MOTION OF HYPERSURFACES BY GAUSS CURVATURE

Ben Andrews

We consider \( n \)-dimensional convex Euclidean hypersurfaces moving with normal velocity proportional to a positive power \( \alpha \) of the Gauss curvature. We prove that hypersurfaces contract to points in finite time, and for \( \alpha \in \left( \frac{1}{n+2}, \frac{1}{n} \right] \) we also prove that in the limit the solutions evolve purely by homothetic contraction to the final point. We prove existence and uniqueness of solutions for non-smooth initial hypersurfaces, and develop upper and lower bounds on the speed and the curvature independent of initial conditions. Applications are given to the flow by affine normal and to the existence of non-spherical homothetically contracting solutions.

1. Introduction.

Motivation for the study of hypersurfaces moving by their Gauss curvature comes from several sources:

1.1. Tumbling stones. W.J. Firey introduced the Gauss curvature flow in 1974, as a model of the wearing process undergone by a pebble on a beach [Fi]. Consider a stone which occupies an open, bounded convex region of \( \mathbb{R}^{n+1} \) at time \( t = 0 \). The stone tumbles, and collides with a hyperplane (the beach) with random orientation. We assume for simplicity that the amount of material removed in a collision at a point \( x \) of the stone depends only on the normal direction \( \nu_x \) (thus allowing some anisotropy in the material of the stone). The number of collisions with a region \( B \) of the surface of the stone is proportional to the measure of the set \( \nu(B) = \{ \nu_x : x \in B \} \subset S^n \) of normal directions to \( B \). This is equal to \( \int_B K_x d\mathcal{H}^n(x) \), where \( K_x \) is the Gauss curvature of the hypersurface at \( x \). The rate at which the stone wears away at a point \( x \) is given by \( \rho(\nu_x)K_x \) for some positive function \( \rho \) on \( S^n \), and we have the evolution equation

\[
\dot{x} = -\rho(\nu_x)K_x \nu_x.
\]

1.2. Affine geometry: Inner parallel surfaces and the affine normal flow. K. Leichtweiss introduced the notion of inner parallel surfaces for a convex body in affine geometry in the paper [Le]. Given a convex region \( \Omega \), the idea is to construct a family of related regions \( P_t \Omega \), by a procedure which is well-defined in the setting of affine geometry — that is, if we perform
an area-preserving affine transformation $L$ to get a new region $L\Omega$, then $P_t(L\Omega) = L(P_t\Omega)$.

The procedure is as follows: For each direction $z \in S^n$, there exists a unique supporting hyperplane $H_z = \{ \langle z, y \rangle = h(z) \}$ to $\Omega$ with normal direction $z$ pointing outward from $\Omega$. The notion of parallel hyperplanes is well-defined in affine geometry, as is the notion of the volume of a region. Hence we can choose a unique hyperplane $H_{z,t}$ parallel to $H_z$ such that the volume of the part of $\Omega$ between $H_{z,t}$ and $H_z$ is equal to $t^{(n+2)/2}$ (in the case where $\Omega$ is the region above a paraboloid, this choice of exponent ensures that $H_{z,t}$ moves at constant speed). Equivalently, we can define a function $h_t(z)$ by the requirement

$$\text{Vol} \left( \{ y \in \Omega : h_t(z) \leq \langle y, z \rangle \leq h(z) \} \right) = t^{(n+2)/2}. \tag{2}$$

Then we define $P_t\Omega$ to be the convex set

$$\bigcap_{z \in S^n} \{ y \in \mathbb{R}^{n+1} : \langle z, y \rangle \leq h_t(z) \}. \tag{3}$$

In contrast to the corresponding situation in Euclidean geometry, this procedure does not define a semi-group: If we begin with a region $\Omega$, construct the regions $P_t\Omega$, and use them to construct the regions $P_tP_t\Omega$, then these are not in general given by $P_t\Omega$ for any $t'$. To remedy this we consider $P_{t/n}\Omega$, obtained by following the above construction repeatedly over small intervals, and take the limit $n \to \infty$ of infinitesimally small steps to obtain a region $\tilde{P}_t\Omega$. This defines a deformation which is clearly well-defined in affine geometry, and satisfies the semi-group property $\tilde{P}_t\tilde{P}_{t'} = \tilde{P}_{t+t'}$.

To find an explicit description of this deformation in the case where $\Omega$ is smooth and strictly convex, we consider the regions $P_t\Omega$ in the limit of small $t$: Fix $z$, and choose coordinates for $\mathbb{R}^{n+1}$ such that $e_1, \ldots, e_n$ span the supporting hyperplane $H_z$ of $\Omega$, and the supporting point is at the origin. Then $M = \partial \Omega$ is locally a graph in these coordinates:

$$x_{n+1} = -\frac{1}{2} \sum_{i,j=1}^n h_{ij} x_i x_j + O(|x|^3),$$

where $h_{ij}$ is the second fundamental form at the supporting point. There exists a volume-preserving linear transformation which fixes the $e_{n+1}$ direction and brings $M$ locally to the form

$$x_{n+1} = -\frac{1}{2} K^{1/n} \sum_{i=1}^n x_i^2 + O(|x|^3)$$

where $K = \det h_{ij}$ is the Gauss curvature at the supporting point. Then

$$\text{Vol} \left( \{ y \in \Omega : h(z) - d \leq \langle y, z \rangle \leq h(z) \} \right) = 2^{n/2} \omega_n K^{-1/2} d^{(n+2)/2} + O(d^{(n+3)/2})$$
where \( \omega_n \) is the volume of the unit ball in \( \mathbb{R}^n \), and hence the requirement (2) implies that

\[
h_t(z) = h(z) - \frac{K_1^n}{2^n \omega_n^{\frac{3}{2}}} t + O(t^{3/2}).
\]

It follows that the limiting deformation is given by the equation

\[
\dot{x} = -c_n K_1^n \nu_x.
\]

This evolution equation is the simplest invariant flow in affine differential geometry; up to reparametrisation it is the motion of a hypersurface in the direction of its affine normal vector. This has been considered in [ST1, ST2] for the case of convex curves in the plane, and in [A4] more generally. For nonconvex curves results were recently obtained in [AST].

1.3. Image analysis. Many fundamental problems in image analysis have been approached using geometric flows: An image represented by a grey-scale density function \( u \) can be processed to remove noise by smoothing the level sets of \( u \) with a parabolic flow. Various candidates have been considered, but in [AGLM] axioms were proposed which included the natural requirement of affine invariance. This leads to the evolution Equation (4). In the case of nonconvex hypersurfaces this is no longer parabolic, and various authors (see [AGLM], [CS], [NK]) have considered the generalization (for two-dimensional surfaces)

\[
\dot{x} = -(\text{sgn} H) \max\{K, 0\}^{1/4} \nu
\]

where \( H \) is the mean curvature. Applications of plane curve evolution equations to image analysis and computer vision are described in [AGLM], [OST], and [ST1]-[ST3].

1.4. Gradient flows of the mean width. The width of a convex region \( \Omega \) in a direction \( z \in S^n \) is defined by

\[
w(z) = \sup_{y_1, y_2 \in \Omega} \langle y_1 - y_2, z \rangle = h(z) + h(-z)
\]

where \( h(z) = \sup_{y \in \Omega} \langle y, z \rangle \) is the support function of \( \Omega \). The mean width \( V_1(\Omega, \varphi) \) with respect to a measure \( \varphi d\mu \) on \( S^n \) is given (up to a constant factor) by integrating the width over all directions \( z \in S^n \):

\[
V_1(\Omega) = \int_{S^n} w(z) \varphi(z) d\mu(z) = \int_{S^n} h(z) (\varphi(z) + \varphi(-z)) d\mu(z) = \int_{\partial \Omega} K h \tilde{\varphi} d\mathcal{H}^n,
\]
where $\tilde{\varphi}(z) = \varphi(z) + \varphi(-z)$. The first variation formula for the mean width can be calculated as follows: Consider a smooth family $\Omega_t$ of convex regions, with support functions $h_t(z)$ such that

$$\frac{\partial}{\partial t} h_t(z) \bigg|_{t=0} = f(z),$$

then

$$\frac{d}{dt} V_1(\Omega_t) \bigg|_{t=0} = \int_{S^n} \tilde{\varphi}(z) f(z) d\mu = \int_{\partial \Omega_0} K f \tilde{\varphi} d\mathcal{H}^n.$$

We consider the flow of steepest descent of the mean width in $L^p$ spaces on $\partial \Omega$ — that is, we seek that variation $f$ for which $V_1(\Omega_t, \varphi)$ decreases fastest amongst all variations with the same $L^p$ norm ($\int_{\partial \Omega} |f|^p \sigma(\nu) d\mathcal{H}^n)^{1/p}$ ($\sigma$ is a positive smooth function on $S^n$): By the Hölder inequality we have for $p > 1$

$$\left| \int_{\partial \Omega_0} K f \tilde{\varphi} d\mathcal{H}^n \right| \leq \left( \int_{\partial \Omega_0} |f|^p \sigma d\mathcal{H}^n \right)^{1/p} \left( \int_{\partial \Omega_0} \left( \frac{K \tilde{\varphi}}{\sigma} \right)^{p/(p-1)} \sigma d\mathcal{H}^n \right)^{1-1/p}$$

with equality if and only if $f = c (K \tilde{\varphi}/\sigma)^{1/(p-1)}$. The flow of steepest descent is therefore

$$(5) \quad \dot{x} = -\rho(\nu_x) K^{1/(p-1)}(\nu_x)$$

where $\rho = (\tilde{\varphi}/\sigma)^{1/(p-1)}$ is a smooth positive function on $S^n$.

1.5. Evolving hypersurfaces and degenerate fully nonlinear PDE.

The evolution equations derived above are included in a large class of parabolic evolution equations for hypersurfaces which have been considered before. Simplest in this class is the mean curvature flow, in which a hypersurface moves in the direction of its inward normal with speed given by the mean curvature. Huisken [Hu] showed that convex hypersurfaces moving under such equations contract to points in finite time, and that the hypersurfaces become spherical in shape in the process. This argument has since been extended to many processes where convex hypersurfaces move with speeds given by homogeneous degree one, concave or convex monotone symmetric functions of the principal curvatures: Chow considered flows by the $n$th root of the Gauss curvature [Ch1] and the square root of the scalar curvature [Ch2], and the author has considered a general class of such evolution equations [A1]. Corresponding results for flows where the speed has other positive degrees of homogeneity in the curvature seem much harder to prove. The author has treated the special case of flow by the power $1/(n+2)$ of the Gauss curvature, which is the flow by affine normal [A4]. Tso [Ts]
and Chow [Ch1] have shown that hypersurfaces moving with speed equal to any positive power of the Gauss curvature contract to points in finite time.

The Gauss curvature flows form a convenient class of examples of parabolic equations with varying degeneracy: For large $\alpha$ they become more degenerate, and for small $\alpha$ they become singularly parabolic. Intermediate values of $\alpha$ are singular in some situations and degenerate in others. The precise effect of such degeneracy or singularity on the regularity of solutions is extremely complicated. In particular, it would be interesting to know how irregular solutions can be, how the regularity estimates depend on time (particularly where the initial solution is highly irregular), and how solutions behave in the neighbourhood of degenerate or singular regions.

There are several other important families of PDE for which similar questions can be asked — in particular, natural families of parabolic equations with varying degeneracy include the porous medium equations

$$\dot{u} = \Delta(|u|^{m-1}u),$$

and the $p$-harmonic heat flows

$$\dot{u} = \nabla \cdot (|\nabla u|^{p-2}\nabla u),$$

for which there is also a natural generalization to $p$-harmonic maps between Riemannian manifolds. The Gauss curvature flows can be considered a geometric analogue of the porous medium equations.

In the case of curves in the plane, more complete results are known: Gage [Ga1]-[Ga2] and Hamilton [GH] showed that convex curves contract to points in finite time and become round under the curve shortening flow (where the speed of motion equals the curvature), and Grayson [Gr] extended this by showing that any embedded curve eventually becomes convex. This was extended to include anisotropic analogues of the curve-shortening flow by Gage [Ga3] in the convex case, and by Oaks [Oa] for nonconvex curves. The author considered equations of varying degeneracy in the convex case [A2], [A8], and obtained optimal estimates on the regularity of solutions, including their initial behaviour. The particular case of the affine normal flow has also been extended to nonconvex curves [AST].

2. The result.

Our main aim in this paper is to prove results about the regularity and limiting behaviour of solutions of the Gauss curvature flows of the form

$$\frac{dx}{dt} = -\rho(\nu(x))K(x)^{\alpha}\nu(x),$$

where $\alpha$ is in the range $[1/(n + 2), 1/n]$. These particular exponents arise as follows in the proof: We first prove (in Section 4 of the paper) that the solutions of Gauss curvature flows have isoperimetric ratio bounded as long
as they exist, provided \( \alpha \) is greater than the critical value \( 1/(n + 2) \). This value is sharp — the flow with \( \rho \equiv 1 \) and \( \alpha = 1/(n + 2) \) is the affine normal flow, for which solutions converge to ellipsoids of arbitrary eccentricity [A4], and so the isoperimetric ratio tends to stay bounded but does not generally improve; in a separate paper [A9] we prove that for exponents smaller than this (or equal to this if \( \rho \) is nonconstant) there are solutions which have isoperimetric ratios approaching infinity. Second, we prove (in Sections 5 and 6) that if the hypersurface has bounded isoperimetric ratio, and \( \alpha \leq 1/n \), then a short time later the moving hypersurfaces are strictly convex and have bounded curvature. The exponent \( 1/n \) is again sharp, as there are solutions of the Gauss curvature flow for any \( \alpha > 1/n \) which remain non-strictly convex and are not \( C^\infty \) — in fact any initial convex hypersurface which includes a planar piece will behave this way. This phenomenon was first noted by Richard Hamilton for the case \( \alpha = 1 \) [Ha1]. We describe such behaviour more fully in Section 12 of this paper.

