is some $a \in A(p,r)$ with $d(a,[p,x]) \leq \delta$ and $d(a,[p,z]) \leq \delta \leq \epsilon/3$. Since $r > d(p,x) + \delta$, we must have $a \in X_I^0$, and so $d(x,a) \leq \delta + \epsilon/3 \leq 2\delta/3$. It follows that $d(x,[p,z]) \leq 2\epsilon/3 + \epsilon/3 = \epsilon$, and again we have $z \in C(p,x,\epsilon)$. We have shown that $W(p,r,\delta)\{x\} \subseteq C(p,x,\epsilon)$.

Case (3): $x \in X_I^\infty$.

Given $r > 0$, take $y \in [p,x]$ with $d(p,y) = r$. Then $C(p,r,\epsilon)\{x\} \subseteq W(p,r,\epsilon)\{x\}$ for all $\epsilon > 0$.

Conversely, suppose $y \in [p,x]$. Let $r = d(p,y)$, and let $\delta = \delta(r) = \frac{1}{2} d(X_I^0, [p,y]) > 0$. Suppose $\epsilon \in (0,\delta)$, and $z \in W(p,r,\epsilon)\{x\}$. Then, there is some $a \in A(p,r)$ with $d(a,[p,y]) \leq \epsilon \leq \delta$ and $d(a,[p,z]) \leq \epsilon$. If $a \in X_I^0$, then $d(a,[p,y]) > \delta$ and so $d(a,[x,y]) \leq \epsilon \leq 2\epsilon$. If $a \notin X_I^0$, then $d(p,a) \geq r$, and so again, $d(a,[x,y]) \leq 2\epsilon$. Applying CAT(0), we find that $d(y,[p,z]) \leq 3\epsilon$ and so $z \in C(p,y,3\epsilon)$. Thus $W(p,r,\epsilon)\{x\} \subseteq C(p,y,3\epsilon)$. □

It follows that the uniform base $\mathscr{W}_p$ defines the unique uniformity on X_C inducing the topology $\tau(X_C)$. In particular, $\mathscr{W}_p$ and $\mathscr{W}_q$ are equivalent for all $p, q \in X$.

5.3. Continuity of geodesics. By Lemma 4.1.4, any pair of points $(x, y) \in (X_C \times X_C) \backslash (X_I^\infty \times X_I^\infty)$ may be joined by a unique geodesic $[x, y]$.

Lemma 5.3.1. *Each geodesic $[x, y]$ is closed X_C.*

Proof. We can assume $x \neq y$. Choose $p \in [x, y] \backslash \{x, y\}$. If $z_n \in [x, y]$ is any sequence, it is easily seen that some subsequence converges in $\tau(X_C, p)$ to a limit in $[x, y]$. □

We give $\mathscr{C}(X_C)$ the Hausdorff topology as described in Section 5.2. We give $X_C \times X_C$ the product topology $\tau(X_C) \times \tau(X_C)$.

Proposition 5.3.2. *The map $[(x, y) \mapsto [x, y]] : (X_C \times X_C) \backslash (X_I^\infty \times X_I^\infty) \longrightarrow \mathscr{C}(X_C)$ is continuous.*

Proof. We distinguish six cases.

Case (1): $x, y \in X$.

This follows from Proposition 3.4.2.

Case (2): $x, y \in X_I^0$ and $x \neq y$.

Fix some $p \in [x, y] \cap X$. Suppose $r > 0$ and $\epsilon > 0$. Let U, V be the neighbourhoods of x, y respectively, given by Lemma 5.1.1, so that if $u \in U$ and $v \in V$, then we can find $u', v' \in [u, v]$ with $d(x, u') \leq \epsilon/2$, $d(y, v') \leq \epsilon/2$ and $u' \in [u, v']$. From the convexity of the distance function (Proposition 3.4.6), we have that $[u', v'] \subseteq N([x, y], \epsilon/2) \subseteq W(p, r, \epsilon)[x, y]$ and $[x, y] \subseteq N([u', v'], \epsilon/2) \subseteq W(p, r, \epsilon)[u, v]$. (Figure 5b.) Suppose $z \in [u, u']$. Again, by convexity, we have $d(u', [p, z]) \leq \epsilon/2$, and so $d(x, [p, z]) \leq \epsilon$. Thus $z \in C(p, x, \epsilon) \subseteq W(p, r, \epsilon)\{x\}$. Therefore, $[u, u'] \subseteq W(p, r, \epsilon)\{x\}$. Similarly, $[v, v'] \subseteq W(p, r, \epsilon)\{y\}$. We have shown that

$$[u, v] \subseteq W(p, r, \epsilon)[x, y]$$

and

$$[x, y] \subseteq W(p, r, \epsilon)[u, v].$$

In other words, $[u, v] \in P(W(p, r, \epsilon))\{[x, y]\}$. Now, the sets

$$P(W(p, r, \epsilon))\{[x, y]\}$$

as $\epsilon \to 0$ and $r \to \infty$ form a neighbourhood base for $[x, y]$ in the Hausdorff topology on $\mathscr{C}(X_C)$. This deals with Case (2).

Case (3): $x = y \in X_I^0$.

Choose any $p \in X$, and suppose $\epsilon > 0$ and $r > 0$. By Lemma 4.3.2, there is some $\delta_0 > 0$ such that if $a, z \in X_C$ with $d(x, a) \leq \delta_0$ and $z\hat{a}p \geq \pi/3$,

then $d(x,[p,z]) \leq \epsilon$. Let $\delta = \min(\delta_0, \epsilon/3)$. Suppose $u, v \in C(p,x,\delta)$, and $(u,v) \notin X_I^\infty \times X_I^\infty$. We claim that $[u,v] \subseteq C(p,x,\epsilon)$.

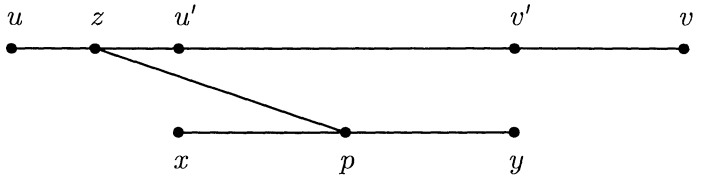


Figure 5b.

To see this, choose $a \in [p,u]$ and $b \in [p,v]$ with $d(x,a) \leq \delta$ and $d(x,b) \leq \delta$, and suppose $z \in [u,v]$. (Figure 5c.) If $d(a,z) \leq 2\delta$, then $d(x,z) \leq \delta + 2\delta \leq \epsilon$, and so $z \in N(x,\epsilon) \subseteq C(p,x,\epsilon)$. Similarly if $d(b,z) \leq 2\delta$. Thus, we can suppose that $d(a,z) \geq 2\delta$ and $d(b,z) \geq 2\delta$, and so, by the Angle Comparison Theorem, we have that $a\hat{z}b \leq \pi/3$. Thus, without loss of generality, we can suppose that $u\hat{z}a \geq \frac{1}{2}(\pi - \pi/3) = \pi/3$. Thus, again by angle comparison, $u\hat{a}z \leq \pi - \pi/3 = 2\pi/3$, and so $z\hat{a}p \geq \pi/3$. It follows that $d(a,[p,z]) \leq \epsilon$, and so $z \in C(p,x,\epsilon)$. This proves the claim that $[u,v] \subseteq C(p,x,\epsilon)$.

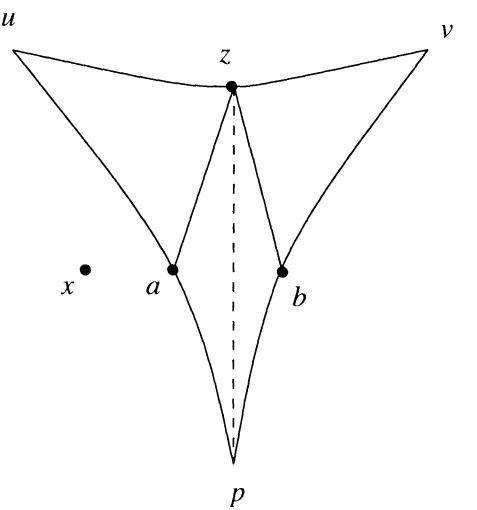


Figure 5c.

Now, for all $r > 0$, we have $C(p,x,\epsilon) \subseteq W(p,r,\epsilon)\{x\}$. Since $W(p,r,\epsilon)$ is symmetric, we have $x \in W(p,r,\epsilon)[u,v]$, and so $[u,v] \in P(W(p,r,\epsilon))\{\{x\}\}$. As $\epsilon \to 0$ and $r \to \infty$, the sets $P(W(p,r,\epsilon))\{\{x\}\}$ form a neighbourhood base for $\{x\} = [x,x]$ in the Hausdorff topology on $\mathscr{C}(X_C)$.

Case (4): $x \in X$ and $y \in X_I^0$.

This is similar to Case (2).

Case (5): $x \in X_I^0$ and $y \in X_I^\infty$.

Fix $p \in [x,y]\backslash\{x,y\}$, and suppose $\epsilon > 0$ and $r > 0$. Choose $z \in [p,y]$ with $d(p,z) \geq r$. By Lemma 5.1.3, we can find neighbourhoods U, V about x, y respectively, such that if $u \in U$, $v \in V$ and $(u,v) \notin X_I^\infty \times X_I^\infty$, then there exist $u', v' \in [u,v]$ with $d(x,u') \leq \epsilon/2$, $d(y,v') \leq \epsilon/2$ and $u' \in [u,v']$. Arguing as in Case (2), we see that $[u',v'] \subseteq N([x,y],\epsilon/2)$, $[x,z] \subseteq N([u,v],\epsilon/2)$, $[u,u'] \subseteq C(p,x,\epsilon)$, $[v,v'] \subseteq C(p,z,\epsilon)$ and $[z,y] \subseteq C(p,v',\epsilon/2)$. Now $x, z \in A(p,r)$ and so $[u,v] \in P(W(p,r,\epsilon))\{[x,y]\}$.

Case (6): $x \in X$ and $y \in X_I^\infty$.

This is similar to case (5). □

6. Visibility.

In this Chapter, we assume that X satisfies properties (A), (B) and (C), where (C) is the statement:

(C) There exist $p_0 \in X$, and $L_0, R_0 > 0$ such that if $x \in X$ with $d(p_0,x) \geq R_0$, then $\kappa(x) \leq -1/L_0^2 d(p_0,x)^2$.

It follows immediately that if we fix any $L \in (0, L_0)$, then for all $p \in X$, there is some $R = R(p)$ such that if $d(p,x) \geq R$, then $\kappa(x) \leq -1/L^2 d(p,x)$. We aim to show that, with these hypotheses, X is a visibility manifold, and that geodesics vary continuously on $X_C \times X_C$.

6.1. Convergence of asymptotic geodesics. Suppose $y \in X_I^\infty$, and $h : X \cup X_I^0 \longrightarrow \mathbb{R}$ is a horofunction about y. (Section 4.2.) Suppose $b_0, b_1 \in X \cup X_I^0$ with $h(b_0) = h(b_1)$. Let $\beta_i : [0,\infty) \longrightarrow X \cup X_I^0$ be the geodesic ray $[b_i, y]$. Thus $h(\beta_0(t)) = h(\beta_1(t)) = h(b_0) + t$ for all $t \in [0,\infty)$.

Lemma 6.1.1. *$d(\beta_0(t), \beta_1(t)) \to 0$ as $t \to \infty$.*

In fact, we show that $d(\beta_0(t), \beta_1(t)) \leq A(t+\lambda)^{-\mu}$ where $\mu > 0$ is fixed, and $A, \lambda \geq 0$ depend on b_0 and b_1.

Proof. We can assume that $b_0, b_1 \in X$. Join b_0 to b_1 by a smooth path $\gamma : [0,1] \longrightarrow X$. Let $t_0 = \max\{h(\gamma(u)) \mid u \in [0,1]\}$. Let B be the horoball $X \cap h^{-1}([t_0,\infty))$, and let S be the bounding horosphere $X \cap h^{-1}(t_0)$. Let $\rho : X \backslash \operatorname{int} B \longrightarrow S$ be the nearest-point retraction. Now, the path $h \circ \gamma : [0,1] \longrightarrow S$ joins $\beta(t_0)$ to $\beta(t_1)$, and, by Lemma 4.2.2, is C^2. Thus, without loss of generality, we can assume that $b_0, b_1 \in S = h^{-1}(0)$, and that b_0 and b_1 can be joined by a C^2 path $\gamma : [0,1] \longrightarrow S$.

Now, for each $u \in [0,1]$, let $\beta_u : [0,\infty) \longrightarrow X$ be the geodesic ray based at $\gamma(u)$ tending to y. Define $\beta : [0,\infty) \times [0,1] \longrightarrow X$ by $\beta(t,u) = \beta_u(t)$. By Lemma 4.1.5, β is C^2. Note that $h(\beta(t,u)) = t$ for all $(t,u) \in [0,\infty) \times [0,1]$.

Also $\frac{\partial\beta}{\partial u}(t,u) = \operatorname{grad}\beta(h(t,u))$. Thus $\left\langle \frac{\partial\beta}{\partial t}(t,u), \frac{\partial\beta}{\partial u}(t,u)\right\rangle = 0$ for all (t,u). In other words β is a normalised ruled map in the sense of Section 3.1 (except that it is only C^2 and not smooth, though this is more than enough). For a fixed u, the map $\left[t \mapsto \frac{\partial\beta}{\partial u}(t,u)\right]$ is a Jacobi field along β_u. Thus the map $[t \mapsto J(t,u)]$ is convex, where $J(t,u) = \left|\frac{\partial\beta}{\partial u}(t,u)\right|$. Given $t \in [0,\infty)$ write $\beta^t : [0,1] \longrightarrow X$ for the C^2 transversal path $[u \mapsto \beta(t,u)]$. Thus $\operatorname{length}\beta^t = \int_0^1 J(t,u)du$. Now, for all $u_1, u_2 \in [0,1]$ the function $d(\beta(t,u_1),\beta(t,u_2))$ is monotonically non-increasing in t. Thus, for any fixed subinterval $I \subseteq [0,1]$, the rectifiable lengths of the paths $\beta^t|I$ are non-increasing in t. Now, $\operatorname{length}(\beta^t|I) = \int_I J(t,u)du$. We deduce that for all $u \in [0,1]$ the map $[t \mapsto J(t,u)]$ is non-increasing.

Now choose $p \in X$, and let $R = R(p)$. Thus, if $d(p,x) \geq R$, then we have $\kappa(x) \leq -1/L^2 d(p,x)^2$. Let $\lambda = \max\{d(p,\gamma(u)) \mid u \in [0,1]\}$. Thus $t - \lambda \leq d(p,\beta(t,u)) \leq t + \lambda$. Without loss of generality, we can assume that $d(p,\beta(t,u)) \geq R$ for all (t,u), and so $\kappa(\beta(t,u)) \leq -1/L^2(t+\lambda)^2$.

From the formula in Section 4.2, we find that $J(t,u) \leq J(0,u)\,(1+\frac{t}{\lambda})^{-\mu}$ where $\mu = (\sqrt{1+4L^2}) - 1 > 0$. Thus

$$\begin{aligned} d(\beta_0(t),\beta_1(t)) \leq \operatorname{length}\beta^t &= \int_0^1 J(t,u)du \\ &\leq \left(1+\frac{t}{\lambda}\right)^{-\mu} \operatorname{length}\gamma = A(t+\lambda)^{-\mu}, \end{aligned}$$

where $A = \lambda^\mu \operatorname{length}\gamma$. In particular $d(\beta_0(t),\beta_1(t)) \to 0$ as $t \to \infty$. □

6.2. Bi-infinite geodesics. A *bi-infinite geodesic* is a geodesic $\beta : \mathbb{R} \longrightarrow X$ parameterised by arc-length. We say that β joins $x \in X_I^\infty$ to $y \in X_I^\infty$ if $\beta(-t) \to x$ and $\beta(t) \to y$ as $t \to \infty$. Clearly the points x and y are determined by β. We refer to them as the "endpoints" of β. Since $d(\beta(-t),\beta(t)) = 2|t|$, the rays $[t \mapsto \beta(-t)]$ and $[t \mapsto \beta(t)]$ for $t \geq 0$ are not asymptotic. Thus the endpoints of β must be distinct. Moreover, the endpoints determine β up to reparameterisation:

Lemma 6.2.1. *Suppose that the bi-infinite geodesics $\alpha, \beta : \mathbb{R} \longrightarrow X$ have the same endpoints. Then, there is some $t_0 \in \mathbb{R}$ such that $\beta(t) = \alpha(t+t_0)$.*

Proof. Let $y \in X_I^\infty$ be the common endpoint so that $\alpha \to y$ and $\beta \to y$ as $t \to \infty$. Let h be a horofunction about y. There is some $t_0 \in \mathbb{R}$ such that $h(\alpha(t+t_0)) = h(\beta(t))$ for all $t \in \mathbb{R}$. By Lemma 6.1.1, we have $d(\alpha(t+t_0),\beta(t)) \to 0$ as $t \to \infty$. Also $d(\alpha(t+t_0),\beta(t))$ is bounded as $t \to -\infty$. By Proposition 3.5.6, the map $[t \mapsto d(\alpha(t+t_0),\beta(t))]$ is convex, and thus identically zero. □

We next want to establish the existence of a bi-infinite geodesic joining any pair of distinct points of X_I^∞.

Lemma 6.2.2. *Suppose $p \in X$. Then, for all $\theta > 0$, there exists $r > 0$ such that $x, y \in X \cup X_I^\infty$, then either $d(p,[x,y]) \leq r$ or else $x\hat{a}y \leq \theta$.*

Proof. Let $R = R(p)$ and $L > 0$ be the constants defined at the start of this chapter. Let $r = R\max(1, e^{2\pi L^2/\theta})$. Suppose, for contradiction, that $d(p,[x,y]) \geq r$, and $x\hat{p}y \geq \theta$. We form a ruled surface T by joining p to each point $w \in [x,y]$ with a geodesic $[p,w]$ (c.f. Lemma 4.3.2). Thus T is a non-positively curved 3-gon with vertices p, x and y. By Gauss-Bonnet, we have $-\int_T \kappa(z)d\omega(z) \leq \pi$ where $d\omega$ is the area element of T. As in Lemma 4.3.2, we obtain the contradiction:

$$\begin{aligned}\pi &\geq -\int_T \kappa(z)d\omega(z) \geq \int_R^r \frac{\theta t}{L^2 t^2}dt \\ &= \frac{\theta}{L^2}\log(r/R) \geq 2\pi.\end{aligned}$$

□

Proposition 6.2.3. *If $x, y \in X_I^\infty$, and $x \neq y$, then there is a bi-infinite geodesic joining x to y.*

Proof. Fix any $p \in X$. Thus $x\hat{p}y > 0$. Choose sequences $x_n \in [p,x] \cap X$ and $y_n \in [p,y] \cap X$ with $x_n \to x$ and $y_n \to y$. By Lemma 6.2.2, we can find points $z_n \in [x_n, y_n]$ with $d(p,z_n)$ bounded. Since $(X_C, \tau(X_C))$ is compact metrisable, we can assume that z_n converges to a point $z \in X_C$. By the lower-semicontinuity of the distance function (Proposition 5.1.4), we see that $d(p,z) < \infty$ and so $z \in X \cup X_I^0$. Thus, by Lemma 4.1.4, we can construct the geodesics $[z,x]$ and $[z,y]$.

Now choose any $a \in [z,x]\backslash\{z,x\}$ and $b \in [z,y]\backslash\{z,y\}$. We claim that $d(a,z) + d(z,b) = d(a,b)$. By Proposition 5.3.2, the geodesic $[z_n, x_n]$ tends to $[z,x]$ in the Hausdorff topology. Since the metric topology on X agrees with that induced by $\tau(X_C)$, we have, in particular, that $d(a,[x_n,z_n]) \to 0$. Similarly $d(b,[y_n,z_n]) \to 0$. Thus we can find $a_n \in [x_n,z_n]$ and $b_n \in [y_n,z_n]$ with $d(a,a_n) \to 0$ and $d(b,b_n) \to 0$. Now $d(a_n,z_n) + d(z_n,b_n) = d(a_n,b_n)$ and so the claim follows. Thus, since $[a,b]$ is the unique geodesic from a to b, we have that $z \in [a,b]$. It follows that $z \in X$, and $[x,z] \cup [z,y]$ gives a bi-infinite geodesic joining x to y. □

If $x, y \in X_I^\infty$ and $x \neq y$, we write $[x,y] = \{x,y\} \cup \text{image}\,\beta$, where β is the unique (up to parameterisation) geodesic joining x to y. It is easily seen that $[x,y]$ is closed in $(X_C, \tau(X_C))$. Note that $[x,y] = [y,x]$. We write $[x,x] = \{x\}$.

6.3. Continuity of geodesics.

Lemma 6.3.1. *Suppose $p \in X$, and $\kappa(p) < 0$. Then for all $\epsilon > 0$, there exists $\delta > 0$ such that if $x, y \in X_C \backslash \{p\}$ with $x\hat{p}y \geq \pi - \delta$, then $d(p, [x, y]) \leq \epsilon$.*

Proof. By continuity of κ, we have constants $h > 0$ and $k > 0$ such that $N(p, h) \subseteq X$ and $\kappa(z) \leq -k$ for all $z \in N(p, h)$. Given $\epsilon \in (0, h)$, let $\delta = \min(\pi/2, k\pi h^2/4)$. Suppose that $x, y \in X_C \backslash \{p\}$ are distinct with $d(p, [x, y]) \geq \epsilon$. Let $\theta = x\hat{p}y$. We claim that $\theta \leq \pi - \delta$. We can suppose that $\theta \geq \pi/2$. For the moment, assume that $x, y \in X \cup X_I^0$. We form a ruled surface by joining each $w \in [x, y]$ to p by the geodesic $[p, w]$ (c.f. Lemma 4.3.2). Integrating the curvature, we find that

$$\pi - \theta \geq \int_0^h k\left(\frac{\pi}{2}t\right) dt = k\pi h^2/4 \geq \delta.$$

Thus $\theta \leq \pi - \delta$ as required.

We can deal with the general case by taking the sequences $x_n, y_n \in [x, y] \cap X$ with $x_n \to x$ and $y_n \to y$, and noting that $\overrightarrow{px_n} \to \overrightarrow{px}$ and $\overrightarrow{py_n} \to \overrightarrow{py}$. □

We give $X_C \times X_C$ the product topology, and give $\mathscr{C}(X_C)$ the Hausdorff topology.

Proposition 6.3.2. *The map $[(x, y) \mapsto [x, y]] : X_C \times X_C \longrightarrow \mathscr{C}(X_C)$ is continuous.*

Proof. Note that Lemmas 5.1.1 and 5.1.3 generalise easily to the case where $(u, v) \in X_I^\infty \times X_I^\infty$, with essentially the same proofs. Thus the argument of Proposition 5.3.2 works to show that the map $[(x, y) \mapsto [x, y]]$ extended to all of $X_C \times X_C$ is continuous at each point $(x, y) \in (X_C \times X_C) \backslash (X_I^\infty \times X_I^\infty)$. It thus remains to show that it is continuous at each point $(x, y) \in X_I^\infty \times X_I^\infty$. There are two cases.

Case (1): $x \neq y$.

Fix some $p \in [x, y] \cap X$ with $\kappa(p) < 0$. Suppose $\epsilon > 0$ and $r > 0$. Let $\delta > 0$ be the constant given by Lemma 6.3.1, and set $\eta = \min(\epsilon, r \sin(\delta/4))$. Choose points $a \in [p, x]$ and $b \in [p, y]$, with $d(p, a) = d(p, b) = r + 2\epsilon$. Let $U = C(p, a, \eta)$ and $V = C(p, b, \eta)$. If $u \in U$ and $v \in V$, then by the Angle Comparison Theorem (Proposition 3.4.4), we find that $x\hat{p}u \leq \delta/2$ and $y\hat{p}v \leq \delta/2$. Thus $u\hat{p}v \geq \pi - \delta$ and so $d(p, [u, v]) \leq \epsilon$. Thus, there is some $q \in [u, v]$ with $d(p, q) \leq \epsilon$. If $u \in X_I^\infty$, then $[p, u]$ and $[q, u]$ are asymptotic, and so, since $d(a, [p, u]) \leq \eta \leq \epsilon$, we can find $u' \in [q, u]$ with $d(a, u') \leq 2\epsilon$. If $u \in X \cup X_I^0$, we can apply The Angle Comparison Theorem to

find such a u'. Similarly, we can find $v' \in [q, v]$ with $d(b, v') \leq 2\epsilon$. Note that $d(p, u') \geq r$ and $d(p, v') \geq r$. By convexity of the distance function, we have $[u', v'] \subseteq N([a, b], 2\epsilon)$ and $[a, b] \subseteq N([u', v'], 2\epsilon)$. Also $[x, a] \subseteq C(p, u', 2\epsilon)$ and $[y, b] \subseteq C(p, v', 2\epsilon)$. If $z \in [u, u']$, then by angle comparison, applied to zpq, we see that $d(u', [p, z]) \leq 2\epsilon$, and so $d(a, [p, z]) \leq 4\epsilon$. This shows that $[u, u'] \subseteq C(p, a, 4\epsilon)$. Similarly, $[v, v'] \subseteq C(p, b, 4\epsilon)$. Since $a, b, u', v' \in A(p, r)$, we have that $[x, y] \subseteq W(p, r, 4\epsilon)[u, v]$ and $[u, v] \subseteq W(p, r, 4\epsilon)[u, v]$. In other words, $[u, v] \in P(W(p, r, 4\epsilon))\{[x, y]\}$. As $\epsilon \to 0$ and $r \to \infty$, the sets $P(W(p, r, 4\epsilon))\{[x, y]\}$ form a neighbourhood base for $[x, y]$ in the Hausdorff topology on $\mathscr{C}(X_C)$.

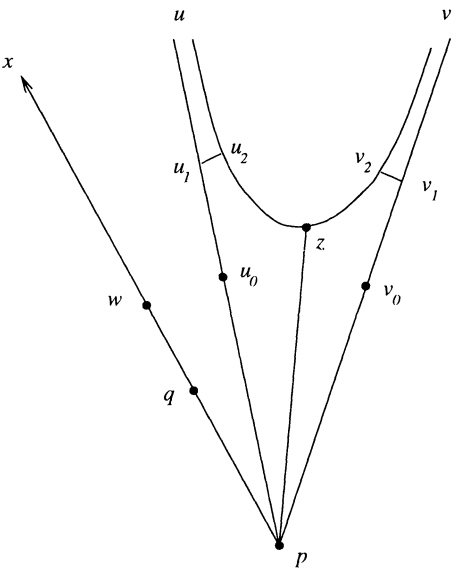


Figure 6.

Case (2): $x = y$.

Choose any point $p \in X$. Suppose $p \in X$. Suppose $r_0 > 0$ and $\epsilon > 0$. Let $q \in [p, x]$ be the point with $d(p, q) = r_0$. By the continuity of the logarithm map (Section 4.5), there is some $\theta > 0$ such that if $q' \in X \cup X_I^0$ with $d(p, q') = r_0$ and $q\hat{p}q' \leq 2\theta$, then $q' \in N(q, \epsilon)$. Thus if $z \in X_C$ with $d(p, z) \geq r_0$ and $x\hat{p}z \leq 2\theta$, then $z \in C(p, q, \epsilon)$.

Given $\theta > 0$, and $p \in X$, let $r > 0$ by the constant given by Lemma 6.2.2. Choose any $\eta > 0$ and let $R = \max(r_0 + 4\eta, r + 5\eta, \eta \operatorname{cosec} \theta)$. Let $w \in [p, y]$ be the point with $d(p, w) = r$. Thus, by angle comparison, if $u \in C(p, w, \eta)$ then $y\hat{p}u \leq \theta$.

Now suppose that $u, v \in C(p, w, \eta)$. Choose $u_0 \in [p, u]$ and $v_0 \in [p, v]$ with $d(w, u_0) \leq \eta$ and $d(w, v_0) \leq \eta$. Suppose $z \in [u, v] \setminus \{u, v\}$. If $u \in X_I^\infty$,

then $[u_0, u]$ and $[z, u]$ are asymptotic, and so we can find $u_1 \in [u_0, u]$ and $u_2 \in [z, u]$ with $d(u_1, u_2) \leq \eta$ (Lemma 6.1.1). If $u \in X_I^0$, take $u_1 = u_2 = u$. Similarly, we find $v_1 \in [v_0, v]$ and $v_2 \in [z, v]$ with $d(v_1, v_2) \leq \eta$. (Figure 6.) Thus

$$\begin{aligned} 2d(p,z) &\geq d(p,u_1) + d(p,v_1) - d(z,u_1) - d(z,v_1) \\ &\geq d(p,u_0) + d(p,v_0) + (d(u_0,u_1) + d(v_0,v_1) - d(u_2,v_2)) - 2\eta \\ &\geq 2d(p,w) - 8\eta, \end{aligned}$$

and so $d(p,z) \geq d(p,w) - 4\eta \geq \max(r_0, r+\eta)$. Since z is arbitrary, we see that $d(p,[u,v]) \geq r + \eta$. Given this, we see in particular that $d(p,[z,u_2]) \geq r + \eta$ and so $d(p,[z,u_1]) \geq r$. Thus $zpu_1 = z\hat{p}u \leq \theta$. Since also $x\hat{p}u \leq \theta$ we have $x\hat{p}z \leq 2\theta$. Since $d(p,z) \geq r$, it follows that $z \in C(p,q,\epsilon)$.

We have shown that if $u, v \in C(p,w,\eta)$, then $[u,v] \subseteq C(p,q,\epsilon)$. We deduce that $[u,v] \in P(W(p,r_0,\epsilon))\{\{x\}\}$. As $r_0 \to \infty$ and $\epsilon \to 0$, these sets form a neighbourhood base for $\{x\} = [x,x]$ in the Hausdorff topology on $\mathscr{C}(X_C)$. □

References

[AbS] U. Abresch and V. Schroeder, *Graph manifolds, ends of negatively curved spaces and the hyperbolic* 120*-cell space*, J. Diff. Geom., **35** (1992), 299–336.

[Ah] L.V. Ahlfors, *Curvature properties of Teichmüller space* , J. d'Anal. Math., **9** (1961), 161–176.

[Alek] A.D. Aleksandrov, *Ruled surfaces in metric spaces*, Vestnik Leningrad Univ. **12** (1957), 5–26 (Russian).

[AlexB] S.B. Alexander and R.L.Bishop, *The Hadamard-Cartan theorem in locally convex metric spaces*, L'Enseign. Math. **36** (1990), 309–320.

[An] M.T. Anderson, *The Dirichlet problem at infinity for manifolds of negative curvature*, J. Diff. Geom., **18** (1983), 701–721.

[BaGS] W. Ballmann, M. Gromov and V.Schroeder, *Manifolds of non-positive curvature*, Progress in Maths., **61**, Birkhäuser (1985).

[BiO] R.L. Bishop and B. O'Neill, *Manifolds of negative curvature*, Trans. Amer. Math. Soc., **145** (1969), 1–49.

[Bo] B.H. Bowditch, *Some results on the geometry of convex hulls in manifolds of pinched negative curvature*, Comment. Math. Helv., **69** (1994), 49–81.

[BrH] M.R. Bridson and A. Haefliger, *Non-positively curved metric spaces*, in preparation.

[EO] P. Eberlein and B. O'Neill, *Visibility manifolds*, Pacific J. Math., **46** (1973), 45–110.

[GH] E. Ghys, P.de la Harpe (ed.), *Sur les groupes hyperboliques d'après Mikhael Gromov*, Progress in Maths., **83**, Birkhäuser (1990).

[Ha] J. Hass, *Bounded 3-manifolds admit negatively curved metrics with concave boundary*, J. Diff. Geom., **40** (1994), 449–459.

[HeI] E. Heintze and H.C. Im Hof, *Geometry of horospheres*, J. Diff. Geom., **12** (1977), 481–491.

[K] J.L. Kelley, *General topology*, Graduate Texts in Maths., **21**, Springer-Verlag (reprint of Van Nostrand edition 1955).

[S] M. Spivak, *A comprehensive introduction to differential geometry*, Publish or Perish (1979).

[Tra] S. Trapani, *On the determinant of the bundle of meromorphic differentials on the Deligne-Mumford compactification of the moduli space of Riemann surfaces*, Math. Annalen, **293** (1992), 481–459.

[Tro] A. Tromba, *On a natural algebraic affine connection on the space of almost complex structures and the curvature of Teichmüller space with respect to its Weil-Peterssen metric*, Manuscripta Math., **56** (1986), 475–497.

[W1] S.A. Wolpert, *Non-completeness of the Weil-Peterssen metric for Teichmüller space*, Pacific J. Math., **61** (1975), 573–577.

[W2] ———, *Geodesic length functions and the Neilsen problem*, J. Diff. Geom., **25** (1987), 275–296.

Received July 30, 1993 and revised June 12, 1995. This article was prepared at the University of Melbourne, with the support of an A.R.C. Fellowship.

UNIVERSITY OF SOUTHAMPTON
HIGHFIELD, SOUTHAMPTON SO17 1BJ
GREAT BRITAIN
E-mail address: bhb@maths.soton.ac.uk

Current address:
UNIVERSITY OF MELBOURNE
PARKVILLE, VICTORIA 3052
AUSTRALIA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

THE QUASI-LINEARITY PROBLEM FOR C^*-ALGEBRAS

L.J. BUNCE AND J.D. MAITLAND WRIGHT

Let $\mathcal{A}$ be a C^*-algebra with no quotient isomorphic to the algebra of all two-by-two matrices. Let μ be a quasi-linear functional on $\mathcal{A}$. Then μ is linear if, and only if, the restriction of μ to the closed unit ball of $\mathcal{A}$ is uniformly weakly continuous.

Introduction.

Throughout this paper, $\mathcal{A}$ will be a C^*-algebra and A will be the real Banach space of self-adjoint elements of $\mathcal{A}$. The unit ball of A is A_1 and the unit ball of $\mathcal{A}$ is $\mathcal{A}_1$. We do not assume the existence of a unit in $\mathcal{A}$.

Definition. A *quasi-linear functional* on A is a function $\mu : A \to \mathbb{R}$ such that, whenever B is an abelian subalgebra of A, the restriction of μ to B is linear. Furthermore μ is required to be bounded on the closed unit ball of A.

Given any quasi-linear functional μ on A we may extend it to $\mathcal{A}$ by defining

$$\tilde{\mu}(x + iy) = \mu(x) + i\mu(y)$$

whenever $x \in A$ and $y \in A$. Then $\tilde{\mu}$ will be linear on each maximal abelian *-subalgebra of $\mathcal{A}$. We shall abuse our notation by writing 'μ' instead of '$\tilde{\mu}$'.

When $\mathcal{A} = \mathrm{M}_2(\mathbb{C})$, the C^* -algebra of all two-by-two matrices over $\mathbb{C}$, there exist examples of quasi-linear functionals on $\mathcal{A}$ which are not linear.

Definition. A local *quasi-linear functional* on A is a function $\mu : A \to \mathbb{R}$ such that, for each x in A, μ is linear on the smallest norm closed subalgebra of A containing x. Furthermore μ is required to be bounded on the closed unit ball of A.

Clearly each quasi-linear functional on A is a local quasi-linear functional. Surprisingly, the converse is false, even when A is abelian (see Aarnes [**2**]). However when A has a rich supply of projections (e.g. when $\mathcal{A}$ is a von Neumann algebra) each local quasi-linear functional is quasi-linear [**3**].

The solution of the Mackey-Gleason Problem shows that every quasi-linear functional on a von Neumann algebra $\mathcal{M}$, where $\mathcal{M}$ has no direct summand of Type I_2, is linear [**4**, **5**, **6**]. This was first established for positive quasi-linear functionals by the conjunction of the work of Christensen [**7**] and

Yeadon [**11**], and for σ-finite factors by the work of Paschciewicz [**10**]. All build on the fundamental theorem of Gleason [**8**].

Although quasi-linear functionals on general C^* -algebras seem much harder to tackle than the von Neumann algebra problem, we can apply the von Neumann results to make progress. In particular, we prove:

Let $\mathcal{A}$ be a *C^*-algebra with no quotient isomorphic to* $\mathrm{M}_2(\mathbb{C})$. *Let μ be a (local) quasi-linear functional on A. Then μ is linear if, and only if, the restriction of μ to A_1, is uniformly weakly continuous.*

1. Preliminaries: Uniform Continuity.

Let X be a real or complex vector space. Let $\mathcal{F}$ be a locally convex topology for X. Let V be a $\mathcal{F}$ -open neighbourhood of 0. We call V *symmetric* if V is convex and, whenever $x \in V$ then $-x \in V$.

Let B be a subset of X. A scalar valued function on X, μ, is said to be *uniformly continuous* on B, with respect to the $\mathcal{F}$ -topology, if, given any $\epsilon > 0$, there exists an open symmetric neighbourhood of $0, V$, such that whenever $x \in B$, $y \in B$ and $x - y \in V$ then

$$|\mu(x) - \mu(y)| < \epsilon.$$

Lemma 1.1. *Let X be a Banach space and let $\mathcal{F}$ be any locally convex topology for X which is stronger than the weak topology. Let μ be any bounded linear functional on X. Then μ is uniformly $\mathcal{F}$-continuous on X.*

Proof. Choose $\epsilon > 0$. Let

$$\begin{aligned} V &= \{x \in X : |\mu(x)| < \epsilon\} \\ &= \mu^{-1}\{\lambda : |\lambda| < \epsilon\}. \end{aligned}$$

Then V is open in the weak topology of X. Hence V is a symmetric $\mathcal{F}$-open neighbourhood of o such that $x - y \in V$ implies

$$|\mu(x) - \mu(y)| = |\mu(x - y)| < \epsilon.$$

□

Lemma 1.2. *Let X be a subspace of a Banach space Y. Let $\mathcal{G}$ be a locally convex topology for Y which is weaker than the norm topology. Let $\mathcal{F}$ be the relative topology induced on X by $\mathcal{G}$. Let B be a subset of X and let C be the closure of B in Y, with respect to the $\mathcal{G}$ -topology. Let $\mu : B \to \mathbb{C}$ be uniformly continuous on B with respect to the $\mathcal{F}$ -topology. Then there exists*

a function $\overline{\mu} : C \to \mathbb{C}$ which extends μ and which is uniformly $\mathcal{G}$-continuous. Furthermore, if μ is bounded on B then $\overline{\mu}$ is bounded on C.

Proof. Since $\mathcal{F}$ is the relative topology induced by $\mathcal{G}$, μ is uniformly $\mathcal{G}$-continuous on B. Let K be the closure of $\mu[B]$ in $\mathbb{C}$. Then K is a complete metric space. So, see [**9**, page 125], μ has a unique extension to $\overline{\mu} : C \to K$ where $\overline{\mu}$ is uniformly $\mathcal{G}$-continuous.

If μ is bounded on B then K is bounded and so $\overline{\mu}$ is bounded on C. □

Lemma 1.3. *Let X be a Banach space. Let X_1 be the closed unit ball of X and let X_1^{**} be closed unit ball of X^{**}. Let $\mu : X_1 \to \mathbb{C}$ be a bounded function which is uniformly weakly continuous. Then μ has a unique extension to $\overline{\mu} : X_1^{**} \to \mathbb{C}$ where $\overline{\mu}$ is bounded and uniformly weak*-continuous.*

Proof. Let $\mathcal{G}$ be the weak*-topology on X^{**}. For each $\phi \in X^*$

$$X \cap \{x \in X^{**} : |\phi(x)| < 1\} = \{x \in X : |\phi(x)| < 1\}.$$

So $\mathcal{G}$ induces the weak topology on X. So μ is uniformly $\mathcal{G}$-continuous on X_1. Since X_1 is dense in X_1^{**}, with respect to the $\mathcal{G}$-topology, it follows from Lemma 1.2 that $\overline{\mu}$ exists and has the required properties. □

2. Algebraic Preliminaries.

Lemma 2.1. *Let $\mathcal{B}$ be a non-abelian C^*-subalgebra of a von Neumann algebra $\mathcal{M}$, where $\mathcal{M}$ is of Type I_2. Then $\mathcal{B}$ has a surjective homomorphism onto $\mathrm{M}_2(\mathbb{C})$, the algebra of all two-by-two complex matrices.*

Proof. We have $\mathcal{M} = \mathrm{M}_2(\mathbb{C})\overline{\otimes}C(S)$ where S is hyperstonian. For each $s \in S$ there is a homomorphism π_S from $\mathcal{M}$ onto $\mathrm{M}_2(\mathbb{C})$ defined by

$$\pi_S \begin{Bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{Bmatrix} = \begin{Bmatrix} x_{11}(s) & x_{12}(s) \\ x_{21}(s) & x_{22}(s) \end{Bmatrix}.$$

Clearly, if $\pi_S[\mathcal{B}]$ is abelian for every s then $\mathcal{B}$ is abelian. So, for some s, $\pi_S[\mathcal{B}]$ is a non-abelian*-subalgebra of $\mathrm{M}_2(\mathbb{C})$ and so equals $\mathrm{M}_2(\mathbb{C})$. □

Lemma 2.2. *Let π be a representation of a C^*-algebra $\mathcal{A}$ on a Hilbert space H. Let $\mathcal{M} = \pi[\mathcal{A}]''$ where the von Neumann algebra $\mathcal{M}$ has a direct summand of Type I_2. Then $\mathcal{A}$ has a surjective homomorphism onto $\mathrm{M}_2(\mathbb{C})$.*

Proof. Let e be a central projection of $\mathcal{M}$ such that $e\mathcal{M}$ is of Type I_2. Since $\pi[\mathcal{A}]$ is dense in $\mathcal{M}$ in the strong operator topology, $e\pi[\mathcal{A}]$ is dense in $e\mathcal{M}$. Since $e\mathcal{M}$ is not abelian neither is $e\pi[\mathcal{A}]$. So, by the preceding lemma, $e\pi[\mathcal{A}]$, and hence $\mathcal{A}$, has a surjective homomorphism onto $\mathrm{M}_2(\mathbb{C})$. □

3. Linearity.

We now come to our basic theorem.

Theorem 3.1. *Let $\mathcal{A}$ be a C^*-algebra which has no quotient isomorphic to $M_2(\mathbb{C})$. Let π be a representation of $\mathcal{A}$ on a Hilbert space H. Let $\mathcal{M}$ be the closure of $\mathcal{A}$ in the strong operator-topology of $L(H)$. Let μ be a local quasi-linear functional on $\pi[A]$, which is uniformly continuous on the closed unit ball of $\pi[A]$ with respect to the topology induced on $\pi[A]$ by the strong operator topology of $L(H)$. Then μ is linear.*

Proof. We may suppose, by restricting to a closed subspace of H if necessary, that $\pi[\mathcal{A}]$ has an upward directed net converging, in the strong operator topology to the identity of H. Clearly $\pi[\mathcal{A}]$ has no quotient isomorphic to $M_2(\mathbb{C})$ for, otherwise, $M_2(\mathbb{C})$ would be a quotient of $\mathcal{A}$.

So, to simplify our notation we shall suppose that $\mathcal{A} = \pi[\mathcal{A}] \subset L(H)$.

Let $\mathcal{M}$ be the double commutant of $\mathcal{A}$ in $L(H)$. Let M_1 be the set of all self-adjoint elements in the unit ball of M. Then, by the Kaplansky Density Theorem, A_1 is dense in M_1 with respect to the strong operator-topology of $L(H)$.

Then, by Lemma 1.2, there exists $\overline{\mu} : M_1 \to \mathbb{C}$ such that $\overline{\mu}$ is an extension of $\mu \mid A_1$ and such that $\overline{\mu}$ is continuous with respect to the strong operator topology. Since $\mu[A_1]$ is bounded so, also, is $\overline{\mu}[M_1]$.

We know that for each $a \in A_1$ and each $t \in \mathbb{R}$,

$$\mu(ta) = t\mu(a).$$

We extend the definition of $\overline{\mu}$ to the whole of M by defining

$$\overline{\mu}(x) = \|x\|\overline{\mu}\left(\frac{1}{\|x\|}x\right)$$

whenever $x \in M$ with $\|x\| > 1$. It is then easy to verify that if (a_λ) is a bounded net in A which converges to x in the strong operator topology of $L(H)$ then

$$\mu(a_\lambda) \to \overline{\mu}(x).$$

Also, whenever $(x_n)(n = 1, 2..)$ is a bounded sequence in M, converging to x in the strong operator topology, then

$$\overline{\mu}(x_n) \to \overline{\mu}(x).$$

Let x be a fixed element of M and let (a_λ) be a bounded net in A which converges to x in the strong operator topology. Then, for each positive whole number n, $a_\lambda^n \to x^n$ in the strong operator topology. So $\mu(a_\lambda^n) \to \overline{\mu}(x^n)$.

Let ϕ_1, ϕ_2 be polynomials with real coefficients and zero constant term. Then, since μ is a local quasi-linear functional,

$$\mu\{\phi_1(a_\lambda)\} + \mu\{\phi_2(a_\lambda)\} = \mu\{(\phi_1 + \phi_2)(a_\lambda)\}.$$

Now

$$\phi_1(a_\lambda) \to \phi_1(x), \phi_2(a_\lambda) \to \phi_2(x).$$

and

$$(\phi_1 + \phi_2)(a_\lambda) \to (\phi_1 + \phi_2)(x)$$

in the strong operator topology. So

$$\overline{\mu}\{\phi_1(x)\} + \overline{\mu}\{\phi_2(x)\} = \overline{\mu}\{\phi_1(x) + \phi_2(x)\}.$$

Let $N(x)$ be the norm-closure of the set of all elements of the form $\phi(x)$, where ϕ is a polynomial with real coefficients and zero constant term. Then, since each norm convergent sequence is bounded and strongly convergent, $\overline{\mu}$ is linear on $N(x)$.

Let $p_1, p_2, ...p_n$ be orthogonal projections in M.

Let

$$x = p_1 + \frac{1}{2}p_2 + ... + \frac{1}{2^{n-1}}p_n + \frac{1}{2^n}\{1 - p_1 - p_2 - ... - p_n\}.$$

Then $(x^k)(k = 1, 2, ...)$ converges in norm to p_1. So p_1 is in $N(x)$. Then

$$\{(2x - 2p_1)^k\}\,(k = 1, 2, ...)$$

converges in norm to p_2. Similarly, $p_3, p_4, ...p_n$ and $1 - p_1 - p_2 - ... - p_n$ are all in $N(x)$.

Let $\nu(p) = \overline{\mu}(p)$ for each projection p in M. Then ν is a bounded finitely additive measure on the projections of M.

Since $\mathcal{A}$ has no quotient isomorphic to $M_2(\mathbb{C})$, it follows from Lemma 2.2 that $\mathcal{M}$ has no direct summand of Type I_2. Hence, by Theorem A of [**4**] or [**6**], ν extends to a bounded linear functional on $\mathcal{M}$, which we again denote by ν. From the argument of the preceding paragraph, $\overline{\mu}$ and ν coincide on finite (real) linear combinations of orthogonal projections. Hence by norm-continuity and spectral theory, $\overline{\mu}(x) = \nu(x)$ for each $x \in M$. Thus μ is linear. □

As an application of the above theorem, we shall see that when a quasi-linear functional μ has a "control functional", it is forced to be linear. We need a definition.

Definition. Let ϕ be a positive linear functional in $\overset{*}{\mathcal{A}}$ and let μ be a quasi-linear functional on $\mathcal{A}$. Then μ is said to be *uniformly absolutely continuous with respect to* ϕ if, given any $\epsilon > 0$ there can be found $\delta > 0$ such that, whenever $b \in A_1$ and $c \in A_1$ and $\phi((b-c)^2) < \delta$, then $|\mu(b) - \mu(c)| < \epsilon$.

Corollary 3.2. *Let $\mathcal{A}$ be a C^*-algebra which has no quotient isomorphic to $\mathrm{M}_2(\mathbb{C})$. Let μ be a local quasi-linear functional on $\mathcal{A}$ which is uniformly absolutly continuous with respect to ϕ, where ϕ is a positive linear functional in $\mathcal{A}^*$. Then μ is linear.*

Proof. Let (π, H) be the universal representation of $\mathcal{A}$ on its universal representation space H. We identify $\mathcal{A}$ with its image under π and identify $\pi[\mathcal{A}]''$ with $\mathcal{A}^{**}$.

Let ξ be a vector in H which induces ϕ, that is,

$$\phi(a) = \langle a\xi, \xi\rangle \text{for each } a \in \mathcal{A}.$$

Choose $\epsilon > 0$. Then, by hypothesis, there exists $\delta > 0$ such that, whenever $b \in A_1$ and $c \in A_1$ with

$$\|(b-c)\xi\|^2 < \delta$$

then

$$|\mu(b) - \mu(c)| < \epsilon.$$

So μ is uniformly continuous on A_1, with respect to the strong operator topology of $\mathrm{L}(H)$. Hence, by the preceding theorem μ is linear. □

Theorem 3.3. *Let $\mathcal{A}$ be a C^*-algebra with no quotient isomorphic to $\mathrm{M}_2(\mathbb{C})$. Let μ be a (local) quasi-linear functional on A. Then μ is a bounded linear functional if, and only if, μ is uniformly weakly continuous on the unit ball of A.*

Proof. By Lemma 1.1 each bounded linear functional on A is uniformly weakly continuous. We now assume that μ is uniformly weakly continuous on A_1. Let (π, H) be the universal representation of $\mathcal{A}$. Let $\mathcal{M} = \pi[\mathcal{A}]''$. Then A^{**} can be identified with $\mathcal{M}$ and A^{**} with M.

By Lemma 1.3 there exists a function $\overline{\mu} : M_1 \to \mathbb{C}$ which is uniformly continuous with respect to the weak*-topology on M_1 and such that $\overline{\mu}|A_1$ coincides with $\mu|A_1$.

The weak*-topology on M_1 coincides with the weak-operator topology of $\mathrm{L}(H)$, restricted to M_1. This is weaker than the strong operator-topology restricted to M_1. So $\overline{\mu}$ is uniformly continuous on M_1 with respect to the strong operator topology of $\mathrm{L}(H)$. Thus μ is uniformly continuous on A_1

with respect to the strong operator topology of $\mathrm{L}(H)$. Then, by Theorem 3.1, μ is linear. □

References

[1] J.F. Aarnes, *Quasi-states on C^*-algebras*, Trans. Amer. Math. Soc., **149** (1970), 601-625.

[2] J.F. Aarnes, (pre-print).

[3] C.A. Akemann and S.M. Newberger, *Physical states on a C^*-algebra*, Proc. Amer. Math. Soc., **40** (1973), 500.

[4] L.J. Bunce and J.D.M. Wright, *The Mackey-Gleason Problem*, Bull. Amer. Math. Soc., **26** (1992), 288-293.

[5] L.J. Bunce and J.D.M. Wright, *Complex Mesures on Projections in von Neumann Algebras*, J. London. Math. Soc., **46** (1992), 269-279.

[6] L.J. Bunce and J.D.M. Wright, *The Mackey-Gleason Problem for Vector Measures on Projections in Von Neumann Algebras*, J. London. Math. Soc., **49** (1994), 131-149.

[7] E. Christensen, *Measures on Projections and Physical states*, Comm. Math. Phys., **86** (1982), 529-538.

[8] A.M. Gleason, *Measures on the closed subspaces of a Hilbert space*, J. Math. Mech., **6** (1957), 885-893.

[9] J.L. Kelley, *General Topology*, Van Nostrand, (1953).

[10] A. Paszkiewicz, *Measures on Projections in W^*-factors*, J. Funct. Anal., **62** (1985), 295-311.

[11] F.W. Yeadon, *Finitely additive measures on Projections in finite W^*-algebras*, Bull. London Math. Soc., **16** (1984), 145-150.

Received June 25, 1993.

THE UNIVERSITY OF READING
READING RG6 2AX, ENGLAND

AND

ISAAC NEWTON INSTITUTE FOR MATHEMATICAL SCIENCES
20 CLARKSON ROAD
CAMBRIDGE, U.K.

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

DISTORTION OF BOUNDARY SETS UNDER INNER FUNCTIONS (II)

JOSÉ L. FERNÁNDEZ, DOMINGO PESTANA AND JOSÉ M. RODRÍGUEZ

We present a study of the metric transformation properties of inner functions of several complex variables. Along the way we obtain fractional dimensional ergodic properties of classical inner functions.

1. Introduction.

An inner function is a bounded holomorphic function from the unit ball $\mathbb{B}_n$ of $\mathbb{C}^n$ into the unit disk Δ of the complex plane such that the radial boundary values have modulus 1 almost everywhere. If E is a non empty Borel subset of $\partial\Delta$, we denote by $f^{-1}(E)$ the following subset of the unit sphere $\mathbb{S}_n$ of $\mathbb{C}^n$

$$f^{-1}(E) = \left\{\xi \in \mathbb{S}_n : \lim_{r\to 1} f(r\xi) \text{ exist and belongs to } E\right\} .$$

The classical lemma of Löwner, see e.g. [**R**, p. 405], asserts that inner functions f, with $f(0) = 0$, are measure preserving transformations when viewed as mappings from $\mathbb{S}_n$ to $\partial\Delta$, i.e. if E is a Borel subset of $\partial\Delta$ then $|f^{-1}(E)| = |E|$, where in each case $|\cdot|$ means the corresponding normalized Lebesgue measure.

In this paper we extend this result to fractional dimensions as follows:

Theorem 1. *If f is inner in the unit disk Δ, $f(0) = 0$, and E is a Borel subset of $\partial\Delta$, we have:*

$$\operatorname{cap}_\alpha\left(f^{-1}(E)\right) \geq \operatorname{cap}_\alpha(E)\,, \qquad 0 \leq \alpha < 1\,.$$

Moreover, if E is any Borel subset of $\partial\Delta$ with $\operatorname{cap}_\alpha(E) > 0$, equality holds if and only if either f is a rotation or $\operatorname{cap}_\alpha(E) = \operatorname{cap}_\alpha(\partial\Delta)$.

Moreover, it is well known, see [**N**], that if f is not a rotation then f is ergodic, i.e., there are no nontrivial sets A, with $f^{-1}(A) = A$ except for a set of Lebesgue measure zero. This also has a fractional dimensional parallel.

Corollary. *With the hypotheses of Theorem* 1, *if* f *is not a rotation and if the symmetric difference between* E *and* $f^{-1}(E)$ *has zero* α*-capacity, then either* $\operatorname{cap}_\alpha(E) = 0$ *or* $\operatorname{cap}_\alpha(E) = \operatorname{cap}_\alpha(\partial\Delta)$.

Theorem 2. *If* f *is inner in the unit ball of* $\mathbb{C}^n$, $f(0) = 0$, *and* E *is a Borel subset of* $\partial\Delta$, *we have:*

$$\operatorname{cap}_{2n-2+\alpha}\left(f^{-1}(E)\right) \geq K(n,\alpha)^{-1} \operatorname{cap}_\alpha(E)\,, \qquad 0 < \alpha < 1\,,$$

and

$$\frac{1}{\operatorname{cap}_{2n-2}\left(f^{-1}(E)\right)} \leq 1 + (2n-2)\log\frac{1}{\operatorname{cap}_0(E)}\,, \qquad (n > 1)\,.$$

Corollary. *In particular, for any inner function* f, *we have that*

$$\operatorname{Dim}\left(f^{-1}(E)\right) \geq 2n - 2 + \operatorname{Dim}(E)\,,$$

where Dim *denotes Hausdorff dimension.*

Here $\operatorname{cap}_\alpha$ and cap_0 denote, respectively, α-dimensional Riesz capacity and logarithmic capacity. We refer to [**C**], [**KS**] and [**L**] for definitions and basic background on capacity.

For background and some applications of these results we refer to [**FP**] where it is shown that Theorem 1 holds with some constants depending on α.

The outline of this paper is as follows: In Section 2 we obtain an integral expression for the α-energy that is used in Section 3, where Theorems 1 and 2 are proved. Section 4 contains some further results for the case $n = 1$. In Section 5, we prove an analogous distortion theorem, with Hausdorff measures replacing capacities. Section 6 discusses an open question and some partial results concerning distortion of subsets of the disc.

We would like to thank José Galé and Francisco Ruíz-Blasco for some helpful conversations concerning the energy functional. Also, we would like to thank David Hamilton for suggesting that the right constant in Theorem 1 is 1 (see [**H**]), and the referee for some valuable comments.

2. An integral expression for the α-energy.

In this section we obtain an expression of the α-energy of a signed measure μ in Σ_{N-1} (the unit sphere of $\mathbb{R}^N$) as an L^2-norm of its Poisson extension. This approach is due to Beurling [**B**].

If μ is a signed measure on Σ_{N-1}, and $0 \leq \alpha < N-1$, then the α-energy $I_\alpha(\mu)$ of μ is defined as

$$I_\alpha(\mu) = \iint_{\Sigma_{N-1}\times\Sigma_{N-1}} \Phi_\alpha(|x-y|)\, d\mu(x)\, d\mu(y)\,,$$

where

$$\Phi_\alpha(t) = \begin{cases} \log\dfrac{1}{t}\,, & \text{if } \alpha = 0\,, \\ \dfrac{1}{t^\alpha}\,, & \text{if } 0 < \alpha < N-1\,. \end{cases}$$

Recall that if E is a closed subset of Σ_{N-1}, then

$$(\mathrm{cap}_\alpha(E))^{-1} = \inf\{I_\alpha(\mu) :\ \mu \text{ a probability measure supported on } E\}\,,$$

for $0 < \alpha < N-1$,

$$\log\frac{1}{\mathrm{cap}_0(E)} = \inf\{I_0(\mu) :\ \mu \text{ a probability measure supported on } E\}\,,$$

and that the infimum is attained by a unique probability measure μ_e which is called the *equilibrium distribution* of E.

If E is any Borel subset of Σ_{N-1}, then the α-capacity of E is defined as

$$\mathrm{cap}_\alpha(E) = \sup\{\mathrm{cap}_\alpha(K) :\ K \subset E,\ K \text{ compact}\}\,.$$

We recall Choquet's theorem that all Borel sets are *capacitables*, i.e.

$$\mathrm{cap}_\alpha(E) = \inf\{\mathrm{cap}_\alpha(O) : E \subset O,\ O \text{ open}\}\,.$$

As we shall remark later on, for a general Borel set E of Σ_{N-1}, one has

$$\frac{1}{\mathrm{cap}_\alpha(E)} = \inf\{I_\alpha(\mu) :\ \mu \text{ a probability measure, } \mu(E) = 1\}\,,$$

and analogously for the logarithmic capacity.

We first need to obtain the expansion of the integral kernel Φ_α in terms of the spherical harmonics. We refer to [**SW**, Chap. IV] for details about spherical harmonics; we shall follow its notations.

Let $\mathcal{H}_k$ be the real vector space of the spherical harmonics of degree k in $\mathbb{R}^N$ $(N > 1)$. If a_k is the dimension of $\mathcal{H}_k$, we have

$$a_0 = 1, \quad a_1 = N, \quad a_k = \frac{N+2k-2}{k}\binom{N+k-3}{k-1}\,. \qquad \text{[SW, p. 145]}$$

If Σ_{N-1} denotes the unit sphere of $\mathbb{R}^N$, the space $L^2(\Sigma_{N-1}, d\xi)$ can be decomposed as

$$L^2(\Sigma_{N-1}, d\xi) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k \,,$$

where $d\xi$ is the usual Lebesgue measure (not normalized).

If ξ, η belongs to Σ_{N-1}, $Z_\eta^k(\xi)$ will denote the zonal harmonic of degree k with pole η, and if $\{Y_1^k, \ldots, Y_{a_k}^k\}$ is any orthonormal basis of $\mathcal{H}_k$, we have

$$Z_\eta^k(\xi) = \sum_{m=1}^{a_k} Y_m^k(\xi) Y_m^k(\eta) = Z_\xi^k(\eta). \qquad \text{[SW, p. 143]}$$

The zonal harmonics can be expressed in terms of the ultraspherical (or Gegenbauer) polynomials P_k^λ which are defined by the formula

$$(1 - 2rt + r^2)^{-\lambda} = \sum_{k=0}^{\infty} P_k^\lambda(t) r^k \,,$$

where $|r| < 1$, $|t| \leq 1$ and $\lambda > 0$.

We have [SW, p. 149], if $N > 2$,

$$Z_\eta^k(\xi) = C_{k,N} P_k^{(N-2)/2}(\xi \cdot \eta) \,.$$

It is easy to compute the constants $C_{k,N}$. First, if ω_{N-1} denotes the Lebesgue measure of Σ_{N-1}, then

$$\left\| Z_\eta^k \right\|_2^2 = \frac{a_k}{\omega_{N-1}}, \qquad \text{[SW, p. 144]}$$

while, on the other hand,

$$\begin{aligned} \frac{a_k}{\omega_{N-1}} &= C_{k,N}^2 \int_{\Sigma_{N-1}} \left| P_k^{(N-2)/2}(\xi \cdot \eta) \right|^2 d\xi \\ &= C_{k,N}^2 \, \omega_{N-2} \int_{-1}^{1} \left| P_k^{(N-2)/2}(t) \right|^2 (1 - t^2)^{(N-3)/2} \, dt \,. \end{aligned}$$

Now, the polynomials $P_k^{(N-2)/2}(t)$ form an orthogonal basis of

$$L^2\left([-1, 1],\ (1 - t^2)^{(N-3)/2} \, dt\right)$$

[SW, p. 151], [AS, p. 774], and

$$\left\| P_k^{(N-2)/2} \right\|_2^2 = \frac{\pi \, 2^{4-N} \Gamma(k + N - 2)}{k! \, (2k + N - 2) \Gamma\left(\frac{N-2}{2}\right)^2}, \qquad \text{[AS, p. 774]}$$

where $\Gamma(\cdot)$ denotes the Euler's Gamma function, and, therefore

$$C_{k,N}^2 = \frac{a_k}{\omega_{N-1}\,\omega_{N-2}} \left\| P_k^{(N-2)/2} \right\|_2^{-2} = \frac{(N+2k-2)^2}{16\,\pi^N} \Gamma\left(\frac{N-2}{2}\right)^2 .$$

Hence

$$C_{k,N} = \frac{N+2k-2}{4\pi^{N/2}} \Gamma\left(\frac{N-2}{2}\right),$$

and

$$Z_\eta^k(\xi) = \frac{N+2k-2}{4\pi^{N/2}} \Gamma\left(\frac{N-2}{2}\right) P_k^{(N-2)/2}(\xi\cdot\eta) .$$

The case $N = 2$ is slightly different. In this case we can take $P_k^0 \equiv T_k$, the Chebyshev's polynomials defined in $[-1,1]$ by

$$T_k(\cos\theta) = \cos k\theta .$$

It is known that these polynomials form an orthogonal basis of

$$L^2\left([-1,1],\ (1-t^2)^{-1/2}\,dt\right) .$$

In this particular case, if $\xi = e^{i\theta}, \eta = e^{i\psi}$, then $\xi\cdot\eta = \cos(\theta-\psi)$, and

$$\begin{aligned} Z_\eta^k(\xi) = \frac{1}{\pi}\cos k(\theta-\psi) &= \frac{1}{\pi} T_k(\cos(\theta-\psi)) \\ &= \frac{1}{\pi} P_k^0(\xi\cdot\eta), \qquad k = 1,2,\dots, \\ Z_\eta^0(\xi) = \frac{1}{2\pi} &= \frac{1}{2\pi} P_0^0(\xi\cdot\eta) . \end{aligned}$$

Therefore,

$$C_{k,2} = \begin{cases} \dfrac{1}{\pi}, & \text{if } k > 0, \\ \dfrac{1}{2\pi}, & \text{if } k = 0. \end{cases}$$

We can now write down the expansion of the kernel $\Phi_\alpha(|x-y|)$ in a Fourier series of Gegenbauer's polynomials. Fix, first, α, with $0 < \alpha < N-1$. If we denote by $g(t)$ the function

$$g(t) = \left(\frac{1}{2-2t}\right)^{\alpha/2},$$

then we can express the kernel Φ_α in terms of g as

$$\Phi_\alpha(|\xi-\eta|) = \Phi_\alpha\left(\sqrt{|\xi|^2 - 2\xi\cdot\eta + |\eta|^2}\right) = g(\xi\cdot\eta) .$$

Now, develop $g(t)$ as a Fourier series

$$g(t) = \sum_{k=0}^{\infty} g_k P_k^{(N-2)/2}(t), \quad \text{where} \quad g_k \left\| P_k^{(N-2)/2} \right\|_2^2 = \left\langle g, P_k^{(N-2)/2} \right\rangle,$$

and conclude

$$\Phi_\alpha(|\xi - \eta|) = g(\xi \cdot \eta) = \sum_{k=0}^{\infty} g^k Z_\eta^k(\xi), \tag{1}$$

where $g^k C_{k,N} = g_k$. Hereafter F will denote the usual hypergeometric function

$$F(a, b; c; t) = \sum_{m=0}^{\infty} \frac{(a)_m (b)_m}{(c)_m} \frac{t^m}{m!},$$

where

$$(u)_m = u(u+1)\dots(u+m-1) = \frac{\Gamma(u+m)}{\Gamma(u)}.$$

The polynomials $P_k^{(N-2)/2}$ can be expressed in terms of F [**AS**, p. 779].

If $N > 2$,

$$P_k^{(N-2)/2}(t) = \binom{k+N-3}{k} F(-k, k+N-2; (N-1)/2; (1-t)/2).$$

Then,

$$\left\langle g, P_k^{(N-2)/2} \right\rangle = \binom{k+N-3}{k} \int_{-1}^{1} F(-k, k+N-2; (N-1)/2; (1-t)/2) \cdot (2-2t)^{-\alpha/2} (1-t^2)^{(N-3)/2} \, dt.$$

Therefore

$$\left\langle g, P_k^{(N-2)/2} \right\rangle = 2^{N-2-\alpha} \binom{k+N-3}{k} \int_0^1 s^{-1+(N-1-\alpha)/2} (1-s)^{-1+(N-1)/2} \cdot F(-k, k+N-2; (N-1)/2; s) \, ds.$$

Using the relationship

$$P_k^{(N-2)/2}(-t) = (-1)^k P_k^{(N-2)/2}(t), \qquad [\mathbf{SW}, \text{p. } 149], [\mathbf{AS}, \text{p. } 775]$$

we have

$$\left\langle g, P_k^{(N-2)/2} \right\rangle = 2^{N-2-\alpha} \binom{k+N-3}{k} (-1)^k \int_0^1 s^{-1+(N-1-\alpha)/2} (1-s)^{-1+(N-1)/2} \cdot F(-k, k+N-2; (N-1)/2; 1-s)\, ds\,.$$

Term by term integration of the series defining F gives

$$\int_0^1 s^{a-1}(1-s)^{b-1} F(-k, c; b; 1-s)\, ds = B(a,b) F(-k, c; a+b; 1)\,,$$

where $B(\cdot,\cdot)$ is the Euler's Beta function. Moreover, it is easy to see that ([**AS**, p. 556])

$$\begin{aligned} F(-k, c; a+b; 1) &= \frac{\Gamma(a+b)\Gamma(a+b-c+k)}{\Gamma(a+b+k)\Gamma(a+b-c)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a+b+k)} (-1)^k \frac{\Gamma(1+c-a-b)}{\Gamma(1+c-a-b-k)}\,, \end{aligned}$$

and so

$$(-1)^k \int_0^1 s^{a-1}(1-s)^{b-1} F(-k, c; b; 1-s)\, ds = \frac{\Gamma(a)\Gamma(b)\Gamma(1+c-a-b)}{\Gamma(a+b+k)\Gamma(1+c-a-b-k)}\,.$$

This gives

$$\left\langle g, P_k^{(N-2)/2} \right\rangle = 2^{N-2-\alpha} \binom{k+N-3}{k} \frac{\Gamma\left(\frac{N-1-\alpha}{2}\right)\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(k+\frac{\alpha}{2}\right)}{\Gamma\left(N-1-\frac{\alpha}{2}+k\right)\Gamma\left(\frac{\alpha}{2}\right)}\,,$$

and

$$\begin{aligned} g_k &= \frac{\left\langle g, P_k^{(N-2)/2} \right\rangle}{\left\| P_k^{(N-2)/2} \right\|_2^2} \\ &= 2^{N-3-\alpha} \frac{N+2k-2}{\sqrt{\pi}} \frac{\Gamma\left(\frac{N-1-\alpha}{2}\right)\Gamma\left(\frac{N}{2}-1\right)\Gamma\left(k+\frac{\alpha}{2}\right)}{\Gamma\left(N-1-\frac{\alpha}{2}+k\right)\Gamma\left(\frac{\alpha}{2}\right)}\,. \end{aligned}$$

Therefore,

$$g^k = g_k C_{k,N}^{-1} = 2^{N-1-\alpha}\pi^{(N-1)/2}\frac{\Gamma\left(\frac{N-1-\alpha}{2}\right)\Gamma\left(k+\frac{\alpha}{2}\right)}{\Gamma\left(N-1-\frac{\alpha}{2}+k\right)\Gamma\left(\frac{\alpha}{2}\right)}, \tag{2}$$

if $N > 2$. On the other hand, if $N = 2$, the k-th Chebyshev's polynomial is $T_k(t) = F(-k, k; 1/2; (1-t)/2)$, (see [**AS**, p. 779]), and

$$\langle g, P_k^0\rangle = \int_{-1}^{1}(2-2t)^{-\alpha/2}F(-k,k;1/2;(1-t)/2)(1-t^2)^{-1/2}\,dt\,.$$

Using the above computations when $N = 2$, we have that

$$\langle g, P_k^0\rangle = 2^{-\alpha}\pi^{1/2}\frac{\Gamma\left(\frac{1-\alpha}{2}\right)\Gamma\left(k+\frac{\alpha}{2}\right)}{\Gamma\left(1-\frac{\alpha}{2}+k\right)\Gamma\left(\frac{\alpha}{2}\right)}.$$

Moreover it is easy to see, [**AS**, p. 774], that

$$\|P_k^0\|_2^2 = \begin{cases}\dfrac{\pi}{2}, & \text{if } k>0\,,\\ \pi\,, & \text{if } k=0\,,\end{cases}$$

and also that $C_{k,2}^{-1} = 2\,\|P_k^0\|_2^2$.

Then

$$g^k = \frac{\langle g, P_k^0\rangle}{\|P_k^0\|_2^2}C_{k,2}^{-1}\,,$$

and so (2) is also satisfied in this case ($N = 2$). Therefore we have proved the following:

Lemma 1. *For all $N \in \mathbb{N}$, $N > 1$ and $0 < \alpha < N-1$,*

$$\Phi_\alpha(|\xi-\eta|) = \sum_{k=0}^{\infty} g^k Z_\eta^k(\xi)\,,$$

where

$$g^k = 2^{N-1-\alpha}\pi^{(N-1)/2}\frac{\Gamma\left(\frac{N-1-\alpha}{2}\right)\Gamma\left(k+\frac{\alpha}{2}\right)}{\Gamma\left(N-1-\frac{\alpha}{2}+k\right)\Gamma\left(\frac{\alpha}{2}\right)}.$$

Now we can express the α-energy of a measure μ in terms of its Poisson extension P_μ.

Lemma 2. *If μ is a signed measure supported on Σ_{N-1}, we have:*

(i) *If $0 < \alpha < N-1$, then*

$$I_\alpha(\mu) = C(N,\alpha)\int_0^1 \left\{ \int_{\Sigma_{N-1}} |P_\mu(r\xi)|^2 \, d\xi \right\} r^{\alpha-1}(1-r^2)^{N-2-\alpha}\, dr\,,$$

with

$$C(N,\alpha) = \frac{4\pi^{N/2}}{\Gamma\left(\frac{\alpha}{2}\right)\Gamma\left(\frac{N-\alpha}{2}\right)}\,.$$

(ii) *If $m = \mu(\Sigma_{N-1})$, then*

$$I_0(\mu) = \omega_{N-1}\int_0^1 \int_{\Sigma_{N-1}} \left| P_\mu(r\xi) - \frac{m}{\omega_{N-1}} \right|^2 d\xi\,(1-r^2)^{N-2}\frac{dr}{r} + \frac{m^2}{2}\left[\frac{\Gamma'}{\Gamma}\left(\frac{N}{2}\right) - \frac{\Gamma'}{\Gamma}(N-1)\right].$$

In particular, if $N = 2$,

$$I_0(\mu) = 2\pi\int_0^1 \int_0^{2\pi} \left| P_\mu(re^{i\theta}) - \frac{m}{2\pi} \right|^2 d\theta\frac{dr}{r}\,.$$

Proof. Let $\left\{\mu_j^k\right\}$, $k \geq 0$, $1 \leq j \leq a_k$, be the Fourier coefficientes of μ, i.e.,

$$\mu \sim \sum_{k=0}^{\infty}\sum_{j=1}^{a_k} \mu_j^k Y_j^k\,.$$

Recall that P_μ is defined by

$$P_\mu(r\xi) = \int_{\Sigma_{N-1}} p(\eta, r\xi)\, d\mu(\eta)\,,$$

where $p(\eta, r\xi)$ is the classical (normalized) Poisson kernel

$$p(\eta, r\xi) = \frac{1}{\omega_{N-1}}\frac{1-r^2}{|\eta - r\xi|^N}\,.$$

We have [**SW**, p. 145]

$$p(\eta, r\xi) = \sum_{k=0}^{\infty} r^k Z_\eta^k(\xi) = \sum_{k,j} r^k Y_j^k(\eta) Y_j^k(\xi)\,.$$

Now, Plancherel's theorem gives that

$$P_\mu(r\xi) = \sum_{k,j} r^k \mu_j^k Y_j^k(\xi)\,.$$

Using again Plancherel's theorem we obtain that

$$\int_{\Sigma_{N-1}} |P_\mu(r\xi)|^2\, d\xi = \sum_{k,j} r^{2k} \left|\mu_j^k\right|^2\,,$$

and so if we denote by Λ the right hand side in (i), we have that

$$\Lambda = C(N,\alpha) \sum_{k,j} \left|\mu_j^k\right|^2 \int_0^1 r^{2k+\alpha-1}(1-r^2)^{N-2-\alpha}\, dr\,,$$

and, substituting $r^2 = t$, we get that

$$\Lambda = \frac{C(N,\alpha)}{2} \sum_{k,j} \frac{\Gamma\left(k+\frac{\alpha}{2}\right)\Gamma(N-1-\alpha)}{\Gamma\left(k+N-1-\frac{\alpha}{2}\right)} \left|\mu_j^k\right|^2 = \sum_{j,k} g^k \left|\mu_j^k\right|^2\,.$$

Note that we have used the known duplication formula for the Gamma function in the last equality.

On the other hand, by (1),

$$\Phi_\alpha(|\xi-\eta|) = \sum_{k=0}^{\infty} g^k Z_\eta^k(\xi) = \sum_{k,j} g^k Y_j^k(\eta) Y_j^k(\xi)\,,$$

and using Plancherel's theorem we obtain that

$$\int_{\Sigma_{N-1}} \Phi_\alpha(|\xi-\eta|)\, d\mu(\eta) = \sum_{k,j} g^k \mu_j^k Y_j^k(\xi)\,,$$

$$I_\alpha(\mu) = \sum_{k,j} g^k \left|\mu_j^k\right|^2 = \Lambda\,.$$

This finishes the proof of (i).

In order to prove (ii) observe that

$$\int_{\Sigma_{N-1}} \left|P_\mu(r\xi) - \frac{m}{\omega_{N-1}}\right|^2 d\xi + \frac{m^2}{\omega_{N-1}} = \int_{\Sigma_{N-1}} |P_\mu(r\xi)|^2\, d\xi\,.$$

Integrating this equality we have that

$$I_\alpha(\mu) = C(N,\alpha) \int_0^1 \int_{\Sigma_{N-1}} \left|P_\mu(r\xi) - \frac{m}{\omega_{N-1}}\right|^2 d\xi\, r^{\alpha-1}(1-r^2)^{N-2-\alpha}\, dr + m^2\, U(\alpha)\,,$$

where

$$U(\alpha) = \frac{\Gamma(N/2)\Gamma(N-1-\alpha)}{\Gamma\left((N-\alpha)/2\right)\Gamma(N-1-\alpha/2)},$$

and hence

$$\lim_{\alpha\to 0} \frac{I_\alpha(\mu) - m^2\, U(\alpha)}{\alpha}$$
$$= \omega_{N-1} \int_0^1 \int_{\Sigma_{N-1}} \left| P_\mu(r\xi) - \frac{m}{\omega_{N-1}} \right|^2 d\xi\, (1-r^2)^{N-2} \frac{dr}{r}.$$

On the other hand,

$$\lim_{\alpha\to 0} \frac{I_\alpha(\mu) - m^2\, U(\alpha)}{\alpha} = \lim_{\alpha\to 0} \frac{I_\alpha(\mu) - m^2}{\alpha} - m^2 \lim_{\alpha\to 0} \frac{U(\alpha) - 1}{\alpha}$$
$$= I_0(\mu) - m^2 U'(0),$$

and

$$U'(0) = \frac{1}{2}\left[\frac{\Gamma'}{\Gamma}\left(\frac{N}{2}\right) - \frac{\Gamma'}{\Gamma}(N-1)\right].$$

This finishes the proof of Lemma 2. □

3. Distortion of α-capacity.

We need the following lemmas.

Lemma 3. *Let μ be a finite positive measure in $\partial\Delta$, and let f be an inner function. Then, there exists a unique positive measure $\widetilde{\nu}$ in $\mathbb{S}_n$ such that $P_\mu \circ f = P_{\widetilde{\nu}}$ and*

$$\widetilde{\nu}\left(f^{-1}(\text{support } \mu)\right) = \widetilde{\nu}(\mathbb{S}_n).$$

Moreover, if $f(0) = 0$, then

$$\frac{1}{\omega_{2n-1}} \widetilde{\nu}(\mathbb{S}_n) = \frac{1}{2\pi} \mu(\partial\Delta).$$

Proof. It is essentially the same proof as that of Lemma 1 of [**FP**], but see Lemma 10 below for further details.

A different normalization is useful; choosing $\nu = (2\pi/\omega_{2n-1})\widetilde{\nu}$, one obtai

$$P_\nu = \frac{2\pi}{\omega_{2n-1}} P_\mu \circ f \qquad \text{and} \qquad \nu(\mathbb{S}_n) = \mu(\partial\Delta).$$

The following is well known

Lemma 4. (Subordination principle). *Let $f : \mathbb{B}_n \longrightarrow \Delta$ be a holomorphic function such that $f(0) = 0$, and let $v : \Delta \longrightarrow \mathbb{R}$ be a subharmonic function. Then*

$$\frac{1}{\omega_{2n-1}} \int_{\mathbb{S}_n} v(f(r\xi))\, d\xi \leq \frac{1}{2\pi} \int_0^{2\pi} v(re^{i\theta})\, d\theta .$$

It will be relevant later on to recall the well known fact that, in the case $n = 1$, equality in Lemma 4 holds for a given r, $0 < r < 1$, if and only if either v is harmonic in $\Delta_r = \{|z| < r\}$ or f is a rotation. Note also that there is no such equality statement when $n > 1$ since in higher dimensions the extremal functions in Schwarz's lemma are not so clearly determined (see e.g. [**R**, p. 164]).

Lemma 5. *Let μ be a signed measure on $\partial\Delta$, f an inner function with $f(0) = 0$, and ν a signed measure on $\mathbb{S}_n$ such that*

$$P_\nu = (2\pi/\omega_{2n-1}) P_\mu \circ f .$$

Then

(i) *If $n = 1$ and $0 \leq \alpha < 1$, then*

$$I_\alpha(\nu) \leq I_\alpha(\mu) .$$

(ii) *If $n > 1$ and $0 < \alpha < 1$, then*

$$I_{2n-2+\alpha}(\nu) \leq K(n, \alpha) I_\alpha(\mu) ,$$

where

$$K(n, \alpha) = \frac{(n-1)!\, \Gamma\left(\frac{\alpha}{2}\right)}{\Gamma\left(n - 1 + \frac{\alpha}{2}\right)} .$$

If $\alpha = 0$ and $m = \mu(\partial\Delta) = \nu(\mathbb{S}_n)$, we have

$$I_{2n-2}(\nu) \leq (2n-2) I_0(\mu) + m^2 .$$

The measure ν is obtained from Lemma 3 by splitting μ into its positive and negative parts. Note that for fixed α,

$$K(n, \alpha) \sim n^{1-\alpha/2}\, \Gamma\left(\frac{\alpha}{2}\right) , \qquad \text{as } n \to \infty ,$$

while for fixed $n > 1$

$$K(n, \alpha) \sim \frac{C_n}{\alpha} , \qquad \text{as } \alpha \to 0 .$$

Let us observe also that $K(n,\alpha)$ takes the value 1 for $n=1$.

Proof. Since $|P_\mu - \frac{m}{2\pi}|^2$ and $|P_\mu|^2$ are subharmonic, we obtain by subordination, Lemma 4, that if $n=1$ and $\alpha=0$

$$\int_0^{2\pi}\left|P_\nu-\frac{m}{2\pi}\right|^2 d\theta=\int_0^{2\pi}\left|P_\mu(f)-\frac{m}{2\pi}\right|^2 d\theta\le\int_0^{2\pi}\left|P_\mu-\frac{m}{2\pi}\right|^2 d\theta\,,$$

and if $n\ge 1$, $0<\alpha<1$, that

$$\int_{\mathbb{S}_n}|P_\nu|^2\,d\xi=\left(\frac{2\pi}{\omega_{2n-1}}\right)^2\int_{\mathbb{S}_n}|P_\mu(f)|^2\,d\xi\le\frac{2\pi}{\omega_{2n-1}}\int_0^{2\pi}|P_\mu|^2 d\theta\,. \tag{3}$$

In the first case, we obtain

$$I_0(\nu)\le I_0(\mu)$$

by integrating with respect to $2\pi\,dr/r$ and applying Lemma 2, part (ii).

In the second case, using Lemma 2, part (i), and Lemma 4 with $v=|P_\mu|^2$, we have that

$$\begin{aligned}I_{2n-2+\alpha}(\nu)&=C(2n,2n-2+\alpha)\int_0^1\left\{\int_{\mathbb{S}_n}|P_\nu(r\xi)|^2\,d\xi\right\}r^{2n-2+\alpha-1}\frac{dr}{(1-r^2)^\alpha}\\&\le\frac{C(2n,2n-2+\alpha)}{C(2,\alpha)}\,C(2,\alpha)\\&\quad\cdot\frac{2\pi}{\omega_{2n-1}}\int_0^1\left\{\int_0^{2\pi}|P_\mu(re^{i\theta})|^2\,d\theta\right\}r^{\alpha-1}\frac{dr}{(1-r^2)^\alpha}\\&=K(n,\alpha)I_\alpha(\mu)\,,\end{aligned}$$

where

$$K(n,\alpha)=\frac{(n-1)!\,\Gamma\left(\frac{\alpha}{2}\right)}{\Gamma\left(n-1+\frac{\alpha}{2}\right)}\,.$$

Finally, since $\nu(\mathbb{S}_n)=m$,

$$\int_{\mathbb{S}_n}\left|P_\nu(r\xi)-\frac{m}{\omega_{2n-1}}\right|^2 d\xi=\int_{\mathbb{S}_n}|P_\nu(r\xi)|^2\,d\xi-\frac{m^2}{\omega_{2n-1}}\,,$$

and so, Lemma 2 gives, if $n>1$, that

$$I_{2n-2}(\nu)=m^2+\frac{4\pi^n}{(n-2)!}\int_0^1\int_{\mathbb{S}_n}\left|P_\nu(r\xi)-\frac{m}{\omega_{2n-1}}\right|^2 d\xi\,r^{2n-3}\,dr\,.$$

By Lemmas 3 and 4, we get that

$$\begin{aligned}\int_{\mathbb{S}_n}\left|P_\nu(r\xi)-\frac{m}{\omega_{2n-1}}\right|^2 d\xi &= \int_{\mathbb{S}_n}\left|\frac{2\pi}{\omega_{2n-1}}P_\mu(f(r\xi))-\frac{m}{\omega_{2n-1}}\right|^2 d\xi\\ &= \left(\frac{2\pi}{\omega_{2n-1}}\right)^2\int_{\mathbb{S}_n}\left|P_\mu(f(r\xi))-\frac{m}{2\pi}\right|^2 d\xi\\ &\le \frac{2\pi}{\omega_{2n-1}}\int_0^{2\pi}\left|P_\mu(re^{i\theta})-\frac{m}{2\pi}\right|^2 d\theta\,.\end{aligned}$$

Therefore

$$\begin{aligned}I_{2n-2}(\nu) &\le m^2+\frac{4\pi^n}{(n-2)!}\int_0^1\frac{2\pi}{\omega_{2n-1}}\int_0^{2\pi}\left|P_\mu(re^{i\theta})-\frac{m}{2\pi}\right|^2 d\theta\,\frac{dr}{r}\\ &= m^2+\frac{4\pi^n}{(n-2)!}\frac{1}{\omega_{2n-1}}I_0(\mu)\\ &= m^2+(2n-2)I_0(\mu)\,.\end{aligned}$$

The proof of Lemma 5 is finished. □

Finally, we can prove

Theorem 1. *If f is inner in the unit disk Δ, $f(0)=0$, and E is a Borel subset of $\partial\Delta$, we have:*

$$\operatorname{cap}_\alpha\left(f^{-1}(E)\right)\ge\operatorname{cap}_\alpha(E)\,,\qquad 0\le\alpha<1\,.$$

Moreover, if E is any Borel subset of $\partial\Delta$ with $\operatorname{cap}_\alpha(E)>0$, equality holds if and only if either f is a rotation or $\operatorname{cap}_\alpha(E)=\operatorname{cap}_\alpha(\partial\Delta)$.

Notice the following consequence concerning invariant sets. It is well known that an inner function f with $f(0)=0$, which is not a rotation, is ergodic with respect to Lebesgue measure, see e.g. [**P**]. As a consequence of the above, it is also ergodic with respect to α-capacity. More precisely,

Corollary. *With the hypotheses of Theorem 1, if f is not a rotation and if the symmetric difference between E and $f^{-1}(E)$ has zero α-capacity, then either $\operatorname{cap}_\alpha(E)=0$ or $\operatorname{cap}_\alpha(E)=\operatorname{cap}_\alpha(\partial\Delta)$.*

In higher dimensions we have

Theorem 2. *If f is inner in the unit ball of $\mathbb{C}^n$, $f(0)=0$, and E is a Borel subset of $\partial\Delta$, we have:*

$$\operatorname{cap}_{2n-2+\alpha}\left(f^{-1}(E)\right)\ge K(n,\alpha)^{-1}\operatorname{cap}_\alpha(E)\,,\qquad 0<\alpha<1\,,$$

and

$$\frac{1}{\mathrm{cap}_{2n-2}\,(f^{-1}(E))} \leq 1 + (2n-2)\log\frac{1}{\mathrm{cap}_0(E)}\,, \qquad (n>1)\,.$$

Proof of Theorems 1 *and* 2. To prove the inequalities in the theorems we may assume that E is closed. Assume first that $n=1,\ 0<\alpha<1$. Let us denote by μ_e the α-equilibrium probability distribution of E, and let ν be the probability measure such that $P_\nu = P_{\mu_e}\circ f$. By Lemma 5,

$$I_\alpha(\nu) \leq I_\alpha(\mu_e) = (\mathrm{cap}_\alpha(E))^{-1}\,. \tag{4}$$

But, from Lemma 3, $\nu\,(f^{-1}(E)) = 1$, and so

$$I_\alpha(\nu) = \iint_{f^{-1}(E)\times f^{-1}(E)} \Phi_\alpha(|z-w|)\,d\nu(z)\,d\nu(w)\,.$$

Now, let $\{K_n\}$ be an increasing sequence of compacts subsets in $\partial\Delta$, $K_n \subset f^{-1}(E)$, such that $\nu(K_n)\nearrow 1$. Then, for each $n\geq 1$,

$$\begin{aligned} I_\alpha(\nu) &= \iint_{f^{-1}(E)\times f^{-1}(E)} \Phi_\alpha(|z-w|)\,d\nu(z)\,d\nu(w) \\ &\geq \nu\,(K_n)^2 \iint_{K_n\times K_n} \Phi_\alpha(|z-w|)\,\frac{d\nu(z)}{\nu\,(K_n)}\,\frac{d\nu(w)}{\nu\,(K_n)} \\ &\geq \nu\,(K_n)^2\ (\mathrm{cap}_\alpha\,(K_n))^{-1} \\ &\geq \nu\,(K_n)^2\ (\mathrm{cap}_\alpha\,(f^{-1}(E)))^{-1}\,, \end{aligned}$$

and consequently

$$I_\alpha(\nu) \geq (\mathrm{cap}_\alpha\,(f^{-1}(E)))^{-1}\,. \tag{5}$$

The inequality in Theorem 1 follows now from (4) and (5).

The cases $n>1$ (Theorem 2) and $n=1,\ \alpha=0$ are completely analogous.

Proof of the equality statement of Theorem 1. First we prove it assuming that E is closed, to show the ideas that we will use to demonstrate the general case.

Suppose that $0<\alpha<1$. We have seen that

$$\frac{1}{\mathrm{cap}_\alpha\,(f^{-1}(E))} \leq I_\alpha(\nu) \leq I_\alpha\,(\mu_e) = \frac{1}{\mathrm{cap}_\alpha(E)}\,.$$

Therefore, if E and $f^{-1}(E)$ have the same α-capacity, then

$$I_\alpha(\nu) = I_\alpha\,(\mu_e)\,,$$

and this is possible only if for all $r \in (0,1)$,

$$\int_0^{2\pi} \left|P_{\mu_e}\left(re^{i\theta}\right)\right|^2 d\theta = \int_0^{2\pi} \left|P_{\mu_e}\left(f\left(re^{i\theta}\right)\right)\right|^2 d\theta .$$

This can occur only if either f is a rotation or $|P_{\mu_e}|^2$ is harmonic. In the latter case, we obtain that μ_e is normalized Lebesgue measure, or equivalently that $\mathrm{cap}_\alpha(E) = \mathrm{cap}_\alpha(\partial\Delta)$. Since E is closed, it follows that $E = \partial\Delta$.

In order to prove the general case we need a characterization of the α-capacity of E when E is not closed (see Lemma 6 below). We begin by recalling some facts about convergence of measures.

We will say that a sequence of signed measures $\{\sigma_n\}$ with supports contained in a compact set K converges w^* to a signed measure σ if

$$\int h(x)\, d\sigma_n(x) \underset{n\to\infty}{\longrightarrow} \int h(x)\, d\sigma(x) , \qquad \text{for all } \ h \in C(K) .$$

Here, the w^*-convergence refers to the duality between the space of signed measures on K and the space $C(K)$ of continuous functions with support contained in K.

In this Section, we will denote by $\mathcal{M}_\alpha(K)$ $(0 \le \alpha < 1)$ the vector space of all signed measures whose support is contained in the set K and whose α-energy is finite. $\mathcal{M}_\alpha(\mathbb{C})$ or $\mathcal{M}_\alpha(\overline{\Delta})$ is denoted simply by $\mathcal{M}_\alpha$, and $\mathcal{M}_\alpha^+$ denotes the corresponding cone of positive measures.

The positivity properties of I_α [**L**, p. 79-80] allow us to define an inner product in $\mathcal{M}_\alpha$ (for $0 < \alpha < 1$) and e.g. in $\mathcal{M}_0(\{|z| = 1/2\})$ (for $\alpha = 0$) as follows

$$\langle \sigma, \gamma \rangle = \iint \Phi_\alpha(|x-y|)\, d\sigma(x) d\gamma(y) .$$

Observe that the associated norm verifies

$$\|\sigma\|^2 = I_\alpha(\sigma) .$$

In the next lemma we collect some useful information concerning the above inner product.

Lemma 6.

(i) *If $0 < \alpha < 1$, K is a compact subset of $\mathbb{C}$, $\{\sigma_n\}$ is a Cauchy sequence (with respect to the inner product) in $\mathcal{M}_\alpha^+(K)$ and $\sigma_n \overset{w^*}{\longrightarrow} \sigma$, then*

$$\|\sigma_n - \sigma\| \longrightarrow 0 , \qquad \text{as } \ n \to \infty .$$

(ii) *If E is any Borel subset of K, then*

$$\frac{1}{\mathrm{cap}_\alpha(E)} = \inf\left\{ I_\alpha(\mu) : \ \ \mu \ \text{ a probability measure, } \ \mu(E) = 1 \right\} ,$$

and there exists a probability measure μ_e supported on $\overline{E}$ such that

$$\frac{1}{\operatorname{cap}_\alpha(E)} = I_\alpha(\mu_e)\,.$$

In fact, if K_n is an increasing sequence of compact subsets of E such that

$$\operatorname{cap}_\alpha(K_n) \nearrow \operatorname{cap}_\alpha(E)\,,$$

and if μ_n is the equilibrium distribution of K_n, then

$$\mu_n \xrightarrow{w^*} \mu_e \qquad \textit{and} \qquad \|\mu_n - \mu_e\| \longrightarrow 0\,,$$

as $n \to \infty$.

These statements remain true in the case $\alpha = 0$, if K is a compact subset of Δ.

Lemma 6 is contained in [**L**, p. 82, 89, 145] if $0 < \alpha < 1$. The case $\alpha = 0$ is similar, though we need the restriction $K \subset \Delta$ so that $\|\cdot\|$ is a norm [**L**, p. 80].

Now we are ready to finish the proof of Theorem 1. Let E be a Borel subset of $\partial\Delta$ such that

$$\operatorname{cap}_\alpha\left(f^{-1}(E)\right) = \operatorname{cap}_\alpha(E) > 0\,. \tag{6}$$

We choose an increasing sequence of compact sets $K_n \subset E$ such that $\operatorname{cap}_\alpha(K_n) \nearrow \operatorname{cap}_\alpha(E)$. Let μ_n be the α-equilibrium measure of K_n and let μ_e be the probability measure supported on $\overline{E}$ given by Lemma 6. We have

$$\mu_n \xrightarrow{w^*} \mu_e \qquad \text{and} \qquad I_\alpha(\mu_n) \searrow I_\alpha(\mu_e)\,,$$

as $n \to \infty$. In fact,

$$\|\mu_n - \mu_e\| \to 0\,, \qquad \text{as} \quad n \to \infty\,.$$

Let ν_n be the probability measure, with $\nu_n(f^{-1}(K_n)) = 1$, such that $P_{\nu_n} = P_{\mu_n} \circ f$ (see Lemma 3). We can suppose after extracting a subsequence if necessary, that ν_n converges w^* to a probability measure ν on $\overline{f^{-1}(E)}$. Since the Poisson kernel is continuous in Δ we obtain, by using the w^*-convergence, that

$$P_{\mu_n} \to P_{\mu_e} \quad \text{and} \quad P_{\nu_n} \to P_\nu\,, \qquad \text{as} \quad n \to \infty\,,$$

pointwise. Therefore $P_\nu = P_{\mu_e} \circ f$, which in particular shows that ν is a probability measure supported on $f^{-1}(\overline{E})$.

Claim. $I_\alpha(\nu_n) \to I_\alpha(\nu)$ *as* $n \to \infty$.