By combining these results, we obtain the following:

**Theorem 1.** For any open bounded convex region \( \Omega_0 \), any smooth positive function \( \rho : S^n \to \mathbb{R} \), and any \( \alpha \in (1/(n + 2), 1/n] \), there exists a family of embeddings \( x : S^n \times [0,T) \to \mathbb{R}^{n+1} \) satisfying (6), unique up to composition with an arbitrary time-independent diffeomorphism, such that \( M_t = x(S^n, t) \) converges in Hausdorff distance to the boundary of the region \( \Omega_0 \) as \( t \) approaches zero. \( x \) is smooth and strictly convex for \( t > 0 \) and converges to a point \( p \in \mathbb{R}^{n+1} \) as \( t \) approaches \( T \). Furthermore, the hypersurfaces

\[
\tilde{M}_t = \left( \frac{\text{Vol}(S^n)}{\text{Vol}(M_t)} \right)^{1/(n+1)} (M_t - p)
\]

converge in \( C^\infty \) as \( t \) approaches \( T \), to a smooth, strictly convex limit hypersurface \( \tilde{M}_T \) for which \( \langle x, \nu \rangle = c\rho(\nu)K^\alpha \) for some \( c > 0 \).

This means that the evolving hypersurface contracts to a point, and asymptotically approaches a solution which evolves purely by homothetically scaling about this limiting point.

We also have the following generalisation for smaller \( \alpha \):

**Theorem 2.** The result of Theorem 1 also holds for a solution of (6) with any \( \alpha \in (0, 1/n] \), provided the isoperimetric ratios of the evolving hypersurfaces remain bounded.

We deduce in Section 7 the first part of Theorem 1, that solutions contract to points in finite time (in fact we prove this for any \( \alpha > 0 \)). This was proved previously for isotropic cases by Tso [Ts] and Chow [Ch1]. In Section 8 we prove that solutions exist starting from singular initial hypersurfaces, and immediately become smooth and strictly convex if \( \alpha \leq 1/n \) — note that we make no regularity assumptions about the initial hypersurface, other than
those implied by its convexity. We also prove in Theorem 15 the existence of unique viscosity solutions for $\alpha > 1/n$, although this is not required for the proof of Theorems 1 and 2. In Section 9 we digress from the main argument of the paper to apply the regularity and convexity estimates in a simple new proof of the convergence theorem for the affine normal flow. In Section 11 of the paper we use Theorem 2 to deduce the existence of non-spherical homothetic solutions of Equation (6) for constant $\rho$ and suitable $\alpha$ between 0 and $1/(n + 2)$.

We remark that Urbas [U2] has considered noncompact solutions of isotropic equations of the form (6), in particular proving the existence of solutions which evolve by homothetically expanding or translating.

3. Notation and preliminaries.

The inradius $r_-$ of an open convex region is the supremum of the radii of all balls contained in it, and the circumradius $r_+$ is the infimum of the radii of balls containing it. In this paper we refer to the ratio $r_-/r_+$ as the isoperimetric ratio of the body.

For a convex region with boundary given by a smooth embedding $x : M \to \mathbb{R}^{n+1}$, we have an outward unit normal vector field $\nu : M \to S^n \subset \mathbb{R}^{n+1}$, which we use to define the Weingarten map $W_x : T_xM \to T_xM$ by the formula

$$W(u) = D_u\nu \in T_{\nu(x)}S^n \simeq T_xM$$

for any $x \in M$ and $u \in T_xM$. The eigenvalues $\lambda_1, \ldots, \lambda_n$ of $W(x)$ are the principal curvatures of $M$ at $x$. The elementary symmetric functions $E_j$ of these are defined by

$$E_j = \frac{1}{(n_j)} \sum_{1 \leq i_1 < i_2 < \ldots < i_j \leq n} \left( \prod_{k=1}^{j} \lambda_{i_k} \right).$$

In particular, $K = E_n$ is the Gauss curvature, and $H = E_1$ is the mean curvature.

The covariant derivative $\nabla$ on the hypersurface is given by the formula

$$\nabla_u v = D_u v + \langle W(u), v \rangle \nu.$$
Alternatively, in the case where $M$ is strictly convex and smooth, the support function can be used to define a canonical embedding $\bar{x}$ of $S^n$ with image equal to $M$:

$$\bar{x}(z) = h(z)z + \tilde{\nabla}_ih(z)\tilde{g}^{ij}\tilde{\nabla}_jz.$$  

This has the property that the outward normal to $M$ at the point $\bar{x}(z)$ is equal to $z$, for each $z \in S^n$.

The Weingarten map can also be recovered directly from $h$:

$$\langle W^{-1}(u), v \rangle = \tilde{\nabla}_u \tilde{\nabla}_vh + \langle u, v \rangle h$$

where $\tilde{\nabla}$ is the covariant derivative on $S^n$, and we identify $T_xM$ and $T_{\nu(x)}S^n$.

For convenience we will denote by $r_{ij}$ the corresponding symmetric bilinear form, the eigenvalues of which are the principal radii of curvature $r_i = \lambda_i^{-1}$, $i = 1, \ldots, n$:

$$r_{ij} = \tilde{\nabla}_i\tilde{\nabla}_j h + \bar{g}_{ij} h.$$  

$r_{ij}$ satisfies a Codazzi-type identity:

$$\tilde{\nabla}_k r_{ij} = \tilde{\nabla}_k \tilde{\nabla}_i \tilde{\nabla}_j h + \bar{g}_{ij} \tilde{\nabla}_kh$$

Differentiating (11), commuting derivatives, and applying (11) again to the result, we obtain a version of the Simons' identity for the second derivatives of the second fundamental form $[\text{Si}]:$

$$\bar{\nabla}(\tilde{\nabla}_i \tilde{\nabla}_j)r_{kl} = \tilde{\nabla}(\tilde{\nabla}_k \tilde{\nabla}_l)r_{ij} + \bar{g}_{ij} r_{kl} - \bar{g}_{kl} r_{ij}$$

where the brackets denote symmetrisation.

For convenience, we will denote by $S_k$ the $k$th elementary symmetric function of the eigenvalues of $r_{ij}$. In particular, $S_n = K^{-1}$. $S_k$ may be considered as a function of the components of the matrix $r_{ij}$, and we denote by $\dot{S}_k^{ij}$ and $\ddot{S}_k^{ijpq}$ the first and second derivatives:

$$\dot{S}_k^{ij} = \frac{\partial S_k}{\partial r_{ij}}$$

and

$$\ddot{S}_k^{ijpq} = \frac{\partial^2 S_k}{\partial r_{ij} \partial r_{pq}}.$$  

$\dot{S}_k$ is a positive definite symmetric bilinear form provided $r$ is positive definite, and $S_k^{1/k}$ is a concave function of the components of $r$ for $k = 1, \ldots, n$ [Mi].
The support function allows the degenerate parabolic system of equation (6) to be re-written as a parabolic scalar equation (see [U1], [A1], [A4]):

\[
\frac{dh(z)}{dt} = -\rho(z) \det (\nabla^2 h + \text{Id})^{-\alpha}.
\]

In particular, this implies the existence of a smooth solution of Equation (6) for a short time for any smooth, strictly convex initial hypersurface.

In a region where a family of hypersurfaces moving under Equation (6) can be represented as graphs \(x_{n+1} = u_t(x_1, \ldots, x_n)\) for some convex functions \(u_t\), we can work with an equivalent scalar parabolic equation for the functions \(u_t\):

\[
\frac{\partial}{\partial t} u = \tilde{\rho}(Du) \frac{(\det D^2 u)^\alpha}{(1 + |Du|^2)^{\alpha(n+2) - 1}}
\]

where

\[
\tilde{\rho}(Du) = \rho \left( \frac{\sum_{i=1}^n e_i D_i u - e_{n+1}}{\sqrt{1 + |Du|^2}} \right).
\]

We note some elementary features of Eq. (6): First, the speed of motion is given by a homogeneous function of the curvatures, and this homogeneity leads to a scaling property of solutions. Specifically, if \(x : M \times [0, T] \to \mathbb{R}^{n+1}\) satisfies Eq. (6), then for each \(\lambda > 0\) another solution \(x_\lambda : M \times [0, \lambda^{1+n\alpha}T] \to \mathbb{R}^{n+1}\) is given by

\[
x_\lambda(p, t) = \lambda x(p, \lambda^{-(1+n\alpha)}t).
\]

This also implies corresponding scaling invariance properties for the solutions of Equations (13) and (14).

A second important property of solutions of Eq. (6) is the comparison principle: If \(\{M_t^{(i)}\}, i = 1, 2\) are two families of smooth, strictly convex hypersurfaces moving under Eq. (6), and \(M_0^{(1)} \cap M_0^{(2)} = \emptyset\), then \(M_t^{(1)} \cap M_t^{(2)} = \emptyset\) for all \(t > 0\) in the common interval of existence. A local version also holds: If \(\{M_t^{(i)}\}, i = 1, 2\) are families of smooth, strictly convex hypersurfaces with boundary, \(M_0^{(1)} \cap M_0^{(2)} = \emptyset\), and \(M_t^{(1)} \cap \partial M_t^{(2)} = M_t^{(2)} \cap \partial M_t^{(1)} = \emptyset\) for \(t \in [0, T]\), then \(M_t^{(1)} \cap M_t^{(2)} = \emptyset\) for \(t \in [0, T]\).

### 4. Monotone quantities and diameter bounds.

In this section we prove that whenever \(\alpha > 1/(n+2)\), the evolving hypersurfaces have bounded isoperimetric ratios for as long as the solution exists. The main tool used here is an integral estimate known as the entropy estimate, which was proved for the case \(\alpha = 1\) by Chow [Ch1], and for other \(\alpha\) by the author [A3].
We define an integral quantity $Z_{\rho, \alpha}$ for any given $\alpha$ and $\rho$ by

$$Z_{\rho, \alpha} = \text{Vol}(M)^{n/(n+1)} \left( \frac{1}{\int_{S^n} \rho \, d\mu} \int_M \rho K^\alpha \, dH^n \right)^{1/(\alpha-1)}$$

if $\alpha \neq 1$, and

$$Z_{\rho, 1} = \text{Vol}(M)^{n/(n+1)} \exp \left\{ \frac{1}{\int_{S^n} \rho \, d\mu} \int_M \rho \log K \, dH^n \right\}$$

if $\alpha = 1$. For convenience, we also denote by $Z^\#_{\alpha}$ the same quantity with $\rho \equiv 1$.

**Theorem 3.** For any smooth, strictly convex solution of Equation (6),

$$\frac{d}{dt} Z_{\rho, \alpha} \leq 0$$

with equality if and only if the equation $\langle x, \nu \rangle = c \rho(\nu) K^\alpha$ holds for some $c > 0$ and some choice of origin in $\mathbb{R}^{n+1}$.

This integral bound will be combined with the following estimate to deduce isoperimetric ratio bounds for solutions of the flow:

**Theorem 4.** For any smooth, strictly convex hypersurface $M^n$ in $\mathbb{R}^{n+1}$,

$$\frac{r_+(M)}{r_-(M)} \leq C(\alpha, \rho) Z_{\rho, \alpha}^{(n+1)\beta(\alpha)}$$

for some positive constant $\beta(\alpha)$, provided $\alpha > 1/(n + 2)$.

**Proof.** We begin with a bound in terms of $Z_{z, \alpha}$:

Following [Ha2], we begin by obtaining a lower bound on the $n$-dimensional areas of projections of $M$ onto hyperplanes: Given a direction $z_0 \in S^n$, the area of the projection on to the plane with normal $z_0$ is given by

$$A_{z_0} = \frac{1}{2} \int_{S^n} |\langle z, z_0 \rangle| S_n \, d\mu.$$ 

We apply the Hölder inequality to bound this from below, as follows:

$$A_{z_0} = \frac{1}{2} \int_{S^n} |\langle z, z_0 \rangle| S_n \, d\mu$$

$$\geq \frac{1}{2} \left( \int_{S^n} S_n^{1-\alpha} \, d\mu \right)^{1/(1-\alpha)} \left( \int_{S^n} |\langle z, z_0 \rangle|^{1-1/\alpha} \, d\mu \right)^{\alpha/(\alpha-1)}$$

provided $\alpha \in (0, 1)$. The integral $\int_{S^n} |\langle z, z_0 \rangle|^\beta \, d\mu$ is bounded for $\beta > -1$. Hence for $\alpha \in (1/2, 1)$ we have

$$A_{z_0} \geq C \left( \int_{S^n} S_n^{1-\alpha} \, d\mu \right)^{1/(1-\alpha)} = CV^{n/(n+1)} Z_{z, \alpha}^{-1}.$$ 

For $\alpha \geq 1$ this inequality still holds, because by the Hölder inequality $Z_{z, \alpha}$ is increasing in $\alpha$. 
Finally, we consider the case where $\alpha \in (1/(n+2), 1/2]$: The Hölder inequality gives
\[
Z_{\alpha,\alpha} \geq Z_{\frac{n+1}{3}(\frac{3}{4}-\alpha)}^{-\frac{\alpha-1/(n+2)}{\alpha-1/(n+2)}} Z_{\frac{n+1}{3}(\frac{3}{4}-\alpha)}^{-\frac{n-1}{n+2}}.
\]

The affine isoperimetric inequality (see [B], §26 and §73, [Sa] or [A4], Theorem 7.1) implies that $Z_{\frac{n+1}{3},1/(n+2)} \geq 1$. Hence by applying the bound on $A_{z_0}$ in terms of $Z_{\frac{n+1}{3},1}$, which we know from the cases treated above, we have
\[
A_{z_0} \geq CV^{n/(n+1)} Z_{\frac{n+1}{3},1}^{-1} \geq CV^{n/(n+1)} Z_{\frac{n+1}{3},1}^{-\frac{n-1}{n+2}}.
\]

Hence for each $\alpha \in (1/(n+2), \infty)$ we have $A_{z_0} \geq C Z_{\frac{n+1}{3},1}^{-\beta(\alpha)}$ for some constant $C$ and some positive exponent $\beta(\alpha)$.

Next we deduce a bound on the maximum width of $M$ (the largest distance between parallel supporting hyperplanes): Let $z_0$ be the normal direction of a pair of parallel supporting hyperplanes for $M$ at maximal separation. Then the points of contact of $M$ with these two planes are joined by a segment with length equal to the maximum width of $M$, and which is entirely contained in $M$. Choosing the origin to be at the centre of this segment, we have $h(z_0) = h(-z_0) = w_+/2$ where $w_+$ is the maximum width of $M$, and $h(z) \geq |\langle z, z_0 \rangle h(z_0)|$ for all $z \in S^n$. But then the enclosed volume of $M$ is computed by:
\[
\text{Vol}(M) = \frac{1}{n+1} \int_{S^n} hS_n d\mu \geq \frac{1}{n+1} \int_{S^n} h(z_0) |\langle z, z_0 \rangle| S_n d\mu = \frac{w_+}{2(n+1)} A_{z_0}.
\]

Hence $w_+ \leq 2(n+1) \text{Vol}(M)/A_{z_0} \leq C Z_{\frac{n+1}{3},1}^{-\beta(\alpha)} \text{Vol}(M)^{1/(n+1)}$.

Note that $\text{Vol}(M) \leq w_- w_+^n$ where $w_-$ is the minimum width of $M$, since $M$ in contained between $(n+1)$ pairs of parallel planes in any set of orthonormal directions; and, in particular, in the case where one of the pairs of planes is at minimal separation. Then the separation of all the other pairs is bounded by $w_+$. It follows that $w_- \geq Vw_-^n \geq C Z_{\frac{n+1}{3},1}^{-n\beta(\alpha)} \text{Vol}(M)^{1/(n+1)}$.

This gives a bound on the ratio of the minimum and maximum widths of the hypersurface, and this is sufficient to bound the isoperimetric ratio (see for example [A1], Lemma 5.4).

Finally, we consider the anisotropic cases $\rho \neq \text{const.}$: For $\alpha < 1$, $Z_{\rho,\alpha}$ is comparable to $Z_{\frac{n+1}{3},1}$:
\[
\inf_{S^n} \rho^{1/(1-\alpha)} Z_{\frac{n+1}{3},1} \leq Z_{\rho,\alpha} \leq \sup_{S^n} \rho^{1/(1-\alpha)} Z_{\frac{n+1}{3},1}.
\]

For $\alpha \geq 1$ the desired inequality results from the monotonicity of $Z_{\rho,\alpha}$ as a function of $\alpha$, a consequence of the Hölder inequality. \qed
We have shown in particular that for $\alpha > 1/(n+2)$, any solution of Equation (6) with smooth, strictly convex initial data has uniformly bounded isoperimetric ratio on the entire interval of its existence.

5. Displacement and speed bounds.

In this section we prove that the ratio of the maximum and minimum values of the speed remains uniformly bounded for as long as the solution exists.

We will first deduce upper bounds on the displacement of the hypersurfaces, by using spheres enclosed within $M_0$ as barriers:

**Theorem 5.** For any $\alpha > 0$ and smooth positive $\rho$, and any smooth, strictly convex solution $\{M_t\}_{t>0}$ of Eq. (13),

$$h(z,t) \geq h(z,0) - C \frac{r_+(M_0)}{r_-(M_0)} t^{\frac{1}{1+n\alpha}}$$

for all $t \in (0, Cr_-(M_0)^{1+n\alpha}]$ in the interval of existence of the solution, where $C$ and $C'$ depend only on $\alpha$ and $\rho$.

**Proof.** Choose the origin at the centre of a ball of radius $r_-(M_0)$ enclosed by $M_0$. Fix $z \in S^n$, and define for each $\varepsilon \in (0, 1]$ a sphere

$$S_\varepsilon = \{y \in \mathbb{R}^{n+1} : |y - (1-\varepsilon)\bar{x}(z)| = \varepsilon r_-(M_0)\}.$$

Then $S_\varepsilon$ is contained in the convex hull of $\bar{x}(z)$ and $B_{r_-(M_0)}(0)$, so by convexity is contained in $M_0$.

Any family of spheres of the form $S_{r(t)}(p)$ with $p \in \mathbb{R}^{n+1}$ and

$$r(t) = (r(0)^{1+n\alpha} - \sup \rho(1+n\alpha)t)^{1/(1+n\alpha)}$$

satisfies $\langle \dot{x}, \nu \rangle \leq -\rho K^\alpha$, and hence act as barriers for solutions of Eq. (6), by the comparison principle.

This gives an estimate on the support function of $M_t$ in direction $z$: $S_\varepsilon$ produces a barrier which shrinks to its centre at time $\varepsilon^{1+n\alpha} r_-(M)^{1+n\alpha}/((1+n\alpha) \sup \rho)$, and we have

$$h(z,t) - h(z,0) \geq -\left(\frac{(1+n\alpha) \sup \rho t}{r_-(M)}\right)^{\frac{1}{1+n\alpha}} h(z,0) \geq -C(\alpha, \rho) \frac{r_+(M_0)}{r_-(M_0)} t^{\frac{1}{1+n\alpha}}.$$
Theorem 6. For any smooth, strictly convex solution \( \{M_t\}_{[0,T]} \) of Eq. (13) with \( R_- \leq r_-(M_t) \leq r_+(M_t) \leq R_+ \) for \( t \in [0,T] \),
\[
\rho(z)S_n^{-\alpha} \leq C(n, \alpha, \rho) \left( R_-^{-\alpha} + \left( \frac{R_+}{R_-} \right)^{\frac{\alpha n}{1+n\alpha}} \right).
\]

Proof. From the definition (10) and the evolution Equation (13) we obtain
\[
\frac{\partial}{\partial t} r_{ij} = - \left( \bar{\nabla}_i \bar{\nabla}_j (\rho S_n^{-\alpha}) + \bar{g}_{ij} \rho S_n^{-\alpha} \right).
\]
Since \( S_n = \det r_{ij} \), this implies
\[
\frac{\partial}{\partial t} (\rho S_n^{-\alpha}) = \alpha \rho S_n^{-1} \dot{S}_n \left( \bar{\nabla}_k \bar{\nabla}_l (\rho S_n^{-\alpha}) + \bar{g}_{kl} \rho S_n^{-\alpha} \right).
\]
We also have
\[
\frac{\partial}{\partial t} h(z,t) = - \rho S_n^{-\alpha} = \alpha \rho S_n^{-1} \dot{S}_n \left( \bar{\nabla}_k \bar{\nabla}_l h + \bar{g}_{kl} h \right) - (1+n\alpha) \rho S_n^{-\alpha}.
\]
Combining Eqs. (17) and (18), we obtain for \( q = \frac{\rho S_n^{-\alpha}}{h-R_-/2} \)
\[
\frac{\partial}{\partial t} q = \alpha \rho S_n^{-1} \dot{S}_n \left( \bar{\nabla}_k \bar{\nabla}_l q + \frac{2\alpha \rho S_n^{-1} \dot{S}_n}{h-R_-/2} \bar{\nabla}_k h \bar{\nabla}_l q \right)
- q^2 \left( \alpha R_- H/2 - (1+n\alpha) \right)
\]
where \( H = \sum_{i=1}^n r_i^{-1} = n S_n^{-1}/S_n \geq n S_n^{-1/n} \). By the maximum principle, this implies the following inequality for \( Q = \sup_{S_n} q \):
\[
\frac{dQ}{dt} \leq -Q^2 \left( C(n,\alpha,\rho) R_- R_+^{\frac{1}{1+n\alpha}} Q^\frac{1}{1+n\alpha} - (1+n\alpha) \right)
\]
and we deduce
\[
Q \leq \max \left\{ \frac{C(\alpha)}{R_-^{-\frac{n\alpha}{1+n\alpha}} R_+^{\frac{1}{1+n\alpha}} R_-^{-\frac{\alpha n}{1+n\alpha}} R_+^{\frac{1}{1+n\alpha}}}, C'(\alpha) R_-^{-\frac{n\alpha}{1+n\alpha}} t^{-\frac{n\alpha}{1+n\alpha}} \right\}.
\]
From the definition of \( Q \) and the estimate \( h \leq 2R_+ \), we have
\[
\rho S_n^{-\alpha} \leq \max \left\{ C(n,\alpha,\rho) R_-^{-\frac{n\alpha}{1+n\alpha}}, C'(n,\alpha,\rho) \left( \frac{R_+}{R_-} \right)^{\frac{\alpha n}{1+n\alpha}} t^{-\frac{n\alpha}{1+n\alpha}} \right\}.
\]
□

We now proceed to obtain lower bounds on the speed and displacement. It is in this estimate that we require \( \alpha \leq 1/n \). The argument combines barrier arguments with a Harnack inequality (proved for isotropic Gauss curvature flows by Chow [Ch3] and for more general flows by the author [A7]).
The Harnack estimate can be stated as follows for any solution of Equation (13) (see [A7], Theorem 5.6):

**Theorem 7.** For any smooth, strictly convex solution of (13) on \( S^n \times [0, T) \),
\[
\frac{d}{dt} \left( \rho S_n^{-\alpha t^{\frac{n\alpha}{n+1}}} \right) \geq 0
\]
everywhere on \( S^n \times (0, T) \).

Our lower speed estimate is the following:

**Theorem 8.** For \( \alpha < 1/n \) the following holds for any smooth, strictly convex solution of Eq. (13):
\[
h(z, t) \leq h(z, 0) - C(\rho, n, \alpha)r_+(M_0)^{-2n\alpha t^{1-n\alpha}}
\]
and
\[
\rho(z)S_n(z, t)^{-\alpha} \geq C'(\rho, n, \alpha)r_+(M_0)^{-2n\alpha t^{1-n\alpha}}
\]
for \( 0 < t < C^n(n, \rho, \alpha)r_+(M_0)^{1+n\alpha} \). For \( \alpha = 1/n \) we have instead the estimates for each \( \gamma > 0 \)
\[
h(z, t) \leq h(z, 0) - C(n, \rho, \gamma)r_+(M_0)^{1+2\gamma t^{-\gamma}}e^{-C'(\rho, n, \gamma)r_+(M_0)^{2n}t^{-n}}
\]
and
\[
\rho(z)S_n(z, t)^{-1/n} \geq C(n, \rho, \gamma)r_+(M_0)^{2\gamma - 1}t^{-\gamma}e^{-C'(\rho, n, \gamma)r_+(M_0)^{2n}t^{-n}}
\]
for \( 0 < t < C^n(n, \rho, \gamma)r_+(M_0)^{2} \).

**Proof.** For \( n = 1 \) these estimates are proved in [A2], Theorem II2.4. Suppose \( n \geq 2 \).

In the case \( \alpha < 1/n \) it suffices to use large spheres as barriers: Fix \( z \in S^n \). Then \( M_0 \) is enclosed in a hemispherical region obtained by intersecting the sphere of radius \( 2r_+(M_0) \) centred at \( \bar{x}(z) \) with the half-space \( \{ y \in \mathbb{R}^{n+1} : (y, z) \leq h(z) \} \). For any \( \varepsilon < 2r_+(M) \) this hemispherical region is enclosed by the sphere \( S_{\varepsilon} \) of radius \((\varepsilon^2 + 4r_+(M)^2)/(2\varepsilon)\) centred at the point \( \bar{x}(z) - (4r_+(M)^2 - \varepsilon^2)/(2\varepsilon)z \) (this sphere is chosen to have support function in direction \( z \) equal to \( h(z) + \varepsilon \)). We consider the evolution of these spheres for suitably small \( \varepsilon \): Since \( \rho \geq \inf_{S^n} \rho \), a sphere of radius \( r \) evolves in time \( t \) to be contained inside a sphere of radius \((r^{1+n\alpha} - (1+n\alpha)\inf \rho t)^{1/(1+n\alpha)}\) about the same centre, which is enclosed by the sphere of radius \( r - \inf \rho r^{-n\alpha}t \). In particular, this applies for each of the sphere \( S_\varepsilon \), and by the comparison principle \( M_t \) is also enclosed by this smaller sphere. This implies the inequality
\[
h(z, t) - h(z, 0) \leq \varepsilon - (\inf \rho) \left( \frac{\varepsilon^2 + 4r_+(M)^2}{2\varepsilon} \right)^{-n\alpha} t
\]
\[
\leq \varepsilon (1 - \inf \rho(4r_+(M)^2)^{-n\alpha} \varepsilon^{n\alpha-1} t).
\]
In particular, choosing
\[
\varepsilon = \left( \frac{t \inf \rho}{2^{1+2\alpha} r_+(M_0)^{2\alpha}} \right)^{1/(1-\alpha)}
\]
we obtain
\[
h(z, t) - h(z, 0) \leq - \left( \frac{\inf \rho}{2^{1+2\alpha} r_+(M_0)^{2\alpha}} \right)^{1/(1-\alpha)} t^{1/(1-\alpha)}
\]
for \( t \leq C_0 = C_0(\rho, n, \alpha) r_+(M_0)^{1+\alpha} \).