Since ν_n *is a probability measure on* $f^{-1}(E)$, *Lemma* 6 *guarantees that*

$$\frac{1}{\operatorname{cap}_\alpha(f^{-1}(E))} \leq I_\alpha(\nu_n)\,,$$

and so, by letting $n \to \infty$, *and using that* $P_\nu = P_{\mu_e} \circ f$ (*by Lemma* 5) *we obtain that*

$$\frac{1}{\operatorname{cap}_\alpha(f^{-1}(E))} \leq I_\alpha(\nu) \leq I_\alpha(\mu_e) = \frac{1}{\operatorname{cap}_\alpha(E)}\,.$$

From (6), *we deduce that* $I_\alpha(\nu) = I_\alpha(\mu_e)$. *Finally, we can reason as in the case of* E *being closed and conclude that either* f *is a rotation or* μ_e *is normalized Lebesgue measure, i.e.,* $\operatorname{cap}_\alpha(E) = \operatorname{cap}_\alpha(\partial\Delta)$.

Proof of the Claim. Consider first the case $0 < \alpha < 1$. Since $P_{\nu_p - \nu_n} = P_{\mu_p - \mu_n} \circ f$, by Lemma 5 we obtain that

$$\|\nu_p - \nu_n\|^2 = I_\alpha(\nu_p - \nu_n) \leq I_\alpha(\mu_p - \mu_n) = \|\mu_p - \mu_n\|^2 \underset{p,n\to\infty}{\longrightarrow} 0\,.$$

Therefore $\{\nu_n\}$ is a Cauchy sequence in the norm and so, by Lemma 6, we have that

$$\|\nu_n - \nu\| \to 0 \qquad \text{and} \qquad I_\alpha(\nu_n) \to I_\alpha(\nu)$$

as $n \to \infty$.

For $\lambda > 0$, and $A \subset \mathbb{C}$, we will denote by λA the set $\lambda A = \{\lambda z : z \in A\}$.

If E is a Borel subset of $\partial\Delta$, then $\frac{1}{2}E$ is a Borel subset of $\{|z| = 1/2\}$. Also, if σ is a probability measure in $\partial\Delta$, we will denote by σ^* the probability measure in $\{|z| = 1/2\}$ defined by

$$\sigma(A) = \sigma^*\left(\frac{1}{2}A\right)\,, \tag{7}$$

for A a Borel subset of $\partial\Delta$. It is clear that

$$I_0(\sigma^*) = I_0(\sigma) + \log 2\,. \tag{8}$$

Now, in order to prove the case $\alpha = 0$, let μ_n^* and ν_n^* be the measures defined from μ_n and ν_n by (7). Then using again Lemma 5 and (8) we have that

$$\begin{aligned}\|\nu_p^* - \nu_n^*\|^2 &= I_0(\nu_p^* - \nu_n^*) = I_0(\nu_p - \nu_n) + \log 2 \\ &\leq I_0(\mu_p - \mu_n) + \log 2 = \left\|\mu_p^* - \mu_n^*\right\|^2 \underset{p,n\to\infty}{\longrightarrow} 0\,.\end{aligned}$$

Therefore $\{\nu_n^*\}$ is a Cauchy sequence in the norm and again by Lemma 6, we obtain that

$$\|\nu_n^* - \nu^*\| \to 0 \qquad \text{and} \qquad I_0(\nu_n^*) \to I_0(\nu^*)$$

as $n \to \infty$. It follows, from (8) that

$$I_0(\nu_n) \to I_0(\nu)\,, \qquad \text{as} \quad n \to \infty\,.$$

□

4. Some further results on distortion of capacity in the case n = 1.

First we show that Theorem 1 is sharp. In what follows $|\cdot|$ will denote not normalized Lebesgue measure in $\partial\Delta$ (i.e. $|\partial\Delta| = 2\pi$).

Proposition 1. $\mathrm{cap}_\alpha\,(f^{-1}(E))$ *can take any value between* $\mathrm{cap}_\alpha(E)$ *and* $\mathrm{cap}_\alpha(\partial\Delta)$. *More precisely, given* $0 < s \le t < \mathrm{cap}_\alpha(\partial\Delta)$ *there exist a Borel subset* E *of* $\partial\Delta$ *and an inner function* f *with* $f(0) = 0$ *such that* $\mathrm{cap}_\alpha(E) = s$ *and* $\mathrm{cap}_\alpha\,(f^{-1}(E)) = t$.

In order to prove this, we need the following lemma whose proof will given later.

Lemma 7. *Let* I *be any closed interval in* $\partial\Delta$ *with* $|I| > 0$, *and let* B *be a finite union of closed intervals in* $\partial\Delta$ *such that* $|B| = |I|$. *Then there exists an inner function* f *such that*

$$f(0) = 0 \qquad \textit{and} \qquad f^{-1}(I) = B\,.$$

In fact, if $0 < |I| < 2\pi$, *then* f *is unique.*

Remark. It is natural to wonder if this lemma holds in higher dimensions, more precisely: Is it true that given an interval I in $\partial\Delta$ and a Borel subset B of $\mathbb{S}_n$ such that

$$\frac{|B|}{\omega_{2n-1}} = \frac{|I|}{2\pi}\,,$$

there is an inner function $f : \mathbb{B}_n \longrightarrow \Delta$ such that $f^{-1}(I) \stackrel{\circ}{=} B$?

It is not possible to construct such f by using the Ryll-Wojtaszczyk polynomials (see [**R1**]), since in that case the following stronger result would be true too:

Given E, I subsets of $\partial\Delta$ with $|E| = |I|$ and $N \in \mathbb{N}$, there exists an inner function $f : \Delta \longrightarrow \Delta$ such that

$$E = f^{-1}(I), \qquad \text{and} \qquad f^{(j)}(0) = 0, \quad \text{if } j \leq N.$$

But it is easy to see, as a consequence of Lemma 8, that in general this is not possible.

Proof of Proposition 1. Let I be a closed interval in $\partial\Delta$ centered at 1 and such that $\text{cap}_\alpha(I) = s$. Consider the function $g(z) = z^2$. Then (see e.g. [**FP**] or Proposition 3 below),

$$s = \text{cap}_\alpha(I) < \text{cap}_\alpha\left(g^{-1}(I)\right) < \cdots < \text{cap}_\alpha\left(g^{-k}(I)\right) \underset{k\to\infty}{\longrightarrow} \text{cap}_\alpha(\partial\Delta).$$

Therefore, if $t = \text{cap}_\alpha\left(g^{-k}(I)\right)$ for some k, we are done.

Note that $g^{-k}(I)$ consists of 2^k closed intervals of length $|I|/2^k$ and centered at the points $z_{j,k} = e^{2\pi j i/2^k}$ $(j = 1, \ldots, 2^k)$.

If $\text{cap}_\alpha\left(g^{-(k-1)}(I)\right) < t < \text{cap}_\alpha\left(g^{-k}(I)\right)$ a simple continuity argument shows that there exist a finite union B of 2^k closed intervals in $\partial\Delta$ of total length $|I|$ with $\text{cap}_\alpha(B) = t$.

Finally, applying Lemma 7 to the pair I, B we obtain an inner function f with $f(0) = 0$ and $f^{-1}(I) = B$. □

Proof of Lemma 7. Let u be the Poisson integral of the characteristic function of B, and let $\tilde{u}$ be its conjugate harmonic function chosen such that $\tilde{u}(0) = 0$. Since $u(0) = |B|/2\pi$ the holomorphic function $F = u + i\tilde{u}$ transforms Δ into the strip $S = \{\omega : 0 < \text{Re}\,\omega < 1\}$. Notice that F has radial boundary values except for a finite number of points, and F applies the interior of B into $\{\omega : \text{Re}\,\omega = 1\}$ and $\partial\Delta \setminus B$ into $\{\omega : \text{Re}\,\omega = 0\}$.

Now, let G be the Riemann mapping of S chosen such that

$$G(|B|/2\pi) = 0.$$

G transforms $\{\omega : \text{Re}\,\omega = 1\}$ onto an interval J of $\partial\Delta$. On the other hand, the function $h = G \circ F$ is clearly an inner function, $h(0) = 0$ and $h^{-1}(I) = B$. By composing h with an appropriate rotation we finish the proof of the existence statement.

To show the uniqueness of f, it is sufficient to prove the following

Lemma 8. *If A is any Borel subset of $\partial\Delta$, such that $\int_A e^{-i\theta} d\theta \neq 0$, and f, g are inner functions with $f(0) = g(0) = 0$ such that*

$$f^{-1}(A) \overset{\circ}{=} g^{-1}(A),$$

then $f \equiv g$.

Here $\overset{\circ}{=}$ *denotes equality up to a set of zero Lebesgue measure.*

Proof. Let $F : \Delta \longrightarrow \{\omega : 0 < \operatorname{Re}\omega < 1\}$ be the holomorphic function given by

$$F(z) = \frac{1}{2\pi} \int_A \frac{e^{i\theta} + z}{e^{i\theta} - z}\, d\theta .$$

F is univalent in a neighbourhood of 0, because

$$F'(0) = \frac{1}{\pi} \int_A e^{-i\theta} d\theta \neq 0 .$$

Now, observe that $\operatorname{Re}(F \circ f) = \operatorname{Re}(F \circ g)$ almost everywhere on $\partial\Delta$. Since $\operatorname{Re}(F \circ f)$ and $\operatorname{Re}(F \circ g)$ are bounded harmonic functions it follows that $F \circ f = F \circ g + ic$ in Δ, where c is a real constant. Since $f(0) = g(0)$, we deduce that $F \circ f = F \circ g$ which proves the lemma because F is univalent in a neighbourhood of 0. □

Observe that, in particular, the condition $\int_A e^{-i\theta} d\theta \neq 0$ is satisfied e.g. if A is any interval in $\partial\Delta$ with $0 < |A| < 2\pi$.

The condition $\int_A e^{-i\theta} d\theta \neq 0$ is not only a technicality. If A is k-symmetrical (i.e., there exists a subset $A_0 \subset A$, with $A_0 \subset [0, 2\pi/k]$, such that $A \overset{\circ}{=} A_0 \cup (A_0 + 2\pi/k) \cup (A_0 + 4\pi/k) \cup \cdots \cup (A_0 + 2\pi(k-1)/k)$), and $\int_A e^{-ik\theta} d\theta \neq 0$, then $f = \omega g$, where ω is a k-th root of unity. To see this, one can use Lemma 8 with the functions $h \circ f$, $h \circ g$ and the set $h(A)$, where $h(z) = z^k$.

Also, note that if A is the union of two intervals in $\partial\Delta$, then $f = \pm g$, because $\int_A e^{-i\theta} d\theta = 0$ implies that A is 2-symmetrical.

Notice that if the function g in Lemma 8 were the identity, and $0 < |A| < 2\pi$, then, by ergodicity, we would have that f is a rotation of rational angle. This, together with the above remark, could suggest that perhaps the following statement was true:

If A is any Borel subset of $\partial\Delta$, such that $0 < |A| < 2\pi$, and f, g are inner functions with $f(0) = g(0) = 0$ such that

$$f^{-1}(A) \overset{\circ}{=} g^{-1}(A) ,$$

then $f \equiv \lambda g$ with $|\lambda| = 1$.

But this is false as the next example shows: Let B be the following Blaschke product

$$B(z) = z\, \frac{2z - 1}{2 - z} .$$

By applying a theorem of Stephenson [**S**, Theorem 3] to the pair $B, -B$, one obtains two inner functions f and g with $f(0) = g(0) = 0$, such that

$$B \circ f = -B \circ g .$$

But then $(B(f))^2 = (B(g))^2$, and so, if we had $f = \lambda g$, we could conclude that $B(z) = -B(\lambda z)$. But, since $B'(0) \neq 0$, we had $\lambda = -1$, i.e., $B(z) = -B(-z)$, a contradiction.

The following is well known, at least for $\alpha = 0$, see for instance [**A**, p. 35-36] where it is credited to Beurling.

Proposition 2. *Let $0 \leq \alpha < 1$. If I is any interval in $\partial\Delta$, then I has the minimum α-capacity between all the Borel subsets of $\partial\Delta$ with the same Lebesgue measure than I.*

Proof. Let E be a Borel set such that $|E| = |I|$. A standard approximation argument shows that for all $\varepsilon > 0$ there exists a finite union B_ε of closed intervals such that

$$||E| - |B_\varepsilon|| < \varepsilon \qquad \text{and} \qquad |\operatorname{cap}_\alpha(E) - \operatorname{cap}_\alpha(B_\varepsilon)| < \varepsilon\,.$$

Let I_ε be a closed interval with the same center than I and such that $|I_\varepsilon| = |B_\varepsilon|$. By Lemma 7, we can find an inner function f_ε such that

$$f_\varepsilon(0) = 0 \qquad \text{and} \qquad f_\varepsilon^{-1}(I_\varepsilon) = B_\varepsilon\,.$$

Therefore, by Theorem 1,

$$\operatorname{cap}_\alpha(E) + \varepsilon \geq \operatorname{cap}_\alpha(B_\varepsilon) \geq \operatorname{cap}_\alpha(I_\varepsilon)\,,$$

but $\operatorname{cap}_\alpha(I_\varepsilon) \to \operatorname{cap}_\alpha(I)$ as $\varepsilon \to 0$. □

The following proposition is not unexpected since ergodic theory says that $f^{-k}(E)$ is well spread on $\partial\Delta$. Hereafter $f^k = f \circ \cdots \circ f$ denotes the k-iterate of f and $f^{-k} = (f^k)^{-1}$.

Proposition 3. *If $f : \Delta \longrightarrow \Delta$ is inner but not a rotation, $f(0) = 0$, $0 \leq \alpha < 1$ and E is a Borel subset of $\partial\Delta$ with $\operatorname{cap}_\alpha(E) > 0$, then*

$$\operatorname{cap}_\alpha\left(f^{-k}(E)\right) \to \operatorname{cap}_\alpha(\partial\Delta) \qquad \text{as} \quad k \to \infty\,.$$

The proof of this result is an easy consequence of the following lemma.

Lemma 9. *With the hypotheses of Proposition 3, if μ is any probability measure on E with finite α-energy and if ν_k is the probability measure in $f^{-k}(E)$ such that $P_{\nu_k} = P_\mu \circ f^k$, then*

$$I_\alpha(\nu_k) \longrightarrow I_\alpha\left(\frac{|\cdot|}{2\pi}\right) \qquad \text{as} \quad k \to \infty\,.$$

With this, we have

$$\frac{1}{\mathrm{cap}_\alpha\left(f^{-k}(E)\right)} \leq I_\alpha(\nu_k) \longrightarrow I_\alpha\left(\frac{|\cdot|}{2\pi}\right) = \frac{1}{\mathrm{cap}_\alpha(\partial\Delta)}$$

giving us the conclusion of Proposition 3.

Proof of Lemma 9. We will prove it for $0 < \alpha < 1$; the case $\alpha = 0$ being similar.

By Lemma 2 (i), we have with an appropriate function g_α that

$$I_\alpha(\sigma) = \int_0^1 \int_0^{2\pi} \left|P_\sigma(re^{i\theta})\right|^2 \, d\theta \, g_\alpha(r) \, dr$$

for any probability measure σ on $\partial\Delta$.

Using (3) we have for all $r \in (0,1)$ that

$$\int_0^{2\pi} \left|P_{\nu_k}(re^{i\theta})\right|^2 d\theta \leq \int_0^{2\pi} \left|P_\mu(re^{i\theta})\right|^2 d\theta \,.$$

Since μ has finite α-energy, the right hand side in the last inequality, as a function of r, belongs to $L^1(g_\alpha(r)\,dr)$. Therefore, by using the Lebesgue's dominated convergence theorem, we would be done if we show that

$$\int_0^{2\pi} \left|P_{\nu_k}(re^{i\theta})\right|^2 d\theta \longrightarrow \frac{1}{2\pi} \qquad \text{as} \quad k \to \infty \,, \tag{9}$$

for each r with $0 < r < 1$. But, by Schwarz's lemma, and since f is not a rotation, $\left|f^k\left(re^{i\theta}\right)\right| \longrightarrow 0$ as $k \to \infty$, uniformly on θ for r fixed. Therefore, for each r, $P_{\nu_k}\left(re^{i\theta}\right) = P_\mu\left(f^k\left(re^{i\theta}\right)\right) \longrightarrow 1/2\pi$, as $k \to \infty$, uniformly on θ, and this implies (9). □

Even in the case when $\mathrm{cap}_\alpha(E) = 0$, the sets $f^{-k}(E)$ are well spread on $\partial\Delta$.

Proposition 4. *If $f : \Delta \longrightarrow \Delta$ is an inner function (but not a rotation) with $f(0) = 0$, E is any non empty Borel subset of $\partial\Delta$, and μ is any probability measure on E, then for some absolute constant C and a positive constant A that only depends on $|f'(0)|$, we have that*

$$\left|\nu_k(I) - \frac{|I|}{2\pi}\right| < C\, e^{-Ak} \,,$$

for each interval $I \subset \partial\Delta$. In particular,

$$\nu_k \longrightarrow \frac{|\cdot|}{2\pi}$$

in the usual weak-$*$ topology.

Here ν_k is the probability measure concentrated in $f^{-k}(E)$ such that $P_{\nu_k} = P_\mu \circ f^k$.

Proof. The proof is similar to that of Lemma 3 in [**P**], but using here the fact that $P_{\nu_k} = P_\mu \circ f^k$ instead of Lemma 1 in [**P**]. □

Proposition 5. *If $f : \mathbb{B}_n \longrightarrow \Delta$ is inner, then f assumes in $\partial\mathbb{B}_n$ all the values in $\partial\Delta$.*

Proof. Let $f : \mathbb{B}_n \longrightarrow \Delta$ be an inner function. It is enough to prove that $f^{-1}\{1\} \neq \varnothing$. But,

$$u := \operatorname{Re}\left(\frac{1+f}{1-f}\right) = \frac{1-|f|^2}{|1-f|^2} > 0, \qquad \text{in } \mathbb{B}_n\,. \tag{10}$$

Therefore, u is harmonic and positive in $\mathbb{B}_n$ and so there exists a positive measure in $\mathbb{S}_n$ such that

$$\operatorname{Re}\left(\frac{1+f}{1-f}\right) = P_\mu\,.$$

By (10) P_μ tends radially to 0 a.e. with respect to Lebesgue measure, since f is inner and (by Privalov's theorem, (see e.g., [**R**, Theorem 5.5.9])) f can assume the value 1 at most in a set of zero Lebesgue measure. Then, the Radon-Nikodym derivative of μ with respect to Lebesgue measure is zero a.e., and so μ is a singular measure.

By Lemma 11 it follows that $P_\mu \to +\infty$ in a set of full μ-measure. But this is the same to say that $f(re^{i\theta}) \to 1$ in that set. □

When the inner function f has order $k \geq 1$ at 0, we can improve Theorem 1 in the case α=0.

Theorem 3. *If $f : \Delta \longrightarrow \Delta$ is inner,*

$$f(0) = f'(0) = \cdots = f^{(k-1)}(0) = 0\,, \qquad f^{(k)}(0) \neq 0\,, \qquad (k \geq 1)\,,$$

and E is a Borel subset of $\partial\Delta$, then

$$\operatorname{cap}_0\left(f^{-1}(E)\right) \geq \left(\operatorname{cap}_0(E)\right)^{1/k}\,. \tag{11}$$

Moreover, if $\operatorname{cap}_0(E) > 0$, equality holds if and only if either $f(z) = \lambda z^k$, with $|\lambda| = 1$, or $\operatorname{cap}_0(E) = \operatorname{cap}_0(\partial\Delta)$.

Proof. For such a function f, Schwarz's lemma says us that $|f(z)| \leq |z|^k$, with equality only if $f(z) = \lambda z^k$ with $|\lambda| = 1$. With this in mind, the

subordination principle says now (see e.g. [**HH**]) that if v is a subharmonic function in Δ, then

$$\int_0^{2\pi} v\left(f\left(re^{i\theta}\right)\right)\, d\theta \leq \int_0^{2\pi} v\left(r^k e^{i\theta}\right)\, d\theta\,,$$

with equality for a given r only if v is harmonic in $\{|z| < r\}$ or f is a rotation of z^k.

Now, in order to prove (11), we can assume that E is closed. If μ_e is the equilibrium probability distribution of E and ν is the probability measure in $f^{-1}(E)$ such that $P_\nu = P_\mu \circ f$, then

$$\begin{aligned} I_0(\nu) &= 2\pi \int_0^1 \int_0^{2\pi} \left| P_{\mu_e}\left(f\left(re^{i\theta}\right)\right) - \frac{1}{2\pi}\right|^2 d\theta\, \frac{dr}{r} \\ &\leq 2\pi \int_0^1 \int_0^{2\pi} \left| P_{\mu_e}\left(r^k e^{i\theta}\right) - \frac{1}{2\pi}\right|^2 d\theta\, \frac{dr}{r}\,. \end{aligned}$$

Substituting $r^k = t$, we obtain that

$$I_0(\nu) \leq \frac{1}{k}\, I_0(\mu_e)\,.$$

This finishes the proof of (11). The equality statement can be proved in the same way as that of Theorem 1. □

Remark. For other α's $(0 < \alpha < 1)$ we can show

$$\frac{1}{\mathrm{cap}_\alpha\left(f^{-1}(E)\right)} - \frac{1}{\mathrm{cap}_\alpha(\partial\Delta)} \leq \frac{C_\alpha}{k^{1-\alpha}} \left(\frac{1}{\mathrm{cap}_\alpha(E)} - \frac{1}{\mathrm{cap}_\alpha(\partial\Delta)}\right)$$

where C_α is a constant depending only on α.

We expect $C_\alpha = 1$, but we have not been able to show this.

5. Distortion of α-content.

The following is an extension of Löwner's lemma.

Theorem 4. *If $f : \mathbb{B}_n \longrightarrow \Delta$ is inner, $f(0) = 0$ and E is a Borel subset of $\partial\Delta$, then, for $0 < \alpha \leq 1$,*

(i) $$M_{2n-2+\alpha}\left(f^{-1}(E)\right) \geq C_{n,\alpha}\, M_\alpha(E)$$

and

(ii) $$\mathcal{M}_{2(n-1+\alpha)}\left(f^{-1}(E)\right) \geq C'_{n,\alpha}\, M_\alpha(E)\,.$$

Here M_β and $\mathcal{M}_\beta$ denote, respectively, β-dimensional content with respect to the euclidean metric and with respect to the metric in $\mathbb{S}_n$ given by

$$d(a,b) = |1 - \langle a,b\rangle|^{1/2},$$

where $\langle a,b\rangle = \sum a_j \bar{b}_j$ is the inner product in $\mathbb{C}^n$. This metric is equivalent to the Carnot-Carathèodory metric in the Heisenberg group model for $\mathbb{S}_n$. We refer to [**R**] for details about this metric.

Recall that in a general metric space (X,d) the α-content of a set $E \subset X$ is defined as

$$M_\alpha(E) = \inf\left\{\sum_i r_i^\alpha : \ E \subset \bigcup_i B_d(x_i, r_i)\right\}.$$

Observe that, as a consequence of Theorem 4, one obtains

Corollary. *If $f : \mathbb{B}_n \longrightarrow \Delta$ is inner and E is a Borel subset of $\partial\Delta$, then*

$$\mathrm{Dim}\,(f^{-1}(E)) \geq 2n - 2 + \mathrm{Dim}(E)$$

and

$$\mathcal{D}im\,(f^{-1}(E)) \geq 2n - 2 + 2\,\mathrm{Dim}(E)$$

where Dim *and* $\mathcal{D}im$ *denote, respectively, Hausdorff dimension with respect to the euclidean metric and the metric d.*

In order to prove Theorem 4 we will prove a lemma about Poisson integrals. We need to consider the classical Poisson kernel (not normalized)

$$P(\xi,z) = \frac{1-|z|^2}{|\xi - z|^{2n}} \qquad (z \in \mathbb{B}_n\,,\ \xi \in \mathbb{S}_n)\,,$$

and the invariant Poisson kernel

$$Q(\xi,z) = \frac{(1-|z|^2)^n}{|1-\langle z,\xi\rangle|^{2n}} \qquad (z \in \mathbb{B}_n\,,\ \xi \in \mathbb{S}_n)\,.$$

Of course, they coincide if $n = 1$. In this section if ν is a positive measure in $\mathbb{S}_n$, we will denote by P_ν the function

$$P_\nu(z) = \int_{\mathbb{S}_n} P(\xi,z)\, d\nu(\xi)$$

and by Q_ν the invariant Poisson extension of ν

$$Q_\nu(z) = \int_{\mathbb{S}_n} Q(\xi,z)\, d\nu(\xi)\,.$$

Lemma 10. *Let μ be a finite positive measure in $\partial\Delta$, and let $f : \mathbb{B}_n \longrightarrow \Delta$ be an inner function. Then, there exists a finite measure $\nu \geq 0$ in $\mathbb{S}_n$ such that $P_\mu \circ f = P_\nu$, and if ν has singular part σ and continuous part γ, and we denote by A the set*

$$A = \{\xi \in \mathbb{S}_n : \ P_\sigma(r\xi) \to +\infty, \ \text{as } r \to 1\}$$

and by B the set

$$B = \left\{\xi \in \mathbb{S}_n : \ \exists \lim_{r\to 1} f(r\xi) = f(\xi), \ |f(\xi)| = 1 \ \text{ and } \ \lim_{r\to 1} P_\gamma(r\xi) > 0\right\},$$

then A has full σ-measure, B has full γ-measure and

$$A \cup B \subset f^{-1}(\text{support}\,\mu)$$

and so

$$\nu\left(f^{-1}(\text{support}\,\mu)\right) = \|\nu\|.$$

The same is true if we replace P_ν by $Q_{\nu'}$ $(P_\mu \circ f = Q_{\nu'})$ and A, B by the following sets

$$A' = \{\xi \in \mathbb{S}_n : \ Q_{\sigma'}(r\xi) \to +\infty, \ \text{as } r \to 1\},$$

and

$$B' = \left\{\xi \in \mathbb{S}_n : \ \exists \lim_{r\to 1} f(r\xi) = f(\xi), \ |f(\xi)| = 1 \ \text{ and } \ \lim_{r\to 1} Q_{\gamma'}(r\xi) > 0\right\},$$

where σ' and γ' denote, respectively, the singular and the continuous part of ν'.

Proof. We will prove the lemma only for the measure ν', since the proof of the result for ν is similar and standard.

Let $U : \Delta \longrightarrow \mathbb{C}$ be a holomorphic function such that $\text{Re}\,U = P_\mu$. Then $U \circ f$ is also holomorphic and so $\text{Re}\,(U \circ f) = P_\mu \circ f$ is pluriharmonic, i.e. harmonic and $\mathcal{M}$-harmonic (see e.g. [**R**, Theorem 4.4.9]). Therefore there exist finite positive measures ν and ν' in $\mathbb{S}_n$ such that

$$P_\mu \circ f = P_\nu, \qquad P_\mu \circ f = Q_{\nu'}.$$

Let us denote by E the support of μ. If $\xi \in A'$, then $|f(r\xi)| \to 1$ as $r \to 1$. The curve $\{f(r\xi) : \ 0 \leq r < 1\}$ in Δ must end on a unique point $e^{i\psi} = f(\xi) \in \Delta$, since otherwise we would have $P_\mu \equiv +\infty$ on a set of positive Lebesgue measure. Now, $e^{i\psi} \in E$, since otherwise P_μ vanishes continuously at $e^{i\psi}$. Therefore $A' \subset f^{-1}(E)$. Similarly one sees that $B' \subset f^{-1}(E)$.

The set A' has full σ'-measure since by the inequality (14), that we will prove later,

$$\left\{\xi \in \mathbb{S}_n : \underline{D}\,\sigma'(\xi) = \infty\right\} \subset A',$$

where

$$\underline{D}\,\sigma'(\xi) = \liminf_{r\to 0} \frac{\sigma'(B_d(\xi,r))}{|B_d(\xi,r)|},$$

and the set $\{\xi : \underline{D}\,\sigma'(\xi) = \infty\}$ has full σ'-measure (see Lemma 11 below). Let us observe that ([**R**, p. 67])

$$|B_d(\xi,r)| \sim r^{2n}.$$

The set B' has full γ'-measure, since as $r \to 1$

$$Q_{\gamma'}(r\xi) \longrightarrow \frac{d\gamma'}{dL} \qquad \text{a.e.}$$

with respect to Lebesgue measure L (see, e.g., [**R**, Theorem 5.4.9]) and $\left\{\frac{d\gamma'}{dL} > 0\right\}$ has full γ'-measure. □

Lemma 11. *Suppose that μ is a singular positive Borel measure (with respect to Lebesgue measure) in $\mathbb{S}_n$. Then*

$$\underline{D}\,\mu(x) = \infty \qquad a.e.\ \ \mu.$$

Proof. Let $\mathcal{A}$ be a Borel set such that $|\mathcal{A}| = 0$, and μ is concentrated on $\mathcal{A}$. Define for $\alpha > 0$

$$\mathcal{A}_\alpha = \left\{x \in \mathcal{A} : \underline{D}\,\mu(x) < \alpha\right\}.$$

It is enough to prove that $\mu(\mathcal{A}_\alpha) = 0$, and by regularity that $\mu(K) = 0$ for all K compact subset of $\mathcal{A}_\alpha$.

Fix $\varepsilon > 0$. Since $K \subset \mathcal{A}_\alpha \subset \mathcal{A}$, $|K| = 0$ and so there exists an open set V with $K \subset V$ and $|V| < \varepsilon$ ($|\cdot|$ denotes Lebesgue measure).

Now, for each $x \in K$, we can find $r_x > 0$ such that

$$\frac{\mu(B_d(x,r_x))}{|B_d(x,r_x)|} < \alpha \qquad \text{and} \qquad B_d(x,r_x/3) \subset V.$$

The family $\{B_d(x,r_x/3) : x \in K\}$ covers K, hence we can extract a finite subcollection Φ that also covers K. Now, using a Vitaly-type lemma (see, e.g., [**R**, Lemma 5.2.3]), we can find a disjoint subcollection Γ of Φ such that

$$K \subset \bigcup_{\Gamma} B_d(x_i, r_{x_i}).$$

Note that as a consequence of Proposition 5.1.4 in [**R**] we have that

$$\Theta_d := \sup_{\delta} \frac{|B_d(x, r_x)|}{|B_d(x, r_x/3)|} < \infty .$$

Therefore

$$\begin{aligned} \mu(K) &\leq \sum_{\Gamma} \mu\left(B_d(x_i, r_{x_i})\right) < \alpha \sum_{\Gamma} |B_d(x_i, r_{x_i})| \\ &< \Theta_d \, \alpha \sum_{\Gamma} |B_d(x_i, r_{x_i}/3)| \leq \Theta_d \, \alpha \, |V| < \Theta_d \, \alpha \, \varepsilon . \end{aligned}$$

□

Proof of Theorem 4. We will prove only (ii), since (i) is obtained in a similar way.

Assume, as we may, that E is a closed subset of $\partial\Delta$ and $M_\alpha(E) > 0$. Then, see e.g. [**T**, p. 64], there exists a positive mass distribution on E of finite total mass, such that: (a) $\mu(E) = M_\alpha(E)$, (b) $\mu(I) \leq C_\alpha |I|^\alpha$ for any open interval I, where C_α is a constant independent of E. A standard estimate shows that

$$P_\mu(z) \leq \frac{C_\alpha}{(1 - |z|)^{1-\alpha}} , \qquad (z \in \Delta) , \tag{12}$$

with C_α a new constant. Let $\nu' \geq 0$ be a measure in $\mathbb{S}_n$ such that $P_\mu \circ f = Q_{\nu'}$. Schwarz's lemma (see e.g. [**R**, Theorem 8.1.2]) and (12) give the corresponding inequality for ν':

$$Q_{\nu'}(z) \leq \frac{C_\alpha}{(1 - \|z\|)^{1-\alpha}} , \qquad (z \in \mathbb{B}_n) . \tag{13}$$

We claim that for each $z \in \mathbb{B}_n$

$$Q_{\nu'}(z) \geq C_n \frac{\nu'(B_d(\xi, (2(1 - \|z\|))^{1/2}))}{(1 - \|z\|)^n} , \qquad (z \in \mathbb{B}_n) , \tag{14}$$

where $\xi = z/\|z\|$ and $B_d(\xi, R)$ denotes the d-ball with center ξ and radius R.

Assuming (14) for the moment and using (13), we obtain that

$$\nu'(B_d(\xi, R)) \leq C_{n,\alpha} \, R^{2(n-1+\alpha)} , \qquad (\xi \in \mathbb{S}_n , R > 0) . \tag{15}$$

If we cover the set $A' \cup B'$ (see Lemma 14) with d-balls of radii R_i, we see by (15) that

$$\nu'(A' \cup B') \leq C_{n,\alpha} \sum_i R_i^{2(n-1+\alpha)}$$

and so

$$\begin{aligned}\|\nu'\| = \nu'(A' \cup B') &\leq C_{n,\alpha}\, \mathcal{M}_{2(n-1+\alpha)}(A' \cup B') \\ &\leq C_{n,\alpha}\, \mathcal{M}_{2(n-1+\alpha)}\left(f^{-1}(E)\right) .\end{aligned}$$

So, since $f(0) = 0$,

$$M_\alpha(E) = \|\mu\| = \|\nu'\| \leq C_{n,\alpha}\, \mathcal{M}_{2(n-1+\alpha)}\left(f^{-1}(E)\right) .$$

Therefore, in order to finish the proof, it remains only to prove (14). Observe first that we can assume that $\xi = e_1 = (1, 0, \ldots, 0)$ since d is invariant under the unitary transformations of $\mathbb{S}_n$ for the inner product $\langle \cdot, \cdot \rangle$. Now, if $z = re_1$, write $\delta^2 = 2(1-r)$. If $\eta \in B_d(e_1, \delta)$, then

$$|1 - r\eta_1| \leq |1 - \eta_1| + |\eta_1|(1-r) \leq 3(1-r) .$$

Hence, if $\eta \in B_d(e_1, \delta)$

$$Q(\eta, z) = \left(\frac{1-r^2}{|1-r\eta_1|^2}\right)^n \geq \frac{9^{-n}}{(1-r)^n} .$$

Since Q is invariant under the action of the unitary group for the inner product $\langle \cdot, \cdot \rangle$ in $\mathbb{S}_n$, we obtain that if $z = r\xi$ and $\eta \in B_d(\xi, \delta)$, then

$$Q(\eta, z) \geq \frac{9^{-n}}{(1-r)^n} .$$

Finally,

$$Q_{\nu'}(z) \geq \int_{B_d(\xi,\delta)} Q(\eta, z)\, d\nu'(\eta) \geq 9^{-n} \frac{\nu'(B_d(\xi, \delta))}{(1-r)^n} .$$

□

6. Distortion of subsets of the disc.

We have discussed how inner functions distort boundary sets. There are some results on how they distort subsets of Δ. On the one hand Hamilton [**H**] has shown that

Theorem H. *For all Borel subsets E of Δ,*

$$H_\alpha\left(f^{-1}(E)\right) \geq H_\alpha(E) , \qquad 0 < \alpha \leq 1 ,$$

where H_α denotes α-Hausdorff measure.

One naturally expects the following to be true:

If $f : \Delta \longrightarrow \Delta$ is inner, $f(0) = 0$ and E is a Borel subset of Δ, then

$$\mathrm{cap}_\alpha\left(f^{-1}(E)\right) \geq \mathrm{cap}_\alpha(E)\,.$$

This we can prove only if $\alpha = 0$. The idea comes from [**P1**, p. 336].

Theorem 5. *Let $f : \Delta \longrightarrow \Delta$ be an inner function. If for some $k \geq 1$*

$$f(0) = f'(0) = \cdots = f^{(k-1)}(0) = 0\,, \qquad f^{(k)}(0) \neq 0\,,$$

then,

$$\mathrm{cap}_0\left(f^{-1}(E)\right) \geq \left(\mathrm{cap}_0(E)\right)^{1/k}\,,$$

for all Borel subsets of Δ. Moreover, this inequality is sharp.

Sketch of proof. By approximation, it is enough to prove it if E is closed and f is a finite Blaschke product. Let f be

$$f(z) = z^k \prod_{j=1}^{d} e^{i\nu_j} \frac{z - a_j}{1 - \bar{a}_j z}\,.$$

Denote by g_E, g_F the Green's functions of the unbounded connected component of $\hat{\mathbb{C}} \setminus E$ and $\hat{\mathbb{C}} \setminus F$ (here $F = f^{-1}(E)$) with pole at ∞. Therefore,

$$g_E(z) - \log|z| = \log \frac{1}{\mathrm{cap}_0(E)} + O\left(|z|^{-1}\right)\,,$$

$$g_F(z) - \log|z| = \log \frac{1}{\mathrm{cap}_0(F)} + O\left(|z|^{-1}\right)\,,$$

as $|z| \to \infty$. Moreover, since $k \geq 1$

$$g_E(f(z)) - k\log|z| + \log \prod_{j=1}^{d} |a_j| = \log \frac{1}{\mathrm{cap}_0(E)} + O\left(|z|^{-1}\right)\,,$$

as $|z| \to \infty$. It is easy to see that

$$g_E(f(z)) - \sum_{j=1}^{d} g_F\left(z, \bar{a}_j^{-1}\right)$$

is harmonic in the unbounded connected component of $\mathbb{C} \setminus \left(F \cup \left(\cup_{j=1}^{d} \{\bar{a}_j^{-1}\}\right)\right)$ and it is bounded at the points $\bar{a}_j^{-1}$ (here $g_F(z, \bar{a}_j^{-1})$ denotes the Green's

function of the unbounded connected component of $\hat{\mathbb{C}} \setminus F$ with pole at $\bar{a}_j^{-1}$). Therefore, the function

$$G(z) = \frac{1}{k} g_E(f(z)) - g_F(z) - \frac{1}{k} \sum_{j=1}^{d} g_F\left(z, \bar{a}_j^{-1}\right) \tag{16}$$

is harmonic and bounded in the unbounded connected component of $\hat{\mathbb{C}} \setminus F$. Since $G = 0$ on the outer boundary of F, it follows that $G \equiv 0$.

Now, by using the symmetry of Green's function, we have that

$$g_F\left(z, \bar{a}_j^{-1}\right) \longrightarrow g_F\left(\bar{a}_j^{-1}\right), \qquad \text{as } |z| \to \infty,$$

and so, from (16),

$$\log \frac{1}{\operatorname{cap}_0(E)} - \log \prod_{j=1}^{d} |a_j| - k \log \frac{1}{\operatorname{cap}_0(F)} - \sum_{j=1}^{d} g_F\left(\bar{a}_j^{-1}\right) = 0. \tag{17}$$

On the other hand, since $F \subset \Delta$, the maximum principle says that

$$g_F(z) \geq g_\Delta(z) = \log |z|, \qquad |z| > 1.$$

Hence, from (17), we obtain that

$$\log \frac{1}{\operatorname{cap}_0(E)} - \log \prod_{j=1}^{d} |a_j| - k \log \frac{1}{\operatorname{cap}_0(F)} \geq \sum_{j=1}^{d} \log |a_j|^{-1},$$

and the inequality in the theorem follows.

Finally, to show that the inequality is sharp one simply has to consider the function $f(z) = z^k$. □

References

[A] L.V. Ahlfors, *Conformal invariants*, McGraw-Hill, 1973.

[AS] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical functions*, Dover Pub. Inc., nine-th printing, 1970.

[B] A. Beurling, *Ensembles exceptionnels*, Acta Math., **72** (1939), 1-13.

[C] L. Carleson, *Selected problems on exceptional sets*, Van Nostrand, 1967.

[FP] J.L. Fernández and D. Pestana, *Distortion of boundary sets under inner functions and applications*, Indiana Univ. Math. J., **41** (1992), 439-447.

[H] D. Hamilton, *Distortion of sets by inner functions*, Proc. Amer. Math. Soc., **117** (1993), 771-774.

[HH] R.R. Hall and W.K. Hayman, *A problem in the theory of subordination*, J. d'Analyse, **60** (1993), 99-111.

[KS] J.P. Kahane and R. Salem, *Ensembles parfaits et Sèries Trigonomètriques*, Hermann, 1963.

[L] N.S. Landkof, *Foundations of modern potential theory*, Springer-Verlag, 1972.

[N] J.H. Neuwirth, *Ergodicity of some mappings of the circle and the line*, Israel J. Math., **31** (1978), 359-367.

[P] Ch. Pommerenke, *Ergodic properties of inner functions*, Math. Ann., **256** (1981), 43-50.

[P1] ———, *Univalent functions*, Vandenhoek und Ruprecht, Göttingen, 1975.

[R] W. Rudin, *Function theory in the unit ball of* $\mathbb{C}^n$, Springer-Verlag, 1980.

[R1] ———, *New constructions of functions holomorphic in the unit ball of* $\mathbb{C}^n$, Lecture presented at the NSF-CBMS Regional Conference hosted by Michigan State University (1985).

[S] K. Stephenson, *Analytic functions and hypergroups of function pairs*, Indiana Univ. Math. J., **31** (1982), 843-884.

[SW] E. Stein and G. Weiss, *Introduction to Fourier Analysis on euclidean spaces*, Princeton University Press, 1971.

[T] M. Tsuji, *Potential theory in Modern Function theory*, Chelsea, 1959.

Received July 29, 1993 and revised September 12, 1994.

UNIVERSIDAD AUTÓNOMA DE MADRID
28049 MADRID, SPAIN
E-mail address: pando@ccuam3.sdi.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
28049 MADRID, SPAIN
E-mail address: madom@ccuam3.sdi.uam.es

AND

UNIVERSIDAD CARLOS III DE MADRID
BUTARQUE, 15
LEGANÉS, 28911 MADRID, SPAIN

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

IRREDUCIBLE NON-DENSE $A_1^{(1)}$-MODULES

V.M. FUTORNY

We study the irreducible weight non-dense modules for Affine Lie Algebra $A_1^{(1)}$ and classify all such modules having at least one finite-dimensional weight subspace. We prove that any irreducible non-zero level module with all finite-dimensional weight subspaces is non-dense.

1. Introduction.

Let $A = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$ and $\mathcal{G} = \mathcal{G}(A)$ is the associated Kac-Moody algebra over the complex numbers $\mathbf{C}$ with Cartan subalgebra $H \subset \mathcal{G}$, 1-dimensional center $\mathbf{C}c \subset H$ and root system Δ.

A $\mathcal{G}$-module V is called a *weight* if $V = \bigoplus_{\lambda \in H^*} V_\lambda$, $V_\lambda = \{v \in V \mid hv = \lambda(h)v$ for all $h \in H\}$. If V is an irreducible weight $\mathcal{G}$-module then c acts on V as a scalar. We will call this scalar the *level* of V, For a weight $\mathcal{G}$-module V, set $P(V) = \{\lambda \in H^* \mid V_\lambda \neq 0\}$.

Let $Q = \sum_{\varphi \in \Delta} \mathbf{Z}\varphi$. It is clear that if a weight $\mathcal{G}$-module V is irreducible then $P(V) \subset \lambda + Q$ for some $\lambda \in H^*$. An irreducible weight $\mathcal{G}$-module V is called *dense* if $P(V) = \lambda + Q$ for some $\lambda \in H^*$, and *non-dense* otherwise.

Irreducible dense modules whose weight spaces are all one-dimensional were classified by S. Spirin [**1**] for the algebra $A_1^{(1)}$ and by D. Britten, F. Lemire, F. Zorzitto [**2**] in the general case. It follows from [**2**] that such modules exist only for algebras $A_n^{(1)}$, $C_n^{(1)}$. V. Chari and A. Pressley constructed a family of irreducible integrable dense modules with all infinite-dimensional weight spaces. These modules can be realized as tensor product of standard highest weight modules with so-called loop modules [**3**].

In the present paper we study irreducible non-dense weight $\mathcal{G}$-modules. We use Kac [**4**] as a basic reference for notation, terminology and preliminary results. Our main result is the classification of all irreducible non-dense $\mathcal{G}$-modules having at least one finite-dimensional weight subspace. This includes, in particular, all irreducible highest weight modules. Moreover, we show that this classification includes all irreducible modules of non-zero level whose weight spaces are all finite- dimensional.

The paper is organized as follows. In Section 3 we study generalized Verma modules $M_\alpha^\varepsilon(\lambda,\gamma)$, α is a real root, $\lambda \in H^*$, $\gamma \in \mathbf{C}$, $\varepsilon \in \{+,-\}$ which do not necessarily have a highest weight (cf. [**5**]). By making use of the generalized Casimir operator and generalized Shapovalov form we obtain the criteria of irreducibility for the modules $M_\alpha^\varepsilon(\lambda,\gamma)$ without highest weight (Theorem 3.11).

In Section 4 we classify all irreducible $\mathbf{Z}$-graded modules for the Heisenberg subalgebra $G \subset \mathcal{G}$ with at least one finite-dimensional graded component. Irreducible G- modules with trivial action of c were described earlier in [**6**]. Let $\delta \in \Delta$ such that $\mathbf{Z}\delta - \{0\}$ is the set of all imaginary roots in Δ. Following [**6**] we introduce in Section 5 the category $\tilde{\mathcal{O}}(\alpha)$ of weight $\mathcal{G}$-modules $\tilde{V}$ such that $P(\tilde{V}) \subset \bigcup_{i=1}^{\ell} \{\lambda_i - k\alpha + n\delta \mid k,n \in \mathbf{Z}, k \geq 0\}$ where $\lambda_i \in H^*$, but without any restriction on the action of the center (unlike in [**6**] where the trivial action of the center is required). The irreducible objects in $\tilde{\mathcal{O}}(\alpha)$ are the unique quotients of $\mathcal{G}$-modules $M_\alpha(\lambda,V)$, where $\lambda \in H^*$, V is irreducible $\mathbf{Z}$-graded G-module. Modules $M_\alpha(\lambda,\mathbf{C})$, with $\lambda(c) = 0$ were studied in [**7-9**]. If $\lambda(c) \neq 0$ and at least one graded component of V is finite-dimensional then the module $M_\alpha(\lambda,V)$ is irreducible [**8**, **9**]. In Section 6 we classify all irreducible non-dense $\mathcal{G}$-modules with at least one finite-dimensional weight subspace (Theorem 6.2). It turns out that these modules are the quotients of the modules of type $M_\alpha^\varepsilon(\lambda,\gamma)$ or $M_\alpha(\lambda,V)$. Moreover, any irreducible $\mathcal{G}$-module of non-zero level whose weight spaces are all finite- dimensional is the quotient of $M_\alpha^\varepsilon(\lambda,\gamma)$ for some real root α, $\lambda \in H^*$, $\gamma \in \mathbf{C}$, $\varepsilon \in \{+,-\}$ (Theorem 6.3).

2. Preliminaries.

We have the root space decomposition for $\mathcal{G}$: $\mathcal{G} = H \oplus \sum_{\varphi \in \Delta} \mathcal{G}_\varphi$, where $\dim \mathcal{G}_\varphi = 1$ for all $\varphi \in \Delta$. Denote by $\mathcal{U}(\mathcal{G})$ the universal enveloping algebra of $\mathcal{G}$, by W the Weyl group and by $(\ ,\ )$ the standard non-degenerate symmetric bilinear form on $\mathcal{G}$ [**4**, Theorem 3.2]. Let Δ^{re} be the set of real roots in Δ and Δ^{im} be the set of imaginary roots in Δ. Fix $\alpha \in \Delta^{re}$ and consider a subalgebra $\mathcal{G}(\alpha) \subset \mathcal{G}$ generated by $\mathcal{G}_\alpha$ and $\mathcal{G}_{-\alpha}$. Then $\mathcal{G}(\alpha) \simeq sl(2)$ and we fix in $\mathcal{G}(\alpha)$ a standard basis e_α, $e_{-\alpha}$, $h_\alpha = [e_\alpha, e_{-\alpha}]$ where $[h_\alpha, e_{\pm\alpha}] = \pm 2e_{\pm\alpha}$. We will use the following realization of $\mathcal{G}$:

$$\mathcal{G} = \mathcal{G}(\alpha) \otimes \mathbf{C}[t,t^{-1}] \oplus \mathbf{C}c \oplus \mathbf{C}d$$

with $[x \otimes t^n + ac + bd,\ y \otimes t^m + a_1 c + b_1 d] = [x,y] \otimes t^{n+m} + bmy \otimes t^m - b_1 nx \otimes t^n + n\delta_{n,-m}(x,y)c$, for all x, $y \in \mathcal{G}(\alpha)$, $a, b, a_1, b_1 \in \mathbf{C}$. Then $H = \mathbf{C}h_\alpha \oplus \mathbf{C}c \oplus \mathbf{C}d$.

Denote by δ the element of H^* defined by: $\delta(h_\alpha) = \delta(c) = 0$ and $\delta(d) = 1$. Then $\Delta^{im} = \mathbf{Z}\delta - \{0\}$ and $\pi = \{\alpha, \delta - \alpha\}$ is a basis of Δ. Let $\Delta_+ = \Delta_+(\pi)$ be the set of all positive roots with respect to π. The root system Δ can be described in the following way: $\Delta = \{\pm\alpha + n\delta \mid n \in \mathbf{Z}\} \cup \{n\delta \mid n \in \mathbf{Z} - \{0\}\}$. We have $\mathcal{G}_{\pm\alpha+n\delta} = \mathcal{G}_{\pm\alpha} \otimes t^n$, $n \in \mathbf{Z}$, $\mathcal{G}_{n\delta} = \mathbf{C}h_\alpha \otimes t^n$, $n \in \mathbf{Z} - \{0\}$. Set $e_{\alpha+n\delta} = e_\alpha \otimes t^n$, $e_{-\alpha+n\delta} = e_{-\alpha} \otimes t^n$, $n \in \mathbf{Z}$, $e_{m\delta} = h_\alpha \otimes t^m$, $m \in \mathbf{Z} - \{0\}$. Then $[e_{k\delta}, e_{m\delta}] = 2k\delta_{k,-m}c$, $[e_{k\delta}, e_{\pm\alpha+n\delta}] = \pm 2e_{\pm\alpha+(n+k)\delta}$, $[e_{\alpha+k\delta}, e_{-\alpha+m\delta}] = \delta_{k,-m}(h_\alpha + kc) + (1 - \delta_{k,-m})e_{(k+m)\delta}$ for any $k, m \in \mathbf{Z}$.

For a Lie algebra $\mathcal{A}$, $S(\mathcal{A})$ will denote the corresponding symmetric algebra. We will identify the algebra $\mathcal{U}(H) = S(H)$ with the ring of polynomials $\mathbf{C}[H^*]$ and denote by σ the involutive antiautomorphism on $\mathcal{U}(\mathcal{G})$ such that $\sigma(e_\alpha) = e_{-\alpha}$, $\sigma(e_{\delta-\alpha}) = e_{\alpha-\delta}$. Set $\mathcal{N}_+ = \sum_{\varphi\in\Delta_+} \mathcal{G}_\varphi$, $\mathcal{N}_- = \sum_{\varphi\in\Delta_+} \mathcal{G}_{-\varphi}$.

3. Generalized Verma modules.

The center of $\mathcal{U}(\mathcal{G}(\alpha))$ is generated by the Casimir element $z_\alpha = (h_\alpha + 1)^2 + 4e_{-\alpha}e_\alpha$. Denote

$$\mathcal{N}_\alpha^+ = \sum_{\varphi\in\Delta_+-\{\alpha\}} \mathcal{G}_\varphi, \qquad \mathcal{N}_\alpha^- = \sum_{\varphi\in\Delta_+-\{\alpha\}} \mathcal{G}_{-\varphi},$$

$$T_\alpha = S(H) \otimes \mathbf{C}[z_\alpha], \qquad E_\alpha^\varepsilon = (H + \mathcal{G}(\alpha)) \oplus \mathcal{N}_\alpha^\varepsilon,\ \varepsilon \in \{+, -\}.$$

Let $\lambda \in H^*, \gamma \in \mathbf{C}$. Consider the 1-dimensional T_α-module $\mathbf{C}v_\lambda$ with the action $(h \otimes z_\alpha^n)v_\lambda = h(\lambda)\gamma^n v_\lambda$ for any $h \in S(H)$, and construct an $H + \mathcal{G}(\alpha)$-module

$$V(\lambda, \gamma) = \mathcal{U}(\mathcal{G}(\alpha) + H) \bigotimes_{T_\alpha} \mathbf{C}v_\lambda.$$

It is clear that the module $V(\lambda, \gamma)$ has a unique irreducible quotient $V_{\lambda,\gamma}$.

Proposition 3.1.

(i) *If V is an irreducible weight $H + \mathcal{G}(\alpha)$-module then $V \simeq V_{\lambda,\gamma}$ for some $\lambda \in H^*$, $\gamma \in \mathbf{C}$.*

(ii) *$V_{\lambda,\gamma} \simeq V_{\lambda',\gamma'}$ if and only if $\gamma = \gamma'$, $\lambda' = \lambda + n\alpha$, $n \in \mathbf{Z}$, $\gamma \neq (\lambda(h_\alpha) + 2\ell + 1)^2$ for all integers ℓ, $0 \leq \ell < n$ if $n \geq 0$ or for all integers ℓ, $n \leq \ell < 0$ if $n < 0$.*

Proof. This is essentially the classification of irreducible weight $sl(2)$-modules. □

Let $\lambda \in H^*$, $\gamma \in \mathbf{C}$, $\varepsilon \in \{+, -\}$. Consider $V_{\lambda,\gamma}$ as E_α^ε-module with trivial action of $\mathcal{N}_\alpha^\varepsilon$ and construct the $\mathcal{G}$-module

$$M_\alpha^\varepsilon(\lambda, \gamma) = \mathcal{U}(\mathcal{G}) \bigotimes_{\mathcal{U}(E_\alpha^\varepsilon)} V_{\lambda,\gamma}$$

associated with $\alpha, \lambda, \gamma, \varepsilon$.

The module $M_\alpha^\varepsilon(\lambda, \gamma)$ is called a generalized Verma module. Notice that $V_{\lambda,\gamma}$ does not have to be finite-dimensional.

Proposition 3.2.

(i) $M_\alpha^\varepsilon(\lambda, \gamma)$ *is a free* $\sigma(\mathcal{U}(\mathcal{N}_\alpha^\varepsilon))$- *module with all finite-dimensional weight subspaces.*

(ii) $M_\alpha^\varepsilon(\lambda, \gamma)$ *has a unique irreducible quotient,* $L_\alpha^\varepsilon(\lambda, \gamma)$.

(iii) $M_\alpha^\varepsilon(\lambda, \gamma) \simeq M_{\pm\alpha}^{\varepsilon'}(\lambda', \gamma')$ *if and only if* $\varepsilon = \varepsilon'$, $\gamma = \gamma'$, $\lambda' = \lambda + n\alpha, n \in \mathbf{Z}$ *and* $\gamma \neq (\lambda(h_\alpha) + 2\ell + 1)^2$ *for all* $\ell \in \mathbf{Z}$, $0 \leq \ell < n$ *if* $n \geq 0$ *or for all* $\ell \in \mathbf{Z}$, $n \leq \ell < 0$ *if* $n < 0$.

Proof. Follows from the construction of $\mathcal{G}$- module $M_\alpha^\varepsilon(\lambda, \gamma)$ and Proposition 3.1. □

Let $R_\lambda = \{(\lambda(h_\alpha) + 2\ell + 1)^2 \mid \ell \in \mathbf{Z}\}$. Recall that V is called a highest weight module with respect to $\mathcal{N}_+$ and with highest weight $\lambda \in H^*$ if $V = \mathcal{U}(\mathcal{G})v$, $v \in V_\lambda$ and $V_{\lambda+\varphi} = 0$ for all $\varphi \in \Delta_+(\pi)$. Proposition 3.2, (iii) implies that $M_\alpha^\varepsilon(\lambda, \gamma)$ and $L_\alpha^\varepsilon(\lambda, \gamma)$ are highest weight modules with respect to some choice of basis of Δ and, therefore, are the quotients of Verma modules [**4**], if and only if $\gamma \in R_\lambda$. The theory of highest weight modules was developed in [**4**, **10**].

Corollary 3.3.

(i) *Let* V *be an irreducible weight* $\mathcal{G}$*-module,* $0 \neq v \in V_\lambda$ *and* $\mathcal{N}_\alpha^\varepsilon v = 0$. *Then* $V \simeq L_\alpha^\varepsilon(\lambda, \gamma)$ *for some* $\gamma \in \mathbf{C}$.

(ii) *Let* $\lambda \notin R_\lambda$. $L_\alpha^\varepsilon(\lambda, \gamma) \simeq L_{\alpha'}^{\varepsilon'}(\lambda', \gamma')$ *if and only if* $\varepsilon = \varepsilon'$, $\alpha' = \alpha$ *or* $\alpha' = -\alpha$, $\gamma = \gamma'$, $\lambda' = \lambda + n\alpha$, $n \in \mathbf{Z}$ *and* $\gamma \neq (\lambda(h_\alpha) + 2\ell + 1)^2$ *for all* $\ell \in \mathbf{Z}$, $0 \leq \ell < n$ *if* $n \geq 0$ *or for all* $\ell \in \mathbf{Z}$, $n \leq \ell < 0$ *if* $n < 0$.

Proof. Since V is irreducible $\mathcal{G}$- module, $V' = \mathcal{U}(\mathcal{G}(\alpha))v$ is an irreducible $\mathcal{G}(\alpha)$-module and $V \simeq \sigma(\mathcal{U}(\mathcal{N}_\alpha^\varepsilon))V'$. Then V is a homomorphic image of $M_\alpha^\varepsilon(\lambda, \gamma)$ for some $\gamma \in \mathbf{C}$ and, thus, $V \simeq L_\alpha^\varepsilon(\lambda, \gamma)$ which proves (i). (ii) follows from Proposition 3.2, (iii). □

From now on we will consider the modules $M_\alpha^+(\lambda, \gamma)(= M(\lambda, \gamma))$. All the results for the modules $M_\alpha^-(\lambda, \gamma)$ can be proved analogously. Set $z = z_\alpha$. For $\lambda \in H^*$, $\gamma \in \mathbf{C}$ and integer $n \geq 0$ we denote by $z(n)$ the restriction of z to the subspace $M(\lambda, \gamma)_{\lambda - n(\delta - \alpha)}$.

Proposition 3.4. *If* $\gamma \neq (\lambda(h_\alpha) + 2\ell + 1)^2$ *for all* $0 \leq \ell < 2n$ *then* $\operatorname{Spec} z(n) = \{(2k \pm \sqrt{\gamma})^2 \mid k \in \mathbf{Z}, 0 \leq k \leq n\}$.

Proof. Denote $V_n = M(\lambda, \gamma)_{\lambda - n(\delta - \alpha)}$, $n > 0$. One can easily show that $V_n = e_{\alpha-\delta} V_{n-1} + e_{-\delta} e_\alpha V_{n-1} + e_{-\alpha-\delta} e_\alpha^2 V_{n-1}$. Let $V_{n-1} = \oplus V_{n-1}(\tau), \tau \in \mathbf{C}$,

where $V_{n-1}(\tau) = \{v \in V_{n-1} \mid \exists N : (z(n-1)-\tau)^N v = 0\}$. Then the subspace $e_{\alpha-\delta}V_{n-1}(\tau) + e_{-\delta}e_\alpha V_{n-1}(\tau) + e_{-\alpha-\delta}e_\alpha^2 V_{n-1}(\tau) \subset V_n$ is $z(n)$- invariant and $z(n)$ has on it the eigenvalues τ and $(2 \pm \sqrt{\tau})^2$, thanks to the condition $\gamma \neq (\lambda(h_\alpha) + 2\ell + 1)^2$, $0 \leq \ell < 2n$, which implies that $z(n)$ has eigenvalues $(2k \pm \sqrt{\gamma})^2$, $0 \leq k \leq n$. □

Corollary 3.5. *If $\gamma \notin R_\lambda$ then e_α and $e_{-\alpha}$ act injectively on $M(\lambda, \gamma)$.*

Proof. If $\gamma \notin R_\lambda$ then $\operatorname{Spec} z(n) \bigcap R_{\lambda - n\beta} = \emptyset$ for all integer $n \geq 0$ by Proposition 3.4 and, therefore, e_α and $e_{-\alpha}$ act injectively on $M(\lambda, \gamma)$. □

Fix $\rho \in H^*$ such that $(\rho, \alpha) = 1$, $(\rho, \delta) = 2$. Since $M(\lambda, \gamma)$ is a restricted module, i.e. for every $v \in M(\lambda, \gamma)$, $\mathcal{G}_\varphi v = 0$ for all but a finite number of positive roots φ, we have well-defined action of a generalized Casimir operator Ω on $M(\lambda, \gamma)$ [**4**]:

$$\Omega v = (\mu + 2\rho, \mu)v + 2 \sum_{\varphi \in \Delta_+} \overline{e}_{-\varphi} e_\varphi v, \; v \in M(\lambda, \gamma)_\mu,$$

where $\overline{e}_{-\varphi} \in \mathcal{G}_{-\varphi}$, $(\overline{e}_{-\varphi}, e_\varphi) = 1$, $\varphi \in \Delta_+$. Set $\tilde{\Omega} = 2\Omega + id$.

Let $s_\alpha \in W$, $s_\alpha(\mu) = \mu - (\mu, \alpha)\alpha$, $\mu \in H^*$.

Lemma 3.6. *For a $\mathcal{G}$-module $M(\lambda, \gamma)$*

$$\tilde{\Omega} = [(\lambda + 2\rho + s_\alpha(\lambda + 2\rho), \lambda) + \gamma]id.$$

Proof. Follows from [**4**, Th.2.6] and definition of $\tilde{\Omega}$. □

Lemma 3.7. *Let $n > 0, \beta = \delta - \alpha$, $0 \neq v \in M(\lambda, \gamma)_{\lambda - n\beta}$, $\gamma \neq (\lambda(h_\alpha) + 2\ell + 1)^2$ for all $0 \leq \ell < 2n$ and $\mathcal{N}_\alpha^+ v = 0$. Then $k^2\gamma = (n(\lambda(c) + 2) - k^2)^2$ for some $k \in \mathbf{Z}$, $0 \leq k \leq n$.*

Proof. It follows from Lemma 3.6 that $z(n)v = \gamma' v$ and

$$(\lambda - n\beta + 2\rho + s_\alpha(\lambda - n\beta + 2\rho), \lambda - n\beta) + \gamma' = (\lambda + 2\rho + s_\alpha(\lambda + 2\rho), \lambda) + \gamma$$

which implies

$$\gamma' = \gamma + 4n(\lambda(c) + 2).$$

But, $\gamma' = (2k \pm \sqrt{\gamma})^2$ for some $k \in \mathbf{Z}$, $0 \leq k \leq n$ by Proposition 3.4. Therefore, $k^2\gamma = (n(\lambda(c) + 2) - k^2)^2$ which completes the proof. □

Corollary 3.8. *Let $\lambda \in H^*$, $\gamma \in \mathbf{C} - R_\lambda$. If $k^2\gamma \neq (n(\lambda(c) + 2) - k^2)^2$ for all $n, k \in \mathbf{Z}$, $n > 0$, $0 \leq k \leq n$ then $\mathcal{G}$-module $M(\lambda, \gamma)$ irreducible.*

Proof. If the $\mathcal{G}$-module $M(\lambda, \gamma)$ has a non-trivial submodule M, then M contains a non-zero vector v of weight $\lambda - n(\delta - \alpha)$, $n > 0$, such that $\mathcal{N}_\alpha^+ v = 0$. Now, the statement follows from Lemma 3.7. □

Consider the following decomposition of $\mathcal{U}(\mathcal{G})$:

$$\mathcal{U}(\mathcal{G}) = (\mathcal{N}_\alpha^- \mathcal{U}(\mathcal{G}) + \mathcal{U}(\mathcal{G})\mathcal{N}_\alpha^+) \oplus T_\alpha \mathbf{C}[e_\alpha]e_\alpha \oplus T_\alpha \mathbf{C}[e_{-\alpha}]e_{-\alpha} \oplus T_\alpha.$$

Let j be the projection of $\mathcal{U}(\mathcal{G})$ to T_α. Introduce the generalized Shapovalov form F, a symmetric bilinear form on $\mathcal{U}(\mathcal{G})$ with values in T_α, as follows (cf. [11]): $F(x,y) = j(\sigma(x)y)$, $x, y \in \mathcal{U}(\mathcal{G})$. The algebra $\mathcal{U}(\mathcal{G})$ is Q-graded: $\mathcal{U}(\mathcal{G}) = \bigoplus_{\eta \in Q} \mathcal{U}(\mathcal{G})_\eta$. It is clear that $F(\mathcal{U}(\mathcal{G})_{\eta_1}, \mathcal{U}(\mathcal{G})_{\eta_2}) = 0$ if $\eta_1 \neq \eta_2$. Denote $\mathcal{U}(\mathcal{N}_-)_{-\eta} = \mathcal{U}(\mathcal{N}_-) \bigcap \mathcal{U}(\mathcal{G})_{-\eta}$ and let F_η be a restriction of F to $\mathcal{U}(\mathcal{N}_-)_{-\eta}$.

For $\lambda \in H^*$, $\gamma \in \mathbf{C}$, consider the linear map $\theta_{\lambda,\gamma} : T_\alpha \to \mathbf{C}$ defined by $\theta_{\lambda,\gamma}(h \otimes z^n) = h(\lambda)\gamma^n$ for any $h \in S(H)$, $n \in \mathbf{Z}_+$.

Set $\lambda_k = \lambda + k\alpha$, $k \in \mathbf{Z}$. Let $\mu = \lambda - n(\delta - \alpha) \in P(M(\lambda,\gamma))$, $n \in \mathbf{Z}_+$ and $\gamma \neq (\lambda(h_\alpha) + 2s + 1)^2$ for all integer s, $0 \leq s < 2n$. Then $\lambda_{2n} \in P(M(\lambda,\gamma))$, $M(\lambda,\gamma)_{\lambda_{2n}} = \mathbf{C}v_n$ and $M(\lambda,\gamma)_\mu = \mathcal{U}(\mathcal{N}_-)_{-n(\alpha+\delta)}v_n$. Set $F^{(n)} = F_{n(\alpha+\delta)}$. We define a a bilinear $\mathbf{C}$-valued form F^0_μ on $M(\lambda,\gamma)_\mu$ as follows:

$$F^0_\mu(u_1 v_n,\ u_2 v_n) = \theta_{\lambda_{2n},\gamma}\left(F^{(n)}(u_1,\ u_2)\right),\ u_1,\ u_2 \in \mathcal{U}(\mathcal{N}_-)_{-n(\alpha+\delta)}.$$

One can see that $\dim L(\lambda,\gamma)_\mu = \operatorname{rank} F^0_\mu$.

Lemma 3.9. *Let $\lambda \in H^*$, $\gamma \in \mathbf{C} - R_\lambda$. The following conditions are equivalent:*

(i) *$M(\lambda,\gamma)$ is irreducible.*

(ii) *$F^0_{\lambda - n(\delta-\alpha)}$ is non-degenerate for all integers $n > 0$.*

(iii) *$\theta_{\lambda_{2n},\gamma}(\det F^{(n)}) \neq 0$ for all integers $n > 0$.*

Proof. Follows from the Corollary 3.5. □

Consider in T_α the following polynomials: $f_{m,k} = k^2 z - (m(c+2) - k^2)^2$, $g_s = z - (h_\alpha + 2s + 1)^2$, $s, m, k \in \mathbf{Z}$, $0 \leq k \leq m$. Lemma 3.7 implies that if $\theta_{\lambda,\gamma}(g_s) \neq 0$ for all $s \in \mathbf{Z}$, $0 \leq s < 2n$ and $\theta_{\lambda_{2m},\gamma}(f_{m,k}) \neq 0$ for all $m, k \in \mathbf{Z}$, $0 < m \leq n$, $0 \leq k \leq m$, then $M(\lambda,\gamma)_{\lambda - n(\delta-\alpha)} = L(\lambda,\gamma)_{\lambda - n(\delta-\alpha)}$ and $\theta_{\lambda_{2n},\gamma}(\det F^{(n)}) \neq 0$. We conclude that the polynomial $\det F^{(n)}$ is not identically equal to zero and has its zeros in the union of zeros of polynomials $f_{m,k}$, $0 < m \leq n$, $0 \leq k \leq m$, g_s, $0 \leq s \leq 2n$. Therefore, $\det F^{(n)}$ is a product of factors of type $f_{m,k}$ and g_s.

Lemma 3.10. *Let $n, m \in \mathbf{Z}$, $n > 0$, $0 < m \leq n$. Then $f_{m,k}$ is a factor of $\det F^{(n)}$ if and only if k is a divisor of m or $k = 0$.*

Proof. Assume that k is a divisor of m or $k = 0$. Set $r = 2n + 2m + k$. Consider $\lambda \in H^*$ and $\gamma \in \mathbf{C} - \mathbf{Z}$ such that $\theta_{\lambda,\gamma}(f_{m,k}) = \theta_{\lambda,\gamma}(g_r) = 0$. For integer $s \geq 0$

set $\nu_s = \lambda_{-s} = \lambda - s\alpha$. Then $\theta_{\nu_s,\gamma}(f_{m,k}) = \theta_{\nu_s,\gamma}(g_{r+s}) = 0$ and $\nu_s(h_\alpha) \notin \mathbf{Z}$, which implies that $\theta_{\nu_s,\gamma}(g_\ell) \neq 0$ for all $\ell \in \mathbf{Z}$, $\ell < r+s$. Thus, the form $F^0_{\nu_s - i\beta}$, $\beta = \delta - \alpha$ is defined for all $s \geq 0$, $0 < i \leq n$ and $M(\nu_s, \gamma) \simeq M(\lambda_r)$, $s \geq 0$ by Proposition 3.2, (iii), where $M(\lambda_r)$ is the Verma module with highest weight $\lambda_r = \lambda + r\alpha$. Therefore, $M(\nu_s,\gamma)_{\nu_s - i\beta} \simeq M(\lambda_r)_{\nu_s - i\beta}$, $0 < i \leq n$ as T_α-modules. The operator $z(m)$ has eigenvectors w_s^+, $w_s^- \in M(\lambda_r)_{\nu_s - m\beta}$ with eigenvalues $\gamma^+ = (\lambda(h_\alpha)+4(n+m+k)+1)^2$ and $\gamma^- = (\lambda(h_\alpha)+4(n+m)+1)^2$ respectively. Since $\theta_{\nu_s,\gamma}(f_{m,k}) = 0$, then

$$\gamma^* = \gamma + 4m(\lambda(c) + 2) \in \{\gamma^+, \gamma^-\}$$

and

$$(\nu_s+2\rho+s_\alpha(\nu_s+2\rho),\nu_s)+\gamma = (\nu_s-m\beta+2\rho+s_\alpha(\nu_s-m\beta+2\rho),\nu_s-m\beta)+\gamma^*.$$

Let $w_s^* \in \{w_s^+, w_s^-\}$ and $z(m)w_s^* = \gamma^* w_s^*$. Then

$$\tilde{\Omega} w_s^* = [(\nu_s - m\beta + 2\rho + s_\alpha(\nu_s - m\beta + 2\rho), \nu_s - m\beta) + \gamma^*]w_s^*$$

by Lemma 3.6. But, $w_s^* \in M(\lambda_r)$ and

$$\tilde{\Omega}\, w_s^* = (2(\lambda_r + 2\rho, \lambda_r) + 1)w_s^*$$

by Corollary 2.6 in [**4**]. Hence

$$2(\lambda_r + 2\rho, \lambda_r) + 1 = (\nu_s - m\beta + 2\rho + s_\alpha(\nu_s - m\beta + 2\rho), \nu_s - m\beta) + \gamma^*$$

and

$$(\lambda_r + 2\rho, \lambda_r) = (\lambda_r + 2\rho - \tau^*, \lambda_r - \tau^*)$$

where $\tau^* = m\delta - k\alpha$ if $\gamma^* = \gamma^+$ and $\tau^* = m\delta + k\alpha$ if $\gamma^* = \gamma^-$. If k divides m or $k = 0$ then τ^* is a quasiroot and $D = Hom_{\mathcal{G}}(M(\lambda_r - \tau^*), M(\lambda_r)) \neq 0$ [**10**, Prop. 4.1].

Let $0 \neq \chi \in D$. Then $\chi(M(\lambda_r - \tau^*)) \cap M(\lambda_r)_{\nu_s - n\beta} \neq 0$ and therefore, $\theta_{\lambda_{2n-s},\gamma}(\det F^{(n)}) = 0$ for any integer $s \geq 0$. It implies that if $\lambda \in H^*$, $\gamma \in \mathbf{C} - \mathbf{Z}$ and $\theta_{\lambda,\gamma}(f_{m,k}) = 0$ then $\theta_{\lambda,\gamma}(\det F^{(n)}) = 0$. Thus, $f_{m,k}$ is a factor of $\det F^{(n)}$. Conversely, suppose that $f_{n,k}$ is a factor of $\det F^{(n)}$, $k \neq 0$ and k is not a divisor of n. Let $r = 4n+k$. Consider a pair $(\lambda, \gamma) \in H^* \times (\mathbf{C} - \mathbf{Z})$ such that $\theta_{\lambda,\gamma}(f_{n,k}) = \theta_{\lambda,\gamma}(g_r) = 0$ but $\theta_{\lambda,\gamma}(f_{p,q}) \neq 0$ for all $0 < p < n$, $0 \leq q \leq p$ (such λ and γ always exist). Then $\theta_{\lambda,\gamma}(\det F^{(n)}) = 0$ and the Verma module $M(\lambda_r)$ has an irreducible subquotient with highest weight $\lambda_r - \tau^*$, where τ^* is one of $n\delta + k\alpha$, $n\delta - k\alpha$. But, this contradicts the Theorem 2 in [**10**]. Therefore, $f_{n,k}$ can not be a factor of $\det F^{(n)}$ if $k \neq 0$ and k is not a divisor of n.

Let now $0 < m < n$, $0 < k < m$, k is not a divisor of m and $f_{m,k}$ is a factor of $\det F^{(n)}$. Consider a pair $(\lambda, \gamma) \in H^* \times \mathbf{C}$ such that $\theta_{\lambda,\gamma}(f_{m,k}) = 0$, $\theta_{\lambda,\gamma}(f_{p,q}) \neq 0$ for all $p, q \in \mathbf{Z}$, $0 < p \leq n$, $0 \leq q \leq p$, $(p,q) \neq (m,k)$ and $\theta_{\lambda,\gamma}(g_s) \neq 0$ for all $s \in \mathbf{Z}$. As it was shown above $f_{m,k}$ is not a factor of $\det F^{(m)}$ which implies that $\theta_{\lambda_{2m},\gamma}(\det F^{(m)}) \neq 0$. Now it follows from Lemma 3.7 that $M(\lambda,\gamma)_{\lambda - n\beta} = L(\lambda,\gamma)_{\lambda - n\beta}$ and $\theta_{\lambda_{2n},\gamma}(\det F^{(n)}) \neq 0$. But, this contradicts the assumption that $f_{m,k}$ is a factor of $\det F^{(n)}$. The Lemma is proved. □

For $n \in \mathbf{Z}$, $n > 0$ denote $X_n = \{0\} \cup \{k \in \mathbf{Z}_+ \mid \frac{n}{k} \in \mathbf{Z}\}$.

Theorem 3.11. *Let $\lambda \in H^*$, $\gamma \in \mathbf{C} - R_\lambda$. $\mathcal{G}$-module $M(\lambda,\gamma)$ is irreducible if and only if $k^2\gamma \neq (n(\lambda(c)+2) - k^2)^2$ for all $n \in \mathbf{Z}$, $n > 0$, $k \in X_n$.*

Proof. Follows from Lemmas 3.9 and 3.10. □

4. Irreducible representations of the Heisenberg subalgebra.

Consider the Heisenberg subalgebra $G = \mathbf{C}c \oplus \sum_{k \in \mathbf{Z} - \{0\}} \mathcal{G}_{k\delta} \subset \mathcal{G}$. It is a $\mathbf{Z}$-graded algebra with $deg\, c = 0$, $deg\, e_{k\delta} = k$. This gradation induces a $\mathbf{Z}$-gradation on the universal enveloping algebra $\mathcal{U}(G)$: $\mathcal{U}(G) = \bigoplus_{i \in \mathbf{Z}} \mathcal{U}_i$.

In this section we study the irreducible $\mathbf{Z}$-graded G- modules. The central element c acts as a scalar on each such module. In general, we say that a G-module V is a module of level $a \in \mathbf{C}$ if c acts on V as a multiplication by a.

4.1. G-Modules of non-zero level. Let $G_+ = \sum_{k>0} \mathcal{G}_{k\delta}$, $G_- = \sum_{k<0} \mathcal{G}_{k\delta}$. For $a \in \mathbf{C}^* = \mathbf{C} - \{0\}$, let $\mathbf{C}v_a$ be the 1- dimensional $G_\varepsilon \oplus \mathbf{C}c$-module for which $G_\varepsilon v_a = 0$, $cv_a = av_a$, $\varepsilon \in \{+,-\}$. Consider the G-module

$$M^\varepsilon(a) = \mathcal{U}(G) \bigotimes_{\mathcal{U}(G_\varepsilon \oplus \mathbf{C}c)} \mathbf{C}v_a$$

associated with a and ε.

The module $M^\varepsilon(a)$ is a $\mathbf{Z}$-graded: $M^\varepsilon(a) = \sum_{i \in \mathbf{Z}} M^\varepsilon(a)_i$ where

$$M^\varepsilon(a)_i = (\sigma(\mathcal{U}(G_\varepsilon)) \cap \mathcal{U}_i) \otimes v_a.$$

Proposition 4.1.

(i) *The G-module $M^\varepsilon(a)$ is irreducible.*

(ii) *$M^\varepsilon(a)$ is a $\sigma(\mathcal{U}(G_\varepsilon))$-free module.*

(iii) $\dim M^{\varepsilon}(a)_i = P(|\, i\,|)$ *where* $P(n)$ *is a partition function.*

Proof. (ii) and (iii) follow directly from the definition of $M^{\varepsilon}(a)$. Since $a \neq 0$ one can easily show that for any non-zero $u \in \sigma(\mathcal{U}(G_{\varepsilon}))$ there exists $u' \in \mathcal{U}(G_{\varepsilon})$ such that $0 \neq u'uv_a \in M^{\varepsilon}(a)_0$ which implies (i) and completes the proof. □

Lemma 4.2. *If* V *is a* $\mathbf{Z}$*-graded* G*-module of level* $a \in \mathbf{C}^*$ *and* $\dim V_i < \infty$ *for at least one* $i \in \mathbf{Z}$ *then*

$$\operatorname{Spec} e_{\delta}e_{-\delta} \,|_V \subset \{2ma \mid m \in \mathbf{Z}\}\,.$$

Proof. Let $v \in V_j$ be a non-zero eigenvector of $e_{\delta}e_{-\delta}$ with eigenvalue b and $b \neq 2ma$ for all $m \in \mathbf{Z}$. Since $a \neq 0$, if $e_{n\delta}v = 0$ then $e_{-n\delta}v \neq 0$, $n \in \mathbf{Z} - \{0\}$. Denote $Y = \{n \in \mathbf{Z} - \{0,1\} \mid e_{n\delta}v \neq 0\}$. We may assume without lost of generality that $j = i$ and $|\, Y \cap \mathbf{Z}_+ \,| = \infty$. Elements e_{δ} and $e_{-\delta}$ act injectively on the subspace spanned by $e_{\delta}^k v$, $e_{-\delta}^k v$, $k \in \mathbf{Z}$. Then, for each $k \in Y \cap \mathbf{Z}_+$, $e_{\delta}e_{-\delta}(e_{k\delta}v) = be_{k\delta}v$ and $0 \neq e_{-\delta}^k e_{k\delta}v \in V_i$. Set $w_k = e_{-\delta}^k e_{k\delta}v$. Then $e_{\delta}e_{-\delta}w_k = (b + 2ka)w_k$, $k \in Y \cap \mathbf{Z}_+$. This contradicts the assumption that $\dim V_i < \infty$. Therefore, $b = 2ma$ for some $m \in \mathbf{Z}$. □

For a $\mathbf{Z}$-graded G-module V and $j \geq 0$ denote by $V^{[j]}$ the $\mathbf{Z}$-graded G-module with $(V^{[j]})_i = V_{i-j}$, $i \in \mathbf{Z}$.

We describe now all irreducible $\mathbf{Z}$-graded G-modules of non-zero level with finite-dimensional components.

Proposition 4.3.

(i) *Let* V *be an irreducible* $\mathbf{Z}$*-graded* G*-module of level* $a \in \mathbf{C}^*$ *such that* $\dim V_i < \infty$ *for at least one* $i \in \mathbf{Z}$. *Then* $V^{[j]} \simeq M^{\varepsilon}(a)$ *for some* $\varepsilon \in \{+,-\}$, $j \in \mathbf{Z}$.

(ii) $\operatorname{Ext}^1((M^{\varepsilon}(a))^{[j]}, M^{\varepsilon'}(a)) = 0$ *for any* $j \in \mathbf{Z}$, $\varepsilon, \varepsilon' \in \{+,-\}$.

Proof. (i) By Lemma 4.2 $\operatorname{Spec} X \,|_V \subset \{2ma \mid m \in \mathbf{Z}\}$ where X stands for $e_{\delta}e_{-\delta}$. Let $V_i \neq 0$, n be an integer with maximal absolute value such that $2na \in \operatorname{Spec} X \,|_{V_i}$ and let $0 \neq v \in V_i$, $Xv = 2nav$. Assume that $n > 0$. Then $e_{k\delta}v = 0$ for all $k > 1$. Indeed, if $e_{k\delta}v \neq 0$ for some $k > 1$ then $X(e_{k\delta}v) = e_{k\delta}Xv = 2na\, e_{k\delta}v$ and $2(n+k)a$ is an eigenvalue of X on V_i which contradicts the assumption. Therefore, $e_{k\delta}v = 0$ for all $k > 1$. Consider the element $\tilde{v} = e_{\delta}^{n-1}v \neq 0$. Then $e_{-\delta}e_{\delta}\tilde{v} = e_{k\delta}\tilde{v} = 0$, $k > 1$. If $e_{\delta}\tilde{v} \neq 0$ then $v_p = e_{\delta}^p\tilde{v} \neq 0$, $e_{k\delta}v_p = 0$ and, hence $e_{-k\delta}v_p \neq 0$ for all $p > 0$, $k > 1$. This would imply that $\dim V_i = \infty$. Therefore, $e_{\delta}\tilde{v} = 0$ and $V = \mathcal{U}(G)\tilde{v} \simeq M^{+}(a)$ up to a shifting of gradation. If $n \leq 0$ then, clearly,

$V \simeq M^-(a)$ up to a shifting of gradation. Suppose that $V_i = 0$ but, for example, $V_{i-1} \neq 0$. Then $e_{k\delta}v = 0$ for any non-zero $v \in V_{i-1}$ for all $k > 0$ and thus $V = \mathcal{U}(G)v \simeq M^+(a)$ up to a shifting of gradation. This completes the proof of (i).
(ii) Follows from the proof of (i) and Proposition 4.1, (ii). □

Lemma 4.4. *Every finitely-generated* $\mathbf{Z}$*-graded* G*-module* V *of level* $a \in \mathbf{C}^*$ *such that* $\dim V_i < \infty$ *for at least one* $i \in \mathbf{Z}$ *has a finite length.*

Proof. If $V_i = 0$ then statement follows from Proposition 4.3. Let $V_i \neq 0$, n be an integer with maximal absolute value such that $2na \in \operatorname{Spec} e_\delta e_{-\delta}|_{V_i}$ and v be a corresponding eigenvector. It follows from the proof of Proposition 4.3, (i) that $V' = \mathcal{U}(G)v \simeq M^\varepsilon(a)$ up to a shifting of gradation. Consider a G-module $\tilde{V} = V/V'$. Then $\dim \tilde{V}_i < \dim V_i$ and we can complete the proof by induction on $\dim V_i$. □

Now we are in the position to establish the completely reducibility for for finitely-generated G-modules of non-zero level with finite-dimensional components.

Proposition 4.5. *Every finitely-generated* $\mathbf{Z}$*-graded* G*-module* V *of a non-zero level such that* $\dim V_i < \infty$ *for at least one* $i \in \mathbf{Z}$ *is completely reducible.*

Proof. Follows from Lemma 4.4 and Proposition 4.3. □

4.2. G**-modules of level zero.** The irreducible G-modules of level zero are classified by V. Chari [**6**]. We recall this classification.

Let $\tilde{G} = \mathcal{U}(G)/\mathcal{U}(G)c$ and let $g : \mathcal{U}(G) \to \tilde{G}$ be the canonical homomorphism. For $r > 0$ consider a $\mathbf{Z}$-graded ring $L_r = \mathbf{C}[t^r, t^{-r}]$, $deg\, t = 1$ and denote by P_r the set of graded ring epimorphisms $\Lambda : \tilde{G} \to L_r$ with $\Lambda(1) = 1$. Let $L_0 = \mathbf{C}$ and $\Lambda_0 : \tilde{G} \to \mathbf{C}$ is a trivial homomorphism such that $\Lambda_0(1) = 1$, $\Lambda_0(g(e_{k\delta})) = 0$ for all $k \in \mathbf{Z} - \{0\}$. Set $P_0 = \{\Lambda_0\}$.

Given $\Lambda \in P_r$, $r \geq 0$ define a G-module structure on L_r by:

$$e_{k\delta}t^{rs} = \Lambda(g(e_{k\delta}))t^{rs}, \quad k \in \mathbf{Z} - \{0\}, \quad ct^{rs} = 0, s \in \mathbf{Z}.$$

Denote this G-module by $L_{r,\Lambda}$.

Proposition 4.6.

(i) *Let* V *be an irreducibe* $\mathbf{Z}$*-graded* G*-module of level zero. Then* $V \simeq L_{r,\Lambda}$ *for some* $r \geq 0$, $\Lambda \in P_r$ *up to a shifting of gradation.*

(ii) $L_{r,\Lambda} \simeq L_{r',\Lambda'}$ *if and only if* $r = r'$ *and there exists* $b \in \mathbf{C}^*$ *such that* $\Lambda(g(e_{k\delta})) = b^k\Lambda'(g(e_{k\delta}))$, $k \in \mathbf{Z} - \{0\}$.

Proof. (i) is essentially Lemma 3.6 in [**6**]; (ii) follows from [**6**, Prop. 3.8]. □

Remark 4.7. All the results of Section 4, except Proposition 4.1 (iii), are hold for the Heisenberg subalgebra of an arbitrary Affine Lie Algebra.

5. The category $\tilde{\mathcal{O}}(\alpha)$.

Let $\alpha \in \pi$. Following [**6**] we define category $\tilde{\mathcal{O}}(\alpha)$ to be the category of weight $\mathcal{G}$-modules M satisfying the condition that there exist finitely many elements $\lambda_1, ..., \lambda_r \in H^*$ such that $P(M) \subseteq \bigcup_{i=1}^{r} D(\lambda_i)$ where

$$D(\lambda_i) = \{\lambda_i + k\alpha + n\delta \mid k, n \in \mathbf{Z},\ k \leq 0\}.$$

Notice that the trivial action of c, as in [**6**], is no longer required. It is clear that $\tilde{\mathcal{O}}(\alpha)$ is closed under the operations of taking submodules, quotients and finite direct sums.

Denote $B_\alpha = \sum_{n \in \mathbf{Z}} \mathcal{G}_{\alpha+n\delta}$. Then $\mathcal{G} = B_{-\alpha} \oplus (H + G) \oplus B_\alpha$.

Let V be an irreducible $\mathbf{Z}$-graded G-module of level $a \in \mathbf{C}$ and let $\lambda \in H^*$, $\lambda(c) = a$. Then we can define a $B = (H + G) \oplus B_\alpha$-module structure on V by setting: $hv_i = (\lambda + i\delta)(h)v_i$, $B_\alpha v_i = 0$ for all $h \in H$, $v_i \in V_i$, $i \in \mathbf{Z}$.

Consider the $\mathcal{G}$-module

$$M_\alpha(\lambda, V) = \mathcal{U}(\mathcal{G}) \bigotimes_{\mathcal{U}(B)} V$$

associated with α, λ, V.

Proposition 5.1.

(i) *The $\mathcal{G}$-module $M_\alpha(\lambda, V)$ is $S(B_{-\alpha})$- free.*

(ii) *$M_\alpha(\lambda, V)$ has a unique irreducible quotient $L_\alpha(\lambda, V)$.*

(iii) $P(M_\alpha(\lambda, V)) = (D(\lambda) - \{\lambda + n\delta \mid n \in \mathbf{Z}\}) \cup P(V) \subset D(\lambda)$.

(iv) *$M_\alpha(\lambda, V) \simeq M_{\alpha'}(\lambda', V')$ if and only if $\alpha' \in \{\alpha + n\delta \mid n \in \mathbf{Z}\}$ and there exists $i \in \mathbf{Z}$ such that $\lambda = \lambda' + i\delta$ and $V^{[i]} \simeq V'$ as graded G-modules.*

Proof. Follows from the construction of $\mathcal{G}$- module $M_\alpha(\lambda, V)$. □

Now we describe the classes of isomorphisms of irreducible modules in $\tilde{\mathcal{O}}(\alpha)$.

Proposition 5.2.

(i) *Let $\tilde{V}$ be an irreducible object in $\tilde{\mathcal{O}}(\alpha)$. Then there exist $\lambda \in H^*$ and an irreducible G- module V such that $\tilde{V} \simeq L_\alpha(\lambda, V)$.*

(ii) $L_\alpha(\lambda, V) \simeq L_\alpha(\lambda', V')$ *if and only if there exists* $i \in \mathbf{Z}$ *such that* $\lambda = \lambda' + i\delta$ *and* $V^{[i]} \simeq V'$ *as graded* G*-modules.*

Proof. One can see that $\tilde{V}$ contains a non-zero element $v \in \tilde{V}_\lambda$ such that $B_\alpha v = 0$. Then $V = \mathcal{U}(G)v$ is an irreducible $\mathbf{Z}$-graded G- module and $\tilde{V} \simeq \mathcal{U}(B_{-\alpha})V$. This implies that $\tilde{V}$ is a homomorphic image of $M_\alpha(\lambda, V)$ and, therefore, is isomorphic to $L_\alpha(\lambda, V)$, which proves (i). Part (ii) follows from Proposition 5.1, (iv). □

Lemma 5.3. *If* $0 < \dim L_\alpha(\lambda, V)_\mu < \infty$ *for some* $\mu \in H^*$ *then* $\dim V_i < \infty$ *for all* $i \in \mathbf{Z}$.

Proof. If $\lambda(c) = 0$ then $V^{[j]} \simeq L_{r,\Lambda}$ for some $r \geq 0$, $\Lambda \in P_r$, $j \in \mathbf{Z}$ by Proposition 4.6 and, hence $\dim V_i \leq 1$ for all $i \in \mathbf{Z}$. Let $\lambda(c) = a \in \mathbf{C}^*$ and $V^{[j]} \simeq M^\varepsilon(a)$, for any $j \in \mathbf{Z}$, $\varepsilon \in \{+, -\}$. By Proposition 4.3, (i), $\dim V_i = \infty$ for all i. If $a \in \mathbf{Q}_+$ ($a \notin \mathbf{Q}_+$ respectively) then $\lambda(h_\alpha) - na \notin \mathbf{Z}_+$ for all integer $n \geq n_0$ ($n \leq n_0$ respectively) and for some $n_0 \in \mathbf{Z}$. Thus, $e_{\alpha - n\delta} e_{-\alpha + n\delta}$ acts injectively on $L_\alpha(\lambda, V)$ for all $n \geq n_0$ ($n \leq n_0$ respectively) which implies that $\dim L_\alpha(\lambda, V)_\mu = \infty$. But, this contradicts the assumption. We conclude that $V^{[j]} \simeq M^\varepsilon(a)$ for some $j \in \mathbf{Z}$, $\varepsilon \in \{+, -\}$ and $\dim V_i < \infty$ for all $i \in \mathbf{Z}$. □

Theorem 5.4. *Let* $\tilde{V} \in \tilde{\mathcal{O}}(\alpha)$ *be an irreducible.*

(i) [**6**] *If* $\tilde{V}$ *is of level zero then* $\tilde{V} \simeq L_\alpha(\lambda, L_{r,\Lambda})$ *for some* $\lambda \in H^*$, $\lambda(c) = 0$, $r \geq 0$, $\Lambda \in P_r$.

(ii) *If* $\tilde{V}$ *is of level* $a \in \mathbf{C}^*$ *and* $\dim \tilde{V}_\mu < \infty$ *for at least one* $\mu \in P(\tilde{V})$ *then* $\tilde{V} \simeq L_\alpha(\lambda, M^\varepsilon(a))$ *for some* $\lambda \in H^*$, $\lambda(c) = a$, $\varepsilon \in \{+, -\}$.

Proof. (i) follows from Propositions 5.2 and 4.6, while (ii) follows from Lemma 5.3, Propositions 5.2 and 4.3. □

In some cases we can describe the structure of modules $L_\alpha(\lambda, V)$.

Let $\lambda(c) = 0$, $r = 0$, $\Lambda = \Lambda_0$, $L_{0,\Lambda_0} \simeq \mathbf{C}$. Set $\tilde{M}(\lambda) = M_\alpha(\lambda, \mathbf{C})$. Notice that $\tilde{M}(\lambda) \simeq S(B_{-\alpha})$ as vector spaces and, therefore, $P(\tilde{M}(\lambda)) = \{\lambda - n\alpha + k\delta \mid k, n \in \mathbf{Z},\ n > 0\} \cup \{\lambda\}$ and

$$\dim \tilde{M}(\lambda)_{\lambda - n\alpha + k\delta} = \infty, n > 1, \dim \tilde{M}(\lambda)_\lambda = \dim \tilde{M}(\lambda)_{\lambda - \alpha + k\delta} = 1, k \in \mathbf{Z}.$$

Proposition 5.5.

(i) $L_\alpha(\lambda, \mathbf{C}) \simeq \tilde{M}(\lambda)$ *if and only if* $\lambda(h_\alpha) \neq 0$.

(ii) *If* $\lambda(h_\alpha) = 0$ *then* $L_\alpha(\lambda, \mathbf{C})$ *is a trivial one-dimensional module.*

Proof. Proposition follows from [**7**, Proposition 6.2] and is also proved in [**8**]. □

Let $\lambda(c) = a \in \mathbf{C}^*$. Set $M^\varepsilon(\lambda, a) = M_\alpha(\lambda, M^\varepsilon(a))$. We have

$$P(M^\varepsilon(\lambda, a)) = \{\lambda - k\alpha + n\delta \mid k, n \in \mathbf{Z},\ k > 0\} \cup \{\lambda - \varepsilon n\delta \mid n \in \mathbf{Z}_+\}$$

and

$$\dim M^\varepsilon(\lambda, a)_{\lambda - k\alpha + n\delta} = \infty,\ k > 0, n \in \mathbf{Z}, \dim M^\varepsilon(\lambda, a)_{\lambda - \varepsilon n\delta} = P(n),\ n \in \mathbf{Z}_+.$$

Proposition 5.6. [**8**, **9**] $L_\alpha(\lambda, M^\varepsilon(a)) \simeq M^\varepsilon(\lambda, a)$.

Recall, that $\mathcal{G}$-module $\tilde{V}$ is called *integrable* if $e_{\pm\alpha}$ and $e_{\pm(\delta-\alpha)}$ act locally nilpotently on $\tilde{V}$. All irreducible integrable $\mathcal{G}$- modules in $\tilde{\mathcal{O}}(\alpha)$ of level zero were classified in [**6**]. In fact, they are the only integrable modules in $\tilde{\mathcal{O}}(\alpha)$.