This estimate on the change in the support function can be converted to an estimate on the speed using the Harnack estimate from Theorem 7: Applying the estimate on the time interval \([t/2, t]\), we have
\[
Cr_+(M_0)^{-2^{\alpha} n^{\alpha+1}} \leq h(z, t/2) - h(z, t)
= \rho(z) \int_{t/2}^t S_n(z, \tau)^{-\alpha} d\tau
\leq t/2 \sup_{\tau \in [t/2, t]} (\rho S_n^{-\alpha}) .
\]

Theorem 7 then gives
\[
t^{\alpha/n+1} \rho(z) S_n(z, t)^{-\alpha} \geq (t/2)^{\alpha/n+1} \sup_{[t/2, t]} \rho S_n^{-\alpha}
\]
and hence
\[
\rho(z) S_n(z, t)^{-\alpha} \geq C' r_+(M_0)^{-2^{\alpha} n^{\alpha+1}} t^{1/(1-\alpha)-1} .
\]

In the case \(1/n\) the sphere barriers are not sufficient, and we instead work with graphical barriers. The displacement bound is a consequence of the following:

**Lemma 9.** Suppose \(\alpha = 1/n\). If \(M_0\) has bounded isoperimetric ratio, and lies in the region \(x_{n+1} \geq 0\) and within the ball \(\|x\| \leq R\), then \(M_t\) lies inside the region
\[
x_{n+1} \geq C_1 R^{1+2\gamma} t^{-\gamma} \left( e^{-C_2 R^{2n-2}(r-2R)^2 t^{-n}} + e^{-C_2 R^{2n-2}(r+2R)^2 t^{-n}} \right)
\]
for \(0 \leq t \leq C_3 R^2\), where \(r^2 = \sum_{i=1}^n x_i^2\), and \(C_1, C_2\) and \(C_3\) are constants depending only on \(n\), \(\rho\), and \(\gamma\).

**Proof.** We will show that the boundary of the region described is a graphical subsolution of the evolution Equation (14). In the special case of a radially symmetric function, we have
\[
K = \frac{u''(u')^{n-1}}{r^{n-1}(1 + (u')^2)^{(n+2)/2}} ,
\]
and so
\[(19)\]
\[
\dot{u} - \dot{\rho}(Du)K^{1/n}\sqrt{1 + |Du|^2} \leq \dot{u} - \inf \rho(u'')^{1/n} \left( \frac{u'}{r} \right)^{1-1/n} \left( 1 + (u')^2 \right)^{-1/n}.
\]

A direct computation shows that the function
\[
u(r, t) = C_1 R^{1-2\gamma} e^{-C_2 R^{2n-2}(r-2R)^2t^{-n}} + e^{-C_2 R^{2n-2}(r+2R)^2t^{-n}}
\]
makes the right-hand side of (19) non-positive on the region \( r < R, t \leq C_3 R^2 \), for any \( \gamma \geq 0 \), where \( C_1, C_2, \) and \( C_3 \) depend on \( n, \rho \) and \( \gamma \). Since the boundary of this region cannot intersect the hypersurface \( M_t \), the comparison principle applies.

This gives the bounds in the theorem, since we can rotate and translate the solution to bring the initial supporting hyperplane to the hyperplane \( x_{n+1} = 0 \), with \( M_0 \) satisfying the conditions of Lemma 9 with \( R = 2r_+(M_0) \). Thus \( h(z, 0) = 0 \). For positive sufficiently small \( t \), Lemma 9 gives
\[
h(z, t) = -\inf_{M_t} x_{n+1} \leq -u(0, t)
\]
as required.

Similar barriers can also be constructed for each \( \alpha < 1/n \).

The speed bound follows using Theorem 7 as for the previous cases.

We remark that the estimate for \( \alpha < 1/n \) does not rely at all on the particular structure of the Gauss curvature flows — the same result holds for any strictly parabolic flow with speed homogeneous of degree less than 1 in the curvatures. \( \square \)

6. Curvature control.

In this section we prove that the ratio of the maximum and minimum principal curvatures remains bounded throughout the evolution, given the upper and lower speed bounds of the previous section. Our argument is an application of the parabolic maximum principle to the evolution equation for the curvature.

In the case \( n = 1 \) the speed bounds above and below already give complete control on the curvatures. For the rest of this section we assume \( n \geq 2 \).

**Theorem 10.** Suppose \( h : S^n \times [0, T] \) is a solution of Eq. (13) for which the isoperimetric ratio is bounded and the speed is bounded above and below — that is, there exist constants \( C_1, C_2 \) such that
\[
0 < C_1 \leq S_n(z, t) \leq C_2
\]
for every \( z \in S^n \) and \( t \in [0, T] \). Then there exist positive constants \( C_3 \) and \( C_4 \) depending only on \( C_1, C_2, \rho, n \) and \( \alpha \) such that
\[
\lambda_i(z, t) \geq \min\{C_3 t^{n-1}, C_4\}
\]
for all \( i \in \{1, \ldots, n\} \), \( z \in S^n \), and \( t \in [0, T] \).

**Proof.** We begin by computing the evolution equation for the matrix \( r_{ij} = \nabla_i \nabla_j h + \bar{g}_{ij} h \) under Eq. (13):

\[
\frac{\partial}{\partial t} r_{ij} = -\nabla_i \nabla_j (\rho S_n^{-\alpha}) - \bar{g}_{ij} \rho S_n^{-\alpha}
\]

\[
= \alpha \rho S_n^{-(1+\alpha)} \tilde{s}^{kl}_n \nabla_i \nabla_k \tau_{kl} - \alpha (1 + \alpha) \rho S_n^{-(2+\alpha)} \nabla_i S_n \nabla_j S_n
\]

\[
+ \alpha \rho S_n^{-(2+\alpha)} \tilde{s}^{klmn}_n \nabla_i \tau_{kl} \nabla_j \tau_{mn} - \bar{g}_{ij} \rho S_n^{-\alpha} - S_n^{-\alpha} \nabla_i \nabla_i \rho
\]

\[
+ \alpha S_n^{-(1+\alpha)} \nabla_i \rho \nabla_j S_n + \alpha S_n^{-(1+\alpha)} \nabla_j \rho \nabla_i S_n.
\]

In the first term here we apply the identity (12), to yield:

\[
\frac{\partial}{\partial t} r_{ij} = \alpha \rho S_n^{-(1+\alpha)} \tilde{s}^{kl}_n \nabla_i \nabla_k \tau_{ij} - \alpha (1 + \alpha) \rho S_n^{-(2+\alpha)} \nabla_i S_n \nabla_j S_n
\]

\[
+ \alpha \rho S_n^{-(2+\alpha)} \tilde{s}^{klmn}_n \nabla_i \tau_{kl} \nabla_j \tau_{mn}
\]

\[
+ (n \alpha - 1) \rho S_n^{-\alpha} \bar{g}_{ij} - \alpha \rho S_n^{-(1+\alpha)} \tilde{s}^{kl}_n \bar{g}_{ij}
\]

\[
+ S_n^{-\alpha} \nabla_i \nabla_j \rho + \alpha S_n^{-(1+\alpha)} \nabla_i \rho \nabla_j S_n + \alpha S_n^{-(1+\alpha)} \nabla_j \rho \nabla_i S_n.
\]

We wish to obtain an upper bound for the eigenvalues of \( r_{ij} \), so the second term on the first line and the last term on the second line are good terms since they are negative. The first term of the first line is an elliptic operator, and so is non-positive at a point and direction where a maximum eigenvalue occurs. The first term of the second line we estimate using the concavity of the \( n \)th root of the determinant as a function of the components of the matrix, which is equivalent to the inequality

\[
\left( \tilde{s}^{klmn}_n - \frac{n-1}{n S_n} \tilde{s}^{kl}_n \tilde{s}^{mn}_n \right) \xi_{kl} \xi_{mn} \leq 0
\]

for any symmetric matrix \( \xi \). Finally, the first term on the last line is bounded, and the other two terms of the last line can be estimated in terms of the good second term of the first line:

\[
| \nabla_i \rho \nabla_i S_n | \leq C \varepsilon | \nabla_i S_n |^2 + C \varepsilon^{-1}
\]

for any \( \varepsilon > 0 \). Combining these estimates, we obtain

\[
\frac{\partial}{\partial t} r_{ij} \leq \alpha \rho S_n^{-(1+\alpha)} \tilde{s}^{kl}_n \nabla_i \nabla_k \tau_{ij} + C S_n^{-\alpha} \bar{g}_{ij} - \alpha \rho S_n^{-(1+\alpha)} \tilde{s}^{kl}_n \bar{g}_{ij} \tau_{ij}.
\]

The last term here will allow us to obtain an estimate independent of initial data: We have \( S_n \bar{g}_{kl} = S_{n-1} / S_n \), and the Newton inequalities \([\text{Mi}]\) give

\[
S_{n-1} \geq S_n^{\frac{n-2}{n-1}} S_1^{\frac{1}{n-1}} \geq C S_n^{\frac{n-2}{2}} \xi_{\text{max}}.
\]
Now work at a point and time where a maximum eigenvalue is attained, and suppose \( i = j \) and \( e_i \) is the eigenvector of \( r \) with the largest eigenvalue. Then the first term in the evolution equation is negative, and
\[
\frac{\partial r_{\text{max}}}{\partial t} \leq CS_{\cdot}^{-\alpha} \bar{g}_{ij} - CS_{\cdot}^{-\alpha} \left( \frac{n}{\pi - 1} \right)^{\frac{n}{n-1}} \frac{n}{n} \leq CS_{\cdot}^{-\alpha} \left( \bar{g}_{ij} - CS_{\cdot}^{-\alpha} \left( \frac{n}{\pi - 1} \right)^{\frac{n}{n-1}} \right).
\]
Given the bound below on \( S_{\cdot}^{-\alpha} \), the bracket is negative provided \( r_{\text{max}} \) is sufficiently large. For the same reason the coefficient in front of the bracket does not become small, and we have for \( r_{\text{max}} \) sufficiently large
\[
d \frac{d}{dt} r_{\text{max}} \leq -C \frac{n}{(n-1)}.\]
The result now follows by the parabolic maximum principle and comparison with the solution of the ordinary differential equation \( du/dt = -C u^{n/(n-1)} \).

Next we observe that this automatically provides an upper bound on the principal curvatures:

**Proposition 11.** Let \( W \) be a positive definite symmetric matrix for which \( W \geq \varepsilon \text{Id} \) and \( \det W \leq C \). Then \( W \leq C \varepsilon^{-n/(n-1)} \text{Id} \).

**Proof.** Number the eigenvalues \( \lambda_1 \) of \( W \) in ascending order: \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \). Then
\[
\lambda_n = \frac{K}{\lambda_1 \lambda_2 \cdots \lambda_{n-1}} \leq \frac{K}{\lambda_1^{n-1}},
\]
where \( K = \lambda_1 \cdots \lambda_n = \det W \). \( \square \)

In particular, the upper speed bound of Theorem 6 and the lower curvature bound of Theorem 10 imply an upper curvature bound:

**Corollary 12.** Under the conditions of Theorem 10 there exist constants \( C_5 \) and \( C_6 \) such that
\[
W \leq \max\{C_5 t^{-(n-1)^2}, C_6\} \text{Id}.
\]

### 7. Convergence to a point.

In this section we prove that any solution of Eq. (6) with a smooth, strictly convex initial hypersurface converges to a point in finite time. In the special case of isotropic flows (\( \rho \equiv 1 \)) this was proved by K.S. Chou [Ts] for \( \alpha = 1 \) and by Ben Chow [Ch1] for other \( \alpha \). While we only need the result for \( \alpha \leq 1/n \), we give a proof which works for larger \( \alpha \) as well.

**Theorem 13.** For any \( \alpha > 0 \) and positive \( \rho \in C^\infty(S^n) \), and any smooth, strictly convex hypersurface \( M_0 \subset \mathbb{R}^{n+1} \), the hypersurfaces \( M_t \) given by the solution of Eq. (6) exist for a finite time \( T \) and converge in Hausdorff distance to \( p \in \mathbb{R}^{n+1} \) as \( t \) approaches \( T \).
Proof. The maximal time of existence must be finite: By the comparison principle, if $M_0$ is enclosed by a sphere $S^n_{r(0)}(q)$ for some $r > 0$ and $q \in \mathbb{R}^{n+1}$, then for all $t$ in the interval of existence, $M_t$ is enclosed by the sphere $S^n_{r(t)}(q)$, where

$$r(t) = \left(r(0)^{1+n\alpha} - (\inf \rho)(1 + n\alpha)t\right)^{1/(1+n\alpha)}.$$

$r(t)$ converges to zero in finite time, and $M_t$ cannot exist beyond this time.

Consider again the estimate (22) for the evolution of the curvature. We also have the evolution equation

$$\frac{\partial}{\partial t} h = -\rho S_n^{-\alpha}$$

$$= \alpha \rho S_n^{-1(1+\alpha)} \tilde{S}_n^{kl} \tilde{\nabla}_k \tilde{\nabla}_l h + \alpha \rho S_n^{-1(1+\alpha)} \tilde{S}_n^{kl} \tilde{g}_{kl} h - (1 + n\alpha) \rho S_n^{-\alpha}.$$

Combining these, we obtain

$$\frac{\partial}{\partial t} (r_{ij} + Ah \tilde{g}_{ij}) \leq \alpha \rho S_n^{-1(1+\alpha)} \tilde{S}_n^{kl} \tilde{\nabla}_k \tilde{\nabla}_l (r_{ij} + Ah \tilde{g}_{ij})$$

$$+ (C - A(1 + n\alpha)) S_n^{-\alpha} \tilde{g}_{ij}$$

$$+ (Ah - r_{ij}) \alpha \rho S_n^{-1(1+\alpha)} \tilde{S}_n^{kl} \tilde{g}_{kl}.$$

Choose $A = C/(1 + n\alpha)$, so that the last term of the first line vanishes. Also note that since the hypersurfaces are contracting, we have $h \leq h_0 = \sup_{S^n} h(z, 0)$ as long as the solution exists. Therefore we have, writing $q_{ij} = r_{ij} + Ah \tilde{g}_{ij}$,

$$\frac{\partial}{\partial t} q_{ij} \leq \alpha \rho S_n^{-1(1+\alpha)} \tilde{S}_n^{kl} \tilde{\nabla}_k \tilde{\nabla}_l q_{ij} - \alpha \rho S_n^{-1(1+\alpha)} \tilde{S}_n^{kl} \tilde{g}_{kl} (q_{ij} - 2h_0 \tilde{g}_{ij})$$

and hence by the parabolic maximum, the maximum eigenvalue of $q_{ij}$ is decreasing if it is larger than $2h_0$. Since the initial hypersurface is smooth and strictly convex, $q_{ij}$ is bounded at $t = 0$. Therefore we have a uniform bound on $q_{ij}$ and hence $r_{ij}$ throughout the interval of existence.

Suppose the inradius of the hypersurfaces $M_t$ do not converge to zero — that is, the solution exists for a maximal time interval $[0, T)$, but there is some ball of positive radius that remains enclosed by the solution throughout. By the argument in Section 5, the speed remains bounded throughout the interval of existence.

By Proposition 11 this also implies a bound on the curvature, so that the hypersurfaces remain uniformly smooth and strictly convex on the time interval $[0, T)$. It follows that there exists a subsequence of times $\{t_k\}$ converging to $T$ such that $M_{t_k}$ converges in $C^\infty$ to a smooth, strictly convex limit $M_T$. Furthermore, the $C^\infty$ convergence implies that all time derivatives converge, so that in fact $M_t$ approaches $M_T$ in $C^\infty$ as $t$ approaches
Hence we have a smooth solution on \([0, T]\), and the short-time existence result implies that this can be extended beyond \(T\), contradicting the assumption that \(T\) was maximal.

Therefore the inradius converges to zero. Since \(r_{ij}\) is uniformly bounded, this also implies that the circumradius converges to zero, and the hypersurfaces converge to a point. \(\square\)

Theorem 13 implies a bound below on the time of existence of solutions with smooth and strictly convex initial data, in terms of the inradius of the initial hypersurface, \(\rho\), and \(\alpha\): Any sphere which is initially enclosed by the hypersurface acts as a barrier, preventing the hypersurface from contracting to a point too quickly.

8. Short-time existence.

In this section we prove the following:

**Theorem 14.** For any positive \(\rho \in C^\infty(S^n)\), \(\alpha \leq 1/n\), and open bounded convex region \(\Omega_0 \subset \mathbb{R}^{n+1}\), there exists a smooth, strictly convex solution \(x_t : S^n \times (0, T)\) of Eq. (6) which converges to \(M_0 = \partial \Omega_0\) in Hausdorff distance as \(t\) approaches zero. Any other such solution \(y_t : S^n \times (0, T')\) is given by \(x_t \circ \varphi\) for some smooth diffeomorphism \(\varphi\) of \(S^n\).

At the end of the section we also prove the existence and uniqueness of viscosity solutions for arbitrary convex initial data and arbitrary \(\alpha > 0\).

8.1. Existence. In order to construct a solution which approaches \(M_0\) at the initial time, we consider a family of smooth, strictly convex hypersurfaces \(M_0^{(\varepsilon)}\) which approach \(M_0\) in Hausdorff distance as \(\varepsilon\) approaches zero. By Theorem 13, for each \(\varepsilon > 0\) there exists a unique solution \(M_t^{(\varepsilon)}\) of Eq. (6) with initial condition \(M_0^{(\varepsilon)}\), which converges to a point in finite time \(T_\varepsilon > 0\).

Then we have

\[
\begin{align*}
d_H(M_0, M_t^{(\varepsilon)}) &\leq d_H(M_0, M_0^{(\varepsilon)}) + d_H(M_0^{(\varepsilon)}, M_t^{(\varepsilon)}) \\
&\leq d_H(M_0, M_0^{(\varepsilon)}) + Ct^{1/(1+n\alpha)}.
\end{align*}
\]

By Theorem 10 and Corollary 12, the hypersurfaces \(M_t^{(\varepsilon)}\) satisfy bounds above and below on the principal curvatures, uniformly in \(\varepsilon\) over every compact subset of \(S^n \times (0, T)\). It follows from the regularity theory for solutions of uniformly parabolic equations concave in the second derivatives ([K], Theorem 5.5) that there are also bounds on all higher derivatives of the curvatures, uniformly in \(\varepsilon\) over every compact subset of \(S^n \times (0, T)\). It follows from the Arzela-Ascoli theorem that there exists a sequence \(\varepsilon_k\) approaching zero such that \(\{M_t^{(\varepsilon_k)}\}\) converges in \(C^\infty\) to a family of hypersurfaces \(\{M_t\}\) satisfying the same bounds. In particular, \(\{M_t\}\) satisfies Eq. (6) on \(S^n \times \)
(0, T), $M_t$ is smooth and strictly convex for each $t > 0$, and $d_H(M_0, M_t) \leq C t^{1/(1+n\alpha)}$.

8.2. Uniqueness. Suppose we have two solutions $\{M^{(1)}_t\}$ and $\{M^{(2)}_t\}$ of Eq. (6), both converging to $M_0$ in Hausdorff distance as $t \to 0$, and denote by $h^{(i)}_t$ the corresponding support functions. Fix $\varepsilon > 0$. Then there exists $t_0(\varepsilon) > 0$ such that $|h^{(i)}_t(z) - h_0(z)| < \varepsilon$ for $i = 1, 2$ and all $z \in S^n$ and $t \in (0, t_0(\varepsilon))$. Choose a smooth, strictly convex hypersurface $M^{(\varepsilon)}_0$ with support function $h^{(\varepsilon)}_0$ such that $h_0(z) - 2\varepsilon < h^{(\varepsilon)}_0(z) < h_0(z) - \varepsilon$. Without loss of generality we assume that the origin is at the centre of a ball of radius $r_-(M_0)$ enclosed by $M_0$. Then $r_-(M_0) \leq h_0(z) \leq 2r_+(M_0)$. It follows that

$$h^{(\varepsilon)}_0(z) < h_0(z) - \varepsilon < h^{(i)}_t(z) < h_0(z) + \varepsilon < \left(1 + \frac{3\varepsilon}{r_-(M_0) - 2\varepsilon}\right) h^{(\varepsilon)}_0$$

for $i = 1, 2$ and all $z \in S^n$ and $t \in (0, t_0(\varepsilon))$. The comparison principle and the scaling property given by Eq. (15) then imply

$$h^{(\varepsilon)}_\tau(z) < h^{(i)}_{t+\tau}(z) < (1 + \lambda)h^{(\varepsilon)}_{(1+\lambda)^{-1(1+n\alpha)\tau}}(z)$$

for all $\tau \geq 0$ for which these all exist, where $\lambda = \frac{3\varepsilon}{r_-(M_0) - 2\varepsilon}$. Consequently,

$$\left|h^{(2)}_{t+\tau}(z) - h^{(1)}_{t+\tau}(z)\right| \leq (1 + \lambda)h^{(\varepsilon)}_{(1+\lambda)^{-1(1+n\alpha)\tau}}(z) - h^{(\varepsilon)}_\tau(z)$$

$$\leq \lambda h^{(\varepsilon)}_{(1+\lambda)^{-1(1+n\alpha)\tau}}(z) + \left(h^{(\varepsilon)}_{(1+\lambda)^{-1(1+n\alpha)\tau}}(z) - h^{(\varepsilon)}_\tau(z)\right)$$

$$\leq 2r_+(M_0)\lambda + C \left(1 - (1 + \lambda)^{-1(1+n\alpha)}\right) \tau^{1/(1+n\alpha)}$$

$$\leq C\lambda + C(\lambda\tau)^{1/(1+n\alpha)}.$$ 

Here we used Theorem 6 to obtain the second-last line. Now take $t \to 0$. Since $\varepsilon > 0$ is arbitrary and $C$ independent of $\varepsilon$, we have for each $\tau > 0$ and $z \in S^n$

$$h^{(2)}_\tau(z) = h^{(1)}_\tau(z).$$

Note that the proof presented here does not rely strongly on the particular structure of the evolution Equation (6). In particular, the uniqueness argument is valid for any flow by a monotone, positively homogeneous function of curvature, since the bound on the change in the support function given in Theorem 6 also holds for all such evolution equations. The existence argument requires a speed bound and regularity estimates independent of initial data.

We now proceed to the case $\alpha > 1/n$: In this case (as we show in Section 12) one cannot expect to produce smooth solutions from arbitrary convex initial hypersurfaces. Instead we will work with a weaker notion of solution: A family of convex regions $\{\Omega_t\}_{0 < t < T}$ is called a viscosity solution of Eq. (6)
if the following conditions hold: First, for any smooth, strictly convex hypersurface $M_0$ contained in $Ω_{t_0}$ for some $t_0 ∈ (0, T)$, the hypersurfaces $M_t$ given by solving (6) are contained in $Ω_{t_0+t}$ for all $t ∈ [0, T − t_0)$ in the domain of existence of the $M_t$. Second, for any smooth, strictly convex hypersurface $M_0$ which encloses $Ω_{t_0}$ for some $t_0 ∈ (0, T)$, the hypersurfaces $M_t$ enclose $Ω_{t_0+t}$ for all $t ∈ [0, T − t_0)$.

**Theorem 15.** For any smooth positive $ρ ∈ C^∞(S^n)$, $α > 0$, and any open bounded convex region $Ω_0 ⊂ \mathbb{R}^{n+1}$, there exists a unique viscosity solution $\{Ω_t\}_{0 < t < T}$ which converges to $Ω_0$ in Hausdorff distance as $t$ approaches zero. $Ω_t$ converges to a point as $t$ approaches $T$.

**Proof.** We use the same construction as presented in the proof of Theorem 14, producing a solution $\{M_t^{(ε)}\}$ for each $ε > 0$, with $M_0^{(ε)}$ approaching $∂Ω_0$ in Hausdorff distance as $ε$ approaches zero. We specify further that $M_0^{(ε)}$ is contained in $Ω_0$ for all $ε > 0$, and is increasing in $ε$.

For $ε$ sufficiently small, we can choose an origin for $\mathbb{R}^{n+1}$ and radii $R > r > 0$ such that the ball $B_r(0)$ is enclosed by all of the hypersurfaces $M_0^{(ε)}$, and the ball $B_R(0)$ contains all of the hypersurfaces $M_0^{(ε)}$. By the comparison principle, there exists $δ > 0$ such that the ball $B_{r/2}(0)$ is enclosed by all the hypersurfaces $M_t^{(ε)}$ for $t ∈ [0, δ]$. The hypersurfaces also remain enclosed by the ball $B_R(0)$.

It follows that the support functions $h_t^{(ε)}(z)$ are uniformly Lipschitz: By the Formula (9), we have $|\bar{x}|^2 = h^2 + |\nabla h|^2$, and re-arrangement gives $|\nabla h|^2 ≤ |\bar{x}|^2 ≤ R^2$, which is a uniform Lipschitz bound.

Furthermore, the displacement bound and the speed bound of Theorem 6 show that $h_t^{(ε)}$ is Hölder continuous in $t$, uniformly in $ε$ and $z$, and also uniformly Lipschitz on compact subsets of $(0, δ)$. Therefore $h(ε)(z, t)$ is a Hölder continuous function on $S^n × [0, δ)$, uniformly in $ε$. By the Arzela-Ascoli theorem, there exists a sequence $ε_k$ approaching zero which converges to a limit $h(z, t)$ satisfying the same estimates. By the Blaschke selection theorem, each of the functions $h_t = h(., t)$ is the support function of a convex region $Ω_t$, and the same argument as in the proof of Theorem 14 shows that $Ω_t$ approaches $Ω_0$ in Hausdorff distance as $t$ approaches zero.

We need to prove that the family $\{Ω_t\}$ is a viscosity solution. The first condition is easily checked: If $M_0''$ is contained within $Ω_0$, then $M_0''$ is also enclosed by $M_0^{(ε)}$ for $ε > 0$ sufficiently small. By the comparison principle, the resulting solution $M_t''$ is enclosed by $M_t^{(ε)}$ for $t > 0$, and also $M_t^{(ε)}$ is increasing in $ε$ and converges to $∂Ω_t$ as $ε$ approaches zero. Therefore $M_t''$ is contained in $Ω_t$ for $t > 0$. 

The second condition also follows easily: Any hypersurface $M'_0$ which
encloses $\Omega_0$ also encloses all of the hypersurfaces $M_0^{(\varepsilon)}$, and so by the
comparison principle $M'_t$ encloses $M_t^{(\varepsilon)}$ for $\varepsilon > 0$ and $t > 0$, and so also encloses
the limit $\partial \Omega_t$.

The uniqueness statement in Theorem 15 follows exactly as in the proof
of uniqueness in Theorem 14, and the same argument shows that the regions
$\Omega_t$ converge to a point. □


In this section we apply the speed and curvature bounds of the previous
sections to give a short proof of the following theorem:

**Theorem 16.** Let $\alpha = 1/(n + 2)$ and $\rho \equiv 1$. For any convex open region
$\Omega_0 \subset \mathbb{R}^{n+1}$ there exists a smooth family of strictly convex embeddings $x_t : S^n \to \mathbb{R}^{n+1}$ satisfying Eq. (6) for which the Hausdorff distance between
$M_t = x_t(S^n)$ and $M_0$ approaches zero as $t \to 0$. Any other such solution $\{\tilde{x}_t\}$
is related to $\{x_t\}$ by composition with a time-independent diffeomorphism.
$M_t$ converges to a point $p \in \mathbb{R}^{n+1}$ as $t$ approaches a finite time $T$, and

$$\tilde{M}_t = \left( \frac{\text{Vol}(S^n)}{\text{Vol}(M_t)} \right)^{1/(n+1)} (M_t - p)$$

converges in $C^\infty$ to an ellipsoid centred at the origin.

This theorem was proved in the case of smooth, strictly convex $M_0$ in
[A4]. The results of Section 6 allow us to give a proof which works also
for singular initial hypersurfaces. The argument is also considerably simpler
because it avoids the complicated third-derivative estimate which was the
key to the proof in [A4]. On the other hand, we use the result that elliptic
affine hyperspheres are ellipsoids, which was not necessary for the proof in
[A4].

**Proof.** By Section 8, we have a unique solution of Eq. (6) with the given
initial condition. Since this is smooth and strictly convex for $t > 0$, the
result of Theorem 13 implies that this solution converges to $p \in \mathbb{R}^{n+1}$ in
finite time $T$.