Corollary 5.7. *If $\tilde{V}$ is irreducible integrable $\mathcal{G}$-module in $\tilde{\mathcal{O}}(\alpha)$ then $\tilde{V}$ is of level zero.*

Proof. Suppose $\tilde{V}$ is of level $a \neq 0$. Since $\tilde{V}$ is integrable, it follows from Proposition 5.6 that $\tilde{V} \neq L_\alpha(\lambda, M^\varepsilon(a))$, $\varepsilon \in \{+, -\}$. Then $\tilde{V} \simeq L_\alpha(\lambda, V)$ and for any $k \in \mathbf{Z}_+$ there exist $i > k$, $j < -k$ such that $V_i \neq 0$, $V_j \neq 0$. Now the same arguments as in the proof of Lemma 5.3 show that $e_{-\alpha}$ and $e_{\delta-\alpha}$ are not locally nilpotent on such module and, therefore, $\tilde{V}$ has a zero level. □

Remark. (i) The structure of modules $L_\alpha(\lambda, L_{r,\Lambda})$, $r > 0$ is unclear is general. Some examples were considered in [**1**, **12**].
(ii) Most of the results of Section 5 can be generalized for an arbitrary Affine Lie Algebra [**6**, **7**, **12**].

6. Non-dense $\mathcal{G}$-modules.

Definition. An irreducible weight $\mathcal{G}$-module V is called dense if $P(V) = \lambda + Q$ for some $\lambda \in H^*$ and non-dense otherwise.

In this section we classify all irreducible non-dense $\mathcal{G}$- modules with at least one finite-dimensional weight subspace. Our main result is the following Theorem.

Theorem 6.2. *If $\tilde{V}$ is an irreducible non-dense $\mathcal{G}$-module with at least one finite-dimensional weight subspace then $\tilde{V}$ belongs to one of the following disjoint classes:*

(i) *highest weight modules with respect to some choice of π;*

(ii) $L^\varepsilon_\alpha(\lambda, \gamma)$, $\alpha \in \Delta^{re}$, $\lambda \in H^*$, $\gamma \in \mathbf{C} - R_\lambda$, $\varepsilon \in \{+, -\}$;

(iii) $L_\alpha(\lambda, L_{r,\Lambda})$, $\alpha \in \Delta^{re}$, $\lambda \in H^*$, $\lambda(c) = 0$, $r \geq 0$, $\Lambda \in P_r$.

(iv) $L_\alpha(\lambda, M^\varepsilon(a))$, $\alpha \in \Delta^{re}$, $\lambda \in H^*$, $a \in \mathbf{C}^*$, $\lambda(c) = a$, $\varepsilon \in \{+,-\}$.

Moreover, we can describe the irreducible $\mathcal{G}$-modules of non-zero level with finite-dimensional weight subspaces.

Theorem 6.3. *Let $\tilde{V}$ be an irreducible $\mathcal{G}$-module of level $a \neq 0$ with all finite-dimensional weight subspaces. Then $\tilde{V} \simeq L^\varepsilon_\alpha(\lambda, \gamma)$ for some $\alpha \in \Delta^{re}$, $\lambda \in H^*$, $\lambda(c) = a$, $\gamma \in \mathbf{C}$, $\varepsilon \in \{+,-\}$.*

Remark 6.4. Theorems 6.2, 6.3 imply that in order to complete the classification of all weight irreducible $\mathcal{G}$-modules one has to study the following classes:

(i) Modules of type $L_\alpha(\lambda, V)$ where V is a graded irreducible G-module of non-zero level with all infinite- dimensional components.

(ii) Dense $\mathcal{G}$-modules of zero level.

(iii) Dense $\mathcal{G}$-modules of non-zero level with an infinite-dimensional weight subspace.

These classification problems are still open.

The proof of Theorem 6.2 is based on some preliminary results. We start with the following Definition.

Definition 6.5. A subset $P \subset \Delta$ is called closed if $\beta_1, \beta_2 \in P$, $\beta_1 + \beta_2 \in \Delta$ imply $\beta_1 + \beta_2 \in P$. A closed subset $P \subset \Delta$ is called a partition if $P \cap -P = \emptyset$, $P \cup -P = \Delta$.

Lemma 6.6. *Let P be a partition, $P \ni \delta$, $P^{re} = P \cap \Delta^{re}$, $\beta \in \Delta^{re}$.*

(i) *If $\mid P^{re} \cap \{\beta + k\delta \mid k \in \mathbf{Z}_+\} \mid < \infty$ or $\mid P^{re} \cap \{-\beta + k\delta \mid k \in \mathbf{Z}\} \mid < \infty$ then $P^{re} = \{\varphi + n\delta \mid n \in \mathbf{Z}\}$ for some $\varphi \in \Delta^{re}$.*

(ii) *If $\mid P^{re} \cap \{\beta + k\delta \mid k \in \mathbf{Z}\} \mid = \mid P^{re} \cap \{-\beta + k\delta \mid k \in \mathbf{Z}_+\} \mid = \infty$ then $P = \Delta_+(\tilde{\pi})$ for some basis $\tilde{\pi}$ of Δ.*

Proof. Recall that $\Delta = \{\pm\beta + k\delta \mid k \in \mathbf{Z}\} \cup \{n\delta \mid n \in \mathbf{Z} - \{0\}\}$. It follows from [7] that there exist $w \in W$ and $\beta' \in \Delta^{re}$ such that

$$wP = \{\beta' + k\delta \mid k \in \mathbf{Z}\} \cup \{k\delta \mid k > 0\}$$

or

$$wP = \{\beta' + n\delta, -\beta' + k\delta \mid n \geq 0,\ k > 0\} \cup \{k\delta \mid k > 0\} = \Delta_+(\pi')$$

where $\pi' = \{\beta', \delta - \beta'\}$. Then

$$P = \{w^{-1}\beta' + k\delta \mid k \in \mathbf{Z}\} \cup \{k\delta \mid k > 0\}$$

or $P = \Delta_+(w^{-1}\pi')$. This implies the statement of Lemma. □

Definition 6.7. A non-zero element v of a $\mathcal{G}$-module V is called admissible if $\mathcal{N}_\varphi^\varepsilon v = 0$ or $B_\varphi v = 0$, for some $\varphi \in \Delta^{re}$, $\varepsilon \in \{+,-\}$.

Lemma 6.8. *If the $\mathcal{G}$-module V contains a non-zero vector $v \in V_\lambda$ such that $e_\varphi v = 0$ and $\lambda + k\delta \notin P(V)$ for some $\varphi \in \Delta^{re}$, $k \in \mathbf{Z} - \{0\}$ then V contains an admissible vector.*

Proof. We will assume that $k > 0$. The case $k < 0$ can be considered analogously. We prove the Lemma by the induction on k. Let $k = 1$. Then we have $e_{\varphi+m\delta}v = e_\delta v = 0$ for all $m \geq 0$. If $e_{\varphi-i\delta}v = 0$ for all $i > 0$ then $B_\varphi v = 0$ and v is admissible. Let $e_{\varphi-n\delta}v \neq 0$ for some $n > 0$ and $e_{\varphi-i\delta}v = 0$, $0 \leq i < n$. Set $\tilde{v} = e_{\varphi-n\delta}v \neq 0$. Then $e_{\varphi-i\delta}\tilde{v} = e_\delta\tilde{v} = e_{-\varphi+(n+1)\delta}\tilde{v} = 0$, $i < n$ and, thus, $e_\psi\tilde{v} = 0$ for any $\psi \in \tilde{P} = \{\varphi - i\delta, -\varphi + (n+j+1)\delta, (j+1)\delta \mid i < n, j \geq 0\}$. One can see that $\tilde{P} \cup \{-\varphi + n\delta\}$ is a partition and $\tilde{P} = \Delta_+(\tilde{\pi}) - \{\varphi'\}$ for some $\varphi' \in \Delta^{re}$, $\tilde{\pi} = \{\varphi', \delta - \varphi'\}$, by Lemma 6.6. Hence, $\mathcal{N}_{\varphi'}^+\tilde{v} = 0$ which proves the Lemma for $k = 1$.

Assume now that the Lemma is proved for all $0 < k' < k$ and consider two cases:

(i) There exists $n \in \mathbf{Z}$, $0 < n < k$ such that $e_{\varphi+i\delta}v = 0$ for all $0 \leq i < n$ but $e_{\varphi+n\delta}v \neq 0$. Then $e_{\varphi+i\delta}\tilde{v} = e_{-\varphi+(k-n)\delta}\tilde{v} = 0$, $0 \leq i < n$ where $\tilde{v} = e_{\varphi+n\delta}v$ and $e_{-\varphi+(k-n)\delta}\tilde{v} \in V_{\lambda+k\delta} = 0$. If $k - n = 1$ or $k - n > 1$ and $e_{-\varphi+\delta}\tilde{v} = 0$ then $\mathcal{N}_+v = 0$ and $\tilde{v}$ is admissible. Let $k - n > 1$ and $v' = e_{-\varphi+\delta}\tilde{v} \neq 0$. Then $v' \in V_{\lambda'}$, $e_{\varphi'}v' = 0$, $\lambda' + (k-n-1)\delta \notin P(V)$ where $\lambda' = \lambda + (n+1)\delta$, $\varphi' = -\varphi + (k-n)\delta$ and V has an admissible element by the induction hypotheses.

(ii) Let $e_{\varphi+i\delta}v = 0$ for all $0 \leq i \leq k$. Since $e_{k\delta}v = 0$ we have $e_{\varphi+i\delta}v = 0$ for all $i \geq 0$. If $\tilde{v}_m = e_{m\delta}v \neq 0$ for some $0 < m < k$ then $\tilde{v}_m \in V_{\lambda'}$, $\lambda' = \lambda + m\delta$, $e_\varphi\tilde{v}_m = 0$, $\lambda' + (k-m)\delta \notin P(V)$ and we can apply induction. Assume that $\tilde{v}_m = 0$ for all $0 < m < k$. Then we have $e_{\varphi+i\delta}v = e_{m\delta}v = 0$, $i \geq 0$, $0 < m \leq k$. If $e_{\varphi-j\delta}v = 0$ for all $j > 0$ then $B_\varphi v = 0$ and v is admissible. Otherwise, let n be a minimal positive integer such that $\tilde{v} = e_{\varphi-n\delta}v \neq 0$. Then $e_{\varphi-j\delta}\tilde{v} = e_{-\varphi+(n+k)\delta}\tilde{v} = e_{i\delta}\tilde{v} = 0$, $i \geq 0$, $j < n$. Assume that $e_{-\varphi+(n+1)\delta}\tilde{v} = 0$. We have $e_\psi\tilde{v} = 0$ for any $\psi \in \tilde{P} = \{\varphi - j\delta, -\varphi + (n+m)\delta, m\delta \mid j < n, m > 0\}$. The set $\tilde{P} \cup \{-\varphi + n\delta\}$ is a partition, $\mid \tilde{P}^{re} \cap \{\varphi + i\delta \mid i \geq 0\} \mid = \mid \tilde{P}^{re} \cap \{-\varphi + i\delta \mid i > 0\} \mid = \infty$ and, therefore, $\tilde{P} = \Delta_+(\tilde{\pi}) - \{\varphi'\}$ for some $\varphi' \in \Delta^{re}$, $\tilde{\pi} = \{\varphi', \delta - \varphi'\}$ by Lemma 6.6. We conclude that $\mathcal{N}_{\varphi'}^+\tilde{v} = 0$ and $\tilde{v}$ is admissible. Finally, suppose that $v' = e_{-\varphi+(n+1)\delta}\tilde{v} \neq 0$. Then $v' \in V_{\lambda'}$, $e_\varphi v' = 0$, $\lambda' + (k-1)\delta \notin P(V)$ where λ' stands for $\lambda + \delta$ and, thus V has an admissible element by the assumption of induction. This completes the proof of Lemma. □

Proposition 6.9. *Let V be an irreducible non-dense $\mathcal{G}$-module. Then V contains an admissible element.*

Proof. Let $\lambda \in P(V)$ and $\lambda + \varphi \notin P(V)$ for some $\varphi \in \Delta$. We can assume that $\varphi \in \Delta^{re}$. Indeed, let $\varphi = \delta$. If $e_\alpha v = e_{\delta-\alpha} v = 0$ for some $0 \neq v \in V_\lambda$, $\alpha \in \Delta^{re}$ then V is a highest weight module with respect to $\{\alpha, \delta - \alpha\}$ and v is admissible. If, for example, $e_\alpha v \neq 0$ then $\lambda' = \lambda + \alpha \in P(V)$ and $\lambda' + (\delta - \alpha) \notin P(V)$. Hence, we can assume that $\lambda + \varphi \notin P(V)$, $\varphi \in \Delta^{re}$. Let $0 \neq v \in V_\lambda$. If $v' = e_{\varphi - n\delta} v \neq 0$ for some $n \in \mathbf{Z} - \{0\}$ then $e_\varphi v' = 0$, $v' \in V_{\tilde{\lambda}}$, $\tilde{\lambda} = \lambda + \varphi - n\delta$, $\tilde{\lambda} + n\delta \notin P(V)$ and Proposition follows from Lemma 6.8. If $e_{\varphi - n\delta} v = 0$ for all $n \in \mathbf{Z}$ then $B_\varphi v = 0$ and v is admissible. □

Corollary 6.10. *If $\tilde{V}$ is an irreducible non-dense $\mathcal{G}$-module then either $\tilde{V} \simeq L^\varepsilon_\alpha(\lambda, \gamma)$ or $\tilde{V} \simeq L_\alpha(\lambda, V)$ for some $\alpha \in \Delta^{re}$, $\lambda \in H^*$, $\gamma \in \mathbf{C}$, $\varepsilon \in \{+, -\}$ and irreducible G- module V.*

Proof. Follows from Proposition 6.9, Corollary 3.3 (i) and Proposition 5.2. □

Now Theorem 6.2 follows from Corollary 6.6 and Theorem 5.4.

Proof of Theorem 6.3. Let $\mu \in P(\tilde{V})$. Consider the $\mathcal{G}$-submodule $V = \mathcal{U}(G)\tilde{V}_\mu \subset \tilde{V}$. Then it follows from Proposition 4.5 that V is completely reducible and moreover each irreducible component is isomorphic to $M^\varepsilon(a)$, $\varepsilon \in \{+, -\}$ up to a shifting of gradation by Proposition 4.3, (i). Denote by V^+ the sum of all irreducible components of V isomorphic to $M^+(a)$ and assume that $V^+ \neq 0$. Let $0 \neq v \in V^+ \cap \tilde{V}_\chi$, $\chi \in P(\tilde{V})$ and $V^+ \cap \tilde{V}_{\chi+\delta} = 0$. We will show that for any $\alpha \in \Delta^{re}$ there exists $m_\alpha \in \mathbf{Z}_+$ such that $e_{\alpha + m\delta} v = 0$ for all $m \geq m_\alpha$. Indeed, let $v_0 = e_\alpha v \neq 0$. Consider the G-module $\mathcal{U}(G)v_0$ which is again completely reducible by Proposition 4.5. If $e_{k\delta} v \neq 0$ for all $k > 0$ then $v_k = e^k_\delta v_0 \neq 0$ for all $k > 0$. But, for big enough k, v_k will belong to the direct sum of irreducible components of $\mathcal{U}(G)v_0$ each of which is isomorphic to $M^-(a)$ up to a shifting of gradation. This contradicts Proposition 4.1, (ii), since $e^2_\delta v_k = 2^{k+2} e_{\alpha + (k+2)\delta} v = 2 e_{2\delta} v_k$. Thus, there exists $m_\alpha \geq 0$ such that $e_{\alpha + m_\alpha \delta} v = 0$ and, therefore, $e_{\alpha + m\delta} v = 0$ for any $m \geq m_\alpha$.

Suppose that $\chi + \delta \in P(\tilde{V})$. Since $\tilde{V}$ is irreducible there exists $0 \neq u \in \mathcal{U}(\mathcal{G})$ such that $0 \neq uv \in \tilde{V}_{\chi+\delta}$. It follows from the discussion above that $e_{n\delta} uv = 0$ for big enough $n \in \mathbf{Z}_+$. The G-submodule $V' = \mathcal{U}(G)uv$ is completely reducible by Proposition 4.5 and since $V^+ \cap \tilde{V}_{\chi+\delta} = 0$, any irreducible component $L \subset V'$ such that $L \cap \tilde{V}_{\chi+\delta} \neq 0$ is isomorphic to $M^-(a)$ up to a shifting of gradation. Hence, $e_{n\delta} \tilde{v} \neq 0$ for any non-zero $\tilde{v} \in V' \cap \tilde{V}_{\chi+\delta}$ by Proposition 4.1, (ii) and $e_{n\delta} uv \neq 0$ in particular. This contradiction implies that $\chi + \delta \notin P(\tilde{V})$ and therefore $\tilde{V}$ is a non-dense

$\mathcal{G}$-module. Applying Theorem 6.2 we conclude that $\tilde{V} \simeq L^{\varepsilon}_{\alpha}(\lambda, \gamma)$ for some $\alpha \in \Delta^{re}$, $\lambda \in H^*$, $\lambda(c) = a$, $\gamma \in \mathbf{C}$, $\varepsilon \in \{+, -\}$ which completes the proof. □

Acknowledgement.

The author gratefully acknowledges the support of the Natural Sciences and Engineering Research Council of Canada.

References

[1] S. Spirin, $\mathbf{Z}^2$-*graded modules with one-dimensional components over the Lie Algebra* $A_1^{(1)}$, Funkts. Anal. Prilozhen., **21** (1987), 84-85.

[2] D. Britten, F. Lemire and F. Zorzitto, *Pointed torsion free Modules of Affine Lie Algebras*, Commun. in Algebra, **18** (1990), 3307-3321.

[3] V. Chari and A. Pressley, *A new family of irreducible, integrable modules for Affine Lie Algebras*, Math. Ann., **277** (1987), 543-562.

[4] V. Kac, *Infinite Dimensional Lie Algebras*, 3rd Edition, Cambridge University Press 1990.

[5] J. Lepowsky, *Generalized Verma modules, loop space cohomology and MacDonald type identities*, Ann. Sci. Ecole Norm. sup., **12** (1979), 169-234.

[6] V. Chari, *Integrable representations of Affine Lie Algebras*, Invent. Math., **85** (1986), 317-335.

[7] H. Jakobsen and V. Kac, *A new class of unitarizable highest weight representations of infinite dimensional Lie algebras*, Lecture Notes in Physics, **226** (1985), 1-20.

[8] V. Futorny, *On Imaginary Verma modules over the affine Lie algebra* $A_1^{(1)}$, Oslo University preprint, 1991-9.

[9] ———, *Irreducible graded* $A_1^{(1)}$*- modules*, Funkts. Anal. Prilozhen., **26** (1992), 73-75.

[10] V. Kac and D. Kazhdan, *Structure of representations with highest weight of infinite-dimensional Lie algebras*, Advances in Math., **34** (1979), 97-108.

[11] N. Shapovalov, *On bilinear form on the universal enveloping algebra of a complex semisimple Lie algebra*, Funkts. Anal. Prilozhen., **6** (1972), 307-312.

[12] V. Futorny, *Imaginary Verma modules for Affine Lie Algebras*, Canad. Math. Bull., to appear.

Received July 12, 1993 and revised December 2, 1993.

KIEV UNIVERSITY
KIEV, 252617, UKRAINE
E-mail address: FUTORNY@UNI-ALG.KIEV.UA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

M-HYPERBOLIC REAL SUBSETS OF COMPLEX SPACES

GIULIANA GIGANTE, GIUSEPPE TOMASSINI AND SERGIO VENTURINI

The aim of this paper is to make a first attempt to study real analytic subsets of complex manifolds (or more generally of complex analytic spaces) from the viewpoint of the theory of metric spaces.

1. Introduction.

Our starting point was inspired by the definition of the so-called Kobayashi pseudodistance on complex manifolds We recall briefly that such a pseudodistance is defined on any complex analytic space M using only the space of all holomorphic maps sending the open unit disk Δ in $\mathbb{C}$ in the space M. Moreover the complex space M is said to be "hyperbolic" if such a pseudodistance actually is a real distance, namely it assigns non vanishing values to pair of distinct points of M. In our situation, we introduce a similar pseudodistance $d_{V,M}$ on any subset of V of a complex analytic space M using the space of all holomorphic maps from Δ to M sending the open interval $I =]-1, 1[$ in V, and we introduce the concept of M-hyperbolicity (cf. Section 2).

We are primarily interested in the case when M is a smooth complex manifold and V is a (closed) real analytic smooth submanifold of M, but the definitions work in this more general context as well.

Any holomorphic map between complex manifolds is distance decreasing when the manifolds are endowed with the Kobayashi distances. Our pseudodistances also fulfill this fundamental property. A unexpected phenomenon is that there are some classes of non holomorphic mappings which enjoy this property. A description of such mappings is given in the Section 3 of the paper. As an application, some hyperbolicity criteria are given, and some Liouville type theorems are proved.

We also extend the construction of the Kobayashi-Royden pseudometric when V is a smooth real analytic submanifold of a complex manifold M (Section 4) and we establish some results on the behaviour of a complex Lie group G acting holomorphically on M and leaving V invariant (Section 5). Moreover we define and study the "geodesics" for such a metric Some examples are given (Section 6).

2. Main definitions.

Let us fix some notations. We denote by I the open real interval $]-1,1[$, and by D the open unit disk in $\mathbb{C}$. The Poincaré hyperbolic distance on D will be denoted by ρ.

We denote by $D(R)$, $0 < R \leq +\infty$, the set of complex number z such that $|z| < R$, and also put $I(R) = D(R) \cap \mathbb{R}$.

Let M be a complex analytic (reduced) complex space and let V be a subset of M. By an *M-analytic arc* in V, or simply an analytic arc in V, we mean a holomorphic map $f : D \to M$ such that $f(I) \subset V$. Given two points p and q in V, an *analytic chain* γ in V joining p and q is given by the following data:

(i) points $a_0, \ldots, a_k$ in I;

(ii) M-analytic arcs $f_1, \ldots, f_k$ in V such that $f_1(a_0) = p$, $f_k(a_k) = q$ and $f_j(a_j) = f_{j+1}(a_j)$ for $j = 1, \ldots, k-1$.

The length of the analytic chain γ is by definition the number

$$\rho(\gamma) = \sum_{j=0}^{k-1} \rho(a_j, a_{j+1}).$$

We denote by $C_{p,q}(V, M)$ the set af all the M-analytic chains in V joining p and q.

Using the analytic arcs so defined we introduce a pseudodistance on V by the formula

$$d_{V,M}(p,q) = \inf\{\rho(\gamma) \mid \gamma \in C_{p,q}(V,M)\},$$

where by definition the second member in the definition is $+\infty$ if the set $C_{p,q}(V,M)$ is empty.

Clearly the function $d_{V,M}(p,q)$ so defined is a pseudodistance that vanishes when $p = q$, it is symmetric in p and q, and satisfies the triangle inequality.

We say that V is *hyperbolic* with respect to M, or simply *M-hyperbolic* if $d_{V,M}(p,q) > 0$ whenever $p \neq q$.

On the other hand we say that V is *M-hyperbolically flat*, or simply *M-flat*, if the pseudodistance $d_{V,M}$ vanishes identically.

In this paper we are interest in the case when V is a real analytic subset (even a real analytic submanifold) of M. Nevertheless the definition makes sense with no additional structure on V.

We begin by noting some elementary properties:

(i) If $V = M$, then $d_{V,M}$ is the usual Kobayashi pseudodistance on M;

(ii) If $M = D$ and $V = I$, then the Schwarz Lemma implies that the pseudodistance $d_{V,M}$ is the restriction to I of the Poincaré distance on D;

(iii) If (V_1, M_1) and (V_2, M_2) are pairs of complex spaces as above and $f : M_1 \to M_2$ is a holomorphic map sending V_1 in V_2, then for every p and q in V_1

$$d_{M_2,V_2}(f(p), f(q)) \leq d_{M_1,V_1}(p, q);$$

(iv) If $\delta : V \times V \to [0, +\infty]$ is a pseudodistance such that

$$\delta(f(t), f(t)) \leq \rho(t, s)$$

for all M-analytic arcs f in V then $\delta \leq d_{V,M}$.

(v) If $M = \mathbb{C}$ and $V = \mathbb{R}$ then $d_{V,M}$ vanishes identically, that is, $\mathbb{R}$ is $\mathbb{C}$-flat; indeed, given $y \in \mathbb{R}$, let f be the analytic arc $z \mapsto nyz$, $n \in \mathbb{N}$; then $f(0) = 0$, $f(1/n) = y$ and hence

$$d_{V,M}(0, y) \leq \rho(0, 1/n).$$

Taking the limit for $n \to +\infty$ we obtain $d_{V,M}(0, y) = 0$.

3. Hyperbolicity and "good" mappings.

We say that an arbitrary map $F : M_1 \to M_2$ between complex spaces is *good*, if, for every holomorphic map $f : D(R) \to M_1$, there exists a holomorphic map $\tilde{f} : D(R) \to M_2$ such that $\tilde{f}(t) = F(f(t))$ for every $t \in I(R)$.

The proofs of the following two Propositions are straightforward.

Proposition 3.1. *Let M_1 and M_2 be complex spaces, V_1 and V_2 be subsets of M_1 and M_2 respectively, and let $F : M_1 \to M_2$ be a good map satisfying $F(V_1) \subset V_2$. Then, for every pair of points p and q in V_1,*

$$d_{V_2,M_2}(F(p), F(q)) \leq d_{V_1,M_1}(p, q).$$

Proposition 3.2. *Let M_1, M_2, V_1, V_2 and F as in the previous Proposition.*

(i) *If V_2 is M_2-hyperbolic and $F_{|V_1}$ is injective, then V_1 is M_1-hyperbolic.*

(ii) *If V_1 is M_1-flat and $F(V_1) = V_2$, then V_2 is M_2-flat.*

Every holomorphic map is clearly good. However there also are not holomorphic good maps:

Proposition 3.3. *The map $F : \mathbb{C}^n \to \mathbb{C}^{2n}$ defined by*

$$z = (z_1, \ldots, z_n) \mapsto F(z) = (z_1, \bar{z_1}, \ldots, z_n, \bar{z_n})$$

is good.

Proof. Let $f : D(R) \to \mathbb{C}^n$ be a holomorphic map. Define $f^* : D(R) \to \mathbb{C}^n$ by the formula

$$f^*(z) = f(\bar{z})), \quad z \in D(R).$$

Clearly f^* is holomorphic and the map $\tilde{f} : D(R) \to \mathbb{C}^{2n}$ given by

$$\tilde{f}(z) = (f_1(z), f_1^*(z), \dots, f_n(z), f_n^*(z)),$$

where f_i and f_i^* are the i-th component respectively of f and f^*, satisfies $\tilde{f}(t) = F(f(t))$ for every $t \in I(R)$. □

Since compositions of good maps are good we immediatly obtain

Proposition 3.4. *Let $H : \mathbb{C}^{2n} \to M$ be a holomorphic (or simply a good) map. Then the maps $F, G : \mathbb{C}^n \to M$ defined by*

$$F(z_1, \dots, z_n) = H(z_1, \bar{z}_1, \dots, z_n, \bar{z}_n),$$
$$G(x_1 + iy_1, \dots, x_n + iy_n) = H(x_1, y_1, \dots, x_n, y_n),$$

are good.

For the projective space we have:

Proposition 3.5. *The map $F : \mathbb{CP}^n \to \mathbb{CP}^\nu$, $\nu = (n+1)^2 - 1$, defined by*

$$w_{ij} = z_i \bar{z}_j, \quad i, j = 0, \dots, n \tag{3.1}$$

is good.

Proof. The assertion follows from the Propositions 3.3 and 3.4, and the fact that for every k there is a one to one correspondence between holomorphic maps $f : D \to \mathbb{CP}^k$ and holomorphic maps $g = (g_0, \dots, g_k) : D(R) \to \mathbb{C}^{k+1}$ satisfying $g_i \not\equiv 0$ for some $i = 0, \dots, k$. □

In order to find hyperbolic spaces the following (almost trivial) remark is useful.

Proposition 3.6. *Let M be a complex space and let V be a subset of M. If N is a closed complex subspace of M containing V, then*

$$d_{V,M} = d_{V,N}.$$

In particular, if N is hyperbolic (as complex space), then V is M-hyperbolic.

Proof. It suffices to show that if $f : D \to M$ is a holomorphic arc in V, then $f(D) \subset N$, that is $f^{-1}(N) = D$. But this is obvious, since $f^{-1}(N)$ is a closed complex subspace of D containing I, and any such a subspace must coincide with D. □

We now give some example of flat spaces.

Proposition 3.7. *Any interval of the real line is flat.*

Proof. It suffices to prove the assertion for the interval $J = [0, 1]$. Indeed, the map $z \mapsto \exp(-z^2)$ shows that $d_{J,\mathbb{C}}(1, t) = 0$ for every $t \in]0, 1]$. Analogously, the map $z \mapsto 1 - \exp(-z^2)$ yields $d_{J,\mathbb{C}}(t, 0) = 0$ for every $t \in [0, 1[$. Finally, one has $d_{J,\mathbb{C}}(1, 0) \leq d_{J,\mathbb{C}}(1, 1/2) + d_{J,\mathbb{C}}(1/2, 0) = 0$. □

As consequence of this Proposition we obtain the following Liouville type Theorem.

Theorem 3.1. *Let V be a subset of a complex space M. If V is M-hyperbolic then every holomorphic map $f : \mathbb{C} \to M$ sending some non-trivial real interval $J \subset \mathbb{R}$ in V is a constant map.*

Proof. Since V is M-hyperbolic and J is $\mathbb{C}$-flat the map f must be constant on J and hence it is constant on all $\mathbb{C}$. □

Other examples of flat space are given in the following three Propositions.

Proposition 3.8. *Any connected subset of a non-singular real conic in $\mathbb{C} = \mathbb{R}^2$ is flat.*

Proof. Since real affine self map of $\mathbb{C}$ are good, any conic is isometric either to the unit circle $x^2 + y^2 = 1$, or to the equilateral hyperbola $xy = 1$, or to the parabola $y = x^2$. Any connected subset of such a conic is the image of an interval of some real line in $\mathbb{C}$ under the maps $z \mapsto \cos(z) + i\sin(z)$, $z \mapsto \exp(z) + i\exp(-z)$, and $z \mapsto z + iz^2$ respectively. □

Proposition 3.9. *The boundary S of the unit ball in $\mathbb{C}^n$ (with respect to the standard euclidean norm) is flat.*

Proof. Given two arbitrary distinct points p and q in S, the complex line L joining p and q intersect S along a circumference, that is, a conic in L, and hence, by the previous Proposition, one has

$$d_{S,\mathbb{C}^n}(p, q) \leq d_{S\cap L,L}(p, q) = 0,$$

and the assertion follows. □

Proposition 3.10. *Every real ellipsoid in $\mathbb{C}^n$ is flat.*

Proof. Indeed the unit ball of $\mathbb{C}^n$ can be mapped onto any real ellipsoid under a suitable real linear map of $\mathbb{C}^n$, and any such map is good. □

And now here are some examples of hyperbolic sets. The following Proposition is immediate consequence of Propositions 3.4 and 3.6.

Proposition 3.11. *Let $V \subset \mathbb{C}^n = \mathbb{R}^{2n}$ be a subset defined by k equations*

$$f_i(x_1, y_1, \dots, x_n, y_n) = 0, \quad i = 1, \dots, k, \tag{3.2}$$

where x_i and y_i are the standard real coordinates in $\mathbb{C}^n$, and $f_1, \dots, f_k$ are real analytic functions defined by real power series converging over all $\mathbb{R}^{2n}$. Let $V_{\mathbb{C}}$ be the subset of $\mathbb{C}^{2n}$ defined by the same set of equations 3.2, *where now x_i and y_i represent the complex coordinates of $\mathbb{C}^{2n}$. Under these hypotheses, if $V_{\mathbb{C}}$ is hyperbolic (as complex space) then V is $\mathbb{C}^n$-hyperbolic.*

Example. Let $z = x + iy$ be the standard coordinate in $\mathbb{C}$. Let $V \subset \mathbb{C}$ the graph of the real function $y = \log(1 + x^2)$. Then V is $\mathbb{C}$-hyperbolic. Indeed according to the previous Proposition it suffices to prove that the complex curve

$$V_{\mathbb{C}} = \{(z, w) \in \mathbb{C}^2 \mid \exp(w) = 1 + z^2\}$$

is hyperbolic. Clearly $V_{\mathbb{C}}$ is regular everywhere, that is it is a closed Riemann surface in $\mathbb{C}^2$. Let denote by $g : V_{\mathbb{C}} \to \mathbb{C}$ the restriction to $V_{\mathbb{C}}$ of the projection map $(z, w) \mapsto z$. The map g is a non constant holomorphic map on $V_{\mathbb{C}}$. Since the exponential function never vanishes, then the map g necessarily omits the values i and $-i$, the zeroes of the function $1 + z^2$. The little Picard Theorem therefore implies that the universal covering of $V_{\mathbb{C}}$ can not be the complex plane, and hence $V_{\mathbb{C}}$ is covered by the unit disc D, that is, $V_{\mathbb{C}}$ must be hyperbolic, as asserted.

The following assertion gives a criterion for $\mathbb{CP}^1$-hyperbolicity.

Proposition 3.12. *Let $V \subset \mathbb{C} \subset \mathbb{CP}^1$ be a subset defined by an equation*

$$f(x, y) = 0, \quad z = x + iy \in \mathbb{C}, \tag{3.3}$$

where $f(x, y)$ is a polynomial in the variables x and y of degree d. Let $\bar{V}$ be the (topological) closure of V in $\mathbb{CP}^1$. Let $V_{\mathbb{C}}$ be the complex curve in $\mathbb{C}^2$ of equation 3.3, *where now x and y are considered as complex coordinates in $\mathbb{C}^2$, and finally let $\bar{V}_{\mathbb{C}}$ be the closure of $V_{\mathbb{C}}$ in $\mathbb{CP}^2$. If $\bar{V}_{\mathbb{C}}$ is hyperbolic then $\bar{V}$ (and hence V also) is $\mathbb{CP}^1$-hyperbolic.*

Proof. Let z_0 and z_1 be homogeneous coordinates in $\mathbb{CP}^1$, that is, $z = x + iy = z_1/z_0$.

Let $g \in \mathbb{C}[X_0, X_1, X_2]$ be the homogeneous polynomial defined by the equation

$$g(X_0, X_1, X_2) = X_0^d f\left(\frac{1}{2X_0}(X_1 + X_2), \frac{1}{2iX_0}(X_1 - X_2)\right).$$

Choosing X_0, X_1, X_2 and X_3 as homogeneous coordinates in $\mathbb{CP}^3$, let W be the quasiprojective algebraic subset of $\mathbb{CP}^3$ defined by

$$\begin{cases} g(X_0, X_1, X_2) = 0 \\ X_0 X_3 - X_1 X_2 = 0 \\ X_0 \neq 0 \end{cases},$$

and let $\bar{W}$ be the closure in $\mathbb{CP}^3$ of W.

Consider now the map $F : \mathbb{CP}^1 \to \mathbb{CP}^3$ defined by

$$\begin{cases} X_0 = z_0 \bar{z}_0 \\ X_1 = \bar{z}_0 z_1 \\ X_2 = z_0 \bar{z}_1 \\ X_3 = z_1 \bar{z}_1 \end{cases}$$

Such a map is injective, and Proposition 3.5 says that the map F so defined is a good map. By construction one clearly has $F(V) \subset W$ and hence $F(\bar{V}) \subset \bar{W}$. By Proposition 3.2, in order to check the $\mathbb{CP}^1$-hyperbolicity of $\bar{V}$, it suffices to prove that the curve $\bar{W}$ is hyperbolic.

It is easy to show that W and $V_{\mathbb{C}}$ are isomorphic (as affine algebraic varieties), and therefore $\bar{W}$ and $\bar{V}_{\mathbb{C}}$ are birationally equivalent as projective algebraic curves. Since hyperbolicity is preserved under birational isomorphisms between (compact algebraic) curves, it follows from our hypotheses that $\bar{W}$ is hyperbolic, as asserted. □

For algebraic varieties of higher dimension hyperbolicity is no longer a birational invariant. So the previous argument does not apply to higher dimensional projective spaces. Nevertheless the following Liouville type Theorem for meromorphic mappings holds:

Proposition 3.13. *Let $V \subset \mathbb{C}^n$ be a subset defined by k equations as in the Proposition* 3.11. *Let $V_{\mathbb{C}}$ be defined as in Proposition* 3.11, *and let $\bar{V}_{\mathbb{C}}$ be the closure in $\mathbb{CP}^{2n}$ of $V_{\mathbb{C}}$. Let $f_1, \ldots, f_n : \mathbb{C} \to \mathbb{C}$ be meromorphic functions. Assume that*

(i) *there exists a non-degenerate interval $J \subset \mathbb{R}$ such that every f_i has no poles on J and $(f_1(t), \ldots, f_n(t)) \in V$ for every $t \in J$;*

(ii) *the complex space $\bar{V}_{\mathbb{C}}$ is hyperbolic.*

Then every f_i is a constant function.

Proof. Let P be the set of all the poles of the functions f_i. The set P is discrete and closed in $\mathbb{C}$. It is easy to check that, for every $i = 1, \ldots, n$, the

function $f_i^*(z) = f(\bar{z}))$ is an entire meromorphic function, and the mapping $F : \mathbb{C} \setminus P \to \mathbb{C}^{2n}$ defined by

$$F(z) = \Big(\frac{1}{2}(f_1(z) + f_1^*(z)), \frac{1}{2i}(f_1(z) - f_1^*(z)), \dots, \frac{1}{2}(f_n(z) + f_n^*(z)), \frac{1}{2i}(f_n(z) - f_n^*(z))\Big)$$

is a holomorphic map sending the real interval J in $V_{\mathbb{C}}$. Since $V_{\mathbb{C}}$ is closed in $\mathbb{CP}^{2n}$, it follows that $F(\mathbb{C} \setminus P) \subset V_{\mathbb{C}}$. Moreover, by hypothesis, $\bar{V}_{\mathbb{C}}$ is a compact hyperbolic complex space. Thus the map F extends throughout all $\mathbb{C}$ (cf. Corollary 3.2. of Chapter VI of [**4**]). Again by the hyperbolicity of $V_{\mathbb{C}}$, the map F must be constant, and this yields our assertion. □

The following Proposition follows immediatly from [**8**, Theorem 3].

Proposition 3.14. *Let M be a complex manifold and let V be a subset of M. Assume that there exists a bounded plurisubharmonic function $u : M \to \mathbb{R}$ of class C^2. If u is strictly plurisubharmonic at every point of V, then V is M-hyperbolic.*

4. Real analytic submanifolds.

In this section we assume that M is a (connected) complex manifold and $V \subset M$ is a (connected) closed real analytic submanifold of M.

Proposition 4.1. *Let $p_0 \in V \subset M$ be a point and let (U, x) be a local real coordinate system on V around p_0. Then there exists a neighbourhood $U' \subset U$ of p_0 and a positive finite constant C such that for every p and q in V one has*

$$d_{V,M}(p, q) \leq C \, \|x(p) - x(q)\| .$$

In particular the function $d_{V,M}$ is continuous in $V \times V$.

Proof. Let m be the real dimension of V. Thus the map x is a real analytic diffeomorphism of U onto $x(U) \subset \mathbb{R}^m$. Put $x_0 = x(p_0)$. Since the map $x^{-1} : x(U) \to V$ is real analytic, there exists a neighbourhood $U' \subset U$ of p_0 and a small ball $B \subset \mathbb{C}^m$ centered at x_0 and a holomorphic map $F : B \to M$ such that $x(U') \subset\subset B$, $F(B \cap \mathbb{R}^m) \subset U \subset V$, and $F(x(p)) = p$ for every $p \in U'$. It follows that if p and q are arbitrarily chosen points of U' then

$$d_{V,M}(p, q) = d_{V,M}(F(x(p)), F(x(q))) \leq d_{B \cap \mathbb{R}^m, B}(x(p), x(q)). \tag{4.1}$$

Since $x(U') \subset\subset B$ it easy to prove, using images under complex affine mappings of the unit disc D, that there exists a constant C such that for every pair of points y' and y'' in $x(U')$ one has

$$d_{B\cap\mathbb{R}^m,B}(y',y'') \leq C\,\|y'-y''\|\,. \tag{4.2}$$

Combining 4.1 and 4.2 our assertion follows. □

The following assertion is an immediate consequence of this proposition.

Proposition 4.2. *If V is M-hyperbolic then the distance $d_{V,M}$ induces the topology of V.*

Proof. As $d_{V,M}$ is continuous we only have to prove that for every $p_0 \in V$ the open balls $B(r) = \{p \in V \mid d_{V,M}(p,q) < r\}$ form a fundamental system of neighbourhoods of p_0.

Let U be an arbitrary neighbourhood of p_0. We need to prove that there exists a ball $B(\varepsilon)$ contained in U for some $\varepsilon > 0$. Pick a connected neigbourhood U' of x_0 contained in U with compact boundary $S = \partial U'$. Every analytic chain in V connecting p_0 and an arbitrary point q in $V \setminus U'$ must intersect the boundary S of U' and therefore one has

$$\inf_{p\in V\setminus U} d_{V,M}(p_0,p) \geq \inf_{p\in V\setminus U'} d_{V,M}(p_0,p) \geq \inf_{p\in S} d_{V,M}(p_0,p) = \varepsilon > 0,$$

where the last inequality follows from the M-hyperbolicity of V, the continuity of $d_{V,M}$ and the compactess of S. But this implies that $B(\varepsilon) \subset U' \subset U$, as asserted. □

We now introduce a pseudometric on $V \subset M$ which generalizes the contruction of the Kobayashi-Royden pseudometric on complex manifolds, and then we will prove that its integrated form is the pseudodistance $d_{V,M}$.

Let us fix some notation. For every $p \in V$ we identify the real tangent space of M at p with the holomorphic tangent space of M at p, so that the (real) tangent space T_pV of V at p will be identified with a subspace of the holomorphic tangent space $T_p^{\mathbb{C}}M$ of M at p. For later use we denote by $\mathbb{C}T_pV$ the smallest complex vector subpace of $T_p^{\mathbb{C}}M$ containing T_pV.

If $f : D \to M$ is a holomorphic map sending I in V, for every $t \in I \subset D$ we then denote by $f'(t)$ either the image of the (real) tangent vector $\partial/\partial t$ under the differential of $f_{|I}$ at t, or the image of the holomorphic tangent vector $\partial/\partial z$ under the (holomorphic) differential of f at t.

With this notation, for every $p \in V$ and every $\xi \in T_pV$ we define $[F_{V,M}](p,\xi)$ as the infimum of the positive real numbers $a > 0$ for which there exists an M-analytic arc f in V such that $f(0) = p$ and $f'(0) = a^{-1}\xi$.

It is easy to check that all properties $(i), \ldots, (v)$ of the Section 2 stated for the pseudodistance $d_{V,M}$, with the necessary modifications hold for the pseudometric $[F_{V,M}]$. Moreover one sees that this pseudometric decreases under differentiable good mappings, and that the analogous estimate to that in Proposition 4.1 can also be given for this pseudodistance.

Up to now very little can be said about the regularity of $[F_{V,M}]$. Denoting the (real) tangent bundle of V by TV with its usual topological structure, the best result we can prove is the following:

Proposition 4.3. *The pseudometric $[F_{V,M}] : TV \to [0, +\infty[$ is a Borel function.*

Proof. Denote $[F_{V,M}]$ simply by F. We will prove our assertion finding a decreasing sequence of lower semicontinuous pseudometrics $F_n : TV \to [0, +\infty[$ such that for every $p \in V$ and $\xi \in T_pV$ one has

$$F(p, \xi) = \inf_n F_n(p, \xi). \tag{4.3}$$

Fix a complete hermitian metric h on M and denote by d its associated distance. For every $n \in \mathbb{N}$ let denote by $\mathcal{A}_n$ the class of all analytic arcs f in V satisfying $d(f(z), f(w)) \leq n \|z - w\|$ for every z and w in D. Let F_n be the pseudometric defined as the pseudometric F but using analytic arcs in $\mathcal{A}_n$ instead of all analytic arcs in V. As consequence of the Ascoli Theorem, by the completeness of the metric h and the closure of M, it follows that if f_ν is an arbitrary sequence of analytic arcs in $\mathcal{A}_n$ such that the sequence $f_\nu(0)$ converges to some point $p \in V$, then a subsequence of f_ν converges uniformly on all compact subsets of D to an analytic arc $f \in \mathcal{A}_n$ such that $f(0) = p$. Moreover the derivatives at 0 of such a subsequence converge to $f'(0)$. It is then an easy matter to derive the lower semicontinuity of the pseudometric F_n from this fact.

Let now $p \in V$ and $\xi \in T_pV$ be given. Let f be an analytic arc in V such that $f(0) = p$ and $f'(0) = a^{-1}\xi$. For every $\varepsilon > 0$ small put $f_\varepsilon(z) = f((1 - \varepsilon)z)$, $z \in D$. Then $f_\varepsilon \to f$ uniformely on compact subsets of D, and each f_ε belongs to $\mathcal{A}_n$, for some $n = n(\varepsilon)$. All this clearly implies the formula 4.3. The proof is so completed. □

If $\gamma : [0, 1] \to V$ is an absolutely continuous curve, the length of γ (with respect to the pseudometric $[F_{V,M}]$) is the number

$$\int_0^1 [F_{V,M}](\gamma(s), \dot{\gamma}(s))ds.$$

The integrated form $\bar{d}_{V,M}(p, q)$ of the pseudometric $[F_{V,M}]$ is the infimum of the lengths of the absolutely continuous curves $\gamma : [0, 1] \to V$ such that $\gamma(0) = p$ and $\gamma(1) = q$.

Proposition 4.4. *The pseudodistance $d_{V,M}$ and the integrated form of the pseudometric $[F_{V,M}]$ coincide.*

Proof. It is a direct consequence of the Theorem 2.1 of [**9**]. □

5. Group actions.

In this section M will stand for a complex manifold, V for a closed real analytic submanifold of M, and G for a complex Lie group of holomorphic transformation of M. We denote by $G(V)$ the subgroup of G of the transformations which leave the submanifold V invariant. Being V closed in M, then $G(V)$ is a closed subgroup of G, and therefore is a (real) Lie group. We also denote by $\mathbf{g}$ and $\mathbf{g}(V)$ the Lie algebras respectively of G and of $G(V)$, and by J the complex structure of $\mathbf{g}$.

Theorem 5.1. *If $G(V)$ acts transitively on V, then V is M-flat.*

Proof. Let $p \in V$. Then there is a neighbourhood U of p in V such that every $q \in V$ belongs to a real one parameter subgroup $t \mapsto \exp(tX)$, for some $X \in \mathbf{g}(V)$, which extends holomorphically to a entire holomorphic map by $\mathbb{C} \ni z \mapsto f(z) = \exp(zX)$. Clearly $f(\mathbb{R}) \subset V$, and therefore $d_{V,M}(p,q) = 0$. The triangle inequality then implies that $d_{V,M}$ vanishes everywhere, that is V is M-flat. □

Theorem 5.2. *If $G(V)$ acts effectively on V and V is M-hyperbolic then $G(V)$ is discrete.*

Proof. It suffices to prove that $\mathbf{g} = 0$. Pick $X \in \mathbf{g}$. Consider the real one-parameter subgroups

$$t \mapsto \exp(tX), \quad t \mapsto \exp(tJX).$$

We have $[X, JX] = 0$ and consequentely these two one-parameter subgroups generate a complex one-parameter subgroup H of G. Thus, taking $\mathbb{C}$, the universal covering of H, we obtain a holomorphic action $\mathbb{C} \times M \to M$ which extends the real action on V given by $(t,p) \mapsto \exp(tX)p$. Then, from the Theorem 3.1 it follows that $\exp(tX)p = p$ for every $t \in \mathbb{R}$, $p \in V$ and this implies $X = 0$, because $G(V)$ by hypothesis acts effectively on V. □

Corollary 5.1. *If V is M-hyperbolic and $\dim_{\mathbb{R}} G(V) > 0$, then G acts trivially on V. In particular, if there is a point $p_0 \in V$ such that $\mathbb{C}T_{p_0}V = T_{p_0}M$, then G acts trivially on M.*

Corollary 5.2. *Let M be compact, V be M-hyperbolic and suppose that there is a point p_0 such that $\mathbb{C}T_{p_0}V = T_{p_0}M$. Denote with $\mathrm{Aut}(M)$ the group*

of all the holomorphic automorphisms of M. Then the set

$$\{\sigma \in \mathrm{Aut}(M) \mid \sigma(V) \subset V\}$$

is a discrete subgroup of $\mathrm{Aut}(M)$.

Proof. Indeed $\mathrm{Aut}(M)$ is a complex Lie group wich acts on M effectively. □

Example. Let $M = \mathbb{C}^2$; then $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$ acts on $\mathbb{C}^2$ by

$$(z, w) \mapsto (\lambda z, w + (\lambda^2 - 1)z^2).$$

Let

$$V = \{(t, t^2) \mid t \in \mathbb{R}\}, \quad V' = \{(t + it, 2it^2) \mid t \in \mathbb{R}\}.$$

Then $G(V) = G(V') = \mathbb{R}^* = \mathbb{R}\setminus\{0\}$ acts effectively on V and V' respectively. Observe that V and V' in this example are flat.

6. Geodesics.

Let M be complex space and V be a subset of M. We say that an analytic arc $f : \Delta \to M$ such that $f(I) \subset V$ is a *M-geodesic* if it is a local isometry with respect to the distances $d_{I,\Delta}$ and $d_{V,M}$, that is, for every $t_0 \in I$ there exists a open interval $J \subset I$ containing t_0 such that

$$d_{V,M}(f(t), f(s)) = d_{I,\Delta}(t, s)$$

for every t and s in J. With abuse of language we also call *M-geodesic* in V a one dimensional real submanifold of M contained in V which is the image of the interval I under a M-geodesic $f : \Delta \to M$ in V.

Remark. If M is a hyperbolic Riemann surface and $V = M$ then the distance $d_{V,M}$ is the distance associated to a Hermitian metric h_M, and a M-geodesic in V is a holomorphic map $f : \Delta \to M$ such that $f_{|I}$ is a geodesic with respect to the metric h_M.

The following Proposition on geodesics on Riemann surfaces is useful for finding geodesics.

Proposition 6.1. *Let M be an hyperbolic irreducible complex curve, that is an irreducible complex space of (complex) dimension 1, and let M_r be the set of regular points of M. Let $\varphi : M \to M$ be an antiholomorphic map and let X be the set of the fixed points of φ. Then each connected component of X contained in M_r is (the image of) a geodesic of M.*

Proof. Let X_0 be a connected component of X contained in M_r and let $x_0 \in X_0$. Let $\pi : \tilde{M} \to M$ be the normalization of M and let $\tilde{x}_0 \in \tilde{M}$ be the

unique point such that $\pi(\tilde{x}_0) = x_0$. Let $f : \Delta \to \tilde{M}$ be a universal covering of $\tilde{M}$ such that $f(0) = \tilde{x}_0$ and let $\sigma : \Delta \to \Delta$ be the unique continuous map such that $\sigma(0) = 0$ and $\pi \circ f \circ \sigma = \varphi \circ \pi \circ f$. Then X_0 is the image under $\pi \circ f$ of the set Z of the fixed point set of σ. But σ is an antiholomorphic automorphism of Δ such that $\sigma(0) = 0$ and hence there exists $\theta \in \mathbb{R}$ such that

$$\sigma(z) = e^{i\theta}\bar{z}.$$

Thus the set Z is the intersection of Δ and a straight (real) line through the origin, and therefore it is a geodesic in Δ (for the Poincaré metric of Δ). Since both the covering map f and the restriction of π to $\pi^{-1}(M_r)$ are (local) isometries for the Kobayashi distance, the set X_0 also is a geodesic in M, as asserted. □

Example. Let $X \subset \mathbb{C}^2$ be the image of the periodic map $f : \mathbb{R} \to \mathbb{C}^2$ defined by

$$f(t) = \left(e^{it}, \frac{e^{-it}}{(e^{it} - 2)(2e^{it} - 1)}\right).$$

Then X is a $\mathbb{C}^2$ geodesic. Indeed let $M = \mathbb{C} \setminus \{0, 1/2, 2\}$ and let $F : M \to \mathbb{C}^2$ be the map defined by

$$F(z) = \left(z, \frac{1}{z(z - 2)(2z - 1)}\right).$$

Then F is a holomorphic embedding of M into $\mathbb{C}^2$ and X is the image under F of $S \subset M$, the unit circle in $\mathbb{C}$. Hence it suffices to prove that S is a geodesic in M (for the Kobayashi metric). But this follows immediatly from the previous proposition, observing that S is the fixed point set of the antiholomorphic automorphism $\varphi : M \to M$ defined by

$$\varphi(z) = 1/\bar{z}.$$

Proposition 6.2. *Let $V \subset \mathbb{C}^n = \mathbb{R}^{2n}$ be a subset defined by k real equations as in Proposition* 3.11. *Assume furthermore that V is a real smooth submanifold of (real) dimension one. If V is $\mathbb{C}^n$-hyperbolic then each connected component of V is a $\mathbb{C}^n$-geodesic.*

Proof. Let $F : \mathbb{C}^n \to \mathbb{C}^{2n}$ be defined by

$$z = (x_1 + iy_1, \ldots, x_n + iy_n) \mapsto (x_1, y_1, \ldots, x_n, y_n).$$

Let $V_{\mathbb{C}} \subset \mathbb{C}^{2n}$ be defined as in the Proposition 3.11. Let $L : \mathbb{C}^{2n} \to \mathbb{C}^n$ be the holomorphic map defined by

$$(x_1, y_1, \dots, x_n, y_n) \mapsto (x_1 + iy_1, \dots, x_n + iy_n).$$

Obviously $L(F(z)) = z$ for every $z \in \mathbb{C}^n$. Thus, given $z, w \in V$, one has

$$\begin{aligned} (6.1) \qquad d_{V,\mathbb{C}^n}(z,w) &\geq d_{F(V),\mathbb{C}^{2n}}(F(z),F(w)) \\ &\geq d_{L(F(V)),\mathbb{C}^n}(L(F(z)),L(F(w))) = d_{V,\mathbb{C}^n}(z,w). \end{aligned}$$

It follows that the map $L : F(V) \to V$ is an isometry with respect to the distances $d_{F(V),\mathbb{C}^{2n}}$ and $d_{V,\mathbb{C}^n}$, and hence in order to prove our assertion it suffices to prove that each connected component of $F(V)$ is a $\mathbb{C}^{2n}$-geodesic.

Let $F(V_0)$ be a connected component of $F(V)$, where V_0 is a connected component of V, and let W be the smallest complex analytic subspace of $\mathbb{C}^{2n}$ containing $F(V_0)$. Since W is closed in $\mathbb{C}^{2n}$ then, by Proposition 3.6, one has

$$d_{F(V_0),\mathbb{C}^{2n}}(F(z),F(w)) = d_{F(V_0),W}(F(z),F(w)).$$

Since V is $\mathbb{C}^n$-hyperbolic, by 6.1 it follows that W is not flat for the Kobayashi metric, and hence, since W is a complex one dimensional curve, it is hyperbolic.

Let $\varphi : \mathbb{C}^{2n} \to \mathbb{C}^{2n}$ the map defined by

$$(x_1, y_1, \dots, x_n, y_n) \mapsto (\bar{x}_1, \bar{y}_1, \dots, \bar{x}_n, \bar{y}_n).$$

Since $V_{\mathbb{C}}$ is defined by real equations, the space W is invariant under φ. Clearly the restriction of the map φ to W is an antiholomorphic automorphism of W. We end the proof observing that $F(V_0)$ is a connected component of the fixed point set in W of the map φ and hence the Proposition 6.1 applies. □

References

[1] M. Abate, *Iteration theory of holomorphic maps on taut manifolds*, Mediterrean Press, Cosenza, 1989.

[2] T. Franzoni and E. Vesentini, *Holomorphic maps and invariant distances*, North Holland, 1972.

[3] F. Forstneric and J.P. Rosay, *Localization of the Kobayashi metric and the boundary continuity of proper holomorphic mappings*, Math. Ann., **279** (1987), 239-252.

[4] S. Kobayashi, *Hyperbolic manifolds and holomorphic mappings*, Dekker, New York, 1970.

[5] ———, *Intrinsic distances, measures and geometric function theory*, Bull. Amer. Math. Soc., **82** (1976), 357-416.

[6] S. Lang, *An introduction to complex hyperbolic spaces*, Springer, New York, 1987.

[7] H. Royden, *Remarks on the Kobayashi metric*, In *Several Complex Variables* II, Lect. Notes in Math., **189**, Springer, Berlin, 1971, 125-137.

[8] N. Sibony, *A class of hyperbolic manifolds*, In *Recent Developments in Several Complex Variables*, Ann. of Math. Study, **100**, 1981, 357-372.

[9] S. Venturini, *Pseudodistances and pseudometrics on real and complex manifolds*, Ann. Mat. Pura e Appl., Srie (IV), **154** (1989), 385-402.

[10] E. Vesentini, *Variations on a theme of Carathéodory*, Ann. Sc. Norm. Super. Pisa, **6** (1979), 39-68.

Received July 14, 1993.

UNIVERSITÀ DI PARMA
VIA DELL'UNIVERSITÀ, 12
43100 PARMA, ITALY

SCUOLA NORMALE SUPERIORE
PIAZZA DEI CAVALIERI, 7
56126 PISA, ITALY

AND

UNIVERSITÀ DI BOLOGNA
PIAZZA DI PORTA S. DONATO, 5
40127 BOLOGNA, ITALY

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

VALUES OF BERNOULLI POLYNOMIALS

ANDREW GRANVILLE[1] AND ZHI-WEI SUN[2]

Dedicated to Emma Lehmer

Let $B_n(t)$ be the nth Bernoulli polynomial. We show that $B_{p-1}(a/q) - B_{p-1} \equiv q(U_p - 1)/2p \pmod p$, where U_n is a certain linear recurrence of order $[q/2]$ which depends only on a, q and the least positive residue of $p \pmod q$. This can be re-written as a sum of linear recurrence sequences of order $\le \phi(q)/2$, and so we can recover the classical results where $\phi(q) \le 2$ (for instance, $B_{p-1}(1/6) - B_{p-1} \equiv (3^p-3)/2p + (2^p-2)/p \pmod p$). Our results provide the first advance on the question of evaluating these polynomials when $\phi(q) > 2$, a problem posed by Emma Lehmer in 1938.

Introduction.

It has long been known that the nth Bernoulli polynomial $B_n(t)$, where

$$B_n(t) = \sum_{j=0}^{n} \binom{n}{j} B_{n-j} t^j$$

and B_k, the kth Bernoulli number, defined by the power series

$$\frac{x}{e^x - 1} = \sum_{k \ge 0} B_k \frac{x^k}{k!} ,$$

take 'special' values at certain rational numbers with small denominators:

$$B_n(1) = B_n(0) = B_n, \quad \text{for } n \ne 1 \tag{1}$$

$$B_n\left(\frac{1}{2}\right) = (2^{1-n} - 1)B_n;$$

[1]The first author is an Alfred P. Sloan Research Fellow and a Presidential Faculty Fellow. Also supported, in part, by the National Science Foundation.

[2]The second author was supported by the National Natural Science Foundation of the People's Republic of China.

and for all even $n \geq 2$,

$$(2) \qquad \begin{aligned} B_n\left(\frac{1}{3}\right) = B_n\left(\frac{2}{3}\right) &= \frac{1}{2}\,(3^{1-n}-1)B_n, \\ B_n\left(\frac{1}{4}\right) = B_n\left(\frac{3}{4}\right) &= \frac{1}{2}\,(4^{1-n}-2^{1-n})B_n, \\ B_n\left(\frac{1}{6}\right) = B_n\left(\frac{5}{6}\right) &= \frac{1}{2}\,(6^{1-n}-3^{1-n}-2^{1-n}+1)B_n. \end{aligned}$$

It is not known if $B_n(a/q)$ has as simple a 'closed form' for any other rational a/q with $1 \leq a \leq q-1$ and $(a,q)=1$, though this has long been considered an interesting question.

Following work of Friedmann and Tamarkin [**FT**], Emma Lehmer [**Lh**, 1938] considered Bernoulli numbers and polynomials modulo primes and prime powers, and showed amongst other things that (1) and (2) imply

(3)

$$\begin{aligned} B_{p-1}\left(\frac{1}{2}\right) - B_{p-1} &\equiv \frac{2^p-2}{p} \pmod p \\ B_{p-1}\left(\frac{1}{3}\right) - B_{p-1} \equiv B_{p-1}\left(\frac{2}{3}\right) - B_{p-1} &\equiv \frac{1}{2}\frac{(3^p-3)}{p} \pmod p \\ B_{p-1}\left(\frac{1}{4}\right) - B_{p-1} \equiv B_{p-1}\left(\frac{3}{4}\right) - B_{p-1} &\equiv \frac{3}{2}\frac{(2^p-2)}{p} \pmod p \\ B_{p-1}\left(\frac{1}{6}\right) - B_{p-1} \equiv B_{p-1}\left(\frac{5}{6}\right) - B_{p-1} &\equiv \frac{1}{2}\frac{(3^p-3)}{p} + \frac{2^p-2}{p} \pmod p. \end{aligned}$$

The "Fermat quotients", $(2^p-2)/p$ and $(3^p-3)/p$ play a central rôle in the study of the first case of Fermat's Last Theorem (see Ribenboim's elegant account [**Ri**]), and this connection with Bernoulli polynomials has recently been explored in much greater depth by Skula [**Sk**] (see also [**Gr**]).

However, until now, no progress has been made in extending the table of intriguing congruences given in (3). This is the intention here. (It should be mentioned that recent papers of H. C. Williams [**W1**, **W2**], of G. Andrews [**An**] as well as of the second author and his twin brother Zhi-Hong Sun [**SS**], each come close to doing this.)

Before stating our main result, which is of a somewhat technical nature, let's discuss the next class of examples after (3). The two important things to note about (3) are that,

(i): We've evaluated $B_{p-1}(\frac{a}{q}) - B_{p-1} \pmod p$ where $\phi(q) = 1$ or 2 (ϕ is Euler's totient function);

(ii): Each of the terms of the right hand side, like 2^p, 3^p, are numbers taken from a first-order linear recurrence sequence ($u_{n+1} = 2u_n$ and $u_{n+1} = 3u_n$ respectively).

This is the viewpoint we need to generalize. We shall show, for $q > 2$, that $B_{p-1}(\frac{a}{q}) - B_{p-1} \pmod p$ is congruent to a sum of multiples of terms, each of which are numbers taken from a kth-order linear recurrence sequence with

$$k \le \phi(q)/2.$$

Thus the next class of examples are those q for which $\phi(q) = 4$, namely $q = 5, 8, 10, 12$. We shall show that, for $1 \le a \le q-1$ with $(a,q) = 1$ (there being four such integers a), we have, when prime p does not divide q,

(4)

$$B_{p-1}\left(\frac{a}{5}\right) - B_{p-1} \equiv \frac{5}{4}\left\{\left(\frac{ap}{5}\right)\frac{1}{p}F_{p-(\frac{5}{p})} + \frac{5^{p-1}-1}{p}\right\} \pmod p$$

$$B_{p-1}\left(\frac{a}{8}\right) - B_{p-1} \equiv \left\{2\left(\frac{8}{ap}\right)\frac{1}{p}G_{p-(\frac{8}{p})} + 4\,\frac{(2^{p-1}-1)}{p}\right\} \pmod p$$

$$B_{p-1}\left(\frac{a}{10}\right) - B_{p-1} \equiv \frac{15}{4}\left(\frac{ap}{5}\right)\frac{1}{p}F_{p-(\frac{5}{p})} + \frac{5}{4}\cdot\frac{5^{p-1}-1}{p} + \frac{2(2^{p-1}-1)}{p} \pmod p$$

$$B_{p-1}\left(\frac{a}{12}\right) - B_{p-1} \equiv 3\left(\frac{12}{a}\right)\frac{1}{p}H_{p-(\frac{12}{p})} + \frac{3(2^{p-1}-1)}{p} + \frac{3}{2}\frac{(3^{p-1}-1)}{p} \pmod p$$

where $(-)$ is the Jacobi symbol, and we define the following second-order linear recurrence sequences:

$$F_0 = 0,\ F_1 = 1,\ \text{and } F_{n+2} = F_{n+1} + F_n \quad \text{for all } n \ge 0$$
$$G_0 = 0,\ G_1 = 1,\ \text{and } G_{n+2} = 2G_{n+1} + G_n \quad \text{for all } n \ge 0$$
$$H_0 = 0,\ H_1 = 1,\ \text{and } H_{n+2} = 4H_{n+1} - H_n \quad \text{for all } n \ge 0.$$

($\{F_n\}$ is, of course, the Fibonacci sequence.)

In general we fix residue classes a and $b \pmod q$, with $(ab, q) = 1$. Then, for each divisor d of q, there exists a recurrence sequence $u_n = u_n(d, a, b)$ of order $D = \phi(d)/2$, with characteristic polynomial

$$\prod_{\substack{1\le j\le d/2\\(j,d)=1}}\left(X - 2 + e^{2i\pi j/d} + e^{-2i\pi j/d}\right) = X^D - \sum_{i=0}^{D-1} f_i X^{D-1-i},$$

so that

$$u_{n+D} = f_0 u_{n+D-1} + f_1 u_{n+D-2} + \cdots + f_{D-1}u_n$$

for all $n \ge 0$. The values of $u_0, \cdots, u_{D-1}$ depend on a and $b \pmod d$ and are somewhat complicated to describe – see Section 2 for precise details.

Our main result is that, for any $(a,q) = 1$, $1 \le a \le q$,

$$(5)\quad B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \equiv \sum_{d|q}\frac{1}{2p}\left\{u_p(d;a,b) - (\phi(d) - \mu(d))\right\} \pmod p$$

where b is the least positive residue of $p \pmod q$ (and μ is the Möbius function), provided prime p does not divide q. Each term in the sum is a p-unit.

Our formula involves such an awkward sum of recurrence sequences though each appears "naturally" in

(6)
$$\sum_{d|q} \mu\left(\frac{q}{d}\right)\left(B_{p-1}\left(\frac{a_d}{d}\right) - B_{p-1}\right) \equiv \frac{1}{2p}\{u_p(q;a,b) - (\phi(q) - \mu(q))\} \pmod p$$

where a_d is the least positive residue of $a \bmod d$. Indeed this is the formula we shall prove and then (5) is deduced by summing (6) over divisors of q.

We are unable to answer the question as to whether it is possible to give such a congruence for $B_{p-1}(\frac{a}{q}) - B_{p-1}$ involving only lower order recurrence sequences. Indeed this seems difficult, unless one can give a complete characterization of all linear recurrence sequences $(X_n)_{n\ge 0}$ for which $X_p \equiv 0 \bmod p^2$ for all but finitely many primes p. However we do not even know how to decide this for $X_n = 2^n - 2$.

However, it is easily shown that any sum of recurrence sequences can be written as one recurrence sequence, though of higher order. Thus (5) can be rewritten

(7)
$$B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \equiv \frac{q}{2p}\{U_p(q;a,b) - 1\} \pmod p$$

where, now, U_n has characteristic polynomial

$$\prod_{1\le j\le q/2}\left(X - 2 + e^{2i\pi j/q} + e^{-2i\pi j/q}\right).$$

Again it is complicated to compute the values of U_n for small n.

It is tempting to provide one "concrete" example for arbitrarily large q. We will now completely describe $U_p(q;a,b)$ in the case that $a \equiv \pm b \pmod q$ (that is $a \equiv \pm p \pmod q$) and q is odd:

Theorem. *If q is an odd integer ≥ 3 and $1 \le a \le q$ with $(a,q) = 1$, then*

(8)
$$B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \equiv \frac{q}{2p}\{x_p - 1\} \pmod p$$

whenever $p \equiv \pm a \pmod q$ where $\{x_n\}_{n\ge 0}$ is the $\frac{q-1}{2}$th order recurrence sequence given by

$$x_n = \frac{1}{2}\binom{2n}{n}, \qquad 0 \le n \le \frac{q-1}{2};$$

and for $D = \frac{q-1}{2}$ *we have*

$$x_{n+D} = \frac{f_{D-1}}{(D-1)!}\, x_{n+D-1} - \frac{f_{D-2}}{(D-2)!}\, x_{n+D-2} + \cdots \pm \frac{f_0}{0!}\, x_n$$

where

$$f_k = \sum_{0 \le j \le [\frac{1}{2}(D-k)]} \frac{(D-j)!}{j!(D-2j-k)!} \quad (-1)^j 2^{D-2j-k}.$$

Since this is the simplest general case, we hope the reader understands why we suppress so many details in this introduction!

Finally we give the first example with $\phi(q) = 6$, namely $q = 7$: Here we have that, for $1 \le a \le 6$, and any odd prime $p \ne 7$,

$$B_{p-1}\left(\frac{a}{7}\right) - B_{p-1} \equiv \frac{7}{2p}\,\{U_p(7; a, b) - 1\} \pmod{p}$$

where $b = 1, 2$ or 3 with $b \equiv \pm p \pmod 7$, and U_n satisfies the recurrence relation

$$U_{n+3} = 7U_{n+2} - 14U_{n+1} + 7U_n.$$

The values of U_1, U_2, U_3 are given in the table below:

$\pm a$	$\pm b$	U_1	U_2	U_3
2	1	1	2	5
3	2	2	7	26
1	3	2	6	19
3	1	1	2	6
1	2	3	11	41
2	3	2	5	13
a	a	1	3	10

Analogous results can be given for generalized Bernoulli numbers (for Dirichlet characters) since they may be expressed in terms of values of Bernoulli polynomials. It is perhaps more obvious that there should be simple expressions for these since they can be described in terms of p-adic L-functions which, in turn, can be written in a number of elegant ways. The case of quadratic characters has been examined in [**KS**] and [**W2**], and here we give a somewhat different proof of a result proved there:

Suppose that q is a prime $\equiv 1 \pmod 4$. Let h_q and ε_q be the class number and fundamental unit, respectively, of the real quadratic field $\mathbf{Q}(\sqrt{q})$. It is well-known that $\varepsilon_q^{(p-(\frac{p}{q}))} = U + p\sqrt{q}V$ for some integers U and V, where

$\left(\frac{p}{q}\right)$ is the Legendre symbol. Thus the generalized Bernoulli polynomial

$$(9)\quad B_{p-1,\left(\frac{\cdot}{q}\right)} := \sum_{a=1}^{q-1}\left(\frac{a}{q}\right)\left\{B_{p-1}\left(\frac{a}{q}\right) - B_{p-1}\right\} \equiv -2\left(\frac{p}{q}\right) qh_qV \pmod p.$$

The organization of the paper is as follows: In the next section we shall develop basic identities and results about Bernoulli polynomials that we shall require in our proofs. In Section 2 we shall see how the values of Bernoulli polynomials can be expressed in terms of certain functions of roots of unity. This leads to the proof of a number of the cases mentioned in the introduction; though, because of the computations needed, we give the complete proof of the Theorem in Section 4, and the complete proof of (4) in Section 5. In Section 3 we develop the analogous formulae for those generalized Bernoulli numbers with quadratic characters, which leads to (9) above.

We thank Emma Lehmer, Hugh Williams and the anonymous referee for many useful comments.

1. The (regular) theory of Bernoulli polynomials.

The nth Bernoulli number B_n is defined by the power series

$$(1.1)\qquad \frac{x}{e^x-1} = \sum_{n\ge 0} B_n \frac{x^n}{n!}.$$

The nth Bernoulli polynomial $B_n(t)$ is defined by the power series

$$(1.2)\qquad \frac{xe^{tx}}{e^x-1} = \sum_{n\ge 0} B_n(t)\frac{x^n}{n!}$$

so that $B_n(0) = B_n$ and

$$(1.3)\qquad B_n(t) = \sum_{j=0}^{n}\binom{n}{j} B_j t^{n-j}.$$

Perhaps the most important property of Bernoulli polynomials is that

$$(1.4)\qquad B_n(t+1) - B_n(t) = nt^{n-1} \quad \text{for all } n \ge 1$$

as is easily deduced from (1.2). From (1.4) we notice that $B_n(1) = B_n(0) = B_n$ for all $n \ne 1$, and that it is "easy" to deduce the value of $B_n(t)$ for any real number t, once we understand the value of $B_n(t)$ for t in the interval $[0,1)$.

It is thus of interest to determine $B_n(t)$ for 'special' values of t in $[0,1)$, for instance those rational t with small denominator. We already have

$$B_n(0) = B_n(1) = B_n \quad \text{for } n \neq 1,$$

and from the identity

$$\frac{2xe^x}{e^{2x}-1} = 2\ \frac{x}{e^x-1} - \frac{2x}{e^{2x}-1}$$

we easily deduce that

$$B_n\left(\frac{1}{2}\right) = (2^{1-n}-1)\,B_n,$$

and thus we have proved (1). We next observe that

$$B_n(1-t) = (-1)^n B_n(t) \tag{1.5}$$

from the identity

$$\frac{xe^{(1-t)x}}{e^x-1} = \frac{(-x)e^{t(-x)}}{e^{(-x)}-1},$$

so we study only $t \in (0, \frac{1}{2})$.

The next important observation is due to Lerch [**Lr**]: By taking the identity

$$\frac{qxe^{ax}}{e^{qx}-1} + \frac{qxe^{(a+1)x}}{e^{qx}-1} + \frac{qxe^{(a+2)x}}{e^{qx}-1} + \cdots + \frac{qxe^{(a+q-1)x}}{e^{qx}-1} = \frac{qxe^{ax}}{e^x-1}$$

we obtain

$$B_n\left(\frac{a}{q}\right) + B_n\left(\frac{a+1}{q}\right) + B_n\left(\frac{a+2}{q}\right) + \cdots + B_n\left(\frac{a+q-1}{q}\right) = q^{1-n}B_n(a) \tag{1.6}$$

and, in particular if $a = 0$,

$$B_n + B_n\left(\frac{1}{q}\right) + B_n\left(\frac{2}{q}\right) + \cdots + B_n\left(\frac{q-1}{q}\right) = q^{1-n}B_n. \tag{1.7}_q$$

In order to remove those $B_n(j/q)$ in which j/q is not in lowest terms we may use the standard Möbius inversion formula, as follows: Take $\sum_{d|q} \mu(d)(1.7)_{q/d}$ for $q \geq 3$, so that

$$\sum_{\substack{j=0 \\ (j,q)=1}}^{q-1} B_n\left(\frac{j}{q}\right) = \left(\sum_{d|q} \mu(d)\left(\frac{q}{d}\right)^{1-n}\right) B_n. \tag{1.8}$$

Using (1.5) we have, for all $q \geq 3$ and n even,

$$\sum_{\substack{1\leq j<q/2\\(j,q)=1}} B_n\left(\frac{j}{q}\right) = \frac{1}{2}\, q^{1-n} \prod_{p\mid q} (1-p^{n-1})\, B_n. \tag{1.9}$$

Taking $q = 3, 4$ and 6 we deduce (2).

The seven values $\frac{1}{2}$, $\frac{1}{3}$, $\frac{2}{3}$, $\frac{1}{4}$, $\frac{3}{4}$, $\frac{1}{6}$, $\frac{5}{6}$ are the only rationals with small denominators for which such "straightforward" values of $B_n(t)$ are known, with $0 < t < 1$. It has, however, been recently observed [**AM**] that $B_n(t) - B_n$ shares one surprising property with polynomials which have integer coefficients: namely that $q^n(B_n(a/q) - B_n)$ is an integer whenever a and q are non-zero integers.

One of the most important, and elegant, applications of these valuations is to the study of Bernoulli polynomials modulo p for p prime. The Von Staudt-Clausen theorem asserts that

$$pB_{2k} \equiv -1 \pmod{p}$$

whenever $2k$ is divisible by $p-1$. In 1850 Eisenstein observed the following (easily proved) congruences:

$$\frac{(ab)^{p-1}-1}{p} \equiv \frac{a^{p-1}-1}{p} + \frac{b^{p-1}-1}{p} \pmod{p}$$

and

$$\frac{a^{1-(p-1)}-a}{p} \equiv -\frac{(a^p-a)}{p} \pmod{p}.$$

Thus we deduce (3) from (2) with $n = p-1$. Such congruences fit elegantly into the general overview of the first case of Fermat's Last Theorem (see Chapter 8 of [**Ri**]).

Actually, by the same method, we can transform (1.9) to read, for any even $n \geq 2$,

$$\sum_{\substack{1\leq j<q/2\\(j,q)=1}} \left(B_n\left(\frac{j}{q}\right) - B_n\right) = \frac{qB_n}{2} \sum_{d\mid q} \frac{\mu(d)}{d} \left(\left(\frac{q}{d}\right)^{-n} - 1\right).$$

Taking $n = p-1$ we thus obtain

$$\begin{aligned}\sum_{\substack{1\leq j<q/2\\(j,q)=1}} \left(B_{p-1}\left(\frac{j}{q}\right) - B_{p-1}\right) &\equiv \frac{1}{2} \sum_{d\mid q} \mu(d)\, \frac{\{(q/d)^p - (q/d)\}}{p} \\ &\equiv \frac{\phi(q)}{2} \left(\frac{q^{p-1}-1}{p} + \sum_{\substack{l\mid q\\ l \text{ prime}}} \frac{l^{p-1}-1}{p(l-1)}\right) \pmod{p}\end{aligned}$$

However such a formula allows us to evaluate $B_n(a/q)$ only for particular values of a coprime to q, provided $\phi(q) \leq 2$. It is the main purpose of this paper to determine the value of

$$B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \pmod p.$$

2. Working with roots of unity.

Key Proposition. *If $1 \leq a \leq q$ and odd prime p does not divide q then*

(2.1)
$$B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \equiv \sum_{\substack{\gamma^q=1 \\ \gamma\neq 1}} \left(\frac{\gamma^a - 2 + \gamma^{-a}}{\gamma^p - 2 + \gamma^{-p}}\right)\left(\frac{(1-\gamma)^p - 1 + \gamma^p}{p}\right) \pmod p.$$

Proof. If $\gamma^q = 1$ then

$$\frac{(1-\gamma)^p - 1 + \gamma^p}{p} = \sum_{j=1}^{p-1} \frac{1}{p}\binom{p}{j}(-1)^j\gamma^j \equiv -\sum_{j=1}^{p-1}\frac{\gamma^j}{j} \pmod p$$

since

$$\frac{1}{p}\binom{p}{j}(-1)^j = \frac{(-p)(1-p)(2-p)\cdots(j-1-p)}{p\ \ 1\ \ 2\cdots(j-1)\cdot j} \equiv -\frac{1}{j} \pmod p.$$

We also have

$$\frac{\gamma^a - 2 + \gamma^{-a}}{\gamma^p - 2 + \gamma^{-p}} = m + \sum_{i=1}^{m}(m-i)(\gamma^{ip} + \gamma^{-ip})$$

by substituting $x = \gamma^p$ and $m \equiv a/p \bmod q$ into the identity

(2.2)
$$\frac{x^m - 2 + x^{-m}}{x - 2 + x^{-1}} = m + \sum_{i=1}^{m}(m-i)(x^i + x^{-i}).$$

Therefore the righthand side of (2.1) is

$$\equiv -\sum_{\gamma^q=1}\left\{m + \sum_{i=1}^{m}(m-i)(\gamma^{ip} + \gamma^{-ip})\right\}\sum_{j=1}^{p-1}\frac{\gamma^j}{j} \pmod p$$

$$\equiv -q\left\{m\sum_{\substack{0<j<p \\ q\mid j}}\frac{1}{j} + \sum_{i=1}^{m}(m-i)\left(\sum_{\substack{0<j<p \\ q\mid ip+j}}\frac{1}{ip+j} - \sum_{\substack{0<j<p \\ q\mid ip-j}}\frac{1}{ip-j}\right)\right\} \pmod p,$$

using the fact, for $\gamma = 1$, that $\sum_{j=1}^{p-1} 1/j \equiv 0 \pmod p$. Now, since $ip < ip+j < (i+1)p$ and $(i-1)p < ip-j < ip$, we replace $q/(ip \pm j)$ by $1/k$ so that the above is

$$\equiv -\left\{ m \sum_{0<k<\frac{p}{q}} \frac{1}{k} + \sum_{i=1}^{m} (m-i) \left(\sum_{\frac{ip}{q}<k<(i+1)\frac{p}{q}} \frac{1}{k} - \sum_{\frac{(i-1)p}{q}<k<\frac{ip}{q}} \frac{1}{k} \right) \right\} \pmod p$$

$$\equiv - \sum_{0<k<\frac{mp}{q}} \frac{1}{k} \pmod p$$

$$\equiv (p-1) \sum_{0\le k\le \frac{mp-a}{q}} k^{p-2} \pmod p.$$

But this equals the coefficient of $x^{p-1}/(p-1)!$ in

$$x \sum_{k=0}^{(mp-a)/q} e^{kx} = x \, \frac{e^{(mp-a+q)/qx} - 1}{e^x - 1}$$

which is $B_{p-1}\left(\frac{mp-a+q}{q}\right) - B_{p-1}$ by (1.2). However the Von Staudt-Clausen Theorem tells us that p divides the denominator of B_n if and only if $p-1$ divides n; and so, by (1.3), the denominators of the coefficients of $B_{p-1}(t) - B_{p-1}$ are not divisible by p. Therefore

$$B_{p-1}\left(\frac{mp-a+q}{q}\right) - B_{p-1} \equiv B_{p-1}\left(\frac{-a+q}{q}\right) - B_{p-1} \pmod p$$

$$\equiv B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \pmod p,$$

by (1.5), and the Proposition follows. □

Corollary 1. *If $1 \le a \le q-1$ and odd prime p does not divide q then*

$$(2.3)\quad B_{p-1}\left(\frac{a}{q}\right) - B_{p-1}$$

$$\equiv \frac{1}{2} \sum_{\substack{\gamma^q=1 \\ \gamma\ne 1}} \left(1 - \frac{\gamma^a + \gamma^{-a}}{2}\right) \frac{1}{p} \left(\frac{(2-\gamma-\gamma^{-1})^p}{2-\gamma^p-\gamma^{-p}} - 1 \right) \pmod p .$$

Proof. It is evident that

$$(1-\gamma)^p \equiv 1 - \gamma^p \equiv 2^{p-1}(1-\gamma^p) \pmod p.$$

Therefore

$$0 \equiv \{(1-\gamma)^p - 2^{p-1}(1-\gamma^p)\}^2/(2\gamma)^p \pmod{p^2}$$
$$= -\left(1-\frac{\gamma+\gamma^{-1}}{2}\right)^p + (1-\gamma)^p + (1-\gamma^{-1})^p - 2^{p-1}\left(1-\frac{\gamma^p+\gamma^{-p}}{2}\right).$$

Thus

$$\frac{(1-\gamma)^p + (1-\gamma^{-1})^p - 2 + \gamma^p + \gamma^{-p}}{p}$$
$$\equiv \frac{\left(1-\frac{\gamma+\gamma^{-1}}{2}\right)^p - \left(1-\frac{\gamma^p+\gamma^{-p}}{2}\right)}{p} + \frac{2^{p-1}-1}{p}\left(1-\frac{\gamma^p+\gamma^{-p}}{2}\right) \pmod{p}.$$

Now, adding each term to its conjugate in (2.1) we get the following congruence modulo p:

$$B_{p-1}\left(\frac{a}{q}\right) - B_{p-1}$$
$$\equiv \frac{1}{2}\sum_{\substack{\gamma^q=1\\ \gamma\neq 1}}\left(1-\frac{\gamma^a+\gamma^{-a}}{2}\right)\left\{\frac{1}{p}\left(\frac{\left(1-\frac{\gamma+\gamma^{-1}}{2}\right)^p}{\left(1-\frac{\gamma^p+\gamma^{-p}}{2}\right)}-1\right)+\frac{2^{p-1}-1}{p}\right\}.$$

Since the two terms in the final brackets are both units mod p we may multiply the first by $2^{p-1} \equiv 1 \pmod{p}$ to get

$$\frac{1}{p}\left(\frac{(2-\gamma-\gamma^{-1})^p}{2-\gamma^p-\gamma^{-p}} - 2^{p-1}\right) + \left(\frac{2^{p-1}-1}{p}\right)$$
$$\equiv \frac{1}{p}\left(\frac{(2-\gamma-\gamma^{-1})^p}{2-\gamma^p-\gamma^{-p}} - 1\right) \pmod{p}.$$

The result follows. □

The next result follows immediately by applying Möbius inversion to (2.3) and associating the γ and γ^{-1} terms.

Corollary 2. *If $q \geq 3$, $1 \leq a \leq q$ and odd prime p does not divide q then*

$$\sum_{d|q}\mu\left(\frac{q}{d}\right)B_{p-1}\left(\frac{a_d}{d}\right)$$
$$\equiv \frac{1}{2}\sum_{\substack{j=1\\ (j,q)=1}}^{q/2}(2-(w^{ja}+w^{-ja}))\,\frac{1}{p}\left(\frac{(2-w^j-w^{-j})^p}{2-w^{jp}-w^{-jp}}-1\right) \pmod{p} \tag{2.4}$$

where $w = e^{2i\pi/q}$ *and* a_d *is the least positive residue of* $a \pmod d$.

Next note that if $(a,q)=1$ then

$$\sum_{\substack{j=1\\(j,q)=1}}^{q/2} (2 - w^{ja} - w^{-ja}) = \phi(q) - \sum_{\substack{i=1\\(i,q)=1}}^{q} w^i = \phi(q) - \mu(q).$$

Thus if we define

$$u_n(q;a,b) := \sum_{\substack{j=1\\(j,q)=1}}^{q/2} \left(\frac{2 - w^{ja} - w^{-ja}}{2 - w^{jb} - w^{-jb}}\right)(2 - w^j - w^{-j})^n,$$

where a, b are taken $\pmod q$, then by Corollary 2,

$$\sum_{d|q} \mu\left(\frac{q}{d}\right) B_{p-1}\left(\frac{a_d}{d}\right) \equiv \frac{1}{2p}\left\{u_p(q;a,p \bmod q) - (\phi(q) - \mu(q))\right\} \pmod p.$$

Now u_n so defined is a recurrence sequence with characteristic polynomial

$$F_q(X) := \prod_{\substack{j=1\\(j,q)=1}}^{q/2} (X - 2 + w^j + w^{-j}) .$$

Note that

$$\begin{aligned} F_q((1-X^{-1})(1-X)) &= \prod_{\substack{j=1\\(j,q)}}^{q/2} \left(-\frac{1}{X}\right)(X - w^j)(X - w^{-j}) \\ &= (-X^{-1})^{\phi(q)/2}\phi_q(X) \end{aligned}$$

where $\phi_q(X)$ is the qth cyclotomic polynomial.

If $F(X) = X^D - \sum_{i=0}^{D-1} f_i X^i$ where $D = \phi(q)/2$, then

$$u_{n+D} = f_{D-1}u_{n+D-1} + f_{D-2}u_{n+D-2} + \cdots + f_0 u_n \quad \text{for all } n \ge 0.$$

We get the same recurrence relation for all u_n with a given q, but the starting values, $u_0, u_1, \cdots, u_{D-1}$, are different.

Let's define

$$V_n(q;k) = \sum_{\substack{j=1\\(j,q)=1}}^{q} w^{jk}(2 - w^j - w^{-j})^n.$$

This satisfies the same recurrence relation. Moreover since for $m \equiv a/b \pmod q$ we have

$$\frac{w^{ja} - 2 + w^{-ja}}{w^{jb} - 2 + w^{-jb}} = \sum_{k=-m}^{m} (m - |\, k \,|) w^{jbk},$$

thus

$$u_n(q; bm \bmod q, b) = \frac{1}{2} \sum_{k=-m}^{m} (m - |\, k \,|) V_n(q; bk \bmod q);$$

so we may find the starting values, $u_0, \cdots, u_{D-1}$ given those of $V_0, \cdots, V_{D-1}$. Now, for $0 \le n < \phi(q)/2 = D$, we have

$$\begin{aligned} V_n(q;k) &= \sum_{\substack{j=1 \\ (j,q)=1}}^{q} w^{jkb}(2 - w^j - w^{-j})^n = \sum_{\substack{j=1 \\ (j,q)=1}}^{q} (-1)^n w^{j(bk-n)}(1 - w^j)^{2n} \\ &= \sum_{m=0}^{2n} \binom{2n}{m} (-1)^{n+m} \sum_{\substack{j=1 \\ (j,q)=1}}^{q} w^{j(bk+m-n)} \\ &= \sum_{m=0}^{2n} \binom{2n}{m} (-1)^{n+m} \sum_{d|q, d|m+bk-n} \mu\left(\frac{q}{d}\right) d \\ &= \sum_{d|q} \mu\left(\frac{q}{d}\right) d \sum_{\substack{m=0 \\ m \equiv n - bk \bmod d}}^{2n} \binom{2n}{m} (-1)^{n+m}, \end{aligned}$$

since

$$\sum_{\substack{j=1 \\ (j,q)=1}}^{q} w^{jl} = \sum_{\substack{d|q \\ q|dl}} q \frac{\mu(d)}{d} = \sum_{r|(l,q)} \mu\left(\frac{q}{r}\right) r,$$

taking $r = q/d$. This is computable (though not too beautiful!).

3. Generalized Bernoulli numbers.

For any even character $\chi \pmod q$ define

$$B_{p-1,\chi} = \sum_{a=0}^{q-1} \chi(a) \left(B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \right).$$

Assume that q is prime, so that from Corollary 2 we have for $w = e^{2i\pi/q}$,

$$\begin{aligned} &4pB_{p-1,\chi} \\ &\equiv \sum_{j=1}^{q-1} \left(\sum_{a=0}^{q-1} \chi(a)(2 - w^{ja} - w^{-ja}) \right) \left(\frac{(2 - w^j - w^{-j})^p}{2 - w^{pj} - w^{-pj}} - 1 \right) \pmod{p^2}. \end{aligned}$$

However

$$\sum_{a=0}^{q-1}\chi(a)(2-w^{ja}-w^{-ja}) = \begin{cases} -2\overline{\chi}(j)\sum_{b=0}^{q-1}\chi(b)w^b & \text{for } \chi \text{ non-principal} \\ 2q & \text{for } \chi \text{ principal.} \end{cases}$$

If χ is principal we thus obtain from $(1.7)_q$,

$$2q\sum_{j=1}^{q-1}\left(\frac{(2-w^j-w^{-j})^p}{2-w^{pj}-w^{-pj}}-1\right) \equiv 4pB_{p-1,\chi} = 4pB_{p-1}(q^{1-(p-1)}-q)$$
$$\equiv 4(q^p-q) \pmod{p^2},$$

using the Von Staudt-Clausen theorem. On the other hand if χ is even and non-principal then, for $g(\chi)=\sum_{1\le b\le q}\chi(b)w^b$, we have

$$4pB_{p-1,\chi} \equiv -2g(\chi)\sum_{j=0}^{q-1}\overline{\chi}(j)\left(\frac{(2-w^j-w^{-j})^p}{2-w^{pj}-w^{-pj}}-1\right) \pmod{p^2}.$$

As an example we'll consider χ, the real non-principal character $\pmod q$; that is $\chi(a)=\left(\frac{a}{q}\right)$, the Legendre symbol. We will need q to be $1 \pmod 4$ to ensure that χ is an even character. Then

$$\sum_{a=1}^{q-1}\left(\frac{a}{q}\right)\left\{B_{p-1}\left(\frac{a}{q}\right)-B_{p-1}\right\} = B_{p-1,(\frac{a}{q})} = \frac{-1}{2p}\,g\left(\left(\frac{\cdot}{q}\right)\right)\Sigma_q$$

$$\text{where} \qquad \Sigma_q \equiv \sum_{j=0}^{q-1}\left(\frac{j}{q}\right)\left(\frac{(2-w^j-w^{-j})^p}{2-w^{pj}-w^{-pj}}-1\right) \pmod{p^2}.$$

We will examine Σ_q using p-adic logarithms (see Chapter 5 of [**Wa**] for definitions): Since

$$\frac{(2-w^j-w^{-j})^p}{2-w^{pj}-w^{-pj}}-1 \equiv \log_p\left(\frac{(2-w^j-w^{-j})^p}{2-w^{pj}-w^{-pj}}\right) \pmod{p^2},$$

we deduce that

$$\begin{aligned}\Sigma_q &\equiv \sum_{j=0}^{q-1}\left(\frac{j}{q}\right)\log_p\left(\frac{(2-w^j-w^{-j})^p}{2-w^{pj}-w^{-pj}}\right) \pmod{p^2}\\ &\equiv \log_p\left(\frac{\prod_{j=1}^{q-1}(2-w^j-w^{-j})^{p\left(\frac{j}{q}\right)}}{\prod_{i=1}^{q-1}(2-w^i-w^{-i})^{\left(\frac{pi}{q}\right)}}\right) \pmod{p^2}\\ &\equiv \log_p\left(\prod_{i=1}^{q-1}(2-w^i-w^{-i})^{\left(\frac{i}{q}\right)\left(p-\left(\frac{p}{q}\right)\right)}\right) \pmod{p^2}\\ &\equiv 2\ \log_p\left(\prod_{i=1}^{q-1}(1-w^i)^{\left(\frac{i}{q}\right)\left(p-\left(\frac{p}{q}\right)\right)}\right) \pmod{p^2}\ ,\end{aligned}$$

since $q \equiv 1 \pmod 4$. Now, as Dirichlet discovered (see Ex. 4.6. of [**Wa**]),

$$\prod_{i=1}^{q-1}(1-w^i)^{\left(\frac{i}{q}\right)} = \varepsilon_q^{2h(\sqrt{q})}$$

where $\varepsilon_q, h(\sqrt{q})$ are the fundamental unit and class number of $Q(\sqrt{q})$, respectively. Thus

$$\begin{aligned}\sum_{a=1}^{q-1}\left(\frac{a}{q}\right)\left\{B_{p-1}\left(\frac{a}{q}\right)-B_{p-1}\right\}&\\ \equiv\frac{-1}{p}\ g\left(\left(\frac{\cdot}{q}\right)\right)h\left(\sqrt{q}\right)\log_p\left(\varepsilon_q^{2\left(p-\left(\frac{p}{q}\right)\right)}\right) &\pmod{p}.\end{aligned}$$

So, as $\varepsilon_q = u + v\sqrt{q}$ where $u^2 - v^2 q = -1$, then

$$\varepsilon_q^p \equiv u^p + v^p\sqrt{q}^p \equiv u + v\left(\frac{q}{p}\right)\sqrt{q} \pmod{p}$$

so that $\varepsilon_q^{p-\left(\frac{p}{q}\right)} \equiv \left(\frac{p}{q}\right) \pmod{p}$ and thus $\varepsilon_q^{2(p-\left(\frac{p}{q}\right))} \equiv 1 \pmod{p}$. Suppose that

$$\varepsilon_q^{2\left(p-\left(\frac{p}{q}\right)\right)} = 1 + pu' + pv'\sqrt{q} \pmod{p^2}.$$

Then

$$1 = \varepsilon_q^{2(p-\left(\frac{p}{q}\right))}\overline{\varepsilon}_q^{2(p-\left(\frac{p}{q}\right))} \equiv (1+pu')^2 - (pv'\sqrt{q})^2 \equiv 1 + 2pu' \pmod{p^2},$$

so that p divides u'. So if

$$x_n = \frac{\varepsilon_q^n - \overline{\varepsilon}_q^n}{2v\sqrt{q}} \quad \text{then} \quad \varepsilon_q^{2(p-\left(\frac{p}{q}\right))} \equiv 1 + vx_{2\left(p-\left(\frac{p}{q}\right)\right)}\sqrt{q} \pmod{p^2}.$$

Therefore $\log_p(\varepsilon_q^{2(p-(\frac{p}{q}))}) \equiv vx_{2(p-(\frac{p}{q}))}\sqrt{q} \pmod{p^2}$, and since $g\left(\left(\frac{\cdot}{q}\right)\right) = \sqrt{q}$ (which was proved first by Gauss), we have

$$(3.1)\quad \sum_{a=1}^{q-1}\left(\frac{a}{q}\right)\left\{B_{p-1}\left(\frac{a}{q}\right) - B_{p-1}\right\} \equiv \frac{-q}{p}\, h(\sqrt{q})vx_{2\left(p-\left(\frac{p}{q}\right)\right)} \pmod{p};$$

this is equivalent to (9).

4. Proof of the Theorem.

Take $p \equiv \pm a \pmod{q}$ in Corollary 1 to get

$$B_{p-1}\left(\frac{a}{q}\right) - B_{p-1} \equiv \frac{1}{4p}\sum_{\gamma^q=1}\{(2-\gamma-\gamma^{-1})^p - (2-\gamma^p-\gamma^{-p})\} \pmod{p}.$$

Now $\frac{1}{2}\sum_{\gamma^q=1}(2-\gamma^p-\gamma^{-p}) = q$. Thus, for $x_n = \frac{1}{2q}\sum_{\gamma^q=1}(2-\gamma-\gamma^{-1})^n$, we obtain (8). Now if $0 \le n \le \frac{q-1}{2}$ then

$$\begin{aligned} x_n &= \frac{1}{2q}\sum_{\gamma^q=1}(-\gamma^{-1})^n(1-\gamma)^{2n} \\ &= \frac{1}{2}\sum_{m=0}^{2n}\binom{2n}{m}(-1)^{n+m}\,\frac{1}{q}\sum_{\gamma^q=1}\gamma^{m-n} \\ &= \frac{1}{2}\binom{2n}{n}. \end{aligned}$$

If $w = e^{2i\pi/q}$ then the characteristic polynomial for x_n is

$$\prod_{j=1}^{(q-1)/2}(X - 2 + w^j + w^{-j}).$$

The anonymous referee noted that this polynomial seems to be closely related to the Chebyshev polynomial of the first kind; and we should be able to determine its coefficients directly from known results. Although we agree with this opinion we have been unable to do so. To compute the coefficients we thus proceed as follows: First note that

$$\begin{aligned} &\sum_{j\ge 0}(-1)^j\binom{m-j}{j}(X+X^{-1})^{m-2j} \\ &\quad = \sum_{j\ge 0}(-1)^j\binom{m-j}{j}\sum_{i\ge 0}\binom{m-2j}{i}X^{m-2j-2i} \\ &\quad = \sum_{k\ge 0}X^{m-2k}\left(\sum_{j=0}^{k}\binom{m-j}{j}\binom{m-2j}{k-j}(-1)^j\right) \end{aligned}$$

taking $i + j = k$. The inner sum

$$= \sum_{j=0}^{k} \binom{k}{j}\binom{m-j}{k}(-1)^j$$

$$= \sum_{j=0}^{k} \text{coeff of } T^j \text{ in } (1-T)^k * \text{ coeff of } T^{m-k-j} \text{ in } (1-T)^{-k-1}$$

$$= \text{ coeff of } T^{m-k} \text{ in } (1-T)^{-1} = \begin{cases} 1 & \text{if } k \le m \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$\sum_{j\ge 0}(-1)^j\binom{m-j}{j}(X+X^{-1})^{m-2j} = \sum_{\substack{i=-m \\ i=m \pmod 2}}^{m} X^i \,.$$

So define

$$F_q(y) := \sum_{j\ge 0}(-1)^j\binom{\frac{q-1}{2}-j}{j}(2-y)^{\frac{q-1}{2}-2j} + \sum_{i\ge 0}(-1)^i\binom{\frac{q-3}{2}-i}{i}(2-y)^{\frac{q-3}{2}-2i}. \tag{4.1}$$

Then $F_q(y)$ is a polynomial in y of degree $\frac{q-1}{2}$. For any k, $1 \le k \le \frac{q-1}{2}$,

$$\begin{aligned} F_q(2-w^k-w^{-k}) &= \sum_{j\ge 0}(-1)^j\binom{\frac{q-1}{2}-j}{j}(w^k+w^{-k})^{\frac{q-1}{2}-2j} \\ &\quad + \sum_{i\ge 0}(-1)^i\binom{\frac{q-3}{2}-i}{i}(w^k+w^{-k})^{\frac{q-3}{2}-2i} \\ &= \sum_{l=-(\frac{q-1}{2})}^{\frac{q-1}{2}} w^{kl} = 0 \end{aligned}$$

by (4.1). Thus $F_q(y)$ is our characteristic polynomial, and

$$F_q(y) = \sum_{k=0}^{\frac{q-1}{2}} \frac{f_k}{k!}(-y)^k \qquad \text{where}$$

$$f_k = \sum_{j=0}^{\frac{1}{2}(\frac{q-1}{2}-k)} \frac{(\frac{q-1}{2}-j)!}{j!(\frac{q-1}{2}-2j-k)!}(-1)^j 2^{\frac{q-1}{2}-2j-k}.$$

Actually $F_q(X) = R_{\frac{q-1}{2}}(X)$ where $R_n(X)$ satisfies the recurrence

$$R_m(X) = (2-X)R_{m-1}(X) + R_{m-2}(X).$$

5. Proof of (4).

A *Lucas sequence* $\{x_n\}_{n\ge 0}$ is defined by $x_0 = 0$, $x_1 = 1$ and $x_n = bx_{n-1} - cx_{n-2}$ for all $n \ge 2$, for some integers b and c. As is well-known, if we let $D = b^2 - 4c$ then the roots of the characteristic polynomial $t^2 - bt + c$ of $\{x_n\}$ are $\alpha, \beta = (b \pm \sqrt{D})/2$; and $x_n = (\alpha^n - \beta^n)/(\alpha - \beta)$. Let $y_n = (\alpha^n + \beta^n)$ be the 'companion sequence', which satisfies the same recurrence relation; and we have $\alpha^n, \beta^n = (y_n \pm x_n\sqrt{D})/2$.

We shall be considering these recurrence sequences modulo powers of any prime p that does not divide $2cD$: Now, since p divides $\binom{p}{j}$ except when $j = 0$ or p, we have

$$\left(\frac{b \pm \sqrt{D}}{2}\right)^p \equiv \frac{b^p \pm D^{(p-1)/2}\sqrt{D}}{2^p} \equiv \frac{b \pm (\frac{D}{p})\sqrt{D}}{2} \pmod p.$$

Thus

$$\left(\frac{b \pm \sqrt{D}}{2}\right)^{p-(\frac{D}{p})} \equiv \left(\frac{b \pm (\frac{D}{p})\sqrt{D}}{2}\right)\left(\frac{b \pm \sqrt{D}}{2}\right)^{-(\frac{D}{p})} \equiv c^{\frac{1}{2}(1-(\frac{D}{p}))} \pmod p.$$

Therefore $x_{p-(\frac{D}{p})} \equiv 0 \pmod p$ and

$$c^{p-(\frac{D}{p})} = \alpha^{p-(\frac{D}{p})}\beta^{p-(\frac{D}{p})} = \frac{y^2_{p-(\frac{D}{p})} - Dx^2_{p-(\frac{D}{p})}}{4} \equiv \frac{y^2_{p-(\frac{D}{p})}}{4} \pmod{p^2}.$$

Therefore $y_{p-(\frac{D}{p})} \equiv 2c^{\frac{1}{2}(1-(\frac{D}{p}))}\left(\frac{c^{p-1}+1}{2}\right) \pmod{p^2}$. In fact $c = \pm 1$ in every case below so that

(5.1)
$$\alpha^{p-(\frac{D}{p})}, \beta^{p-(\frac{D}{p})} \equiv \begin{cases} 1 \pm p\sqrt{D}\left(\frac{1}{2p}x_{p-(\frac{D}{p})}\right) \pmod{p^2} & \text{if } c = 1; \\ (\frac{D}{p}) \pm p\sqrt{D}\left(\frac{1}{2p}x_{p-(\frac{D}{p})}\right) \pmod{p^2} & \text{if } c = -1. \end{cases}$$

When $\phi(q) = 4$, we let t be the unique integer in the range $1 < t < q/2$ that is coprime to q. Fix a primitive qth root w of 1, and let $\alpha_j = 2 - w^j - w^{-j}$. By Corollary 2

$$\sum_{d|q} \mu\left(\frac{q}{d}\right) B_{p-1}\left(\frac{a_d}{d}\right) \equiv \frac{1}{2p}\left\{\frac{\alpha_a}{\alpha_p}\,\alpha_1^p + \frac{\alpha_{at}}{\alpha_{pt}}\,\alpha_t^p - (\alpha_a + \alpha_{at})\right\} \pmod p$$

$$\text{(5.2)} \qquad \equiv \frac{1}{2p}\left\{\frac{1}{C'}\left(\alpha_a\alpha_1^{p-(\frac{q}{p})} + \alpha_{at}\alpha_t^{p-(\frac{q}{p})}\right) - B\right\} \pmod p$$

where $B = \alpha_1 + \alpha_t$, $C = \alpha_1\alpha_t$ and $C' = C$ if $\left(\frac{q}{p}\right) = -1$, with $C' = 1$ otherwise.

The cases $q = 5$ and $q = 10$: When $q = 5$ we have $t = 2$, $(x-\alpha_1)(x-\alpha_2) = x^2 - 5x + 5$, so that $B = C = 5$ and we may take $\alpha_j = \frac{1}{2}\sqrt{5}\left(\sqrt{5} + \left(\frac{j}{5}\right)\right)$ for $1 \le j \le 4$. Let $\alpha = (1+\sqrt{5})/2$ and $\beta = (1-\sqrt{5})/2$. By substituting into (5.2) and then using (5.1) (with $b = 1$, $c = -1$ so that $x_n = F_n$) we get

$$\begin{aligned} &B_{p-1}\left(\frac{a}{5}\right) - B_{p-1} \\ &\quad \equiv \frac{5}{2p}\left\{\frac{5^{(p-1)/2}}{2\sqrt{5}}\left(\left(\sqrt{5}+\left(\frac{a}{5}\right)\right)\alpha^{p-\left(\frac{p}{5}\right)} + \left(\sqrt{5}-\left(\frac{a}{5}\right)\right)\beta^{p-\left(\frac{p}{5}\right)}\right) - 1\right\} \\ &\quad \equiv \frac{5}{2p}\left\{5^{(p-1)/2}\left(\left(\frac{p}{5}\right) + p\frac{1}{2p}F_{p-\left(\frac{p}{5}\right)}\left(\frac{a}{5}\right)\right) - 1\right\} \\ &\quad \equiv \frac{5}{4}\left\{\left(\frac{5^{p-1}-1}{p}\right) + \left(\frac{ap}{5}\right)\frac{1}{p}F_{p-\left(\frac{p}{5}\right)}\right\} \pmod p \end{aligned}$$

giving the first congruence in (4), since

$$5^{\frac{p-1}{2}} \equiv \left(\frac{5}{p}\right)\left(1 + \frac{1}{2}\left(5^{p-1}-1\right)\right) \pmod{p^2}.$$

It would be possible to obtain the congruence for $q = 10$ in a similar way. However, by taking $q = 2$ and $a = 1/5$ and $a = 3/5$ in (1.6) we get the identities

$$\begin{aligned} B_{p-1}\left(\frac{1}{10}\right) &= 2^{2-p}B_{p-1}\left(\frac{1}{5}\right) - B_{p-1}\left(\frac{3}{5}\right) \\ B_{p-1}\left(\frac{3}{10}\right) &= 2^{2-p}B_{p-1}\left(\frac{3}{5}\right) - B_{p-1}\left(\frac{4}{5}\right). \end{aligned}$$

By substituting in the first congruence in (4), and by using the Von Staudt-Clausen theorem, we get the third congruence in (4).

The case $q = 8$: Now $t = 3$, $(x-\alpha_1)(x-\alpha_3) = x^2 - 4x + 2$, so that $B = 4$, $C = 2$ and we may take $\alpha_j = \sqrt{2}\left(\sqrt{2} + \left(\frac{8}{j}\right)\right)$ for any odd j. Let $\alpha = (1+\sqrt{2})$ and $\beta = (1-\sqrt{2})$. By substituting into (5.2) and then using (5.1) (with $b = 2$, $c = -1$ so that $x_n = G_n$), we see that the right side of (5.2) is

$$\begin{aligned} &\equiv \frac{2}{p}\left\{\frac{2^{(p-1)/2}}{2\sqrt{2}}\left(\left(\sqrt{2}+\left(\frac{8}{a}\right)\right)\alpha^{p-\left(\frac{8}{p}\right)} + \left(\sqrt{2}-\left(\frac{8}{a}\right)\right)\beta^{p-\left(\frac{8}{p}\right)}\right) - 1\right\} \\ &\equiv \frac{2}{p}\left\{2^{(p-1)/2}\left(\left(\frac{8}{p}\right) + p\frac{1}{p}G_{p-\left(\frac{8}{p}\right)}\left(\frac{8}{a}\right)\right) - 1\right\} \\ &\equiv \left(\frac{2^{p-1}-1}{p}\right) + 2\left(\frac{8}{ap}\right)\frac{1}{p}G_{p-\left(\frac{8}{p}\right)} \pmod p \end{aligned}$$

since $2^{\frac{p-1}{2}} \equiv \left(\frac{8}{p}\right)(1+\frac{1}{2}(2^{p-1}-1)) \pmod{p^2}$. Adding this to the third congruence in (3) gives the second congruence in (4).

The case $q = 12$**:** Now $t = 5$, $(x-\alpha_1)(x-\alpha_3) = x^2 - 4x + 1$, so that $b = B = 4$, $c = C = 1$ and we may take $\alpha_j = 2 + \left(\frac{12}{j}\right)\sqrt{3}$ for $j = 1, 5, 7, 11$; and let $\alpha = \alpha_1$, $\beta = \alpha_2$. Therefore, by using (5.1), the right side of (5.2) is

$$\equiv \frac{1}{2p}\left\{\left(\left(2+\left(\frac{12}{a}\right)\sqrt{3}\right)\alpha^{p-\left(\frac{12}{p}\right)} + \left(2-\left(\frac{12}{a}\right)\sqrt{3}\right)\beta^{p-\left(\frac{12}{p}\right)}\right) - 4\right\}$$
$$\equiv 3\left(\frac{12}{a}\right)\frac{1}{p}H_{p-\left(\frac{12}{p}\right)} \pmod{p}.$$

The final congruence of (4) follows by adding the last two congruences of (3) and subtracting the first.

References

[AM] G. Almkvist and A. Meurman, *Values of Bernoulli polynomials and Hurwitz's zeta function at rational points*, C.R. Math. Rep. Acad. Sci. Canada, **13** (1991), 104–108.

[An] G.H. Andrews, *Some formulae for the Fibonacci sequence with generalizations*, Fib. Quart., **7** (1969), 113–130.

[FT] A. Friedmann and J. Tamarkine, *Quelques formules concernent la théorie de la fonction* $[x]$ *et des nombres de Bernoulli*, J. Reine Angew. Math., **137** (1909), 146–156.

[Gr] A. Granville, *On the Kummer-Wieferich-Skula criteria for the first case of Fermat's Last Theorem*, in 'Advances in Number Theory', ed. F.Q. Gouvêa and N. Yui (Oxford, Midsomer Norton, 1993), 479–498.

[KS] A.A. Kiselev and I.Š. Slavutskiĭ, *On the number of classes of ideals of a quadratic field and its rings*, (Russian), Dokl. Akad. Nauk SSSR, **126** (1959), 1191–1194.

[Lh] E. Lehmer, *On congruences involving Bernoulli numbers and the quotients of Fermat and Wilson*, Ann. of Math., **39** (1938), 350–360.

[Lr] M. Lerch, *Zur Theorie des Fermatschen Quotienten* $(a^{p-1}-1)/p = q(a)$, Math. Ann., **60** (1905), 471–490.

[Ri] P. Ribenboim, 13 *Lectures on Fermat's Last Theorem*, (Springer, New York, 1979).

[Sk] L. Skula, *Fermat's Last Theorem* (1^{st} *Case*) *and the Fermat quotients*, Comm. Math. Univ. Sancti. Pauli, **41** (1992), 35–54.

[SS] Z-H Sun and Z-W Sun, *Fibonacci numbers and Fermat's Last Theorem*, Acta Arith., **60** (1992), 371-388.

[Wa] L.C. Washington, *Introduction to Cyclotomic Fields*, (Springer, New York, 1982).

[W1] H. C. Williams, *A note on the Fibonacci quotient* $F_{p-\epsilon/p}$, Can. Math. Bull., **25** (1982), 366-370.

[W2] ———, *Some formulas concerning the fundamental unit of a real quadratic field*, Disc. Math., **92** (1991), 431-440.

Received July 21,1993.

University of Georgia
Athens, Ga 30602
E-mail address: andrew@math.uga.edu

AND

Nanjing University
Nanjing 210008, P. R. CHINA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

THE UNIQUENESS OF COMPACT CORES FOR 3-MANIFOLDS

LUKE HARRIS AND PETER SCOTT

A compact core for a 3-manifold M is a compact sub-manifold N of M whose inclusion in M induces an isomorphism of fundamental groups. A uniqueness result for compact cores of orientable 3-manifolds is known. The authors show that compact cores are not unique in any reasonable sense for non-orientable 3-manifolds, but they prove a finiteness result about the number of possible cores.

If M is a non-compact 3-manifold with finitely generated fundamental group, then Scott showed in [**Sc1**] that there is a compact sub-manifold N of M with the natural map $\pi_1(N) \to \pi_1(M)$ an isomorphism. See [**R-S**] for a simpler proof. We call such a sub-manifold a *core* or *compact core* for M. In [**McC-Mi-Sw**], McCullough, Miller and Swarup showed that if N_1 and N_2 are irreducible compact cores of a $\mathbb{P}^2$-irreducible 3-manifold M, then N_1 and N_2 are homeomorphic. In this paper, we seek to generalize this to the case when M and its compact cores have no irreducibility restrictions. Of course, we cannot any longer expect to prove that two cores of M are homeomorphic, because the Poincaré conjecture is not resolved. Thus one core for M might be the connected sum of another core with a homotopy sphere. Also we can obtain new cores by removing a 3-ball from a core or by replacing a connected summand of a core which is a 2-sphere bundle over the circle by a disc bundle over the circle . However, we give an example showing that even if one works modulo the equivalence relation on cores generated by the above operations then uniqueness does not hold. We also show that there are only finitely many different cores in a given 3-manifold up to the equivalence relation of almost homeomorphism which we define in §1. We end by using this finiteness result to prove a natural finiteness result for the boundary of a 3-manifold which has finitely generated fundamental group. The result is the following.

Theorem 3.2. *Let M be a 3-manifold with finitely generated fundamental group. Then*

(i) *There are only finitely many boundary components F of M with $\operatorname{Im}(\pi_1(F) \to \pi_1(M))$ not trivial or infinite cyclic,*

(ii) *There are only finitely many boundary components F of M with* $\mathrm{Im}\,(\pi_1(F) \to \pi_1(M))$ *infinite cyclic and with essential core a Möbius band,*

(iii) *Of those components of the boundary F_i with* $\mathrm{Im}(\pi_1(F_i) \to \pi_1(M))$ *infinite cyclic and with essential core an annulus, there are only finitely many conjugacy classes in* $\pi_1(M)$ *of* $\mathrm{Im}(\pi_1(F_i) \to \pi_1(M))$.

In a separate paper [**H-S**], we use Theorem 3.2 to extend earlier results of Brin, Johannson and Scott [**BJS**] on compact totally peripheral 3-manifolds to the non-compact case.

The work in this paper is part of the Liverpool Ph.D. thesis of Luke Harris completed under the supervision of Peter Scott in 1988. Since then Harris obtained a job not in the academic world and has never had time to prepare this for publication. Finally, Scott agreed to prepare this for publication, to avoid the complete disappearance of his work.

§1. Preliminaries and the example.

Definition. Let M and N be compact 3-manifolds. Then M and N are *almost homeomorphic* if they are homeomorphic up to connected sum with compact simply connected manifolds (3-balls and fake 3-spheres) and up to replacing $\mathbb{P}^2 \times I$'s with fake $\mathbb{P}^2 \times I$'s.

We start with our example to show non-uniqueness for cores. Let M be a 3-manifold with finitely generated fundamental group, and with core N in $\overset{\circ}{M}$. Suppose that $M - \overset{\circ}{N}$ has at least two components R_1 and R_2, and let F_1 and F_2 be the components of ∂N which lie in R_1 and R_2. Let X denote the solid torus. Note that similar examples can be constructed if X is any compact manifold with at least one boundary component not the 2-sphere. We can form the connected sum $M\#X$ in several ways, depending on the choice of 3-balls in M and in X. However, $M\#X$ is independent of this choice.

If we form $M\#X$ by selecting a 3-ball in M which lies in the region R_1, then a natural selection of core N_1 for $M\#X$ is N with a 1-handle attached to F_1. If we select the 3-ball in R_2, then the core N_2 could be N with a 1-handle attached to F_2. We could also select the 3-ball to lie in N, in which case the natural core N_0 would be $N\#X$. Note that this is homeomorphic to N with the interior of a trivial solid torus removed, where trivial means that the solid torus lies in a 3-ball in N and is unknotted there. Also note that N_1 and N_2 are each a boundary connected sum of N and X.

So long as F_1 and F_2 are not homeomorphic, and not 2-spheres, these three cores are non-homeomorphic in a fairly non-trivial way. They have different

boundary, for example, and the difference in boundary is not caused just by adding or removing 2-sphere components. Thus they are not almost homeomorphic. Further, so long as F_1 and F_2 have genus at least two, these cores cannot be equivalent under the coarser equivalence relation obtained from the operation of replacing a summand which is a sphere bundle over the circle by a solid torus. However, they do have certain similarities. All contain pieces homeomorphic with X and N, at least up to connected sum with 3-balls, and these pieces are connected by 1-handles and $S^2 \times I$'s.

We cannot avoid this problem even if we insist that one of the cores be embedded in the other, since in the example given it is possible to embed either of N_1 and N_2 inside N_0. To see how to embed N_1 in N_0, for example, consider a simple closed curve α on F_1 which bounds a disc D in F_1 containing the endpoints of the 1-handle of N_1. We can also find a disc E in R_1 with boundary α, such that $E \cup D$ defines a 2-sphere in R_1 which is parallel to the boundary of the 3-ball we removed from R_1 to form $M\#X$. Then N_1 together with a regular neighbourhood of E is homeomorphic to N minus a trivial solid torus and hence homeomorphic to N_0.

It is clear that this example is not a special case, and that any connected sum between two 3-manifolds with cores having non-spherical boundary may have a number of non-homeomorphic cores, constructed similarly to N_0, N_1 and N_2 above.

Now we will need a few definitions. The first four are after Scott in [**Sc2**].

Definition. A sub-manifold X of a 3-manifold M is *incompressible* if ∂X is incompressible in M.

Remark. Then the natural map $\pi_1(X) \to \pi_1(M)$ is injective.

Definition. A 3-manifold N is *weakly irreducible* if the manifold $\hat{N}$ obtained from N by attaching a 3-ball to every boundary 2-sphere of N is irreducible.

Definition. A *chunk* in a 3-manifold is a sub-manifold X of M which is connected, compact, incompressible and weakly irreducible.

Remark. With this definition, a (punctured) 2-sphere bundle over the circle is not a chunk.

Definition. Let M be a 3-manifold, with fundamental group $G = \pi_1(M)$, and let H be a finitely generated indecomposable subgroup of G. Then a *chunk* in M for H is a chunk X in M such that $\pi_1(X)$ contains a conjugate (in G) of H.

Observe that if N is a compact core of a 3-manifold M, then we can decompose N into chunks by cutting along a maximal family of 2-spheres embedded in N and then cutting along compressing discs for the boundaries of the resulting pieces. Thus N can be viewed as a collection of chunks

embedded in M joined together by 1-handles and $S^2 \times I$'s.

Clearly if $\pi_1(M) = F_r * G_1 * G_2 * \ldots * G_s$ where F_r is free and the G_i are indecomposable and not infinite cyclic, then there must be exactly s chunks X_i for a core N that have non trivial fundamental group, with $\pi_1(X_i) = G_i$ up to conjugacy after reordering. Thus the non trivial chunks for any two cores are in 1-1 correspondence. The main result of this section is that these chunks are almost homeomorphic.

Theorem 1.1. *Let X' and X be chunks in a 3-manifold M with finitely generated fundamental group, and suppose that $\pi_1(X')$ and $\pi_1(X)$ are both conjugate to H, an indecomposable factor not $\mathbb{Z}$ in a free product decomposition of $\pi_1(M)$. Then X' and X are almost homeomorphic.*

Remark. Such chunks cannot contain fake 3-balls, but may contain fake $\mathbb{P}^2 \times I$'s.

As a first step, we prove the following special case of Theorem 1.1.

Lemma 1.2. *The result of Theorem* 1.1 *holds if either X and X' are disjoint or if one lies in the interior of the other.*

Proof. First consider the case when X lies in the interior of X'. We know that $\pi_1(X)$ and $\pi_1(X')$ are both conjugate to H, and that $\pi_1(X)$ is a subgroup of $\pi_1(X')$. But $H = \pi_1(X)$ is an indecomposable factor of $\pi_1(M)$, and so no conjugate of H can be properly contained in H. Thus we deduce that the inclusion of X in X' induces an isomorphism of fundamental groups.

Consider a component R of $X' - \overset{\circ}{X}$. Clearly $R \cap X = F$, a single boundary component of X. But we must also have $\pi_1(R) = \pi_1(F)$ by van Kampen's theorem, since $\pi_1(X') = \pi_1(X)$. Then the h-cobordism theorem tells us that R must be homeomorphic to $F \times I$ connected sum with 3-balls unless $F \cong \mathbb{P}^2$, in which case R might be a fake $\mathbb{P}^2 \times I$ connected sum with 3-balls. (Note that in fact we cannot have fake 3-balls in R, since non-simply connected chunks do not contain fake 3-balls.) This is true for all components of $X' - \overset{\circ}{X}$, and so we conclude that X' is almost homeomorphic to X.

Next consider the case when X and X' are disjoint. Take the cover M_H of M with fundamental group H. Then X and X' lift into M_H. Thus H is the fundamental group of a graph of groups with $\pi_1(X)$ and $\pi_1(X')$ as vertex groups. As each of these vertex groups equals H, it follows that there is a path between these vertices with all edge and vertex groups equal to H. In particular, for X and X' there are boundary components F and F' respectively with $\pi_1(F) \to \pi_1(X)$ and $\pi_1(F') \to \pi_1(X')$ being isomorphisms. We apply the h-cobordism theorem to see that both X and X' are almost

homeomorphic to products $F \times I \cong F' \times I$. Thus X and X' must be almost homeomorphic. This completes the proof of Lemma 1.2. □

Proof of Theorem 1.1. We may suppose that ∂M is empty, by pushing X and X' away from ∂M and then deleting ∂M. Then we can find a compact sub-manifold K with X and X' embedded in the interior of K. The proof proceeds by altering K in a fairly canonical way until we find a chunk C_1 derived from K with X embedded in C_1. Lemma 1.2 then shows that X and C_1 must be almost homeomorphic. We may also alter K in a slightly different way to obtain a chunk C_1' containing X', and so this chunk must be almost homeomorphic to X'. But because we obtained the chunks in each case in a fairly standard way, we will be able to show that they must themselves be almost homeomorphic, which will complete the proof.

So for the moment, we will consider X only. Consider a family of 2-spheres Σ embedded in K corresponding to a prime decomposition of K. Note that the pieces obtained by splitting along any such family are unique up to homeomorphism and connected sum with 3-balls. In particular, the pieces are unique up to almost homeomorphism. We will alter Σ so that it does not intersect X, but so that it remains a representation of a prime decomposition of K.

Σ intersects ∂X in a collection of embedded circles. ∂X is incompressible, so we can choose an innermost circle C of $\Sigma \cap \partial X$ bounding a disk E in ∂X, with $E \cap \Sigma = \partial E = C$.

Now we cut and paste Σ along C using E, and push the new Σ away from ∂X on both sides. After deleting any redundant 2-spheres, Σ still represents a decomposition of K into primes. We repeat until $\Sigma \cap \partial X$ is empty.

Some components of Σ may lie inside X. We delete these components from Σ and replace them with the spherical boundary components of X, and then again delete any redundant 2-spheres. Since X is weakly irreducible, any 2-sphere embedded in X other than a boundary sphere must be redundant, and thus Σ still represents a decomposition of K into primes. So we may cut K along this new Σ, and X will be contained in one of the pieces.

We now wish to compress the boundary of the pieces, whose union we will continue to call K. We do this by sequentially finding compressing discs for ∂K, lying in either K or $M - \overset{\circ}{K}$. If a disc D lies outside K, then we add a regular neighborhood of the disc to K. If the disc is contained in K, then we cut along it. Note that, so long as the boundaries of the discs and the order they are dealt with is the same, any sequence of compressing discs yields pieces unique up to homeomorphism and connected sum with 3-balls. This is because the components of K are weakly irreducible, and any two discs with the same boundary embedded in an irreducible 3-manifold are isotopic.

Consider a disc D lying in K. Then D intersects ∂X in a collection of embedded circles. ∂X is incompressible, so we can find an innermost circle C bounding a disc E in ∂X with $D \cap E = \partial E = C$.

We cut and paste D along C using E, and discard the 2-sphere component of the result. Then we push D away from ∂X,thus reducing the intersection number of D with ∂X. We can repeat until $D \cap \partial X = \emptyset$ and thus $D \cap X = \emptyset$ since $\partial D \subset K - X$. Note that this does not disturb ∂D, so D is still a compressing disc for ∂K contained in K.

So after compressing the boundary of all the components, we have derived from the original K a collection of chunks, one of which contains X. Call this chunk C_1. Lemma 1.2 shows that X and C_1 are almost homeomorphic.

If there is another chunk C_2 derived from K which also has fundamental group conjugate to H, then Lemma 1.2 again shows that it is also almost homeomorphic to X.

Now all the above construction can be done to find a number of chunks C_i' for K corresponding to X'. As we have already pointed out, we must get the same chunks up to homeomorphism and connected sum with 3-balls as we did with the first construction. Thus C_i' is almost homeomorphic to C_j for some j. It follows that X is almost homeomorphic to X', which completes the proof of Theorem 1.1. □

§2. There are only finitely many compact cores for 3-manifold.

In §1, we gave an example which rules out the possibility of a uniqueness result such as that to be found in the paper [**McC-Mi-Sw**] of McCullough, Miller and Swarup. In this section we show that, up to almost homeomorphism, there are only finitely many cores for any 3-manifold M.

Theorem 2.1. *Let M be a 3-manifold with finitely generated fundamental group. Then, up to almost homeomorphism, there are only finitely many different cores for M.*

Proof. The proof uses the result of the previous section on uniqueness of non-simply connected chunks in a 3-manifold. Since all cores have a decomposition into essentially the same chunks, then these same chunks can be used as the building blocks to construct any core for M. We can then show that there are, up to almost homeomorphism, only finitely many ways to put these chunks together to give a compact 3-manifold with fundamental group equal to $\pi_1(M)$. Of course, if chunks were unique up to homeomorphism, it would be trivial that there could only be a finite number of cores up to homeomorphism.

So first consider the decomposition of a compact core N into chunks, by splitting along 2-spheres and discs. Then N is the union of these chunks together with 1-handles and $S^2 \times I$'s. If M has fundamental group $\pi_1(M) = G = F_r * G_1 * G_2 * \cdots * G_s$ where F_r is free of rank r and the G_i are indecomposable not infinite cyclic, then we can decompose N into exactly s chunks C_i that are not simply connected, and such that $\pi_1(C_i) = G_i$ up to conjugacy in G. We apply the result of §1 to see that, up to almost homeomorphism, all cores for M have exactly the same non simply connected chunks. It will be useful to choose our splitting family of spheres and discs to be minimal in the sense that no proper subfamily splits N in this way.

Using the decomposition mentioned in the introduction, we first cut sequentially along 2-spheres $\{S_j\}$, and then along discs $\{D_k\}$. We can easily arrange that all the discs and spheres can be embedded in N, and are disjoint from one another. Thus the order in which we cut along the spheres and discs is irrelevant. Since the construction of a core from the chunks is essentially the reverse operation to that of cutting along the spheres and discs, we will find it useful to choose a particular order in which to decompose N.

We now wish to organize the decomposition into four steps. In step one, we cut along non-separating 2-spheres. In step two, we cut along discs which correspond to 1-handles attached to spherical boundary components of chunks. In step three, we cut along separating 2-spheres, and finally in step four we split along the remaining discs. We will comment on each stage of the decomposition.

Step 1 : *Cutting along the non-separating* 2*-spheres.*

Let S be such a non-separating 2-sphere. Then we can find a regular neighborhood $S \times I$ for S. Since S is non-separating, we cam find an arc λ in the complement of $S \times I$ joining $S \times 0$ to $S \times 1$. $S \times I$ together with a regular neighborhood of λ defines a punctured 2-sphere bundle over the circle. Thus every non-separating sphere in the family S_i corresponds to a 2-sphere bundle over the circle in a prime decomposition of N.

Step 2 : *Cutting along discs corresponding to* 1*-handles attached to spherical boundary components of chunks.*

This step is much simpler that it sounds. Let D be a disc in the family $\{D_k\}$. We can cut along all the other spheres and discs in the decomposition leaving D until last. Then D corresponds to a 1-handle with ends attached to discs in the boundary of the chunks. Since D is a compressing disc, if one end of the 1-handle is attached to a spherical boundary component S of a chunk, it must be that the other end is also attached to S, since S is separating and the family of discs $\{D_k\}$ is minimal. Any other disc has a corresponding 1-handle with both ends attached to non-spherical boundary

components of chunks.

We can cut along all the discs which correspond to 1-handles with both ends in S, and so we see that N can be thought of as a simpler manifold with boundary including S, with one or more 1-handles attached to S.

Step 3 : *Cutting along the separating* 2*-spheres.*

This needs little attention for the moment. Note that cutting along such a 2-sphere corresponds to decomposing as a connected sum (up to adding or removing 3-balls, anyway).

Step 4 : *Cutting along the remaining discs.*

As we noted in step 2, all such discs must correspond to 1-handles with both ends in non-spherical boundary components of chunks. Of course in particular this means that these 1-handles are not attached to simply connected chunks.

We are now ready to reverse the decomposition process. Let K denote the disjoint union of the non simply connected chunks C_i. Then any core, up to almost homeomorphism, is obtained from K by adding 1-handles, $S^2 \times I$'s and also simply connected compact manifolds. Eventually we construct a compact 3-manifold from K by reversing steps one to four of the decomposition process.

We consider the steps of the decomposition individually, and in reverse order. To start with, K is uniquely determined up to almost homeomorphism. At each stage, we must ensure that there are only finitely many possibilities for K, up to almost homeomorphism.

Step 4:

To reverse this, we add 1-handles to K. Let H be such a 1-handle. Each end of H is connected to a non spherical boundary component of a non simply connected chunk. Thus we do not need to consider simply connected chunks at this stage.

There are only s non-simply connected chunks in K, and each chunk has only finitely many boundary components, and so there are only finitely many ways to attach each end of H, and hence there are only finitely many ways of attaching H to K, since the ends of H and the orientation of H are the only factors in determining the homeomorphism class of the result. Note that we have used the fact that H must be connected to non-spherical boundary components of chunks, and thus it is irrelevant that K is defined only up to almost homeomorphism.

Step 3:

In this step, we add $S^2 \times I$'s to K, but since one end of an $S^2 \times I$ can be attached to a simply connected chunk, we must also add these. Remember

that all the $S^2 \times I$'s in this stage connect different components of K, and the end result of this is connected.

Consider an $S^2 \times I$ with either (or both) ends attached to a simply connected chunk. Then it corresponds to forming a connected sum, with one of the summands being a simply connected compact 3-manifold. Thus, up to almost homeomorphism, we have done nothing.

Now we consider $S^2 \times I$'s which connect non simply connected chunks. Adding such an $S^2 \times I$ corresponds to forming a connected sum, since each end of the $S^2 \times I$ is connected to a different component of K. If K has n non simply connected components, then we must add $n-1$ $S^2 \times I$'s in this step, and however we do this, the result is unique up to almost homeomorphism. In particular, it is irrelevant which 2-sphere boundary components of K we choose to attach an $S^2 \times I$ to, and also it is irrelevant which components of K the $S^2 \times I$ is attached to, since the end result of step 3 is a connected manifold.

Step 2:

In this step, we attach 1-handles which have both ends in the same spherical boundary component of a chunk. Note that if we wish to add one such 1-handle, we have a choice of 2-sphere boundary components on which to attach it, but, assuming orientability, any choice gives the same result up to homeomorphism. If we are adding many 1-handles, we do not care which 2-spheres we attach them to, but only which 1-handles we allow to share the same boundary 2-spheres, which is a combinatorial question.

If we allow non-oriented 1-handles, there are more possibilities, but there are still only finitely many different ways of adding the 1-handles to K, up to almost homeomorphism of the resulting manifold.

Step 1:

As we noted earlier, this is equivalent to forming a connected sum with 2-sphere bundles over the circle, up to almost homeomorphism. Each 2-sphere bundle could be orientable or non-orientable.

We have $\pi_1(M) = G = F_r * G_1 * G_2 * \cdots * G_s$. Thus there are $(r+s-1)$ 1-handles, non-trivial $S^2 \times I$'s and 2-sphere bundles over the circle to be added to the chunks C_i to get a compact connected manifold with fundamental group G. Hence each of the steps one to four must terminate in fewer than $r+s-1$ steps, since the 1-handles, $S^2 \times I$'s and S^2-bundles added in these steps are precisely those needed to get a compact manifold with fundamental group G.

So after $r+s-1$ stages, and with a finite number of possibilities at each stage, we have a compact manifold with fundamental group G. Thus there are, up to almost homeomorphism, only finitely many compact 3-manifolds

that can be formed from the C_i with fundamental group G, and so M can have only finitely many different cores up to almost homeomorphism. This completes the proof of Theorem 2.1. □

§3. Applications of uniqueness of cores.

In this section we use Theorem 2.1 together with results of McCullough [**McC**] on compact cores to deduce information about the boundary of a 3-manifold M with finitely generated fundamental group.

The first result is an obvious deduction from Theorem 2.1.

Corollary 3.1. *Let M be a 3-manifold with finitely generated group, and let X_i be an infinite sequence of cores of M. Then there is a subsequence X_j of X_i such that all members of the sequence X_j are almost homeomorphic.*

Before we get to the main result of this section, we need a definition. Let M be a 3-manifold with finitely generated fundamental group G. Let F be a component of ∂M, and let $H = \mathrm{Im}(\pi_1(F) \to \pi_1(M))$ be the image of the fundamental group of F under the natural induced map into G. Then H is finitely generated, by the result of Jaco in [**Ja**]. Now we can take a compact regular neighborhood of based loops in F representing the generators of H, and add compressing discs in F to get a compact subsurface C of F with $\mathrm{Im}(\pi_1(c) \to \pi_1(M)) = H$, and C incompressible in F. Then:

Definition. With M, F and C as above, we call C an essential core for F.

Remark. C need not be incompressible in M. Also, if $\mathrm{Im}(\pi_1(F) \to \pi_1(M))$ is infinite cyclic, we can choose a simple closed curve on F to represent a generator, and thus we may choose the essential core in this case to be an annulus or a Möbius band. We assume in what follows that we always choose such an essential core when possible.

We can now state the main result of this section. McCullough gives a result equivalent to part (i) and (ii) of this theorem in the case when ∂M is incompressible as a corollary to his main theorem in [**McC**]. See also [**R-S**].

Theorem 3.2. *Let M be a 3-manifold with finitely generated group. Then:*

(i) *There are only finitely many boundary components F of M with $\mathrm{Im}(\pi_1(F) \to \pi_1(M))$ not trivial or infinite cyclic,*

(ii) *There are only finitely many boundary components F of M with $\mathrm{Im}(\pi_1(F) \to \pi_1(M))$ infinite cyclic and with essential core a Möbius band,*

(iii) *Of those components of the boundary F_i with $\mathrm{Im}(\pi_1(F_i) \to \pi_1(M))$ infinite cyclic and with essential core an annulus, there are only finitely many conjugacy classes in $\pi_1(M)$ of $\mathrm{Im}(\pi_1(F_i) \to \pi_1(M))$.*

Proof. Let the boundary components of M be F_i. Then we can find essential cores C_i for the F_i, with the C_i being annuli or Möbius bands when possible. We are only interested in those F_i with non-trivial image in $\pi_1(M)$, so we will assume that none of the C_i are discs. Now the theorem of McCullough [**McC**] tells us that given a (possibly disconnected) compact subsurface C of the boundary of a manifold M, we can find a compact core X for M with $X \cap \partial M = C$. Thus a manifold M, we can find a sequence of cores X_i for M with $X_i \cap \partial M = \bigcup_{j=1}^{i} C_j$.

By taking a subsequence if necessary, we can assume that this sequence is stable, i.e. that all the X_i are almost homeomorphic. Let dX_i denote the union of all the non-spherical boundary components of X_i. Then dX_i is homeomorphic to a fixed closed surface dX, for all i. Now for any union of non-trivial essential cores of boundary components of M, we can find an embedding of these essential cores in dX_i, for some i, and hence an embedding in dX. Since dX has only finitely many components, we immediately see that there can only be a finite number of closed surfaces embedded in dX. So we are only concerned with subsurfaces of dX which are not closed. Also we can assume that the subsurfaces are injective in dX, by adding discs lying in dX to them. Thus none of the subsurfaces is a disc, and none of the components of the complement of the subsurfaces is a disc.

Assume for the moment that dX is connected. Consider its Euler characteristic $\chi(dX)$. Any collection of embedded disjoint subsurfaces $\{C_i : i \leq m\}$ of dX splits dX into a collection of subsurfaces $\{C_i : i \leq n\}$ where $\{C_i : m+1 \leq i \leq n\}$ are the components of the complement of $\{C_i : i \leq m\}$. Then $\chi(dX) = \sum_{i=1}^{n} \chi(C_i)$. None of the C_i are discs (or are closed), so $\chi(C_i) \leq 0 \ \forall i$. Also, any C_i which are not Möbius bands or annuli have negative Euler characteristic, and so dX can contain at most $|\chi(dX)|$ such surfaces. Of course, when dX is not connected this holds for any component of dX. This proves the first part of the theorem.

Similarly dX can contain only a finite number of Möbius bands, this time limited by the rank of $H_1(dX, \mathbb{Z}_2)$, and so also there are only a finite number of boundary components of M with essential core a Möbius band. This completes part two of the theorem.

Consider now essential cores which are annuli. There is no bound on the number of these that can be embedded in a surface dX. However, we can embed only finitely many non-parallel such annuli. Parallel annuli have fundamental groups which are conjugate in $\pi_1(M)$, so this completes the third part of the theorem. □

References

[BJS] M.Brin, K. Johannson and G.P. Scott, *Totally peripheral 3-manifolds*, Pacific J. Math., **118** (1985), 37-51.

[H-S] L. Harris and P. Scott, *Non-compact totally peripheral 3-manifolds*, to appear in Pacific J. Math.

[Ja] W. Jaco, *Lectures on 3-manifold topology*, C.B.M.S. Regional conference series in mathematics number 43.

[McC] D. McCullough, *Compact submanifolds of 3-manifolds with boundary*, Quart. J. Math Oxford (2), **37** (1986), 299-307.

-Mi-Sw] D. McCullough, A. Miller and G.A. Swarup, *Uniqueness of cores of non-compact 3-manifolds*, J. London. Math. Soc. (2), **32** (1985), 548-556.

[R-S] J.H. Rubinstein and G.A. Swarup, *On Scott's core theorem*, Bull London Math. Soc., **22** (1990), 495-498.

[Sc1] G.P. Scott, *compact submanifolds of 3-manifolds*, J. London Math. Soc. (2), **7** (1973), 246-250.

[Sc2] ———, *fundamental groups of non-compact 3-manifolds*, Proc. London Math. Soc. (3), **34** (1977), 303-326.

Received February 15, 1993 and revised July 25, 1993. The second author partially supported by NSF grant DMS 90-03974.

University of Michigan
Ann Arbor, MI 48109
E-mail address: pscott@umich.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

ESTIMATION OF THE NUMBER OF PERIODIC ORBITS

BOJU JIANG

The main theme of this paper is to estimate, for self-maps $f : X \to X$ of compact polyhedra, the asymptotic Nielsen number $N^\infty(f)$ which is defined to be the growth rate of the sequence $\{N(f^n)\}$ of the Nielsen numbers of the iterates of f. The asymptotic Nielsen number provides a homotopy invariant lower bound to the topological entropy $h(f)$. To introduce our main tool, the Lefschetz zeta function, we develop the Nielsen theory of periodic orbits. Compared to the existing Nielsen theory of periodic points, it features the mapping torus approach, thus brings deeper geometric insight and simpler algebraic formulation. The important cases of homeomorphisms of surfaces and punctured surfaces are analysed. Examples show that the computation involved is straightforward and feasible. Applications to dynamics, including improvements of several results in the recent literature, demonstrate the usefulness of the asymptotic Nielsen number.

Introduction.

Motivated by dynamical problems, Nielsen theory of fixed points of self-maps $f : X \to X$ of compact polyhedra was generalized to study periodic points, i.e. solutions to $f^n(x) = x$, where f^n is the n-th iterate. See e.g. [**J1**, §III.4]. As the Nielsen number $N(f)$ is a homotopy invariant lower bound to the number of fixed points of f, the Nielsen number $N(f^n)$ is certainly a lower bound to the number of n-points (i.e. fixed points of the n-th iterate) for any map g homotopic to f.

However, generally speaking, the Nielsen numbers are notoriously difficult to compute. We will demonstrate that the asymptotic growth rate of the sequence $\{N(f^n)\}$ (when $n \to \infty$), which we denote by $N^\infty(f)$, is a more computationally accessible invariant than the sequence itself, yet one that is still useful for dynamics. Although the exact evaluation of $N^\infty(f)$ would be desirable, its estimation is a more realistic goal and, as we shall show, one that is sufficient for many applications.

For an asymptotic study, the first challenge is to develop a unified algebraic formulation for the Nielsen theory of all iterates of f so that we can easily relate the fixed point class data of various f^n. This is why we propose the

Nielsen theory for periodic orbits. The key idea is to work on the mapping torus T_f of f and to count periodic orbits rather than periodic points, then Nielsen equivalent f-orbits on X correspond to freely homotopic closed orbit curves on T_f. (This observation of [**J2**] can actually be traced back to the pioneering work of Fuller [**Fu**] in a different context.) The fixed point data of f^n are organized into the generalized Lefschetz number $L_\Gamma(f^n)$, a homotopy invariant living in the free abelian group generated by the set of conjugacy classes in $\Gamma = \pi_1(T_f)$.

For the sake of practical computation, we assume that a matrix representation ρ of Γ is given. The traces of the ρ-images of the generalized Lefschetz numbers constitute a sequence of complex numbers. The Lefschetz zeta function ζ_f is a generating function for this sequence which turns out to be a rational function easily computable for cellular maps. Our Lefschetz zeta function is the same as that of Fried [**F4**] using matrix representations of $\pi_1(T_f)$, rather than the earlier version in [**F1**] using abelian representations, so that non-abelian information can be better retained. This makes a difference in applications, as shown in §4.3.

Now every zero or pole of ζ_f supplies a convenient lower bound for the asymptotic Lefschetz number $L^\infty(f)$ of f, defined to be the growth rate of the sequence $\{\|L_\Gamma(f^n)\|\}$ of norms of the generalized Lefschetz numbers. On the other hand, the asymptotic Lefschetz number is identified with the asymptotic Nielsen number for some important classes of maps.

The sketch above, of the approach to the estimation of $N^\infty(f)$ that we will present in this paper, is given a more detailed exposition in [**J3**].

The structure of the paper is as follows. §1 establishes the basic Nielsen theory of periodic orbits and introduces the Nielsen numbers, the Lefschetz numbers and the Lefschetz zeta function. §2 defines the asymptotic invariants, discusses the conditions for their equality and their relation to the topological entropy, and provides methods for their estimation. §3 analyses the case of homeomorphisms of compact aspherical surfaces and proposes a theory for homeomorphisms of punctured surfaces which often arise in recent 2-dimensional dynamical systems theory. The examples in §4 serve to illustrate our theory. Some open problems are given in §5.

1. Nielsen theory for periodic orbits.

We first give a brief account of the invariants of Nielsen fixed point theory in §1.1. To simplify the algebra involved, we shall work with the natural semi-flow on the mapping torus described in §1.2. The notion of periodic orbit classes is introduced in §1.3. The Lefschetz numbers and Nielsen numbers are then defined in §1.4, and their invariance shown in §1.5. When a matrix

representation of the fundamental group of the mapping torus is given, we introduce in §1.6 the associated Lefschetz zeta function. Since this will be our main tool for asymptotic estimates, an analysis of our requirement on the representation is given in §1.7. Finally, in §1.8 we introduce the relative invariants.

1.1. Nielsen theory for fixed points. The basis of Nielsen fixed point theory is the notion of a fixed point class.

Let X be a compact connected polyhedron, $f : X \to X$ be a map. The fixed point set $\operatorname{Fix} f := \{x \in X \mid x = f(x)\}$ splits into a disjoint union of *fixed point classes.* Two fixed points are in the same class if and only if they can be joined by a path which is homotopic (relative to end-points) to its own f-image. Each fixed point class $\mathbf{F}$ is an isolated subset of $\operatorname{Fix} f$ hence its index $\operatorname{ind}(\mathbf{F}, f) \in \mathbb{Z}$ is defined. A fixed point class with non-zero index is called *essential.* The number of essential fixed point classes is called the *Nielsen number* $N(f)$ of f. It is a homotopy invariant of f, so that every map homotopic to f must have at least $N(f)$ fixed points. (Cf. [**J1**, p.19].)

Pick a base point $v \in X$ and a path w from v to $f(v)$. Let $G := \pi_1(X, v)$ and let $f_G : G \to G$ be the composition

$$\pi_1(X, v) \xrightarrow{f_*} \pi_1(X, f(v)) \xrightarrow{w_*} \pi_1(X, v).$$

Two elements $g, g' \in G$ are said to be f_G-*conjugate* if there is an $h \in G$ such that $g' = f_G(h)gh^{-1}$. (There are two definitions of f_G-conjugacy in the literature, related by an inversion. The one we use here is the original one of [**R**] and [**We**] which turns out to be more convenient than the other one used in [**J1**, p. 26].) Thus G splits into f_G-conjugacy classes. Let G_f denote the set of f_G-conjugacy classes, and $\mathbb{Z}G_f$ denote the abelian group freely generated by G_f. We use the bracket notation $a \mapsto [a]$ for both projections $G \to G_f$ and $\mathbb{Z}G \to \mathbb{Z}G_f$, where $\mathbb{Z}G$ is the integral group ring of G.

For every $x \in \operatorname{Fix} f$, its *$G$-coordinate* $\operatorname{cd}_G(x, f) \in G_f$ is defined as follows: Pick a path c from v to x. The f_G-conjugacy class in G of the loop $w(f \circ c)c^{-1}$, which is evidently independent of the choice of c, is called the G-coordinate of x. (This also differs from the definition in [**J1**, p. 26] by an inversion.) Two fixed points are in the same fixed point class if and only if they have the same G-coordinates. The *G-coordinate* $\operatorname{cd}_G(\mathbf{F}, f)$ of a fixed point class $\mathbf{F}$ is then defined to be the common G-coordinate of its members.

The *generalized Lefschetz number* is defined ([**R**], [**We**], cf. [**FH**]) as

$$L_G(f) := \sum_{\mathbf{F}} \operatorname{ind}(\mathbf{F}, f) \cdot \operatorname{cd}_G(\mathbf{F}, f) \quad \in \mathbb{Z}G_f, \tag{1.1}$$

the summation being over all (essential) fixed point classes $\mathbf{F}$ of f. When

all fixed points of f are isolated, we also have

$$L_G(f) = \sum_{x \in \mathrm{Fix}\, f} \mathrm{ind}(x,f) \cdot \mathrm{cd}_G(x,f) \quad \in \mathbb{Z}G_f. \tag{1.1'}$$

The Nielsen number $N(f)$ is the number of non-zero terms in $L_G(f)$, and the indices of the essential fixed point classes appear as the coefficients in $L_G(f)$.

The invariant $L_G(f)$ used to be called the Reidemeister trace because it can be computed as an alternating sum of traces on the chain level ([**R**], [**We**]).

Let $p : \widetilde{X}, \tilde{v} \to X, v$ be the universal covering. The deck transformation group is identified with G. Let $\tilde{f} : \widetilde{X} \to \widetilde{X}$ be the lift of f such that the reference path w lifts to a path from $\tilde{v}$ to $\tilde{f}(\tilde{v})$. Then for every $g \in G$ we have $\tilde{f} \circ g = f_G(g) \circ \tilde{f}$ (cf. [**J1**, pp. 24–25]).

Assume that X is a finite cell complex and $f : X \to X$ is a cellular map. Pick a cellular decomposition $\{e_j^d\}$ of X, the base point v being a 0-cell. It lifts to a G-invariant cellular structure on the universal covering $\widetilde{X}$. Choose an arbitrary lift $\tilde{e}_j^d$ for each e_j^d. They constitute a free $\mathbb{Z}G$-basis for the cellular chain complex of $\widetilde{X}$. The lift $\tilde{f}$ of f is also a cellular map. In every dimension d, the cellular chain map $\tilde{f}$ gives rise to a $\mathbb{Z}G$-matrix $\widetilde{F}_d$ with respect to the above basis, i.e. $\widetilde{F}_d = (a_{ij})$ if $\tilde{f}(\tilde{e}_i^d) = \sum_j a_{ij}\tilde{e}_j^d$, $a_{ij} \in \mathbb{Z}G$. Then we have the Reidemeister trace formula

$$L_G(f) = \sum_d (-1)^d \left[\mathrm{tr}\, \widetilde{F}_d\right] \quad \in \mathbb{Z}G_f. \tag{1.2}$$

Remark. The base point v and the path w serve as a reference frame for the G-coordinate. (When v is a fixed point and w is the constant path, the G-coordinate of v is $[1] \in \mathbb{Z}G_f$.) A change of the reference path w would affect the homomorphism f_G, hence also the f_G-conjugacy relation in G and the set G_f where the G-coordinates live. This develops into a considerable mess when we apply the above theory to all the iterates f^n of f, as we are then forced to deal with infinitely many different sets G_{f^n} at the same time. In order to simplify the algebra, we propose the following alternative approach to the coordinates of fixed points.

1.2. The mapping torus. The *mapping torus* T_f of $f : X \to X$ is the space obtained from $X \times \mathbb{R}_+$ by identifying $(x, s+1)$ with $(f(x), s)$ for all $x \in X$, $s \in \mathbb{R}_+$, where $\mathbb{R}_+$ stands for the real interval $[0, \infty)$. On T_f there is a natural semi-flow ("sliding along the rays")

$$\varphi : T_f \times \mathbb{R}_+ \to T_f, \qquad \varphi_t(x,s) = (x, s+t) \text{ for all } t \geq 0.$$

A point $x \in X$ and a positive number $\tau > 0$ determine the *time-τ orbit curve* $\varphi_{(x,\tau)} := \{\varphi_t(x)\}_{0 \le t \le \tau}$ in T_f. We may identify X with the cross-section $X \times 0 \subset T_f$, then the map $f : X \to X$ is just the return map of the semi-flow φ.

Take the base point v of X as the base point of T_f. Let $\Gamma := \pi_1(T_f, v)$. By the van Kampen Theorem, Γ is obtained from G by adding a new generator z represented by the loop $\varphi_{(v,1)}w^{-1}$, and adding the relations $z^{-1}gz = f_G(g)$ for all $g \in G$:

$$\Gamma = \langle\, G, z \mid gz = zf_G(g) \text{ for all } g \in G \,\rangle. \tag{1.3}$$

Remark. Note that the homomorphism $G \to \Gamma$ induced by the inclusion $X \subset T_f$ is not necessarily injective. Its kernel equals $\cup_{m>0} \ker(f_G^m)$, the union of the kernels of all iterates of $f_G : G \to G$. This fact can be proved by a topological argument similar to that of [**J2**, §3].

Notation. Let Γ_c denote the set of conjugacy classes in Γ. Theoretically, it is better to regard Γ_c as the set of free homotopy classes of closed curves in T_f, so that it is independent of the base point. Let $\mathbb{Z}\Gamma$ be the integral group ring of Γ, and let $\mathbb{Z}\Gamma_c$ be the free abelian group with basis Γ_c. We use the bracket notation $\alpha \mapsto [\alpha]$ for both projections $\Gamma \to \Gamma_c$ and $\mathbb{Z}\Gamma \to \mathbb{Z}\Gamma_c$.

1.3. Periodic orbit classes. We intend to study the periodic points of f, i.e. the fixed points of the iterates of f.

We shall call $\operatorname{PP} f := \{\, (x,n) \in X \times \mathbb{N} \mid x = f^n(x) \,\}$ the *periodic point set* of f, where $\mathbb{N}$ denotes the set of natural numbers. A fixed point x of f^n is called an *n-point* of f, and its f-orbit $\{x, f(x), \dots, f^{n-1}(x)\}$ an *n-orbit* of f. The latter is called a *primary* n-orbit if it consists of n distinct points, i.e. if n is the least period of the periodic point x.

A fixed point class $\mathbf{F}^n$ of f^n will be called an *n-point class* of f. Recall from [**J1**, Proposition III.3.3] that $f(\mathbf{F}^n)$ is also an n-point class, and $\operatorname{ind}(f(\mathbf{F}^n), f^n) = \operatorname{ind}(\mathbf{F}^n, f^n)$. Thus f acts as an index-preserving permutation among its n-point classes. We define an *n-orbit class* of f to be the union of an orbit of this action. In other words, two points $x, x' \in \operatorname{Fix} f^n$ are said to be in the same n-orbit class of f if and only if some $f^i(x)$ and some $f^j(x')$ are in the same n-point class of f. The set $\operatorname{Fix} f^n$ splits into a disjoint union of n-orbit classes.

On the mapping torus T_f, observe that $(x,n) \in \operatorname{PP} f$ if and only if the time-n orbit curve $\varphi_{(x,n)}$ is a closed curve. The free homotopy class of the closed curve $\varphi_{(x,n)}$ will be called the *Γ-coordinate* of (x,n), written $\operatorname{cd}_\Gamma(x,n) = [\varphi_{(x,n)}] \in \Gamma_c$.

It follows from [**J2**, §3] that periodic points $(x,n), (x',n') \in \operatorname{PP} f$ have the same Γ-coordinate if and only if $n = n'$ and x, x' belong to the same

n-orbit class of f. Thus we can equivalently define $x, x' \in \operatorname{Fix} f^n$ to be in the same n-orbit class if and only if they have the same Γ-coordinate, and define the Γ-*coordinate* of an n-orbit class $\mathbf{O}^n$ as the common Γ-coordinate of its members, written $\operatorname{cd}_\Gamma(\mathbf{O}^n)$.

Remark. The notion of Γ-coordinate has great algebraic advantage over that of G-coordinate (cf. Remark in §1.1). Ordinary conjugacy classes have replaced the awkward skew-conjugacy classes. The Γ-coordinates $\operatorname{cd}_\Gamma(\mathbf{O}^n)$ are independent of the choice of the base point, and for any n they all live in the same set Γ_c.

An important notion in the Nielsen theory for periodic orbits is that of reducibility. Suppose m is a factor of n and $m < n$. When the n-orbit class $\mathbf{O}^n$ contains an m-orbit class $\mathbf{O}^m$ then $\operatorname{cd}_\Gamma(\mathbf{O}^n)$ is the (n/m)-th power of $\operatorname{cd}_\Gamma(\mathbf{O}^m)$, because for $x \in \mathbf{O}^m$ the closed curve $\varphi_{(x,n)}$ is the closed curve $\varphi_{(x,m)}$ traced n/m times. This motivates the definition that the n-orbit class $\mathbf{O}^n$ is *reducible to period* m if $\operatorname{cd}_\Gamma(\mathbf{O}^n)$ has an (n/m)-th root, and that $\mathbf{O}^n$ is *irreducible* if $\operatorname{cd}_\Gamma(\mathbf{O}^n)$ is *primary* in the sense that it has no nontrivial root.

This notion of reducibility is consistent with that introduced in [**J1**]. An n-orbit class $\mathbf{O}^n$ is reducible to period m if and only if every n-point class $\mathbf{F}^n \subset \mathbf{O}^n$ is reducible to period m in the sense of [**J1**, Definition III.4.2].

1.4. Lefschetz numbers and n-orbit Nielsen numbers. Every n-orbit class $\mathbf{O}^n$ is an isolated subset of $\operatorname{Fix} f^n$. Its *index* is $\operatorname{ind}(\mathbf{O}^n, f^n)$, the index of $\mathbf{O}^n$ with respect to f^n. An n-orbit class $\mathbf{O}^n$ is called *essential* if its index is non-zero.

For each natural number n, we define the (generalized) *Lefschetz number* (with respect to Γ)

$$L_\Gamma(f^n) := \sum_{\mathbf{O}^n} \operatorname{ind}(\mathbf{O}^n, f^n) \cdot \operatorname{cd}_\Gamma(\mathbf{O}^n) \quad \in \mathbb{Z}\Gamma_c, \tag{1.4}$$

the summation being over all n-orbit classes $\mathbf{O}^n$ of f. When every fixed point of f^n is isolated, we also have

$$L_\Gamma(f^n) = \sum_{(x,n)\in \mathrm{PP}\, f} \operatorname{ind}(x, f^n) \cdot [\varphi_{(x,n)}] \quad \in \mathbb{Z}\Gamma_c. \tag{1.4'}$$

Let $N_\Gamma(f^n)$ be the number of non-zero terms in $L_\Gamma(f^n)$. It is the number of essential n-orbit classes, and will be called the *Nielsen number of n-orbits*. Clearly it is a lower bound for the number of n-orbits of f.

Let $NI_\Gamma(f^n)$ be the number of non-zero primary terms in $L_\Gamma(f^n)$. It is the number of irreducible essential n-orbit classes, and will be called the *Nielsen number of irreducible n-orbits*. It is a lower bound for the number of primary n-orbits.

The indices of the essential n-orbit classes appear as the coefficients in $L_\Gamma(f^n)$. Another numerical invariant derived from $L_\Gamma(f^n)$ is its norm. We give a general definition here:

Notation. For any set S let $\mathbb{Z}S$ denote the free abelian group with the specified basis S. The *norm* in $\mathbb{Z}S$ is defined by

$$\left\|\sum_i k_i s_i\right\| := \sum_i |k_i| \in \mathbb{Z} \qquad \text{when the } s_i\text{'s in } S \text{ are all different.} \tag{1.5}$$

The norm $\|L_\Gamma(f^n)\|$ is the sum of absolute values of the indices of all the (essential) n-orbit classes. It equals $\|L_G(f^n)\|$, the sum of absolute values of the indices of all the (essential) n-point classes, because any two n-point classes contained in the same n-orbit class must have the same index. Hence $\|L_\Gamma(f^n)\| \geq N(f^n) \geq N_\Gamma(f^n)$.

Corresponding to the trace formula (1.2), we have the following trace formula:

Theorem 1.1. *Let $\widetilde{F}_d$ be the $\mathbb{Z}G$-matrices defined before* (1.2). *Then*

$$L_\Gamma(f^n) = \sum_d (-1)^d \left[\operatorname{tr}(z\widetilde{F}_d)^n\right] \quad \in \mathbb{Z}\Gamma_c, \tag{1.6}$$

where $z\widetilde{F}_d$ is regarded as a $\mathbb{Z}\Gamma$-matrix.

Proof. Applying the theory of §1.1 to the iterates f^n of f, $n \geq 1$, we get

$$L_G(f^n) := \sum_{\mathbf{F}^n} \operatorname{ind}(\mathbf{F}^n, f^n) \cdot \operatorname{cd}_G(\mathbf{F}^n, f^n) \quad \in \mathbb{Z}G_{f^n}, \tag{1.7}$$

the summation being over all fixed point classes $\mathbf{F}^n$ of f^n. (The reference path for f^n is taken to be the path $w^{(n)} := w(f \circ w) \cdots (f^{n-1} \circ w)$ from v to $f^n(v)$.)

By definition, for $(x, n) \in \operatorname{PP} f$ and for any path c in X from v to x, the Γ-coordinate of (x, n) is the conjugacy class in Γ of the loop $c\varphi_{(x,n)}c^{-1} \sim \varphi_{(v,n)}(f^n \circ c)c^{-1} \sim z^n w^{(n)}(f^n \circ c)c^{-1}$. So

$$\operatorname{cd}_\Gamma(x, n) = z^n \operatorname{cd}_G(x, f^n). \tag{1.8}$$

Now from (1.4) and (1.7) we see

$$L_\Gamma(f^n) = z^n L_G(f^n). \tag{1.9}$$

On the other hand, the trace formula (1.2) gives

$$L_G(f^n) = \sum_d (-1)^d \left[\operatorname{tr} \widetilde{F}_d^{(n)}\right] \quad \in \mathbb{Z}G_{f^n}, \tag{1.10}$$

where $\widetilde{F}_d^{(n)}$ is the matrix of $\tilde{f}^n$. Since $\tilde{f} \circ g = f_G(g) \circ \tilde{f}$ for all $g \in G$, we have

$$\widetilde{F}_d^{(n)} = f_G^{n-1}\widetilde{F}_d \cdot f_G^{n-2}\widetilde{F}_d \cdots f_G\widetilde{F}_d \cdot \widetilde{F}_d. \tag{1.11}$$

Hence by (1.9–11) we obtain

$$\begin{aligned} L_\Gamma(f^n) &= \sum_d (-1)^d \left[\operatorname{tr}(z^n \widetilde{F}_d^{(n)})\right] \\ &= \sum_d (-1)^d \left[\operatorname{tr}(z\widetilde{F}_d)^n\right] \quad \in \mathbb{Z}\Gamma_c. \end{aligned}$$

□

Remark. Occasionally in applications we may use a homomorphism from Γ to a more convenient group Γ', which determines an obvious homomorphism $\mathbb{Z}\Gamma_c \to \mathbb{Z}\Gamma'_c$. Let $L_{\Gamma'}(f^n)$ be the image of $L_\Gamma(f^n)$. Let $N_{\Gamma'}(f^n)$ be the number of non-zero terms in $L_{\Gamma'}(f^n)$, and let $NI_{\Gamma'}(f^n)$ be the number of non-zero primary terms. Then $N_{\Gamma'}(f^n)$ etc. are lower bounds for $N_\Gamma(f^n)$ etc. respectively. This technique is similar to that for fixed points developed in [**J1**, §III.2].

1.5. Invariance properties. The following basic invariance properties are similar (with similar proofs) to that for fixed points (cf. [**J1**, §§I.4–5]).

Homotopy invariance. *Suppose $f \simeq f' : X \to X$ via a homotopy $\{f_t\}_{0\le t\le 1}$. The homotopy gives rise to a homotopy equivalence $T_f, v \simeq T_{f'}, v$ in a standard way. If we identify $\Gamma' = \pi_1(T_{f'}, v)$ with $\Gamma = \pi_1(T_f, v)$ via this homotopy equivalence, then $L_\Gamma(f'^n) = L_\Gamma(f^n)$ for all n, hence also $N_\Gamma(f'^n) = N_\Gamma(f^n)$ and $NI_\Gamma(f'^n) = NI_\Gamma(f^n)$.*

Commutativity. *Suppose $f : X \to Y$ and $g : Y \to X$. Then $T_{g\circ f}$ and $T_{f\circ g}$ are homotopy equivalent in a standard way. If we identify $\Gamma = \pi_1(T_{g\circ f})$ with $\Gamma' = \pi_1(T_{f\circ g})$ in this way, then $L_\Gamma((g\circ f)^n) = L_\Gamma((f\circ g)^n)$ for all n, hence also $N_\Gamma((g\circ f)^n) = N_\Gamma((f\circ g)^n)$ and $NI_\Gamma((g\circ f)^n) = NI_\Gamma((f\circ g)^n)$.*

Homotopy type invariance. *Suppose $h : X \to X'$ is a homotopy equivalence. Suppose $f : X \to X$ and $f' : X' \to X'$ are maps such that the diagram*

$$\begin{array}{ccc} X & \xrightarrow{f} & X \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle h} \\ X' & \xrightarrow{f'} & X' \end{array}$$

commutes up to homotopy. Then $T_{f'}$ is homotopy equivalent to T_f, and when $\Gamma' = \pi_1(T_{f'})$ is suitably identified with $\Gamma = \pi_1(T_f)$, we have $L_\Gamma(f'^n) = L_\Gamma(f^n)$ for all n, hence also $N_\Gamma(f'^n) = N_\Gamma(f^n)$ and $NI_\Gamma(f'^n) = NI_\Gamma(f^n)$.

1.6. Twisted Lefschetz numbers and Lefschetz zeta function. Let R be a commutative ring with unity. Let $\mathrm{GL}_l(R)$ be the group of invertible $l \times l$ matrices in R, and $\mathcal{M}_{l\times l}(R)$ be the algebra of $l \times l$ matrices in R.

Suppose a representation $\rho : \Gamma \to \mathrm{GL}_l(R)$ is given. It extends to a representation $\rho : \mathbb{Z}\Gamma \to \mathcal{M}_{l\times l}(R)$. We define the *$\rho$-twisted Lefschetz number*

$$L_\rho(f^n) := \operatorname{tr}(L_\Gamma(f^n))^\rho = \sum_{\mathbf{O}^n} \operatorname{ind}(\mathbf{O}^n, f^n) \cdot \operatorname{tr}(\mathrm{cd}_\Gamma(\mathbf{O}^n))^\rho \quad \in R \tag{1.12}$$

for every $n \in \mathbb{N}$, the summation being over all n-orbit classes $\mathbf{O}^n$ of f. It is well defined because matrices in a conjugacy class have the same trace. When all fixed points of f^n are isolated, we have

$$L_\rho(f^n) = \sum_{(x,n)\in \mathrm{PP}\, f} \operatorname{ind}(x, f^n) \cdot \operatorname{tr}\left(\varphi_{(x,n)}\right)^\rho \quad \in R. \tag{1.12'}$$

It has the trace formula

$$\begin{aligned} L_\rho(f^n) &= \sum_d (-1)^d \operatorname{tr}\left(\operatorname{tr}(z\widetilde{F}_d)^n\right)^\rho \\ &= \sum_d (-1)^d \operatorname{tr}\left((z\widetilde{F}_d)^\rho\right)^n \quad \in R \end{aligned} \tag{1.13}$$

where for a $\mathbb{Z}\Gamma$-matrix A, its ρ-image A^ρ means the block matrix obtained from A by replacing each element a_{ij} with the $l \times l$ R-matrix a_{ij}^ρ.

We now define the (ρ-twisted) *Lefschetz zeta function* of f to be the formal power series

$$\zeta_\rho(f) := \exp \sum_n L_\rho(f^n) \frac{t^n}{n}. \tag{1.14}$$

It has constant term 1, so it is in the multiplicative subgroup $1 + tR[[t]]$ of the formal power series ring $R[[t]]$.

Clearly $\zeta_\rho(f)$ enjoys the same invariance properties as that of $L_\Gamma(f^n)$. As to its computation, we obtain from (1.13) the following determinant formula:

Theorem 1.2. *$\zeta_\rho(f)$ is a rational function in R.*

$$\zeta_\rho(f) = \prod_d \det\left(I - t(z\widetilde{F}_d)^\rho\right)^{(-1)^{d+1}} \quad \in R(t), \tag{1.15}$$

where I stands for suitable identity matrices.

Proof.

$$\zeta_\rho(f) = \exp \sum_d (-1)^d \sum_n \operatorname{tr}\left((z\widetilde{F}_d)^\rho\right)^n \frac{t^n}{n}$$

$$= \prod_d \det\left(I - t(z\widetilde{F}_d)^\rho\right)^{(-1)^{d+1}}.$$

□

By (1.12), (1.14) and the homotopy invariance, we have the

Twisted version of the Lefschetz fixed point theorem. *Let $f : X \to X$ be a map and $\rho : \pi_1(T_f) \to \mathrm{GL}_l(R)$ be a representation. If f is homotopic to a fixed point free map g, then $L_\rho(f) = 0$. If f is homotopic to a periodic point free map g, then $\zeta_\rho(f) = 1$.*

Remark 1. When $R = \mathbb{Q}$ and $\rho : \Gamma \to \mathrm{GL}_1(\mathbb{Q}) = \mathbb{Q}$ is trivial (sending everything to 1), then $L_\rho(f) \in \mathbb{Z}$ is the ordinary Lefschetz number $L(f)$, and $\zeta_\rho(f)$ is the classical Lefschetz zeta function $\zeta(f) := \exp \sum_n L(f^n)t^n/n$ introduced by Weil (cf. [**Bt**]).

Remark 2. Our Lefschetz zeta function is essentially the same as the twisted Lefschetz function of David Fried. He first introduced it in [**F1**] using f-invariant abelianizations of $\pi_1(X)$, and showed in [**F2**] that it is a certain Reidemeister torsion of the mapping torus T_f. Then in the paper [**F4**] he adopted the Reidemeister torsion approach, with respect to a flat vector bundle (which is equivalent to a matrix representation of the fundamental group).

Example. (Recipe for surfaces with boundary).

Let X be a surface with boundary, and $f : X \to X$ be a map. Suppose $\{a_1, \cdots, a_r\}$ is a free basis for $G = \pi_1(X)$. Then X has the homotopy type of a bouquet X' of r circles which can be decomposed into one 0-cell and r 1-cells corresponding to the a_i's, and f has the homotopy type of a cellular map $f' : X' \to X'$. By the homotopy type invariance of the invariants, we can replace f with f' in computations. The homomorphism $f_G : G \to G$ induced by f and f' is determined by the images $a_i' := f_G(a_i)$, $i = 1, \cdots, r$. By (1.3), the fundamental group $\Gamma = \pi_1(T_f)$ has a presentation

$$\Gamma = \langle\, a_1, \cdots, a_r, z \mid a_i z = z a_i',\ i = 1, \cdots, r \,\rangle. \tag{1.16}$$

As pointed out in [**FH**], the matrices of the lifted chain map $\tilde{f}'$ are

$$\widetilde{F}_0 = (1), \tag{1.17}$$

$$\widetilde{F}_1 = D := \left(\frac{\partial a_i'}{\partial a_j}\right),$$

where D is the Jacobian matrix in Fox calculus (see [**Bi**, §3.1] for an introduction). Then, by (1.6), in $\mathbb{Z}\Gamma_c$ we have

$$L_\Gamma(f) = [z] - \sum_{i=1}^{r} \left[z\frac{\partial a_i'}{\partial a_i}\right], \tag{1.18}$$

$$L_\Gamma(f^n) = [z^n] - [\mathrm{tr}(zD)^n]\,. \tag{1.19}$$

When a representation $\rho : \Gamma \to \mathrm{GL}_l(R)$ is given, by (1.13) and (1.15) we have

$$L_\rho(f) = \mathrm{tr}\, z^\rho - \mathrm{tr}(zD)^\rho \quad \in R, \tag{1.20}$$

$$\zeta_\rho(f) = \frac{\det(I - t(zD)^\rho)}{\det(I - tz^\rho)} \quad \in R(t). \tag{1.21}$$

1.7. A closer look at the representation ρ. A practical difficulty in the use of $L_\rho(f^n)$ and $\zeta_\rho(f)$ is to find a useful ρ which was assumed to be a group homomorphism $\Gamma \to \mathrm{GL}_l(R)$. Can we weaken the assumption on ρ?

Observe from (1.8) that the Γ-coordinate of an n-orbit class can be written as $z^n g$ for some $g \in G$, whereas a general element of Γ has the form $z^k g z^{-l}$ with $g \in G$ and $k, l \geq 0$. The definition (1.12) only requires that $\mathrm{tr}(\mathrm{cd}_\Gamma(\mathbf{O}^n))^\rho \in R$ be well defined, so ρ need to behave well only on a subset of Γ, not on the whole Γ. This motivates the following approach.

Definition. Let Γ_+ be the *monoid* defined by the presentation (1.3)

$$\Gamma_+ := \mathrm{Monoid}\,\langle G, z \mid gz = zf_G(g) \text{ for all } g \in G\,\rangle. \tag{1.22}$$

In other words, as a set,

$$\Gamma_+ = \{\, z^n g \mid n \geq 0, g \in G\,\}. \tag{1.23}$$

The letter z is regarded as a symbol so that Γ_+ is in one-one correspondence with $\mathbb{Z}_+ \times G$, where $\mathbb{Z}_+$ is the monoid of non-negative integers. And the multiplication in Γ_+ is defined by

$$(z^n a)(z^m b) := z^{n+m}\,(f_G(a)b)\,. \tag{1.24}$$

The obvious projection $\eta : \Gamma_+ \to \Gamma$, $z^n g \mapsto z^n g$ is a monoid homomorphism which will often be omitted in notations. Beware that η is not necessarily monomorphic.

Lemma 1.3. *Suppose $z^n a, z^n b \in \Gamma_+$ project to conjugate elements in Γ, where $n > 0$. Then there exist $0 \leq r < n$ and $h \in G$ such that in Γ_+ we have*

$$z^n b = h^{-1} z^{n-r} a z^r h. \tag{1.25}$$

Proof. Suppose $z^n b = \gamma^{-1} z^n a \gamma$ for some $\gamma \in \Gamma$. This γ can be written in the form $\gamma = z^k c z^{-l}$ with $c \in G$ and $k, l \geq 0$. So in Γ we have

$$z^n f_G^l(b) = z^n(z^{-l} b z^l) = z^{-l}(z^n b) z^l = z^{-l}(\gamma^{-1} z^n a \gamma) z^l$$

$$= (c^{-1}z^{-k})z^n a(z^k c) = (c^{-1}z^n)(z^{-k}az^k)c = z^n f_G^n(c^{-1}) f_G^k(a)c,$$

hence $f_G^l(b) = f_G^n(c^{-1}) f_G^k(a)c$.

By the Remark in §1.2, we can find some $m > 0$ such that

$$f_G^{l+m}(b) = f_G^{n+m}(c^{-1}) f_G^{k+m}(a) f_G^m(c) \quad \in G.$$

Increasing m if necessary, we may assume $l + m = pm$ and $k + m = qm + r$, where $p, q > 0$ and $0 \le r < n$.

Let $h = f_G^r(a^{-1}) f_G^{n+r}(a^{-1}) \cdots f_G^{(q-1)n+r}(a^{-1}) f_G^m(c) f_G^{(p-1)n}(b) \cdots f_G^n(b)b \in G$. Then we have

$$\begin{aligned}
&f_G^n(h^{-1}) f_G^r(a)h \\
&\quad = \left[f_G^n(b^{-1}) \cdots f_G^{pn}(b^{-1}) f_G^{n+m}(c^{-1}) f_G^{qn+r}(a) \cdots f_G^{n+r}(a)\right] \cdot f_G^r(a) \\
&\qquad \cdot \left[f_G^r(a^{-1}) f_G^{n+r}(a^{-1}) \cdots f_G^{(q-1)n+r}(a^{-1}) f_G^m(c) f_G^{(p-1)n}(b) \cdots f_G^n(b)b\right] \\
&\quad = f_G^n(b^{-1}) \cdots f_G^{pn}(b^{-1}) f_G^{n+m}(c^{-1}) f_G^{k+m}(a) f_G^m(c) f_G^{(p-1)n}(b) \cdots f_G^n(b)b \\
&\quad = f_G^n(b^{-1}) \cdots f_G^{pn}(b^{-1}) f_G^{l+m}(b) f_G^{(p-1)n}(b) \cdots f_G^n(b)b = b.
\end{aligned}$$

Thus, in Γ_+ we get

$$\begin{aligned}
z^n b &= z^n f_G^n(h^{-1}) f_G^r(a)h = h^{-1} z^n f_G^r(a)h \\
&= h^{-1} z^{n-r} z^r f_G^r(a)h = h^{-1} z^{n-r} a z^r h,
\end{aligned}$$

as required. □

Theorem 1.4. *Suppose a monoid representation $\rho : \Gamma_+ \to \mathcal{M}_{l\times l}(R)$ is given. In other words, suppose we have a group representation $\rho : G \to \mathrm{GL}_l(R)$ and a matrix $z^\rho \in \mathcal{M}_{l\times l}(R)$ satisfying the condition*

$$g^\rho z^\rho = z^\rho (f_G(g))^\rho \qquad \text{for any } g \in G. \tag{1.26}$$

Extend ρ to a ring homomorphism $\rho : \mathbb{Z}\Gamma_+ \to \mathcal{M}_{l\times l}(R)$. Then the theory of §1.6 works.

Proof. The basis of §1.6 is the definition (1.12) of $L_\rho(f^n)$. So it suffices to show that for any $z^n a, z^n b \in \Gamma_+$ that are conjugate in Γ, we have $\mathrm{tr}(z^n a)^\rho = \mathrm{tr}(z^n b)^\rho$.

Let r and h be as in Lemma 1.3, and write a^ρ, b^ρ, h^ρ as A, B, H respectively. Then

$$\mathrm{tr}(z^n b)^\rho = \mathrm{tr}(h^{-1} z^{n-r} a z^r h)^\rho = \mathrm{tr}(H^{-1} Z^{n-r} A Z^r H)$$

$$= \operatorname{tr}(Z^{n-r}AZ^r) = \operatorname{tr}(Z^nA) = \operatorname{tr}(z^na)^\rho.$$

□

Remark. If z^ρ is invertible, ρ will give a group representation $\Gamma \to \mathrm{GL}_l(R)$. The point of the theorem is that we do not require z^ρ to be an invertible matrix.

Example. (Free abelian group).

Suppose G is a (multiplicative) free abelian group with basis $\{g_1, \cdots, g_r\}$, and the homomorphism $f_G : G \to G$ is given by the $r \times r$ integral matrix $A = (a_{ij})$ such that $f_G(g_i) = g_1^{a_{i1}} \cdots g_r^{a_{ir}}$.

Every element $g = g_1^{v_1} \cdots g_r^{v_r} \in G$ corresponds to an integer row-vector $v(g) := (v_1, \cdots, v_r)$. Clearly $v(f_G(g)) = v(g) \cdot A$ for any $g \in G$.

Then the assignments

$$g^\rho = \begin{pmatrix} 1 & v(g) \\ 0 & I \end{pmatrix}, \qquad z^\rho = \begin{pmatrix} 1 & 0 \\ 0 & A \end{pmatrix} \tag{1.27}$$

define a monoid representation $\rho : \Gamma_+ \to \mathcal{M}_{(r+1)\times(r+1)}(\mathbb{Z})$. The verification of the condition (1.26) is trivial.

1.8. Relative invariants mod a subpolyhedron. Let X be a compact connected polyhedron as before, and A be a subpolyhedron. Let $f : X, A \to X, A$ be a self-map of the pair.

A fixed point x of f is *related to* A if there is a path c such that $c \simeq f \circ c : I, 0, 1 \to X, x, A$, where $\simeq$ means homotopic. A fixed point class $\mathbf{F}$ of f will be called a *fixed point class on* $X \setminus A$ if it is not related to A. The number of essential fixed point classes of f on $X \setminus A$ is called the *Nielsen number of the complement*, denoted $N(f; X \setminus A)$. It is a lower bound for the number of fixed points of f on $X \setminus A$, and it is invariant under homotopy of maps $X, A \to X, A$ ([**Z**], cf. [**S**, §2.3]). Obviously $N(f; X \setminus A) \leq N(f)$.

Under the mapping torus point of view, a fixed point x of f is related to A if and only if the corresponding closed orbit curve $\varphi_{(x,1)}$ in T_f is freely homotopic to a closed curve in $T_{f|A}$, the mapping torus of the restriction $f|A : A \to A$ naturally regarded as a subspace of T_f.

The Nielsen theory of periodic orbits for X developed above has a natural relative version for $X \setminus A$. A free homotopy class of closed curves in T_f (i.e. an element of Γ_c) will be called *related to* A if it contains a closed curve in $T_{f|A} \subset T_f$. An *n-orbit class of f on $X \setminus A$* is defined to be an n-orbit class of f whose coordinate is not related to A. The *Nielsen number of the complement* $N_\Gamma(f^n; X \setminus A)$ is the number of essential n-orbit classes of f on $X \setminus A$. The *Lefschetz number of the complement* $L_\Gamma(f^n; X \setminus A) \in \mathbb{Z}\Gamma_c$ is obtained from $L_\Gamma(f^n)$ by deleting the terms related to A. Clearly $\|L_\Gamma(f^n; X \setminus A)\| \leq \|L_\Gamma(f^n)\|$.

2. Asymptotic Nielsen numbers and topological entropy.

The asymptotic behavior of the number of periodic orbits is more important than that number for a specific period n. The former is also often easier to estimate. In §2.1 several asymptotic invariants are defined as growth rates of the Nielsen numbers and Lefschetz numbers. Sufficient conditions for these invariants to be equal are given in §2.2. In §2.3 we propose our method of lower estimation for the asymptotic absolute Lefschetz number via twisted Lefschetz zeta functions. §2.4 provides a method of upper estimation. §2.5 is devoted to the relation between the asymptotic Nielsen number and the topological entropy. The final section §2.6 is an aside discussing the growth rates of some Nielsen type numbers.

2.1. Asymptotic invariants. The *growth rate* of a sequence $\{a_n\}$ of complex numbers is defined by

$$\text{Growth}_{n\to\infty} a_n := \max\left\{1,\ \limsup_{n\to\infty} |a_n|^{1/n}\right\} \tag{2.1}$$

which could be infinity. Note that $\text{Growth} a_n \geq 1$ even if all $a_n = 0$. When $\text{Growth} a_n > 1$, we say that the sequence *grows exponentially.*

We define the *asymptotic Nielsen number* of f to be the growth rate of the Nielsen numbers

$$N^\infty(f) := \text{Growth}_{n\to\infty} N(f^n) = \text{Growth}_{n\to\infty} N_\mathrm{r}(f^n), \tag{2.2}$$

where the second equality is due to the obvious inequality $N_\mathrm{r}(f^n) \leq N(f^n) \leq n \cdot N_\mathrm{r}(f^n)$. And we define the *asymptotic irreducible Nielsen number* of f to be the growth rate of the irreducible Nielsen numbers

$$NI^\infty(f) := \text{Growth}_{n\to\infty} NI_\mathrm{r}(f^n). \tag{2.3}$$

We also define the *asymptotic absolute Lefschetz number*

$$L^\infty(f) := \text{Growth}_{n\to\infty} \|L_\mathrm{r}(f^n)\|. \tag{2.4}$$

All these asymptotic numbers enjoy the invariance properties of §1.5.

The following proposition ensures that these asymptotic invariants are finite positive numbers.

Proposition 2.1.

$$NI^\infty(f) \leq N^\infty(f) \leq L^\infty(f) < \infty. \tag{2.5}$$

Proof. The first two inequalities are from the obvious fact

$$NI_\mathrm{r}(f^n) \leq N_\mathrm{r}(f^n) \leq \|L_\mathrm{r}(f^n)\|.$$

The last one is by Proposition 2.6 below. □

Remark. These asymptotic invariants have obvious generalizations to the relative setting of §1.8.

2.2. Conditions for the equalities $NI^\infty(f) = N^\infty(f) = L^\infty(f)$. To compare $NI^\infty(f)$ with $N^\infty(f)$, we need the following definition.

Definition. An n-orbit class $\mathbf{O}^n$ and all n-point classes contained in it will be called *essentially irreducible* if it is essential and it does not contain any essential m-orbit class for any $m < n$.

Clearly every irreducible essential n-orbit class is essentially irreducible, but not vice versa.

Theorem 2.2. *A sufficient condition for the equality $NI^\infty(f) = N^\infty(f)$ is that f has the following Property of Essential Irreducibility:*

(EI) The number E_n of essentially irreducible n-point classes that are reducible is uniformly bounded in n.

Proof. The case $N^\infty(f) = 1$ is trivial. We assume $N^\infty(f) = 1 + a > 1$. Let E be a bound for E_n.

Let $S_n := \sum_{m \le n} N_\Gamma(f^m)$. Then by [**FLP**, p. 185, Lemma 1], $\mathrm{Growth}_{n\to\infty} S_n = N^\infty(f) = 1 + a$, hence $S_n \le (1 + \frac{5}{4}a)^n$ for sufficiently large n.

We have

$$NI_\Gamma(f^n) \ge N_\Gamma(f^n) - E_n - \sum_{\substack{m|n \\ m<n}} N_\Gamma(f^m) \ge N_\Gamma(f^n) - E - S_{n/2}$$
$$\ge N_\Gamma(f^n)\left(1 - \frac{E + S_{n/2}}{N_\Gamma(f^n)}\right).$$

Pick a subsequence $\{n_j\}$ such that $\lim_{j\to\infty} N_\Gamma(f^{n_j})^{1/n_j} = N^\infty(f)$, so that $N_\Gamma(f^{n_j}) \ge (1 + \frac{3}{4}a)^{n_j}$ for sufficiently large j. Then $S_{n_j/2}/N_\Gamma(f^{n_j}) \le (1 + \frac{1}{4}a)^{-n_j/2}$, so the quantity in the big parentheses approaches 1 when $j \to \infty$. Hence the conclusion. □

Theorem 2.3. *A sufficient condition for the equality $N^\infty(f) = L^\infty(f)$ is that f has the following Property of Bounded Index:*

(BI) The maximum absolute value B_n of the indices of n-point classes $\mathbf{F}^n$ is uniformly bounded in n.

Proof. Suppose B is a bound for B_n. Then $\|L_\Gamma(f^n)\| = \sum_{\mathbf{F}^n} \| \operatorname{ind}(\mathbf{F}^n, f^n) \| \leq BN(f^n)$. Hence $L^\infty(f) \leq N^\infty(f)$. □

Note that both properties (EI) and (BI) are invariant under homotopy. Both are satisfied in many important cases, e.g. when X is a torus of any dimension, or when f is a homeomorphism of a surface X with $\chi(X) < 0$.

2.3. Lower estimation of $L^\infty(f)$ via $\zeta_\rho(f)$.

Proposition 2.4. *Let $R = \mathbb{C}$ and let $\rho : \Gamma_+ \to \mathcal{M}_{l\times l}(\mathbb{C})$ be a monoid representation. Suppose $\{\mu_n\}$ is a sequence such that for every $n \in \mathbb{N}$,*

$$|\operatorname{tr}(z^n g)^\rho| \leq \mu_n \qquad \text{for all } g \in G, \tag{2.6}$$

and $\mu := \operatorname{Growth}_{n\to\infty}\mu_n$. Let w be a zero or a pole of the rational function $\zeta_\rho(f) \in \mathbb{C}(t)$. Then

$$L^\infty(f) \geq \frac{1}{\mu|w|}. \tag{2.7}$$

Proof. Note that $w \neq 0$ because $\zeta_\rho(f)(0) = 1$. We know from complex analysis and (1.14) that $\operatorname{Growth} L_\rho(f^n)$ is the reciprocal of the radius of convergence of the function $\log \zeta_\rho(f)$, hence $\operatorname{Growth} L_\rho(f^n) \geq 1/|w|$.

On the other hand, according to §1.6, the Γ-coordinates of n-orbit classes are in the form $[z^n g]$ with $g \in G$. So we can assume $L_\Gamma(f^n) = \sum_i k_i[z^n g_i]$, where the $[z^n g_i]$'s are different conjugacy classes in Γ. Then $L_\rho(f^n) \in \mathbb{C}$ are bounded by $|L_\rho(f^n)| = |\sum_i k_i \operatorname{tr}(z^n g_i)^\rho| \leq \sum_i |k_i| \cdot |\operatorname{tr}(z^n g_i)^\rho| \leq \mu_n \sum_i |k_i| = \mu_n \|L_\Gamma(f^n)\|$. Hence $\operatorname{Growth} L_\rho(f^n) \leq \mu \cdot L^\infty(f)$.

So we get the formula (2.7). □

Example 1. For homomorphisms of free abelian groups, the representation ρ defined by (1.27) satisfies the assumption of Proposition 2.4 where μ equals the spectral radius of the matrix A. More precisely,

$$\mu = \max\{1, |\lambda_1|, \cdots, |\lambda_r|\}, \tag{2.8}$$

where $\lambda_1, \cdots, \lambda_r$ are the eigenvalues of A.

Example 2. (Maps of the circle).

Let $f : S^1 \to S^1$ be a self-map of the circle and let $d \in \mathbb{Z}$ be its degree. The fundamental group $G = \pi_1(S^1)$ is the infinite cyclic group generated by a, and the homomorphism induced by f is $f_G : G \to G$, $a \mapsto a^d$. By (1.16), the fundamental group $\Gamma = \pi_1(T_f)$ has a presentation $\Gamma = \langle a, z \mid az = za^d \rangle$.

According to (1.27) with $r = 1$, a representation ρ of Γ_+ into $\mathcal{M}_{2\times 2}(\mathbb{Z})$ is defined by specifying

$$a^\rho = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad z^\rho = \begin{pmatrix} 1 & 0 \\ 0 & d \end{pmatrix}. \tag{2.9}$$

An easy computation shows that for any $d \in \mathbb{Z}$, we have

$$D^\rho = \left(\frac{\partial a^d}{\partial a}\right)^\rho = \begin{pmatrix} d & d(d-1)/2 \\ 0 & d \end{pmatrix}, \qquad (zD)^\rho = \begin{pmatrix} d & d(d-1)/2 \\ 0 & d^2 \end{pmatrix}. \tag{2.10}$$

Thus

$$\zeta_\rho(f) = \frac{(1-dt)(1-d^2t)}{(1-t)(1-dt)} = \frac{1-d^2t}{1-t}. \tag{2.11}$$

Hence, by (2.7) and (2.8), we get

$$L^\infty(f) \geq \max\{1, |d|\}. \tag{2.12}$$

Since the trace of a unitary matrix is bounded by its dimension, we get the very useful

Corollary 2.5. *Suppose $\rho : \Gamma \to \mathrm{U}(l)$ is a unitary representation. Let w be a zero or a pole of the rational function $\zeta_\rho(f)$. Then*

$$L^\infty(f) \geq \frac{1}{|w|}. \tag{2.13}$$

2.4. Upper estimation of $L^\infty(f)$. In practice, the initial data of our lower estimation in the last section is the knowledge of the $\mathbb{Z}G$-matrices $\{\widetilde{F}_d\}$ provided by a cellular map, which enables us to compute the Lefschetz zeta function. There is also a simple way to derive an upper bound from the same data.

We first extend the notation (1.5).

Notation. For a matrix $A = (a_{ij})$ in $\mathbb{Z}S$, its *matrix of norms* is defined to be the matrix

$$\|A\| := (\|a_{ij}\|) \tag{2.14}$$

which is a matrix of non-negative integers. (In what follows, the set S will be G or Γ.)