In the proof we use the fact that the evolution equation is invariant under
the action of the special affine group: If $\{M_t\}$ is a family of hypersurfaces
moving under Eq. (6), then $\{L(M_t)\}$ is also such a family, for any special
affine transformation $L$.

Fix $t \in [0, T)$. There exists a special affine transformation $L_t$ such that
$L_t(M_t)$ has

$$C_- \text{Vol}(M_t)^{1/(n+1)} \leq \text{Vol}(S^n)^{1/(n+1)} h_{L_t(M_t)}(z) \leq C_+ \text{Vol}(M_t)^{1/(n+1)},$$

for some absolute constants $C_\pm$. 
We consider the solution \( \tilde{h} \) given by the scaling relation (15):

\[
\tilde{h}(z, \tau) = \left( \frac{\text{Vol}(S^n)}{\text{Vol}(M_t)} \right)^{1/(n+1)} h_L \left( z, t + \left( \frac{\text{Vol}(S^n)}{\text{Vol}(M_t)} \right)^{-2/(n+2)} \tau \right),
\]

where \( h_L \) is the support function of the convex body obtained by applying the special affine transformation \( L \) to the body with support function \( h \). Then \( C_- \leq \tilde{h}(z, 0) \leq C_+ \). By the comparison principle we also have \( \frac{1}{2} C_- \leq \tilde{h}(z, \tau) \leq C_+ \) for \( \tau \in [0, \delta] \), where \( \delta = (C_-)^2/(n+2)(1 - 2^{-2(n+1)/(n+2)})(n+2)/2(n+1) \). Hence on the interval \([\delta/2, \delta]\) there are uniform speed and displacement bounds (by Theorem 6), a uniform lower bound on the speed (by Theorem 8), uniform bounds above and below on the principal curvatures (by Theorem 10 and Corollary 12), and uniform bounds on all higher derivatives of the curvature (by Theorem 5.5 of \([K]\)).

It follows that the original solution satisfies uniform bounds on all quantities which are both scaling invariant and special-affine invariant, on the time interval \([t + \frac{1}{2} C \text{Vol}(M_t)^{2/(n+2)}, t + C \text{Vol}(M_t)^{2/(n+2)}]\), for some absolute constant \( C \). Since \( t \) is arbitrary, we have such bounds on the entire interval \([T/2, T]\).

Therefore there exists a sequence \( \{t_k\} \) approaching \( T \), and a sequence of special affine transformations \( \{L_k\} \), such that the hypersurfaces \( \{L_k(M_{t_k})\} \) converge in \( C^\infty \) to a smooth, strictly convex limit \( \tilde{M}_T \).

Suppose \( \tilde{M}_T \) does not satisfy the condition \( K^{1/(n+2)} = c \langle x, \nu \rangle \) for some \( c > 0 \) and some choice of origin. Then the time derivative of \( Z_{\frac{n}{2}, 1/(n+2)} \) on \( \tilde{M}_T \) is strictly negative, by Theorem 3. By the \( C^\infty \) convergence and scaling, there exists \( k_0, \delta > 0, \) and \( \varepsilon > 0 \) such that whenever \( k \geq k_0 \) we have

\[
Z_{\frac{n}{2}, 1/(n+2)}(M_{t_k} + \delta \text{Vol}(M_{t_k})^{2/(n+2)}) \leq Z_{\frac{n}{2}, 1/(n+2)}(M_{t_k}) - \varepsilon.
\]

But since \( Z_{\frac{n}{2}, 1/(n+2)} \) is non-increasing, this would imply \( Z_{\frac{n}{2}, 1/(n+2)}(\tilde{M}_{t_k}) \to -\infty \) as \( k \to \infty \), which is impossible. Therefore \( \tilde{M}_T \) satisfies the required condition.

By Theorem 1 of \([Ca]\), a smooth, strictly convex hypersurface satisfies the condition \( K^{1/(n+2)} = c \langle x, \nu \rangle \) if and only if it is an ellipsoid.

The stronger convergence statements in the Theorem follow by considering the linearization of the evolution Equation (13) about the space of ellipsoids — a direct calculation shows that this space is strictly linearly stable, so Proposition 9.2.3 of \([Lu]\) and a scaling argument implies that \( M_t \) converges in \( C^\infty \) to the ellipsoid \( \tilde{M}_T \) after rescaling. The details of this argument are given for a related evolution equation in \([A10]\), Propositions 40-41. \( \square \)
Corollary 17. For any open bounded convex region \( \Omega_0 \subset \mathbb{R}^{n+1} \), the following generalised affine isoperimetric inequality holds:

\[
\lim_{t \to 0} \mathcal{Z}_n,\frac{1}{n+2}(\Omega_t) \geq 1
\]

with equality if and only if \( \Omega_0 \) is an ellipse.

In effect this result allows a definition of the affine surface area for convex hypersurfaces which may be non-smooth or non-strictly convex, in such a way that the affine isoperimetric inequality remains true. A related extension of the affine surface area has been given in [Le].

10. Proof of the main Theorems.

In this section we complete the proofs of Theorems 1 and 2. The proof is similar to that presented in the special case of the affine normal flow in the previous section, but is somewhat simpler because we do not require the machinery of affine corrections to obtain bounded isoperimetric ratios.

Section 8 provides us with a unique solution \( \{M_t\} \) of Eq. (6) for a short time, and Theorem 13 ensures that this solution remains smooth until it converges to a point \( p \in \mathbb{R}^{n+1} \) in finite time \( T \). In the case \( \alpha > 1/(n + 2) \), Theorem 4 gives a uniform bound on the isoperimetric ratios of the hypersurfaces \( M_t \) throughout the interval \( [0, T] \). In the case described in Theorem 2, we also have such an estimate by hypothesis.

Theorems 6, 8 and 10 and Corollary 12 therefore imply uniform bounds above and below on the principal curvatures of the rescaled hypersurfaces \( \{\tilde{M}_t\} \) defined by rescaling to fixed enclosed volume about the point \( p \). Theorem 5.5 of [K] and Schauder estimates ([Li], Theorem 4.9) imply uniform bounds on all higher derivatives of curvature. It follows that there exists a subsequence of times \( \{t_k\} \) approaching \( T \) for which the hypersurfaces \( \tilde{M}_{t_k} \) converge in \( C^\infty \) to a smooth, strictly convex limit \( \tilde{M}_T \). By the same argument as that in Section 9, \( \tilde{M}_t \) satisfies the equation \( \langle x, \nu \rangle = c \rho K^\alpha \) for some \( c > 0 \). The convergence in \( C^\infty \) of \( \tilde{M}_t \) to \( \tilde{M}_T \) as \( t \) approaches \( T \) follows from Theorem 2 of [A6].


In this section we give an application of Theorem 2 to prove the existence of homothetically contracting solutions of the isotropic flows

\[
\frac{\partial}{\partial t} x = -K^\alpha \nu
\]

for sufficiently small \( \alpha \). The idea is to consider the evolution of hypersurfaces close to the sphere \( S^n \), possessing suitable symmetries. Precisely, our result is the following:
**Theorem 18.** Let $\Gamma$ be a proper subgroup of $SO(n+1)$ such that for every $z \in S^n$, the orbit of $z$ under $\Gamma$ spans $\mathbb{R}^{n+1}$ (that is, the inclusion of $\Gamma$ in $SO(n+1)$ is an irreducible representation). Let $\lambda$ be the smallest eigenvalue corresponding to a non-trivial $\Gamma$-invariant spherical harmonic $\varphi$. Then for $\alpha \in (0,1/(\lambda - n))$ there exists a non-spherical, $\Gamma$-symmetric, smooth, strictly convex hypersurface satisfying the identity $\langle x, \nu \rangle = K^\alpha$.

**Proof.** The linearization of the normalised isotropic Equation (6) about the sphere solution $h \equiv 1$ is given by

$$\frac{\partial}{\partial t} \eta = \alpha(\Delta + n)u + u$$

where $\Delta$ is the Laplacian on $S^n$. In particular, for $\alpha \in (0,1/(\lambda - n))$ the $h \equiv 1$ solution is strictly unstable in the direction $\varphi$. By [Lu], Theorem 9.1.3, there exists a $\Gamma$-symmetric solution $\{M_t\}$ of Eq. (6) which converges to $h \equiv 1$ as $t \to -\infty$ and diverges exponentially from $h \equiv 1$.

We observe that $\Gamma$-symmetry of a convex hypersurface implies a bound on the isoperimetric ratio:

**Lemma 19.** For any $\Gamma$ satisfying the conditions of Theorem 18, there exists a constant $C$ such that every $\Gamma$-symmetric convex hypersurface $M \subset \mathbb{R}^{n+1}$ satisfies $r_+(M)/r_-(M) \leq C$.

**Proof.** If this is not the case, then we can find a sequence of $\Gamma$-symmetric convex hypersurfaces $M_k$ such that $r_+(M_k) = 1$ and $r_-(M_k) \leq 1/k$. By the Blaschke selection theorem ([Sc], Theorem 2.5.14) we can choose a subsequence $M_{k'}$ which converges in Hausdorff distance to a limit $M_\infty$ which is again $\Gamma$-symmetric but has $r_-(M_\infty) = 0$ and $r_+(M_\infty) = 1$. It follows that $M_\infty$ is contained in a lower-dimensional subspace of $\mathbb{R}^{n+1}$. But this is impossible, because there exists $x \in M_\infty$ with $|x| = 1$; by $\Gamma$-symmetry all of the point $g(x)$ are in $M_\infty$, but these are not contained in any such sub-space.

Therefore we can apply Theorem 2 to the solution $\{M_t\}$, obtaining $C^\infty$ convergence to a limit $M_T$ satisfying the required identity (possibly after scaling to ensure $\kappa = 1$). $M_T$ is $\Gamma$-symmetric, and has $Z$ strictly less than that for the sphere solution, since $Z$ has strictly decreased along the solution $\{M_t\}$. Therefore $M_T$ is non-spherical.

**Corollary 20.** In the case $n = 1$, for each $k \geq 3$ and each $\alpha \in (0,1/(k^2 - 1))$ there exists a non-circular strictly convex smooth curve $C_k$ with $k$-fold symmetry which contracts homothetically under the flow

$$\frac{\partial}{\partial t} x = -\kappa^\alpha n$$

where $\kappa$ is the curvature, and $n$ the unit normal.
Proof. For \( k \geq 3 \) the subgroup \( \Gamma_k \) generated by rotation through \( 2\pi/k \) satisfies the conditions of Theorem 18. The first spherical harmonics symmetric under \( \Gamma_k \) are \( \cos(k\theta) \) and \( \sin(k\theta) \), with corresponding eigenvalue \( \lambda = k^2 \). □

Corollary 21. In the case \( n = 2 \), there exists a homothetically contracting solution of Eq. (6) with tetrahedral symmetry provided \( \alpha \in (0, 1/10) \); there exists one with octahedral symmetry provided \( \alpha \in (0, 1/18) \); and there exists one with icosahedral symmetry provided \( \alpha \in (0, 1/40) \).

Proof. In this case the only subgroups satisfying the required condition are the symmetry groups of the platonic solids. There are three such groups, since the dual solids have the same symmetry group. The first tetrahedrally symmetric spherical harmonic is given by the restriction of the function \( xyz \) to \( S^2 \), and the corresponding eigenvalue is 12. The first octahedrally symmetric spherical harmonic is \( x^4 + y^4 + z^4 - 3x^2y^2 - 3x^2z^2 - 3y^2z^2 \), and the corresponding eigenvalue is 20. Finally, the first icosahedrally symmetric spherical harmonic is

\[
231z^6 - 315z^4(x^2 + y^2) + 105z^2(x^2 + y^2)^2 - 5(x^2 + y^2)^3 \\
+ 42zx^5 - 420zx^3y^2 + 210zxy^4,
\]

and the corresponding eigenvalue is 42. □

Corollary 22. For \( n \geq 3 \), there exists a non-spherical homothetically contracting solution with the symmetry of a regular \((n + 2)\)-simplex for \( \alpha \in (0, 1/(2(n + 3))) \), and one with the symmetries of a regular hypercube for \( \alpha \in (0, 1/3(n + 4)) \).

Proof. The function \( \sum_i x_i^4 - (6/n) \sum_{i\neq j} x_i^2x_j^2 \) has the symmetry of a regular hypercube in \( \mathbb{R}^{n+1} \), and its restriction to \( S^n \) is a spherical harmonic with eigenvalue \( 4(n + 3) \).

For all \( n \geq 1 \) there exists a cubic homogeneous harmonic polynomial \( u_n \) on \( \mathbb{R}^{n+1} \) with the symmetries of a regular simplex. These are given by the recursive definitions

\[
u_1(x_1, x_2) = x_2^3 - 3x_1^2x_2 \\
u_{n+1}(x_1, \ldots, x_{n+2}) = u_n(x_1, \ldots, x_{n+1}) \\
+ \beta_n \left( x_{n+2}^3 - \frac{3}{n + 1} \left( x_1^2 + \cdots + x_{n+1}^2 \right) \right)
\]

where \( \beta_1 = \sqrt{2} \) and

\[
\beta_{k+1} = \beta_k \sqrt{\frac{(k + 1)^3}{k^2(k + 3)}}.
\]

The restriction of \( u_n \) to \( S^n \) is a spherical harmonic with the required symmetries, and the corresponding eigenvalue is \( 3(n + 2) \). □
In [A5] we prove that the only homothetically contracting solutions of flows by positive powers of curvature are those given in Corollary 20. For \( n = 2 \) we expect that there are many more solutions which are not described by Corollary 21. In particular, for small \( \alpha \) there should be many solutions symmetric under each of the platonic symmetry groups, and there should also be solutions symmetric under some other subgroups of \( SO(3) \), such as the subgroup of rotations about a fixed axis, and its discrete subgroups. In the case \( n \geq 3 \) there are of course many more examples of suitable subgroups \( \Gamma \) which we have not mentioned explicitly in Corollary 22.