Proposition 2.6. *Let* $\Phi_d := \|\widetilde{F}_d\|$ *for every dimension* d. *Then*

$$L^{\infty}(f) \leq \max_d\{\text{spectral radius of } \Phi_d\}. \tag{2.15}$$

Proof. By (1.6) and the definitions,

$$\begin{aligned}\|L_\Gamma(f^n)\| &\leq \sum_d \|[\operatorname{tr}(z\widetilde{F}_d)^n]\| \leq \sum_d \|\operatorname{tr}(z\widetilde{F}_d)^n\| \leq \sum_d \operatorname{tr}\|(z\widetilde{F}_d)^n\| \\ &\leq \sum_d \operatorname{tr}\|z\widetilde{F}_d\|^n \leq \sum_d \operatorname{tr}\|\widetilde{F}_d\|^n = \sum_d \operatorname{tr}\,\Phi_d^n.\end{aligned}$$

Hence

$$\begin{aligned}L^{\infty}(f) &= \operatorname{Growth}_{n\to\infty}\|L_\Gamma(f^n)\| \\ &\leq \operatorname{Growth}_{n\to\infty}\sum_d \operatorname{tr}\,\Phi_d^n \\ &= \max_d\{\operatorname{Growth}_{n\to\infty}\operatorname{tr}\,\Phi_d^n\} \\ &= \max_d\{\text{spectral radius of } \Phi_d\}.\end{aligned}$$

□

The use of this Proposition will be illustrated by the examples in §4.

2.5. Topological entropy. The most widely used measure for the complexity of a dynamical system is the topological entropy. (See [**Wa**] for an introduction.) For the convenience of the reader, we include its definition.

Let $f : X \to X$ be a self-map of a compact metric space. For given $\epsilon > 0$ and $n \in \mathbb{N}$, a subset $E \subset X$ is said to be (n, ϵ)-*separated* under f if for each pair $x \neq y$ in E there is $0 \leq i < n$ such that $d(f^i(x), f^i(y)) > \epsilon$. Let $s_n(\epsilon, f)$ denote the largest cardinality of any (n, ϵ)-separated subset E under f. Thus $s_n(\epsilon, f)$ is the greatest number of orbit segments $\{x, f(x), \ldots, f^{n-1}(x)\}$ of length n that can be distinguished one from another provided we can only distinguish between points of X that are at least ϵ apart. Now let

$$\begin{aligned}h(f, \epsilon) &:= \limsup_{n\to\infty} \frac{1}{n} \log s_n(\epsilon, f), \\ h(f) &:= \lim_{\epsilon\to 0} h(f, \epsilon).\end{aligned} \tag{2.16}$$

The number $0 \leq h(f) \leq \infty$, which is easily seen to be independent of the metric d used, is called the *topological entropy of* f.

If $h(f, \epsilon) > 0$ then, up to resolution $\epsilon > 0$, the number $s_n(\epsilon, f)$ of distinguishable orbit segments of length n grows exponentially with n. So $h(f)$

measures the growth rate in n of the number of orbit segments of length n with arbitrarily fine resolution.

A basic relation between periodic points and topological entropy is proved by Ivanov [**I**]. We present a different proof.

Theorem 2.7. *Let $f : X \to X$ be a self-map of a compact connected polyhedron. Then*

$$h(f) \geq \log N^{\infty}(f). \tag{2.17}$$

Proof. Let $\delta > 0$ be such that every loop in X of diameter $< 2\delta$ is contractible. Let $\epsilon > 0$ be a smaller number such that $d(f(x), f(y)) < \delta$ whenever $d(x, y) < 2\epsilon$. Let $E_n \subset X$ be a set consisting of one point from each essential n-point class. Thus $|E_n| = N(f^n)$. By the definition of $h(f)$, it suffices to show that E_n is (n, ϵ)-separated.

Suppose it is not so. Then there would be two n-points $x \neq y \in E_n$ such that $d(f^i(x), f^i(y)) \leq \epsilon$ for $0 \leq i < n$ hence for all $i \geq 0$. Pick a path c_i from $f^i(x)$ to $f^i(y)$ of diameter $< 2\epsilon$ for $0 \leq i < n$ and let $c_n = c_0$. By the choice of δ and ϵ, $f \circ c_i \simeq c_{i+1}$ for all i, so $f^n \circ c_0 \simeq c_n = c_0$. This means x, y in the same n-point class, contradicting the construction of E_n. □

Theorem 2.7 is remarkable in that it does not require smoothness of the map and it provides a common lower bound for the topological entropy of all maps in a homotopy class.

Example 1. (Linear maps on tori).

Let $T^k := \mathbb{R}^k/\mathbb{Z}^k$ be the k-dimensional torus. Let f be an automorphism of T^k defined by an integer matrix A. Then

$$h(f) = \sum_{|\lambda_i|>1} \log |\lambda_i| \tag{2.18}$$

where $\lambda_1, \ldots, \lambda_k$ are eigenvalues of A. (Cf. [**Wa**, p. 203] or [**B1**, Corollary 16].) Note that $\prod_{|\lambda_i|>1} |\lambda_i|$ is the spectral radius of the endomorphism f^* induces by f on the cohomology ring $H^*(T^k, \mathbb{R})$ which is the exterior algebra of the linear space $H^1(T^k, \mathbb{R})$. So, according to [**MP**], it is a lower bound for the entropy of all continuous maps homotopic to f. Thus f has the minimal entropy in its homotopy class.

On the other hand, $N(f^n) = |L(f^n)|$ and $L(f^n) = \det(I - A^n) = \prod_i (1 - \lambda_i^n)$, so that

$$N^{\infty}(f) = \mathrm{Growth}_{n\to\infty} \prod_i |1 - \lambda_i^n| = \begin{cases} 1 & \text{if } L(f) = 0, \\ \prod_{|\lambda_i|>1} |\lambda_i| & \text{otherwise.} \end{cases} \tag{2.19}$$

Observe that f has both Properties (EI) and (BI), so $N^\infty(f) = NI^\infty(f) = L^\infty(f)$. Hence, if $L(f) \neq 0$ then $h(f) = \log L^\infty(f) = \log N^\infty(f) = \log NI^\infty(f)$. If $L(f) = 0$, then all $L(f^n) = N(f^n) = 0$ and $\log L^\infty(f) = \log N^\infty(f) = \log NI^\infty(f) = 0$, but $h(f)$ may be positive.

Example 2. (Pseudo-Anosov maps).

Let X be a compact surface with $\chi(X) < 0$. Let f be a pseudo-Anosov homeomorphism with stretching factor $\lambda > 1$. Then

$$h(f) = \log \lambda = \log N^\infty(f) \tag{2.20}$$

is the minimal entropy in the homotopy class of f ([**FLP**, p. 194] and [**I**]).

2.6. Growth rate of Nielsen type numbers. In Nielsen theory for periodic points, it is well known that $N(f^n)$ is often very poor as a lower bound for the number of fixed points of f^n. A good homotopy invariant lower bound $NF_n(f)$, called the Nielsen type number for n-th iterate, is defined in [**J1**, Definition III.4.8]. Consider any finite set of periodic orbit classes $\{\mathbf{O}^{k_j}\}$ (of varied periods k_j) such that every essential periodic m-orbit class, $m \mid n$, contains at least one class in the set. Then $NF_n(f)$ is the minimal sum $\sum_j k_j$ for all such finite sets.

The definition of $NF_n(f)$ is rather complicated, and if we count periodic orbits instead of periodic points, a good lower bound can be defined in a simpler way. The *Nielsen type number for n-orbits* $NO_n(f)$ is defined to be the total number of essentially irreducible m-orbit classes for all $m \mid n$.

When we count primary n-points, a good lower bound $NP_n(f)$, called the Nielsen type number of least period n, is defined in [**J1**] to be n times the number of irreducible essential n-orbit classes. If we count primary n-orbits, the good bound should be $NI_\Gamma(f^n)$ defined in §1.4.

The following proposition says that as far as asymptotic growth rate is concerned, these Nielsen type numbers are no better than the Nielsen numbers.

Proposition 2.8. *For any map $f : X \to X$,*

$$\operatorname{Growth}_{n\to\infty} NF_n(f) = \operatorname{Growth}_{n\to\infty} NO_n(f) = N^\infty(f), \tag{2.21}$$

$$\operatorname{Growth}_{n\to\infty} NP_n(f) = NI^\infty(f). \tag{2.22}$$

Proof. Let S_n be as in the proof of Theorem 2.2. By the definition of $NF_n(f)$ we see $N_\Gamma(f^n) \leq N(f^n) \leq NF_n(f) \leq \sum_{m|n} N(f^m) \leq \sum_{m|n} m \cdot N_\Gamma(f^m) \leq nS_n$. Similarly for $NO_n(f)$ we have $N_\Gamma(f^n) \leq NO_n(f) \leq \sum_{m|n} N_\Gamma(f^m) \leq S_n$. Hence the formula (2.21). The second formula follows from the equality $NP_n(f) = n \cdot NI_\Gamma(f^n)$. □

3. Periodic orbit classes of surface homeomorphisms.

The results of §§2.2–3 provide us with a method of asymptotic estimation for maps $f : X \to X$ that have both Properties (EI) and (BI). In §3.1 we show that self-homeomorphisms of aspherical surfaces have both these Properties and the asymptotic Nielsen number coincides with the largest stretching factor in the Thurston canonical form. §3.2 is devoted to the development of a Nielsen theory for self-homeomorphisms of punctured surfaces which is very useful in applications.

3.1. Compact aspherical surfaces. Let X be a compact connected aspherical surface and let $f : X \to X$ be a homeomorphism. The main result 3.7 of this section is easy when X is the disc, the annulus, the Möbius strip, the torus or the Klein bottle. So we shall assume $\chi(X) < 0$.

Thurston Theorem ([**T**]). *Every homeomorphism $f : X \to X$ is isotopic to a homeomorphism φ such that either*

(1) *φ is a periodic map, i.e. $\varphi^m = id$ for some m; or*

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

(3) *φ is a reducible map, i.e. there is a system of disjoint simple closed curves $\gamma = \{\gamma_1, \cdots, \gamma_k\}$ in $\operatorname{int} X$ such that γ is invariant by φ (but the γ_i's may be permuted) and γ has a φ-invariant tubular neighborhood U such that each component of $X \setminus U$ has negative Euler characteristic and on each (not necessarily connected) φ-component of $X \setminus U$, φ satisfies (1) or (2).*

The φ above is called the Thurston canonical form of f. In (3) it can be chosen so that some iterate φ^m is a generalized Dehn twist on U. Such a φ, as well as the φ in (1) or (2), will be called *standard* [**JG**, §3.1]. A key observation is that if φ is standard, so are all iterates of φ.

For the convenience of the reader, we list the following information about the fixed point classes of a standard φ (cf. [**JG**, Lemmas 3.6 and 3.4]). The superscripts '+' and '−' indicate that φ preserves or reverses the local orientation at the fixed point class.

Lemma 3.1. *Every fixed point class of a standard φ is connected. The possible types of fixed point classes are listed below, with a description of their local behavior.*

$(1)^{\pm}$ *Isolated fixed point x:*

- $(a)^+$ *$x \in \operatorname{int} M$, φ is conjugate to a rotation in a neighborhood of x;* $\operatorname{ind}(x, \varphi) = 1$.
- $(b)^+$ *$x \in \operatorname{int} M$ is a fixed point of an annular flip-twist;* $\operatorname{ind}(x, \varphi) = 1$.

(c)$^+$ $x \in \operatorname{int} M$ *is a type* $(p,k)^+$ *interior fixed point of a pseudo-Anosov piece;* $\operatorname{ind}(x,\varphi) = 1-p$ *or* 1.

(d)$^-$ $x \in \operatorname{int} M$ *is a type* $(p,k)^-$ *interior fixed point of a pseudo-Anosov piece;* $\operatorname{ind}(x,\varphi) = 1, -1$ *or* 0.

(e)$^-$ $x \in \partial M$ *and* x *is in a type* $(p,k)^-$ *invariant boundary component of some pseudo-Anosov piece;* $\operatorname{ind}(x,\varphi) = 1$ *or* 0.

(2)$^\pm$ *Fixed circle* C*:*

(a)$^+$ $C \subset \operatorname{int} M$ *is a fixed circle of an annular twist;* $\operatorname{ind}(C,\varphi) = 0$.

(b)$^-$ $C \subset \operatorname{int} M$ *and in a neighborhood of* C, φ *is conjugate to the reflection* $(z,t) \mapsto (z, 1-t)$ *on the annulus* $S^1 \times I$ *or the Möbius band* $S^1 \times I/\sim$; $\operatorname{ind}(C,\varphi) = 0$.

(c)$^+$ $C \subset \operatorname{int} M$*; on one side* C *is a type* $(p,0)^+$ *boundary component of some pseudo-Anosov piece, on the other side* C *is a boundary component of an annular twist;* $\operatorname{ind}(C,\varphi) = -p$.

(d)$^+$ $C \subset \partial M$, *and* C *is a type* $(p,0)^+$ *boundary component of some pseudo-Anosov piece;* $\operatorname{ind}(C,\varphi) = -p$.

(3)$^-$ *Fixed arc* A*, contained in some subsurface* B *of* M *on which* φ *acts as an involution. Every endpoint* x *of* A *is either*

(a) $x \in \operatorname{int} M$*; on the outside of* B, x *is in a type* $(p,k)^-$ *invariant boundary component of a pseudo-Anosov piece, or*

(b) $x \in \partial M$.

The possible values of $\operatorname{ind}(A,\varphi)$ *are* $1, -1$ *or* 0.

(4)$^+$ *Fixed subsurface* B *of* M *with* $\chi(B) \le 0$. *The possible forms for a component* C *of* ∂B*:*

(a) $C \subset \operatorname{int} M$*; on the outside of* B, C *is a type* $(p_C,0)^+$ *invariant boundary component of some pseudo-Anosov piece;*

(b) $C \subset \operatorname{int} M$*; on the outside of* B, C *is a boundary component of an annular twist;*

(c) $C \subset \partial M$.

We have $\operatorname{ind}(B,\varphi) = \chi(B) - \sum p_C < 0$, *where the summation is over the components* C *of* ∂B *of type* (a).

Moreover, a fixed point class $\mathbf{F}$ *is related to a boundary component* $C \subset \partial X$ *if and only if* $\mathbf{F}$ *intersects* C.

When we talk about the type of an n-point class $\mathbf{F}^n$ of a standard φ, we mean the type of $\mathbf{F}^n$ as a fixed point class of φ^n.

Definition. An n-point class $\mathbf{F}^n$ is *special* if it is either of type (1c)$^+$ with $p \ge 3$ and $k = 0$, or of types (2c)$^+$, (2d)$^+$ or (4)$^+$.

Corollary 3.2. *Almost all essential* n*-point classes has index* ± 1. *More*

precisely,

(i) *If an n-point class $\mathbf{F}^n$ is special then* $\operatorname{ind}(\mathbf{F}^n, \varphi^n) \leq -1$. *Otherwise* $-1 \leq \operatorname{ind}(\mathbf{F}^n, \varphi^n) \leq 1$.

(ii) *The number of special n-point classes is at most $-2\chi(X)$, and*

$$\sum_{\text{special } \mathbf{F}^n} |\operatorname{ind}(\mathbf{F}^n, \varphi^n) + 1| \leq -2\chi(X).$$

(iii) $\|L_r(\varphi^n)\| + 2\chi(X) \leq N(\varphi^n) \leq \|L_r(\varphi^n)\|$.

(iv) *Suppose $A \subset \partial X$ is a union of $a \geq 0$ boundary circles of X. Then*

$$N(\varphi^n; X \setminus A) \geq \|L_r(\varphi^n)\| + 2\chi(X) - 2a.$$

(v) *If X is orientable and φ preserves orientation, then*

$$N(\varphi^n; X \setminus A) \geq \|L_r(\varphi^n)\| + 2\chi(X).$$

Proof. (i) is clear. (ii) follows from the proof of [**JG**, Theorem 4.1]. (iii) follows from (i) and (ii). (iv) uses the last statement of Lemma 3.1 and the fact that each boundary circle can intersect at most 2 fixed point classes of φ^n, so $N(\varphi^n; X \setminus A) \geq N(\varphi^n) - 2a$. In the orientation preserving case (v), observe that if $\mathbf{F}^n$ intersects ∂X then $\mathbf{F}^n$ is special. □

Corollary 3.3. *Special n-point classes of φ are stable under iteration, and are the only ones that can contain an inessential periodic point class of lower period. More precisely:*

Let $\mathbf{F}^n$ be an n-point class of φ. Let n' be a multiple of n and $\mathbf{F}^{n'}$ be the n'-point class containing $\mathbf{F}^n$. Then

(i) $\operatorname{ind}(\mathbf{F}^n, \varphi^n) \geq \operatorname{ind}(\mathbf{F}^{n'}, \varphi^{n'})$.

(ii) *If $\mathbf{F}^n$ is special, then $\mathbf{F}^n$ and $\mathbf{F}^{n'}$ are equal as subsets of X and*

$$\operatorname{ind}(\mathbf{F}^n, \varphi^n) = \operatorname{ind}(\mathbf{F}^{n'}, \varphi^{n'}).$$

(iii) *If* $\operatorname{ind}(\mathbf{F}^n, \varphi^n) = 0 > \operatorname{ind}(\mathbf{F}^{n'}, \varphi^{n'})$ *then $\mathbf{F}^{n'}$ is special and φ^n reverses the local orientation at $\mathbf{F}^n$.*

Proof. Clear from Lemma 3.1. □

Lemma 3.4. *If an n-point class $\mathbf{F}^n$ of φ is reducible to period m, then it contains some m-point class $\mathbf{F}^m$.*

Remark. The reducibility of $\mathbf{F}^n$ to period m means it "contains" some (possibly empty) m-point class. The point of this Lemma is that it indeed contains some non-empty m-point class.

Proof. It clearly suffices to prove the case $m = 1$. We want to show $\mathbf{F}^n$ contains a fixed point of φ.

By [**J1**, Lemma III.4.6], the reducibility of $\mathbf{F}^n$ to period 1 implies there is a path c from some $x \in \mathbf{F}^n$ to $\varphi(x)$ such that the loop $c(\varphi \circ c) \cdots (\varphi^{n-1} \circ c)$ is contractible. Now $(\varphi \circ c) \cdots (\varphi^{n-1} \circ c)(\varphi^n \circ c)$ is also contractible, hence $c \simeq \varphi^n \circ c$. Applying [**JG**, Lemma 3.4] to the standard map φ^n, we find a path γ in $\mathbf{F}^n$ homotopic to c. It follows that $\varphi(\mathbf{F}^n) = \mathbf{F}^n$, and $\gamma(\varphi \circ \gamma) \cdots (\varphi^{n-1} \circ \gamma)$ is contractible in $\mathbf{F}^n$ (because it is contractible in X and $\pi_1(\mathbf{F}^n)$ injects into $\pi_1(X)$).

According to Lemma 3.1, $\mathbf{F}^n$ is either a point, or a circle, or an arc, or a subsurface B of X with $\chi(B) \leq 0$. In the first or the third case, φ certainly has a fixed point on $\mathbf{F}^n$. In the second case φ is a rotation or a reflection on the circle, no path γ of the above type can exist unless $\varphi|_{\mathbf{F}^n}$ has a fixed point.

It remains to consider the case that $\mathbf{F}^n$ is a subsurface B and $\varphi|_B : B \to B$ is a periodic map. Equip B with a hyperbolic (or Euclidean) metric such that φ is an isometry. Let γ_0 be a shortest path of the above type, i.e. from some point to its φ-image and $\beta_0 := \gamma_0(\varphi \circ \gamma_0) \cdots (\varphi^{n-1} \circ \gamma_0)$ contractible. This β_0 must be a smooth closed geodesic because otherwise γ_0 can be shortened. But a smooth closed geodesic cannot be contractible unless it degenerates to a point. Hence γ_0 is a point, a fixed point of φ. □

Corollary 3.5. *Every homeomorphism $f : X \to X$ has Properties (EI) and (BI), hence $NI^\infty(f) = N^\infty(f) = L^\infty(f)$.*

Proof. Via isotopy we may replace f with a standard φ. By Lemma 3.4 and Corollary 3.3(iii) every reducible essential n-point class contains some essential periodic point class of lower period, except possibly the special ones. By Corollary 3.2(ii) we have Property (EI) with $E = -2\chi(X)$. (When X is orientable and φ preserves orientation, we can even take $E = 0$.)

On the other hand, by Corollary 3.2(i),(ii), f has Property (BI) with $B = 1 - 2\chi(X)$.

Then apply Theorems 2.2 and 2.3. □

Lemma 3.6. *Suppose φ is standard and λ is the largest stretching factor of the pseudo-Anosov pieces ($\lambda := 1$ if there is no pseudo-Anosov piece). Then*

$$h(\varphi) = \log \lambda \qquad \text{and} \qquad N^\infty(\varphi) = \lambda.$$

Proof. Let U be the open regular neighborhood of the k reducing curves in the Thurston theorem, and $\{M_j\}$ be the components of $X \setminus U$. Let λ_j be the stretching factor of φ_j if φ_j is pseudo-Anosov and $\lambda_j = 1$ otherwise. Thus $\lambda = \max_j \lambda_j$.

The topological entropy of a periodic map is 0. By (2.20) we see $h(\varphi_j) = \log \lambda_j$ for all j. Since the topological entropy of a Dehn twist is 0, $h(\varphi|U) = 0$. So the first conclusion follows from the fact that $h(\varphi) = \max\{h(\varphi_j), h(\varphi|U)\}$ (cf. [**Wa**, Theorem 7.5]).

To prove the second conclusion, we need an inequality

$$N(\varphi_j) - 2k \le N(\varphi) \le \sum_j N(\varphi_j) + 2k. \tag{3.1}$$

Let $\mathbf{F}$ be a fixed point class of φ. Observe from Lemma 3.1 that if $\mathbf{F} \subset M_j$, then $\operatorname{ind}(\mathbf{F}, \varphi) = \operatorname{ind}(\mathbf{F}, \varphi_j)$. So if $\mathbf{F}$ is counted in $N(\varphi)$ but not counted in $\sum_j N(\varphi_j)$, it must intersect U. But we see from Lemma 3.1 that a component of U can intersect at most 2 essential fixed point classes of φ. Hence the second inequality.

Let $\mathbf{F}_j$ be a fixed point class of φ_j and let $\mathbf{F}$ be the fixed point class of φ containing $\mathbf{F}_j$. If $\mathbf{F}_j$ makes a contribution to $N(\varphi_j) - N(\varphi)$, $\mathbf{F}$ must intersect U. Via Lemma 3.1 we check that each component of U can contribute at most 2 to $N(\varphi_j) - N(\varphi)$. Hence the first inequality. Thus (3.1) is proved.

Applying (3.1) to φ^n, we have

$$N(\varphi_j^n) - 2k \le N(\varphi^n) \le \sum_j N(\varphi_j^n) + 2k.$$

Taking the growth rate in n, we get

$$N^\infty(\varphi) = \max_j N^\infty(\varphi_j). \tag{3.2}$$

But according to (2.20), $N^\infty(\varphi_j) = \lambda_j$. Hence the second conclusion. □

The above results are summarized in

Theorem 3.7. *Let X be a compact connected surface with $\chi(X) < 0$, and let $f : X \to X$ be a homeomorphism. Let $A \subset \partial X$ be a union of boundary circles such that $f(A) = A$. Then*

$$NI^\infty(f) = NI^\infty(f; X \setminus A) = N^\infty(f) = N^\infty(f; X \setminus A) = L^\infty(f) = \lambda,$$

where λ is the largest stretching factor of the pseudo-Anosov pieces in the Thurston canonical form of f ($\lambda := 1$ if there is no pseudo-Anosov piece).

3.2. Punctured surfaces. Let X be a connected compact surface and let P be a nonempty finite set of points (punctures) in the interior of X. Assume that $\chi(X) - |P| < 0$ where $|P|$ denotes the cardinality of P. Let $f : X, P \to X, P$ be a homeomorphism.

Let Y be the compactification of $X \setminus P$ by blowing up each point of P into its circle of unit tangent vectors. Then Y is a compact surface with $\chi(Y) = \chi(X) - |P| < 0$. The added circles form a set $Q \subset \partial Y$ and $Y \setminus Q = X \setminus P$.

There always exists a homeomorphism $f' : X, P \to X, P$ that is piecewise linear near P and isotopic to f rel P. We can even require that the connecting isotopy is supported in any given small neighborhood of P. See [**E**, Appendix]. We shall call such an f' a local rectification of f. (Indeed f' can be further made smooth near P if one prefers.)

Let $g : Y, Q \to Y, Q$ be the blow-up of $f' \setminus P$, i.e. the homeomorphism extending $f' : Y \setminus Q \to Y \setminus Q$ to Q according to the piecewise differential of f' at P (cf. [**B2**, §2]). Then the homotopy class of g, hence the isotopy class of g also (see [**E**]), is independent of the local rectification f'. Since $G := \pi_1(Y) = \pi_1(Y \setminus Q) = \pi_1(X \setminus P)$, we can identify the automorphism $g_G : \pi_1(Y) \to \pi_1(Y)$ with $f_G : \pi_1(X \setminus P) \to \pi_1(X \setminus P)$.

The relative Nielsen number $N(g; Y \setminus Q)$ is thus independent of the local rectification f'. We define the *punctured Nielsen number* of f to be $N(f \setminus P) := N(g; Y \setminus Q)$. It is a lower bound for the number of fixed points of $f \setminus P$ because if g is the blow up of a rectification f' of f in a sufficiently small neighborhood of P, then every fixed point of f' that is not a fixed point of f must be a fixed point of g related to Q.

There is a direct definition of the punctured Nielsen number $N(f \setminus P)$ without using rectifications. Two fixed points x, x' of f on $X \setminus P$ are said to be in the same punctured fixed point class of f if there is a path c in $X \setminus P$ such that $c \simeq f \circ c : I, 0, 1 \to X \setminus P, x, x'$. A fixed point x of f on $X \setminus P$ is said to be related to P if there is a path c in X such that $c \simeq f \circ c : I, I \setminus \{1\}, 0, 1 \to X, X \setminus P, x, P$. The punctured Nielsen number $N(f \setminus P)$ is defined to be the number of essential punctured fixed point classes of f that are unrelated to P.

The equivalence of the two definitions is not difficult to see. Let U be a regular neighborhood of P in X and let V be a smaller one such that $f(V) \subset U$. Then every fixed point of f on $V \setminus P$ is related to P. Let f' be a rectification of f on V and let g be its blow-up. This $N(g; Y \setminus Q)$ is easily identified with the second definition.

The definition using rectifications is more convenient in computations because it explicitly involves the ordinary Nielsen theory on Y.

But the direct definition sometimes gives us more insight. For example, we can see $N(f \setminus P) \geq N(f) - |P|$. In fact, consider an essential fixed point class of f that does not intersect P. It must be a disjoint union of punctured fixed point classes of f that are unrelated to P, so at least one of these latter classes must be essential, hence the inequality. Similarly, we can see that if $P \subset P'$ and $f : X, P', P \to X, P', P$, then $N(f \setminus P) + |P| \leq N(f \setminus P') + |P'|$.

Remark. $N(f \setminus P)$ is not the same as $N(f; X \setminus P)$. The former is often a much better lower bound for the number of fixed points of f on $X \setminus P$. See the examples in §4.2. Note that the coordinates of the fixed point classes of $f \setminus P$ are not in $\pi_1(T_f)$, but in $\Gamma := \pi_1(T_g) = \pi_1(T_g \setminus T_{g|Q}) = \pi_1(T_f \setminus T_{f|P})$, the fundamental group of the complement of the link $T_{f|P}$ in T_f.

We now turn to the punctured invariants for periodic orbits of f.

Definition. $N_\Gamma(f^n \setminus P) := N_\Gamma(g^n; Y \setminus Q)$, a lower bound for the number of n-orbits of f on $X \setminus P$.

$NI_\Gamma(f^n \setminus P) := NI_\Gamma(g^n; Y \setminus Q)$, a lower bound for the number of primary n-orbits of f on $X \setminus P$.

$N(f^n \setminus P) := N(g^n; Y \setminus Q)$, a lower bound for the number of n-points of f on $X \setminus P$.

$L_\Gamma(f^n \setminus P) := L_\Gamma(g^n; Y \setminus Q)$, the sum of absolute values of the indices of n-orbits of f on $X \setminus P$.

The asymptotic invariant is also defined.

Definition. $N^\infty(f \setminus P) := \mathrm{Growth}_{n\to\infty} N_\Gamma(f^n \setminus P)$.

Theorem 3.8. *$N^\infty(f \setminus P)$ is the common growth rate of various punctured Nielsen numbers:*

$$\begin{aligned} N^\infty(f \setminus P) &= \mathrm{Growth}_{n\to\infty} N(f^n \setminus P) = \mathrm{Growth}_{n\to\infty} NI_\Gamma(f^n \setminus P) \\ &= \mathrm{Growth}_{n\to\infty} \|L_\Gamma(f^n \setminus P)\| = \lambda, \end{aligned}$$

where λ is the largest stretching factor of the pseudo-Anosov pieces in the Thurston canonical form of the punctured homeomorphism $f : X \setminus P \to X \setminus P$ ($\lambda := 1$ if there is no pseudo-Anosov piece).

Proof. Easy from Theorem 3.7 and the definitions. □

To compare $N^\infty(f \setminus P)$ with the topological entropy $h(f)$, we need a lemma.

Lemma 3.9. *There exists a finite regular branched cover $\widehat{X}, \widehat{P} \to X, P$ with branching set P such that every homeomorphism $f : X, P \to X, P$ lifts to a homeomorphism $\hat{f} : \widehat{X}, \widehat{P} \to \widehat{X}, \widehat{P}$.*

Proof. We need the following fact from group theory: Suppose G is a free group of finite rank and $g_1, \cdots, g_k \in G$ are all $\neq 1$. Then there exists a normal subgroup $K \subset G$ with finite index in G, which is invariant under any automorphism of G, and such that all $g_1, \cdots, g_k \notin K$. (See [**LS**, pp. 143, 195–196].)

Let Y, Q be the blow-up of X, P as before. Then X, P is obtained from Y, Q by shrinking every component of Q to a point. Let $G = \pi_1(X \setminus P) = \pi_1(Y)$. Let $k = |P|$ and $g_1, \cdots, g_k \in G$ be elements represented by the circles in Q. Since $\chi(Y) = \chi(X) - |P| < 0$, no boundary curve of Y is contractible in Y, hence every $g_i \neq 1$. Now take a normal subgroup K guaranteed by the above algebraic fact.

Let $q : \widehat{Y} \to Y$ be the regular covering of Y such that $q_*\pi_1(\widehat{Y}) = K$, and let $\widehat{Q} = q^{-1}(Q)$. Let $p : \widehat{X}, \widehat{P} \to X, P$ be obtained from $q : \widehat{Y}, \widehat{Q} \to Y, Q$ by shrinking every component of $\widehat{Q}$ and Q to a point. Then $p : \widehat{X}, \widehat{P} \to X, P$ is a finite regular branched cover. Every point of P is a branching point because a small circle around it cannot be lifted to a circle in $\widehat{X}$. On $X \setminus P$, the covering $p : \widehat{X} \setminus \widehat{P} \to X \setminus P$ also has $p_*\pi_1(\widehat{X} \setminus \widehat{P}) = K$.

A homeomorphism $f : X, P \to X, P$ restricts to $f \setminus P : X \setminus P \to X \setminus P$ which induces an automorphism of G. By the invariance property of $K \subset G$, the homeomorphism $f \setminus P$ lifts to a homeomorphism $\hat{f} \setminus \widehat{P} : \widehat{X} \setminus \widehat{P} \to \widehat{X} \setminus \widehat{P}$. Then compactify to get the required lifting $\hat{f} : \widehat{X} \to \widehat{X}$. □

The following is the analogue of Theorem 2.1 for punctured surfaces.

Theorem 3.10. *For any homeomorphism $f : X, P \to X, P$,*

$$h(f) \geq \log N^{\infty}(f \setminus P).$$

Proof. Let $f' : X, P \to X, P$ and $g : Y, Q \to Y, Q$ be as before. Let $\psi : Y, Q \to Y, Q$ be the standard form of g, and let $\varphi : X, P \to X, P$ be the blow-down of ψ, i.e. shrinking every component of Q back again to a point.

First consider the simpler case that no component C of Q is a 1-prong boundary component of a pseudo-Anosov piece of ψ. Then φ itself is a standard map isotopic to f. So by Theorems 2.1 and 3.7,

$$\begin{aligned} h(f) &\geq \log N^{\infty}(f) = \log \lambda_{\varphi} = \log \lambda_{\psi} \\ &= \log N^{\infty}(g) = \log N^{\infty}(g; Y \setminus Q) = \log N^{\infty}(f \setminus P). \end{aligned}$$

For the general case, let $\widehat{X}, \widehat{P} \to X, P$ be the finite branched cover in Lemma 3.9, and lift f to a homeomorphism $\hat{f} : \widehat{X}, \widehat{P} \to \widehat{X}, \widehat{P}$. When blown up, $\widehat{Y}, \widehat{Q} \to Y, Q$ is an honest regular cover. Let $\hat{g}, \hat{\psi} : \widehat{Y}, \widehat{Q} \to \widehat{Y}, \widehat{Q}$ be the lifts of g, ψ. Now $\hat{\psi}$ is a standard map. If a component $\widehat{C}$ of $\widehat{Q}$ projects to a boundary component C of a pseudo-Anosov piece of ψ, then $\widehat{C}$ has more prongs than C does because every point of P is a branching point. Thus $\hat{\psi}$ belongs to the simpler case already proved, so $h(\hat{f}) \geq \log \lambda_{\hat{\psi}}$. But we know $h(\hat{f}) = h(f)$ by [**B1**, Theorem 17], and $\lambda_{\hat{\psi}} = \lambda_{\psi}$ by construction. Hence by

Theorem 3.7,

$$h(f) \geq \log \lambda_\psi = \log N^\infty(g) = \log N^\infty(f \setminus P).$$

□

4. Examples of asymptotic estimates.

To illustrate our method of estimation, we study some surface homeomorphisms arising in the recent literature of dynamical systems theory. §4.1 improves a well known result of Handel. §4.2 uses an abelian representation to estimate the growth rate of periodic orbits. §4.3 displays the power of non-abelian representations when abelianization does not work. For more applications to dynamics, see [**J3**].

4.1. Orientation reversing homeomorphisms of surfaces. The following theorem strengthens a result of Handel [**H**]:

Theorem 4.1. *If $f : X \to X$ is an orientation reversing homeomorphism of a compact oriented surface of genus g, and if f has orbits with $g+2$ distinct odd periods, then the number of primary n-orbits grows exponentially in n, hence $h(f) > 0$.*

Proof. It is shown in [**H**] that the punctured homeomorphism $f : X \setminus P \to X \setminus P$, where P is the union of the known orbits, has at least one pseudo-Anosov piece in its Thurston canonical form. According to Theorem 3.8, this guarantees $N^\infty(f \setminus P) > 1$, hence $h(f) > 0$ by Theorem 2.7. □

In [**H**] periodic points are discussed under the assumption that f is differentiable at the periodic points in question, which is now deleted, and the conclusion is that f has orbits with infinitely many distinct periods.

4.2. Orientation preserving embeddings of the disk. Let $X = D^2$ be a disk in the plane $\mathbb{R}^2$, and let $f : X \to X$ be an orientation preserving embedding. Suppose $P = \{x_1, \cdots, x_r\}$ is a finite set in the interior of D^2 such that $f(P) = P$. Then $N^\infty(f \setminus P)$ can be estimated once we know the induced automorphism $f_G : G \to G$ where $G := \pi_1(X \setminus P)$. (That f is an embedding rather than a homeomorphism is only a technical problem. By slightly enlarging the disk, we can extend the embedding to a homeomorphism of D^2 so that all the additional periodic points arising from this extension are related (on $D^2 \setminus P$) to the boundary of D^2, hence do not affect the asymptotic Nielsen number by Theorem 3.7.)

Consider the map $f : X, P \to X, P$ studied in [**GST**], where $r = 3$ and G is the free group on 3 standard generators a_1, a_2, a_3. We do not quote the

description of the map f given there in terms of a braid, but only point out that $f_G : G \to G$ is easily seen to be

$$\begin{cases} a_1' := f_G(a_1) = a_1 a_3 a_1^{-1}, \\ a_2' := f_G(a_2) = a_1, \\ a_3' := f_G(a_3) = a_3^{-1} a_2 a_3. \end{cases} \tag{4.1}$$

In [**GST**] it is shown that there exist primary periodic orbits of every period n. We shall give a lower bound to the growth rate of the number of such orbits.

The Jacobian matrix D in Fox calculus is readily calculated.

$$D = \begin{pmatrix} 1 - a_1 a_3 a_1^{-1} & 0 & a_1 \\ 1 & 0 & 0 \\ 0 & a_3^{-1} & -a_3^{-1} + a_3^{-1} a_2 \end{pmatrix}. \tag{4.2}$$

According to (1.16), we have $\Gamma = \langle a_1, a_2, a_3, z \mid a_i z = z a_i', \ i = 1, 2, 3 \rangle$. An obvious way to get a representation of Γ is to abelianize. Thus we obtain a U(1) representation ρ by letting all $a_i \mapsto a$ and $z \mapsto 1$, where a is a unimodular complex number.

Thus

$$(zD)^\rho = \begin{pmatrix} 1 - a & 0 & a \\ 1 & 0 & 0 \\ 0 & a^{-1} & 1 - a^{-1} \end{pmatrix}, \tag{4.3}$$

so that by (1.21), for the blow-up $g : Y, Q \to Y, Q$ of a local rectification of f,

$$\zeta_\rho(g) = \frac{\det(I - t(zD)^\rho)}{\det(I - tz^\rho)} = 1 - (1 - a - a^{-1})t + t^2. \tag{4.4}$$

Take $a = -1$, then we get the zeta function $\zeta_\rho(g) = 1 - 3t + t^2$ and its smallest root is $r = (3 - \sqrt{5})/2$. Hence, by Corollary 2.5, Theorem 3.8 and Theorem 3.10, we get the estimates

$$NI^\infty(f \setminus P) = N^\infty(f \setminus P) = L^\infty(f \setminus P) = L^\infty(g) \geq (3 + \sqrt{5})/2,$$

$$h(f) \geq \log \frac{3 + \sqrt{5}}{2}. \tag{4.5}$$

To obtain an upper bound by Proposition 2.6, note from (1.17) that

$$\|\widetilde{F}_0\| = (1), \qquad \|\widetilde{F}_1\| = \|D\| = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 2 \end{pmatrix}. \tag{4.6}$$

The characteristic polynomial of $\|D\|$ is $(\lambda-1)(\lambda^2-3\lambda+1)$, so (2.14) gives

$$L^\infty(f \setminus P) \leq \text{spectral radius of } \|D\| = (3+\sqrt{5})/2. \tag{4.7}$$

Now that the lower and upper estimates coincide, we have

$$NI^\infty(f \setminus P) = N^\infty(f \setminus P) = L^\infty(f \setminus P) = (3+\sqrt{5})/2. \tag{4.8}$$

Since X is the disk, all $N(f^n) = 1$ and $N^\infty(f) = 1$, so the punctured Nielsen numbers do give better estimates.

Another example on the disk is Smale's horseshoe. In [**F3**] it was shown that for the horseshoe embedding $f : D^2 \to D^2$ there is a 5-orbit $P = \{p_1, \cdots, p_5\}$. Using this orbit as punctures, the argument of [**F3**, §4] can be adapted to show that, for some representation $\rho : \Gamma \to \mathrm{U}(1)$, $\zeta_\rho(f) = 1 + t - t^2 + t^3 + t^4$. Thus by Corollary 2.5 we get $N^\infty(f \setminus P) > 1.72$ and $h(f) > \log 1.72$.

4.3. A homeomorphism of the torus. The following example is taken from [**LM**, Example 2] where it was shown that $h(f) > 0$.

Let X be the torus T^2 represented as $\mathbb{R}^2/\mathbb{Z}^2$, and let the three points $Q_1 = (0,0)$, $Q_2 = (\frac{1}{3}, \frac{1}{3})$ and $Q_3 = (\frac{2}{3}, \frac{2}{3})$ constitute the puncture set P. Let $D_1, D_2 : X, P \to X, P$ be diffeomorphisms of the form

$$D_1 : \begin{cases} x' = x \\ y' = y + B_1(x) \end{cases} \qquad D_2 : \begin{cases} x' = x + B_2(y) \\ y' = y \end{cases} \tag{4.9}$$

where x, y are in the unit interval I, and B_1, B_2 are smooth functions on I with $B_1(\frac{2}{3}) = 1$, $B_1(x) = 0$ for x outside of the open interval $(\frac{1}{2}, \frac{5}{6})$, $B_2(\frac{1}{3}) = 1$, $B_2(y) = 0$ for y outside of the open interval $(\frac{1}{6}, \frac{1}{2})$. Let $f = D_2 \circ D_1 : X, P \to X, P$.

Choose the point $(\frac{1}{2}, \frac{1}{2})$ to be the base point in X. The fundamental group $G = \pi_1(X \setminus P)$ is a free group of rank 4 with generators a_2, a_3, b_1, b_2, where b_1 is represented by the loop $\{(\frac{1}{2}+t, \frac{1}{2})\}_{t\in I}$; b_2 by the loop $\{(\frac{1}{2}, \frac{1}{2}+t)\}_{t\in I}$; a_2 is represented by the square loop with sides on the lines $y = \frac{1}{2}$, $x = \frac{1}{6}$, $y = \frac{1}{6}$ and $x = \frac{1}{2}$; a_3 represented by the square loop with sides on the lines $y = \frac{1}{2}$, $x = \frac{5}{6}$, $y = \frac{5}{6}$ and $x = \frac{1}{2}$.

It is clear that D_1, D_2 act on G by

$$D_1 : \begin{cases} a_2 \mapsto a_2 \\ a_3 \mapsto b_2 a_3 b_2^{-1} \\ b_1 \mapsto a_3^{-1} b_1 \\ b_2 \mapsto b_2, \end{cases} \qquad D_2 : \begin{cases} a_2 \mapsto b_1 a_2 b_1^{-1} \\ a_3 \mapsto a_3 \\ b_1 \mapsto b_1 \\ b_2 \mapsto b_2 b_1 a_2 b_1^{-1}. \end{cases} \tag{4.10}$$

So the automorphism $f_G : G \to G$ induced by f is

$$f_G : \begin{cases} a_2 \mapsto a_2' = b_1 a_2 b_1^{-1} \\ a_3 \mapsto a_3' = (b_2 b_1 a_2 b_1^{-1})\, a_3\, (b_2 b_1 a_2 b_1^{-1})^{-1} \\ b_1 \mapsto b_1' = a_3^{-1} b_1 \\ b_2 \mapsto b_2' = b_2 b_1 a_2 b_1^{-1}. \end{cases} \tag{4.11}$$

It is routine to compute the Fox calculus Jacobian matrix D.

$$D = \begin{pmatrix} b_1 & 0 & 1 - a_2' & 0 \\ (1 - a_3')\, b_2 b_1 & b_2' & (1 - a_3')\, b_2\, (1 - a_2') & 1 - a_3' \\ 0 & -a_3^{-1} & a_3^{-1} & 0 \\ b_2 b_1 & 0 & b_2 - b_2' & 1 \end{pmatrix}. \tag{4.12}$$

Now

$$\Gamma = \langle a_2, a_3, b_1, b_2, z \mid a_2 z = z a_2', a_3 z = z a_3', b_1 z = z b_1', b_2 z = z b_2' \rangle. \tag{4.13}$$

The abelianization does not work here, because the nature of (4.11) would force both a_2, a_3 to become 1, so that any U(1) representation can only give the trivial estimate $N^{\infty}(f \setminus P) \geq 1$. Thus we have to look for non-abelian representations of Γ. Fortunately a simple representation of Γ into the multiplicative group of unimodular quaternions works.

(4.14)
$$\rho : \Gamma \to \mathrm{Sp}(1); \quad z \mapsto i, \quad a_2 \mapsto -1, \quad a_3 \mapsto -1, \quad b_1 \mapsto j, \quad b_2 \mapsto k.$$

Under this representation, the matrix zD becomes

$$(zD)^{\rho} = \begin{pmatrix} k & 0 & 2i & 0 \\ 2 & j & -4j & 2i \\ 0 & i & -i & 0 \\ 1 & 0 & -2j & i \end{pmatrix}. \tag{4.15}$$

But quaternions form a skew-field, not a field. We have to identify Sp(1) with SU(2) via the correspondence

$$1 \mapsto \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad i \mapsto \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad j \mapsto \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad k \mapsto \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}. \tag{4.16}$$

Thus $(zD)^{\rho}$ becomes an 8×8 complex matrix. One then calculates that

$$\det(1 - t(zD)^{\rho}) = (1 + t)^2 (1 + t^2)[(1 - t + t^2)^2 + 3t^2], \tag{4.17}$$

$$\zeta_\rho = (1+t)^2[(1-t+t^2)^2 + 3t^2], \tag{4.18}$$

from which follow the estimates

$$N^\infty(f \setminus P) > 2.29, \qquad h(f) > \log 2.29 \tag{4.19}$$

by Corollary 2.5.

For the upper bound, we have

$$\|\widetilde{F}_0\| = (1), \qquad \|\widetilde{F}_1\| = \|D\| = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 2 & 1 & 4 & 2 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 2 & 1 \end{pmatrix}. \tag{4.20}$$

The characteristic polynomial of $\|D\|$ is $(\lambda^2+1)(\lambda^2-4\lambda+1)$. Then (2.14) gives

$$L^\infty(f \setminus P) \leq \text{spectral radius of } \|D\| = 2+\sqrt{3} < 3.74. \tag{4.21}$$

So the estimate we get is

$$2.29 < N^\infty(f \setminus P) < 3.74. \tag{4.22}$$

5. Questions.

It is clear from our discussion that among the asymptotic invariants, $NI^\infty(f)$ is most interesting geometrically, and $L^\infty(f)$ is most manageable algebraically. So, the conditions for the equalities $NI^\infty(f) = N^\infty(f) = L^\infty(f)$ and the extent to which they can fail are worth further study. (In fact, the author knows of no counter-example to the equality $NI^\infty(f) = N^\infty(f)$.) For example,

Question 5.1. Is it true that a self-homeomorphism (or self-map) f of an aspherical compact polyhedron X always has Properties (EI) and (BI), or at least $NI^\infty(f) = N^\infty(f) = L^\infty(f)$?

A question related to the Entropy Conjecture of Shub is the following:

Question 5.2. For smooth maps of compact manifolds, is it always true that

$$h(f) \geq \log L^\infty(f)?$$

Another natural question is

Question 5.3. Give conditions for $\log N^\infty(f)$ to be the best lower bound for $h(f)$ of all maps homotopic to f. In other words, in the inequality

$$\inf\{h(g) \mid g \simeq f : X \to X\} \geq \log N^\infty(f)$$

when does the equality hold? (The example of torus maps in §2.5 shows that the equality may fail even for very nice maps.)

References

[Bi] J.S. Birman, *Braids, Links, and Mapping Class Groups*, Ann. Math. Studies, vol. 82, Princeton Univ. Press, Princeton, 1974.

[Bt] R. Bott, *On the shape of a curve*, Advances in Math., **16** (1975), 23–38.

[B1] R. Bowen, *Entropy for group endomorphisms and homogeneous spaces*, Tran. Amer. Math. Soc., **153** (1971), 401–413.

[B2] R. Bowen, *Entropy and the fundamental group*, The Structure of Attractors in Dynamical Systems, (N.G. Markley et al., eds.), Lecture Notes in Math., vol. 668, Springer-Verlag, Berlin, Heidelberg, New York, 1978, 21–29.

[E] D.B.A. Epstein, *Curves on 2-manifolds and isotopies*, Acta Math., **115** (1966), 83–107.

[FH] E. Fadell and S. Husseini, *The Nielsen number on surfaces*, Topological Methods in Nonlinear Functional Analysis, (S.P. Singh et al., eds.), Contemp. Math., vol. 21, Amer. Math. Soc., Providence, 1983, 59–98.

[FLP] A. Fathi, F. Laudenbach and V. Poénaru, *Travaux de Thurston sur les surfaces, Séminaire Orsay*, Astérisque, vol. 66–67, Soc. Math. France, Paris, 1979.

[F1] D. Fried, *Periodic points and twisted coefficients*, Geometric Dynamics, (J. Palis Jr. ed.), Lecture Notes in Math., vol. 1007, Springer-Verlag, Berlin, Heidelberg, New York, 1983, 261–293.

[F2] ———, *Homological identities for closed orbits*, Invent. Math., **71** (1983), 419–442.

[F3] ———, *Entropy and twisted cohomology*, Topology, **25** (1986), 455–470.

[F4] ———, *Lefschetz formulas for flows*, The Lefschetz Centennial Conference, Part III: Proceedings on Differential Equations, (A. Verjovsky ed.), Contemp. Math., vol. 58. III, Amer. Math. Soc., Providence, 1987, 19–69.

[Fu] F.B. Fuller, *The treatment of periodic orbits by the methods of fixed point theory*, Bull. Amer. Math. Soc., **72** (1966), 838–840.

[GST] J.-M. Gambaudo, S. van Strien and C. Tresser, *Vers un ordre de Sarkovskii pour les plongements du disque préservant l'orientation*, C. R. Acad. Sci. Paris, Série I, **310** (1990), 291–294.

[H] M. Handel, *The entropy of orientation-reversing homeomorphisms of surfaces*, Topology, **21** (1982), 291–296.

[I] N.V. Ivanov, *Entropy and the Nielsen numbers*, Soviet Math. Dokl., **26** (1982), 63–66.

[J1] B. Jiang, *Lectures on Nielsen Fixed Point Theory*, Contemp. Math., vol. 14, Amer. Math. Soc., Providence, 1983.

[J2] ———, *A characterization of fixed point classes*, Fixed Point Theory and its Applications, (R.F. Brown ed.), Contemp. Math., vol. 72, Amer. Math. Soc., Providence, 1988, 157–160.

[J3] ———, *Nielsen theory for periodic orbits and applications to dynamical systems*, Nielsen Theory and Dynamical Systems, (Ch. McCord ed.), Contemp. Math., vol. 152, Amer. Math. Soc., Providence, 1993.

[JG] B. Jiang and J. Guo, *Fixed points of surface diffeomorphisms*, Pacific J. Math., **160** (1993), 67–89.

[LM] J. Llibre and R.S. MacKay, *Rotation vectors and entropy for homeomorphisms of the torus isotopic to the identity*, Ergod. Th. Dyn. Sys., **11** (1991), 115–128.

[LS] R.C. Lyndon and P.E. Schupp, *Combinatorial Group Theory*, Springer, New York, 1977.

[MP] M. Misiurewicz and F. Przytycki, F. *Entropy conjecture of tori*, Bull. Ac. Pol. Sci., Ser. Math., **25** (1977), 575–578.

[R] K. Reidemeister, *Automorphismen von Homotopiekettenringen*, Math. Ann., **112** (1936), 586–593.

[T] W.P. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc., **19** (1988), 417–431.

[Wa] P. Walters, *An Introduction to Ergodic Theory*, Springer, New York, 1982.

[We] F. Wecken, *Fixpunktklassen*, II, Math. Ann., **118** (1942), 216–234.

[Z] X.-Z. Zhao, *A relative Nielsen number for the complement*, Topological Fixed Point Theory and Applications, (B. Jiang ed.), Lecture Notes in Math., vol. 1411, Springer-Verlag, Berlin, Heidelberg, New York, 1989, 189–199.

Received May 28, 1993 and revised October 22, 1993. Partially supported by NSFC. The author is also grateful to the hospitality of UCLA and Heidelberg University while this paper was being written up, and the partial support by NSF Grant DMS88-57452 and Deutsche Forschungs Gemeinschaft.

PEKING UNIVERSITY
BEIJING 100871, CHINA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

FACTORIZATION OF P-COMPLETELY BOUNDED MULTILINEAR MAPS

CHRISTIAN LE MERDY

Given Banach spaces $X_1, \dots, X_N$, $Y_1, \dots, Y_N, X, Y$ and subspaces $S_i \subset B(X_i, Y_i)$ $(1 \leq i \leq N)$, we study p-completely bounded multilinear maps $A : S_N \times \cdots \times S_1 \to B(X, Y)$. We obtain a factorization theorem for such A which is entirely similar to the Christensen-Sinclair representation theorem for completely bounded multilinear maps on operator spaces. Our main tool is a generalisation of Ruan's representation theorem for operator spaces in the Banach space setting. As a consequence, we are able to compute the norms of adapted multilinear Schur product maps on $B(\ell_p^n)$.

1. Introduction and preliminaries.

1.1. Introduction. In a recent paper, Pisier [**Pi1**] proved that the Wittstock factorization theorem for completely bounded maps (cf. [**Ha**], [**Pa1**], [**Pa2**], [**W**]) has a natural generalization to the more general framework of p-completely bounded maps defined on sets of Banach space operators. The main goal of this paper is to prove a version of the Christensen-Sinclair theorem (cf. [**CS**, **PS**]) in this extensive setting.

Let us first recall the definition of p-complete boundedness as introduced (or suggested) in [**Pi**]. Let $1 \leq p < +\infty$ be a number. Let X, Y be Banach spaces. We denote by $B(X, Y)$ the space of all bounded operators from X into Y. Let $S \subset B(X, Y)$ be a subspace. We denote by $M_{n,m}(S)$ the vector space of all $n \times m$ matrices with entries from S. Any $s = [s_{ij}] \in M_{n,m}(S)$ may be canonically identified with a bounded operator from $\ell_p^m(X)$ into $\ell_p^n(Y)$. Under this identification, s has the following norm:

$$(1.1) \quad \|[s_{ij}]\|_{M_{n,m}(S)} = \sup \left\{ \left(\sum_i \left\| \sum_j s_{ij}(x_j) \right\|^p \right)^{\frac{1}{p}} / x_1, \dots, x_m \in X, \sum_j \|x_j\|^p \leq 1 \right\}.$$

Then the usual concept of complete boundedness has the following natural extension.

Definition 1.1. Let $X_1, \dots, X_N$, $Y_1, \dots, Y_N, X, Y$ be Banach spaces. For each $1 \leq i \leq N$, let $S_i \subset B(X_i, Y_i)$ be a subspace. Let $A : S_N \times \cdots \times S_1 \to B(X, Y)$ be a N-linear map. We will say that A is p-completely bounded if there is a constant $C > 0$ for which the following holds.

For any

$$s_N \in M_{n,k_{N-1}}(S_N), s_{N-1} \in M_{k_{N-1},k_{N-2}}(S_{N-1}), \dots ,$$
$$s_2 \in M_{k_2,k_1}(S_2), s_1 \in M_{k_1,m}(S_1),$$

we have:

$$\Big\| \Big[\sum_{\substack{1 \leq \ell \leq N-1 \\ 1 \leq r_\ell \leq k_\ell}} A(s_{N(i,r_{N-1})}, s_{(N-1)(r_{N-1},r_{N-2})}, \dots , s_{2(r_2,r_1)}, s_{1(r_1,j)}) \Big]_{(i,j)} \Big\|_{M_{n,m}(B(X,Y))}$$
$$\leq C \|s_N\|_{M_{n,k_{N-1}}(S_N)} \|s_{N-1}\|_{M_{k_{N-1},k_{N-2}}(S_{N-1})} \cdots \cdots \|s_1\|_{M_{k_1,m}(S_1)}.$$

Moreover, we denote by $\|A\|_{pcb}$ the least constant $C > 0$ for which this holds.

We will prove that whenever $p \in]1, +\infty[$, a p-completely bounded multilinear map A as above factors as a product of p-completely bounded linear maps defined on each S_i (see Theorem 5.1 for a precise statement). Thus using Pisier's generalization of the Wittstock theorem, we obtain a representation of A which is quite similar to the Christensen-Sinclair representation for a completely bounded multilinear map on operator spaces. This answers the question raised by Pisier in the Final Remark of [**Pi1**]. Note that our result is new only for $N \geq 3$. However, even in the case $N = 2$, we feel that our proof is simpler than Pisier's one.

The recently developped theory of operator spaces (see [**B**, **BP**, **BS**, **ER1**, **ER2**]) has emphasized the role of the Haagerup tensor product in the study of completely bounded multilinear maps. It is now well-known to specialists (see [**B**, Theorem 2.4] for example) that the Christensen-Sinclair theorem may be viewed as a combination of the factorization theorem for completely bounded bilinear forms (which goes back to [**EK**]), Ruan's representation theorem for operator spaces (see [**R**, **ER3**]) and simple properties of the Haagerup tensor product. In this approach, the crucial point is that given two operator spaces, their Haagerup tensor product is again an operator space. This essentialy follows from Ruan's theorem. In order to prove our main Theorem 5.1, we will follow the above scheme. We will especially

prove a generalization of Ruan's theorem (see our Theorem 4.1) which is of independant interest.

Let us now explain the organization of the paper. In the two following subsections, we recall Pisier's result about p-completely bounded linear maps and introduce necessary definitions about matrix normed spaces. In Section 2, we define a generalized Haagerup tensor product $\otimes_h$ adapted to our definition of p-complete boundedness and prove elementary properties which will be needed later. In Section 3, we combine ideas from [**E**], [**ER3**] and [**Pi1**] in order to prove an abstract factorization theorem which is used in the two following sections. Section 4 is devoted to our generalization of Ruan's theorem. We follow the same line of attack as Effros and Ruan [**ER3**]. Our main result explained above is proved in Section 5. In the last Section 6, we investigate some of the properties of our new tensor product $\otimes_h$. We then prove a theorem about multilinear Schur products on $B(\ell_p^n, \ell_p^n)$ which generalizes previous works on this subject (see [**ER4**, **Gr**, **Ha**, **S**] for example).

1.2. Pisier's theorem. We wish to recall Pisier's theorem as stated in [**Pi1**]. It will be formulated in the language of ultraproducts. We first introduce a notation which will be frequently used in this paper.

Definition 1.2. Let E and X be Banach spaces. Let $1 \leq p < +\infty$ be a number. We will write $E \in SQ_p(X)$ provided that E is (isometric to) a quotient of a subspace of an ultraproduct of spaces of the form $L_p(\mu; X)$.

Let X_1, Y_1 be Banach spaces and $S \subset B(X_1, Y_1)$. Consider a number $1 \leq p < +\infty$. Let $(\Omega_j, \mu_j)_{j\in J}$ be a family of measure spaces and let $\mathcal{U}$ be an ultrafilter on the index set J. Let us denote by $\widehat{X_1}$ and $\widehat{Y_1}$ respectively the ultraproducts relative to $\mathcal{U}$ of the families $(L_p(\mu_j\,;\, X_1))_{j\in J}$ and $(L_p(\mu_j\,;\, Y_1))_{j\in J}$. For any $a \in B(X_1, Y_1)$, we may define $\widehat{\pi_j}(a) : L_p(\mu_j\,;\, X_1) \to L_p(\mu_j\,;\, Y_1)$ by setting $(\pi_j(a)f)(w) = a.f(w)$. We denote by $\widehat{\pi}(a) : \widehat{X_1} \to \widehat{Y_1}$ the map associated to the family $(\widehat{\pi}_j(a))_{j\in J}$. Let $N \subset M \subset \widehat{X_1}$ and $N' \subset M' \subset \widehat{Y_1}$ be closed subspaces such that for any $s \in S$, $\widehat{\pi}(s)(M) \subset M'$ and $\widehat{\pi}(s)(N) \subset N'$. Then letting $G = \dfrac{M}{N}$ and $G' = \dfrac{M'}{N'}$, we obtain that $\widehat{\pi}_{/S}$ canonically induces a map $\pi : S \to B(G, G')$. Namely we may set $\pi(s)(m+N) = \widehat{\pi}(s)(m) + N'$ for any $(s, m) \in S \times M$. Such a map will be called a p-representation from S into $B(G, G')$. More precisely we state the following:

Definition 1.3. Let $G \in SQ_p(X_1)$ and $G' \in SQ_p(Y_1)$ be two Banach spaces. Let $\pi : S \to B(G, G')$ be a bounded linear map. We will say that π is a p-representation provided that it may be constructed as above.

Theorem 1.4 ([**Pi1**, Theorem 2.1]). *Let $S \subset B(X_1, Y_1)$, let $A : S \to B(X, Y)$ be a linear map and let C be a constant. The following assertions*

are equivalent:

(i) *A is p-completely bounded and $\|A\|_{pcb} \leq C$.*

(ii) *There are two Banach spaces $G \in SQ_p(X_1), G' \in SQ_p(Y_1)$ and a p-representation $\pi : S \to B(G, G')$ as well as operators $V : X \to G$ and $W : G' \to Y$ with $\|V\|\,\|W\| \leq C$ such that:*

$$\forall\, s \in S, \quad A(s) = W\pi(s)V.$$

1.3. Matrix normed spaces. Let S be a complex normed space. Let us denote by $M_{n,m}(S)$ the vector space of $n \times m$ matrices over S. As usual, we just denote by $M_n(S)$ the space $M_{n,n}(S)$. In the case when $S = \mathbb{C}$, we will simply write $M_{n,m}$ or M_n for these spaces. We will say that S is a matrix normed space provided that we are given norms $\| \ \|_{n,m}$ on each $M_{n,m}(S)$ satisfying $M_1(S) = S$ and:

(i) For any $s \in M_{n,m}(S), s' \in M_{n,k}(S)$,

$$\max\left\{\|s\|_{n,m}, \|s'\|_{n,k}\right\} \leq \|(s, s')\|_{n,m+k}.$$

(ii) For any $s \in M_{n,m}(S), 0 \in M_{n,k}(S)$,

$$\|s\|_{n,m} = \|(s, 0)\|_{n,m+k} = \|(0, s)\|_{n,m+k}.$$

(iii) For any $s \in M_{n,m}(S), s' \in M_{k,m}(S)$,

$$\max\left\{\|s\|_{n,m}, \|s'\|_{k,m}\right\} \leq \left\|\begin{pmatrix} s \\ s' \end{pmatrix}\right\|_{n+k,m}.$$

(iv) For any $s \in M_{n,m}(S), 0 \in M_{k,m}(S)$,

$$\|s\|_{n,m} = \left\|\begin{pmatrix} s \\ 0 \end{pmatrix}\right\|_{n+k,m} = \left\|\begin{pmatrix} 0 \\ s \end{pmatrix}\right\|_{n+k,m}.$$

Actually, these are very weak conditions. They are chosen to ensure two reasonnable properties. First, for any $n, m \leq k$, the canonical embedding of $M_{n,m}(S)$ in $M_k(S)$ is isometric. Secondly, for any $s = [s_{ij}] \in M_{n,m}(S)$,

$$\sup_{i,j} \|s_{ij}\| \leq \|s\|_{n,m} \leq \sum_{i,j} \|s_{ij}\|. \tag{1.2}$$

Thus $M_{n,m}(S)$ and S^{nm} are isomorphic as normed spaces. From now on, we leave the notation $\| \ \|_{n,m}$ and merely denote by $\| \ \|$ the norm on all the spaces $M_{n,m}(S)$. We will have to distinguish a possible property of a matrix

normed space S. For any $s \in M_{n,m}(S)$, $s' \in M_{n',m'}(S)$, we set $s \oplus s' = \begin{pmatrix} s & 0 \\ 0 & s' \end{pmatrix} \in M_{n+n',m+m'}(S)$. We will say that S satisfies $\mathcal{D}_\infty$ whenever the following condition is fulfilled:

$\mathcal{D}_\infty$: For any $s \in M_{n,m}(S), s' \in M_{n',m'}(S)$, $\|s \oplus s'\| = \max\{\|s\|, \|s'\|\}$.

The latter property is one of the characteristic conditions in Ruan's representation theorem for operator spaces. It will play a similar role in our Theorem 4.1.

Let us now introduce some standard definitions and traditional notation. Let S, T be two matrix normed spaces and let $u \in B(S,T)$. We define $u^{(n)} : M_n(S) \to M_n(T)$ by $u^{(n)}([s_{ij}]) = [u(s_{ij})]$. We let $\|u\|_{cb} = \sup_{n \geq 1} \left\|u^{(n)}\right\|$. We say that u is completely bounded (in short c.b.) provided that $\|u\|_{cb} < +\infty$. We denote by $CB(S,T)$ the resulting normed space. We say that u is completely contractive (in short c.c.) when $\|u\|_{cb} \leq 1$ and u is completely isometric provided that for any $n \geq 1$, $u^{(n)}$ is isometric.

2. p-matrix normed spaces.

We introduce a special kind of matrix normed spaces.

Definition 2.1. Let $p, q \in]1, +\infty[$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Let S be a matrix normed space. We will say that S is a p-matrix normed space if it satisfies the condition $\mathcal{D}_\infty$ above and the following:

(2.1) For any $s \in M_{n,m}(S), s' \in M_{n',m}(S)$, $\left\|\begin{pmatrix} s \\ s' \end{pmatrix}\right\|^p \leq \|s\|^p + \|s'\|^p$.

(2.2) For any $s \in M_{n,m}(S)$, $s' \in M_{n,m'}(S)$, $\|(s, s')\|^q \leq \|s\|^q + \|s'\|^q$.

(2.3) For any $s \in M_{n,m}(S), \alpha \in M_{m,1}$, $\|s\alpha\| \leq \|s\| \left(\sum_{j=1}^m |\alpha_j|^p\right)^{1/p}$.

(2.4) For any $s \in M_{n,m}(S), \beta \in M_{1,n}$, $\|\beta s\| \leq \|s\| \left(\sum_{i=1}^n |\beta_i|^q\right)^{1/q}$.

Example 2.2. Let X, Y be Banach spaces. Let $S \subset B(X,Y)$ be a subspace. Let us equip each $M_{n,m}(S)$ with the norm defined by (1.1). Recall that this yields an isometric embedding $M_{n,m}(S) \subset B(\ell_p^m(X), \ell_p^n(Y))$. Then it is not hard to check that S becomes a p-matrix normed space. Let us emphasize for further that given a finite family $(s_j)_{1 \leq j \leq n}$ in S, the corresponding column and row matrices have the following norms:

$$(2.5) \qquad \left\|\begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix}\right\| = \sup\left\{\left(\sum_{j=1}^n \|s_j(x)\|^p\right)^{1/p} \;/\; x \in X,\ \|x\| \leq 1\right\}$$

$$(2.6) \qquad \|(s_1, \cdots, s_n)\| = \sup \left\{ \left(\sum_{i=1}^{n} \|s_i^*(y^*)\|^q \right)^{1/q} \ / \ y^* \in Y^*, \|y^*\| \leq 1 \right\}.$$

Throughout the rest of the paper, we fix a number $p \in]1, +\infty[$ and let $q = \frac{p}{p-1}$ (i.e.: $\frac{1}{p} + \frac{1}{q} = 1$). Given a subspace S of some $B(X,Y)$, we will always assume that it is endowed with its p-matrix normed space structure as defined in Example 2.2. As announced in the introduction, our purpose is to define an adapted variant of the Haagerup tensor product. Although we will be mainly concerned by matrix normed spaces $S \subset B(X,Y)$ as above, it is convenient to work in the slighly more general setting of p-matrix normed spaces. We will only give short proofs of the results listed below since they are all variants of known results of the classical theory of operator spaces. We will use the following well-know fact :

$$(2.7) \qquad \forall\ (a,b) \in \mathbb{R}_+^2,\ ab = \inf \left\{ \frac{\theta^p a^p}{p} + \frac{\theta^{-q} b^q}{q} \ / \ \theta > 0 \right\}.$$

Let S, T be two p-matrix normed spaces. Given $s = [s_{ir}] \in M_{n,k}(S)$ and $t = [t_{rj}] \in M_{k,m}(T)$, we define $s \odot t = \left[\sum_{r=1}^{k} s_{ir} \otimes t_{rj} \right] \in M_{n,m}(S \otimes T)$. For any $z \in M_{n,m}(S \otimes T)$ we set:

$$(2.8) \qquad \|z\|_h = \inf \ \{ \|s\| \ \|t\| \ / \ s \in M_{n,k}(S), t \in M_{k,m}(T), z = s \odot t \}.$$

Proposition-Definition 2.3. The function $\| \ \|_h$ is a norm on each space $M_{n,m}(S \otimes T)$. Endowed with these norms, $S \otimes T$ becomes a p-matrix normed space.

We will denote by $S \otimes_h T$ this p-matrix normed space.

Proof. Let $z = s \odot t$ and $z' = s' \odot t' \in M_{n,m}(S \otimes T)$. Then $z + z' = (s, s') \odot \begin{pmatrix} t \\ t' \end{pmatrix}$. Therefore, applying (2.2) to (s, s'), (2.1) to $\begin{pmatrix} t \\ t' \end{pmatrix}$ and (2.8), we deduce that $\| \ \|_h$ is a semi-norm on $M_{n,m}(S \otimes T)$. It is clear that these semi-norms satisfy all the conditions (i), (ii), (iii), (iv) required in the definition of a matrix normed space. Hence by (1.2), in order to prove that $\| \ \|_h$ is a norm on each $M_{n,m}(S \otimes T)$, it is enough to show that $\| \ \|_h$ is non-degenerate on $S \otimes T$. Let $z = \sum_{r=1}^{N} s_r \otimes t_r \in S \otimes T$. Let $s^* \in S^*, t^* \in T^*$. Since the space S satisfies (2.3), we have: $\left(\sum_{r=1}^{N} |\langle s^*, s_r \rangle|^q \right)^{1/q} \leq \|s^*\| \ \|(s_1, \ldots, s_N)\|$.

Analogoulsy,

$$\left(\sum_{r=1}^{N} |\langle t^*, t_r\rangle|^p\right)^{1/p} \leq \|t^*\| \left\| \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \right\|.$$

Now,

$$|\langle z, s^* \otimes t^*\rangle| \leq \left(\sum_{r=1}^{N} |\langle s^*, s_r\rangle|^q\right)^{1/q} \left(\sum_{r=1}^{N} |\langle t^*, t_r\rangle|^p\right)^{1/p},$$

hence we obtain

$$|\langle z, s^* \otimes t^*\rangle| \leq \|s^*\| \ \|t^*\| \ \|(s_1, \ldots, s_N)\| \left\| \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \right\|.$$

Therefore, $\|z\|_h = 0$ implies $z = 0$ and we are done.

It remains to check the condition $\mathcal{D}_\infty$ and the four properties (2.1) - (2.4). Let $z = s \odot t \in M_{n,m}(S \otimes T)$ and $z' = s' \odot t' \in M_{n',m'}(S \otimes T)$. Then $\begin{pmatrix} z & \\ & z' \end{pmatrix} = (s \oplus s') \odot \begin{pmatrix} t & \\ & t' \end{pmatrix}$. Hence, applying $\mathcal{D}_\infty$ to $s \oplus s'$ and (2.1) to $\begin{pmatrix} t & \\ & t' \end{pmatrix}$ we obtain that $S \otimes T$ satisfies (2.1). The proofs of (2.2), (2.3), (2.4) and $\mathcal{D}_\infty$ are similar, we omit them. □

Remark 2.4. In the case when S and T are operator spaces, the space $S \otimes_h T$ defined above is the usual Haagerup tensor product of S and T.

Remark 2.5. The tensor product $\otimes_h$ is associative. Thus given p-matrix normed spaces $S_1, \ldots, S_N$, we may define unambigously the space $S_N \otimes_h \cdots \otimes_h S_1$. Let us now come back to Example 2.2. Let $N \geq 2$ and $X_1, \ldots, X_N, Y_1, \ldots, Y_N$ be Banach spaces. For any $1 \leq i \leq N$, we give ourselves $S_i \subset B(X_i, Y_i)$. From above, we may consider the p-matrix normed space $S = S_N \otimes_h \cdots \otimes_h S_1$. Let X, Y be two Banach spaces and let $A : S_N \times \cdots \times S_1 \to B(X, Y)$ be a multilinear map. It may be viewed as a linear map $\widehat{A} : S \to B(X, Y)$ as well. Now it is easy to see that A is p-completely bounded in the sense of Definition 1.1 if and only if $\widehat{A}$ is completely bounded. Moreover, $\left\|\widehat{A}\right\|_{cb} = \|A\|_{pcb}$.

Let E be a Banach space. The identification $E = B(\mathbb{C}, E)$ allows us to define a p-matrix normed space structure on E. To conform with the notation used in the operator space theory, we denote by E_c the p-matrix normed space above. Similarly, we denote by E_r^* the p-matrix normed space structure on E^* defined by the identification $E^* = B(E, \mathbb{C})$. Two simple

facts should be noticed:

(2.9) For any $x_1, \dots, x_n \in E$, $\left\| \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right\|_{E_c} = \left(\sum_{j=1}^{n} \|x_j\|^p \right)^{1/p}$.

(2.10) For any $x_1^*, \dots, x_n^* \in E^*$, $\|(x_1^*, \dots, x_n^*)\|_{E_r^*} = \left(\sum_{i=1}^{n} \|x_i^*\|^q \right)$.

We end this section by two simple lemmas about these p-matrix normed spaces.

Lemma 2.6. *Let S be a p-matrix normed space and let E, F be Banach spaces.*

(a) *For any $u : S \to E_c$,*

$$\|u\|_{cb} = \sup \left\{ \left(\sum_{j=1}^{n} \|u(s_j)\|^p \right)^{1/p} \Big/ \left\| \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix} \right\| \leq 1 \right\}.$$

(b) *For any $v : S \to F_r^*$,*

$$\|v\|_{cb} = \sup \left\{ \left(\sum_{i=1}^{n} \|v(s_i)\|^q \right)^{1/q} \Big/ \|(s_1, \dots, s_n)\| \leq 1 \right\}.$$

Proof. Apply (2.9), (2.3) to show (a) and apply (2.10), (2.4) to show (b). □

Let S be a p-matrix normed space and let E, F be Banach spaces. Let $u : S \to B(E, F^{**})$ be a linear map. We can regard u as a trilinear form $\hat{u}$ on $F^* \times S \times E$ by setting:

$$\hat{u}(b^*, s, e) = \langle u(s)(e), b^* \rangle$$

Lemma 2.7. *The map $u \mapsto \hat{u}$ gives rise to the isometric identification*

$$CB(S, B(E, F^{**})) = (F_r^* \otimes_h S \otimes_h E_c)^*.$$

Proof. Apply (2.9) to E and (2.10) to F. □

Remark 2.8. It should be noticed that the one-dimensional vector space $\mathbb{C}$ may be endowed with several different p-matrix structures. Very natural examples may be obtained as follows. We give oursleves a Banach space X. Let us denote by I_X the identity map on X. Then we set

(2.11) $$\mathbb{C}^X = \text{Span } \{I_X\} \subset B(X, X)$$

and this provides us a p-matrix structure on $\mathbb{C}$. We refer to Section 3 below for more about $\mathbb{C}^X$. In the sequel, we keep the notation $\mathbb{C}$ to refer to the p-matrix normed space $\mathbb{C}^{\mathbb{C}}$. Now let S be a p-matrix normed space. We wish to point out two simple facts.

(a) For any linear form $\xi : S \to \mathbb{C}$, $\|\xi\| = \|\xi\|_{cb}$. This is a straightfoward consequence of the assertions (2.3) and (2.4).

(b) Let X, Y be Banach spaces. Let J (resp. J_1, J_2) be the canonical identification map from S onto $\mathbb{C}^Y \otimes_h S \otimes_h \mathbb{C}^X$ (respectively $S \otimes_h \mathbb{C}^X$, $\mathbb{C}^Y \otimes_h S$). Then it follows from (2.3) and (2.4) again that J, J_1, J_2 are isometric. Moreover they are obviously c.c. maps. However, in general, they are not completely isometric. We will come back to this problem in Remark 4.3.

3. An abstract factorization theorem.

Let X be a Banach space. Given $a = [a_{ij}] \in M_{n,m}$, we let

$$\|a\|_{p,X} = \sup \left\{ \left(\sum_{i=1}^{n} \left\| \sum_{j=1}^{m} a_{ij} x_j \right\|^p \right)^{1/p} \right\} \tag{3.1}$$

where the supremum runs over all the $x_1, \ldots, x_m$ in X which satisfy

$$\sum_j \|x_j\|^p \leq 1.$$

To understand the relation between this definition and preceding ones, consider the subspace $S = \mathbb{C}^X \subset B(X, X)$ defined by (2.11). Let $s = a \otimes I_X \in M_{n,m}(S)$. Then the definitions (1.1) and (3.1) obviously give $\|s\| = \|a\|_{p,X}$.

The following criterion of Hernandez will be used several times.

Theorem 3.2 [**He1**, **He2**]. *Let E and X be Banach spaces. Then $E \in SQ_p(X)$ if and only if:*

$$\forall\, a \in M_n,\ \|a\|_{p,E} \leq \|a\|_{p,X}.$$

Proof. We follow [**Pi1**, Section 3] and refer to this for more informations. Let $A : \mathbb{C}^X \to B(E, E)$ be defined by $A(I_X) = I_E$. Then A is c.c. iff $\forall\, a \in M_n$, $\|a\|_{p,E} \leq \|a\|_{p,X}$. Hence the result follows from Pisier's theorem 1.4. Finally we should mention that in the particular case $X = \mathbb{C}$, this result goes back to Kwapien [**K**]. □

In order to prove our Theorem 3.4 below, we will need techniques used by Pisier in the proof of Theorem 1.4. As in [**Pi1**], the following form of the Hahn-Banach theorem will prove useful.

Lemma 3.3. *Let $\wedge$ be a real vector space equipped with a cone $\wedge_+$. Let $\lambda : \wedge \to \mathbb{R}$ be sublinear and let $\mu : \wedge_+ \to \mathbb{R}_+$ be superlinear. Assume that $\mu \leq \lambda$ on $\wedge_+$. Then there is a positive linear form $f : \wedge \to \mathbb{R}$ such that $\mu \leq f$ on $\wedge_+$ and $f \leq \lambda$ on $\wedge$.*

We are now ready to prove the main result of this section.

Theorem 3.4. *Let X_1, X_2, Y_1, Y_2 be Banach spaces and let $T \subset B(X_1, Y_1)$, $Z \subset B(X_2, Y_2)$ be subspaces. Let S be a matrix normed space and $\sigma : Z \times S \times T \to \mathbb{C}$ be a trilinear map. Assume that S satisfies the condition $\mathcal{D}_\infty$ and that for any $z_1, \dots, z_m \in Z$, $s = [s_{ij}] \in M_m(S)$, $t_1, \dots, t_m \in T$:*

$$\left| \sum_{1 \leq i,j \leq m} \sigma(z_i, s_{ij}, t_j) \right| \leq \|s\| \ \|(z_1, \dots, z_m)\| \ \left\| \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix} \right\|. \tag{3.2}$$

Then there exist Banach spaces $E \in SQ_p(Y_1), F \in SQ_p(X_2)$ and three completely contractive maps $\varphi : S \to B(E, F)$, $u : T \to E_c$ and $v : Z \to F_r^*$ such that:

$$\forall\ (z, s, t) \in Z \times S \times T, \quad \sigma(z, s, t) = \langle \varphi(s)u(t), v(z) \rangle.$$

Proof. Let $\wedge$ be the set of all functions $\phi : X_1 \times Y_2^* \to \mathbb{R}$ for which there exist $\alpha > 0, \beta > 0$ such that

$$\forall\ (x_1, y_2^*) \in X_1 \times Y_2^*,\ |\phi(x_1, y_2^*)| \leq \alpha^p \|x_1\|^p + \beta^q \|y_2^*\|^q. \tag{3.3}$$

Then $\wedge$ is a real vector space and the subset $\wedge_+$ of non-negative functions in $\wedge$ is a cone. We will apply Lemma 3.3 in this space. For any $\phi \in \wedge$, we let:

$$\lambda(\phi) = \inf \left\{ \frac{\alpha^p}{p} + \frac{\beta^q}{q} \right\}$$

where the infimum runs over all $(\alpha, \beta) \in \mathbb{R}_+^{*2}$ such that (3.3) holds. This clearly defines a sublinear map $\lambda : \wedge \to \mathbb{R}$. For any $\phi \in \wedge_+$, we let:

$$\mu(\phi) = \sup \left\{ \sum_{1 \leq i,j \leq m} Re\, \sigma(z_i, s_{ij}, t_j) \right\}$$

where the supremum runs over all $m \geq 1$, $z_1, \dots, z_m \in Z$, $s = [s_{ij}] \in M_m(S)$, $t_1, \dots, t_m \in T$ such that $\|s\| \leq 1$ and

$$\forall\ (x_1, y_2^*) \in X_1 \times Y_2^*,\ \phi(x_1, y_2^*) \geq \sum_{j=1}^m \|t_j(x_1)\|^p + \sum_{i=1}^m \|z_i^*(y_2^*)\|^q.$$

We claim that μ is superadditive on $\wedge_+$. To check this, consider $\phi, \phi' \in \wedge_+, (z_i)_{1\le i\le n}$, $(z'_i)_{1\le i\le m}$ in Z, $(t_j)_{1\le j\le n}, (t'_j)_{1\le j\le m}$ in T such that for any $(x_1, y_2^*) \in X_1 \times Y_2^*$:

$$\phi(x_1, y_2^*) \ge \sum_{j=1}^{n} \|t_j(x_1)\|^p + \sum_{i=1}^{n} \|z_i^*(y_2^*)\|^q \qquad \text{and}$$

$$\phi'(x_1, y_2^*) \ge \sum_{j=1}^{m} \left\|t'_j(x_1)\right\|^p + \sum_{i=1}^{m} \left\|z_i'^*(y_2^*)\right\|^q .$$

Then letting

$$(z''_i)_{i\le n+m} = (z_1, \dots, z_n, z'_1, \dots, z'_m)$$

and

$$(t''_j)_{j\le n+m} = (t_1, \dots, t_n, t'_1, \dots, t'_m),$$

we obtain for all x_1, y_2^* :

$$(\phi + \phi')(x_1, y_2^*) \ge \sum_{j=1}^{n+m} \left\|t''_j(x_1)\right\|^p + \sum_{i=1}^{n+m} \left\|z_i''^*(y_2^*)\right\|^q .$$

Now the point is that if we consider $s = [s_{ij}] \in M_n(S)$ and $s' = [s'_{ij}] \in M_m(S)$ with norms less than one and let $s'' = s \oplus s' = [s''_{ij}] \in M_{n+m}(S)$, we have $\|s''\| \le 1$ (by our assumption on S) and

$$\sum_{i,j} Re\, \sigma(z''_i, s''_{ij}, t''_j) = \sum_{i,j} Re\, \sigma(z_i, s_{ij}, t_j) + \sum_{i,j} Re\, \sigma(z'_i, s'_{ij}, t'_j).$$

We thus obtain $\mu(\phi+\phi') \ge \mu(\phi)+\mu(\phi')$ as claimed above. Hence $\mu : \wedge_+ \to \mathbb{R}$ is a non-negative superlinear map. Let us now prove that:

$$\forall\ \phi \in \wedge_+, \quad \mu(\phi) \le \lambda(\phi). \tag{3.4}$$

We give ourselves $(\alpha, \beta) \in \mathbb{R}_+^{*2}$, $(z_i)_{1\le i\le m}$ in Z and $(t_j)_{1\le j\le m}$ in T such that for any $(x_1, y_2^*) \in X_1 \times Y_2^*$:

$$\sum_j \|t_j(x_1)\|^p + \sum_i \|z_i^*(y_2^*)\|^q \le \phi(x_1, y_2^*) \le \alpha^p \|x_1\|^p + \beta^q \|y_2^*\|^q .$$

Then $\left\| \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix} \right\| \le \alpha$ and $\|(z_1, \dots, z_m)\| \le \beta$ by (2.5) and (2.6). Hence for any $s = [s_{ij}] \in M_m(S)$ of norm less than one, we have by (3.2):

$$\left| \sum_{i,j} \sigma(z_i, s_{ij}, t_j) \right| \le \alpha\beta.$$

Therefore $\sum_{i,j} Re\,\sigma(z_i, s_{ij}, t_j) \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}$ whence (3.4).

By Lemma 3.3, we thus obtain a positive linear form $f : \wedge \to \mathbb{R}$ such that:

$$\forall\, \phi \in \wedge,\ f(\phi) \leq \lambda(\phi) \tag{3.5}$$

$$\forall\, \phi \in \wedge_+,\ \mu(\phi) \leq f(\phi). \tag{3.6}$$

We now come to the definitions of E, F, u, v. We proceed with similar constructions as in [**Pi1**].

Let $\mathcal{G}_1$ be the set of all the functions $\psi : X_1 \to Y_1$ for which there exists $\alpha > 0$ such that for any $x_1 \in X_1$, $\|\psi(x_1)\| \leq \alpha \|x_1\|$. Clearly $\mathcal{G}_1$ is a complex vector space. Moreover, for any $\psi \in \mathcal{G}_1$, the function $\widetilde{\psi} : X_1 \times Y_2^* \to \mathbb{R}$ defined by $\widetilde{\psi}(x_1, y_2^*) = \|\psi(x_1)\|^p$ belongs to $\wedge$, hence we may define:

$$N_1(\psi) = f(\widetilde{\psi})^{1/p}.$$

The function N_1 is a semi-norm on $\mathcal{G}_1$. We denote by G_1 the Banach space obtained after passing to the quotient by the kernel of N_1 and completing the resulting normed space.

For any $t \in T$, let us denote by $\psi_t \in \mathcal{G}_1$ the function defined by $\psi_t(x_1) = t(x_1)$. We may define a linear map $u : T \to G_1$ by setting (up to equivalence classes): $u(t) = p^{1/p}\psi_t$.

Let us regard u as a map from T into $(G_1)_c$. Then $\|u\|_{cb} \leq 1$. Indeed for any finite family $(t_1, \ldots, t_n)$ in T:

$$\begin{aligned}
\tfrac{1}{p}\textstyle\sum_{j=1}^n \|u(t_j)\|^p = \sum_{j=1}^n N_1(\psi_{t_j})^p &= f\left(\textstyle\sum_{j=1}^n \widetilde{\psi}_{t_j}\right) \\
&\leq \left\| \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \right\|^p f\left((x_1, y_2^*) \mapsto \|x_1\|^p\right) \quad \text{by} \quad (2.5) \\
&\leq \left\| \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \right\|^p \lambda\left((x_1, y_2^*) \mapsto \|x_1\|^p\right) \quad \text{by} \quad (3.5) \\
&\leq \tfrac{1}{p} \left\| \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \right\|^p .
\end{aligned}$$

Hence the result follows from Lemma 2.6 (a).

In the same manner, we can introduce the vector space $\mathcal{G}_2$ of all the functions $\psi : Y_2^* \to X_2^*$ for which there exists $\beta > 0$ such that for any $y_2^* \in$

Y_2^*, $\|\psi(y_2^*)\| \leq \beta \|y_2^*\|$. Letting $\widetilde{\psi}(x_1, y_2^*) = \|\psi(y_2^*)\|^q$ and $N_2(\psi) = f(\widetilde{\psi})^{1/q}$, we can similarly define a Banach space G_2 from $(\mathcal{G}_2, N_2)$. We then define a map $v : Z \to G_2$ by letting (up to equivalence classes) $v(z)(y_2^*) = q^{1/q} z^*(y_2^*)$. Then using (2.6) and Lemma 2.6 (b), we obtain that $v : Z \to (G_2^*)_r^*$ satisfies $\|v\|_{cb} \leq 1$.

Finally, we set $E = \overline{u(T)}, F = v(Z)^*$ and can consider that we actually have $u : T \to E_c$ and $v : Z \to F_r^*$ with $\|u\|_{cb} \leq 1$ and $\|v\|_{cb} \leq 1$.

In view of Theorem 3.2, we clearly have $G_1 \in SQ_p(Y_1)$ and therefore $E \in SQ_p(Y_1)$. Similarly, we obtain that $G_2 \in SQ_q(X_2^*)$. Thus by a simple duality argument, we deduce that $F \in SQ_p(X_2)$.

In order to complete the proof of Theorem 3.4, it remains to show that for any $z_1, \ldots, z_m \in Z, s = [s_{ij}] \in M_m(S)$, $t_1, \ldots, t_m \in T$, we have:

$$\left|\sum_{i,j} \sigma(z_i, s_{ij}, t_j)\right| \leq \|s\| \left(\sum_j \|u(t_j)\|^p\right)^{1/p} \left(\sum_i \|v(z_i)\|^q\right)^{1/q}. \tag{3.7}$$

Indeed, such an inequality allows to define $\varphi : S \to B(E, F)$ by letting $\langle \varphi(s)u(t), v(z)\rangle = \sigma(z, s, t)$ and proves that φ is c.c.. Let us now check (3.7). By trivial scaling, we may assume that $\|s\| = 1$ and $\sum_{i,j} \sigma(z_i, s_{ij}, t_j) \in \mathbb{R}^+$. We define $\phi \in \wedge_+$ by setting $\phi(x_1, y_2^*) = \sum_j \|t_j(x_1)\|^p + \sum_i \|z_i^*(y_2^*)\|^q$. Then we have:

$$\left|\sum_{i,j} \sigma(z_i, s_{ij}, t_j)\right| \leq \mu(\phi) \leq f(\phi) \quad \text{by} \quad (3.6)$$

$$\leq \frac{1}{p} \sum_j \|u(t_j)\|^p + \frac{1}{q} \sum_i \|v(z_i)\|^q.$$

Since we have $\sum_{i,j} \sigma(z_i, s_{ij}, t_j) = \sum_{i,j} \sigma(\theta^{-1} z_i, s_{ij}, \theta t_j)$ for any $\theta > 0$, the preceding inequality implies for all $\theta > 0$:

$$\left|\sum_{i,j} \sigma(z_i, s_{ij}, t_j)\right| \leq \frac{\theta^p}{p} \sum_j \|u(t_j)\|^p + \frac{\theta^{-q}}{q} \sum_i \|v(z_i)\|^q.$$

From (2.7), we deduce that (3.7) holds. □

4. A generalization of Ruan's representation theorem.

Let X and Y be Banach spaces and let $S \subset B(X, Y)$ be a subspace. For any $n, m \geq 1$, we may define (unambigously) a p-matrix structure on $M_{n,m}(S)$

by letting $M_{k,l}(M_{n,m}(S)) = M_{kn,lm}(S)$ for all $k, l \geq 1$. In other words, this p-matrix structure is given by the canonical embedding $M_{n,m}(S) \subset B(\ell_p^m(X), \ell_p^n(Y))$. Now let X be a Banach space and let $n \geq 1$ be an integer. Recall the definition (2.11). We set :

$$R_n^X = M_{1,n}(\mathbb{C}^X). \tag{4.1}$$

Actually, R_n^X is a p-matrix structure on the Banach space ℓ_q^n.

Indeed for any $t = (t(\ell))_{1 \leq \ell \leq n} \in \ell_q^n$, let $\widehat{t} : \ell_p^n(X) \to X$ be defined by $\widehat{t}((x_\ell)_{\ell \leq n}) = \sum_{\ell=1}^n t(\ell) x_\ell$. Then $\left\|\widehat{t}\right\| = \left(\sum_{\ell=1}^n |t(\ell)|^q\right)^{1/q} = \|t\|$ and we clearly have:

$$R_n^X = \left\{\widehat{t} \ / \ t \in \ell_q^n\right\} \subset B(\ell_p^n(X), X).$$

In the same manner, given a Banach space Y, we set for any $n \geq 1$:

$$C_n^Y = M_{n,1}(\mathbb{C}^Y). \tag{4.2}$$

C_n^Y is a p-matrix structure on ℓ_p^n. For any $z = (z(k))_{1 \leq k \leq n} \in \ell_p^n$, we may let $\widehat{z}(y) = (z(k)y)_{k \leq n} \in \ell_p^n(Y)$ for all $y \in Y$ and:

$$C_n^Y = \left\{\widehat{z} \ / \ z \in \ell_p^n\right\} \subset B(Y, \ell_p^n(Y)).$$

The spaces R_n^X and C_n^Y will be used in the proof of Proposition 4.2.

The following is the main result of this section:

Theorem 4.1. *Let X, Y be Banach spaces and let S be a matrix normed space. The following assertions are equivalent:*

(i) *S satisfies the two following conditions:*

$\mathcal{D}_\infty$: *For any $s \in M_{n,m}(S)$, $s' \in M_{n',m'}(S)$,*

$$\|s \oplus s'\| = \max\left\{\|s\|, \|s'\|\right\}.$$

$\mathcal{M}_{p,Y,X}$: *For any $a \in M_{n,m}$, $s \in M_m(S), b \in M_{m,n}$,*

$$\|asb\| \leq \|a\|_{p,Y} \ \|s\| \ \|b\|_{p,X}.$$

(ii) *There exist Banach spaces $E \in SQ_p(X), F \in SQ_p(Y)$ and a completely isometric map $J : S \to B(E, F)$.*

This statement will allow us to consider any matrix normed space S which satisfy $\mathcal{D}_\infty$ and $\mathcal{M}_{p,Y,X}$ as a subspace of $B(E, F)$ for some suitable Banach spaces E, F. In the particular case when $p = 2$ and $X = Y = \mathbb{C}$, we recover

Ruan's representation theorem [**R**, **ER3**]. However for $1 < p \neq 2 < +\infty$, the particular case $X = Y = \mathbb{C}$ is already new. We will come back to this in the last Section 6. We do not know whether Theorem 4.1 can be extended to the case $p = 1$. Before coming into the proof of Theorem 4.1, note that a matrix normed space S satisfying the condition (i) above for some Banach spaces X and Y is obviously a p-matrix normed space as defined in Section 2. Although we could not find any convincing example, it seems unlikely that the converse is true. The problem arising here is the following: given a p-matrix normed space S, does there exist a couple of Banach spaces X and Y for which $\mathcal{M}_{p,Y,X}$ holds ?

In order to prove Theorem 4.1, we will follow the approach of [**ER3**]. More precisely, we will deduce the non-trivial implication (i) $\Rightarrow$ (ii) from a convenient factorization of the linear forms $\xi \in M_n(S)^*$.

Proposition 4.2. *Let S be a matrix normed space satisfying the assumptions $\mathcal{D}_\infty$ and $\mathcal{M}_{p,Y,X}$. Let $n \geq 1$ and $\xi \in M_n(S)^*$ with $\|\xi\| = 1$.*

Then there exist Banach spaces $E \in SQ_p(X), F \in SQ_p(Y)$ and a completely contractive map $\varphi : S \to B(E,F)$ such that: $\forall\, s \in M_n(S),\ |\xi(s)| \leq \|\varphi^{(n)}(s)\|$.

Proof. We denote by $T = R_n^X$ and $Z = C_n^Y$ the two p-matrix normed spaces defined in (4.1) and (4.2). Fix $\xi \in M_n(S)^*$ with $\|\xi\| = 1$.

Given $z = (z(k))_{k \leq n} \in Z$ and $t = (t(\ell))_{\ell \leq n} \in T$, we denote by $zt \in M_n$ the matrix obtained by the product of the column matrix $\begin{pmatrix} z(1) \\ \vdots \\ z(n) \end{pmatrix}$ with the row matrix $(t(1), \ldots, t(n))$. Namely, we have $zt = [z(k)t(\ell)]$. With the above notation, we define $\sigma = Z \times S \times T \to \mathbb{C}$ by letting $\sigma(z,s,t) = \xi(zt \otimes s)$.

We claim that σ satisfies the assumption (3.2) of Theorem 3.4. In order to show that, consider $z_1, \ldots, z_m \in Z$, $t_1, \ldots, t_m \in T$ and $s = [s_{ij}] \in M_m(S)$. Let $a = [a_{ki}] \in M_{n,m}$ and $b = [b_{j\ell}] \in M_{m,n}$ be defined by $a_{ki} = z_i(k)$ and $b_{j\ell} = t_j(\ell)$. Clearly we have $\sum_{i,j} \sigma(z_i, s_{ij}, t_j) = \xi(asb)$ hence $\left|\sum_{i,j} \sigma(z_i, s_{ij}, t_j)\right| \leq \|asb\|$. Note that from the definitions (4.1) and (4.2), we have $\|a\|_{p,Y} = \|(z_1, \ldots, z_m)\|$ and $\|b\|_{p,X} = \left\| \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix} \right\|$. Therefore, the assumption $\mathcal{M}_{p,Y,X}$ implies that (3.2) holds.

Moreover we assumed that S satisfies $\mathcal{D}_\infty$. Hence we may apply Theorem 3.4 to the trilinear map σ and this yields two Banach spaces $E \in$

$SQ_p(X), F \in SQ_p(Y)$ and three c.c maps $u : T \to E_c, v : Z \to F_r^*$ and $\varphi : S \to B(E,F)$ such that $\sigma(z,s,t) = \langle \varphi(s)u(t), v(z)\rangle$ for all $(z,s,t) \in Z \times S \times T$. Let us denote by $(\eta_j)_{1\le j\le n}$ and $(\nu_i)_{1\le i\le n}$ the canonical bases of T and Z respectively. Then for any $s = [s_{ij}] \in M_n(S)$,

$$\xi(s) = \sum_{1\le i,j\le n} \sigma(\nu_i, s_{ij}, \eta_j) = \sum_{i,j} \langle \varphi(s_{ij})u(\eta_j), v(\nu_i)\rangle$$

hence

$$|\xi(s)| \le \left\|\varphi^{(n)}(s)\right\| \left(\sum_{j=1}^n \|u(\eta_j)\|^p\right)^{1/p} \left(\sum_{i=1}^n \|v(\nu_i)\|^q\right)^{1/q}.$$

Now $\sum_{j=1}^n \|u(\eta_j)\|^p \le 1$ and $\sum_{i=1}^n \|v(\nu_i)\|^q \le 1$ by Lemma 2.6. Hence $|\xi(s)| \le \left\|\varphi^{(n)}(s)\right\|$. This achieves the proof. □

Proof of Theorem 4.1. We assume (i). Let I_n be the unit sphere of $M_n(S)^*$ and let $I = \bigcup_{n\ge 1} I_n$. For any $\xi \in I_n$, we may apply Proposition 4.2 and thus obtain $E_\xi \in SQ_p(X), F_\xi \in SQ_p(Y)$ and a c.c. map $\varphi_\xi : S \to B(E_\xi, F_\xi)$ such that for any $s \in M_n(S)$, $|\xi(s)| \le \left\|\varphi_\xi^{(n)}(s)\right\|$.

Let $E = \bigoplus_{\xi\in I}^p E_\xi$ and $F = \bigoplus_{\xi\in I}^p F_\xi$. Of course we have $E \in SQ_p(X), F \in SQ_p(Y)$. We now define $J : S \to B(E,F)$ by setting

$$J(s)\big((e_\xi)_{\xi\in I}\big) = \big((\varphi_\xi(s)(e_\xi)\big)_{\xi\in I}\big).$$

Since each $J(s)$ acts diagonally we have for any $s \in M_n(S)$:

$$\left\|J^{(n)}(s)\right\| = \sup_{\xi\in I} \left\|\varphi_\xi^{(n)}(s)\right\|.$$

Therefore, J is a completely isometric map. This proves (i) $\Rightarrow$ (ii). The converse implication is obvious. □

Remark 4.3. Let S be a p-matrix normed space. Note that S satisfies $\mathcal{D}_\infty$. Therefore an obvious reformulation of Theorem 4.1 is that the two following are equivalent:

(i) The canonical identification $\mathbb{C}^Y \otimes_h S \otimes_h \mathbb{C}^X = S$ is completely isometric.

(ii) There exist Banach spaces $E \in SQ_p(X), F \in SQ_p(Y)$ and a completely isometric embedding $S \subset B(E,F)$. This complements Remark 2.8 (b).

Remark 4.4. It is not hard to modify the proofs of Proposition 4.2 and Theorem 3.4 in order to settle an isomorphic variant of Theorem 4.1. Consider the three following properties depending on some constants C_1, C_2, C_3.

(a) For any $s_1 \in M_{n_1,m_1}(S_1), s_2 \in M_{n_2,m_2}(S_2), \ldots, s_k \in M_{n_k,m_k}(S_k)$,

$$\|s_1 \oplus \ldots \oplus s_k\| \leq C_1 \max \{\|s_1\|, \ldots, \|s_k\|\}.$$

(b) For any $a \in M_{n,m}, s \in M_m(S), b \in M_{m,n}$,

$$\|asb\| \leq C_2 \|a\|_{p,Y} \|s\| \|b\|_{p,X}.$$

(c) There exist Banach spaces $E \in SQ_p(X), F \in SQ_p(Y)$ and a complete C_3-isomorphic embedding $J : S \to B(E,F)$.

Then, the assertion (c) implies that (a) and (b) hold with $C_1 = C_2 = C_3$. The converse (and more significant result) is that if (a) and (b) hold, then condition (c) is fulfilled with $C_3 = C_1 C_2$.

5. Representation of p-completely bounded multilinear maps.

In this section we show how to deduce a representation theorem for p-c.b. multilinear maps from our previous work. We will give two formulations of this result. Here is the first one:

Theorem 5.1. *Let $X_1, \ldots, X_N, Y_1, \ldots, Y_N, X, Y$ be Banach spaces. For each $1 \leq i \leq N$, let $S_i \subset B(X_i, Y_i)$ be a subspace. Let $S = S_N \otimes_h \cdots \otimes_h S_1$ (see Remark 2.5 for the definition) and let $A : S \to B(X,Y)$ be a c.b. map.*

Then there are Banach spaces K_i $(1 \leq i \leq N-1)$ and c.b. maps $A_1 : S_1 \to B(X, K_1)$, $A_j : S_j \to B(K_{j-1}, K_j)$ $(2 \leq j \leq N-1)$, $A_N : S_N \to B(K_{N-1}, Y)$ such that

$$\|A_1\|_{cb} \cdots \|A_N\|_{cb} \leq \|A\|_{cb}$$

and:

$$\forall\ (s_N, \ldots, s_1) \in S_N \times \cdots \times S_1,$$

$$A(s_N, \ldots, s_1) = A_N(s_N) \circ \cdots \circ A_2(s_2) \circ A_1(s_1).$$

The proof of Theorem 5.1 will rely upon two lemmas which are now simple corollaries of Section 3 and 4.

Lemma 5.2. *Let X_1, Y_1, X_2, Y_2 be Banach spaces and let $T \subset B(X_1, Y_1)$ and $Z \subset B(X_2, Y_2)$ be subspaces. Then there are Banach spaces $E \in SQ_p(X_1)$, $F \in SQ_p(Y_2)$ and a completely isometric map $J : Z \otimes_h T \to B(E,F)$.*

Proof. Let $z \in M_{m,k}(Z), t \in M_{k,m}(T), a \in M_{n,m}, b \in M_{m,n}$. Then $a(z \odot t)b = az \odot tb$ and $\|az\| \leq \|a\|_{p,Y_2} \|z\|$, $\|tb\| \leq \|t\| \|b\|_{p,X_1}$. Hence we may apply Theorem 4.1 with $S = Z \otimes_h T$, $X = X_1$, $Y = Y_2$. □

Lemma 5.3. *The statement of Theorem* 5.1 *holds in the case* $N = 2$, $X = Y = \mathbb{C}$.

Proof. We consider a c.c. map $A : S_2 \otimes_h S_1 \to \mathbb{C}$. Let $S = \mathbb{C}^{Y_1} \subset B(Y_1, Y_1)$ and let $\sigma : S_2 \times S \times S_1 \to \mathbb{C}$ be defined by $\sigma(s_2, I_{Y_1}, s_1) = A(s_2, s_1)$. Then we may clearly apply Theorem 3.4 with $T = S_1$ and $Z = S_2$ and this yields the result. □

Proof of Theorem 5.1. We follow the approach of [**B**, Theorem 2.4]. Since Lemma 5.2 allows us to use induction, we only need to consider the case $N = 2$. We thus consider a c.c. map $A : S_2 \otimes_h S_1 \to B(X, Y)$. Let us define $\widetilde{A} : (Y_r^* \otimes_h S_2) \otimes_h (S_1 \otimes_h X_c) \to \mathbb{C}$ by setting:

(5.1)

$$\forall\, (y^*, s_2, s_1, x) \in Y^* \times S_2 \times S_1 \times X, \quad \widetilde{A}(y^* \otimes s_2, s_1 \otimes x) = \langle A(s_2, s_1)x, y^* \rangle.$$

From the associativity of $\otimes_h$ (see Remark 2.5) and Lemma 2.7, we have $\left\|\widetilde{A}\right\|_{cb} \leq 1$. Apply Lemma 5.2 to $S_1 \otimes_h X_c$ and $Y_r^* \otimes_h S_2$ together with Lemma 5.3. This yields a Banach space K and two completely contractive maps $\widetilde{A_1} : S_1 \otimes_h X_c \to K_c$ and $\widetilde{A_2} : Y_r^* \otimes_h S_2 \to K_r^*$ such that:

$$\forall\, z \in Y_r^* \otimes_h S_2, \forall\, t \in S_1 \otimes_h X_c, \quad \widetilde{A}(z, t) = \langle \widetilde{A_1}(t), \widetilde{A_2}(z) \rangle. \tag{5.2}$$

We now proceed with converse identifications. We define $A_1 : S_1 \to B(X, K)$ and $A_2 : S_2 \to B(K, Y^{**})$ by setting

$$\forall\, (s_1, x) \in S_1 \times X, \quad A_1(s_1)(x) = \widetilde{A_1}(s_1 \otimes x). \tag{5.3}$$

$$\forall\, (s_2, y^*) \in S_2 \times Y^*, \quad (A_2(s_2))^*(y^*) = \widetilde{A_2}(y^* \otimes s_2). \tag{5.4}$$

Clearly, (5.1), (5.2), (5.3), (5.4) imply that for any $(s_2, s_1) \in S_2 \times S_1$,

$$A(s_2, s_1) = A_2(s_2) \circ A_1(s_1).$$

Now it is easy to see that we may as well assume that $K = \overline{A_1(S_1)(X)}$ and then, A_2 is actually a c.c. map from S_2 into $B(K, Y)$. This concludes the proof. □

Remark 5.4. The converse of Theorem 5.1 obviously holds. Namely, given c.c. maps $A_1 : S_1 \to B(X, K_1)$, $A_N : S_N \to B(K_{N-1}, Y)$ and $A_j : S_j \to B(K_{j-1}, K_j)$ $(2 \leq j \leq N-1)$, the map $A : S_N \times \cdots \times S_1 \to B(X, Y)$ defined by $A(s_N, \ldots, s_1) = A_N(s_N) \circ \cdots \circ A_1(s_1)$ provides a c.c. map from $S_N \otimes_h \cdots \otimes_h S_1$ into $B(X, Y)$.

Remark 5.5. In view of Lemma 5.2, we could have been more precise in the statement of Theorem 5.1. For example we may write that for any $1 \leq j \leq N-1$, $K_j \in SQ_p(Y_j)$. However, we shall see in Theorem 5.6 that such an information is not really an improvement.

We now turn back to the terminology of p-completely bounded maps defined in the introduction (see Definition 1.1). Recall that given $S \subset B(X_1, Y_1)$ and two Banach spaces $G \in SQ_p(X_1), G' \in SQ_p(Y_1)$, it made sense to define a notion of p-representation from S into $B(G, G')$ (see Definition 1.3).

Then by an obvious combination of Theorem 5.1, Remark 5.4, Remark 2.5 and Theorem 1.4, we obtain:

Theorem 5.6. *Let $X_1, \dots, X_N, Y_1, \dots, Y_N, X, Y$ be Banach spaces. For each $1 \leq i \leq N$, let $S_i \subset B(X_i, Y_i)$ be a subspace. Let $A : S_N \times \cdots \times S_1 \to B(X, Y)$ be a N-linear map and let C be a constant. The following assertions are equivalent:*

(i) *A is p-completely bounded and $\|A\|_{pcb} \leq C$.*

(ii) *There exist Banach spaces*

$$G_j \in SQ_p(X_j)(1 \leq j \leq N), G'_j \in SQ_p(Y_j)(1 \leq j \leq N),$$

p-representations $\pi_j : S_j \to B(G_j, G'_j)$ $(1 \leq j \leq N)$ and operators $V_0 : X \to G_1, V_N : G'_N \to Y$ and $V_j : G'_j \to G_{j+1}$ $(1 \leq j \leq N-1)$ such that $\|V_0\| \dots \|V_N\| \leq C$ and $\forall\, (s_N, \dots, s_1) \in S_N \times \cdots \times S_1$,

$$A(s_N, \dots, s_1) = V_N \pi_N(s_N) V_{N-1} \dots V_2 \pi_2(s_2) V_1 \pi_1(s_1) V_0.$$

6. Complements.

6.1. Some remarks about $\otimes_h$. The Haagerup tensor product of operator spaces has been extensively studied recently (see [**B**, **BP**, **BS**, **ER2**, **PS**]). A main feature of this tensor product is that it is both injective and projective in the category of operator spaces. It is then natural to study similar properties in our more general framework. We will easily obtain that our tensor product $\otimes_h$ is projective and is not injective. Let us make these statements precise.

Let S be a matrix normed space and let $T \subset S$ be a closed subspace. We may define a norm on each $M_{n,m}\left(\frac{S}{T}\right)$ by setting $M_{n,m}\left(\frac{S}{T}\right) = \frac{M_{n,m}(S)}{M_{n,m}(T)}$. Endowed with these norms, $\frac{S}{T}$ becomes a matrix normed space. Moreover,

if we assume that S is a p-matrix normed space, then $\frac{S}{T}$ is also a p-matrix normed space.

The announced surjectivity of $\otimes_h$ is:

Proposition 6.1. *Let S_1, S_2 be two p-matrix normed spaces. For $i = 1, 2$, let $T_i \subset S_i$ be a closed subspace and let $q_i : S_i \to \frac{S_i}{T_i}$ be the associated quotient map. Consider $Q = q_2 \otimes q_1 : S_2 \otimes_h S_1 \to \frac{S_2}{T_2} \otimes_h \frac{S_1}{T_1}$.*

Then Q is a complete quotient map, i.e. for any $n \geq 1$, $Q^{(n)}$ is a quotient map.

Proof. Mimic the proof of [**ER2**, Proposition 3.1]. □

Remark 6.2. The tensor product $\otimes_h$ is not injective. Indeed let E, F, G be Banach spaces such that $E \subset F$. Let $j : G_r^* \otimes_h E_c \to G_r^* \otimes_h F_c$ be the canonical embedding. We wish to prove that j is not isometric in general. Assume for simplicity that G is reflexive. Then $(G_r^* \otimes_h E_c)^* = B(E, G)$, $(G_r^* \otimes_h F_c)^* = B(F, G)$ and $j^* : B(F, G) \to B(E, G)$ is the restriction map. Therefore, j^* is onto if and only if any bounded linear map from E into G has a bounded linear extension to F. This fails in general and then, j is not even isomorphic in general.

We now fix two Banach spaces X, Y. Let us denote by $\mathcal{C}_{X,Y}$ the class of all p-matrix normed spaces S defined by a completely isometric embedding $S \subset B(E, F)$ for some $E \in SQ_p(X)$ and $F \in SQ_p(Y)$. Note for further the following straightforward consequence of our Theorem 4.1:

$$\frac{S}{T} \in \mathcal{C}_{X,Y} \quad \text{whenever} \quad S \in \mathcal{C}_{X,Y}. \tag{6.1}$$

The end of this subsection is devoted to a convenient identification result about $\mathcal{C}_{X,Y}$. Let S be a p-matrix normed space. Recall from Section 4 that given $z \in C_n^Y$ and $t \in R_m^X$, we may define $zt \in M_{n,m}$ as a matrix product. Thus we can introduce a canonical map

$$J : C_n^Y \otimes_h S \otimes_h R_m^X \to M_{n,m}(S)$$

by letting $J(z \otimes s \otimes t) = zt \otimes s$.

Proposition 6.3. *Assume that $S \in \mathcal{C}_{X,Y}$. Then the above map J induces a completely isometric identification*

$$C_n^Y \otimes_h S \otimes_h R_m^X = M_{n,m}(S). \tag{6.2}$$

Proof. *1st step.* Under our assumption, it is clear from the proof of Proposition 4.2 that the map J is isometric (see also Remark 4.3).

2nd step. We claim that for any $k, N, n \geq, 1$, we have canonical isometric identifications :

$$M_{1,k}\left(C_N^Y \otimes_h C_n^Y\right) = M_{1,k}\left(C_{Nn}^Y\right) \tag{6.3}$$

$$M_{k,1}\left(R_n^X \otimes_h R_N^X\right) = M_{k,1}\left(R_{nN}^X\right). \tag{6.4}$$

Let us check (6.3). We have the following isometric identifications

$$\begin{aligned} M_{1,k}\left(C_N^Y \otimes_h C_n^Y\right) &= C_N^Y \otimes_h C_n^Y \otimes_h R_k^Y \quad \text{by the first step} \\ &= M_{N,k}\left(C_n^Y\right) \quad \text{by the first step} \\ &= M_{1,k}\left(C_{nN}^Y\right) \quad \text{by (4.2)} \end{aligned}$$

whence (6.3). The proof of (6.4) is similar.

3rd step. We now prove that (6.2) is indeed a completely isometric identification. Fix $N \geq 1$. Then we have (isometrically):

$$\begin{aligned} M_N\left(C_n^Y \otimes_h S \otimes_h R_m^X\right) &= C_N^Y \otimes_h C_n^Y \otimes_h S \otimes_h R_m^X \otimes_h R_N^X \quad \text{by the first step} \\ &= C_{Nn}^Y \otimes_h S \otimes_h R_{mN}^X \quad \text{by (6.3) and (6.4)} \\ &= M_{Nn,Nm}(S) \quad \text{by the first step} \end{aligned}$$

and thus $M_N\left(C_n^Y \otimes_h S \otimes_h R_m^X\right) = M_N(M_{n,m}(S))$. □

6.2. Multilinear Schur products on $B(\ell_p^n)$. Although Schur products have been studied for a long time (see [**Gr**, **Be**]), Haagerup [**Ha**] was the first to realize the link between Schur products and the theory of completely bounded maps. Namely he proved that for any Schur product map $\phi : B(\ell_2^n) \to B(\ell_2^n)$, we have $\|\phi\| = \|\phi\|_{cb}$. This approach was lately exploited in [**PPS**]. We refer to this paper for further information. Recently, Effros and Ruan [**ER4**] proved that multilinear Schur products may be naturally defined on $B(\ell_2^n)$ and that their c.b. norms may be easily computed from the Christensen-Sinclair theorem. Moreover, it is not hard to deduce from [**S**] that for such a multilinear Schur product map $\phi : B(\ell_n^2) \times \cdots \times B(\ell_n^2) \to B(\ell_n^2)$, we have $\|\phi\| = \|\phi\|_{cb}$ as in the linear case. In this last subsection, we will indicate how to generalize all these results to multilinear Schur products on $B(\ell_p^n)$.

In the sequel, we will simply denote by R_n and C_n the p-matrix normed spaces $R_n^{\mathbb{C}}$ and $C_n^{\mathbb{C}}$ defined by (4.1) and (4.2). Similarly, the notation SQ_p will stand for $SQ_p(\mathbb{C})$ and $\mathcal{C}$ will stand for $\mathcal{C}_{\mathbb{C},\mathbb{C}}$. Let $(\varepsilon_i)_{1\leq i\leq n}$ and $(\varepsilon'_j)_{1\leq j\leq n}$ be the canonical bases of R_n and C_n respectively. We set:

$$G_n = \text{Span}\left\{\varepsilon_i \otimes \varepsilon'_j \;/\; i \neq j\right\} \subset R_n \otimes_h C_n.$$

Recall that $(R_n \otimes_h C_n)^* = B(\ell_p^n)$ (see Lemma 2.7 for example). In this duality, $G_n^\perp$ is clearly identified with the space of diagonal operators on $B(\ell_p^n)$. Thus $G_n^\perp = \ell_\infty^n$ and therefore we have isometrically:

$$\ell_1^n = \frac{R_n \otimes_h C_n}{G_n}. \tag{6.5}$$

Now the quotient formula (6.5) defines a p-matrix structure on ℓ_1^n (see the Subsection 6.1). In the sequel we will always consider ℓ_1^n as the p-matrix normed space defined above. Note that from Lemma 5.2, we have $R_n \otimes_h C_n \in \mathcal{C}$. Thus by (6.1) we obtain that $\ell_1^n \in \mathcal{C}$. Note also that when $p = 2$, this space is nothing but Max (ℓ_1^n). Thus the following is not really surprising.

Lemma 6.4. *Let E, F be Banach spaces and let $A : \ell_1^n \to B(E,F)$ be a linear map. Assume that $E \in SQ_p$ and $F \in SQ_p$. Then we have $\|A\|_{cb} = \|A\|$.*

Proof. Let $(\eta_i)_{1\le i\le n}$ be the canonical basis of ℓ_1^n. For any $1 \le i \le n$, let $T_i = A(\eta_i) \in B(E,F)$. We define $\widetilde{A} : F_r^* \otimes_h R_n \otimes_h C_n \otimes_h E_c \to \mathbb{C}$ by setting:

$$\forall\ 1 \le i,j \le n,\ \widetilde{A}(f^*, \varepsilon_i, \varepsilon_j', e) = \delta_{ij}\langle T_i(e), f^*\rangle.$$

By Lemma 2.7 and Proposition 6.1, we have $\left\|\widetilde{A}\right\| = \|A\|_{cb}$. Since $E, F \in SQ_p$, Proposition 6.3 implies that $C_n \otimes_h E_c = (\ell_p^n(E))_c$ and $F_r^* \otimes_h R_n = (\ell_p^n(F))_r^*$ completely isometrically. Thus by Lemma 2.7 again:

$$(F_r^* \otimes_h R_n \otimes_h C_n \otimes_h E_c)^* = M_n(B(E, F^{**})).$$

Under this identification, $\widetilde{A}$ becomes the diagonal matrix $\begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_n \end{pmatrix}$. Therefore $\left\|\widetilde{A}\right\| = \mathrm{Sup}_{i\le n} \|T_i\|$. Since $\|A\| = \mathrm{Sup}_{i\le n} \|T_i\|$, the result follows. □

We now turn to multilinear Schur products. Let $N \ge 1$ and let $n_0, \dots, n_N$ be some fixed positive integers. We give ourselves a finite family of complex numbers $a = (a_{i_N,\dots,i_0})_{\substack{0\le j\le N \\ 1\le i_j\le n_j}}$. Note that any $m(j) \in B\left(\ell_p^{n_{j-1}}, \ell_p^{n_j}\right)$ has a canonical matrix representation $m(j) = [m(j)_{i_j,i_{j-1}}]_{i_j,i_{j-1}}$ with respect to the canonical bases of $\ell_p^{n_{j-1}}$ and $\ell_p^{n_j}$. We define the N-linear Schur product

$$\Phi_a : B\left(\ell_p^{n_{N-1}}, \ell_p^{n_N}\right) \otimes_h \cdots \otimes_h B\left(\ell_p^{n_1}, \ell_p^{n_2}\right) \otimes_h B\left(\ell_p^{n_0}, \ell_p^{n_1}\right) \to B\left(\ell_p^{n_0}, \ell_p^{n_N}\right)$$

associated to a as follows. For any $1 \le j \le N$, let $m(j) = [m(j)_{i_j,i_{j-1}}]_{i_j,i_{j-1}} \in B\left(\ell_p^{n_{j-1}}, \ell_p^{n_j}\right)$. Then we set

$$\Phi_a(m(N), \dots, m(1)) = \left[\sum_{\substack{1\le j\le N-1 \\ 1\le i_j\le n_j}} a_{i_N,\dots,i_0}\ m(N)_{i_N,i_{N-1}} \dots m(2)_{i_2,i_1}\ m(1)_{i_1,i_0}\right]_{i_N,i_0} \in B\left(\ell_p^{n_0}, \ell_p^{n_N}\right).$$

We now introduce another map naturally asociated to a. For any $0 \leq j \leq N$, let us denote by $(\eta_{i_j})_{1 \leq i_j \leq n_j}$ the canonical basis of $\ell_1^{n_j}$. Then we define $\varphi_a : \ell_1^{n_N} \otimes_h \ldots \otimes_h \ell_1^{n_1} \otimes_h \ell_1^{n_0} \to \mathbb{C}$ by setting $\varphi_a(\eta_{i_N}, \ldots, \eta_{i_0}) = a_{i_N,\ldots,i_0}$. We are now ready to state our last result. We keep the notation above.

Theorem 6.5. *The following are equivalent.*

(i) $\|\Phi_a\| \leq 1$

(ii) $\|\Phi_a\|_{cb} \leq 1$

(iii) $\|\varphi_a\| \leq 1$

(iv) *There are Banach spaces $K_1, \ldots, K_N$ which are all in SQ_p and there are linear contractions $T_{i_0} : \mathbb{C} \to K_1$ $(1 \leq i_0 \leq n_0)$, $T_{i_j} : K_j \to K_{j+1} (1 \leq j \leq N-1, 1 \leq i_j \leq n_j)$, $T_{i_N} : K_N \to \mathbb{C} (1 \leq i_N \leq n_N)$ such that for all $i_0, \ldots, i_N$:*

$$a_{i_N,\ldots,i_0} = T_{i_N} \circ \cdots \circ T_{i_1} \circ T_{i_0}.$$

Proof. Recall that for any $0 \leq j \leq N$, the p-matrix normed space $\ell_1^{n_j}$ belongs to $\mathcal{C}$. Thus the equivalence (iii) $\iff$ (iv) follows from Theorem 5.1, Remarks 5.4, 5.5 and Lemma 6.4. Let us now check that (ii) $\iff$ (iii).

Let $S = B\left(\ell_p^{n_{N-1}}, \ell_p^{n_N}\right) \otimes_h \ldots \otimes_h B\left(\ell_p^{n_1}, \ell_p^{n_2}\right) \otimes_h B\left(\ell_p^{n_0}, \ell_p^{n_1}\right)$.

By Proposition 6.3, each $B\left(\ell_p^{n_{j-1}}, \ell_p^{n_j}\right)$ may be completely isometrically identified with $C_{n_j} \otimes_h R_{n_{j-1}}$. Thus by Lemma 2.7, this yields:

$$CB\left(S, B\left(\ell_p^{n_0}, \ell_p^{n_N}\right)\right) = \left(R_{n_N} \otimes_h C_{n_N} \otimes_h \cdots \otimes_h C_{n_1} \otimes_h R_{n_0} \otimes_h C_{n_0}\right)^*.$$

Now since $\otimes_h$ is projective (see Proposition 6.1), $(\ell_1^{n_N} \otimes_h \cdots \otimes_h \ell_1^{n_1} \otimes_h \ell_1^{n_0})^*$ may be viewed as a subspace of $(R_{n_N} \otimes_h C_{n_N} \otimes_h \cdots \otimes_h C_{n_0})^*$. As a consequence, we obtain an isometric embedding $\rho : (\ell_1^{n_N} \otimes_h \cdots \otimes_h \ell_1^{n_0})^* \to CB\left(S, B\left(\ell_p^{n_0}, \ell_p^{n_N}\right)\right)$. Now it is not hard to see that the range of ρ is exactly the set of N-linear Schur products from S into $B\left(\ell_p^{n_0}, \ell_p^{n_N}\right)$ and that $\rho(\varphi_a) = \Phi_a$. This achieves the proof of (ii) $\iff$ (iii).

Since (ii) $\implies$ (i) is obvious, it remains to show that (i) $\implies$ (ii). We follow the approach of [**S**, Theorem 2.1]. First note that given $\beta \in B\left(\ell_p^{n_N}\right)$, $\alpha \in B\left(\ell_p^{n_0}\right)$ and $m(j) \in B\left(\ell_p^{n_{j-1}}, \ell_p^{n_j}\right)$ $(1 \leq j \leq N)$, we may set $\beta(m(N) \otimes \cdots \otimes m(1))\alpha = \beta m(N) \otimes \cdots \otimes m(1)\alpha$. By linearity this allows us to consider the product $\beta s \alpha$ for all $s \in S$. It is easy to check that for any $\alpha_1, \ldots, \alpha_m \in$

$B\left(\ell_p^{n_0}\right), \beta_1, \dots \beta_m \in B\left(\ell_p^{n_N}\right)$, $s = [s_{\ell k}] \in M_m(S)$:

$$\left\| \sum_{1 \le \ell, k \le m} \beta_\ell s_{\ell k} \alpha_k \right\| \le \|s\| \left\| \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} \right\| \|(\beta_1, \dots, \beta_m)\|. \tag{6.6}$$

We now define $D_{n_0} \subset B\left(\ell_p^{n_0}\right)$ (resp. $D_{n_N} \subset B(\ell_p^{n_N})$) as the space of all the diagonal operators on $B\left(\ell_p^{n_0}\right)$ (resp. $B\left(\ell_p^{n_N}\right)$). A main feature of Schur products is that:

$$\forall\ (\beta, s, \alpha) \in D_{n_N} \times S \times D_{n_0}, \quad \Phi_a(\beta s \alpha) = \beta \Phi_a(s) \alpha. \tag{6.7}$$

We are now ready to show that $\|\Phi_a\|_{cb} \le 1$. In order to achieve this, take $s = [s_{\ell k}] \in M_m(S)$ and $x_1, \dots, x_m \in \ell_p^{n_0}$, $y_1^*, \dots, y_m^* \in \left(\ell_p^{n_N}\right)^* = \ell_q^{n_N}$ such that $\|s\| \le 1$ and

$$\sum_{k=1}^m \|x_k\|^p \le 1 \text{ and } \sum_{\ell=1}^m \|y_\ell^*\|^q \le 1. \tag{6.8}$$

We thus have to show that:

$$\left| \sum_{1 \le \ell, k \le m} \langle \Phi_a(s_{\ell k}) x_k, y_\ell^* \rangle \right| \le 1. \tag{6.9}$$

For any $1 \le \ell, k \le m$, write $x_k = (x_k(i_0))_{1 \le i_0 \le n_0}$ and $y_\ell^* = (y_\ell^*(i_N))_{1 \le i_N \le n_N}$. We define $\widehat{x} \in \ell_p^{n_0}$ and $\widehat{y}^* \in \ell_q^{n_N}$ by letting $\widehat{x}(i_0) = \left(\sum_{k=1}^m |x_k(i_0)|^p \right)^{1/p}$ and $\widehat{y}^*(i_N) = \left(\sum_{\ell=1}^m |y_\ell^*(i_N)|^q \right)^{1/q}$. Thus (6.8) imply:

$$\|\widehat{x}\| \le 1 \quad \text{and} \quad \|\widehat{y}^*\| \le 1. \tag{6.10}$$

Now we define $\alpha_k \in D_{n_0}$ as follows. We set $\alpha_k(i_0) = \dfrac{x_k(i_0)}{\widehat{x}(i_0)}$ for any $1 \le i_0 \le n_0$ $\left(\text{with the usual convention } \dfrac{0}{0} = 0\right)$ and we let $\alpha_k = \begin{pmatrix} \alpha_k(1) & & \\ & \ddots & \\ & & \alpha_k(n_0) \end{pmatrix}$.

Similarly we define $\beta_\ell = \begin{pmatrix} \beta_\ell(1) & & \\ & \ddots & \\ & & \beta_\ell(n_N) \end{pmatrix} \in D_{n_N}$ by $\beta_\ell(i_N) = \dfrac{y_\ell^*(i_N)}{\widehat{y}^*(i_N)}$.

Obviously, we have for all $1 \le k, \ell \le m$: $\quad x_k = \alpha_k(\widehat{x})$ and $y_\ell^* = \beta_\ell^*(\widehat{y}^*)$. Hence we have:

$$\begin{aligned}\left|\sum_{1\le \ell,k\le m}\langle \Phi_a(s_{\ell k})x_k, y_\ell^*\rangle\right| &= \left|\sum_{\ell,k}\langle \Phi_a(s_{\ell k})\alpha_k(\widehat{x}), \beta_\ell^*(\widehat{y}^*)\rangle\right| \\ &= \left|\langle \Phi_a\left(\sum_{\ell,k}\beta_\ell s_{\ell k}\alpha_k\right)\widehat{x}, \widehat{y}^*\rangle\right| \quad \text{by (6.7)} \\ &\le \|(\beta_1,\dots,\beta_m\| \left\|\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{pmatrix}\right\| \quad \text{by (6.6) and (6.10).}\end{aligned}$$

Clearly we have

$$\left\|\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_m\end{pmatrix}\right\| = \sup_{1\le i_0\le m}\left(\sum_{k=1}^m |\alpha_k(i_0)|^p\right)^{1/p}.$$

Hence we have

$$\left\|\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{pmatrix}\right\| \le 1.$$

Similarly, $\|(\beta_1,\dots,\beta_m)\| \le 1$ and therefore, (6.9) follows. □

Remark 6.6. In the particular case $N = 1$, the previous factorization theorem can be refined as follows. We give ourselves a family $a = (a_{i,j})_{\substack{1\le i\le n\\ 1\le j\le m}}$ to which we associate a Schur product map $\Phi_a : B(\ell_p^m, \ell_p^n) \to B(\ell_p^m, \ell_p^n)$ as above as well as the linear map $u_a : \ell_1^m \to \ell_\infty^n$ of canonical matrix a. Then the following are equivalent:

(i) $\|\Phi_a\| \le 1$

(ii) The map u_a factors contractively through L_p-spaces, i.e. there exist a measure space (Ω, μ) and linear contractions $T_1 : \ell_1^m \to L_p(\Omega,\mu)$, $T_2 : L_p(\Omega,\mu) \to \ell_\infty^n$ such that $u_a = T_2T_1$.

Indeed by Theorem 6.5, $\|\Phi_a\| \le 1$ if and only if u_a factors contractively through SQ_p-spaces. From the lifting property of ℓ_1 and the extension property of ℓ_∞, this is equivalent to (ii).

The (linear) result mentioned in this remark was learned to me by G. Pisier. It is stated in [**Pi2**, Chapter 5] where explanations on its origine are given.

References

[Be] G. Bennett, *Schur multipliers*, Duke Math. J., **44** (1977), 603-639.

[B] D.P. Blecher, *Tensor product of operator spaces II*, Canadian J. Math., **44** (1992), 75-90.

[BP] D.P. Blecher and V.I. Paulsen, *Tensor product of operator spaces*, J. Funct. Anal., **99** (1991), 262-292.

[BS] D.P. Blecher and R.R. Smith, *The dual of the Haagerup tensor product*, J. London Math. Soc., **45** (1992), 126-144.

[CS] E. Christensen and A.M. Sinclair, *Representations of completely bounded multilinear operators*, J. Funct. Anal., **72** (1987), 151-181.

[E] E.G. Effros, *On multilinear completely bounded module maps*, Contemporary Mathematics, **62** (1987), 163-187.

[EK] E.G. Effros and A. Kishimoto, *Module maps and Hochschild-Johnson cohomology*, Indiana Univ. Math. J., **36** (1987), 257-276.

[ER1] E.G. Effros and Z.-J. Ruan, *A new approach to operator spaces*, Canadian Math. Bull., **34** (1991), 329-337.

[ER2] E.G. Effros and Z.-J. Ruan, *Self-duality for the Haagerup tensor product and Hilbert space factorization*, J. Funct. Anal., **100** (1991), 257-284.

[ER3] E.G. Effros and Z.-J. Ruan, *On the abstract characterization of operator spaces*, Proc. Amer. Math. Soc., **119** (1993), 579-584.

[ER4] E.G. Effros and Z.-J. Ruan, *Multivariable multipliers for groups and their operator algebras*, Proc. of Symposia in Pure Math., **51**, part I (1990), 197-218.

[Gr] A. Grothendieck, *Résumé de la théorie métrique des produits tensoriels topologiques*, Bol da Soc. Mat. Sao Paulo, **8** (1956), 1-79.

[Ha] U. Haagerup, *Decomposition of completely bounded maps on operator algebras*, Unpublished manuscript (1980).

[He1] R. Hernandez, *Espaces L^p, factorisation et produits tensoriels dans les espaces de Banach*, C.R.A.S.A., **296** (1983), 385-388.

[He2] R. Hernandez, *Espaces L^p, factorisation et produits tensoriels dans les espaces de Banach*, Thèse de 3° cycle, Université Paris 6 (1983).

[K] S. Kwapien, *On operators factorizable through L_p-spaces*, Bull. Soc. Math. France Mem., **31-32** (1972), 215-225.

[Pa1] V.I. Paulsen, *Completely bounded maps and dilatations*, Pitman Research notes in Math., **146**, Longman (1986).

[Pa2] V.I. Paulsen, *Completely bounded maps on C^*-algebras and invariant operator ranges*, Proc. Amer. Math. Soc., **86** (1982), 91-96.

[PPS] V.I. Paulsen, S.C. Power, R.R. Smith, *Schur products and matrix completions*, J. Funct. Anal., **85** (1989), 151-178.

[PS] V.I. Paulsen and R.R. Smith, *Multilinear maps and tensor norms on operator systems*, J. Funct. Anal., **73** (1987), 258-276.

[Pi1] G. Pisier, *Completely bounded maps between sets of Banach space operators*, Indiana Univ. Math. J., **39** (1990), 249-277.

[Pi2] G. Pisier, *Similarity problems and completely bounded maps*, Lecture Notes, Springer Verlag, to appear.

[R] Z.-J. Ruan, *Subspaces of C^*-algebras*, J. Funct. Anal., **76** (1988), 217-230.

[S] R.R. Smith, *Completely bounded module maps and the Haagerup tensor product*, J.

Funct. Anal., **102** (1991), 156-175.

[W] G. Wittstock, *Ein operatorwertiger Hahn-Banach Satz*, J. Funct. Anal., **40** (1981), 127-150.

Received July 14, 1993 and revised September 22, 1993.

Laboratoire de Mathematiques
URA CNRS 741
UFR Sciences et Techniques
16, Route de Gray
F - 25030 Besancon Cedex

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

FINITELY GENERATED COHOMOLOGY HOPF ALGEBRAS AND TORSION

JAMES P. LIN

H-spaces X whose mod p cohomology is finitely generated as an algebra are studied. Even generators of infinite height lie in degrees $2p^j$ for $j \geq 0$. If $H^*(X; \mathbb{Z}_p)$ is not finite dimensional, then $H^*(X; \mathbb{Z})$ must have p torsion of all orders if X is two connected.

Introduction.

In this note we begin a study of H-spaces whose cohomology mod p is not finite dimensional, but is finitely generated as an algebra. We study these H-spaces by studying the structure of their Borel decompositions. From the Borel structure theorem, if the mod p cohomology of an H-space is nonfinite, either there are an infinite number of algebra generators or there are elements of infinite height. We prove the following:

Theorem A. *Let A be a mod p cohomology Hopf algebra admitting an action of the Steenrod algebra. If A is finitely generated as an algebra, then the generators of infinite height lie in degrees of the form $2p^j$, for $j \geq 0$.*

In the next theorem we show that nonfinite H-spaces must have unbounded p-torsion in their cohomology.

Theorem B. *Let X be a two-connected H-space and suppose $p^r H^*(X; \mathbb{Z})$ is torsion-free for some $r > 0$. Then if $H^*(X; \mathbb{Z}_p)$ is finitely generated as an algebra, then it is finite dimensional.*

Theorem C. *Let X be a two-connected H-space with $H^*(X; \mathbb{Z}_p)$ finitely generated as an algebra, but not finite dimensional. Then $\beta_1 Q H^{\text{even}}(X; \mathbb{Z}_p) \neq 0$, where β_1 is the first cohomology Bockstein.*

Theorem D. *Let X be a two-connected H-space with $H^*(X; \mathbb{Z}_p)$ finitely-generated as an algebra, but not finite dimensional. Then $H^*(X; \mathbb{Z})$ has p-torsion of all orders.*

Corollary E. *Let X be a homotopy associative H-space with $H^*(X; \mathbb{Z}_p)$ finitely generated as an algebra and primitively generated. If p is an odd*

prime, all even generators lie in degrees $2p^j$, $j \geq 0$, *and for every primitive even generator* x, *of degree greater than* 2, $\beta_1 x \neq 0$.

One should note that the three-connective cover of a Lie group is an example of an H-space whose cohomology mod p is finitely generated as an algebra, but not finite dimensional. Other examples are $K(\mathbb{Z}_p r, 1)s$ and $K(\mathbb{Z}, 2)s$. Presently these are the only examples known to the author. From this, one might conjecture that all H-spaces whose cohomology mod p is finitely generated, but not finite has the mod p cohomology of a product of a finite H-space with copies of three-connective covers of Lie groups and $K(\mathbb{Z}_p r, 1)s$ and $K(\mathbb{Z}, 2)s$.

There are a number of results related to Theorems A, B, C, D. Throughout this discussion, let X be an H-space whose mod p cohomology is finitely generated as an algebra. If the Bockstein β_1 vanishes, Lin [**L1**] shows if p is odd, all even generators lie in degree 2. If X is homotopy commutative and homotopy associative, Slack [**S**] shows X is mod 2 equivalent to a generalized Eilenberg MacLane complex with homotopy groups in degrees 1 and 2. If X is the loops on an a_p space, Lin [**L2**] proves the same result for p odd.

Throughout the entire paper, all spaces will have the homotopy type of a connected CW complex with finitely many cells in each degree. Unless otherwise specified, all cohomology or homology modules will be understood to be with $\mathbb{Z}_p$ coefficients where p is a prime. All Hopf algebras are assumed to be connected and biassociative. We will assume the reader has some familiarity with the concepts of cohomology Hopf algebras. These objects are discussed in detail in [**MS**].

Finally the author would like to express his appreciation to Mike Slack and Clarence Wilkerson for several useful conversations. In particular, Theorem 2.1 is due to Wilkerson. Also the referee considerably shortened the proof of Theorem A.

§1. Elements of Infinite Height.

In this chapter we prove Theorem A using the T-functor technology originally due to Lannes [**L**]. Throughout this chapter, A will be a finitely generated mod p cohomology Hopf algebra over the Steenrod algebra.

Proposition 1.1. *There exists an integer* N *such that the* p^N *powers of elements of* A *form a Hopf subalgebra that has only elements of infinite height.*

Proof. Let $\xi : A \to A$ be the pth power map. Then $\xi^N A$ is a finitely generated Hopf algebra over the Steenrod algebra. If the generators of finite height

have height less than p^N, then $\xi^N A$ is polynomial by the Borel structure theorem. □

Following the notation of Lannes, let $V = (\mathbb{Z}_p)^\ell$ and $BV = K((\mathbb{Z}_p)^\ell, 1)$. Recall

$$H^*(BV; \mathbb{Z}_p) \cong \wedge(x_1, \ldots, x_\ell) \otimes \mathbb{Z}_p[y_1, \ldots, y_\ell]$$

for p odd where the degrees of the x_is is one and the degrees of the y_js is two. For $p = 2$,

$$H^*(BV; \mathbb{Z}_2) = \mathbb{Z}_2[x_1, \ldots, x_\ell] \quad \deg x_i = 1.$$

Proposition 1.2. *Suppose $\xi^N A$ embeds as Hopf algebras into $H^*(BV, \mathbb{Z}_p)$ for some $\ell > 0$. Then the generators of infinite height of A lie in degrees $2p^j$ for $j \geq 0$.*

Proof. If the generators of $\xi^N A$ lie in degrees $2p^k$ for $k > 0$, then the generators of infinite height of A lie in degrees $2p^{k-N}$ so we are done.

Let $g : \xi^N A \to H^*(BV; \mathbb{Z}_p)$ be a Hopf algebra embedding. Then suppose by induction that all generators of degree less than m in $\xi^N A$ lie in degrees twice a power of p. Let B be the subalgebra generated by elements of degree less than m. Then B is a Hopf algebra and $\xi^N A \cong B \otimes \xi^N A//B$ as a left B-module and $H^*(BV; \mathbb{Z}_p) \cong g(B) \otimes H^*(BV; \mathbb{Z}_p)//g(B)$ as a left $g(B)$ module, from [**MiMo**].

It follows that there is an induced monomorphism

$$\xi^N A//B \xrightarrow{\tilde{g}} H^*(BV; \mathbb{Z}_p)//g(B)$$

of Hopf algebras. If x is an algebra generator in degree m, then $0 \neq \tilde{g}\{x\} \in P(H^*(BV; \mathbb{Z}_p)//g(B))$. It follows that degree $\{x\} = 2p^k$ for some $k > 0$. Hence $\deg x = 2p^k$. □

Proposition 1.3. *Let C be a polynomial algebra over $\mathcal{A}(p)$ with finitely many generators of even degrees. There is an embedding of Hopf algebras over the Steenrod algebra $C \to H^*(BV; \mathbb{Z}_p)$ for any appropriately chosen V.*

Proof. By the Adams Wilkerson embedding theorem [**AW**], there is an embedding of domains $f : C \to H^*(BT^n; \mathbb{Z}_p)$. Follow this by the map $H^*(BT^n; \mathbb{Z}_p) \to H^*(B\mathbb{Z}_p^n; \mathbb{Z}_p)$ and call the composite $g : C \to H^*(B\mathbb{Z}_p^n; \mathbb{Z}_p)$. By [**DW**, Thm. 4.1, (1) ⇒ (2)], $T_g^V(C) \cong T_\phi^V(C)$. By [**DW**, Lemma 4.5] $T_\phi^V(C) \cong C$. Finally by [**DMW**, Props. 3.5, 3.6], $T_g^V(C) \to H^*(BV; \mathbb{Z}_p)$ is monic, and hits the smallest $\mathcal{A}(p)$ Hopf algebra containing $g(C)$. □

Proof of Theorem A. Theorem A now follows from Propositions 1.2 and 1.3.

Corollary 1.4.
(1) *$\xi^N A$ is primitively generated.*

(2) *There is a map of Hopf algebras* $A \to H^*(BV; \mathbb{Z}_p)$ *which is monic when restricted to the* p^N *powers of* A.

Proof. $\xi^N A$ may be considered a subHopf algebra of $H^*(BV; \mathbb{Z}_p)$, which is primitively generated. Hence, so is $\xi^N A$. This proves (1). (2) follows since $H^*(BV; \mathbb{Z}_p)$ is an unstable injective in the category of unstable algebras. □

§2. Torsion and Finite Generation.

In this chapter we develop the connection between p-torsion in $H^*(X; \mathbb{Z})$ and finite generation of the Hopf algebra $H^*(X; \mathbb{Z}_p)$, for X an H-space. The following result is due to Wilkerson.

Theorem 2.1. *Let* A *be a differential cohomology algebra mod* p *and suppose* A *is finitely generated as an algebra. Then the homology of* A *is also finitely generated as an algebra.*

Proof. Consider A as a left module over the subring ξA of pth powers. Then the differential d is a ξA-module map and A is finitely generated as a ξA-module. ξA is Noetherian, so A is a Noetherian ξA-module. Hence the cycles $Z(A)$ and boundaries $B(A)$ are also finitely generated ξA modules. Hence the homology $H(A)$ is a finitely generated ξA-module and algebra generators of $H(A)$ may be chosen from $Q(\xi A)$ and the module generators of $Z(A)$. □

Theorem 2.2. *Let* X *be an* H*-space with* $H^*(X; \mathbb{Z}_p)$ *finitely generated as an algebra. If* $p^r H^*(X; \mathbb{Z})$ *is torsion free, than* $\beta_s Q E_s^{\text{even}}$ *is decomposable for* $s \geq 1$ *where* E_s *is the* s*th term of the Bockstein spectral sequence.*

Proof. By a theorem of Browder [**B**] (infinite implications) given $\bar{x} \in QE_s^{2n}$ with $\beta_x \bar{x} = \bar{y} \neq 0$ indecomposable, either there is a new generator $\bar{x}_1 \in QE_s^{2np}$ with $\beta_s \bar{x}_1 = \bar{y}_1 \neq 0$ or $\beta_{s+1}\{x^p\} = \{x^{p-1}y\} \neq 0$ in QE_{s+1}. Now if $p^r H^*(X; \mathbb{Z})$ is torsion free, $E_r = E_\infty$. Let r be the smallest integer such that $E_r = E_\infty$. Then for some $s < r$ there must be an infinite sequence of nontrivial even generators $\bar{x}$, $\bar{x}_1$, $\bar{x}_2$, $\cdots$ in QE_s^{even}. But $H^*(X; \mathbb{Z}_p)$ is finitely generated as an algebra. Hence by Theorem 2.1 each E_s is also finitely generated as an algebra. It follows that we must have $\beta_s QE_s^{\text{even}}$ is decomposable for $s \geq 1$. □

Proof of Theorem B *and Theorem* C. Since $H^*(X; \mathbb{Z}_p)$ is finitely generated as an algebra, the algebra generators of infinite height lie in degrees $2p^j$ for $j \geq 0$ by Theorem A. Further $j \geq 1$ since X is two-connected. Therefore

it suffices to show there are no generators of infinite height. For p odd, by Theorem 1.3.3 of [**L3**], and the fact that $\beta_1 H^{\text{even}}(X; \mathbb{Z}_p)$ is decomposable (by Theorem 2.2), it follows

$$QH^{2p^j}(X; \mathbb{Z}_p) = \mathcal{P}^1 QH^{2p^j-2p+2}(X; \mathbb{Z}_p). \tag{2.1}$$

If $H^*(X; \mathbb{Z}_p)$ has a generator of infinite height, let x be such an element of lowest degree. Then by (2.1) and Theorem A,

$$x = \mathcal{P}^1 y + d \quad \text{where } d \text{ is decomposable.}$$

By choice of x, y and d both have finite height. Hence x has finite height. This is a contradiction, so $H^*(X; \mathbb{Z}_p)$ has no generators of infinite height. The Borel structure theorem implies $H^*(X; \mathbb{Z}_p)$ is finite dimensional.

For $p = 2$, if $H^*(X; \mathbb{Z}_2)$ is not finite dimensional, let x be a Borel generator of infinite height of highest degree, say 2^j. The Adem relations imply for $j > 1$,

$$Sq^{2^j+1} = Sq^{2^j} Sq^1 + Sq^2 Sq^1 Sq^{2^j-2}.$$

We have $j > 1$ since X is two-connected. Then since $Sq^1 = \beta_1$ we have by Theorem 2.2,

$$Sq^1 H^{2^j}(X; Z_2) \quad \text{and} \quad Sq^1 Sq^{2^j-2} H^{2^j}(X; \mathbb{Z}_2) \text{ are decomposable.}$$

We now apply Theorem 1.3 of [**L3**]. Let $B(m)$ be the $A(2)$ subHopf algebra generated by elements of degree less than or equal to m. By induction on m, assume $B(m)$ is finite dimensional, and $\bar{\Delta} x \in B(m) \otimes B(m)$, $x \notin B(m)$. By Theorem 1.3 of [**L3**] there is an element $\phi(x) \in H^{2^{j+1}}(X; \mathbb{Z}_2)$ with

$$\bar{\Delta}\phi(x) = x \otimes x + \text{im } Sq^{2^j} + \text{im } Sq^2 + z \tag{2.2}$$

where $z \in H^* \otimes I(B(m))H^* + H^* I(B(m)) \otimes H^*$.

We will show $\phi(x)$ is a generator of infinite height. This will contradict the fact that x is a generator of highest degree of infinite height.

We first show $\phi(x)$ is indecomposable. Consider the subalgebra B generated by $\xi H^* + \text{im } Sq^1 + \text{im } Sq^2 + B(m)$. Then x has non-zero projection in $Q(H^*(X)//B)$ since $\bar{x}$ is indecomposable and elements of im $Sq^1 + \text{im } Sq^2 + B(m)$ have finite height. Therefore, there is a primitive $t \in H_*(X; \mathbb{Z}_p)$ with

$$\langle t, x \rangle = 1, \quad \langle t, B \rangle = 0.$$

By (2.2), $\langle t^2, \phi(x) \rangle = \langle t, x \rangle^2 = 1$ so $\phi(x)$ is dual to a primitive. Therefore, $\phi(x)$ is indecomposable.

Now z and $Sq^2(H^* \otimes H^*)$ in degree 2^{j+1} both have finite height in $H^* \otimes H^*$, since $B(m)$ has finite height and H^{odd} and $Sq^2 H^{4s+2}$ have finite height for $s \geq 1$, by Theorem A. By (2.2), if ℓ is large,

$$\bar{\Delta}\phi(x)^{2^\ell} = x^{2^\ell} \otimes x^{2^\ell} + \xi^\ell(\text{im } Sq^{2^j}). \tag{2.3}$$

We claim $\phi(x)^{2^\ell} \neq 0$ because $\bar{\Delta}\phi(x)^{2^\ell} \neq 0$. If $\bar{\Delta}\phi(x)^{2^\ell} = 0$ then $x^{2^\ell} = y^{2^{\ell+1}}$ for some y so $(x - y^2)$ has height 2^ℓ. But a Borel generator has the property that the height of x is less than or equal to the height of $x + d$ where d is decomposable. This is a contradiction, so $\bar{\Delta}\phi(x)^{2^\ell} \neq 0$ for all ℓ and $\phi(x)$ is a generator of degree higher than x of infinite height. Further, $\phi(x)$ cannot be changed by decomposables to have finite height. So there must be a Borel generator of degree greater than degree of x of infinite height. This is a contradiction. We conclude $H^*(X; \mathbb{Z}_2)$ is finite dimensional.

To prove Theorem C, notice the above argument shows if $\beta_1 QH^{\text{even}}(X; \mathbb{Z}_p) = 0$ then if $H^*(X; \mathbb{Z}_p)$ is finitely generated as an algebra, there are no elements of infinite height. □

Proof of Theorem D. By Theorem C there exists an even degree generator $x \in H^{2n}(X; \mathbb{Z}_p)$ with $\beta_1 x = y$ where y is indecomposable. By infinite implications we have either E_r is not finitely generated as an algebra or there is an even generator z in E_{r+1} with $\beta_{r+1} z$ indecomposable. By Theorem 2.1 we must have $\beta_{r+1} \neq 0$ for every r. This implies there exists p torsion of all orders. □

Proof of Corollary E. By [**Z1**, Thm. C] $H^*(X; \mathbb{Z}_p)$ is free commutative, so by Theorem A, all even generators lie in degrees $2p^j$. Now suppose there is an even primitive generator $x \in PH^{2p^j}(X; \mathbb{Z}_p)$ with $\beta_1 x = 0$, and $j > 0$. Then the Adem relations imply

$$\beta_1 \mathcal{P}^{p^j} = c_1 \mathcal{P}^{p^j} \beta_1 + c_2 \mathcal{P}^1 \beta_1 \mathcal{P}^{p^j - 1}$$

where $c_1, c_2 \in \mathbb{Z}_p$.

We have $\mathcal{P}^{p^j - 1} x$ is decomposable primitive, hence it is zero since all primitives lie in degrees $2p^\ell$, $\ell \geq 0$. Therefore by [**Z2**], there is a secondary operation $\phi(x)$ with

$$\bar{\Delta}\phi(x) = x \otimes \cdots \otimes x + \text{im } \mathcal{P}^{p^j} + \text{im } \mathcal{P}^1 \beta.$$

Since $x \notin \text{im } \mathcal{P}^1$, there exists a $t \in PH_*(X; \mathbb{Z}_p)$ with

$$\langle t, x \rangle \neq 0 \quad \text{and} \quad \langle t, \text{im } \mathcal{P}^1 \rangle = 0.$$

It follows that

$$\langle t^p, \phi(x)\rangle \neq 0, \quad \text{so} \quad t^p \neq 0.$$

But $H_*(X; \mathbb{Z}_p)$ has no pth powers since $H^*(X; \mathbb{Z}_p)$ is primitively generated. We conclude $\beta_1 x \neq 0$ for every primitive even generator of degree greater than 2. □

References

[AW] J.F. Adams and C. Wilkerson, *Finite H-spaces and algebras over the Steenrod algebra*, Ann. of Math., **111** (1980), 95–143.

[B] W. Browder, *Torsion in H-spaces*, Ann. of Math., **74** (1961), 24–51.

[DMW] W. Dwyer, H. Miller and C. Wilkerson, *Homotopical uniqueness of classifying spaces*, Topology, **31** (1992), 29–45.

[DW] W. Dwyer and C. Wilkerson, *Spaces of null homotopic maps*, Asterisque, **191** (1990), 97–108.

[H] J. Hubbuck, *Finitely generated cohomology Hopf algebras*, Topology, **9** (1970), 205–210.

[L] J. Lannes, *Sur la cohomologie modulo p des p-groupes Abeliens elementaires*, in *Homotopy Theory*, Proc. Durham Symp., 1985, eds. Rees and Jones, Cambridge Univ. Press, Cambridge 1987.

[L1] J. Lin, *H-spaces without simple torsion*, J. Pure Applied Algebra, **11** (1977), 61–66.

[L2] ———, *Loops of H-spaces with finitely generated cohomology rings*, submitted to Topology and Its Applications.

[L3] ———, *Torsion in H-spaces*, I, Ann. of Math., **103** (1976), 457–487.

[MiMo] J. Milnor and J.C. Moore, *On the structure of Hopf algebras*, Ann. Math., **81** (1965), 211–264.

[MS] J.C. Moore and L. Smith, *Hopf algebras and multiplicative fibrations*, I, II, Amer. J. Math., **90** (1968), 752–780, 1113–1150.

[S] M. Slack, *A classification theorem for homotopy commutative H-spaces with finitely generated* mod 2 *cohomology rings*, Memoirs AMS, **92** (1991).

[Z1] A. Zabrodsky, *Implications in the cohomology of H-spaces*, Ill. J. of Math., **16** (1971), 363–375.

[Z2] ———, *Secondary operations in the cohomology of H-spaces*, Ill. J. of Math., **15** (1971), 648–655.

Received June 30, 1993 and revised July 28, 1994.

UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CA 92093-0112

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

THE POSITIVE DIMENSIONAL FIBRES OF THE PRYM MAP

JUAN-CARLOS NARANJO

The fibres of positive dimension of the Prym map are characterized.

Let C be an irreducible complex smooth curve of genus g. Let $\pi : \tilde{C} \longrightarrow C$ be a connected unramified double covering of C.

The *Prym variety* associated to the covering is, by definition, the component of the origin of the Kernel of the norm map

$$P(\tilde{C}, C) = \mathrm{Ker}(Nm_\pi)^0 \subset J\tilde{C}.$$

It is a principally polarized abelian variety (p.p.a.v.) of dimension $g(\tilde{C}) - g = g - 1$.

One defines the Prym map

$$\begin{array}{rccc} P_g : & \mathcal{R}_g & \longrightarrow & \mathcal{A}_{g-1} \\ & (\tilde{C} \xrightarrow{\pi} C) & \longmapsto & P(\tilde{C}, C), \end{array}$$

where $\mathcal{R}_g$ is the coarse moduli space of the coverings π as above and $\mathcal{A}_{g-1}$ stands for the coarse moduli space of p.p.a.v.'s of dimension $g-1$.

It is well-known that this map is generically injective for $g \geq 7$ (Friedman-Smith, Kanev). On the other hand this map is never injective; this is a consequence of the tetragonal construction due to Donagi (see [**Do1**] for a description of the construction). This fact is already implicit in the results of Mumford ([**M**]):

The coarse moduli space $\mathcal{RH}_g$ of unramified double coverings of smooth hyperelliptic curves of genus g has $[\frac{g-1}{2}] + 1$ irreducible components $\mathcal{RH}_{g,t}$, $t = 0, ..., [\frac{g-1}{2}]$. For an element $(\tilde{C}, C) \in \mathcal{RH}_{g,t}$ there exist two hyperelliptic curves

$$p_1 : C_1 \longrightarrow \mathbb{P}^1, \quad p_2 : C_2 \longrightarrow \mathbb{P}^1$$

of genus $g(C_1) = t \leq g - t - 1 = g(C_2)$ such that

a) $\tilde{C} = C_1 \times_{\mathbb{P}^1} C_2$ and

b) $C = \tilde{C}/(\sigma_1 \circ \sigma_2)$, where σ_1 (resp. σ_2) is the involution on $\tilde{C}$ attached to the branched covering $\tilde{C} \longrightarrow C_1$ (resp. $\tilde{C} \longrightarrow C_2$).

Mumford proves (loc. cit. p. 346) that one has an isomorphism of p.p.a.v.

$$P(\tilde{C}, C) \cong JC_1 \times JC_2.$$

Consequently the fibres of the restriction of P_g to $\mathcal{RH}_g$ have positive dimension. In fact $P_g(\mathcal{RH}_{g,t})$ is contained in the product $\mathcal{JH}_t \times \mathcal{JH}_{g-t-1}$, where $\mathcal{JH}_s$ stands for the locus of Jacobians of hyperelliptic curves of genus s. Thus

$$\dim \mathcal{RH}_{g,t} = 2g-1 > \dim \mathcal{JH}_t \times \mathcal{JH}_{g-t-1} = \begin{cases} 2g-4 & \text{if } t \neq 0 \\ 2g-3 & \text{if } t = 0. \end{cases}$$

On the other hand positive dimensional fibres also appear for some coverings of bi-elliptic curves (a curve is called bi-elliptic if it can be represented as a ramified double covering of an elliptic curve).

In this note we characterize the fibres of positive dimension of the Prym map. To state our theorem we need some notation: let $\mathcal{RB}_g$ be the coarse moduli space of the unramified double coverings $\pi : \tilde{C} \longrightarrow C$ such that C is a smooth bi-elliptic curve of genus g. This variety has $[\frac{g-1}{2}] + 2$ irreducible components

$$\mathcal{RB}_g = \left(\bigcup_{t=0}^{[\frac{g-1}{2}]} \mathcal{RB}_{g,t} \right) \cup \mathcal{RB}'_g$$

(see [**N**] for more details).

We obtain:

Theorem. *Assume $g \geq 13$. A fibre of P_g is positive dimensional at $(\tilde{C}, C)$ if and only if C is either hyperelliptic or*

$$(\tilde{C}, C) \in \bigcup_{t \geq 1} \mathcal{RB}_{g,t}.$$

Proof. If C is hyperelliptic we apply the results of Mumford. On the other hand, all the irreducible components of the fibres of $P_{g|\mathcal{RB}_{g,t}}$ are positive dimensional for $t \geq 1$ (see [**N**, §20]). This finishes one implication.

The first step to see the opposite implication is to prove that the curve C is tetragonal (i.e. there exists a g^1_4 on C).

Let $\eta \in JC$ be the two-torsion point characterizing the covering and denote by L the line bundle $\omega_C \otimes \eta$. It is easy to check that L is very ample if C is non-tetragonal. Let Φ_L be the projective embedding of C defined by L.

As in Beauville ([**B**, p. 379]), we replace $\mathcal{R}_g$ and $\mathcal{A}_{g-1}$ by the corresponding functors. Then, the Prym map defines a morphism of functors Pr_g. Our

hypothesis on the fibre of P_g implies that the cotangent map to Pr_g at $(\tilde{C}, C)$ is not surjective. By loc. cit. Prop. (7.5), this map can be shown as the cup-product map

$$S^2H^0(C, L) \longrightarrow H^0(C, L^{\otimes 2})$$

followed by the isomorphism induced in cohomology by $L^{\otimes 2} \cong \omega^{\otimes 2}$. Hence, the non-surjectivity implies that $\Phi_L(C)$ is not a projectively normal curve.

We recall Theorem 1 in [**G-L**]: If L is very ample and

$$\deg(L) \geq 2g + 1 - 2h^1(L) - \mathrm{Cliff}(C),$$

then $\Phi_L(C)$ is projectively normal (where $\mathrm{Cliff}(C)$ is the Clifford index of C).

Since $h^1(L) = 0$ and $\deg(L) = 2g - 2$ one obtains $\mathrm{Cliff}(C) \leq 2$. By using Clifford's Theorem and [**Ma**, Propositions 7 and 8], it follows that the curve either possess a g^1_4 or is plane curve of degree six. The second case contradicts $g \geq 13$.

Thus C is tetragonal. Since $g \geq 13$ the results in [**De**] can be applied: either the fibre is finite (generically, three elements) or we are in one of the following three possibilities: C is either hyperelliptic or bi-elliptic or trigonal.

Assume that C is bi-elliptic. Theorems (9.4), (10.9) and (10.10) in [**N**] states that $P_g^{-1}(P(\tilde{C}, C))$ consists of two points for every $(\tilde{C}, C) \in \mathcal{RB}_{g,0} \cup \mathcal{RB}'_g$, hence

$$(\tilde{C}, C) \in \bigcup_{t \geq 1} \mathcal{RB}_{g,t}.$$

To finish the proof we have to rule out the case: C trigonal. In [**R**], Recillas (cf. also [**Do2**]) establishes an isomorphism

$$\tau : \mathcal{RT}_g \xrightarrow{\cong} \mathcal{M}^{tet,0}_{g-1},$$

where $\mathcal{RT}_g$ is the coarse moduli space of unramified double coverings of trigonal curves and $\mathcal{M}^{tet,0}_{g-1}$ is the moduli space of pairs (X, g^1_4) of tetragonal curves X and a base-point-free tetragonal linear series on X not containing divisors of the form $2x + 2y$. This map satisfies that

$$\tau(\tilde{C}, C) = (X, g^1_4) \Longrightarrow P(\tilde{C}, C) \cong JX \quad \text{(as p.p.a.v.)}.$$

Let us fix $(\tilde{C}, C)$ as above and let $(\tilde{D}, D) \in \mathcal{R}_g$ such that $P(\tilde{D}, D) \cong P(\tilde{C}, C) \cong JX$. Since C is not hyperelliptic, then the singular locus of the theta divisor of $P(\tilde{D}, D)$ has codimension 3 by [**M**, p. 344]. In loc. cit. a list of the Prym varieties with such property appear. We obtain that D is either trigonal or bi-elliptic. Since $P(\tilde{D}, D)$ is the Jacobian of a curve the bi-elliptic case contradicts [**S**].

Hence it suffices to prove that all the fibres of the restriction of P_g to $\mathcal{RT}_g$ are zero dimensional. This follows from the bijection τ. Indeed, a curve X of genus $g \geq 12$ has at most one base-point-free g_4^1 without divisors of the form $2x+2y$; otherwise there exists a map $f : X \longrightarrow \mathbb{P}^1 \times \mathbb{P}^1$ and then either the genus is ≤ 9 or X is bi-elliptic. By [**T**, Lemma (4.3)] the linear series of degree 4 and dimension 1 on a bi-elliptic curve come from g_2^1 linear series on the elliptic curve, thus divisors of the forbidden form appear.

Now the classical Torelli Theorem says that

$$\begin{aligned} \mathcal{M}_{g-1}^{tet,0} &\longrightarrow \mathcal{A}_{g-1} \\ (X, g_4^1) &\longmapsto JX \end{aligned}$$

is injective. Composing with τ we are done. □

Remark. Note that if one drops the hypothesis on the genus, at least one gets that the Clifford index of C is ≤ 2.

References

[B] A. Beauville, *Variétés de Prym et Jacobiennes Intermédiaires*, Ann. Sci. E.N.S., **10** (1977), 309-391.

[De] O. Debarre, *Sur les variétés de Prym des courbes tétragonales*, Ann. Sci. E.N.S., **21** (1988), 545-559.

[Do1] R. Donagi, *The tetragonal construction*, Bull. Amer. Math. Soc., **4** (1981), 181-185.

[Do2] R. Donagi, *The fibres of the Prym map*, Preprint 1992.

[G-L] M. Green and R. Lazarsfeld, *On projective normality of complete linear systems on an algebraic curve*, Invent. Math., **83** (1985), 73-90.

[M] D. Mumford, *Prym varieties, I*, in Contributions to Analysis, New York 1974.

[Ma] G. Martens, *Funktionen von vorgegebener Ordnung auf Komplexen Kurven*, J. reine angew. Math., **320** (1980), 68-85.

[N] J.C. Naranjo, *Prym varieties of bi-elliptic curves*, J. reine angew. Math., **424** (1992), 47-106.

[R] S. Recillas, *Jacobians of curves with g_4^1's are the Prym's of trigonal curves*, Bol. Soc. Mat. Mexicana, **19** (1974), 9-13.

[S] V.V. Shokurov, *Distinguishing Prymians from Jacobians*, Invent. Math., **65** (1981), 209-219.

[T] M. Teixidor, *On translation invariance for W_d^r*, J. reine angew. Math., **385** (1988), 10-23.

Received March 1, 1993. The author was partially supported by the European Science Program, "Geometry of Algebraic Varieties" project, contract no. SC1-0398-C(A) and by the DGICYT no. PS90-0069.

Departament d'Algebra i Geometria
Universitat de Barcelona, Gran Via 585
08007 Barcelona, Spain
E-mail address: naranjo@cerber.ub.es

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

ENTROPY OF A SKEW PRODUCT WITH A Z^2-ACTION

KYEWON KOH PARK

We consider the entropy of a dynamical system of a skew product $\widehat{T}$ on $X_1 x X_2$ where there is a Z^2-action on the fiber X_2. If the Z^2-action comes from a Cellular Automaton map, then the contribution of the fiber to the entropy of the skew product is the directional entropy in the direction of the integral of a skewing function φ from X_1 to Z^2.

1. Introduction.

J. Milnor has defined the notion of directional entropy in the study of dynamics of Cellular Automata [**Mi1**], [**Mi2**]. When the notion is applied to a Z^n action it is considered to be a generalization of the entropy of non co-compact subgroups of Z^n.

In the case of a Z^2-action, we denote the generators of the groups by $\{U, V\}$. Let P be a generating partition under the Z^2-action. We write $P_{i,j} = U^i V^j P$. If a subgroup is generated by $U^p V^q$, then there is a natural way to compute the entropy of $U^p V^q$ as a Z-action on the space. Milnor extended this idea to define the entropy of a vector by embedding Z^2 to the ambient vector space R^2 as follows.

$$h(\vec{v}) = \sup_{B:\text{bounded set}} \overline{\lim}_{t\to\infty} \frac{1}{t} H\left(\bigvee_{(i,j)\in B+[o,t)\vec{v}} P_{i,j} \right).$$

Given a vector $\vec{v}$, we let θ_o be the angle between two vectors $\vec{v}$ and $(1,0)$. Let $w = \frac{1}{\tan\theta_o}$ so that $(w,1)$ is a scalar multiple of the vector $\vec{v}$. It is easy to see that

$$h(\vec{v}) = \lim_{m\to\infty} \lim_{t\to\infty} \frac{1}{t} H\left(\bigvee_{j=0}^{[ty]} \bigvee_{-m+jw<i<m+jw} P_{i,j} \right),$$

where $[a]$ denote the greatest integer $\leq a$.

We note that if $\vec{v} = (p,q)$, then $h(\vec{v}) = h(U^p V^q)$. And it is easy to see that directional entropy is a homogeneous function, that is $h(c\vec{v}) = ch(\vec{v})$ for any $c \in R$.

Directional entropy in the case of a Z^2-action generated by a Cellular Automaton map has been investigated in [**Pa1**, **Pa3**] and [**Si**]. D. Lind

defined a cone entropy, denoted by $h^c(\vec{v})$, of a vector $\vec{v}$. Given a vector $\vec{v} = (x, y)$ and a small angle θ, we consider the vectors $\vec{v}_\theta = (x_\theta, y)$ and $\vec{v}_{-\theta} = (x_{-\theta}, y)$ where x_θ and $x_{-\theta}$ satisfy $\dfrac{y}{x_\theta} = \tan(\theta_o + \theta)$ and $\dfrac{y}{x_{-\theta}} = \tan(\theta_o - \theta)$ respectively. Cone entropy is defined as follows.

$$h^c(\vec{v}) = \lim_{\theta \to 0} \lim_{n \to \infty} \frac{1}{n} H\left(\bigvee_{j=0}^{[ny]} \bigvee_{jx_{-\theta} \le i \le jx_\theta} P_{i,j} \right).$$

From the definition, it is clear that we have $h^c(\vec{v}) \ge h(\vec{v})$.

We say that a Z^2-action is generated by a Cellular Automaton if one of the generators of the Z^2-action, say V, is a block map (a finite code) of U. That is, $(V(x))_i$ depends only on the coordinates $x_{-r}, x_{-r+1}, \ldots, x_r$ [**He**]. We call r the size of the block map V. We will show that in the case of a Z^2-action generated by a Cellular Automaton map, the directional entropy and the cone entropy are the same (Theorem 1).

Let (X_1, ζ_1, μ_1, G) and (X_2, ζ_2, μ_2, H) be two ergodic measure preserving dynamical systems with finite entropy, where G and H denote the respective group. Given an integrable skewing function $\varphi : X_1 \to H$, we define a skew product G-action $\widehat{T}$ on $(X_1 \times X_2, \zeta_1 \times \zeta_2, \mu_1 \times \mu_2)$ such that $\widehat{T}^g(x, y) = (T^g x, F^{\varphi(x)} y)$ where T denotes the G-action of X_1 and F denotes the H-action on X_2. When we have $G = H = Z$, then the entropy of $\widehat{T}$ has been extensively studied by many people (e.g. [**Ab**], [**Ad**], [**Ma**, **Ne**]). It is well known in this case that $h(\widehat{T}) = h(T) + |\int \varphi \, d\mu| h(F)$. The above formula says that, as we expect, the fiber contribution to the entropy is $|\int \varphi \, d\mu| h(F)$.

We investigate the entropy of $\widehat{T}$ when $G = Z$ and $H = Z^2$. Note that the above formula cannot hold when the acting group on the fiber is a more general group, say Z^2. First of all, $\int \varphi \, d\mu$ is in general a vector. Secondly, if the skewing function takes a constant value, say $(1, 1)$, then the fiber contribution should come from the entropy of UV, not necessarily from the whole Z^2-action. We prove that if the fiber Z^2-action is generated by a Cellular Automaton map, then we have the analogous theorem (Theorem 2) to the case when $H = Z$.

We may mention that directional entropy can be also defined in a topological setting. D. Lind constructed an example whose topological entropy does not satisfy the analogue of our Theorem 3 [**Li**]. His example involves a Z^2-action which is not generated by a Cellular Automaton map. It is not clear that Theorem 3 holds for topological entropy when we have a Z^2-action on the fiber generated by a Cellular Automaton map. Lind's example is not interesting in the measure theoretic sense because it has the trivial invariant measure.

We have constructed a counterexample which does not satisfy Theorem 3

[**Pa2**]. For the example we explicitly construct the base transformation and use the Z^2-action due to Thouvenot [**Th**] on the fiber. Both of them are constructed by cutting and staking method. It would be interesting to find out how generally Theorem 3 holds. For example, it is unknown if Theorem 3 is true when we have a topological Markov shift which does not satisfy the condition of Corollary 4. We are more interested in the case when the topological Markov shift has 0-entropy as a Z^2-action.

Although Theorem 2 and 4 are more general than Theorem 1 and 3, we will prove Theorem 1 and 3 because their proofs are easier and more geometric. It is also easy to see the proofs of Theorem 2 and 4 from those of Theorem 1 and 3.

We would like to thank Professor D. Ornstein for helpful discussions and the Referee for many valuable comments.

2. Cone entropy.

Throughout the section we assume that our Z^2-action is generated by a Cellular Automaton map. We denote by $H^m(\vec{v})$

$$\lim_{n\to\infty}\frac{1}{n}H\left(\bigvee_{j=0}^{n-1}\bigvee_{i=-m+jw}^{m+jw}P_{i,j}\right).$$

Note that $H^m(\vec{v})$ is independent of the size of the vector $\vec{v}$. Let τ denote $H(P_{0,0})$.

Lemma 1. *$H^m(\vec{v}) = H^{m'}(\vec{v})$ if $m, m' > 2r + w$.*

Proof. Case 1. $\vec{v}$ is not a scalar multiple of $(1,0)$.

Suppose $m' \geq m$. Clearly from the definition we have $H^{m'}(\vec{v}) \geq H^m(\vec{v})$. Hence it is enough to show $H^{m'}(\vec{v}) \leq H^m(\vec{v})$. Note that

$$\begin{aligned}H^m(\vec{v}) &= \lim_{n\to\infty}\frac{1}{n}H\left(\bigvee_{j=0}^{n-1}\bigvee_{-m+jw\leq i\leq m+jw}P_{i,j}\right)\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1}H\left(\bigvee_{-m+jw\leq i\leq m+jw}P_{i,j}\,\middle|\,\bigvee_{0\leq k<j}\bigvee_{-m+kw\leq i\leq m+kw}P_{i,k}\right)\end{aligned}$$

$$= \lim_{n\to\infty} \frac{1}{n} \left\{ H\left(\bigvee_{-m\le i\le m} P_{i,0} \right) \right.$$
$$+ \sum_{j=1}^{n-1} H\left(\bigvee_{jw\le i\le (j-1)w+r} P_{i,j} \,\middle|\, \bigvee_{0\le k<j} \bigvee_{kw\le i\le 2m+kw} P_{i,k} \right)$$
$$+ \sum_{j=1}^{n-1} H\left(\bigvee_{(j-1)w-r\le i\le jw} P_{i,j} \,\middle|\, \bigvee_{1\le k<j} \bigvee_{-2m+kw\le i\le kw} P_{i,k} \right.$$
$$\left.\left. \bigvee \bigvee_{-2m+jw\le i\le -2m+(j-1)w+r} P_{i,j} \right) \right\}.$$

We make the following observations:

(1)

$$\lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{-m\le i\le m} P_{i,0} \right) = 0 = \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{-m'\le i\le m'} P_{i,0} \right).$$

(2)

$$H\left(\bigvee_{jw\le i<(j-1)w+r} P_{i,j} \,\middle|\, \bigvee_{0\le k<j} \bigvee_{kw\le i\le 2m+kw} P_{i,k} \right)$$
$$\ge H\left(\bigvee_{jw\le i<(j-1)w+r} P_{i,j} \,\middle|\, \bigvee_{0\le k<j} \bigvee_{kw<i\le 2m'+kw} P_{i,k} \right),$$

because we condition on more information.

(3) By the same reason, we have

$$H\left(\bigvee_{(j-1)w-r\le i\le jw} P_{i,j} \,\middle|\, \bigvee_{0\le k<j} \bigvee_{-2m+kw<i<kw} P_{i,k} \right.$$
$$\left. \bigvee \bigvee_{-2m+jw\le i\le -2m+(j-1)w+r} P_{i,j} \right)$$
$$\ge H\left(\bigvee_{(j-1)w-r\le i\le jw} P_{i,j} \,\middle|\, \bigvee_{0\le k<j} \bigvee_{-2m'+kw\le i<kw} P_{i,k} \right.$$
$$\left. \bigvee \bigvee_{-2m'+jw\le i\le -2m'+(j-1)w+r} P_{i,j} \right).$$

These observations together with the formula for $H^m(\vec{v})$ above shows $H^{m'}(\vec{v}) \leq H^m(\vec{v})$.

Case 2. $\vec{v} = \eta(1,0)$ for some real η.

We analogously denote by $H^m(\vec{v})$

$$\lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{i=0}^{[n\eta]} \bigvee_{j=-m}^{m} P_{i,j}\right).$$

We note that

$$\begin{aligned}
H^{m'}(\vec{v}) &= \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{i=0}^{[n\eta]}; \bigvee_{j=-m}^{m+2(m'-m)} P_{i,j}\right) \\
&= \lim_{n\to\infty} \frac{1}{n} \left(H\left(\bigvee_{i=0}^{[n\eta]} \bigvee_{j=-m}^{m} P_{i,j}\right) \right. \\
&\quad \left. + H\left(\bigvee_{i=1}^{[n\eta]} \bigvee_{j=m+1}^{m+2(m'-m)} P_{i,j} \,\middle|\, \bigvee_{i=0}^{[n\eta]} \bigvee_{j=-m}^{m} P_{i,j}\right)\right) \\
&\leq H^m(\vec{v}) + \lim_{n\to\infty} \frac{1}{n} \sum_{j=m+1}^{m+2(m'-m)} H\left(\bigvee_{i=0}^{[n\eta]} P_{i,j} \,\middle|\, \bigvee_{i=0}^{[n\eta]} P_{i,j-1}\right) \\
&\leq H^m(\vec{v}) + \lim_{n\to\infty} \frac{1}{n} \sum_{j=m+1}^{m+2(m'-m)} H\left(\bigvee_{i=0}^{r} P_{i,j} \bigvee \bigvee_{i=[n\eta]-r}^{[n\eta]} P_{i,j}\right) \\
&\leq H^m(\vec{v}) + \lim_{n\to\infty} \frac{1}{n} \sum_{j=m+1}^{m+2(m'-m)} 2r\tau \\
&= H^m(\vec{v}) + \lim_{n\to\infty} \frac{4r\tau(m'-m)}{n} \\
&= H^m(\vec{v}).
\end{aligned}$$

Since we have $H^{m'}(\vec{v}) \geq H^m(\vec{v})$ by definition, the proof is complete. □

Corollary 1. *If $\vec{v}$ is not a scalar multiple of $(1,0)$, then we have*

$$\begin{aligned}
&\left| \frac{1}{n} H\left(\bigvee_{j=0}^{n-1} \bigvee_{-m+jw\leq i\leq m+jw} P_{i,j}\right) - \frac{1}{n} H\left(\bigvee_{j=0}^{n-1} \bigvee_{-m'+jw\leq i\leq m'+jw} P_{i,j}\right) \right| \\
&\leq \frac{1}{n} H\left(\bigvee_{m<|i|\leq m'} P_{i,0}\right) \\
&\leq \tau \frac{2(m'-m)}{n}.
\end{aligned}$$

Theorem 1. $h^c(\vec{v}) = h(\vec{v})$.

Proof. It is enough to show that $h^c(\vec{v}) - h(\vec{v})$ is small. If $\vec{v} = (x, y)$ where $y \neq 0$, then by rescaling, we may assume that $\vec{v} = (x, 1)$. Given any $\varepsilon > 0$, there exists θ such that if $\kappa \leq \theta$, then

(i)

$$\lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{0\leq j<n} \bigvee_{jx_\kappa \leq i \leq jx_{-\kappa}} P_{i,j} \right) < h^c(\vec{v}) + \varepsilon$$

(ii) $|x_{-\theta} - x_\theta| < \gamma$ where γ satisfies that $\gamma\tau < \varepsilon$. There exists m_0 such that if $m \geq m_0$, then

$$\lim \frac{1}{n} H\left(\bigvee_{j=0}^{n-1} \bigvee_{-m+jx\leq i\leq m+jx} P_{i,j} \right) = h(\vec{v}).$$

We choose n_o such that if $n \geq n_o$, then we have

(iii)

$$h(\vec{v}) - \varepsilon < \frac{1}{n} H\left(\bigvee_{0\leq j\leq n-1} \bigvee_{-m_o+jx\leq i\leq m_o+jx} P_{i,j} \right) \leq h(\vec{v}) + \varepsilon,$$

(iv)

$$h^c(\vec{v}) - 2\varepsilon < \frac{1}{n} H\left(\bigvee_{o\leq j\leq n-1} \bigvee_{jx_\theta \leq i\leq jx_{-\theta}} P_{i,j} \right) \leq h^c(\vec{v}) + 2\varepsilon,$$

(v)

$$\frac{1}{n} H\left(\bigvee_{o\leq j<K} \bigvee_{-m_o+jx<i<m_o+jx} P_{i,j} \right) < \varepsilon, \text{ where}$$

$$K = \max\{j : j|x_\theta - x| < m_o \text{ and } j|x_{-\theta} - x| < m_o\},$$

and

(vi)

$$\frac{1}{n} H\left(\bigvee_{o\leq j<n} \bigvee_{jx_\theta \leq i\leq jx_{-\theta}} P_{i,j} \right) \geq \frac{1}{n} H\left(\bigvee_{0\leq j<n} \bigvee_{-m_o+jx\leq i\leq m_o+jx} P_{i,j} \right).$$

We compute

$$
\begin{aligned}
&|h^c(\vec{v}) - h(\vec{v})| \\
&\leq \left| \frac{1}{n} H\left(\bigvee_{o\leq j<n} \bigvee_{jx_\theta \leq \imath \leq jx_{-\theta}} P_{i,j} \right) \right. \\
&\quad \left. - \frac{1}{n} H\left(\bigvee_{o\leq j<n} \bigvee_{-m_o - jx \leq \imath < m_o + jx} P_{i,j} \right) \right| + 3\varepsilon \\
&\leq \left| \frac{1}{n} H\left(\bigvee_{o\leq j<n} \bigvee_{n(x_\theta - x)+jx \leq i \leq n(x_{-\theta}-x)+jx} P_{i,j} \right) \right. \\
&\quad \left. - \frac{1}{n} H\left(\bigvee_{o\leq j<n} \bigvee_{-m_o + jx \leq i < m_o + jx} P_{i,j} \right) \right| + 3\varepsilon \\
&\leq \frac{1}{n} H\left(\bigvee_{n(x_\theta - x)\leq i \leq n(x_{-\theta}-x)} P_{i,o} \right) + 3\varepsilon \\
&\leq \frac{1}{n} \gamma n \tau + 3\varepsilon.
\end{aligned}
$$

Hence we have

$$|h(\vec{v}) - h^c(\vec{v})| < 4\varepsilon.$$

In the case of $\vec{v} = (x, o)$, it is not hard to see that the idea of the second part of the proof of Lemma 1 combined with the idea of the proof above will give the desired result. □

Theorem 2. *If* $\sum_{m=0}^{\infty} H\left(P_{0,1} \middle| \bigvee_{-m\leq i\leq m} P_{\imath,0}\right)$ *is finite, then we have* $h^c(\vec{v}) = h(\vec{v})$.

Proof. We note that if we choose M so that

$$\sum_{m=M}^{\infty} H\left(P_{0,1} \middle| \bigvee_{-m\leq i\leq m} P_{i,0}\right) < \varepsilon,$$

then we get

$$\sum_{k=-m+M}^{m-M} H\left(P_{k,1} \middle| \bigvee_{-m\leq i\leq m} P_{i,0}\right) < 2\varepsilon,$$

for all $m > M$. Using this, it is easy to see that if $m_2 \geq m_1 \geq M$, we have that for any n,

$$\frac{1}{n}H\left(\bigvee_{j=0}^{[ny]}\bigvee_{i=m_2+jw}^{-m_2+jw} P_{i,j}\right) < \frac{1}{n}H\left(\bigvee_{j=0}^{[ny]}\bigvee_{i=-m_1+jw}^{m_1+jw} P_{i,j}\right) + 2\varepsilon + \frac{m_2-m_1}{n}\tau$$

where $\frac{m_2-m_1}{n}\tau$ comes from the difference between $\frac{1}{n}H\left(\bigvee_{i=-m_1}^{m_1} P_{i,0}\right)$ and $\frac{1}{n}H\left(\bigvee_{i=-m_2}^{m_2} P_{i,0}\right)$.

Hence for a given $\varepsilon > 0$, there exist m_o as in Theorem 1 such that for a sufficiently large n,

$$\begin{aligned}
&|h^c(\vec{v}) - h(\vec{v})| \\
&\leq \left|\frac{1}{n}H\left(\bigvee_{o\leq j<n}\bigvee_{n(x_\theta-x)+jx\leq i\leq n(x_{-\theta}-x)+jx} P_{i,j}\right)\right. \\
&\quad \left. -\frac{1}{n}H\left(\bigvee_{o\leq j<n}\bigvee_{-m_o+jx\leq i<m_o+jx} P_{i,j}\right)\right| + 3\varepsilon \\
&\leq \frac{1}{n}\gamma n\tau + 2\varepsilon + 3\varepsilon.
\end{aligned}$$

□

Corollary 2. *If V is a finitary code with finite expected code length, then $h^c(\vec{v}) = h(\vec{v})$.*

Proof. It is easy to see that a finitary code with finite expected code length satisfies the condition of Theorem 2. See [**Pa3**]. □

3. Main Theorem.

Let $\lambda = \mu_1 \times \mu_2$. We denote $\sum_{i=0}^{n-1} \varphi_k(T^i z)$ by $\varphi_k^n(z)$ for $k = 1$ or 2 and $z \in X_1$. Given two partitions, β_1 and β_2, we write $\beta_1 \leq \beta_2$ if β_2 is a finer partition than β_1.

Theorem 3. *$h(\widehat{T}) = h(T) + h(\vec{v})$ where $\vec{v} = \int \varphi \, d\mu = (\int \varphi_1 \, d\mu, \int \varphi_2 \, d\mu)$.*

Proof. Since $\int \varphi \, d\mu$ is finite, as in the case of a Z-valued skewing function, there exists φ' which is bounded and cohomologous to φ. Hence we may assume that φ is bounded. Let $|\varphi_1(z)| \leq L$ and $|\varphi_2(z)| \leq L$. Suppose $\vec{v} = \int \varphi \, d\mu = (x, y)$ where $y \neq 0$. We let α denote the generating partition

of the base. Let β denote a partition of X_2. Both of the partitions α and β can be considered in a natural way to be a partition of $X_1 \times X_2$. For a given $z \in X_1$, we denote the set $\{(z,u) : u \in X_2\}$ by I_z.

Since

$$\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i(\alpha \vee \beta)\right) = \frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\alpha\right) + \frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta \middle| \bigvee_{i=0}^{n-1}\hat{T}^i\alpha\right)$$

and

$$\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta \middle| \bigvee_{i=0}^{n-1}\hat{T}^i\alpha\right) = \int \frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta | I_z\right) d\mu,$$

we have

$$\begin{aligned}
\sup_{\beta} h\left(\hat{T}, \alpha^\vee\beta\right) &= \sup_{\beta} \lim_{n\to\infty} \frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i(\alpha^\vee\beta)\right) \\
&= h\left(\hat{T},\alpha\right) + \sup_{\beta_m} \lim_{n\to\infty} \int \frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m | I_z\right) d\mu \\
&= h\left(\hat{T},\alpha\right) + \lim_{m\to\infty}\lim_{n\to\infty} \int \frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m | I_z\right) d\mu,
\end{aligned}$$

where β_m denote the partition $\bigvee_{i=-m}^{m}\bigvee_{j=0}^{L-1} P_{i,j}$.

We denote $\lim_{n\to\infty}\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m|I_z\right)$ by $h_z\left(\hat{T},\beta_m\right)$.

As in Lemma 1, it is not hard to see that for sufficiently large m and m', we have

$$h_z\left(\hat{T},\beta_m\right) = h_z\left(\hat{T},\beta_{m'}\right).$$

We will show that for sufficiently large m,

$$\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m|I_z\right) \to h(\vec{v}) \text{ as } n\to\infty, \text{ for a.e. } z \in X_1.$$

We denote by x_ℓ the x-intercept of a line in $\mathbb{R}^2$ passing through $\varphi^\ell(z)$ with the same slope as $\vec{v}$. Let

$$\begin{aligned}
s_n &= \max\{x_1, \ldots, x_n\} \text{ and} \\
t_n &= \min\{x_1, \ldots, x_n\}.
\end{aligned}$$

Given $\varepsilon > 0$, let k_o be the integer such that if $k \geq k_o$, then we have

(i)

$$\left| h(\vec{v}) - \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{0}^{[ny]} \bigvee_{i=-k+jw}^{k+jw} P_{i,j} \right) \right| < \varepsilon.$$

Given any $\delta > 0$ and $\varepsilon > 0$, there exists n_o such that if $n \geq n_o$, then we have

(ii)

$$\mu E_1 = \mu \left\{ z : \left| \int \varphi \, d\mu - \frac{1}{n} \varphi^n(z) \right| < \delta \right\} > 1 - \varepsilon,$$

(iii)

$$\left| h(\vec{v}) - \frac{1}{n} H\left(\bigvee_{j=0}^{[ny]} \bigvee_{i=-k_o+jw}^{k_o+jw} P_{i,j} \right) \right| < \varepsilon,$$

(iv)

$$\mu E_2 = \mu \left\{ z : \left| h_z\left(\widehat{T}, \beta_{k_o}\right) - \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} \widehat{T}^i \beta_{k_o} | I_z \right) \right| < \varepsilon \right\} > 1 - \varepsilon,$$

(v) $k_o < \dfrac{\varepsilon}{2} n_o$,

and

(vi) $|s_n - t_n| < 2n\delta$.

We choose $\delta < \varepsilon^2$ and choose n_o satisfying (ii)-(vi) above. We fix m_o such that $k_o < (\varepsilon/2)n_o < m_o < \varepsilon n_o$. For notational convenience, we write m and n instead of m_o and n_o respectively. We note that

$$\begin{aligned} &\bigvee_{j=0}^{n-1} \widehat{T}^j \beta_m \text{ on } I_Z \\ &\leq \beta_m \vee F^{\varphi(z)}(\beta_m) \vee F^{\varphi^2(z)}(\beta_m) \vee \ldots \vee F^{\varphi^{n-1}(z)}(\beta_m) \text{ on } I_z \\ &\leq \bigvee_{j=0}^{\varphi_2^{n-1}(z)+L-1} \bigvee_{i=t_n-m+jw}^{s_n+m+jw} P_{i,j} \text{ on } I_Z. \end{aligned}$$

Since t_n and s_n satisfy that

$$|(t_n + m) - (s_n - m)| = |2m + t_n - s_n| > |2m - 2n\delta| > k_o$$

and

$$|(s_n + m) - (t_n - m)| < \varepsilon n,$$

if $z \in E_1$, then by our Corollary and (ii), we have

$$\begin{aligned}
&\left|\frac{1}{n}H\left(\bigvee_{j=0}^{\varphi_2^{n-1}(z)+L-1}\bigvee_{i=t_n-m+jw}^{s_n+m+jw}P_{i,j}\right)-\frac{1}{n}H\left(\bigvee_{j=0}^{[ny]}\bigvee_{i=-k_o+jw}^{k_o+jw}P_{i,j}\right)\right| \\
&<\frac{1}{n}H\left(\bigvee_{i=t_n-m}^{s_n+m}P_{i,0}\right)+\frac{1}{n}H\left(\bigvee_{j=q_1}^{q_2}\bigvee_{i=t_n-m+jw}^{s_n+m+jw}P_{i,j}\right) \\
&<\frac{1}{n}\tau\varepsilon n+\frac{1}{n}(q_2-q_1)\tau(w+2r) \\
&<\tau(\varepsilon+\delta(w+2r)),
\end{aligned}$$

where $q_1=\min\{[ny],\ \varphi_2^{n-1}(z)+L-1\}$ and $q_2=\max\{[ny],\ \varphi_2^{n-1}(z)+L-1\}$.