12. Hypersurfaces with planar or cylindrical parts.

In this section we demonstrate that the result of Theorem 1 no longer holds for any \( \alpha > 1/n \). Specifically, we show that any solution starting from a hypersurface containing a planar region must still contain a planar region for small positive times. We also consider the behaviour of hypersurfaces which contain regions which are cylindrical or locally have the form \( M^{n-k} \times \mathbb{R}^k \) for some \( k > 0 \).

**Theorem 23.** Suppose \( M_0 \) is a compact convex hypersurface, and \( F_0 \) a subset of \( M_0 \) which has the form \( N_0^{n-k} \times U_0^k \), where \( N_0 \) is a smooth convex hypersurface in \( \mathbb{R}^{n+1-k} \) and \( U_0 \) is an open subset of \( \mathbb{R}^k \). Then for any smooth positive function \( \rho \) on \( S^n \) and any \( \alpha > 1/k \), the viscosity solution \( \{M_t\} \) starting from \( M_0 \) contains an open subset \( F_t \) of \( F_0 \) for \( t > 0 \) sufficiently small.

Conversely, suppose \( \Omega_0 \) in a bounded open convex set in \( \mathbb{R}^{n+1} \) for which

\[
\sigma_{n-k} = \inf_{x \in \partial \Omega_0} \sup_{P, \Gamma} \left\{ \kappa : y \in \Omega_0 \Rightarrow \langle x - y, P^\perp \rangle \geq \frac{1}{2} \kappa |\pi_{\Gamma}(x-y)|^2 \right\} > 0
\]

where the supremum is over all supporting hyperplanes \( P \) of \( \Omega_0 \) which contain \( x \) and all \( n-k \)-dimensional affine subspaces \( \Gamma \) of \( P \) through \( x \), and \( P^\perp \) is the unit normal to \( P \) which points outward from \( \Omega_0 \). Then for any \( \rho \) and any \( \alpha \in (0, 1/k] \) the viscosity solution \( \{\Omega_t\} \) of (6) with initial condition \( \Omega_0 \) is smooth and strictly convex for \( t > 0 \) and remains so until it contracts to a point.

When \( \Omega_0 \) is smooth, \( \sigma_{n-k} \) is strictly positive if and only if \( E_{n-k}(W) \) is strictly positive, or equivalently if and only if the sum of the smallest \( k+1 \) principal curvatures is strictly positive at every point.

**Proof.** To prove the first part of the Theorem, we will construct barriers with cylindrical symmetry, described by embeddings of the form \( \varphi : S^{n-k} \times B^k_R(0) \) with \( \varphi(z,x) = (u(|x|)z, x) \in \mathbb{R}^{n+1-k} \times \mathbb{R}^k \simeq \mathbb{R}^{n+1} \). We consider the case where \( u \) is concave and decreasing in \( |x| \). The metric and second fundamental
form are given by
\[ g_{zi} = u^2 \hat{g}_{ij}, \]
\[ g_{xz} = 0, \]
\[ g_{xi} = \delta_{ij} + \frac{x_i x_j}{|x|^2} (u')^2 \]
and
\[ \mathcal{W}_{zi}^z = \frac{\delta_i^j}{u \sqrt{1 + (u')^2}} \]
\[ \mathcal{W}_{zi}^x = 0 \]
\[ \mathcal{W}_{xi}^x = -\frac{u''}{(1 + (u')^2)^{3/2}} \frac{x_i x_j}{|x|^2} - \frac{u'}{|x| \sqrt{1 + (u')^2}} \left( \delta_i^j - \frac{x_i x_j}{|x|^2} \right). \]
The Gauss curvature is given by the expression
\[ K = (-1)^k \frac{u''(u')^{k-1}}{u^{n-k} |x|^{k-1}(1 + (u')^2)^{(n+2)/2}} \]
and the unit normal is
\[ \nu = \frac{z - u' \hat{x}}{\sqrt{1 + (u')^2}} \]
where \( \hat{x} = x/|x| \).

It follows that any function \( u(r, t) \) satisfying the inequality
\[ \dot{u} \leq -\rho (u')^{\alpha - 1}(u')^{\alpha(k-1)} |x|^{-\alpha(k-1)} (1 + (u')^2)^{-\frac{1-\alpha(1+2)}{2}} \]
has a cylindrical graph which acts as an inner barrier for convex solutions of Eq. (6). We proceed to construct such barriers.

**Lemma 24.** For given \( k, \rho, \) and \( \alpha > 1/k \) there exist constants \( c_1, c_2 \) and \( c_3 \) such that for any \( \lambda > 0 \) and \( u_0 > 0, R > 0, \) the function
\[ u = u_0 \left( 1 - c_1(k, \alpha) \left( \frac{|x|}{R} + c_2(\rho, k, \alpha) \left( \frac{u_0}{R} \right)^{2k\alpha} \frac{t}{u_0^{1+n\alpha} - 1} \right)^{\frac{k+1-1/\alpha}{k-1/\alpha}} \right) \]
is \( C^2 \) and satisfies (25) on the region
\[ |x| \leq 2R, \quad 0 \leq t \leq c_3(\rho, k, \alpha) u_0^{1+n\alpha} \left( \frac{R}{u_0} \right)^{2k\alpha}, \]
provided \( \alpha \in (1/k, 1/(k-1)) \). If \( \alpha \geq 1/(k-1) \) then \( u \) is not \( C^2 \) but acts as a barrier for smooth, strictly convex solutions of Eq. (6), hence also for the viscosity solutions constructed in Theorem 15.
Proof. The case \( \alpha \in (1/k, 1/(k - 1)) \) follows by direct computation. In the case \( \alpha \geq 1/(k - 1) \) the same computation gives the result except when \( |x| = R - c_2(u_0/R)^{2n/k}R/\rho_0^{1+n\alpha} t \), where \( u \) is not \( C^2 \). A smooth convex hypersurface lying outside the graph of \( u \) and meeting the graph at such a point must have \( K = 0 \), so the barrier condition is verified.

To use these barriers in the comparison principle, we need to check that the viscosity solution \( \{ u \} \) stays away from the boundary of the barrier produced in Lemma 24. Fix \( x \) in the interior of \( M_0 \), \( \varepsilon > 0 \) small, and choose \( u_0 \) smaller than the smallest radius of curvature of \( \sigma_0 \), and place the origin at the point \( x - (u_0 + \varepsilon)\nu_N \). Choose \( R \) sufficiently small so that the distance from \( x \) to the boundary of \( F_0 \) is at least \( 3R \). Let \( M_0 \) be the hypersurface given by

\[
\sqrt{\sum_{i=1}^{n-k} x_i^2} = u_0 \left( 1 - c_1(k, \alpha) \left( \frac{\sqrt{\sum_{i=n-k+1}^{n} x_i^2}}{R} - 1 \right) \right) + \frac{\sqrt{\sum_{i=n-k+1}^{n} x_i^2}}{R} \left( u_0 \right)^{2n/k} R \left( \frac{t}{\rho_0^{1+n\alpha} u_0} - 1 \right) + \frac{\sqrt{\sum_{i=n-k+1}^{n} x_i^2}}{R} \left( u_0 \right)^{2n/k} R \left( \frac{t}{\rho_0^{1+n\alpha} u_0} - 1 \right)
\]

for \( \sum_{i=n-k+1}^{n} x_i^2 \leq 4R^2 \). Then \( M_0 \) lies entirely within \( M_0 \). Furthermore, for each \( y \in \partial M_0 \), the sphere \( B_{c(\rho, k, \alpha) \min\{u_0, R\}}(y) \) is enclosed by \( M_0 \) for some constant \( c \). Define \( M_t \) to be the hypersurface

\[
\sqrt{\sum_{i=1}^{n-k} x_i^2} = u_0 \left( 1 - c_1 \left( \frac{\sqrt{\sum_{i=n-k+1}^{n} x_i^2}}{R} + c_2 \left( u_0 \right)^{2n/k} R \left( \frac{t}{\rho_0^{1+n\alpha} u_0} - 1 \right) \right) \right) + \frac{\sqrt{\sum_{i=n-k+1}^{n} x_i^2}}{R} \left( u_0 \right)^{2n/k} R \left( \frac{t}{\rho_0^{1+n\alpha} u_0} - 1 \right)
\]

for \( \sum_{i=n-k+1}^{n} x_i^2 \leq 4R^2 \). For \( t < c(\rho, n, k, \alpha) \min\{u_0, R\}^{1+n\alpha} \) we have

\[
\partial M_t \subset \bigcup_{y \in \partial M_0} B_{r(t)}(y)
\]

where \( r(t)^{1+n\alpha} = \left( c \min\{u_0, R\}^{1+n\alpha} - (1 + n\alpha) \sup \rho t \right) \). The comparison principle implies that each of these balls is enclosed by \( M_t \), and therefore that \( \partial M_t \) does not meet \( M_t \). It follows by the comparison principle that \( M_t \) remains entirely enclosed by \( M_t \) on this time interval, and therefore that \( h(\nu_x, t) \geq h(\nu_x, t) - \varepsilon \) for all \( \varepsilon > 0 \), so that \( x \) has not moved during this time interval.

Now we proceed to the second part of the Theorem, the case of \( \alpha \leq 1/k \). This also proceeds using barrier constructions. Let \( x \in \partial \Omega_0 \). Since \( \sigma_{n-k} > 0 \), there exists a hyperplane \( P \) supporting \( \Omega_0 \) at \( x \) and an \( n - k \)-dimensional subspace \( \Gamma \) of \( P \) such that \( \Omega_0 \) is contained in the region

\[
\langle x - y, P^\perp \rangle \geq \frac{1}{4} \sigma_{n-k} |\pi_\Gamma(x - y)|^2.
\]
Ω₀ is also clearly contained in the ball $B_{2r_+(Ω₀)}(x)$. We now construct barriers:

**Lemma 25.** Fix $k$, $ρ$, $α ∈ (0, 1/k)$ and constants $R > 0$ and $u₀ > 0$. If every $y ∈ Ω₀$ satisfies

$$
|y| ≤ R, \quad \sum_{i=1}^{n-k} y_i^2 ≤ u₀^2
$$

then for $0 < t ≤ c(ρ, k, n, α) \min\{u₀, R\}^{1+α}$ we have for every $y = (Y, η) ∈ Ωₜ ⊂ ℝ^{n+1} × ℝ^k ≃ ℝ^{n+1},$

$$
|Y| ≤ u₀ - ct^{1/kα} R^{1-kα} \frac{α-1}{α(n-k)} \left( (4R - |η|)^{\frac{α(k+1)}{1-kα}} + (4R + |η|)^{-\frac{α(k+1)}{1-kα}} \right)
$$

where $c$ depends only on $ρ$, $k$, $n$, and $α$. If $α = 1/k$ and every $y ∈ Ω₀$ satisfies (26), then for $0 < t ≤ c(ρ, k, n) \min\{u₀, R\}^{1+1/k}$ and $y = (Y, η) ∈ Ωₜ$ we have

$$
|Y| ≤ u₀ - c₁R \exp \left( -c₂R^{2k} u₀^{n-k} t^{-k} \right) \cosh \left( c₂R^{2k-1} u₀^{n-k} |η| t^{-k} \right)
$$

where $c₁$ and $c₂$ depend only on $ρ$, $k$, and $n$.

**Proof.** A direct calculation shows that these regions are given by cylindrical graphs satisfying the inequality

$$
\dot{u} ≥ -\inf ρ(-u'')^α(-u')^{α(k-1)} u^{-α(n-k)} |x|^{-α(k-1)} (1 + (u')^2)^{-\frac{1-α(n+2)}{2}}
$$

which therefore act as outer barriers for solutions of Eq. (6). □

It follows that the support function at every point must change: If we fix $z ∈ S^ₙ$, and place a cylindrical barrier outside the hypersurface $M₀$ and passing through $x$, then we have

$$
h(z, t) ≤ h(z, 0) - ct^{1/kα} R^{1-kα} \frac{α-1}{α(n-k)} u₀^{\frac{α(n-k)}{1-kα}}
$$

for $α < 1/k$, and

$$
h(z, t) ≤ h(z, 0) - c₁R \exp \left( -c₂R^{2k} u₀^{n-k} t^{-k} \right)
$$

when $α = 1/k$. Theorem 7 then gives lower bounds on the speed $ρS^{n-α}$ uniformly in $z$ for each $t > 0$, by the same argument as given in the proof of Theorem 8. Theorem 10 implies bounds below on all the principal curvatures at each positive time, and Corollary 12 gives upper bounds on the principal curvatures. Theorem 5.5 of [K] implies that the solution hypersurface is smooth and strictly convex for sufficiently small positive times, and the result of Theorem 23 follows from Theorem 13. □
References


Received October 28, 1998. This research was partially supported by an Australian Research Council QEII Fellowship.

AUSTRALIAN NATIONAL UNIVERSITY
A.C.T. 0200
AUSTRALIA
E-mail address: andrews@maths.anu.edu.au
SUMMATION OF FORMAL SOLUTIONS OF A CLASS OF LINEAR DIFFERENCE EQUATIONS

B.L.J. Braaksma, B.F. Faber, and G.K. Immink

Dedicated to the memory of W.A. Harris, Jr

We consider difference equations $y(s+1) = A(s)y(s)$, where $A(s)$ is an $n \times n$-matrix meromorphic in a neighborhood of $\infty$ with $\det A(s) \not\equiv 0$. In general, the formal fundamental solutions of this equation involve gamma-functions which give rise to the critical variable $s \log s$ and a level $1^+$. We show that, under a mild condition, formal fundamental matrices of the equation can be summed uniquely to analytic fundamental matrices represented asymptotically by the formal fundamental solution in appropriate domains.

The method of proof is analogous to a method used to prove multi-summability of formal solutions of ODE’s. Starting from analytic lifts of the formal fundamental matrix in half planes, we construct a sequence of increasingly precise quasi-functions, each of which is determined uniquely by its predecessor.