Hence we have

$$\begin{aligned}
&\left|\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m|I_z\right)-h(\vec{v})\right| \\
&\le\left|\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m|I_z\right)-\frac{1}{n}H\left(\bigvee_{j=0}^{[ny]}\bigvee_{i=-k_o+jw}^{k_o+jw}P_{i,j}\right)\right|+\varepsilon \\
&\le\left|\frac{1}{n}H\left(\bigvee_{j=0}^{\varphi_2^{n-1}(z)+L-1}\bigvee_{i=t_n-m+jw}^{s_n+m+jw}P_{i,j}\right)-\frac{1}{n}H\left(\bigvee_{j=0}^{[ny]}\bigvee_{i=-k_o+jw}^{k_o+jw}P_{i,j}\right)\right|+\varepsilon \\
&\le\tau(\varepsilon+\delta(w+2r))+\varepsilon.
\end{aligned}$$

Let $E=E_1\cap E_2$. If $z\in E$, then by our choice of m and Corollary 1, we have

$$\begin{aligned}
&\left|\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m|I_z\right)-h_z\left(\hat{T},\ \beta_m\right)\right| \\
&\le\left|\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_m|I_z\right)-\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_k|I_z\right)\right| \\
&\quad+\left|\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\hat{T}^i\beta_k|I_z\right)-h_z\left(\hat{T},\ \beta_k\right)\right|+\left|h_z\left(\hat{T},\ \beta_k\right)-h_z\left(\hat{T},\ \beta_m\right)\right| \\
&\le\varepsilon+\varepsilon+\frac{1}{n}m\tau<\varepsilon(2+\tau).
\end{aligned}$$

Since φ_1 and φ_2 are bounded, it is easy to see that there exists ω such that $\left|h_z\left(\hat{T},\ \beta\right)\right|<\omega$ for all β and all z. We may also assume that $h(\vec{v})$ is

bounded above by ω. Now we compute

$$\begin{aligned}
&\left|\sup_{\beta}\int h_z\left(\widehat{T},\ \beta\right)\ d\mu - h(\vec{v})\right| \\
&\leq \int_E \left|h_z\left(\widehat{T},\ \beta_m\right) - h(\vec{v})\right|\ d\mu + \sup_{\beta}\int_{E^c}\left|h_z\left(\widehat{T},\ \beta\right) - h(\vec{v})\right|\ d\mu + \varepsilon \\
&\leq \int_E \left|h_z\left(\widehat{T},\ \beta_m\right) - \frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\widehat{T}^i\beta_m|I_z\right)\right|\ d\mu \\
&\quad + \int_E \left|\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\widehat{T}^i\beta_m|I_z\right) - h(\vec{v})\right|\ d\mu \\
&\quad + \sup_{\beta}\int_{E^c}\left|h_z\left(\widehat{T},\ \beta\right) - h(\vec{v})\right|\ d\mu + \varepsilon \\
&\leq \varepsilon(2+\tau) + \tau(\varepsilon + \delta(w+2r)) + \varepsilon + 4\omega\varepsilon + \varepsilon \\
&\leq \varepsilon(4 + 2\tau + \tau(w+2r) + 4\omega).
\end{aligned}$$

In the case when $\vec{v} = \int\varphi = \eta(1,0)$ for some real number η, we need to argue differently. We may assume $\eta > 0$. We construct φ' which is cohomologous to φ as follows. Let $\varphi' = (\varphi_1', \varphi_2')$.

(i) φ_1' takes the values $[\eta]-1$, $[\eta]$ and $[\eta]+1$
φ_2' takes the values $-1, 0, 1$.

(ii) In an orbit of a point, φ_2' value, 1 or -1, follows its value 0.

(iii) We use the ergodic theorem to construct φ_2' so that it takes the value 0 for all z's except a set of small measure.

Hence we may assume that φ satisfies these properties.

We let $\beta_m = \bigvee_{i=0}^{[\eta]}\bigvee_{j=-m}^{m} P_{i,j}$. Recall that r denote the size of the block map. As in the previous case,we choose m_o so that if $m \geq m_o$, then

(i) $m_o \geq 10r$,

(ii) $|h(\vec{v}) - H^m(\vec{v})| < \varepsilon$,

(iii) $\mu\left\{z : \left|\sup_\beta \int h_z\left(\widehat{T},\ \beta\right) - h_z\left(\widehat{T},\ \beta_m\right)\right| < \varepsilon\right\} > 1-\varepsilon$.
We fix $m \geq m_o$. We choose n_o so that if $n \geq n_o$, then

(iv) $\mu\left\{z : \left|\frac{1}{n}H\left(\bigvee_{i=0}^{n-1}\widehat{T}^i\beta_m|I_z\right) - h_z\left(\widehat{T},\beta_m\right)\right| < \varepsilon\right\} > 1-\varepsilon$,

(v) $\mu\left\{z : \left|\frac{1}{n}\sum_{i=0}^{n-1}\varphi\left(T^i(z)\right) - \int\varphi\ d\mu\right| < \varepsilon\right\} > 1-\varepsilon$,
and

(vi) $\mu\left\{z : \left|\frac{1}{n}\sum_{i=0}^{k}\varphi_2\left(T^i(z)\right)\right| < \varepsilon \text{ for all } 0 \leq k < n\right\} > 1-\varepsilon$.

Let E denote the set satisfying the above conditions, (iii), (iv), (v) and

(vi). We have $\mu E > 1 - 4\varepsilon$. Let $z \in E$.

Let

$$u = \max\left\{\sum_{i=0}^{k} \varphi_2\left(T^i(z)\right) : k = 0, 1, \dots, n-1\right\}$$

and

$$v = \min\left\{\sum_{i=0}^{k} \varphi_2\left(T^i(z)\right) : k = 0, 1, \dots, n-1\right\}.$$

Since $\eta > 0$, there exists $i_o = \max\{k : \varphi_1^k(z) \le i\}$ for a.e. $z \in X_1$. We denote by $\Psi_2^i(z)$

$$\max\left\{\sum_{\zeta=0}^{k} \varphi_2\left(T^\zeta(z)\right) : 0 \le k \le i_o,\ i - [\eta] \le \varphi_1^k(z) \le i\right\}.$$

Now we compute

$$\begin{aligned}
&\frac{1}{n} H\left(\bigvee_{j=-m}^{m} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) \\
&= \frac{1}{n} H\left(\bigvee_{j=-m+u}^{m+u} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) \\
&\le \frac{1}{n} H\left(\bigvee_{i=0}^{\varphi_1^n(z)} \bigvee_{j=-m+\Psi_2^i(z)}^{u+m} P_{i,j}\right) \\
&\le \frac{1}{n}\left(H\left(\bigvee_{i=0}^{\varphi_1^n(z)} \bigvee_{j=-m+\Psi_2^i(z)}^{m+\Psi_2^i(z)} P_{i,j}\right) + 2 \cdot 2(\varepsilon n) \cdot \tau \cdot r\right) \\
&\le \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} \widehat{T}^i \beta_m | I_z\right) + 4\varepsilon r \tau.
\end{aligned}$$

The second to the last inequality is clear because by the condition (i) on m_o, we have

$$\begin{aligned}
&H\left(\bigvee_{i=0}^{\varphi_1^n(z)} \bigvee_{j=-m+\psi_2^i(z)}^{u+m} P_{i,j} \,\middle|\, \bigvee_{i=0}^{\varphi_1^n(z)} \bigvee_{j=-m+\psi_2^i(z)}^{m+\psi_2^i(z)} P_{i,j}\right) \\
&\le H\left(\bigvee_{j=m}^{u+m} \bigvee_{i=0}^{r-1} P_{i,j} \bigvee \bigvee_{j=v+m}^{u+m} \bigvee_{i=\varphi_1^n(z)-r+1}^{\varphi_1^n(z)} P_{i,j}\right)
\end{aligned}$$

$$\leq u \cdot r \cdot \tau + (u - v) \cdot r \cdot \tau.$$

Since the following inequality is also true

$$\begin{aligned}
&\frac{1}{n} H\left(\bigvee_{i=0}^{n-1} \widehat{T}^i \beta_m | I_z\right) \\
&\leq \frac{1}{n} H\left(\bigvee_{j=-m+v}^{m+u} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) \\
&= \frac{1}{n} H\left(\bigvee_{j=-m}^{m+u-v} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) \\
&\leq \frac{1}{n} H\left(\bigvee_{j=-m}^{m} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) + \frac{2}{n}(u - v) r \cdot \tau \\
&= \frac{1}{n} H\left(\bigvee_{j=-m+u}^{m+u} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) + 4\varepsilon r \tau,
\end{aligned}$$

we have

$$\left| \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} \widehat{T}^i \beta_m | I_z\right) - \frac{1}{n} H\left(\bigvee_{-m+u}^{m+u} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) \right| < 4\varepsilon r \tau.$$

We note that

$$\frac{1}{n} H\left(\bigvee_{-m+u}^{m+u} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right) = \frac{\varphi_1^n(z)}{n} \frac{1}{\varphi_1^n(z)} H\left(\bigvee_{-m+u}^{m+u} \bigvee_{i=0}^{\varphi_1^n(z)} P_{i,j}\right)$$

converges to $h(\vec{v})$.

As in the case of $\vec{v} = \int \varphi \, d\mu = (x, y)$ where $y \neq 0$, it is now clear that

$$\left| \sup_{\beta} \int h_z\left(\widehat{T}, \beta\right) d\mu - h(\vec{v}) \right|$$

can be made arbitraily small. □

Similarly we can prove the following theorem.

Theorem 4. *If* $\sum_{m=0}^{\infty} H\left(P_{0,1} \middle| \bigvee_{-m \leq i \leq m} P_{i,0}\right)$ *is finite, then we have* $h(\widehat{T}) = h(T) + h(\vec{v})$ *where* $\vec{v} = \int \varphi \, d\mu = \left(\int \varphi_1 \, d\mu, \int \varphi_2 \, d\mu\right)$.

The following Corollaries are also almost immediate from the proof of Theorem 3.

Corollary 3. *If* $\sum_{m=0}^{\infty} H\left(P_{0,1} \left| \bigvee_{-k\le j\le k} \bigvee_{-m\le i\le m} P_{i,j}\right.\right)$ *is finite for some* k, *then we have* $h(\widehat{T}) = h(T) + h(\vec{v})$ *where* $\vec{v}$ *is given as above.*

Corollary 4. *If a fiber* Z^2*-action,* F, *satisfies the condition of Corollary* 3 *after a linear transformation by a matrix* A *in* $SL(2, Z)$, *that is,* $A \circ F$ *satisfies the condition, then we have the above formula in Corollary* 3 *for the entropy.*

References

Ab, Ro] L.M. Abramov and V.A. Rohlin, *The entropy of a skew product of measure preserving transformations*, AMS Translations, Ser. 2.

[Ad] R. Adler, *A note on the entropy of skew product transformations*, Am. Math. Soc., **4** (1963), 665-669.

[He] G.A. Hedlund, *Endomorphisms and automorphisms of the shift dynamical system*, Math. Syst. Theor., **3** (1969), 320-375.

[Li] D. Lind, personal communication.

Ma, Ne] S.B. Marcus and S. Newhouse, *Measure of maximal entropy for a class of skew products*, Springer Lect. Notes Math., **729** (1979), 105-125.

[Mi1] J. Milnor, *On the entropy geometry of cellular automata*, Complex Systems, **2** (1988), 357-386.

[Mi2] ———, *Directional entropies of cellular automation-maps*, Nato ASI Series, vol. F20, (1986), 113-115.

[Pa1] K.K. Park, *On the continuity of directional entropy*, Osaka J. Math., **31** (1994), 613-628.

[Pa2] ———, *A counter example of the entropy of the skew product*, preprint.

[Pa3] ———, *Continuity of directional entropy for a class of* Z^2*-actions*, J. Korean Math. Soc., **32** (1995), 573-582.

[Si] Y. Sinai, *An answer to a question by J. Milnor*, Comment. Math. Helv., **60** (1985), 173-178.

[Th] J.P. Thouvenot, personal communication.

Received February 10, 1993 and revised June 24, 1993. This research has been supported in part by NSF DMS 8902080 and GARC-KOSEF.

AJOU UNIVERSITY
SUWON, 441-749
KOREA

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

COMMUTING CO-COMMUTING SQUARES AND FINITE DIMENSIONAL KAC ALGEBRAS

TAKASHI SANO

A relationship between finite dimensional Kac algebras and specified commuting co-commuting squares is discussed. The Majid's bicrossproduct Kac algebra is explained in our context.

1. Introduction.

The theory of Kac algebras (Hopf algebras) has been drawing considerable attention (see [**6**] for the reference), and in fact many intensive studies have been made recently. ([**1**, **18**, **19**, **34**, **35**, **36**], etc.) On the other hand, the announcement by A.Ocneanu ([**20**, **21**]) brought us a new aspect in the theory of Kac algebras : it is his claim (proved in [**4**, **17**] and also [**28**]) that, for an irreducible inclusion of factors $M \supset N$ with finite index and depth $= 2$, M is described as the crossed product algebra of N by an outer action of a finite dimensional Kac algebra. Hence, we investigate Kac algebras from the Jones index theoretical point of view.

The purpose of this paper is to find a finite dimensional Kac algebra via the index theory : let $L \supset K$ be an irreducible inclusion of factors with finite index. Suppose that, for an intermediate subfactor M, both inclusions $L \supset M$ and $M \supset K$ are of depth 2. Although the inclusion $L \supset K$ does not always satisfy the depth 2 condition, it can be proved that this pair is of depth 2 if these factors L, M, K, and another intermediate subfactor N form a commuting co-commuting square. Details will be explained in §2 after recalling basic facts on commuting co-commuting squares. Another criterion for the inclusion $L \supset K$ to be of depth 2 is also obtained. Examples are given in §3.

The author would like to express his sincere gratitude to Professor Shigeru Yamagami for helpful advice (in fact the present work was motivated by [**33**]) and to Professor Hideki Kosaki for fruitful discussions and constant encouragement. He is grateful to the referee for many useful comments.

2. Main results.

Let

$$\begin{array}{ccc} L & \supset & M \\ \cup & & \cup \\ N & \supset & K \end{array}$$

be a quadruple of type II_1 factors satisfying $[L : K] < \infty$. (For the standard facts on the index theory, see [**8**, **11**, **23**, **25**, **26**].) It is said to be a commuting square if $E_M^L(N) \subseteq K$, where E_M^L is the conditional expectation from L to M. (See [**8**] for other equivalent conditions.) A quadruple (L, M, N, K) is said to be a co-commuting square if the quadruple

$$\begin{array}{ccc} K' & \supset & M' \\ \cup & & \cup \\ N' & \supset & L', \end{array}$$

or equivalently, that of basic extensions

$$\begin{array}{ccc} \langle L, e_K^L \rangle & \supset & \langle L, e_M^L \rangle \\ \cup & & \cup \\ \langle L, e_N^L \rangle & \supset & L \end{array}$$

on the standard space $L^2(L)$ is a commuting square. Here, e_M^L, e_N^L, and e_K^L are relevant Jones projections (see [**27**, **30**, **31**]). For a commuting co-commuting square (L, M, N, K), we have $K = M \cap N$ and $L = M \vee N$. (In [**27**], a quadruple satisfying these equations is called a quadrilateral and, for a quadrilateral (L, M, N, K), $\mathrm{Ang}(M, N) = Op - \mathrm{ang}(M, N) = \{\frac{\pi}{2}\}$ corresponds to the commuting co-commuting condition.)

For a commuting square, we have characterization of co-commutativity ([**27**, Corollary 7.1] and [**26**, Proposition 1.1.5]).

Proposition 2.1. *Let (L, M, N, K) be a commuting square of type II_1 factors satisfying $[L : K] < \infty$. Then the following are equivalent :*

(1) *(L, M, N, K) is co-commuting.*

(2) $L = M \cdot N = \{\sum_{i \in F} m_i n_i ; F \text{ is a finite set}, m_i \in M, n_i \in N\}$.

(3) $[L : M] = [N : K]$.

(4) *A Pimsner-Popa basis for $N \supset K$ is also that for $L \supset M$.*

Remark that, in [**26**], a commuting square satisfying (one of) the above conditions is called "non-degenerate" and (1),(2) of the following proposition are mentioned in [**26**, Proposition 1.1.6] (see also [**9**, Proposition 2.3]). We will see them for the completeness of this article.

Proposition 2.2. *Let (L, M, N, K) be a commuting co-commuting square of type II_1 factors satisfying $[L : K] < \infty$. Then,*

(1) *$\langle M, e_K^M \rangle \supset M$ is conjugate to $\langle M, e_N^L \rangle \supset M$.*

(2) *The quadrilateral $(\langle L, e_N^L \rangle, \langle M, e_N^L \rangle, L, M)$ is also commuting co-commuting.*

(3) *$\langle L, e_K^L \rangle$ is identified with the Jones extension for $\langle L, e_N^L \rangle \supset \langle M, e_N^L \rangle$.*

Proof. (1) While the condition $\sum_i a_i e_K^M b_i = 0$ $(a_i, b_i \in M)$ is equivalent to $\sum_i a_i E_K^M(b_i c) = 0$ for $c \in M$ on $L^2(M)$, the condition $\sum_i a_i e_N^L b_i = 0$ $(a_i, b_i \in M)$ means $0 = \sum_i a_i E_N^L(b_i c d) = \sum_i a_i E_N^L(b_i c) d = \sum_i a_i E_K^M(b_i c) d$ for $c \in M, d \in N$ on $L^2(L)$ thanks to Proposition 2.1.(2) and the commuting square condition. Hence, we may consider the map $\phi : \langle M, e_K^M \rangle \to \langle M, e_N^L \rangle$ defined by $\phi(\sum_i a_i e_K^M b_i) = \sum_i a_i e_N^L b_i$. It is easy to see that this map ϕ gives an isomorphism between them and $\phi|_M = \mathrm{id}$.

(2) follows from [**8**, Corollary 4.2.3], [**11**, Proposition 3.1.7], and Proposition 2.1.

(3) Since the commuting square condition means $e_M^L e_N^L = e_K^L$, we have $\langle \langle L, e_N^L \rangle, e_M^L \rangle = \langle L, e_K^L \rangle$. We will show that $\langle L, e_K^L \rangle = \langle \langle L, e_N^L \rangle, e_M^L \rangle$ is the Jones extension for $\langle L, e_N^L \rangle \supset \langle M, e_N^L \rangle$. The commuting square condition implies $[e_M^L, x] = 0$ for $x \in \langle M, e_N^L \rangle$. And for the conditional expectation $E^{\langle L, e_K^L \rangle}_{\langle L, e_N^L \rangle}$, by [**27**, Lemma 7.2], we have $E^{\langle L, e_K^L \rangle}_{\langle L, e_N^L \rangle}(e_M^L) = \frac{1}{[L:M]}$. Therefore, we get the conclusion by [**24**, Proposition 1.2.(2)]. □

Thus, we have extensions of a commuting co-commuting square (L, M, N, K) in compatible ways.

For an irreducible inclusion, we have a refined estimation of the dimension of relative commutant algebras as in [**8**, Theorem 4.6.3] (cf. [**11**, Corollary 2.2.3]). We will see this in terms of sectors ([**10**, **12**, **14**, **15**, **16**]).

Lemma 2.1. *For an irreducible inclusion $M \supset N$ of type II_1 factors satisfying $[M : N] < \infty$,*

$$\dim(M_k \cap N') \leq [M : N]^k,$$

where $N \subset M = M_0 \subset M_1 \subset \cdots$ is the Jones tower.

Proof. We only treat the case $k = 2$ since a similar proof will work for any k. We may assume that M and N are properly infinite and isomorphic (by [**16**, Lemma 2.3]) and denote N by $\rho(M)$ for an endomorphism $\rho \in End(M)$. Consider the irreducible decompositions : $\bar{\rho}\rho = \sum_j m_j \alpha_j, \bar{\rho}\rho\bar{\rho} = \sum_{j,k} m_j n_{jk} \beta_k$ $(\alpha_j \bar{\rho} = \sum_k n_{jk} \beta_k)$, where $\bar{\rho}$ is the conjugate sector of ρ. By

the Frobenius reciprocity, we have $\beta_k\rho \geq \sum_j n_{jk}\alpha_j, \alpha_j\bar{\rho} \geq m_j\bar{\rho}$. Combining these, we get

$$\begin{aligned}[M:N]^2 = ([M:N]_0^2 =)\, d(\rho)^4 &= \sum_{j,k} m_j n_{jk} d(\beta_k\rho) \\ &\geq \sum_{j,k} m_j n_{jk} \sum_{j'} n_{j'k} d(\alpha_{j'}) \\ &\geq \sum_k \left(\sum_j m_j n_{jk} \right)^2 = \dim(M_2 \cap N')\end{aligned}$$

thanks to the additivity and the multiplicativity of the statistical dimension d. □

As a corollary of [**8**, Theorem 4.6.3], we have the following ([**10**, Proposition 4.2]) :

Corollary 2.1. *Let $M \supset N$ be an irreducible inclusion of type II_1 factors with finite index. Then the following are equivalent :*

(1) *The inclusion $M \supset N$ is of depth 2.*

(2) $\dim(M_1 \cap N') = [M:N]$.

(3) *$M_2 \cap N'$ is a factor.*

We give another lemma to prove main results.

Lemma 2.2. *Let $(P, Q, R, \mathbf{C})$ be a commuting square of finite dimensional algebras. Then we have*

$$\dim P \geq \dim Q \cdot \dim R.$$

Proof. Let us take a linear basis $\{x_1, x_2, \cdots, x_m\}$ for R and a Pimsner-Popa basis $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ for $Q \supset \mathbf{C}$ with respect to the conditional expectation E from P to R ($m := \dim R, n := \dim Q$). Then $x_i\lambda_j^* (\neq 0)$ are linearly independent ; suppose that $\sum_{i,j} a_{ij} x_i \lambda_j^* = 0$ for $a_{ij} \in \mathbf{C}$. Since $0 = \left(\sum_{i,j} a_{ij} x_i \lambda_j^*\right) \lambda_k = E\left(\sum_{i,j} a_{ij} x_i \lambda_j^* \lambda_k\right) = \sum_{i,j} a_{ij} x_i E\left(\lambda_j^* \lambda_k\right) = \sum_i a_{ik} x_i$ for any k, we have $a_{ik} = 0$. Hence, we get $\dim P \geq nm$. □

For a given commuting co-commuting square, we can get a kind of tiling by double sequences $\{M_{ij}\}_{i,j=0,1,2,\cdots}$ of subfactors (see [**22**, **32**]). By looking at a tiling, we have two criteria for an irreducible inclusion $L \supset K$ to be of depth 2 :

Theorem 2.1. *Let (L, M, N, K) be a commuting co-commuting square of type II_1 factors satisfying $[L : K] < \infty$ and $L \cap K' = \mathbf{C}$. If both inclusions $L \supset M$ and $M \supset K$ are of depth 2, then so is the inclusion $L \supset K$.*

Proof. Let us denote extensions by $\{M_{ij}\}_{i,j=0,1,\cdots}$ such that

$$\begin{aligned}(M_{11}, M_{10}, M_{01}, M_{00}) = (L, M, N, K), M_{22} &= \langle M_{11}, e_{00}^{11} \rangle, \\ M_{21} = \langle M_{11}, e_{01}^{11} \rangle, M_{20} &= \langle M_{10}, e_{01}^{11} \rangle, \\ M_{12} = \langle M_{11}, e_{10}^{11} \rangle, M_{33} &= \langle M_{22}, e_{11}^{22} \rangle, \\ M_{32} = \langle M_{22}, e_{12}^{22} \rangle, M_{31} &= \langle M_{21}, e_{12}^{22} \rangle, \\ M_{30} = \langle M_{20}, e_{12}^{22} \rangle, &\quad \text{and so on.}\end{aligned}$$

Here, e_{kl}^{ij} means the Jones projection for the inclusion $M_{ij} \supset M_{kl}$.

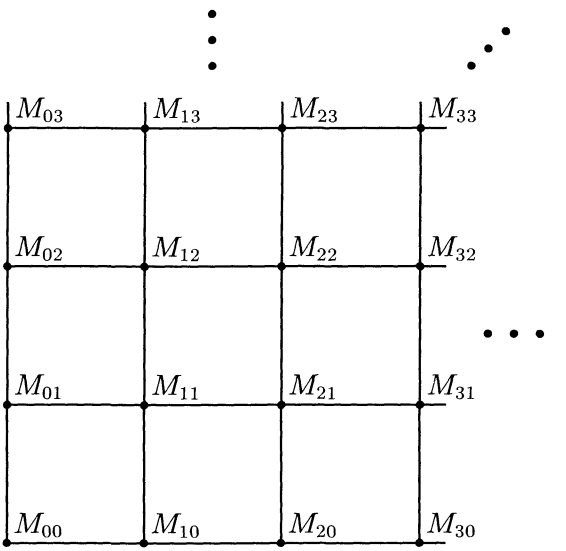


Figure 1.

Clearly, we have $M_{22} \cap M_{00}' \supset M_{20} \cap M_{00}', M_{12} \cap M_{10}'$. But it can be shown that

$$M_{22} \cap M_{00}' = (M_{20} \cap M_{00}') \cdot (M_{12} \cap M_{10}').$$

Let us think of the following commuting square (for the conditional expectations of the restriction of the canonical trace on M_{22}) :

$$\begin{array}{ccc} M_{22} \cap M_{00}' & \supset & M_{12} \cap M_{10}' \\ E \cup & & \cup \\ M_{20} \cap M_{00}' & \supset & \mathbf{C}. \end{array}$$

Here, we remark that $(M_{22}, M_{20}, M_{12}, M_{10})$ forms a commuting square. It follows from Corollary 2.1 that $\dim(M_{20} \cap M'_{00}) = [M : K](=: m)$ and $\dim(M_{12} \cap M'_{10}) = [L : M](=: n)$. Applying Lemma 2.2 to this square, we get $\dim(M_{22} \cap M'_{00}) \geq mn = [L : K]$. Combining this with Lemma 2.1, we have that $\dim(M_{22} \cap M'_{00}) = [L : K]$ and $M_{22} \cap M'_{00} = (M_{20} \cap M'_{00}) \cdot (M_{12} \cap M'_{10})$. Therefore, we get the conclusion by Corollary 2.1. □

Theorem 2.2. *Let (L, M, N, K) be a commuting co-commuting square of type II_1 factors satisfying $[L : K] < \infty$ and $L \cap K' = \mathbf{C}$. If both inclusions $M \supset K$ and $N \supset K$(or $L \supset M$ and $L \supset N$) are of depth 2, then so is the inclusion $L \supset K$.*

Proof. Let us keep the same notation as in the proof of Theorem 2.1. It is sufficient to consider the case that $M \supset K$ and $N \supset K$ are of depth 2 since another case can be proved by looking at the extension $(M_{22}, M_{21}, M_{12}, M_{11})$. For the commuting square $(M_{22} \cap M'_{00}, M_{22} \cap M'_{20}, M_{20} \cap M'_{00}, \mathbf{C})$, we remark that

$$M_{22} \cap M'_{20} \cong M_{02} \cap M'_{00}$$

by Proposition 2.2.(3) and Takesaki duality between $M_{21} \supset M_{20}$ and $M_{01} \supset M_{00}$, which follows from a similar argument in [**23**, Proposition 1.5] about a common Pimsner-Popa basis for $M_{10} \supset M_{00}$ and $M_{11} \supset M_{01}$. Applying Lemma 2.2 and 2.1 to the commuting square $(M_{22} \cap M'_{00}, M_{22} \cap M'_{20}, M_{20} \cap M'_{00}, \mathbf{C})$, we get that $M_{22} \cap M'_{00} = (M_{22} \cap M'_{20}) \cdot (M_{20} \cap M'_{00})$, and $\dim(M_{22} \cap M'_{00}) = [L : K]$. Therefore, we get the theorem by Corollary 2.1. □

Remark. Let $(L, M, N, K) = (M_{11}, M_{10}, M_{01}, M_{00})$ be a commuting co-commuting square as in Theorem 2.2. The Majid's bicrossproduct method corresponds to looking at the quadruple $(M_{21}, M_{20}, M_{11}, M_{10})$ and the relative commutant algebra $M_{32} \cap M'_{10}$.

3. Examples.

In this section, we will explain two examples. The first one is considered in [**33**, Proposition].

(1) Let G be a finite group with two subgroups A, B satisfying $G = AB$ and $A \cap B = \{e\}$. Let γ be an outer action of G on a type II_1 factor P. Then we have

Proposition 3.1. *The inclusion of crossed product algebras*

$$(L :=)(P \otimes l^\infty(G/B)) \rtimes G \supset P \rtimes A(=: K)$$

is irreducible and of depth 2, *where the action of* G *on* $l^\infty(G/B)$ *is induced by the left translation.*

Proof. Let us consider the commuting co-commuting square

$$((P \otimes l^\infty(G/B)) \rtimes G, (P \otimes l^\infty(G/B)) \rtimes A =: M, P \rtimes G =: N, P \rtimes A).$$

Since $(P \otimes l^\infty(G/B)) \rtimes G \cap (P \rtimes A)' = l^\infty(G/B)^A$, the assumption $G = AB$ corresponds to the irreducibility of the inclusion $L \supset K$. Considering Takesaki duality between $L \supset M$ and $P \rtimes B \supset P$ as in the proof of Theorem 2.1, and Proposition 2.2.(1) for $M \supset K (\supset P)$, we also see that $L \supset M$ and $M \supset K$ are of depth 2. Hence, applying Theorem 2.1 to this square (L, M, N, K), we get the conclusion. □

Remark. The Jones tower and the tower of relative commutant algebras can be explicitly written down as in [**3**, **13**, **29**] ; the Jones tower is

$$K = P \rtimes A \subset (P \otimes l^\infty(G/B)) \rtimes G = L$$

$$\subset (P \otimes B(l^2(G/B)) \otimes l^\infty(G/A)) \rtimes G =: L_1$$

$$\subset (P \otimes B(l^2(G/B)) \otimes l^\infty(G/B) \otimes B(l^2(G/A))) \rtimes G =: L_2$$

$$\cdots .$$

And the tower of relative commutant algebras is

$$\mathbf{C} = K \cap K' \subset L \cap K' = l^\infty(G/B)^A = \mathbf{C}$$

$$\subset L_1 \cap K' = (B(l^2(G/B)) \otimes l^\infty(G/A))^A$$

$$\subset L_2 \cap K' = \{B(l^2(G/B)) \otimes l^\infty(G/B) \otimes B(l^2(G/A))\}^A$$

$$\cong B(l^2(G/B)) \otimes B(l^2(G/A))$$

$$\cdots .$$

Hence, we also see that the depth of $L \supset K$ is 2. Next we recall the matched pair ([**18**, **19**]) ; because of the uniqueness of the decomposition of an element in $G = AB = BA$, we can represent ab for $a \in A, b \in B$ as

$$ab = \alpha_a(b)\beta_{b^{-1}}(a^{-1})^{-1} \in BA.$$

The associative law implies

$$\alpha_{aa'}(b) = \alpha_a(\alpha_{a'}(b)), \alpha_a(bb') = \alpha_a(b)\alpha_{\beta_{b^{-1}}(a^{-1})^{-1}}(b'),$$

$$\beta_{bb'}(a) = \beta_b(\beta_{b'}(a)), \beta_b(aa') = \beta_b(a)\beta_{\alpha_{a^{-1}}(b^{-1})^{-1}}(a')$$

for $a, a' \in A, b, b' \in B$. Therefore, the matched pair (A, B, α, β) in [18, Theorem 2.3] appears. (Here, we remark that if we write $ab = \gamma_a(b^{-1})^{-1}\delta_{b^{-1}}(a)(\in BA)$, then the matched pair of another type (A, B, γ, δ) in [**19**] is obtained, but in this article we would like to treat the former one for our purpose.)

For the matched pair (A, B, α, β), we have a finite dimensional Kac algebra of Majid's type ([**19**]) ; the bicrossproduct Kac algebra consists of the crossed product algebra $Q := l^\infty(B) \rtimes_\alpha A$ on $l^2(B) \otimes l^2(A)$ (and others, see below) generated by $m_f \otimes 1$ (simply denoted by $f \otimes 1 = f$) for $f \in l^\infty(B)$ and $u_a \otimes \lambda_a$ (simply denoted by λ_a) for $a \in A$, where m_f is the pointwise multiplication operator on $l^2(B)$, the action α of A on $l^\infty(B)$ is induced by the action α of A on B ; $\alpha_a(f)(b) = f(\alpha_{a^{-1}}(b))$ for $f \in l^\infty(B)$, u_a is the implementing unitary on $l^2(B)$ such that $(u_a\xi)(b) = \xi(\alpha_{a^{-1}}(b))$ for $\xi \in l^2(B)$, and λ_a is the left regular translation; $(\lambda_a\xi)(a') = \xi(a^{-1}a')$ for $\xi \in l^2(A)$.

We know the Kac algebra structure of this crossed product algebra and its dual Kac algebra ([**19**]) ; for the crossed product algebra $Q = l^\infty(B) \rtimes_\alpha A$, the comultiplication Γ, the antipode κ, and the Haar weight ψ are described by :

$$\Gamma(\chi_b) = \sum_{b'b''=b} \chi_{b'} \otimes \chi_{b''},$$

$$\Gamma(\lambda_a) = \sum_b \chi_b \lambda_a \otimes \lambda_{\beta_{b^{-1}}(a)},$$

$$\kappa(\chi_b) = \chi_{b^{-1}},$$

$$\kappa(\lambda_a) = \sum_b \chi_b \lambda_{\beta_b(a^{-1})},$$

$$\psi\left(\sum_a f_a \lambda_a\right) = \frac{1}{|B|} \sum_b f_e(b)$$

for $f_a \in l^\infty(B)$, and χ_b is the characteristic function on $b \in B$. And we have

$$(l^\infty(B) \rtimes_\alpha A)\hat{} = B_\beta \ltimes l^\infty(A),$$

where the right-hand side is generated by $1 \otimes f (f \in l^\infty(A))$ and $\lambda_b \otimes v_b$ $(b \in B)$ on $l^2(B) \otimes l^2(A)((v_b\xi)(a) = \xi(\beta_{b^{-1}}(a)),\ \xi \in l^2(A))$.

The above Kac algebra $\mathbf{K} = (l^\infty(B) \rtimes_\alpha A (= Q), \Gamma, \kappa, \psi)$ has a left action ([**5**]) on the factor $K = P \rtimes_\gamma A$; let us write two generators $\pi(p)(p \in P)$ and $\lambda'_a (a \in A)$ of $P \rtimes_\gamma A$:

$$(\pi(p)\xi)(a') = \gamma_{a'^{-1}}(p)\xi(a'), (\lambda'_a\xi)(a') = \xi(a^{-1}a')$$

for $\xi \in l^2(A, L^2(P))$.

Lemma 3.1. *The following map* $\delta_K : K \to K \otimes Q$ *gives a left action of the Kac algebra* $\mathbf{K}$ *on* $K = P \rtimes A$:

$$\delta_K(\pi(p)) = \sum_b \pi(\gamma_{b^{-1}}(p)) \otimes \chi_b,$$
$$\delta_K(\lambda'_a) = \sum_b \lambda'_{\beta_{b^{-1}}(a)} \otimes \chi_b \lambda_a.$$

This lemma follows from direct computation, hence the author leaves its proof to the reader.

So far, we are now ready to give the theorem.

Theorem 3.1. *The factor* $(P \otimes l^\infty(G/B)) \rtimes G$ *is described as the crossed product algebra of* $P \rtimes A$ *by the left action* δ_K *in Lemma* 3.1 *of the Majid's bicrossproduct algebra* $\mathbf{K} = (l^\infty(B) \rtimes A, \Gamma, \kappa, \psi)$.

Proof. We may think that three kinds of generators $\tilde{\pi}(p)$ $(p \in P), \tilde{\pi}(f)$ $(f \in l^\infty(G/B))$, and $\tilde{\lambda}_g$ $(g \in G)$ of the crossed product algebra $(P \otimes l^\infty(G/B)) \rtimes G$ look like

$$(\tilde{\pi}(p)\xi)(aB, g') = \gamma_{g'^{-1}}(p)\xi(aB, g'),$$
$$(\tilde{\pi}(f)\xi)(aB, g') = f(aB)\xi(aB, g'),$$
$$(\tilde{\lambda}_g\xi)(aB, g') = \xi(g^{-1}aB, g^{-1}g')$$

for $\xi \in l^2(G/B \times G, L^2(P))$. Identifing $G/B \times G$ with $A \times B \times A$ by

$$(a'B, g = ba) \leftrightarrow (a, b, a'),$$

we may write these generators acting on $L^2(P) \otimes l^2(A) \otimes l^2(B) \otimes l^2(A)$ such as

$$(\tilde{\pi}(p)\xi)(a, b, a') = \gamma_{(ba)^{-1}}(p)\,\xi(a, b, a'),$$
$$(\tilde{\pi}(f)\xi)(a, b, a') = f(a')\,\xi(a, b, a'),$$
$$\left(\tilde{\lambda}_{\tilde{a}}\xi\right)(a, b, a') = \xi\left(\beta_{b^{-1}}(\tilde{a})^{-1}a, \alpha_{\tilde{a}^{-1}}(b), \tilde{a}^{-1}a'\right),$$
$$\left(\tilde{\lambda}_{\tilde{b}}\xi\right)(a, b, a') = \xi\left(a, \tilde{b}^{-1}b, \beta_{\tilde{b}^{-1}}(a')\right)$$

for $f \in l^\infty(A), \tilde{a} \in A, \tilde{b} \in B$ and $\xi \in l^2(A \times B \times A, L^2(P))$. On the other hand, the crossed product algebra of N by the (outer) action δ_K of $l^\infty(B) \rtimes A$ is generated by $\delta_K(K) \vee 1 \otimes (l^\infty(B) \rtimes A)\hat{} = \delta_K(K) \vee 1 \otimes (B \ltimes l^\infty(A))$. (See [**5**].) It is easy to see that $\delta_K(\pi(p)) = \tilde{\pi}(p), \delta_K(\lambda'_a) = \tilde{\lambda}_a, 1 \otimes \lambda_b = \tilde{\lambda}_b$, and $1 \otimes f = \tilde{\pi}(f)$. Therefore, we are done. □

(2) Let $M \supset N$ be an irreducible inclusion of type II_1 factors satisfying $[M : N] < \infty$ and depth 2, and G be a finite group with an outer action γ on both M and N. Moreover, suppose that $(M \rtimes G) \cap N' = \mathbf{C}$. (This condition is equivalent to strong outerness of the action γ for $M \supset N$.) Then we have the depth 2 inclusion $(M \otimes l^\infty(G)) \rtimes G \supset N \rtimes G$. In fact, this inclusion is contained in the commuting co-commutig square $((M \otimes l^\infty(G)) \rtimes G, (N \otimes l^\infty(G)) \rtimes G, M \rtimes G, N \rtimes G)$. Similar argument as in Proposition 3.1 implies that the assumption in Theorem 2.1 for the inclusions $(M \otimes l^\infty(G)) \rtimes G \supset (N \otimes l^\infty(G)) \rtimes G (\cong M \supset N$ by Takesaki duality) and $(N \otimes l^\infty(G)) \rtimes G \supset N \rtimes G$ holds, hence we have the conclusion. (Cf. the orbifold construction [**3**, **7**] and also [**2**].)

References

[1] S. Baaj and G. Skandalis, *Unitaries multiplicatifs et dualité pour les produits croisés de C^*-algèbres,* Ann. Sci. École Norm. Sup., **26** (1993), 425-488.

[2] J. De Cannière, *Produit croisé d'une algèbre de Kac par un groupe localement compact,* Bull. Soc. Math. France, **107** (1979), 337-372.

[3] M. Choda and H. Kosaki, *Strongly outer actions for an inclusion of factors*, J. Funct. Anal., **122** (1994), 315-332.

[4] M.-C. David, *Paragroupe d'Adrian Ocneanu et algèbre de Kac*, to appear in Pacific J. Math.

[5] M. Enock, *Produit croisé d'une algèbre de von Neumann par une algèbre de Kac,* J. Funct. Anal., **26** (1977), 16-47.

[6] M. Enock and J.-M. Schwartz, *Kac Algebras and Duality of Locally Compact Groups,* Springer-Verlag, 1992.

[7] D. E. Evans and Y. Kawahigashi, *Orbifold subfactors from Hecke algebras*, Comm. Math. Phys., **165** (1994), 445-484.

[8] F. Goodman, de la Harpe and V.F.R. Jones, *Coxeter Dynkin Diagrams and Towers of Algebras,* Springer-Verlag, 1989.

[9] D. Guido and R. Longo, *Relativistic invariance and charge conjugation in quantum field theory,* Comm. Math. Phys., **148** (1992), 521-551.

[10] M. Izumi, *Applications of fusion rules to classification of subfactors,* Publ. RIMS, Kyoto Univ., **27** (1991), 953-994.

[11] V.F.R. Jones, *Index of subfactors,* Invent. Math., **72** (1983), 1-25.

[12] H. Kosaki, *Automorphisms in the irreducible decomposition of sectors,* in "Quantum and Non-Commutative Analysis", Kluwer Academic Publishers, 1993.

[13] H. Kosaki and T. Sano, *Non-splitting inclusions of factors of type* III_0, preprint.

[14] R. Longo, *Index of subfactors and statistics of quantum fields* I, Comm. Math. Phys., **126** (1989), 217-247.

[15] ———, *Index of subfactors and statistics of quantum fields* II, Comm. Math. Phys., **130** (1990), 285-309.

[16] ———, *Minimal index and braided subfactors,* J. Funct. Anal., **109** (1992), 98-112.

[17] ———, *A duality for Hopf algebras and for subfactors* I, Comm. Math. Phys., **159**

(1994), 133-150.

[18] S. Majid, *Physics for algebraists: non-commutative and non-cocommutative Hopf algebras by a bicrossproduct construction,* J. Algebra, **130** (1990), 17-64.

[19] ———, *Hopf-von Neumann algebra bicrossproducts, Kac algebra bicrossproducts, and the classical Yang Baxter equations,* J. Funct. Anal., **95** (1991), 291-319.

[20] A. Ocneanu, *Quantized groups, string algebras, and Galois theory,* in "Operator Algebras and Applications Vol. II", Cambridge Univ. Press, 1988.

[21] ———, *Quantum Symmetry, Differential Geometry of Finite Graphs, and Classification of Subfactors,* Univ. of Tokyo Seminary Notes, 1991.

[22] S. Okamoto, *Private communication.*

[23] M. Pimsner and S. Popa, *Entropy and index of subfactors,* Ann. Sci. École Norm. Sup., **19** (1986), 57-106.

[24] ———, *Iterating the basic construction,* Trans. Amer. Math. Soc., **310** (1988), 127-133.

[25] S. Popa, *Classification of subfactors: reduction to commuting squares,* Invent. Math., **101** (1990), 19-43.

[26] ———, *Classification of amenable subfactors of type* II, Acta Math., **172** (1994), 163-256.

[27] T. Sano and Y. Watatani, *Angles between two subfactors,* J. Operator Th., **32** (1994), 209-241.

[28] W. Szymański, *Finite index subfactors and Hopf algebra crossed products,* Proc. Amer. Math. Soc., **416** (1994), 519-528.

[29] A. Wasserman, *Coactions and Yang-Baxter equations for ergodic actions and subfactors,* in "Operator Algebras and Application Vol. II", Warwick, London Math. Soc. Lecture Notes Series 136, Cambridge Univ. Press, 1988.

[30] Y. Watatani, *Lattices of intermediate subfactors,* preprint.

[31] Y. Watatani and J. Wierzbicki, *Commuting squares and relative entropy for two subfactors,* J. Funct. Anal., **133** (1995), 329-341.

[32] J. Wierzbicki, *An estimation of the depth from an intermediate subfactor,* Publ. RIMS, Kyoto Univ., **30** (1994), 1139-1144.

[33] S. Yamagami, *Vector bundles and bimodules,* in "Quantum and Non-Commutative Analysis", Kluwer Academic Publishers, 1993.

[34] T. Yamanouchi, *Bicrossproduct Kac algebras, bicrossproduct groups and von Neumann algebras of Takesaki's type,* Math. Scand., **71** (1992), 252-260.

[35] ———, *The intrinsic group of Majid's bicrossproduct Kac algebras,* Pacific J. Math., **159** (1993), 185-199.

[36] ———, *An example of Kac algebra actions on von Neumann algebras,* Proc. Amer. Math. Soc., **119** (1993), 503-511.

Received June 23, 1993 and revised August 25,1994.

Faculty of Science
Yamagata University
Yamagata, 990, Japan

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

SECOND ORDER ORDINARY DIFFERENTIAL EQUATIONS WITH FULLY NONLINEAR TWO POINT BOUNDARY CONDITIONS

H.B. THOMPSON

We establish existence results for two point boundary value problems for second order ordinary differential equations of the form $y'' = f(x, y, y')$, $x \in [0,1]$, where f is continuous and there exist lower and upper solutions. First we consider boundary conditions of the form $G((y(0), y(1)); (y'(0), y'(1))) = 0$, where G is continuous and fully nonlinear. We introduce compatibility conditions between G and the lower and upper solutions. Assuming these compatibility conditions hold and, in addition, f satisfies assumptions guarenteeing a'priori bounds on the derivatives of solutions we show that solutions exist. In the case the lower and upper solutions are constants one of our results is closely related to a result of Gaines and Mawhin. Secondly we consider boundary conditions of the form $(y(i), y'(i)) \in \mathcal{J}(i)$, $i = 0, 1$ where the $\mathcal{J}(i)$ are closed connected subsets of the plane. We introduce various compatibility type conditions relating the $\mathcal{J}(i)$ and the lower and upper solutions and show each is sufficient to construct a compatible G which defines these boundary conditions. Thus our existence results apply. Almost all the standard boundary conditions considered in the literature assuming upper and lower solutions are, or can be, defined by compatible G and their associated existence results follow from ours; in many cases we can improve these results by deleting some of their assumptions.

1. Introduction.

In this paper we consider two point boundary value problems for second order ordinary differential equations of the form

$$y'' = f(x, y, y'), \text{ for all } x \in [0,1], \tag{1.1}$$

where $f : [0,1] \times \mathbb{R}^2 \to \mathbb{R}$ continuous. By a solution of (1.1) we mean a twice continuously differentiable function y satisfying (1.1) everywhere. The first class of boundary conditions we will consider are of the form

$$0 = G((y(0), y(1)); (y'(0), y'(1))), \tag{1.2}$$

where $G = (g^0, g^1), g^i : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, $i = 1, 2$ are continuous and fully nonlinear. We will refer to boundary conditions of the form (1.2) as fully nonlinear boundary conditions. The second class of boundary conditions we will consider are of the form

$$(y(i), y'(i)) \in \mathcal{J}(i) \text{ for } i = 0, 1, \tag{1.3}$$

where $\mathcal{J}(i)$ are continuua. We will refer to boundary conditions of the form (1.3) as boundary set conditions.

We always assume that lower and upper solutions $\alpha \leq \beta$, respectively, exist for (1.1) (see Definition 1 below).

In paragraph 2 we introduce some notation and definitions.

In paragraph 3 we introduce the central notion of compatibility of the boundary conditions G with the lower and upper solutions. In the literature when lower and upper solutions are assumed to exist and the Picard, Neumann or Periodic boundary conditions are considered the assumptions usually made are equivalent to compatibility. We show by simple examples that if the boundary conditions are not compatible with the lower and upper solutions, then solutions need not exist.

In paragraph 4, we present our main existence results. If the boundary conditions G are compatible with α and β and f satisfies additional assumptions guarenteeing a'priori bounds for y' for solutions y of (1.1), then there exist solutions y of (1.1) and (1.2) satisfying $\alpha \leq y \leq \beta$ on $[0, 1]$.

In paragraph 5 we briefly describe the results of Gaines and Mawhin [**16**] and show their relationship to ours.

In paragraph 6 we consider problem (1.1) and (1.3). We introduce two types of compatibility of the boundary sets $\mathcal{J}(i)$, $i = 0, 1$ with the lower and upper solutions. These are satisfied by the usual boundary sets conditions considered in the literature. Given compatible boundary sets $\mathcal{J}(i)$, $1 = 0, 1$ we show that there exists compatible G such that (1.2) implies (1.3). Thus our existence results apply to such boundary set conditions.

Lower and upper solutions exist for (1.1) if f satisfies suitable monotonicity and or growth conditions (see Ako [**2**, **3**], Baxley [**6**], Gaines [**15**], Gaines and Mawhin [**16**], Palamides [**30**], and Jackson and Palamides [**22**]).

A'priori bounds on y' follow if, for example, f satisfies either the Bernstein–Nagumo growth condition with respect to y' (see Bernstein [**12**], Nagumo [**27**]) or it's one sided generalisations (see Baxley [**6**]) or the Scorza Dragoni–Zwirner growth condition with respect to (x, y, y') or the Nagumo–Knobloch–Ako–Schmitt condition (see Ako [**2**], Nagumo [**28**, **29**], Knobloch [**23**, **24**], Schmitt [**31**]) or a Lyapunov condition (see George and Sutton [**18**]).

To prove existence of solutions we modify the differential equation and turn the modified equation into an integral equation, couple it with the two

equations defining the boundary conditions and regard the resultant as an operator equation defined on $C^1([0,1]) \times \mathbb{R}^2$, where the boundary values are regarded as independent variables with values in $\mathbb{R}^2$. We use Schauder degree theory employing homotopies and the reduction theorem to show that the Schauder degree of the operator is the Brouwer degree of a suitable mapping associated with the compatible G on the domain which contains the boundary values.

Methods used in the literature to establish existence results include shooting with initial values, shooting with boundary values, the Schauder fixed point theorem, and Schauder degree theory. Often these methods are applied to a modified problem whose solutions are solutions of the unmodified problem. Some variants of shooting with initial values use the Kneser–Hukuhara continuum theorem and/or Ważewski's retract method and their refinements. Initial value shooting has been commonly used to prove existence for fully nonlinear boundary conditions of the form (1.2) and has been the only method used for boundary set conditions of the form (1.3).

Gaines and Mawhin [**16**] and Guenther Granas and Lee [**19**] give very general existence theorems via coincidence degree, respectively topological transversality, provided associated one parameter families of boundary value problems have no solutions on the boundary of a suitable domain in a suitable function space. These one parameter families are used to construct homotopies. Gaines and Mawhin and Granas Guenther and Lee go on to show that a substantial number of the earlier existence results follow from their general theorems. Also they go on to give many and substantial new results and a coherent framework for viewing old and new results. Gaines and Mawhin [**16**], Theorem V.34 establish an existence result for systems of equations which, in the special case of a single equation, is closely related to the special case of our result above of constant α and β. They apply their result [**16**], Theorem V.37 to a single equation with non constant α and β and a restricted class of G using a modification argument, modifying G. The interdependence required between the boundary conditions and the lower and upper solutions to guarentee existence of solutions is not clear from their work.

In a forth coming paper we extend our result to systems including [**16**], Theorem V.34 as a special case. Also, Gaines and Mawhin [**16**] discuss a'priori bounding of solutions from a geometric perspective giving a new insight into the role of many of these conditions.

Ako was one of the first authors to obtain existence results for nonlinear boundary conditions. He used shooting with the boundary values of minimal solutions as he did not assume uniqueness for the Picard problem; $(y'(0), y'(1))$ is a continuous function of $(y(0), y(1))$ if solutions of the Picard

problem are unique.

Existence results for boundary set conditions assuming lower and upper solutions have been obtained by a number of authors (see, for example, Jackson and Klassen [**21**] and Bebernes and Fraker [**9**], and their references).

The literature on problem (1.1) and (1.2) is vast and for further information we refer the interested reader to the excellent monographs by Bailey, Waltman and Shampine [**4**], Bernfeld and Lakshmikantham [**11**], Gaines and Mawhin [**16**], Guenther, Granas and Lee [**19**], Hartman [**20**], and Mawhin [**26**] and their references.

The contributions this work makes are twofold. First we introduce the compatibility conditions. These conditions are concrete conditions involving the given data which can be easily checked and are satisfied by just about every concrete existence result in the literature. They permit the construction of the one parameter families of boundary value problems used to construct the homotopies in an appropiate function space; both the homotopies and the function space are unusual and clearly demonstrate the role of the compatibility conditions.

Second, we show that the boundary set conditions of the form (1.3) usually considered in the literature are special cases of fully nonlinear boundary conditions.

Most existence results in the literature for (1.1) together with (1.2) or (1.3) which assume lower and upper solutions exist follow as a corollary to our results. In many cases our results can be used to significantly improve upon these results. This is especially true for results concerning fully nonlinear boundary conditions. Some results in the literature concerned both with linear and with nonlinear boundary conditions which assume growth conditions but do not assume explicitly the existence of lower and upper solutions can be obtained from ours by constructing lower and upper solutions compatible with the boundary conditions. In such cases the construction of the lower and upper solutions and the verification of compatibility is usually easier than the given direct proofs of existence. This is true, for example, of some of the results in Baxley [**6**]. Also the central notion of compatibility extends to Carathéodory f with α and β having absolutely continuous first derivatives, to systems with lower and upper solutions, to single equations and systems with lower and upper solutions replaced by other surfaces a'priori bounding solutions. We will discuss these extensions of our ideas and further applications of our results and their extensions in forthcoming papers.

2. Background Notation and Definitions.

In order to state our results we need some notation.

We denote the closure of a set T by $\bar{T}$ and its boundary by ∂T. As usual, $C^m(A;B)$ denotes the space of m times continuously differentiable functions from A to B endowed with the maximum norm. In the case of continuous functions we abreviate this to $C(A;B)$. In the case $B=\mathbb{R}$ we omitt the B. If A is a bounded open subset of $\mathbb{R}^n$, $p\in\mathbb{R}^n$, $f\in C(\bar{A};\mathbb{R}^n)$ and $p\notin f(\partial A)$ we denote the Brouwer degree of f on A at p by $d(f,A,p)$. It is common in the proofs of existence of solutions of two point boundary value problems for (1.1) to modify f. We will do this making use of the following functions (see [**33**]).

If $c\le d$ are given let $\pi:\mathbb{R}\to[c,d]$ be the retraction given by

$$\pi(y,c,d)=\max\{\min\{d,y\},c\}. \tag{2.1}$$

For each $\epsilon>0$, let $K\in C(\mathbb{R}\times(0,\infty);[-1,1])$ satisfy

1. $K(\cdot,\epsilon)$ is an odd function,
2. $K(t,\epsilon)=0$ iff $t=0$ and
3. $K(t,\epsilon)=1$ for all $t\ge\epsilon$.

If $c\le d$ and $\epsilon>0$ are given, let $T\in C(\mathbb{R})$ be given by

$$T(y,c,d,\epsilon)=K(y-\pi(y,c,d),\epsilon). \tag{2.2}$$

Let

$$\mathcal{Q}(x,t)=\begin{cases}(1-x)t, & \text{for } 0\le t\le x\le 1\\ (1-t)x, & \text{for } 0\le x\le t\le 1,\end{cases}$$

and $w(y_0,y_1)(x)=y_0(1-x)+y_1x$. Let $X=C^1([0,1])\times\mathbb{R}^2$ with the usual product norm. Define $\mathcal{C}:C([0,1])\to C^1([0,1])$ by

$$\mathcal{C}(\phi)(x)=-\int_0^1\mathcal{Q}(x,t)\phi(t)dt,$$

for all $\phi\in C([0,1])$ and $x\in[0,1]$. Clearly $\mathcal{C}$ is completely continuous.

Definition 1. We call α (β) a lower (upper) solution for (1.1) if α (β) $\in C^2([0,1])$, and

$$\begin{aligned}&\alpha''(x)\ge f(x,\alpha(x),\alpha'(x)),\ \text{for all } x\in[0,1]\\ &(\beta''(x)\le f(x,\beta(x),\beta'(x)),\ \text{for all } x\in[0,1]).\end{aligned} \tag{2.3}$$

If the inequality in (2.3) is strict then we call α (β) a strict lower (upper) solution for (1.1). As mentioned earlier we assume that $\alpha\le\beta$ and set

$\beta_M = \max\{\beta(x) : x \in [0,1]\}$ and $\alpha_m = \min\{\alpha(x) : x \in [0,1]\}$. We will call the pair non–degenerate if $\Delta = (\alpha(0), \beta(0)) \times (\alpha(1), \beta(1))$ is nonempty. We set

$$\bar{\omega} = \{(x,y) \in [0,1] \times \mathbb{R} : \alpha(x) \leq y \leq \beta(x)\}. \tag{2.4}$$

We will discuss the degenerate case Δ empty later.

Lower solutions are used with maximum principle arguments to obtain a'priori bounds on solutions. For a discussion of the relationship between them and subfunctions and some indication of how the smoothness conditions can be relaxed see Thompson [**35**]. As mentioned earlier our central idea leads to existence results for those f for which there are a'priori bounds on y' for solutions y satisfying $\alpha \leq y \leq \beta$. There are two well known conditions which we employ in our existence results either of which guarentee a'priori bounds on y' for solutions. The first is the Bernstein–Nagumo condition.

Definition 2. Let $\alpha \leq \beta$ be lower and upper solutions for (1.1) on $[0,1]$. We say f satisfies the Bernstein–Nagumo condition if there exists $h \in C([0,\infty);(0,\infty))$ and $N > 0$ such that

$$|f(x,y,p)| \leq h(|p|), \text{ for all } (x,y) \in [0,1] \times [\alpha(x), \beta(x)] \text{ and} \tag{2.5}$$

$$\int_{\sigma}^{N} \frac{s\,ds}{h(s)} > \beta_M - \alpha_m \tag{2.6}$$

where $\sigma = \max\{|\beta(1) - \alpha(0)|, |\beta(0) - \alpha(1)|\}$. We say f satisfies the strengthened Bernstein–Nagumo condition if (2.6) is replaced by

$$\int^{\infty} \frac{s\,ds}{h(s)} = \infty. \tag{2.7}$$

The second condition for guarenteeing a'priori bounds on y' for solutions we call the Nagumo–Knobloch–Schmitt condition.

Definition 3. Let $\alpha \leq \beta$ be lower and upper solutions for (1.1) on $[0,1]$. We say f satisfies the Nagumo–Knobloch–Schmitt condition relative to α and β if there exists $\Phi < \Upsilon \in C^1([0,1] \times \mathbb{R})$ such that

$$\begin{aligned} &f(x,y,\Phi(x,y)) \geq \Phi_x(x,y) + \Phi_y(x,y)\Phi(x,y) \text{ and} \\ &f(x,y,\Upsilon(x,y)) \leq \Upsilon_x(x,y) + \Upsilon_y(x,y)\Upsilon(x,y), \end{aligned} \tag{2.8}$$

for all $(x,y) \in \bar{\omega}$.

See Gaines and Mawhin [**16**] for some discussion of the relationship between these conditions.

3. Nonlinear Boundary Conditions and Compatibility.

Definition 4. We call the vector field $\Psi = (\psi^0, \psi^1) \in C(\bar{\Delta}; \mathbb{R}^2)$ strongly inwardly pointing on Δ if for all $(C, D) \in \partial\Delta$

$$\begin{aligned} &\psi^0(\alpha(0), D) > \alpha'(0),\ \psi^0(\beta(0), D) < \beta'(0) \text{ and} \\ &\psi^1(C, \alpha(1)) < \alpha'(1),\ \psi^1(C, \beta(1)) > \beta'(1). \end{aligned} \tag{3.1}$$

We call Ψ inwardly pointing if the strict inequalities are replaced by weak inequalities.

Definition 5. Let $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$. We say G is strongly compatible with α and β if for all strongly inwardly pointing Ψ on $\bar{\Delta}$

$$\mathcal{G}(C, D) \neq 0 \text{ for all } (C, D) \in \partial\Delta \text{ and} \tag{3.2}$$

$$d(\mathcal{G}, \Delta, 0) \neq 0, \tag{3.3}$$

where

$$\mathcal{G}(C, D) = G((C, D); \Psi(C, D)) \text{ for all } (C, D) \in \bar{\Delta}. \tag{3.4}$$

We say G is compatible with α and β if there is a sequence $G_i \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$ strongly compatible with α and β and converging uniformly to G on compact subsets of $\bar{\Delta} \times \mathbb{R}^2$.

In what follows where there is a strongly inwardly pointing vector field clearly defined from the context $\mathcal{G}$ will denote the vector field defined by (3.4).

Remark 6. If G is (strongly) compatible with α and β, then the Brouwer degree (3.3) is independent of the strongly inwardly pointing vector field Ψ. To see this for strongly compatible G let Ψ_i, $i = 1, 2$ be two such vector fields. Then setting $\Psi(C, D, \theta) = \theta\Psi_1(C, D) + (1 - \theta)\Psi_2(C, D)$ and $\mathcal{H}(C, D, \theta) = G((C, D); \Psi(C, D, \theta))$ on $\bar{\Delta} \times [0, 1]$, $\mathcal{H}$ is a homotopy for the Brouwer degree (3.3).

It is not difficult to see that strongly inwardly pointing vector fields always exist. Moreover, if (3.2) holds for all inwardly pointing vector fields we may choose

$$\begin{aligned} \psi^0(C, D) &= \beta'(0)\frac{C - \alpha(0)}{\beta(0) - \alpha(0)} + \alpha'(0)\frac{\beta(0) - C}{\beta(0) - \alpha(0)} \text{ and} \\ \psi^1(C, D) &= \beta'(1)\frac{D - \alpha(1)}{\beta(1) - \alpha(1)} + \alpha'(1)\frac{\beta(1) - D}{\beta(1) - \alpha(1)}, \end{aligned}$$

when computing the Brouwer degree (3.3).

For the Picard (also called Dirichlet) boundary conditions

$$\begin{aligned} &g^0((y(0),y(1));(y'(0),y'(1))) = y(0) - A = 0 \text{ and} \\ &g^1((y(0),y(1));(y'(0),y'(1))) = y(1) - B = 0, \end{aligned} \tag{3.5}$$

while for the Neumann boundary conditions

$$\begin{aligned} &g^0((y(0),y(1));(y'(0),y'(1))) = y'(0) - A = 0 \text{ and} \\ &g^1((y(0),y(1));(y'(0),y'(1))) = y'(1) - B = 0, \end{aligned} \tag{3.6}$$

and for the Periodic boundary conditions

$$\begin{aligned} &g^0((y(0),y(1));(y'(0),y'(1))) = y(0) - y(1) = 0 \text{ and} \\ &g^1((y(0),y(1));(y'(0),y'(1))) = y'(1) - y'(0) = 0. \end{aligned} \tag{3.7}$$

For these boundary conditions the compatibility conditions become the familiar ones usually assumed in the presence of lower and upper solutions; that is,

$$\alpha(0) \leq A \leq \beta(0),\ \alpha(1) \leq B \leq \beta(1), \tag{3.8}$$

$$\alpha'(0) \geq A,\ \beta'(0) \leq A,\ \alpha(1) \leq B,\ \beta(1) \geq B, \tag{3.9}$$

and

$$\alpha(0) = \alpha(1),\ \beta(0) = \beta(1),\ \alpha'(0) \geq \alpha'(1),\ \beta'(0) \leq \beta'(1), \tag{3.10}$$

respectively. We prove this in the case of periodic boundary conditions.

Lemma 8. *The periodic boundary conditions are (strongly) compatible iff* (3.10) *holds.*

Proof. Let Ψ be an strongly inwardly pointing vector field on $\bar{\Delta}$.

Assume that (3.10) is satisfied, let $G = (g^0, g^1)$ be given by (3.7) and $(C, D) \in \partial\Delta$. If $C = D = \alpha(0)$ then $\mathcal{G}^1 = \psi^1(C, D) - \psi^0(C, D) < \alpha'(1) - \alpha'(0) \leq 0$. If $C = \alpha(0) < D$ then $\mathcal{G}^0 = C - D < 0$. Similar inequalities hold for the other cases $(C, D) \in \partial\Delta$. Thus $\mathcal{G} \neq 0$ for $(C, D) \in \partial\Delta$. Let $\gamma(x) = (\alpha(x)+\beta(x))/2$, $\mathcal{H}(C, D, \theta) = (1-2\theta)\mathcal{G}(C, D)+2\theta(\mathcal{G}^0(C, D), D-\gamma(1)/2)$, for $\theta \in [0, 1/2]$ and $\mathcal{H}(C, D, \theta) = (2 - 2\theta)(\mathcal{G}^0(C, D), D - \gamma(1)/2) + (2\theta - 1)(C - \gamma(0)/2, D - \gamma(1)/2)$, for $\theta \in [1/2, 1]$. Since $\mathcal{H}$ is a homotopy for Brouwer degree $d(\mathcal{G}(\cdot), \Delta, 0) = d(\mathcal{H}(\cdot, 0), \Delta, 0) = d(\mathcal{H}(\cdot, 1), \Delta, 0) = 1 \neq 0$. Thus G is strongly compatible and hence compatible.

Assume now that G is given by (3.7) and that G is strongly compatible with α and β. We show that (3.10) is satisfied.

We may assume that $[\alpha(0),\beta(0)] \cap [\alpha(1),\beta(1)] \neq \emptyset$ otherwise for any strongly inwardly pointing vector field Ψ setting $\mathcal{G} = G((C,D)); \Psi(C,D)$, $\mathcal{G}(C,D) \neq 0$ for all $(C,D) \in \bar{\Delta}$ and $d(\mathcal{G},\Delta,0) = 0$, a contradiction. Assume that $\alpha(0) \neq \alpha(1)$ and in particular that $\alpha(0) < \alpha(1) = D = C \leq \beta(0)$. Thus we may choose a strongly inwardly pointing vector field Ψ as follows. First choose ψ^1 then define ψ^0 such that Ψ is strongly inwardly pointing and $\psi^0(C,D) < \min\{\beta'(0), \psi^1(C,D)\}$ for all $C > \gamma(0)/2$. Thus $\mathcal{G}(C,D) \neq 0$ for all $(C,D) \in \bar{\Delta}$ and again $d(\mathcal{G},\Delta,0) = 0$, a contradiction. The other cases that $\alpha(0) \neq \alpha(1)$ and $\beta(0) \neq \beta(1)$ follow by similar arguments.

Assume now that $\alpha'(1) > \alpha'(0)$. Define a strongly inwardly pointing vector field Ψ as follows. Let $\alpha'(1) > \psi^1(C,\alpha(1)) > \alpha'(0)$ for all $C \in [\alpha(0),\beta(0)]$ and $\psi^1(C,\beta(1)) > \beta'(1)$ for all $C \in [\alpha(0),\beta(0)]$. Using the Teitze extension theorem (see Dugundji [**13**]) we may extend ψ^1 as a continuous function to $\bar{\Delta}$. By continuity and compactness we may choose $\epsilon > 0$ such that $\psi^1(C,\alpha(1)) - \epsilon > \alpha'(0)$ for all $C \in [\alpha(0),\beta(0)]$. Let $\Xi = (\xi^0,\xi^1)$ be a strongly inwardly pointing vector field. Set $\psi^0(C,D) = \min\{\xi^0(C,D), \psi^1(D,C) - \epsilon\}$. Thus Ψ is strongly inwardly pointing and $\mathcal{G}^1 \neq 0$ on $\bar{\Delta}$, where $\mathcal{G}$ is defined by (3.4). Thus $d(\mathcal{G}(\cdot),\Delta,0) = 0$, a contradiction. The other case $\beta'(0) > \beta'(1)$ leads similarly to a contradiction. Thus (3.10) holds. If G is compatible then there exist strongly compatible G_i and the result follows. □

The proof that the Picard, respectively Neumann, boundary conditions are compatible iff (3.8), respectively (3.9), is satisfied is simpler, follows similar lines and hence is omitted.

Remark 8. In the following two examples G is not strongly compatible since condition (3.3) fails. In the first there are solutions of (1.1) and (1.2) lying between α and β and in the second there are no such solutions.

We choose $R > 0$ and

$$G \in C([-R,R]^2 \times \mathbb{R}^2; \mathbb{R}^2)$$

such that

$$G((C,D);(P,Q)) = G((C,D);(S,T))$$

for all

$$((C,D);(P,Q)),((C,D);(S,T)) \in [-R,R]^2 \times \mathbb{R}^2, \quad \mathcal{G} \neq 0$$

on $\partial(-R,R)^2$ and $d(\mathcal{G},(-R,R)^2,0) = 0$; $\mathcal{G}$ is independent of the choice of strongly inwardly pointing vector field Ψ. Let f be identically zero and $-\alpha = R = \beta$.

For the first example we choose G so that $G((C,D);(P,Q)) = 0$, for some $(C,D) \in (-R,R)^2$.

For the second example we choose G so that $G \neq 0$ for any $(C, D) \in (-R, R)^2$.

4. Existence of Solutions.

Theorem 1. *Assume that there exist non-degenerate lower and upper solutions $\alpha \leq \beta$ for (1.1), that f satisfies the Bernstein–Nagumo condition and that $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$ is compatible with α and β, then problem (1.1) and (1.2) has a solution y lying between α and β.*

Proof. Assume first that G is strongly compatible with α and β.

We modify the differential equation for y not between α and β to obtain a second pair of lower and upper solutions. We reformulate the problem as a coupled system of integral and boundary condition equations and show that a solution of the modified problem lies in the region where f is unmodified and hence is the required solution. We use Schauder degree theory to prove existence for the modified problem and compute the degree using a homotopy; the modification is chosen to facilitate the construction of a suitable homotopy.

Choose $L, \epsilon > 0$ such that

$$\int_\sigma^L \frac{s\,ds}{h(s) + \epsilon} > \beta_M - \alpha_m + 2\epsilon \tag{4.1}$$

where $L > \max\{|\alpha'(x)|, |\beta'(x)| : x \in [0, 1]\}$. Let

$$\begin{aligned} j(x, y, p) &= f(x, \pi(y, \alpha(x), \beta(x)), \pi(p - L, L)), \text{ and} \\ k(x, y, p) &= (1 - |T(y, \alpha(x), \beta(x), \epsilon)|) j(x, y, p) + \\ &\quad T(y, \alpha(x), \beta(x), \epsilon)(|j(x, y, p)| + \epsilon), \end{aligned}$$

where π and T are given by (2.1) and (2.2), respectively. Thus k is a bounded continuous function on $[0, 1] \times \mathbb{R}^2$ and satisfies

$$|k(x, y, p)| \leq h(|p|) + \epsilon,$$

for all p with $|p| \leq L$.

Consider

$$y'' = k(x, y, y'), \text{ for all } x \in [0, 1] \tag{4.2}$$

together with (1.2). It suffices to show that problem (4.2) and (1.2) has a solution y satisfying $\alpha \leq y \leq \beta$ and $|y'| \leq L$ on $[0, 1]$ since f and k coincide in this region.

Let

$$\alpha_\epsilon = \alpha_m - \epsilon$$
$$\beta_\epsilon = \beta_M + \epsilon$$

Now

$$\alpha_\epsilon'' = 0 > -(|j(x, \alpha_\epsilon, 0)| + \epsilon)$$
$$= k(x, \alpha_\epsilon, \alpha_\epsilon'), \text{ for all } x \in [0,1]$$

so α_ϵ is a lower solution for (4.2). Similarly β_ϵ is an upper solution for (4.2).

Also, by (4.1),

$$\int_\sigma^L \frac{s\,ds}{h(s)+\epsilon} > \beta_\epsilon - \alpha_\epsilon. \tag{4.3}$$

Suppose that y is a solution of (4.2) and $(y(0), y(1)) \in \bar{\Delta}$. We show that y is a solution of (1.1). We show that $\alpha \le y \le \beta$ on $[0,1]$. Suppose for example that $y(t) < \alpha(t)$ for some $t \in [0,1]$. From the boundary conditions and continuity we may assume that $\alpha - y$ attains its positive maximum at $t \in (0,1)$. Thus $\alpha'(t) = y'(t)$ so that $|y'(t)| < L$ and $\alpha''(t) \le y''(t)$. From the definition of k we have

$$y''(t) = k(t, y(t), y'(t))$$
$$< f(t, \alpha(t), \alpha'(t)) \le \alpha''(t)$$

a contradiction. Similarly $y \le \beta$ on $[0,1]$. Now $|y'| < L$ on $[0,1]$ by the standard argument and y is the required solution.

Let $\Omega_\epsilon = \{y \in C^1([0,1]) : \alpha_\epsilon < y < \beta_\epsilon, |y'| < L, \text{ on } [0,1]\}$ and $\Gamma_\epsilon = \Omega_\epsilon \times \Delta$.

Define $\mathcal{K} : C^1([0,1]) \to C([0,1])$ at $x \in [0,1]$ by

$$\mathcal{K}(\phi)(x) = k(x, \phi(x), \phi'(x)).$$

Define $\mathcal{H} : \bar{\Gamma}_\epsilon \times [0,1] \to X$ by

$$\mathcal{H}(\phi, C, D, \theta) = (\phi + C\mathcal{K}(\phi) - w(C,D), \mathcal{S}(\phi, C, D, \theta))$$

for $\frac{2}{3} \le \theta \le 1$,

$$\mathcal{H}(\phi, C, D, \theta) = (\phi + C3(\theta - 1/3)\mathcal{K}(\phi) - w(C,D), \mathcal{G}(C,D))$$

for $\frac{1}{3} \le \theta \le \frac{2}{3}$, and

$$\mathcal{H}(\phi, C, D, \theta) = (\phi - 3\theta w(C,D) - (1-3\theta)(\alpha_\epsilon + \beta_\epsilon)/2, \mathcal{G}(C,D))$$

for $0 \le \theta \le \frac{1}{3}$, where

$$\mathcal{S}(\phi, C, D, \theta) = G((C,D); (3(\theta - 2/3)(\phi'(0), \phi'(1)) + 3(1-\theta)\Psi(C,D))).$$

Clearly $\mathcal{H}$ is completely continuous. It is easy to see that y is a solution of (4.2) and (1.2) with $(y, y(0), y(1)) \in \bar{\Gamma}_\epsilon$ iff $\mathcal{H}(y, y(0), y(1), 1) = 0$. If there is a solution with $(y, y(0), y(1)) \in \partial\Gamma_\epsilon$ then we are through. Suppose there is no solution in $\partial\Gamma_\epsilon$. We show that $\mathcal{H}$ is a homotopy for Schauder degree on Γ_ϵ at 0. To see this assume there are solutions of $\mathcal{H}(y, C, D, \theta) = 0$ with $\theta \in [0, 1]$ and $(y, C, D) \in \partial\Gamma_\epsilon$. We consider the cases $\theta \in [2/3, 1]$ and $[1/3, 2/3)$; the case $\theta \in [0, 1/3)$ is trivial.

Consider the case $\theta \in [2/3, 1]$. By assumption there is no solution with $\theta = 1$. Assume there is a solution (y, C, D) with $\theta \in [2/3, 1)$. As before $\alpha(x) \leq y(x) \leq \beta(x)$ on $[0, 1], y(0) = C$ and $y(1) = D$.

Assume that $(y(0), y(1)) \in \partial\Delta$. If $y(0) = \alpha(0)$, then $y'(0) \geq \alpha'(0)$. Thus $3(\theta - 2/3)y'(0) + 3(1 - \theta)\psi^0(y(0), y(1)) > \alpha'(0)$ and $\mathcal{S}(y, y(0), y(1), \theta) \neq 0$, a contradiction. Similarly the other cases $(y(0), y(1)) \in \partial\Delta$ lead to a contradiction.

Assume that $y \in \partial\Omega_\epsilon$. Again, by a standard argument, $|y'| < L$ on $[0, 1]$. Assume that $y(t) = \alpha_\epsilon(t)$ for some $t \in [0, 1]$. From the boundary conditions we see that $t \in (0, 1)$ and thus $y'(t) = \alpha'_\epsilon(t) = 0$ while $y''(t) \geq \alpha''_\epsilon(t) = 0$. From the definition of k we have

$$\begin{aligned} y''(t) &= k(t, y(t), y'(t)) \\ &= -(|f(t, \alpha(t), 0)| + \epsilon) < 0, \end{aligned}$$

a contradiction. Similarly the assumption $y(t) = \beta_\epsilon(t)$ for some $t \in [0, 1]$ leads to a contradiction. Thus there are no solutions of $\mathcal{H}(y, C, D, \theta) = 0$ with $\theta \in [2/3, 1]$ and $(y, C, D) \in \partial\Gamma_\epsilon$.

Assume that $\theta \in [1/3, 2/3)$. Since Ψ is strongly inwardly pointing and G is strongly compatible, by (3.2) there are no solutions (y, C, D) with $(C, D) \in \partial\Delta$. The proof that the case $y \in \partial\Omega_\epsilon$ leads to a contradiction is similar to that for $\theta \in [2/3, 1)$.

Thus $\mathcal{H}$ is a homotopy for the Schauder degree and since $\mathcal{H}(\cdot, 0) = (I - c, \mathcal{G})$ where I is the identity on $C^1([0, 1])$ and $c \in \Omega_\epsilon$ is a constant it follows that

$$\begin{aligned} d(\mathcal{H}(\cdot, 1), \Gamma_\epsilon, 0) &= d(\mathcal{H}(\cdot, 0), \Gamma_\epsilon, 0) \\ &= d(\mathcal{G}, \Delta, 0) \neq 0. \end{aligned}$$

Suppose now that G is compatible with α and β. Then there is a sequence $\{G_i\}_{i=1}^\infty$ strongly compatible with α and β and converging uniformly to G on compact subsets of $\mathbb{R}^2 \times \mathbb{R}^2$ to G. Let y_i be the corresponding solutions. By compactness there is a subsequence of the y_i converging in $C^2([0, 1])$ to the desired solution. □

Remark 9. The Bernstein–Nagumo growth condition can be generalised to: There exist $h \in C([0, \infty); (0, \infty))$, $\bar{h} \in C([\alpha_m, \beta_M]; (0, \infty))$ and $r \in$

$C([0,1];(0,\infty))$ such that

$$|f(x,y,p)| \le h(|p|)\bar{h}(y) + r(x), \text{ for all } (x,y) \in [0,1] \times [\alpha(x),\beta(x)] \text{ and}$$
$$\int_\sigma^\infty \frac{s\,ds}{h(s)} > \int_{\alpha_m}^{\beta_M} \bar{h}(s)ds\,s + K\int_0^1 r(x)dx,$$

where $K = \sup\{s/h(s) : s \in [\sigma, L]\}$.

See Scorza Dragoni [**32**], Zwirner [**36**] and Thompson [**34**].

Remark 10. In the case Δ is degenerate we have to modify the result. Suppose, for example, that $\alpha(0) = \beta(0)$. Then we set $\Delta = (\alpha(1), \beta(1))$ and change the other conditions as follows.

The vector field $\Psi \in C(\bar{\Delta})$ is said to be strongly inwardly pointing on Δ if

$$\Psi(\alpha(1)) < \alpha'(1), \Psi(\beta(1)) > \alpha'(1).$$

Let $G \in C(\bar{\Delta} \times \mathbb{R})$ and $\alpha \le \beta$ be lower and upper solutions for (1.1), respectively. We say G is strongly compatible with α and β if for all strongly inwardly pointing Ψ on Δ

$$\begin{aligned} &\mathcal{G}(D) \quad \neq 0 \text{ for all } D \in \partial\Delta \text{ and} \\ &d(\mathcal{G}, \Delta, 0) \neq 0, \end{aligned}$$

where $\mathcal{G}(D) = G(D, \Psi(D))$. We define compatible as before. Theorem 1 and its proof are modified in the obvious way.

Our results do not apply to the case $\alpha(0) = \beta(0)$ and $\alpha(1) = \beta(1)$ since there are no solutions if our compatibility conditions are extended in the natural way.

As mentioned earlier our central idea leads to existence results for those f for which there are a'priori bounds on y' for solutions y satisfying $\alpha \le y \le \beta$. We now discuss the case where f satisfies the Nagumo–Knobloch–Schmitt condition.

Theorem 2. *Assume that there exist nondegenerate lower and upper solutions $\alpha \le \beta$ for* (1.1), *that f satisfies the Nagumo–Knobloch–Schmitt condition, that $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$ is compatible with α and β, that $\alpha'(x) \ge \Phi(x, \alpha(x))$ and $\beta'(x) \le \Upsilon(x, \beta(x))$ on $[0,1]$ and that $G((C,D);(E,F)) = 0$ only if $E \in [\Phi(0,C), \Upsilon(0,C)]$. Then problem* (1.1) *and* (1.2) *has a solution y lying between α and β.*

Proof. Again we modify f. First choose

$$L > \max\{|\Phi(x,y)|, |\Upsilon(x,y)|, |\alpha'(x)|, |\beta'(x)| : (x,y) \in \bar{\omega}\},$$

where $\bar{\omega} = \{(x,y) \in [0,1] \times \mathbb{R} : \alpha(x) \leq y \leq \beta(x), x \in [0,1]\}$. Let

$$l(x,y,p) = \begin{cases} \max\{f(x,y,\Phi(x,y)) + (\Phi(x,y) - p), f(x,y,p)\}, & \text{for } p \leq \Phi(x,y) \\ \min\{f(x,y,\Upsilon(x,y)) + (\Upsilon(x,y) - p), f(x,y,p)\}, & \text{for } p \geq \Upsilon(x,y) \\ f(x,y,p), & \text{otherwise} \end{cases}$$

and

$$m(x,y,p) = l(x,y,\pi(p,-L,L)).$$

Thus α and β are lower and upper solutions for

$$y'' = m(x,y,y') \text{ for all } x \in [0,1]. \tag{4.4}$$

It is easy to see that m satisfies the conditions of Theorem 1 and thus there is a solution y of problem (4.4) and (1.2) satisfying $\alpha \leq y \leq \beta$. To show that this is a solution of our problem it suffices to show that $\Phi(x,y) \leq y' \leq \Upsilon(x,y)$. From the boundary conditions there are no solutions for $y'(0) \notin [\Phi(0,y(0)), \Upsilon(0,y(0))]$. Suppose that $y'(t) < \Phi(t,y(t))$ for some $t \in (0,1]$. By continuity and the definition of L we may choose t and $u \in (0,t)$ such that $-L < y'(x) < \Phi(x,y(x))$ for all $x \in (u,t]$ and $y'(u) = \Phi(u,y(u))$. Now

$$\begin{aligned} (y'(x) - \Phi(x,y(x)))' &= m(x,y(x),y'(x)) \\ &\quad - \Phi_x(x,y(x)) - \Phi_y(x,y(x))\Phi(x,y(x)) \\ &> f(x,y(x),\Phi(x,y(x))) \\ &\quad - \Phi_x(x,y(x)) - \Phi_y(x,y(x))\Phi(x,y(x)) \\ &\geq 0, \end{aligned}$$

a contradiction. Thus $\Phi(x,y) \leq y'$. Similarly the $y' \leq \Upsilon(x,y)$ and the result follows. □

Remark 11. The conditions $G((C,D);(E,F)) = 0$ only if

$$E \in [\Phi(0,C), \Upsilon(0,C)],$$

(2.8) guarentees the solution y satisfies $\Phi(x,y(x)) \leq y'(x) \leq \Upsilon(x,y(x))$. There are other ways to guarentee this as for example in the case of periodic boundary conditions where we may replace the inequality signs in (2.8) by not equals to signs (see for example Schmitt [**31**]).

5. Comparison with Gaines and Mawhin.

Gaines and Mawhin consider problem (1.1) and (1.2), where $f : [0,1]\times\mathbb{R}^{2n} \to \mathbb{R}^n$, $G : \mathbb{R}^{2n} \times \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ and $y, y' \in \mathbb{R}^n$, making the following definitions.

Definition [**16,** Definitions V.2 and V.18]. A set $\tilde{G} \subseteq \mathbb{R}^n$ is called autonomous curvature bounded with respect to (1.1) if for each $y_0 \in \partial\tilde{G}$ there exists $V(y) = V(y_0; y)$ such that $V \in C^2(\mathbb{R}^n)$, $\tilde{G} \subseteq \{y \in \mathbb{R}^n : V(y) < 0\}$, $V(y_0) = 0$ and

$$p^t V_{yy}(y_0)p + V_y(y_0)f(x, y_0, p) > 0$$

for all $p \in \mathbb{R}^n$ satisfying $V_y(y_0)p = 0$ and $x \in (0,1)$.

Theorem [**16**, Theorem V.34]. *Let $\tilde{G}$ be a convex autonomous curvature bounded set relative* (1.1) *such that $0 \in \tilde{G}$ and for each $y_0 \in \partial\tilde{G}$, $V_{yy}(y_0)$ is positive semi-definite. Assume that there exists nondecreasing $h \in C^1([0,\infty);(0,\infty))$ such that*

$$\begin{aligned} &\lim_{t\to\infty} \tfrac{t^2}{h(t)} = \infty \text{ and} \\ &|f(x,y,p)| \le h(|p|) \text{ for all } (x,y) \in G_1, \end{aligned}$$

where $G_1 = [0,1] \times \tilde{G}$. Let $T = \sup\{|y(x)| : x \in [0,1]\}$ and $M \ge 8T$ be chosen such that $\dfrac{t^2}{h(t)} > 4T$ for all $t > M$. Assume that if y is a solution of

$$\begin{aligned} y'' &= \lambda f(x, y, y') \text{ for all } x \in [0,1] \\ 0 &= G((y(0), y(1)); (y'(0), y'(1))) \end{aligned} \tag{5.1}$$

with $\lambda \in (0,1)$, $y(x) \in \bar{\tilde{G}}$ for $x \in [0,1]$ and $|y'(x)| \le M$, then $y(0), y(1) \notin \partial\tilde{G}$. Moreover assume that

$$\begin{aligned} &\mathcal{G}_1(C,D) \ne 0 \text{ for all } (D,C) \in \partial\Delta_0 \\ &d(\mathcal{G}_1, \Delta_0, 0) \ne 0 \end{aligned} \tag{5.2}$$

where $\mathcal{G}_0(C,D) = G((C, D+C); (D,D))$ for all $(C,D) \in \bar{\Delta}_0$ and $\Delta_0 = \{(C,D) \in \mathbb{R}^{2n} : (C, D+C) \in \tilde{G}\}$. Then there exists a solution y of (1.1) *and* (1.2) *with $(x,y) \in G_1$.*

The assumptions on h are standard for systems where it is well known that the Bernstein–Nagumo condition is no longer sufficient to guarentee a'priori bounds on the derivative of solutions. For some more recent variants see Fabry [**14**], for example, and the references cited there.

The essential problems with this remarkable result are concerned with how to extend it to the case $\tilde{G}_1$ is not autonomous but varies with x. What is the

appropriate replacement for Δ_0 bearing in mind the requirements imposed by the method of proof? In such an extension how would one check that (5.1) is satisfied; this is not difficult in the autonomous case? Such an extension is necessary if the result is to be applied to the range of problems currently considered in the literature as we show in a paper forshaddowed above. In the context of systems as opposed to single equations such an extension may not have seemed so important since other technicalities involved in proving such results have hitherto required the autonomous bounding set.

There is a close relationship between conditions (5.2) and the compatibility condition as the following theorem and example demonstrate.

Theorem 3. *Let G be as in* [**16**], *Theorem* V.34,

$$\Psi(C,D) = (D-C, D-C) \text{ for all } (C,D) \in \bar{\Delta} \tag{5.3}$$

where $\Delta = \tilde{G}^2$ and let $\mathcal{G}(C,D)$ be given by (3.4). *Then*

$$d(\mathcal{G}_0, \Delta_0, 0) = d(\mathcal{G}, \Delta, 0).$$

Proof. This follows by the Leray Multiplication Theorem (see Lloyd [**25**, Theorem 2.3.1]) letting the homeomorphism $\iota : \mathbb{R}^{2n} \to \bar{\Delta}$ be given by $\iota(C,D) = (C, D-C)$. □

In the following examples compatibility fails however there exist strongly inwardly pointing vector fields for which (3.2) holds. In the first example there exists solutions and in the second there are no solutions. Moreover the first example highlights a difference between our assumptions and those of Gaines and Mawhin. Let

$$y'' = y, \tag{5.4}$$

$\beta = 1 = -\alpha$. There is a solution y with

$$(y(0), y'(0)) = (-1, a),\ (y(1), y'(1)) = (1, b) \text{ and } a, b > 2.$$

Choose

$$g^0((-1,D);(P,Q)) = -1 \text{ if } P \leq 1 + a/2,\ g^0((1,D);(P,Q)) = 1,$$

$$g^1((C,1);(P,Q)) = 1 \text{ if } Q \leq 1 + b/2,\ g^1((C,-1);(P,Q)) = -1$$

and $g^i((-1,1);(a,b)) = 0$, $i = 0,1$; G can be extended to all of $\bar{\Delta} \times \mathbb{R}^2$ as continuous functions using Teitze's Theorem. It is easy to see from Lemma 14 below that (3.2) holds. It is easy to see that Theorem V.34 applies

and there are solutions. Moreover it is easy to see that there is a strongly inwardly pointing vector field such that (3.2) holds but (3.3) does not hold for all strongly inwardly pointing vector fields.

To produce an example with no solutions is easy. Just let

$$W = \{((y(0), y(1)); (y'(0), y'(1))) : \text{ there is a solution of (5.4) with these values } \}.$$

Set $g^0((C, D); (P, Q)) = 1$ on W. Let $g^0((-1, D); (P, Q)) = -1$ for $P \leq b/2$ and $g^0((1, D); (P, Q)) = 1$ for $P \geq -b/2$, where $b = (\cosh 1 - 1)/(\sinh 1)$. Again G can be extended to all of $\bar{\Delta} \times \mathbb{R}^2$ as continuous functions using Teitze's Theorem. Again, by construction there is a strongly inwardly pointing vector field Ψ such that (3.2) holds and (3.3) does not hold for all strongly inwardly pointing vector fields.

It is easy to see from the above that our compatibility condition is a strengthening of (5.2) and the trade off is that we do not require that (5.1) holds. Apart from the fact that [**16**], Theorem V.34 applies to systems whereas Theorem 1 applies to a single equation this is the essential difference between these results.

For the convenience of the reader and to highlight the difficulty in applying Theorem V.34 to nonconstant α and β we state [**16**], Theorem V.37 which is the result in [**16**] closest to our Theorem 1.

Theorem [**16**, Theorem V.37]. *Assume that there exist strict lower and upper solutions $\alpha < \beta$ for* (1.1), *g^0 is independent of D, Q, nondecreasing in P, g^1 is independent of C, P, nondecreasing in Q and G satisfies*

$$g^0((\beta(0), D); (\beta'(0), Q)) < 0 \,, g^0((\alpha(0), D); (\alpha'(0), Q)) > 0$$
$$g^1((C, \beta(1)); (P, \beta'(1))) > 0 \,, g^1((C, \alpha(1)); (P, \alpha'(1))) < 0.$$

Assume that f satisfies the strengthened Bernstein–Nagumo condition, where $h \in C^1([0, \infty); (0, \infty))$. Then (1.1) *and* (1.2) *has a solution.*

The interested reader is referred to [**16**] and [**19**] for other results of a similar nature.

6. Compatibility for Boundary Set Conditions.

In this section we consider problem (1.1) and (1.3) again assuming that there exist lower and upper solutions $\alpha \leq \beta$, respectively and look for solutions y lying between α and β.

Problems of the form (1.1) together with (1.2) and with (1.3) have been considered by many authors. Shooting methods have been used combined

variously with the maximum principle, with the Jordan separation theorem, the Kneser–Hukuhara continuum theorem and/or the Ważewski retraction theorem. Often these have been refined in the process. See Baxley [**6**], Baxley and Brown [**5**], Bebernes and Fraker [**9**], Bebernes and Wilhelmsen [**7**], Bernfeld and Palamides [**10**], Jackson and Klassen [**21**], Jackson and Palamides [**22**], Palamides [**30**] and their references.

We show that the results of Bebernes and Fraker [**9**] can be derived from our results. In a forthcoming paper we systematically show how all these results involving set valued boundary conditions can be obtained from our results. At first sight this may seem suprising since our results are derived from Schauder degree theory while the others are derived from variants of shooting. The key idea in shooting is a generalised Jordan separation theorem while the key idea in the retract method is that the boundary of the sphere is not a retract of the sphere. Both of these can be derived from degree theory thus a connection of this nature is not suprising but to be expected.

In order to state our results we need some notation (see Bebernes and Fraker [**9**]). For $x \in [0,1]$ let $C(x) = \{(x,y,p) \in \bar{\omega} \times \mathbb{R}\}$, $S_\alpha(x) = \{(x,y,p) \in C(x) : y = \alpha(x)\}$, and $S_\beta(x) = \{(x,y,p) \in C(x) : y = \beta(x)\}$.

Definition 12. We say the pair of sets $\{\mathcal{J}(0), \mathcal{J}(1)\} \subset \mathbb{R}^2$ is strongly compatible, respectively compatible, for (1.1), α and β if there exists $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$ which is strongly compatible, respectively compatible, for (1.1), α and β and such that $G((C,D);(E,F)) \neq 0$ for all $(C,E,D,F) \notin \mathcal{J}(0) \times \mathcal{J}(1)$.

Definition 13. Let $\mathcal{J}(i) \subset \mathbb{R}^2$, $i = 0$ or 1 be a closed connected subset of $[\alpha(i), \beta(i)] \times \mathbb{R}$. We say it is of compatible type 1 if there is $(\alpha(i), u(i)) \in \mathcal{J}(i)$, where $(-1)^i(\alpha'(i) - u(i)) \geq 0$, and there is $(\beta(i), u(i)) \in \mathcal{J}(i)$, where $(-1)^i(u(i) - \beta'(i)) \geq 0$. We say it is of compatible type 2 if for every $p \in \mathbb{R}$ there is $y \in [\alpha(i), \beta(i)]$ such that $(y,p) \in \mathcal{J}(i)$. If it is of compatible type 1 or 2 we say simply it is of compatible type.

Theorem 4. *Let the sets $\mathcal{J}(i) \subset \mathbb{R}^2$ $i = 0,1$ be of compatible type, then the pair $\{\mathcal{J}(0), \mathcal{J}(1)\}$ is compatible for (1.1), α and β.*

We will need the following Lemma in the proof of Theorem 4.

Lemma 14. *Let $M = (m^0, m^1) \in C(\bar{\Delta}; \mathbb{R}^2)$ satisfy $m^0(\alpha(0), D) \leq 0$, $m^0(\beta(0), D) \geq 0$ and $m^1(C, \alpha(1)) \leq 0, m^1(C, \beta(1)) \geq 0$ for all $(C,D) \in \bar{\Delta}$ and $M(C,D) \neq 0$ for all $(C,D) \in \partial\Delta$, then $d(M, \Delta, 0) \neq 0$.*

This follows since $\mathcal{S} = \theta M + (1-\theta)(I-p)$ is a homotopy of M with $I-p$ where I is the identity on $\mathbb{R}^2$ and $p \in \Delta$ is any point.

Proof of Theorem 4. Let $\mathcal{J}(i)$ $i = 0, 1$ be of compatible type. We define $G = (g^0, g^1)$ as follows.

Let $\mathcal{O} = [\alpha(0), \beta(0)] \times \mathbb{R} \setminus \mathcal{J}(0)$. Then $\mathcal{O}$ is a relatively open subset of $[\alpha(0), \beta(0)] \times \mathbb{R}$. Since $\mathcal{J}(0)$ is of compatible type, by the generalised Jordan curve theorem, (see Lloyd [**25**]) $\mathcal{O} = U \cup V \cup W$ where U is the union of all components of $\mathcal{O}$ which intersect $L_0 = \{\alpha(0)\} \times [\alpha'(0), \infty)$, V is the union of all components of $\mathcal{O}$ which intersect $L_1 = \{\beta(0)\} \times (-\infty, \beta'(0)]$ and $W = \mathcal{O} \setminus \{U \cup V\}$. Now $U \neq \emptyset \neq V$. Set

$$g^0((C, D); (E, F)) = \begin{cases} \operatorname{dist}((C, E), \mathcal{J}(0)), & \text{for all } (C, E) \in V \cup W \\ -\operatorname{dist}((C, E), \mathcal{J}(0)), & \text{for all } (C, E) \in U \cup \mathcal{J}(0). \end{cases}$$

It is easy to see that g^0 is continuous, $g^0((\alpha(0), D); (E, F)) \leq 0$ for all $E \geq \alpha'(0)$ and that $g^0((\beta(0), D); (E, F)) \geq 0$ for all $F \geq \beta'(0)$. Similarly we define g^1 using $\mathcal{J}(1)$. To see that G is compatible we set $g_i^0((C, D); (E, F)) = g^0((C, D); (E, F)) + (C - (\alpha(0) + \beta(0))/2)/i)$ and similarly approximate g^1. Let Ψ be a strongly inwardly pointing vector field on $\bar{\Delta}$. Thus $\mathcal{G}_i(C, D) = g_i((C, D); \Psi(C, D))$ satisfies $\mathcal{G}_i^0(\alpha(0), D) < 0$, $\mathcal{G}_i^0(\beta(0), D) > 0$, $\mathcal{G}_i^1(C, \alpha(1)) < 0$ and $\mathcal{G}_i^1(C, \beta(1)) > 0$ for all $(C, D) \in \bar{\Delta}$. The result follows by Lemma 14. □

The next result is an immediate consequence of Theorems 1 and 4.

Corollary 15 [**9**, Theorem 2]. *Assume that there exist lower and upper solutions, $\alpha \leq \beta$, respectively, for* (1.1), *that f satisfies a Bernstein–Nagumo condition and that the sets $\mathcal{J}(i)$, $i = 0, 1$ are of compatible type. Then there is a solution of* (1.1) *and* (1.3) *lying between α and β.*

We show now how the results of Bebernes and Fraker follow from ours.

Corollary 16 [**9**, Theorem 1]. *Assume that there exist lower and upper solutions $\alpha \leq \beta$ for* (1.1) *and that for any $B_2 \geq 0$ and $t_0 \in (0, 1]$ there is $N(B_2) > 0$ such that any solution of* (1.1) *with $|y'(0)| \leq B_2$ and $\alpha(x) \leq y(x) \leq \beta(x)$ for all $x \in [0, t_0]$ satisfies $|y'(x)| \leq N(B_2)$ for all $x \in [0, t_0]$. If $\mathcal{J}(0)$ is compact and of compatible type* 1 *and $\mathcal{J}(1)$ is of compatible type* 2, *then problem* (1.1) *and* (1.3) *has a solution lying between α and β.*

Proof. By compactness there is $B_2 > 0$ such that $(y(0), y'(0)) \in \mathcal{J}(0)$ implies that $|y'(0)| \leq B_2$. By assumption we may choose N such that $|y'| < N$ for all solutions y of (1.1) with $(x, y) \in \bar{\omega}$ on $[0, 1]$ and L such that $L > \max\{|\alpha'(x)|, |\beta'(x)|, N : x \in [0, 1]\}$. Let $j(x, y, p) = f(x, y, \pi(p; -L, L))$ for all $(x, y, p) \in [0, 1] \times \mathbb{R}^2$. Consider

$$y'' = j(x, y, y') \quad \text{for all } x \in [0, 1] \tag{6.1}$$

together with (1.3). Now α and β are lower and upper solutions for (6.1) so by Corollary 15 there is a solution y for this problem between α and β. Assume that $|y'| \geq L$ for some $x \in [0,1]$. Set $t = \min\{x \in [0,1] : |y'(x)| \leq L\}$. By continuity and the choice of L, we have $t > 0$ and $|y'(x)| \leq L$ for all $x \in [0,t]$. Thus y is a solution of (1.1) on $[0,t]$ lying between α and β but $|y'(t)| > N$ a contradiction. Thus $|y'| < L$ on $[0,1]$ and y is the required solution. □

In the above result Bebernes and Fraker assume that there exist strict lower and upper solutions.

The above result can be sharpened as follows.

Definition 17. Set

(6.3)
$$S_0 = \{(y,-N) : \alpha(0) \leq y \leq \beta(0)\} \cup \{(\alpha(0),p) : \alpha'(0) \geq p \geq -N\},$$

(6.4)
$$S_2 = \{(y,N) : \alpha(0) \leq y \leq \beta(0)\} \cup \{(\beta(0),p) : \beta'(0) \leq p \leq N\},$$

(6.5)
$$S_1 = \{(y,-M) : \alpha(1) \leq y \leq \beta(1)\} \cup \{(\beta(1),p) : \beta'(1) \geq p \geq -M\} \text{ and}$$

(6.6)
$$S_3 = \{(y,M) : \alpha(1) \leq y \leq \beta(1)\} \cup \{(\alpha(1),p) : \alpha'(1) \leq p \leq M\}.$$

Corollary 18 [**9**, Theorem 3]. *Assume that there exist lower and upper solutions $\alpha \leq \beta$ for* (1.1) *and that f satisfies a Bernstein–Nagumo condition. Further assume that $\mathcal{J}(i) \subseteq C(i)$, $i = 0,1$ are closed connected sets satisfying $\mathcal{J}(0) \cap \{S_0 \cup S_2\} \neq \emptyset$ and $\mathcal{J}(1) \cap \{S_1 \cup S_3\} \neq \emptyset$, where the S_i are given by* (6.3) *to* (6.6) *and $M = N$ is given by* (2.5). *Then there is a solution y lying between α and β.*

Proof. This follows since either $(\beta(0),u) \in \mathcal{J}(0)$ for some $u \geq \beta'(0)$ or $(y,N) \in \mathcal{J}(0)$ for some $y \in [\alpha(0),\beta(0))$ and we add the straight line segment joining (y,N) to $(\beta(0),N)$ to $\mathcal{J}(0)$. Similarly, either $(\alpha(0),u) \in \mathcal{J}(0)$ for some $u \leq \alpha'(0)$ or $(y,-N) \in \mathcal{J}(0)$ for some $y \in (\alpha(0),\beta(0)]$ and we add the straight line segment joining $(y,-N)$ to $(\alpha(0),-N)$ to $\mathcal{J}(0)$. Similarly we modify $\mathcal{J}(1)$ as above. Thus the modified $\mathcal{J}(i)$ are of compatible type and, by Corollary 15, there exists a solution for (1.1) and (1.3). This solution satisfies $|y'| < N$ and hence is the required solution. □

There is a typographical error in Bebernes and Fraker [**9**], Theorem 3 and a counterexample can be constructed to the result as they have stated it. To explain the error we need the following notation. Let f satisfy the

strengthened Bernstein–Nagumo condition, and $\lambda = \max\{\sigma, |\alpha'(x)|, |\beta'(x)| : x \in [0,1]\}$. Define $N(t)$ by

$$\int_{\lambda}^{N(t)} \frac{s}{\phi(s)} ds = \max\{\beta(u) : u \in [0,t]\} - \min\{\alpha(u) : u \in [0,t]\},$$

and N by

$$N = \min\{N(t) : t \in [0,1]\}. \tag{6.7}$$

Bebernes and Fraker take N from (6.7) and $M = N(1)$. The min should be max in (6.7).

References

[1] K. Ako, *Subfunctions for ordinary differential equations,* J. Fac. Sci. Univ. Tokyo, **I,9** (1965), 17–43.

[2] ———, *Subfunctions for ordinary differential equations* II, Funckialaj Ekvacioj, **10** (1967), 145–162.

[3] ———, *Subfunctions for ordinary differential equations* III, Funckialaj Ekvacioj, **11** (1968), 111–129.

[4] P.B. Bailey, L.F. Shampine and P.E. Waltman, *Nonlinear Two Point Boundary Value Problems,* Academic Press, New York, 1974.

[5] J.V. Baxley and S.E. Brown, *Existence and uniqueness for two–point boundary value problems,* Proc. Roy. Soc. Edinburgh Sect. A, **88** (1981), 219–234.

[6] John V. Baxley, *Existence theorems for nonlinear second order boundary value problems,* J. Differential Equations, **85** (1990), 125–150.

[7] J.W. Bebernes and R. Wilhelmsen, *A technique for solving two dimensional boundary value problems,* SIAM J. Appl. Math., **17** (1969), 1060–1064.

[8] ———, *A general boundary value technique,* J. Differential Equations, **8** (1970), 404–415.

[9] J.W. Bebernes and Ross Fraker, *A 'priori bounds for boundary sets,* Proc. Amer. Math. Soc., **29** (1971), 313–318.

[10] S.R. Bernfeld and P.K. Palamides, *A Topological Method for Vector–Valued and nth–Order Nonlinear Boundary Value Problems,* J. Math. Anal. Appl., **102** (1984), 488–508.

[11] S.R.Bernfeld and V. Lakshmikantham, *An Introduction to Nonlinear Boundary Value problems,* Academic Press, New York, 1974.

[12] S.N. Bernstein, *Sur les équations du calcul des variations,* Ann. Sci. Ecole Norm. Sup., **29** (1912), 438–448.

[13] J. Dugundji, *Topology,* Allyn and Bacon, Boston, 1966.

[14] C. Fabry, *Nagumo Conditions for Systems of Second Order Differential Equations,* J. Math. Anal. Appl., **107** (1985), 132–143.

[15] R.E. Gaines, *A priori bounds and upper and lower solutions for nonlinear second–order boundary value problems,* J. differential Equations, **12** (1972), 291–312.

[16] R.E. Gaines and J.L. Mawhin, *Coincidence Degree and Nonlinear Differential Equations Lecture Notes in Mathematics No. 568,* Springer–Verlag, New York, 1977.

[17] R.E. Gaines and J. Santanilla, *A Coincidence theorem in convex sets with applications to periodic solutions of ordinary differential equations,* Rocky Mountain J. Math., **12** (1982), 669–678.

[18] J.H. George and W.G. Sutton, *Application of Liapunov theory to boundary value problems,* Proc. Amer. Math., **25** (1970), 666–671.

[19] A. Granas, R.B. Guenther and J.W. Lee, *Nonlinear boundary value problems for ordinary differential equations,* Dissertationes Mathematicae, Warsaw, 1985.

[20] P. Hartman, *Ordinary Differential Equations,* Wiley, New York, 1964.

[21] Lloyd K. Jackson and Gene Klassen, *A variation of the topological method of Ważewski,* SIAM J. Appl. Math., **20** (1971), 124–130.

[22] L.K. Jackson and P.K. Palamides, *An existence theorem for a nonlinear two point boundary value problem,* J. Differential Equations, **53** (1984), 43–66.

[23] H.W. Knobloch, *Eine neue Methode zur Approximations periodischer Lösungen nichtlinearer Differentialgleichungen zweiter Ordnung,* Math. Z., **82** (1963), 177–197.

[24] ———, *On the existence of periodic solutions for second order vector differential equations,* J. Differential Equations, **9** (1971), 67–85.

[25] N.G. Lloyd, *Degree Theory* Cambridge University Press, London, 1978.

[26] J.L. Mawhin, *Topological Degree Methods in Nonlinear Boundary Value Problems Regional Conf. Series in Math., No. 40* Amer. math. Soc., Providence R.I., 1979.

[27] M. Nagumo, *Über die Differentialgleichungen* $y'' = f(x, y, y')$, Proc. Phys–Math. Soc. Japan, **19** (1937), 861–866.

[28] ———, $y'' = f(x, y, y')$ *no Kyôkaichi Mondai ni suite* I, Kansû Hôteishiki, **13** (1941), 27–34.

[29] ———, $y'' = f(x, y, y')$ *no Kyôkaichi Mondai ni suite* II, Kansû Hôteishiki, **17** (1942), 37–44.

[30] P.K. Palamides, *A topological method and its application a general boundary value problem,* J. Nonlinear Analysis, Theory, Methods & Applications, **7** (1983), 1101–1114.

[31] K. Schmitt, *Periodic solutions of second order nonlinear differential equations,* Math. Z., **98** (1967), 200–207.

[32] G. Scorza Dragoni, *Intorno a un criterio di esistenza per un problema di valori ai limiti,* Rend. Accad. Naz. Lincei, **(6) 28** (1938), 317–325,

[33] H.B. Thompson, *Existence of solutions for a two point boundary value problem,* Rend. Circ. Mat. Palermo, **(2) XXXV** (1986), 261–275.

[34] ———, *Minimal solutions for two point boundary value problems,* Rend. Circ. Mat. Palermo, **(2) XXXVII** (1988), 261–281.

[35] ———, *Differentiability properties of subfunctions for second order ordinary differential equations,* Pacific J. Math., **140** (1989), 181–207.

[36] G. Zwirner, *Un criterio di esistenza per un problema di valori al contorno per equazioni differenziali del secondo ordine,* Rend. Sem. Mat. Univ. Padova, **10** (1939), 55–64.

Received January 1, 1993. This work is partially supported by a grant from Sonderforschungsbereich 72, the University of Bonn. I would like to thank Professor S. Hilderbrandt for his support and hospitality.

UNIVERSITY OF QUEENSLAND,
BRISBANE, QUEENSLAND, AUSTRALIA 4072,
E-mail address: hbt@maths.uq.oz.au

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

SECOND ORDER ORDINARY DIFFERENTIAL EQUATIONS WITH FULLY NONLINEAR TWO POINT BOUNDARY CONDITIONS II

H.B. THOMPSON

We establish existence results for two point boundary value problems for second order ordinary differential equations of the form $y'' = f(x, y, y')$, $x \in [0,1]$, where f satisfies the Carathéodory measurability conditions and there exist lower and upper solutions. We consider boundary conditions of the form $G((y(0), y(1)); (y'(0), y'(1))) = 0$ for fully nonlinear, continuous G and of the form $(y(i), y'(i)) \in \mathcal{J}(i)$, $i = 0, 1$ for closed connected subsets $\mathcal{J}(i)$ of the plane. We obtain analogues of our results for continuous f. In particular we introduce compatibility conditions between the lower and upper solutions and : (i) G; (ii) the $\mathcal{J}(i)$, $i = 0, 1$. Assuming these compatibility conditions hold and, in addition, f satisfies assumptions guarenteeing a'priori bounds on the derivatives of solutions we show that solutions exist. As an application we generalise some results of Palamides.

1. Introduction.

In this paper we consider two point boundary value problem for second order ordinary differential equations of the form

$$y'' = f(x, y, y'), \text{ for almost all } x \in [0,1] \tag{1.1}$$

where $f : [0,1] \times \mathbb{R}^2 \to \mathbb{R}$ satisfies the Carathéodory conditions. By a solution of (1.1) we mean a function y with y' absolutely continuous and y satisfying (1.1) almost everywhere. The first class of boundary conditions we will consider are of the form

$$0 = G((y(0), y(1)); (y'(0), y'(1))), \tag{1.2}$$

where $G : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ is continuous. We call boundary conditions of this form fully nonlinear boundary conditions. The second class of boundary conditions we will consider are of the form

$$(y(i), y'(i)) \in \mathcal{J}(i) \text{ for } i = 0, 1, \tag{1.3}$$

where $\mathcal{J}(i)$ are continuua. We will call boundary conditions of this form boundary set conditions.

We always assume that lower and upper solutions $\alpha \leq \beta$, respectively, exist for (1.1) (see Definition 1 below). We prove analogues of our existence results for the case f is continuous.

In paragraph 2 we introduce some notation, definitions and preliminary results. We define lower and upper solutions which are the natural analogues of those for continuous f. These cannot be used directly in maximum principle arguments. We define strong lower and strong upper solutions which can be used directly in maximum principle arguments and show how lower and upper solutions can be approximated by strong lower and strong upper solutions, respectively, for an approximating differential equation. We introduce the central notion of compatibility of the boundary conditions G with the lower and upper solutions. In the literature when lower and upper solutions are assumed to exist and the Picard, Neumann or Periodic boundary conditions are considered the assumptions usually made are equivalent to compatibility (see [**29**]).

In paragraph 3, we present our main existence results. If the boundary conditions G are compatible with α and β and f satisfies additional assumptions guarenteeing a'priori bounds for y' for solutions y of (1.1), then there exist solutions y of (1.1) and (1.2) satisfying $\alpha \leq y \leq \beta$ on $[0,1]$. The existence proofs follow the same general lines as in the case that f is continuous (see [**29**]) but with an additional and more subtle modification argument (see [**28**]).

In paragraph 4 we give some applications generalising some results of Palamides [**24**].

In paragraph 5 we consider problem (1.1) and (1.3). We recall two types of compatibility of the boundary sets $\mathcal{J}(i)$, $i = 0, 1$ with the lower and upper solutions (see the author [**29**]). These are satisfied by the usual boundary sets conditions considered in the literature. We give some existence results for problem (1.1) and (1.3) when the boundary sets are compatible.

The compatibility conditions are concrete conditions involving the given data which can be easily checked and are satisfied by just about every concrete existence result in the literature. Most existence results in the literature for (1.1) together with (1.2) or (1.3) which assume lower and upper solutions exist follow as a corollary to our results. In many cases our results can be used to significantly improve upon these results. This is especially true for results concerning fully nonlinear boundary conditions as can be seen for example in the application to Theorem 2.1 of Palamides [**24**] given in paragraph 4. Also the central notion of compatibility extends to systems with lower and upper solutions, to single equations and systems with lower

and upper solutions replaced by other surfaces a'priori bounding solutions. We will discuss these extensions of our ideas and further applications of our results and their extensions in forthcoming papers.

The literature on problem (1.1) and (1.2) is vast and for further information we refer the interested reader to the excellent monographs by Bailey, Waltman and Shampine [**2**], Bernfeld and Lakshmikantham [**9**], Gaines and Mawhin [**11**], Guenther, Granas and Lee [**12**], Hartman [**13**], and Mawhin [**19**] and their references.

2. Background Notation, Definitions and Results.

In order to state our results we need some notation.

As usual we say a function $f : [0,1] \times \mathbb{R}^2 \to \mathbb{R}$ satisfies the Carathéodory conditions if

1. $f(\cdot, y, p)$ is measurable for each $(y,p) \in \mathbb{R}^2$
2. $f(x, \cdot, \cdot)$ is continuous for almost all $x \in [0,1]$
3. to each $l > 0$ there is an integrable function $r_l : [0,1] \to \mathbb{R}$ satisfying $|f(x,y,p)| \leq r_l(x)$ for all $(y,p) \in [-l,l]^2$ and almost all $x \in [0,1]$.

As usual, we denote the closure of a set T by $\bar{T}$ and its boundary by ∂T. We denote the space of m times continuously differentiable functions from A to B endowed with the maximum norm by $C^m(A;B)$. In the case of continuous functions we abreviate this to $C(A;B)$. We denote the space of measurable functions from A to B endowed with the usual norm by $L^m(A;B)$. In the case $B = \mathbb{R}$ we omitt the B. We denote by $W^{2,1}([a,b])$ the Sobolev space of functions $y : [a,b] \to \mathbb{R}$ with y' absolutely continuous and $y'' \in L^1([a,b])$ with the usual norm. If A is a bounded open subset of $\mathbb{R}^n$, $p \in \mathbb{R}^n$, $f \in C(\bar{A}; \mathbb{R}^n)$ and $p \notin f(\partial A)$ we denote the Brouwer degree of f on A at p by $d(f,A,p)$. It is common in the proof of existence of solutions of two point boundary value problems for (1.1) to modify f. We do this making use of the following functions (see [**27**]).

If $c \leq d$ are given let $\pi : \mathbb{R} \to [c,d]$ be the retraction given by

$$\pi(y,c,d) = \max\{\min\{d,y\},c\}. \tag{2.1}$$

For each $\epsilon > 0$, let $K \in C(\mathbb{R} \times (0,\infty); [-1,1])$ satisfy

1. $K(\cdot,\epsilon)$ is an odd function,
2. $K(t,\epsilon) = 0$ iff $t = 0$ and
3. $K(t,\epsilon) = 1$ for all $t \geq \epsilon/4$.

If $c \leq d$ and $\epsilon > 0$ are given, let $T \in C(\mathbb{R})$ be given by

$$T(y,c,d,\epsilon) = K(y - \pi(y,c,d),\epsilon). \tag{2.2}$$

Let

$$\mathcal{Q}(x,t) = \begin{cases} (1-x)t, & \text{for } 0 \le t \le x \le 1 \\ (1-t)x, & \text{for } 0 \le x \le t \le 1, \end{cases}$$

and $w(y_0, y_1)(x) = y_0(1-x) + y_1 x$. Let $X = C^1([0,1]) \times \mathbb{R}^2$ with the usual product norm. Define $\mathcal{C} : C([0,1]) \to C^1([0,1])$ by

$$\mathcal{C}(\phi)(x) = -\int_0^1 \mathcal{Q}(x,t)\phi(t)dt, \quad \text{for all } (\phi, C, D) \in X.$$

Clearly $\mathcal{C}$ is completely continuous.

Definition 1. We call $\alpha(\beta)$ a lower (upper) solution for (1.1) if $\alpha \in W^{2,1}([0,1])$ $(\beta \in W^{2,1}([0,1]))$ and

$$\begin{aligned} &\alpha''(x) \ge f(x, \alpha(x), \alpha'(x)), \text{ for almost all } x \in [0,1], \\ &(\beta''(x) \le f(x, \beta(x), \beta'(x)), \text{ for almost all } x \in [0,1]). \end{aligned} \tag{2.3}$$

If α and β are lower and upper solutions for (1.1) on $[0,1]$ we will assume that $\alpha \le \beta$ and set $\omega = \{(x,y) \in [0,1] \times \mathbb{R}^2 : \alpha(x) < y < \beta(x)\}$ and $\bar{\omega} = \{(x,y) \in [0,1] \times \mathbb{R}^2 : \alpha(x) \le y \le \beta(x)\}$. We will call the pair nondegenerate if $\Delta = (\alpha(0), \beta(0)) \times (\alpha(1), \beta(1))$ is nonempty. Let $\pi_\Delta : \mathbb{R}^2 \to \bar{\Delta}$ be the retraction given by

$$\pi_\Delta(C,D) = (\pi(C, \alpha(0), \beta(0)), \pi(D, \alpha(1), \beta(1))).$$

Lower and upper solutions themselves cannot be used in maximum principle arguments consequently we introduce strong lower and upper solutions (cf Ako [**1**]).

Definition 2. We call $\alpha \in W^{2,1}([0,1])$ a strong lower solution for (1.1) if to each $c \in (0,1)$ there is an open interval $I_c \subseteq (0,1)$ satisfying $c \in I_c$ and a $\delta(c) > 0$ such that

$$\alpha''(x) > f(x, u(x), v(x)), \text{ for almost all } x \in I_c, \tag{2.4}$$

where $u, v : I_c \to \mathbb{R}$ are measurable and

$$(u(x), v(x)) \in (\alpha(x) - \delta, \alpha(x)] \times (\alpha'(c) - \delta, \alpha'(c) + \delta).$$

Similarly we define a strong upper solution β by substituting β in place of α and reversing the inequalities above; in this case

$$(u(x), v(x)) \in [\beta(x), \beta(x) + \delta) \times (\beta'(c) - \delta, \beta'(c) + \delta).$$

Let $\alpha \le \beta$ be lower and upper solutions, respectively. In the next lemma we show that there exist strong lower and strong upper solutions for an

approximating equation which approximate the lower and upper solutions, respectively.

Define $g : [0,1] \times \mathbb{R}^2 \to \mathbb{R}$ by

$$g(x,y,p) = f(x, \pi(y, \alpha(x), \beta(x)), p) \text{ for all } x \in [0,1]. \tag{2.5}$$

Lemma 3. *Let $f : [0,1] \times \mathbb{R}^2 \to \mathbb{R}$ satisfy the Carathéodory measurability conditions. Let $\alpha \leq \beta$ be lower and upper solutions, respectively and g be given by* (2.5). *Given $\epsilon \in (0,1)$ there is a continuous function $\kappa_\epsilon : [0,1] \times \mathbb{R}^2 \to \mathbb{R}$ such that $|\kappa_\epsilon(x,y)| \leq \epsilon$ for all $(x,y) \in [0,1] \times \mathbb{R}$. Moreover, setting*

$$j(x,y,p) = g(x,y,p + \kappa_\epsilon(x,y)), \text{ for all } (x,y,p) \in [0,1] \times \mathbb{R}^2,$$

there exist strong lower and strong upper solutions α^ϵ and β^ϵ, respectively, for

$$y'' = j(x,y,y'), \text{ for almost all } x \in [0,1] \tag{2.6}$$

satisfying

$$\alpha(x) - \epsilon/2 \leq \alpha^\epsilon(x) \leq \alpha(x) - \epsilon/6 \leq \beta(x) + \epsilon/6 \leq \beta^\epsilon(x) \leq \beta(x) + \epsilon/2,$$

for all $x \in [0,1]$.

The proof of [**28**, Lemma 7] by the author applies. Moreover we see that $\alpha'(0) = \alpha^{\epsilon\prime}(0)$, $\beta'(0) = \beta^{\epsilon\prime}(0)$ and $|\alpha'(x) - \alpha^{\epsilon\prime}(x)|$, $|\beta'(x) - \beta^{\epsilon\prime}(x)| < \epsilon/2$ on $[0,1]$.

We associate with these strong lower and strong upper solutions α^ϵ and β^ϵ the function $\gamma : \mathbb{R} \to \mathbb{R}$ given by

$$\gamma(D, \alpha^\epsilon, \beta^\epsilon) =$$
$$\begin{cases} (\alpha(1) - D)(\alpha'(1) - \alpha^{\epsilon\prime}(1))/(\alpha(1) - \alpha^\epsilon(1)), & \text{for } D \leq \alpha(1) \\ 0, & \text{for } \alpha(1) \leq D \leq \beta(1) \\ (D - \beta(1))(\beta'(1) - \beta^{\epsilon\prime}(1))/(\beta^\epsilon(1) - \beta(1)), & \text{for } D \geq \beta(1). \end{cases}$$

Definition 4. Let $\alpha \leq \beta$ be lower and upper solutions for (1.1) on $[0,1]$. We say f satisfies the Bernstein–Nagumo–Zwirner condition if there exist $L > 0$, $h \in C([0,\infty); (0,\infty))$, $\bar{h} \in L^1([\alpha_m, \beta_M]; (0,\infty))$ and $r \in L^1([0,1]; (0,\infty))$ such that

$$|f(x,y,p)| \leq h(|p|)\bar{h}(y) + r(x), \text{ for all } (x,y) \in [0,1] \times (\alpha(x), \beta(x)) \text{ and} \tag{2.7}$$

$$\int_\sigma^L \frac{s\,ds}{h(s)} > \int_{\alpha_m}^{\beta_M} \bar{h}(s)\,ds\,s + K \int_0^1 r(x)\,dx, \tag{2.8}$$

where $K = \sup\{s/h(s) : s \in [\sigma, L]\}$, $\beta_M = \max\{\beta(x) : x \in [0,1]\}$, $\alpha_m = \min\{\alpha(x) : x \in [0,1]\}$ and $\sigma = \max\{|\beta(1) - \alpha(0)|, |\beta(0) - \alpha(1)|\}$.

See Bernstein [**10**], Nagumo [**20**], Scorza Dragoni [**26**], Zwirner [**30**] and Thompson [**28**].

Remark 5. In the special case $\bar{h} = 1$ and $r = 0$ this has been called the Bernstein–Nagumo condition by some authors (see Granas et al [**12**]).

For the convenience of the reader and the sake of completeness we recall some notation and definitions from [**29**].

Definition 6. We call the vector field $\Psi = (\psi^0, \psi^1) \in C(\bar{\Delta}; \mathbb{R}^2)$ strongly inwardly pointing on $\bar{\Delta}$ if for all $(C, D) \in \partial\Delta$

$$\begin{aligned} &\psi^0(\alpha(0), D) > \alpha'(0), \psi^0(\beta(0), D) < \beta'(0) \text{ and} \\ &\psi^1(C, \alpha(1)) < \alpha'(1), \psi^1(C, \beta(1)) > \alpha'(1). \end{aligned} \tag{2.9}$$

We call Ψ inwardly pointing if the strict inequalities are replaced by weak inequalities.

Definition 7. Let $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$. We say G is strongly compatible with α and β if for all strongly inwardly pointing Ψ on $\bar{\Delta}$

$$\mathcal{G}(C, D) \neq 0 \text{ for all } (C, D) \in \partial\Delta \text{ and} \tag{2.10}$$

$$d(\mathcal{G}, \Delta, 0) \neq 0, \tag{2.11}$$

where

$$\mathcal{G}(C, D) = G((C, D); \Psi(C, D)) \text{ for all } (C, D) \in \bar{\Delta}. \tag{2.12}$$

We say G is compatible with α and β if there is a sequence $\{G_i\}_{i=1}^{\infty}$ strongly compatible with α and β which converges to G uniformly on compact subsets of $\bar{\Delta} \times \mathbb{R}^2$.

3. Existence of Solutions.

Theorem 1. *Assume that f satisfies the Carathéodory measurability conditions, that there exist nondegenerate lower and upper solutions $\alpha \leq \beta$ for* (1.1) *and f satisfies the Bernstein–Nagumo–Zwirner condition. If $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$ is compatible with α and β, then problem* (1.1) *and* (1.2) *has a solution y lying between α and β.*

Proof. Assume first that G is strongly compatible with α and β.

We approximate the lower and upper solutions by strong lower and strong upper solutions α^ϵ and β^ϵ, respectively, for the approximating differential equation (2.6). We modify this equation for y not between α and β to

obtain a second pair of constant strong lower and strong upper solutions α_ϵ and β_ϵ, respectively, satisfying $\alpha_\epsilon < \alpha^\epsilon < \alpha \leq \beta < \beta^\epsilon < \beta_\epsilon$. We also modify the boundary conditions so that they are compatible with α^ϵ and β^ϵ. We reformulate the approximating problem as a coupled system of integral and boundary condition equations and show that a solution of the modified problem lies in the region where j is unmodified and hence is solution of the approximating equation and modified boundary conditions. We obtain the required solution by using compactness to find a subsequence converging to the desired solution. We use Schauder degree theory to prove existence for the modified problem and compute the degree using a homotopy; it is easier to construct a suitable homotopy for the modified equation and boundary conditions.

Extend $\bar{h}$ to $\mathbb{R}$ by $\bar{h}(y) = \bar{h}(\pi(y, \alpha_m, \beta_M))$. By (2.8) and the Monontone convergence theorem there exist $\epsilon_0 > 0$ such that

$$\int_{\sigma+2\epsilon_0}^{L} \frac{sd\,s}{h(s)+\epsilon_0} > \int_{\alpha_m-\epsilon_0}^{\beta_M+\epsilon_0} \bar{h}(s)d\,s + K\int_0^1 (r(x)+\epsilon_0)d\,x. \tag{3.1}$$

Choose $N > \max\{|\alpha'(x)|, |\beta'(x)|, L : x \in [0,1]\} + \epsilon_0$, and let

$$k(x,y,p) = j(x, \pi(y, \alpha^\epsilon(x), \beta^\epsilon(x)), \pi(p, -N, N)), \text{ and} \tag{3.2}$$

$$\begin{aligned} l(x,y,p) =& (1 - |T(y, \alpha^\epsilon(x), \beta^\epsilon(x), \epsilon)|)k(x,y,p) + \\ & T(y, \alpha^\epsilon(x), \beta^\epsilon(x), \epsilon)(|k(x,y,p)| + \epsilon), \end{aligned} \tag{3.3}$$

where π and T are given by (2.1) and (2.2), respectively. Let $\alpha_\epsilon = \alpha_m - \epsilon$ and $\beta_\epsilon = \beta_M + \epsilon$. Thus l satisfies the Carathéodory conditions on $[0,1] \times \mathbb{R}^2$ and by continuity, for almost all $x \in [0,1]$ and ϵ small enough

$$|l(x,y,p)| \leq (h(|p|) + \epsilon_0)\bar{h}(y) + (r(x) + \epsilon_0), \tag{3.4}$$

for all $(y,p) \in [\alpha_\epsilon, \beta_\epsilon] \times [-N, N]$.

Consider

$$y'' = l(x, y, y'), \text{ for almost all } x \in [0,1] \tag{3.5}$$

together with

$$G(\pi_\Delta(y(0), y(1)); (y'(0), y'(1) + \gamma(y(1), \alpha^\epsilon, \beta^\epsilon))) = 0. \tag{3.6}$$

Suppose that (3.5) and (3.6) has a solution y^ϵ satisfying $\alpha^\epsilon \leq y^\epsilon \leq \beta^\epsilon$ and $|y^{\epsilon\prime}| \leq L$ on $[0,1]$. Then by compactness there is a subsequence y^{ϵ_i} converging in $W^{2,1}([0,1])$ to y, say, as ϵ_i converges to 0 and y is the required solution. To see this proceed as follows. First $\alpha \leq y \leq \beta$ and $|y'| \leq L$ on $[0,1]$. Since

κ_{ϵ_i} converges uniformly to 0, letting ϵ_i go to 0 in the the integral equation satisfied by y^{ϵ_i} and noting that f and l coincide for $(x,y) \in \bar{\omega}$ it follows that y is a solution of the (1.1). Now $\gamma(y^{\epsilon_i}(1), \alpha^{\epsilon_i}, \beta^{\epsilon_i})$ converges to 0 and thus y satisfies the boundary conditions.

We show that there is such a solution y^ϵ. First

$$\begin{aligned}\alpha_\epsilon'' &= 0 > -(|k(x,u(x),v(x))| + \epsilon) \\ &= l(x,u(x),v(x)), \text{ for almost all } x \in [0,1], u(x) \leq \alpha_m - 2\epsilon/3.\end{aligned}$$

Thus α_ϵ is a strong lower for

$$y'' = \theta l(x,y,y') \text{ for almost all } x \in [0,1] \tag{3.7}$$

for each $\theta \in (0,1]$. Similarly β_ϵ is a strong upper solution for (3.7).

Let $\bar{\omega}^\epsilon = \{(x,y^\epsilon) \in [0,1]\times\mathbb{R} : \alpha^\epsilon(x) \leq y^\epsilon \leq \beta^\epsilon(x)\}$ and $\Delta^\epsilon = (\alpha^\epsilon(0), \beta^\epsilon(0))\times(\alpha^\epsilon(1), \beta^\epsilon(1))$.

Suppose that y^ϵ is a solution of (3.7) with $\alpha_\epsilon \leq y^\epsilon \leq \beta_\epsilon$ on $[0,1]$ and $\theta \in [0,1]$. Then $|y^{\epsilon\prime}| < L$ by the standard argument (see for example the author [**28**]). Suppose that y^ϵ is a solution of (3.5) and (3.6) satisfying $\alpha_\epsilon \leq y^\epsilon \leq \beta_\epsilon$ and $(y^\epsilon(0), y^\epsilon(1)) \in \bar{\Delta}^\epsilon$. We show that $\alpha^\epsilon \leq y^\epsilon \leq \beta^\epsilon$ on $[0,1]$ and hence y^ϵ is the required solution. Suppose for example that $y^\epsilon(t) < \alpha^\epsilon(t)$ for some $t \in [0,1]$. From the boundary conditions and continuity we may assume that $\alpha^\epsilon - y^\epsilon$ attains its positive maximum at $t \in (0,1)$. Thus $\alpha^{\epsilon\prime}(t) = y^{\epsilon\prime}(t)$. From (3.3) and the continuity of $\alpha^{\epsilon\prime}$ and $y^{\epsilon\prime}$ there is an interval $J_t \subseteq I_t$ such that we have

$$\begin{aligned}y^{\epsilon\prime\prime}(x) &= l(x, y^\epsilon(x), y^{\epsilon\prime}(x)) \\ &< l(x, \alpha^\epsilon(x), \alpha^{\epsilon\prime}(x)) < \alpha^{\epsilon\prime\prime}(x) \text{ for almost all } x \in J_t\end{aligned}$$

a contradiction. Similarly $y^\epsilon \leq \beta^\epsilon$ on $[0,1]$. Thus y^ϵ is the required solution.

We show that y^ϵ exists. As the proof is similar to that in [**29**, Theorem 1] we only sketch the proof highlighting the differences.

Let $\Omega_\epsilon = \{y \in C^1([0,1]) : \alpha_\epsilon < y < \beta_\epsilon, |y'| < N, \text{ on } [0,1]\}$ and $\Gamma_\epsilon = \Omega_\epsilon \times \Delta^\epsilon \subseteq X$.

Define $\mathcal{L} : C^1([0,1]) \to L^1([0,1])$ by

$$\mathcal{L}(\phi) = l(\cdot, \phi, \phi').$$

Let Ψ be a strongly inwardly pointing vector field on $\bar{\Delta}$ and let $\mathcal{A} : \Delta^\epsilon \to \mathbb{R}^2$ be given by

$$\mathcal{A}(C,D) = G(\pi_\Delta(C,D); \Psi(\pi_\Delta(C,D))). \tag{3.8}$$

Then $\mathcal{A} \neq 0$ on $\partial\Delta^\epsilon$, since G is strongly compatible with α and β. Define $\mathcal{H} : \bar{\Gamma}_\epsilon \times [0,1] \to X$ by

$$\begin{aligned}
\mathcal{H}(\phi, C, D, \theta) &= (\phi + \mathcal{CL}(\phi) - w(C,D), \mathcal{V}(\phi, C, D, \theta)) \text{ for } \theta \in [2/3, 1], \\
\mathcal{H}(\phi, C, D, \theta) &= (\phi + 3(\theta - 1/3)\mathcal{CL}(\phi) - w(C,D), \mathcal{A}(C,D)) \\
&\qquad \text{for } \theta \in [1/3, 2/3] \text{ and} \\
\mathcal{H}(\phi, C, D, \theta) &= (\phi - 3\theta w(C,D) - (1 - 3\theta)(\alpha_\epsilon + \beta_\epsilon)/2, \mathcal{A}(C,D)) \\
&\qquad \text{for } \theta \in [0, 1/3], \text{ where} \\
\mathcal{V}(\phi, C, D, \theta) &= G(\pi_\Delta(C,D); \mathcal{S}(\phi, C, D, \theta)) \text{ and} \\
\mathcal{S}(\phi, C, D, \theta) &= 3(\theta - 2/3)(\phi'(0), \phi'(1) + \gamma(D, \alpha^\epsilon, \beta^\epsilon)) \\
&\qquad + 3(1 - \theta)\Psi(\pi_\Delta(C,D)).
\end{aligned}$$

Clearly $\mathcal{H}$ is completely continuous. Now $\mathcal{H}(y^\epsilon, y^\epsilon(0), y^\epsilon(1), 1) = 0$ iff y^ϵ is a solution of (3.5) and (3.6) with $(y^\epsilon, y^\epsilon(0), y^\epsilon(1)) \in \bar{\Gamma}_\epsilon$. If there is a solution $(y^\epsilon, y^\epsilon(0), y^\epsilon(1))$ in $\partial\Gamma_\epsilon$ we are through.

Assume that there is no such solution. We show that $\mathcal{H}$ is a homotopy for Schauder degree on Γ_ϵ at 0. Assume that there is a solution (y^ϵ, C, D) of $\mathcal{H}(y^\epsilon, C, D, \theta) = 0$ in $\partial\Gamma_\epsilon$ with $\theta \in [0,1)$. From the formula for $\mathcal{H}$ we see that $\theta \notin [0, 1/3)$.

Assume that $\theta \in [2/3, 1)$. As in the case $\theta = 1$, $\alpha^\epsilon(x) \leq y^\epsilon(x) \leq \beta^\epsilon(x)$ on $[0,1]$, $y^\epsilon(0) = C$ and $y^\epsilon(1) = D$.

Assume $(y^\epsilon(0), y^\epsilon(1)) \in \partial\Delta^\epsilon$. If $y^\epsilon(1) = \alpha^\epsilon(1)$, then $y^{\epsilon\prime}(1) \leq \alpha^{\epsilon\prime}(1)$. This leads to the contradiction that $\mathcal{S}(y^\epsilon, y^\epsilon(0), y^\epsilon(1), \theta) < \alpha'(1)$ and $\mathcal{V}(y^\epsilon, y^\epsilon(0), y^\epsilon(1), \theta) \neq 0$. The other cases $(y^\epsilon(0), y^\epsilon(1)) \in \partial\Delta^\epsilon$ follow similarly.

Now $\alpha_\epsilon < y^\epsilon < \beta_\epsilon$ on $(0,1)$ since α_ϵ and β_ϵ are strong lower and strong upper solutions, respectively, for (3.5). Thus $y^\epsilon \notin \partial\Omega_\epsilon$ and there are no solutions of $\mathcal{H}(y^\epsilon, C, D, \theta) = 0$ with $\theta \in [2/3, 1]$ and $(y^\epsilon, C, D) \in \partial\Gamma_\epsilon$.

Assume that $\theta \in [1/3, 2/3)$. Since $\mathcal{A}(C,D) \neq 0$ on $\partial\Delta^\epsilon$ there are no solutions (y^ϵ, C, D) with $(C,D) \in \partial\Delta^\epsilon$. The proof that $y^\epsilon \notin \partial\Omega_\epsilon$ is similar to that for $\theta \in [2/3, 1)$.

Thus $\mathcal{H}$ is a homotopy for the Schauder degree. Now $\mathcal{H}(\phi, C, D, 0) = (\phi - c, \mathcal{A}(C,D))$ where $c \in \Omega_\epsilon$ is a constant and $\mathcal{A} \neq 0$ in $\bar{\Delta}^\epsilon \setminus \Delta$. Thus by the Homotopy invariance, Reduction and Excision properties of Schauder degree

$$\begin{aligned}
d(\mathcal{H}(\cdot, 1), \Gamma_\epsilon, 0) &= d(\mathcal{H}(\cdot, 0), \Gamma_\epsilon, 0) \\
&= d(\mathcal{A}, \Delta^\epsilon, 0) \\
&= d(\mathcal{A}, \Delta, 0) \neq 0.
\end{aligned}$$

Thus there is a solution y^ϵ of (3.5) and (3.6) and hence a solution y of (1.1) and (1.2).

Suppose now that G is compatible with α and β. Then there is a sequence $\{G_i\}_{i=1}^{\infty}$ strongly compatible with α and β and converging uniformly to G on compact subsets of $\bar{\Delta} \times \mathbb{R}^2$. Let y_i be the corresponding solutions. By compactness there is a subsequence of the y_i converging in $W^{2,1}([0,1])$ to the desired solution. □

Remark 8. In the case Δ is degenerate we have to modify the result. Let $\alpha \leq \beta$ be lower and upper solutions for (1.1), respectively and suppose, for example, that $\alpha(0) = \beta(0)$. Then we set $\Delta = (\alpha(1), \beta(1))$ and change the other conditions as follows.

We call the vector field $\Psi \in C(\bar{\Delta})$ strongly inwardly pointing on $\bar{\Delta}$ if for all $D \in \partial\Delta$

$$\Psi(\alpha(1)) < \alpha'(1), \Psi(\beta(1)) > \beta'(1).$$

We call Ψ inwardly pointing if the strict inequalities are replaced by weak ones.

Let $G \in C(\bar{\Delta} \times \mathbb{R})$ and $\mathcal{G}(D) = G(D, \Psi(D))$ for all $D \in \bar{\Delta}$. We say G is strongly compatible with α and β if for all strongly inwardly pointing Ψ on $\bar{\Delta}$

$$\mathcal{G}(D) \neq 0 \text{ for all } D \in \partial\Delta \text{ and}$$
$$d(\mathcal{G}, \Delta, 0) \neq 0.$$

We define compatible as before. Theorem 1 and its proof are modified in the obvious way. In the degenerate case $\alpha(0) = \beta(0)$ and $\alpha(1) = \beta(1)$ strong compatibility implies that there are no solutions to the problem.

As mentioned earlier our central idea leads to existence results provided f is such that there are a'priori bounds on y' for solutions y satisfying $\alpha \leq y \leq \beta$. We now discuss the case where f satisfies the Nagumo–Knobloch–Schmitt condition.

Definition 9. Let $\alpha \leq \beta$ be lower and upper solutions for (1.1) on $[0,1]$. We say f satisfies the Nagumo–Knobloch–Schmitt conditions relative to α and β if there exists $\Phi < \Upsilon \in C^1([0,1] \times \mathbb{R})$ such that

$$f(x, y, \Phi(x,y)) \geq \Phi_x(x,y) + \Phi_y(x,y)\Phi(x,y) \text{ and} \tag{3.9}$$
$$f(x, y, \Upsilon(x,y)) \leq \Upsilon_x(x,y) + \Upsilon_y(x,y)\Upsilon(x,y) \tag{3.10}$$

for almost all $x \in [0,1]$ and all $y \in [\alpha(x), \beta(x)]$.

See Nagumo [**21**, **22**], Knobloch [**16**, **17**] and Schmitt [**25**].

Theorem 2. *Assume that there exist nondegenerate lower and upper solutions $\alpha \leq \beta$ for* (1.1), *that f satisfies the Nagumo–Knobloch–Schmitt condition, that $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$ is compatible with α and β, that $\alpha'(x) \geq \Phi(x, \alpha(x))$ and $\Upsilon(x, \beta(x)) \geq \beta'(x)$ almost everywhere and moreover $G((C, D); (E, F)) = 0$ only if $E \in [\Phi(0, C), \Upsilon(0, C)]$. Then problem* (1.1) *and* (1.2) *has a solution y lying between α and β.*

Proof. Again we modify f. Choose

$$N > \max\{|\Phi(x, y)|, |\Upsilon(x, y)|, |\alpha'(x)|, |\beta'(x)| : (x, y) \in \bar{\omega}\}$$

and let

$$l(x, y, p) = \begin{cases} \max\{f(x, y, \Phi(x, y)) + (\Phi(x, y) - p), f(x, y, p)\}, & \text{for } p \leq \Phi(x, y) \\ \min\{f(x, y, \Upsilon(x, y)) + (\Upsilon(x, y) - p), f(x, y, p)\}, & \text{for } p \geq \Upsilon(x, y) \\ f(x, y, p), & \text{otherwise} \end{cases}$$

and

$$m(x, y, p) = l(x, y, \pi(p, -N, N)).$$

Thus α and β are lower and upper solutions for

$$y'' = m(x, y, y') \text{ for almost all } x \in [0, 1]. \tag{3.11}$$

It is easy to see that m satisfies the conditions of Theorem 1 and thus there is a solution y of problem (3.11) and (1.2) satisfying $\alpha \leq y \leq \beta$. To show that this is a solution of our problem it suffices to show that $\Phi(x, y) \leq y' \leq \Upsilon(x, y)$. From the boundary conditions there are no solutions for $y'(0) \notin [\Phi(0, y(0)), \Upsilon(0, y(0))]$. Suppose that $y'(t) < \Phi(t, y(t))$ for some $t \in (0, 1]$. By continuity and the definition of N we may choose t and $u \in (0, t)$ such that $-N < y'(x) < \Phi(x, y(x))$ for all $x \in (u, t]$ and $y'(u) = \Phi(u, y(u))$. Now

$$\begin{aligned} (y'(x) - \Phi(x, y(x)))' &= m(x, y(x), y'(x)) \\ &\quad - \Phi_x(x, y(x)) - \Phi_y(x, y(x))\Phi(x, y(x)) \\ &> f(x, y(x), \Phi(x, y(x))) \\ &\quad - \Phi_x(x, y(x)) - \Phi_y(x, y(x))\Phi(x, y(x)) \geq 0, \end{aligned}$$

a contradiction. Thus $\Phi(x, y) \leq y'$. Similarly the $y' \leq \Upsilon(x, y)$ and the result follows. □

Remark 10. The conditions $G((C, D); (E, F)) = 0$ only if

$$E \in [\Phi(0, C), \Upsilon(0, C)],$$

(3.9) and (3.10) guarentee the solution y satisfies

$$\Phi(x, y(x)) \leq y'(x) \leq \Upsilon(x, y(x)).$$

There are other ways to guarentee this as for example in the case of periodic boundary conditions where we may replace the inequality signs in (3.9) and (3.10) by not equals to signs. See for example Schmitt [**25**].

4. Applications.

To show that the boundary conditions (1.2) are compatible we must show that (2.11) holds. Usually this follows easily from the properties of Brouwer degree (see, for example, Lloyd [**18**]) however the following lemma often suffices.

Lemma 11. *Let* $M = (M^0, M^1) \in C(\bar{\Delta}; \mathbb{R}^2)$ *satisfy*

$$M^0(\alpha(0), D) \leq 0, M^0(\beta(0), D) \geq 0$$

and

$$M^1(C, \alpha(1)) \leq 0, M^1(C, \beta(1)) \geq 0$$

for all $(C, D) \in \bar{\Delta}$ *and*

$$M(C, D) \neq 0$$

for all $(C, D) \in \partial\Delta$, *then* $d(M, \Delta, 0) \neq 0$.

This follows since $\mathcal{S} = \theta M + (1-\theta)(I-p)$ is a homotopy of M with $I-p$ where I is the identity on $\mathbb{R}^2$ and $p \in \Delta$ is any point.

Problems of the form (1.1) and (1.2), usually for the case f is continuous, have been considered by many authors. Shooting methods have been used combined variously with the maximum principle, with the Jordan separation theorem, the Kneser–Hukuhara continuum theorem and/or the Ważewski retraction theorem. Often these have been refined in the process. See Baxley [**4**], Baxley and Brown [**3**], Bernfeld and Palamides [**8**], Jackson and Klassen [**14**], Jackson and Palamides [**15**], Palamides [**23**, **24**] and their references. In [**24**] Palamides used an extension of Ważewski's retraction principle involving the Kneser–Hukuhara continuum theorem and the maximum principle to prove the following existence result.

Theorem 2.1 of [**24**]. *Let* f *satisfy the Carathéodory conditions and for each fixed* $p \in \mathbb{R}$ *and almost all* $x \in [0,1]$ *let* $f(x, \cdot, p)$ *be nondecreasing for* $y \in [\alpha(x), \beta(x)]$, *where* $\alpha(x) = -m + \gamma(x)$ *and* $\beta(x) = m - \gamma(x)$, $\gamma(x) = (1-[1+K\nu m^\nu x]^{(\nu-1)/\nu})/(K(\nu-1)m^{\nu-1})$ *and* $K > 0, m > 0, \nu = 2k+1$, $k =$

$1, 2, \ldots$, *and* $\hat{m} = -\gamma'(1)$ *are such that: For each* $\rho > 0$ *there exists a nondecreasing real-valued function* $M(\rho)$ *such that*

$$|f(x, y, p_2) - f(x, y, p_1)| \leq M(\rho)|p_2 - p_1|, \text{ for } |p_2 - p_1| \leq \rho$$
$$|f(x, 0, p)| \leq Kp^{\nu+1}, \text{ for } |p| \geq \hat{m}$$

for almost all $x \in [0, 1]$ *and all* $y \in [\alpha(x), \beta(x)]$, *where* $M(\rho) \leq K\rho^{\nu}$, *for* $\rho \geq \hat{m}$. *Let* f *satisfy the Knobloch–Nagumo–Schmitt condition. Assume that there exists* $\theta_0 \in (\pi/4, \pi/2)$ *such that* $G((C, D); (P, Q))$ *is continuous for* $(C, P, D, Q) \in E \times \hat{E}$, *where*

$$E = \left\{(C, C\tan\theta) : -m \leq C \leq m \text{ and } \frac{\pi}{4} \leq \theta \leq \theta_0\right\} \text{ and}$$
$$\hat{E} = [\alpha(1), \beta(1)] \times [q_m, q_M]$$

where $q_m = \min\{\Phi(1, y) : \alpha(1) \leq y \leq \beta(1)\}$ *and* $q_M = \max\{\Upsilon(1, y) : \alpha(1) \leq y \leq \beta(1)\}$. *Moreover assume that:*

1. *For each fixed* $C \in [-m, m], \phi \in [\pi/4, \theta_0]$ *and* $(D, Q) \in \hat{E}$

$$g^1((y, \alpha(1)); (y\tan\phi, -\hat{m})), \quad g^1((C, z); (C\tan\phi, Q))$$

and $g^1(y, D); (C\tan\phi, q)$ *are nondecreasing functions with respect to the corresponding variables* y, z *or* q *but* $g^1((C, z); (C\tan\phi, q))$ *is (strictly) increasing with respect to both variables* z *and* q *and furthermore for each* $y \in [-m, m]$ *and* $\theta \in [\pi/4, \theta_0]$ *we have*

$$(4.1) \qquad g^1((y, \alpha(1)); (y\tan\theta, -\hat{m})) \leq 0 \leq g^1((y, \beta(1)); (y\tan\theta, \hat{m})).$$

2. *For each point* $y \in [-m, m]$ *we have*

$$\Phi(0, y) \leq y, y\tan\theta_0 \leq \Upsilon(0, y).$$

3. *For each pair of points* (y_1, z_1, q_1) *and* $(y_2, z_2, q_2) \in [-m, m] \times \hat{E}$ *we have*

$$(4.2) \qquad g^0((y_1, z_1); (y_1, q_1))g^0((y_2, z_2); (y_2\tan\theta_0, q_2)) \leq 0.$$

Then problem (1.1) *and* (1.2) *has a solution* y *such that for all* $x \in [0, 1]$ $\alpha(x) \leq y(x) \leq \beta(x)$ *and* $\Phi(x, y(x)) \leq y'(x) \leq \Upsilon(x, y(x))$.

We indicate how this result can be generalised after first showing how it follows from our Theorem 2.

Outline. As in [**24**], $\alpha < \beta$ are lower and upper solutions for (1.1). There are two cases to consider. The first case is $g^0((C, D); (C, Q))$ not identically 0 on $[\alpha(0), \beta(0)] \times \hat{E}$. We modify G without changing its zero set and also

denote the modification by G. Then we extend G to $\bar{\Delta}\times\mathbb{R}^2$ without changing its zero set when $(D,Q)\in\hat{E}$ so that the extension is compatible with α and β. Then solutions of the new problem are solutions.

Replace $g^0((C,D);(P,Q))$ by $Cg^0((C,D);(P,Q))$. Let $s_i:[\alpha(0),\beta(0)]\to\mathbb{R}$, $i=1,2$ be defined by

$$-s_1(-C)=s_2(C)=\begin{cases}C\tan\theta_0, & \text{for } C\geq 0\\ C, & \text{for } C<0.\end{cases}$$

By (4.2), replacing g^0 by $-g^0$, if necessary, we may assume that

$$g^0((C,D);(s_2(C),Q))\leq 0$$

and

$$g^0((C,D);(s_1(C),Q))\geq 0$$

for all $(D,Q)\in\hat{E}$. We extend g^0 as a continuous function to $\bar{\Delta}\times\mathbb{R}^2$ satisfying $g^0((C,D);(P,Q))<0$ for all $P>s_2(C)$ and $g^0((C,D);(P,Q))>0$ for all $P<s_1(C)$ for all $(D,Q)\in[\alpha(1),\beta(1)]\times\mathbb{R}$. By (4.1), and monotonicity, $g^1((C,\alpha(1));(P,Q))\leq 0$ for all $q_m\leq Q\leq\alpha'(1)$ and $g^1((C,\beta(1));(P,Q))\geq 0$ for all $q_M\geq Q\geq\beta'(1)$ and all $(C,P)\in E$; it is not difficult to show from the definition of lower and upper solutions and the Knobloch–Nagumo–Schmitt condition that $q_m<\alpha'(1)$ and $\beta'(1)<q_M$. Extend g^1 to a continuous function on $\bar{\Delta}\times\mathbb{R}^2$ so that $g^1((C,\alpha(1));(P,Q))\leq 0$ for all $Q\leq\alpha'(1)$ and $g^1((C,\beta(1));(P,Q))\geq 0$ for all $Q\geq\beta'(1)$. It is easy to see that G now has the required properties. We show that $\alpha'(x)\geq\Phi(x,\alpha(x))$ and $\Upsilon(x,\beta(x))\geq\beta'(x)$ almost everywhere so that Theorem 2 applies and a solution exists.

Now $\alpha'(0)=-m=\alpha(0)$ so $\alpha'(0)\geq\Phi(0,\alpha(0))$ by assumption (2). Suppose that $z(t)=\alpha'(t)-\Phi(t,\alpha(t))<0$ for some $t\in(0,1]$. By continuity we may choose $u\in[0,t)$ such that $z(u)=0$ and $z<0$ on $(u,t]$. By (2.3), (2.8) and the lipschitz condition on f there is a constant k such that

$$\begin{aligned}z'(x)&\geq f(x,\alpha(x),\alpha'(x))-\Phi_x(x,\alpha(x))-\Phi_y(x,\alpha(x))\alpha'(x)\\&\geq f(x,\alpha(x),\alpha'(x))-f(x,\alpha(x),\Phi(x,\alpha(x)))\\&\qquad-\Phi_y(x,\alpha(x))(\alpha'(x)-\Phi(x,\alpha(x)))\\&\geq kz(x)\end{aligned}$$

for almost all $x\in[u,t]$, a contradiction. Thus $\alpha'(x)\geq\Phi(x,\alpha(x))$ almost everywhere. Similarly $\Upsilon(x,\beta(x))\geq\beta'(x)$ almost everywhere.

The second case is $g^0((C,D);(C,Q))$ identically 0 on $[\alpha(0),\beta(0)]\times\hat{E}$. In this case we replace E by the set $E_1=\{(C,C):C\in[\alpha(0),\beta(0)]\}$ and $s_1(C)=s_2(C)=C$. The rest of the proof remains unchanged. □

Remark 12. In the second case of the proof above the boundary condition g^0 admits the solution $y(0) = y'(0)$ and in our proof we have effectively replaced g^0 by the simpler $y(0) - y'(0)$ and our solution satisfies this.

In the above proof the monotonicity assumptions on g^1 are used only to guarentee that $g^1((C, \alpha(1)); (P, Q)) \leq 0$ for all $q_m \leq Q \leq \alpha'(1)$ and $g^1((C, \beta(1)); (P, Q)) \geq 0$ for all $q_M \geq Q \geq \beta'(1)$ and all $(C, P) \in E$ and hence can be relaxed. In Palamides's proof they are required in a shooting argument and it is not clear how they can be weakened.

We do not need either the local lipschitz or monotonicity conditions on f required in Palamides's proof for application of the maximum principle in a shooting argument. We used the local lipschitz condition only along $(x, \alpha(x), \alpha'(x))$ and $(x, \beta(x), \beta'(x))$ and only to show that $\Upsilon(x, \beta(x)) \geq \beta'(x)$ and $\alpha'(x) \geq \Phi(x, \alpha(x))$ almost everywhere. We used the monotonicity condition on f only in the construction of the lower and upper solutions. Palamides also used the monotonicity condition on f in the construction of the lower and upper solutions.

Moreover the other results of [**24**] also follow from our Theorems 1 and 2; in the statement of Theorem 2.2 of [**24**] conditions on g have been omitted although the intended conditions are clear.

We illustrate the improvement our results represent over [**24**] by modifying the example presented there.

Example. Let

$$f(x, y, p) = -\sin x - (\cos x - y^2) 2 \sin y - p^5 \text{ for } x \in [0, 1],$$

$$g^0((y(0), y(1)); (y'(0), y'(1))) = -y'(0) + y(0) + (y^2(1) - y'^2(1))/10 \text{ and}$$

$$g^1((y(0), y(1)); (y'(0), y'(1))) = y(0) + y'(0) + 6y(1) + y'(1) + \sin(y(0) - y'(1)).$$

Thus we have translated the x interval so it is now $[0, 1]$ and modified f in the y variable so that it is no longer monotonic with respect to y. In view of our remark above Palamides's example has a solution with $y(0) = y'(0)$ so we have modified g^0 to avoid this. Also we modified g^1 to avoid monotonicity.

To see that there is a solution we apply Theorem 2 to a modified problem. We let $\beta(x) = \pi/2 = -\alpha(x)$ and $\Upsilon(x, y) = 2 = -\Phi(x, y)$. It is easy to check that the $\alpha < \beta$ are lower and upper solutions and that the Knobloch–Nagumo–Schmitt condition is satisfied. As above we set $\hat{E} = [-\pi/2, \pi/2] \times [-2, 2]$ but replace E by $E_2 = \hat{E}$. It is easy to check that

if Ψ is a strongly inwardly pointing vector field with $|\Psi| \leq 2$ then conditions (2.10) and (2.11) are satisfied. We modify G for $(y'(0), y'(1)) \notin [-2,2]^2$ by projecting $(y'(0), y'(1))$ in the obvious way so that the modified G is strongly compatible with α and β. Thus, by Theorem 2 there is a solution y of the modified problem. However given the bounds on y and on y' the solution lies in the region where G was not modified. Thus y is the required solution.

Notice that from our analysis of Theorem 2.1 of [**24**] and the above it is clear we could have obtained existence of solutions for the example of [**24**] from our Theorem 2 using constant α, β, Υ and Φ.

5. Boundary Set Conditions.

In this section we consider problem (1.1) and (1.3) again assuming that there exist lower and upper solutions $\alpha \leq \beta$, respectively, and look for solutions y lying between α and β.

Problems of the form (1.1) and (1.3) for the case f is continuous have been considered by many authors. Shooting methods have been used combined with with the Jordan separation theorem (see Bebernes and Fraker [**7**] and Bebernes and Wilhelmsen [**5**, **6**] and their references).

We show that analogues of the results of Bebernes and Fraker [**7**] for the case f is continuous can be derived from our results.

In order to state our results we need some notation (see Bebernes and Fraker [**7**]). For $x \in [0,1]$ let $C(x) = \{(x,y,p) \in \bar{\omega} \times \mathbb{R}\}$, $S_\alpha(x) = \{(x,y,p) \in C(x) : y = \alpha(x)\}$, and $S_\beta(x) = \{(x,y,p) \in C(x) : y = \beta(x)\}$. For the convenience of the reader and the sake of completeness we recall the definition of compatibility of boundary sets (see [**29**]).

Definition 13. We say the pair of sets $\{\mathcal{J}(0), \mathcal{J}(1)\} \subset \mathbb{R}^2$ is strongly compatible, respectively compatible, for (1.1), α and β if there exists $G \in C(\bar{\Delta} \times \mathbb{R}^2; \mathbb{R}^2)$ which is strongly compatible, respectively compatible, for (1.1), α and β and such that $G((C,D);(E,F)) \neq 0$ for all $(C,E,D,F) \notin \mathcal{J}(0) \times \mathcal{J}(1)$.

Definition 14. Let $\mathcal{J}(i) \subset [\alpha(i), \beta(i)] \times \mathbb{R}$, $i = 0$ or 1 be a closed connected set. We say it is of compatible type 1 if there is $(\alpha(i), u(i)) \in \mathcal{J}(i)$, where $(-1)^i(\alpha'(i) - u(i)) \geq 0$, and there is $(\beta(i), u(i)) \in \mathcal{J}(i)$, where $(-1)^i(u(i) - \beta'(i)) \geq 0$. We say it is of compatible type 2 if for every $p \in \mathbb{R}$ there is $y \in [\alpha(i), \beta(i)]$ such that $(y,p) \in \mathcal{J}(i)$. If it is of compatible type 1 or 2 we say simply it is of compatible type.

Theorem 3 [**29**, Theorem 4]. *Let the sets $\mathcal{J}(i) \subset \mathbb{R}^2$, $i = 0, 1$ be of compatible type, then the pair $\{\mathcal{J}(0), \mathcal{J}(1)\}$ is compatible for* (1.1), *α and β.*

The next result is an immediate consequence of Theorems 1 and 3.

Theorem 4. *Assume that there exist lower and upper solutions, $\alpha \leq \beta$, respectively, for* (1.1), *that f satisfies a Bernstein–Nagumo–Zwirner condition and and that the sets $\mathcal{J}(i)$, $i = 0, 1$ are of compatible type. Then there is a solution of* (1.1) *and* (1.3) *lying between α and β.*

We now state the analogue for measurable f of [**7**, Theorem 1].

Theorem 5. *Assume that there exist lower and upper solutions $\alpha \leq \beta$ for* (1.1) *and that for any $B_2 \geq 0$ and $t_0 \in (0, 1]$ there is $N(B_2) > 0$ such that any solution of* (1.1) *with $|y'(0)| \leq B_2$ and $\alpha(x) \leq y(x) \leq \beta(x)$ for all $x \in [0, t_0]$ satisfies $|y'(x)| \leq N(B_2)$ for all $x \in [0, t_0]$. If $\mathcal{J}(0)$ is compact and of compatible type 1 and $\mathcal{J}(1)$ is of compatible type 2, then problem* (1.1) *and* (1.3) *has a solution lying between α and β.*

Proof. By compactness there is $B_2 > 0$ such that $(y(0), y'(0)) \in \mathcal{J}(0)$ implies that $|y'(0)| \leq B_2$. By assumption we may choose N such that $|y'| < N$ for all solutions y of (1.1) with $(x, y) \in \bar{\omega}$ on $[0, 1]$ and L such that $L > \max\{|\alpha'(x)|, |\beta'(x)|, N : x \in [0, 1]\}$. Let

$$j(x, y, p) = f(x, y, \pi(p; -L, L))$$

for all $(x, y, p) \in [0, 1] \times \mathbb{R}^2$. Consider

$$y'' = j(x, y, y') \quad \text{for all } x \in [0, 1] \tag{5.1}$$

together with (1.3). Now α and β are lower and upper solutions for (5.1) so by Theorem 4 there is a solution y for this problem between α and β. Assume that $|y'| \geq L$ for some $x \in [0, 1]$. Set $t = \min\{x \in [0, 1] : |y'(x)| \leq L\}$. By continuity and the choice of L, we have $t > 0$ and $|y'(x)| \leq L$ for all $x \in [0, t]$. Thus y is a solution of (1.1) on $[0, t]$ lying between α and β but $|y'(t)| > N$ a contradiction. Thus $|y'| < L$ on $[0, 1]$ and y is the required solution. □

The above result can be sharpened as follows.

Definition 15. Set

$$S_0 = \{(y, -L) : \alpha(0) \leq y \leq \beta(0)\} \cup \{(\alpha(0), p) : \alpha'(0) \geq p \geq -L\}, \tag{5.2}$$

$$S_2 = \{(y, L) : \alpha(0) \leq y \leq \beta(0)\} \cup \{(\beta(0), p) : \beta'(0) \leq p \leq L\}, \tag{5.3}$$

$$S_1 = \{(y, -L) : \alpha(1) \leq y \leq \beta(1)\} \cup \{(\beta(1), p) : \beta'(1) \geq p \geq -L\} \text{ and} \tag{5.4}$$

$$S_3 = \{(y, L) : \alpha(1) \leq y \leq \beta(1)\} \cup \{(\alpha(1), p) : \alpha'(1) \leq p \leq L\}. \tag{5.5}$$

We can now state the analogue for measurable f of [**7**, Theorem 3].

Theorem 6. *Assume that there exist lower and upper solutions $\alpha \leq \beta$ for* (1.1) *and that f satisfies a Bernstein–Nagumo–Zwirner condition. Further*

assume that $\mathcal{J}(i) \subseteq C(i)$, $i = 0, 1$ *are closed connected sets satisfying* $\mathcal{J}(0) \cap \{S_0 \cup S_2\} \neq \emptyset$ *and* $\mathcal{J}(1) \cap \{S_1 \cup S_3\} \neq \emptyset$, *where the* S_i *are given by* (5.2) *to* (5.5) *and* $L > \max\{|\alpha'(x)|, |\beta'(x)| : x \in [0,1]\}$ *satisfies* (2.8). *Then there is a solution* y *lying between* α *and* β.

Proof. This follows since either $(\beta(0), u) \in \mathcal{J}(0)$ for some $u \geq \beta'(0)$ or $(y, L) \in \mathcal{J}(0)$ for some $y \in [\alpha(0), \beta(0))$ and we add the straight line segment joining (y, L) to $(\beta(0), L)$ to $\mathcal{J}(0)$. Similarly, either $(\alpha(0), u) \in \mathcal{J}(0)$ for some $u \leq \alpha'(0)$ or $(y, -L) \in \mathcal{J}(0)$ for some $y \in (\alpha(0), \beta(0)]$ and we add the straight line segment joining $(y, -L)$ to $(\alpha(0), -L)$ to $\mathcal{J}(0)$. Similarly we modify $\mathcal{J}(1)$ as above. Thus the modified $\mathcal{J}(i)$ are of compatible type and, by Theorem 3, there exists a solution for (1.1) and (1.3). This solution satisfies $|y'| < L$ and hence is the required solution. □

References

[1] K. Ako, *Subfunctions for ordinary differential equations,* J. Fac. Sci. Univ. Tokyo, **I,9** (1965), 17–43.

[2] P.B. Bailey, L.F. Shampine and P.E. Waltman, *Nonlinear Two Point Boundary Value Problems,* (1974), Academic Press, New York.

[3] J.V. Baxley and S.E. Brown, *Existence and uniqueness for two-point boundary value problems,* Proc. Roy. Soc. Edinburgh Sect. A, **88** (1981), 219–234.

[4] John V. Baxley, *Existence theorems for nonlinear second order boundary value problems,* J. Differential Equations, **85** (1990), 125–150.

[5] J.W. Bebernes and R. Wilhelmsen, *A technique for solving two dimensional boundary value problems,* SIAM J. Appl. Math., **17** (1969), 1060–1064.

[6] ———, *A general boundary value technique,* J. Differential Equations, **8** (1970), 404–415.

[7] J.W. Bebernes and Ross Fraker, *A'priori bounds for boundary sets,* Proc. Amer. Math. Soc., **29** (1971), 313–318.

[8] S.R. Bernfeld and P.K. Palamides, *A Topological method for vector-valued and nth-order nonlinear boundary value problems,* J. Math. Anal. Appl., **102** (1984), 488–508.

[9] S.R. Bernfeld and V. Lakshmikantham, *An Introduction to Nonlinear Boundary Value problems,* (1974), Academic Press, New York.

[10] S.N. Bernstein, *Sur les équations du calcul des variations,* Ann. Sci. Ecole Norm. Sup., **29** (1912), 438–448.

[11] R.E. Gaines and J.L. Mawhin, *Coincidence Degree and Nonlinear Differential Equations,* Lecture Notes in Mathematics No. 568, (1977) Springer–Verlag, New York.

[12] A. Granas, R.B. Guenther and J.W. Lee, *Nonlinear boundary value problems for ordinary differential equations,* (1985), Dissertationes Mathematicae, Warsaw.

[13] P. Hartman, *Ordinary Differential Equations,* (1964), Wiley, New York.

[14] Lloyd K. Jackson and Gene Klassen, *A variation of the topological method of Ważewski,* SIAM J. Appl. Math., **20** (1971), 124–130.

[15] L.K. Jackson and P.K. Palamides, *An existence theorem for a nonlinear two point boundary value problem,* J. Differential Equations, **53** (1984), 43–66.

[16] H.K. Knobloch, *Eine neue Methode zur Approximations periodischer Lösungen nichtlinearer Differentialgleichungen zweiter Ordnung,* Math. Z., **82** (1963), 177–197.

[17] ———, *On the existence of periodic solutions for second order vector differential equations,* J. Differential Equations, **9** (1971), 67–85.

[18] N.G. Lloyd, *Degree Theory,* (1978), Cambridge University Press, London.

[19] J.L. Mawhin, *Topological Degree Methods in Nonlinear Boundary Value Problems Regional Conf. Series in Math.*, No. 40, (1979), Amer. math. Soc., Providence R.I.

[20] M. Nagumo, *Über die Differentialgleichungen* $y'' = f(x, y, y')$, Proc. Phys–Math. Soc. Japan, **19** (1937), 861–866.

[21] ———, $y'' = f(x, y, y')$ *no Kyôkaichi Mondai ni suite I,* Kansû Hôteishiki, **13** (1941), 27–34.

[22] ———, $y'' = f(x, y, y')$ *no Kyôkaichi Mondai ni suite II,* Kansû Hôteishiki, **17** (1942), 37–44.

[23] P.K. Palamides, *Singular points of the consequent mapping,* Ann. Math. Pure Appl., **CXXIX** (1981), 383–395.

[24] ———, *A topological method and its application a general boundary value problem,* J. Nonlinear Analysis, Theory, Methods & Applications, **7** (1983), 1101–1114.

[25] K. Schmitt, *Periodic solutions of second order nonlinear differential equations,* Math. Z., **98** (1967), 200–207.

[26] G. Scorza Dragoni, *Intorno a un criterio di esistenza per un problema di valori ai limiti,* Rend. Accad. Naz. Lincei, **(6) 28** (1938), 317–325.

[27] H.B. Thompson, *Existence of solutions for a two point boundary value problem,* Rend. Circ. Mat. Palermo, **(2) XXXV** (1986), 261–275.

[28] ———, *Minimal solutions for two point boundary value problems,* Rend. Circ. Mat. Palermo, **(2) XXXVII** (1988), 261–281.

[29] ———, *Second order ordinary differential equations with fully nonlinear two point boundary conditions,* Pacific J. Math., **172** (1996), 255-277.

[30] G. Zwirner, *Un criterio di esistenza per un problema di valori al contorno per equazioni differenziali del secondo ordine,* Rend. Sem. Mat. Univ. Padova, **10** (1939), 55–64.

Received February 12, 1993. The author would like to thank CMA at the Australian National University, and in particular Professor N.S. Trudinger for their hospitality and support.

UNIVERSITY OF QUEENSLAND,
BRISBANE, QUEENSLAND, AUSTRALIA 4072
E-mail address: hbt@maths.uq.oz.au

PACIFIC JOURNAL OF MATHEMATICS
Vol. 172, No. 1, 1996

THE FLAT PART OF NON-FLAT ORBIFOLDS

FENG XU

We use integrable lattice models to determine the complete invariants of a series of new finite depth orbifold subfactors from Hecke algebras.

Introduction.

Commuting squares are an efficient way of producing subfactors. One can always ask the question about how to determine the higher relative commutants once a subfactor is produced. The interest in this question is that in many well-known examples, these higher relative commutants which are finite dimensional C* algebras if the index is finite, seem to be the right "Quantum symmetry"([**2**]). In the language of coupling system of Ocneanu([**3**]), the question is to determine the flat part of a connection. In [**1**], we found a necessary and sufficient for a certain class of connections from orbifold construction([**4**]) to be flat. In this paper, we will determine the flat part of those non-flat connections. The exact meaning of this will be explained in Section 1. It turns out that the subfactor constructed from those non-flat connections is the same as a sufactor from a flat orbifold construction, where one does orbifold with respect to a subgroup of the original abelian group. (See the theorem of Section 2.) The paper is organized as follows: In Section 1 we recalled the part of [**1**] we need as well as fixing the notations. In Section 2 we proved the main result. This paper is a continuation of [**1**].

1. Orbifold Constructions in subfactors.

The material of this section is contained in Section 2 of [**1**] and Section 4 of [**4**]. Let G be a connected, simply connected, compact simple lie group with nontrivial center, i.e. $G = SU(N), SO(2N+1), SO(2N), SP(2N), E_6, E_7$. Let Z be a nontrivial subgroup of the center $Z(G)$ of G. Let ϕ be a finite dimensional representation of G. $K \in N$ is a fixed integer(level). In [**5**], a coupling system associated to ϕ is constructed, denoted by $(g_\phi(K), h_\phi(K), B, \tau)$. Here $g_\phi(K)$ is the principal graph constructed out of the fusion graph of ϕ,$h_\phi(K)$ is its dual. If K is such that $Z(0) \in g_\phi^{even} \cap h_\phi^{even}$, since the connection is invariant under the action of the central element (see 2.12 of [**1**]), one can apply orbifold method with respect to Z to this coupling system. To get

a better picture of the orbifold construction, let us take a concrete example, namely, let us describe the orbifold construction of the Wenzl's subfactor as in [**4**]. It corresponds to the case $G = SU(N)$, ϕ is the fundamental representation of SU(N), Z is the cyclic group Z_N and $N|K$. Let

$$P_{++}^n = \left\{\lambda = \sum\nolimits_{1\leq i\leq N-1}\lambda_i\Lambda_i | \lambda_i \geq 1, \sum\nolimits_{1\leq i\leq N-1}\lambda_i \leq n-1\right\},$$

where the Λ_i's are the $N-1$ weights of the fundamental representations and $n = k+N$. For fixed N, we define $\mathcal{A}^n$ as follows. The vertices of $\mathcal{A}^n$ are given by elements of P_{++}^n and its oriented edges are given by N vectors e_i defined by $e_1 = \Lambda_1, e_i = \Lambda_i - \Lambda_{i-1}, i = 1, ..., N-1, e_N = -\Lambda_{N-1}$. We define an action of the cyclic group Z_N as follows. We set $A_0 = \star$, and label the other end vertices of the graph $\mathcal{A}^n$ by $A_1 = A_0 + (n-N)e_1$, $A_2 = A_1 + (n-N)e_2$, ..., $A_{N-1} = A_{N-2} + (n-N)e_{N-1}$. Define a rotation symmetry ρ of the graph by $\rho(A_j + \sum_k c_k e_k) = A_{j+1} + \sum_k c_k e_{k+1}$, where the indices are in Z/NZ and $c_k \in C$. Note that $\rho^N = id$. The connection W(see [**3**]) which is used to embed a small algebra into a big one, is invariant under this action. The vertices of $\mathcal{A}^n$ can be colored by N colors in $Z/NZ = 0, 1, ..., N-1$ so that the starting vertex has color 0 and each oriented edges goes from a color k to a color k+1, $k \in Z/NZ$. $\mathcal{A}_r^n$ is A subgraph which has vertices of colors r and $r+1$, $r \in Z_N$. Let

$$\begin{array}{ccccccc}
C_{0,0} & \subset & C_{0,1} & \subset & C_{0,2} & \cdots \longrightarrow & C_{0,\infty} \\
\cap & & \cap & & \cap & & \cap \\
C_{1,0} & \subset & C_{1,1} & \subset & C_{1,2} & \cdots \longrightarrow & C_{1,\infty} \\
\cap & & \cap & & \cap & & \cap \\
C_{2,0} & \subset & C_{2,1} & \subset & C_{2,2} & \cdots \longrightarrow & C_{2,\infty} \\
\cap & & \cap & & \cap & & \cap \\
\vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

be the double sequences constructed as in [**4**]. Since the connection is invariant under the action of the center one can apply ρ to each $C_{l,m}$ as a *-algebra isomorphism and it is compatible with the inclusions of these algebras. Set $D_{m,n}$ to be the fixed point algebra $C_{m,n}$ of under ρ. We get another double string algebras simply replacing $D_{m,n}$ by $C_{m,n}$. The double sequences constructed here are a little different from that of [**3**]. The source of the string is allowed to be any of the A_j j=0,1,2...N-1, where in [**3**] the source is always A_0. However, the subfactor $C_{0,\infty} \subset C_{1,\infty}$ is the same as the Wezl's subfactor which is the subfactor constructed from similar double sequences but restricts the source to be A_0. This is also clearly explained in [**4**]. The reason is the so-called relative Mcduff properties of subfactors constructed out of the commuting squares. If one takes the projection p corresponding

to A_0, since $p \in C_{0,\infty}$, the relative Mcduff property implies $C_{0,\infty} \subset C_{1,\infty}$ is isomorphic to $pC_{0,\infty}p \subset pC_{1,\infty}p$ which is the standard description of Wenzl's subfactor in [**3**]. Set $D_{l,\infty}$ to be the weak closure of $\cup_m D_{l,m}$ in its GNS-representation with respect to the trace. The subfactor $D_{0,\infty} \subset D_{1,\infty}$ is called orbifold subfactors. By [**3**], the question of determining the flat part of the orbifold subfactor including finding those $\sigma' \in D_{l,0}$ such that they commute with all $\sigma \in D_{0,m}$. Since by the main result of [**1**], the connection is flat when N is odd, but when N is even, the connection is flat iff $2N|K$. When the connection is flat, the higher relative commutants are simply given by sequences of algebras $D_{m,0}, m \geq 0$. If the connection is not flat, the higher relative commutants are strictly smaller subalgebras of $D_{m,0}, m \geq 0$ which commute with all $\sigma \in D_{0,m}$. Hence our question is to determine the higher relative commutants of the orbifold subfactor $D_{0,\infty} \subset D_{1,\infty}$ in the case N is a divisor of K but 2N is not and N is even. Since the orbifold connection in this case is not flat, the orbifold subfactor is called a non-flat orbifold and the question is to determine the higher relative commutants of this orbifold subfactor. This is exactly the meaning of the title of this paper.

2. The higher relative commutants.

Now we are ready to compute the higher relative commutants. For the sake of completeness, let us first state the theorem in its general form. Let G be a connected, simply connected, compact simple lie group with nontrivial center, i.e. $G = SU(N), SO(2N+1), SO(2N), SP(2N), E_6, E_7$. Let Z be a nontrivial subgroup of the center Z(G) of G. Let θ_Z denote the set of fundamental weights of G associated to Z, and $(,)$ the killing form of G. In [**5**], a coupling system associated to ϕ is constructed, denoted by $(g_\phi(K), h_\phi(K), B, \tau)$. Here $g_\phi(K)$ is the principal graph constructed out of the fusion graph of ϕ,$h_\phi(K)$ is its dual. If K is such that $Z(0) \in g_\phi^{even} \cap h_\phi^{even}$, since the connection is invariant under the action of the central element (see 2.12 of [**1**]), one can apply orbifold method with respect to Z to this coupling system. Let ϕ be a finite dimensional representation of G. $K \in N$ is a fixed integer(level).

Theorem. *Let K,Z be as before. Let M be the least natural number such that:* $1/2M(\theta_z, \theta_z) \in \mathbb{Z}, \forall \theta_z \in \theta_Z$ *and let N be the least nature number such that* $M|N \times K$. *Let* Z^N *be the abelian subgroup of Z generated* z^N *for all* $z \in Z$. *Then the orbifold subfactor constructed out of the action of Z is the same as the orbifold subfactor constructed out of the action of* Z^N *which is necessarily flat* (*hence the flat part is easy to determine*).

As in Section 1, we are going to use orbifold subfactor of Wenzl's subfactor

to explain and prove the theorem. The proof of the general case is exactly the same except possible change of notations.

Let us assume we are given the conditions at the end of Section 1, namely, N is a divisor of K but 2N is not and N is even. As explained in Section 1, this is the only interesting case. In this case, the theorem says the orbifold subfactor with respect to the cyclic group Z_N is the same as new orbifold subfactor with respect to the subgroup $Z_{N/2}$ of Z_N which is flat by [1].

Let us first describe this new subfactor in a similar double sequences as that of Section 1.

Let $N = 2N_1$, ρ the action of Z_N as in Section 1. Then the $Z_{N/2}$ action is given by ρ^{2m}, m=0,1... N_1. The orbit of distinguished vertex A_0 under the action of ρ^{2m}, m=0,1... N_1 are vertices A_{2m}, m=0,1... N_1 Let

$$\begin{array}{ccccccc}
C_{0,0} & \subset & C_{0,1} & \subset & C_{0,2} & \cdots \longrightarrow & C_{0,\infty} \\
\cap & & \cap & & \cap & & \cap \\
C_{1,0} & \subset & C_{1,1} & \subset & C_{1,2} & \cdots \longrightarrow & C_{1,\infty} \\
\cap & & \cap & & \cap & & \cap \\
C_{2,0} & \subset & C_{2,1} & \subset & C_{2,2} & \cdots \longrightarrow & C_{2,\infty} \\
\cap & & \cap & & \cap & & \cap \\
\vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

be the double sequences as in Section 1, except that the sources of the strings are restricted to the vertices A_{2m}, m=0,1... N_1. Denote by $\hat{D}_{0,\infty} \subset \hat{D}_{1,\infty}$ the orbifold subfactor under the action of ρ^2. Since the orbifold connection is flat, the higher relative commutants $\hat{D}'_{0,\infty} \cap \hat{D}_{k,\infty}$ is given by $\hat{D}_{k,0}$, $k \geq 0$. Take paths α', β' with the same length on the graph $\mathcal{A}_0^n$ without orientation and with $s(\alpha') = A_0$, $s(\beta') = A_j$, j is even, $r(\alpha') = r(\beta') = C_0$, where C_0 is some vertex of $\mathcal{A}_0^n$. Let $\sigma' = \sum_{l=0}^{N_1-1}(\rho^{2l}(\alpha'), \rho^{2l}(\beta'))$. Note that σ''s of the above form with the length k span $D'_{k,0}$ $k \geq 0$. Similarly another family of higher relative commutants $\hat{D}'_{1,\infty} \cap \hat{D}_{k,\infty}$, $k \geq 1$, are spaned by similar σ''s, except that one takes paths on the graph $\mathcal{A}_{N-1}^n$ which is dual to $\mathcal{A}_0^n$. We will show in the following that the higher relative commutants of the orbifold subfactor with respect to the cyclic group Z_N are isomorphic to that of $\hat{D}_{0,\infty} \subset \hat{D}_{1,\infty}$ which is of finite depth. Hence they are isomorphic subfactors, thus completing the proof of the theorem.

Let us first determine $D'_{0,\infty} \cap D_{k,\infty}$. Since the connection is not flat, the higher relative commutants are strictly smaller subalgebras of $D_{k,0}, k = 0, 1...$ which commute with all $\sigma \in D_{0,k}, k = 0, 1....$ Let α, β be paths with the same length on $\mathcal{A}^n$ and with $s(\alpha) = A_0$, $s(\beta) = A_j$, $r(\alpha) = r(\beta) = B_0$, where B_0 is some vertex of $\mathcal{A}^n$. Set $\sigma = \sum_{l=0}^{N-1}(\rho^l(\alpha), \rho^l(\beta))$. Note that σ's of the above form span $D_{0,k}, k = 0, 1....$ Similarly take paths α',

β' with the same length on the graph $\mathcal{A}_0^n$ without orientation and with $s(\alpha') = A_0$, $s(\beta') = A_j$, $r(\alpha') = r(\beta') = C_0$, where C_0 is some vertex of $\mathcal{A}_0^n$. Let $\sigma'(\alpha', \beta') = \sum_{l=0}^{N-1}(\rho^l(\alpha'), \rho^l(\beta'))$. A general element $\sigma' \in D_{k,0}$ may be expressed as: $\sigma' = \sum_{|\alpha'|=|\beta'|=k} \lambda_{\alpha',\beta'}\sigma'(\alpha', \beta')$, where $\lambda_{\alpha',\beta'}$'s are complex numbers. We have to study under what conditions we get $\sigma\sigma' = \sigma'\sigma$. As in [**4**], it is equivalent to the following :

$$(1)\ \sum\nolimits_{\eta,\eta',\alpha'',\beta'',\alpha',\beta',s(\beta'')=s(\beta')=k} |\lambda_{\alpha'',\beta''} x_{\alpha'',\beta'',\eta,\eta',l} - \lambda_{\alpha',\beta'} y_{\alpha',\beta',\eta,\eta',l}|^2 = 0.$$

Where (α, β) are fixed paths on $\mathcal{A}^n, l \in Z_N$ and

$$x_{\alpha'',\beta'',\eta,\eta',l} = \begin{array}{ccccc} A_l & \xrightarrow{\rho^l(\alpha)} & \cdots & \longrightarrow & B_l \\ {\scriptstyle \rho^l(\alpha'')}\downarrow & & & & \downarrow{\scriptstyle \eta} \\ \vdots & & & & \vdots \\ \downarrow & & & & \downarrow \\ C_l & & & & . \\ \downarrow & & & & \downarrow \\ \vdots & & & & \vdots \\ {\scriptstyle \rho^l(\beta'')}\downarrow & & & & \downarrow{\scriptstyle \eta'} \\ A_{l+k} & \xrightarrow{\rho^{l+k}(\alpha)} & \cdots & \longrightarrow & B_{l+k} \end{array}$$

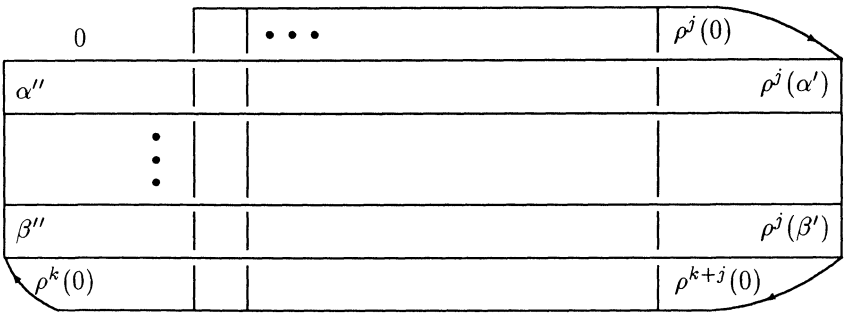
$$y_{\alpha'',\beta'',\eta,\eta',l} = \begin{array}{ccccc} A_{l+j} & \xrightarrow{\rho^l(\beta)} & \cdots & \longrightarrow & B_l \\ {\scriptstyle \rho^{l+j}(\alpha'')}\downarrow & & & & \downarrow{\scriptstyle \eta} \\ \vdots & & & & \vdots \\ \downarrow & & & & \downarrow \\ C_{l+j} & & & & . \\ \downarrow & & & & \downarrow \\ \vdots & & & & \vdots \\ {\scriptstyle \rho^{l+j}(\beta'')}\downarrow & & & & \downarrow{\scriptstyle \eta'} \\ A_{l+k+j} & \xrightarrow{\rho^{l+k}(\beta)} & \cdots & \longrightarrow & B_{l+k} \end{array}$$

We need the following Orthogonality lemma to simplyfy the above expressions.

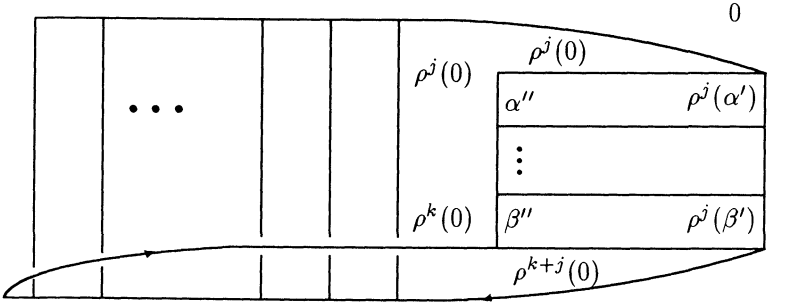
Orthogonality Lemma. *Let* $x_{\alpha'',\beta'',\eta,\eta'}, y_{\alpha,\beta,\eta,\eta'}$ *be as above. If* (α'', β'') *is different from* (α', β'), *Then:*

$$\sum_{\eta,\eta'} x_{\alpha'',\beta'',\eta,\eta'} \bar{y}_{\alpha',\beta',\eta,\eta'} = 0,$$
$$\sum_{\eta,\eta'} y_{\alpha'',\beta'',\eta,\eta'} \bar{y}_{\alpha',\beta',\eta,\eta'} = 0,$$
$$\sum_{\eta,\eta'} x_{\alpha'',\beta'',\eta,\eta'} \bar{x}_{\alpha',\beta',\eta,\eta'} = 0.$$

Proof. We will use the notations from [**5**]. The lemma follows from the topological invariance. In fact, up to nozero constants, $\sum_{\eta,\eta'} x_{\alpha'',\beta'',\eta,\eta'} \bar{y}_{\alpha',\beta',\eta,\eta'}$ is equal to the value of the following diagram (see [**5**] or [**1**]):



As in [**5**], the value of the diagram is invariant under regular isotopy, and equal to the value of the diagram:



If (α'', β'') is different from (α', β'), the coloring of the above diagram is not admissible, hence the value is 0. The other two identities follow by the same kind of argument. Q.E.D. □

Proof of the theorem. By this lemma, (1) can be simplified to be:

$$\sum_{\eta,\eta',\alpha'',\beta'',s(\beta'')=k} |\lambda_{\alpha'',\beta''}|^2 |x_{\alpha'',\beta'',\eta,\eta'} - y_{\alpha'',\beta'',\eta,\eta'}|^2 = 0.$$

By Lemma 2.20 of [**1**],

$$|x_{\alpha'',\beta'',\eta,\eta'} - y_{\alpha'',\beta'',\eta,\eta'}|^2 = 2 - 2Re\exp(2\pi ikjh_{A_1}).$$

Hence (1) is equvialent to : $\exp(2\pi ikjh_{A_1}) = 1$, for all $j \in Z_N$ and for all (α'', β'') such that $|\lambda_{\alpha'',\beta''}|$ is not zero, Where h_{A_1} denotes the conformal weight of A_1(see [**5**]). Thus k is even.

To summarize, we have shown that $D'_{0,\infty} \cap D_{m,\infty}$ is spaned by: $\sigma'(\alpha', \beta') = \sum_{l=0}^{N-1}(\rho^l(\alpha'), \rho^l(\beta'))$, Here α', β' are paths with the same length m on the graph $\mathcal{A}_0^n$ and the source of β' is A_k with k even and the source of α' is A_0. These algebras are clearly isomorphic to $\hat{D}'_{0,\infty} \cap \hat{D}_{m,\infty}$ as described in the beginning of this section. In fact, let p be the projection in $C_{0,\infty}$ corresponding to those even vertices of the graph $\mathcal{A}_0^n$, then p commutes with algebras $D'_{0,\infty} \cap D_{m,\infty}$ and $px = 0, x \in D_{0,\infty} \cap D_{m,\infty}$ iff x=0. Moreover, $p(D'_{0,\infty} \cap D_{m,\infty}) = \hat{D}'_{0,\infty} \cap \hat{D}_{m,\infty}$.

Similarly, one can show that $D'_{1,\infty} \cap D_{m,\infty}$ is isomorphic to $\hat{D}'_{1,\infty} \cap \hat{D}_{m,\infty}$, where $m \geq 1$. Q.E.D. □

We end this section with an example and a remark.

Example. Let $G = SU(2)$, K=4l+2, ϕ the spin 1/2 representation. $Z = Z_2$. We know that the orbifold is not flat. The theorem says the orbifold subfactor is the same as the original subfactor, that is, the flat part of D_n for odd n is $A_{(2n-3)}$ subfactor.

Remark. Morally speaking, if one starts with G/Z , the question of finding the flat part reduces G/Z to G/Z^N which is the correct gauge group as far as the Chern-Simons gauge theory is concerned([**6**]).

References

[1] F. Xu, *Orbifold construction in subfactors*, Comm. Math. Phys., **166** (1994), 237-253.

[2] V. Jones, *Talk on commuting squares.*

[3] A. Ocneanu, *Quantum symmetry, differential geometry of finite graphs and classification of subfactors*, University of Tokyo Seminary Notes 45 (recorded by Y.Kawahigashi), 1991.

[4] Y. Kawahigashi and Evans, *Orbifold subfactors from Hecke algebra*, to appear.

[5] J. de Boer and J. Goeree, *Markov traces and Π_1 factors in conformal field theory*, Comm. Math. Phys., **139** (1991), 267-304.

[6] E. Witten and Dijigraff, Comm. Math. Phys., **127** (1990), 137-150.

[7] A. Ocneanu, *Quantized group string algebras and Galois theory for algebras*, London Math. Soc. Lect. Notes Series, **136** (1988).

[8] H. Wenzl, *Hecke algebras of type A_n and subfactors*, Invent. Math., **92** (1988), 345-383.

Received May 15, 1993 and revised December 3, 1993. I came to the question when I heard several interesting talks in V. Jones' seminars given by A. Wasserman to whom I wish to express my gratitude. I'm grateful to my advisor V. Jones for his constant help and encouragement. I also benefit from talks with Y. Kawahigashi.

UNIVERSITY OF CALIFORNIA
LOS ANGELLES, CA 90095
E-mail address: xufeng@math.ucla.edu

PACIFIC JOURNAL OF MATHEMATICS

Volume 172 No. 1 January 1996

A class of incomplete non-positively curved manifolds 1
BRIAN BOWDITCH

The quasi-linearity problem for C^*-algebras 41
L. J. BUNCE and JOHN DAVID MAITLAND WRIGHT

Distortion of boundary sets under inner functions. II 49
JOSE LUIS FERNANDEZ PEREZ, DOMINGO PESTANA and JOSÉ RODRÍGUEZ

Irreducible non-dense $A_1^{(1)}$-modules 83
VJACHESLAV M. FUTORNY

M-hyperbolic real subsets of complex spaces 101
GIULIANA GIGANTE, GIUSEPPE TOMASSINI and SERGIO VENTURINI

Values of Bernoulli polynomials 117
ANDREW GRANVILLE and ZHI-WEI SUN

The uniqueness of compact cores for 3-manifolds 139
LUKE HARRIS and PETER SCOTT

Estimation of the number of periodic orbits 151
BOJU JIANG

Factorization of p-completely bounded multilinear maps 187
CHRISTIAN LE MERDY

Finitely generated cohomology Hopf algebras and torsion 215
JAMES PEICHENG LIN

The positive-dimensional fibres of the Prym map 223
JUAN-CARLOS NARANJO

Entropy of a skew product with a Z^2-action 227
KYEWON KOH PARK

Commuting co-commuting squares and finite-dimensional Kac algebras 243
TAKASHI SANO

Second order ordinary differential equations with fully nonlinear two-point boundary conditions. I 255
H. BEVAN THOMPSON

Second order ordinary differential equations with fully nonlinear two-point boundary conditions. II 279
H. BEVAN THOMPSON

The flat part of non-flat orbifolds 299
FENG XU