1. Introduction.

This paper is concerned with summability of formal solutions of linear homogeneous difference equations. We consider the system

\[(1.1) \quad y(s+1) = A(s)y(s),\]

where $s$ is a complex variable, $y(s) \in \mathbb{C}^n$, and $A(s)$ an $n \times n$-matrix, meromorphic at infinity, $\det A(s) \not\equiv 0$. For some $p \in \mathbb{N}$ Equation (1.1) has a formal fundamental matrix solution of the form

\[(1.2) \quad \hat{Y}(s) = \hat{H}(s)s^\Lambda e^{G(s)}s^L,\]

with $\hat{H}(s) \in \text{End}(n, \mathbb{C}[[s^{-1/p}]]), \det \hat{H}(s) \not\equiv 0$, $\Lambda = \bigoplus_{j=1}^{m} \lambda_j l_j$ where $\lambda_j \in \mathbb{Z}/p\mathbb{Z}$ and $l_j$ is the $n_j \times n_j$-identity matrix, $G(s) = \bigoplus_{j=1}^{m} g_j(s)l_j$ where $g_j(s) \equiv 0$ or $g_j(s)$ is a polynomial in $s^{1/p}$ of degree at most $p$ with $g_j(0) = 0$, and $L = \bigoplus_{j=1}^{m} L_j$, $L_j = c_j l_j + N_j$ with $c_j \in \mathbb{C}$ and $N_j$ an $n_j \times n_j$-nilpotent matrix, and with $n_1 + n_2 + \cdots + n_m = n$.  
The purpose of this paper is to sum the entries of $\hat{H}(s)$ on certain unbounded domains $D$ in order to obtain uniquely characterizable analytic fundamental matrix solutions

$$Y(s) = H(s)s^{\Lambda_s}e^{G(s)}s^k, \quad \text{with}$$

$$H(s) \sim \hat{H}(s), \ s \to \infty \text{ on } D.$$  

Any solution of (1.1) on $D$ can be written as $Y(s)P(s)$ where $P(s)$ is a 1-periodic $\mathbb{C}^n$-valued function.

If the factor $s^{\Lambda_s}$ does not appear in the formal fundamental matrix, i.e., if all $\lambda_j$’s vanish, the formal fundamental matrix resembles that of a homogeneous linear differential system. Formal power series solutions of meromorphic differential equations can be summed by means of a method known as multisummation. With such an equation one can associate so-called ‘levels’, positive rational numbers $k_1, \ldots, k_r$, and corresponding ‘critical variables’ $s^{k_1}, \ldots, s^{k_r}$, which play a crucial part in the summation process. Multi-summation is a particular case of accelero-summation (see [Eca87]), involving only elementary accelerations. There exist various equivalent definitions of multisummability (see Definition 2). It can be formulated in terms of Borel and Laplace transforms (cf. [MR91]), or in a more abstract way (cf. [MR92]). In [Bal94] Balser presented yet another definition. Multi-summability of solutions of both linear and nonlinear meromorphic differential equations has been proved both by using Borel-Laplace techniques (see [Bra91] and [Bra92]) and in a way based on the definition of Malgrange and Ramis (see [BBRS91], [RS94], [Bal94], [Tov96], and [BIS]).

Two of the most important features that distinguish linear difference equations from linear differential equations are:

(i) The solution space of a homogeneous linear difference equation is linear over the 1-periodic functions instead of $\mathbb{C}$-linear as in the case of homogeneous linear differential equations.

(ii) The occurrence of the factor $s^{\Lambda_s}$, that does not appear in formal solutions of differential equations.

If the factor $s^{\Lambda_s}$ does not appear in the formal fundamental matrix, or, more generally, if all $\lambda_j$’s are equal, then all entries of $H(s)$ are multisummable in all but at most a countably infinite number of directions. This was shown in [BF96] by means of Borel-Laplace techniques in the spirit of the work of Ecalle [Eca85]. With the same techniques multisummability of formal solutions of a class of non-linear difference equations was proved there.

If not all $\lambda_j$’s are equal, some of the entries of $\hat{H}(s)$ may not be multisummable in any direction. This is due to the fact that, in this case, one of the critical variables is $s \log s$, which is not a rational power of $s$.

Following Ecalle (cf. [Eca85]), one might set out to sum the formal solutions by accelero-summation, using Borel and Laplace transforms. For a
particular class of linear difference equations, accelero-summability of the formal solutions was established in [1mm]. Ecalle’s method involves the study of a convolution equation, obtained from the equation satisfied by $\hat{H}$ by means of a formal Borel transformation in the variable $s \log s$, which does not look very inviting. In the present paper we take a different approach, similar to the method employed in [BIS] to sum formal solutions of linear differential equations (cf. Theorem 13). Our starting point is the ‘main asymptotic existence theorem’ for difference equations (Theorem 6), which says that $\hat{H}(s)$ can be lifted on half planes in $C_{\infty}$, bounded by the real or imaginary axis, to an analytic matrix $H(s)$ such that (1.3) defines an analytic fundamental matrix $Y(s)$ of the difference equation. With the equation we associate certain levels $0 < k_1 < \cdots < k_r = 1$, that can be extracted from the formal fundamental solution, as well as a level $1^+$ if not all $\lambda_j$’s are equal (see Definitions 3 and 5). We choose a covering of a neighbourhood of $\infty$ in $C_{\infty}$ by appropriate half planes and, on each half plane a fundamental system of (1.1) represented asymptotically by the formal fundamental system (1.2). In several steps, modifying the solutions by exponentially small functions of increasing order at each subsequent step, we construct a sequence of so-called $k_j$-precise quasi-functions, $j = 1, \ldots, r$. If the equation does not possess a level $1^+$, this procedure yields the multi-sum, or $(k_1, \ldots, k_r)$-sum of the formal solution (Theorem 13).

If the equation does possess level $1^+$, the final step is more delicate than the preceding ones. This is due to the relative ‘closeness’ of the levels 1 and $1^+$ and the transcendental nature of the critical variable $s \log s$. In order to end up with a unique sum, we need to consider domains that are strictly smaller than half planes, but sufficiently large to exclude the existence of flat solutions of the difference equation satisfied by $\hat{H}$, with a dominant factor of the form $s^{(\lambda_i-\lambda_j)}$, with $\lambda_i \neq \lambda_j$. Here we shall consider domains of the type \{ $s \in C_{\infty}$ | $\arg s \in ((h-1)\pi, (h+1)\pi)$, $(-1)^h R \{ s \log s + i \theta \} > 1$ \}, with $\theta \in \mathbb{R}$, $h \in \mathbb{Z}$ (cf. Figures 2-5). On the union of two such domains with the same $h$ we can define a sum $H(s)$ of $\hat{H}(s)$ if a certain generic condition is satisfied (cf. Section 7). By means of (1.3) we obtain a unique analytic fundamental matrix of the difference equation (Theorem 18).

In order to illustrate the particular properties of difference equations with level $1^+$, we end this introduction with a simple example.

**Example 1.** Consider the equation

\begin{equation}
(1.4) \quad h(s+1) - as^{-1}h(s) = s^{-1} \text{ with } a \in \mathbb{R}, \ a > 0
\end{equation}

which can be transformed into the matrix equation

\[
\begin{pmatrix}
y_1 \\
y_2
\end{pmatrix}
(s+1) = \begin{pmatrix}
a/s & 1/s \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
y_1 \\
y_2
\end{pmatrix}
(s).
\]
(1.4) belongs to a class of equations that was discussed in [BH75] and in
great detail by Ecalle in [Eca85, §3.6] and later by Immink in [Imm]. It
has a unique formal solution \( \hat{h} = \sum_{n \geq 1} h_n s^{-n} \). Let \( \hat{u}(t) := \sum_{n \geq 1} \frac{h_n}{(n-1)!} t^{n-1} \),
the formal Borel transform of \( \hat{h} \). The power series \( \hat{u} \) formally satisfies the
convolution equation
\[
e^{-t} u(t) - a(1 \ast u)(t) = 1.
\]
This equation has the unique analytic solution
\[
u(t) = e^{-a e^{e t}}.
\]
Thus \( \hat{u}(t) \) coincides with the Taylor series at \( t = 0 \) of this function and is
actually a convergent power series which extends to a holomorphic function
on \( \mathbb{C} \). The convergence of \( \hat{u} \) implies that \( \hat{h} \) is \( 1 \)-Gevrey. By using Cauchy’s
formula for the coefficients in a convergent Taylor series, one may derive the
more precise estimate
\[
|h_n| \leq K \left( \frac{n}{\log n} \right)^n A^n, \forall n \geq 2, \text{ for some } K, A > 0.
\]
This type of estimate is typical of difference equations possessing a level \( 1^+ \)
(cf. [Imm88]).

The function \( u(t) = e^{-a e^{e t}} \) is bounded in the left half plane and, conse-
quently, \( \hat{h}(s) \) is \( 1 \)-summable in all directions in \( (\frac{\pi}{2}, \frac{3\pi}{2}) \). The \( 1 \)-sum \( h_\ell(s) \) is
analytic on the sector \( 0 < \arg s < 2\pi \) and is a solution of (1.4).

In the right half plane \( u(t) \) grows faster than exponentially of any order
on the horizontal strips \( \{ t \in \mathbb{C} \mid \Re t > 0, \Im t \in (-\pi/2, \pi/2) \text{ mod } 2\pi \} \), but on
the strips
\[
\{ t \in \mathbb{C} \mid \Re t > 0, \Im t \in (\pi/2, 3\pi/2) \text{ mod } 2\pi \}
\]
it decreases faster than exponentially of any order. Hence, the functions
\[
(1.5) \quad h_{r,n}(s) := \int_{C_n} e^{-st} u(t) dt, \quad n \in \mathbb{Z},
\]
with \( C_n \) a path from 0 to \( +\infty + i\theta \), \( \theta \in (\pi/2+2n\pi, 3\pi/2+2n\pi) \) (see Figure 1)
are well defined and satisfy (1.4).
The functions $h_{r,n}$ all have the asymptotic expansion $\hat{h}$ as $s \to \infty$, $-\frac{\pi}{2} < \arg s < \frac{\pi}{2}$, uniformly on closed subsectors (cf. also [BH75]). However, it can be shown that the $h_{r,n}$ are not 1-sums. See also [vdPS97, Chapter 11].

In order to characterize these solutions by means of their asymptotic behaviour, we have to consider this asymptotic behaviour on regions other than sectors, namely regions of the form

$$D(\theta) := \{ s \in \mathbb{C} \mid \Re\{s(\log s + i\theta)\} > 1\}, \theta \in \mathbb{R},$$

see Figures 2-5.

![Figure 2. Region $D(0)$.](image1)

![Figure 3. Region $D(\frac{\pi}{2})$.](image2)

![Figure 4. Region $D(\pi)$.](image3)

![Figure 5. The regions ‘rotate’ clockwise with increasing $\theta$.](image4)
Proposition. For any \( \theta \in (2n\pi, 2(n+1)\pi) \), there exist \( K, A > 0 \) such that
\[
|h_{r,n}(s) - \sum_{n=1}^{N-1} h_n s^{-n}| \leq K A^N (N!) |s|^{-N}, \forall s \in D(\theta), \forall N \in \mathbb{N}.
\]

This proposition has been proved by Borel-Laplace methods in [Fab97]. According to a theorem by Immink in [Imm96], \( h_{r,n} \) is uniquely determined by the above property.

2. Preliminaries.

By Arg\( z \) we denote the principal argument of \( z \in \mathbb{C} \setminus \{0\} \); we take Arg\( z \) \( \in (\pi, -\pi] \). The Riemann surface of the logarithm will be denoted by \( C_\infty \).

For \( \alpha, \beta \in \mathbb{R} \) we denote by \( S(\alpha, \beta) \) the open sector \( \{ s \in C_\infty \ | \ \alpha < \arg s < \beta \} \), and by \( S[\alpha, \beta] \) the closed sector \( \{ s \in C_\infty \ | \ \alpha \leq \arg s \leq \beta \} \). Similarly, \( S[\alpha, \beta) \) and \( S(\alpha, \beta] \) denote half-open sectors. For \( \mu \in \mathbb{Z} \) we define
\[
H_\mu := S((\mu - 1)\pi/2, (\mu + 1)\pi/2); \ |\mathcal{H}_\mu| := S(\mu - 1)\pi/2, (\mu + 1)\pi/2).
\]

Throughout this paper, by an upper half plane, a fourth quadrant, etcetera, we understand a lift of the upper half plane, the fourth quadrant, etcetera, from the complex plane to the Riemann surface of the logarithm. A sector will always be a sector of \( C_\infty \) with vertex at the origin.

By definition, a neighbourhood of \( \infty \) in a sector \( S \) (\( S \) not necessarily open) is an open subset \( U \) of \( S \), such that, for any closed subsector \( S' \) of \( S \) with aperture \( \leq \pi \), we can find \( s_0 \in S \) such that \( s_0 + S' \subset U \). In particular, \( R e^{i\mu\pi/2} + H_\mu, \mu \in \mathbb{Z}, R > 0, \) is a neighbourhood of \( \infty \) both in \( H_\mu \) and in \( \mathcal{H}_\mu \).

If we write \( f(s) = O(g(s)) \) or \( f(s) = o(g(s)) \) as \( s \to \infty \) on a sector \( S \), we mean that \( f \) and \( g \) are functions defined on a neighbourhood \( U \) of \( \infty \) in \( S \), and that the \( O \) or \( o \) relation holds uniformly, as \( s \to \infty \), on the intersection of \( U \) and any closed subsector of \( S \).

Similarly, if \( \hat{f}(s) = \sum_{j \geq 0} a_j s^{-j/p} \) where \( p > 0 \), and if \( S \) is a sector, then \( f(s) \sim \hat{f}(s), s \to \infty \) on \( S \), means the following: \( f \) is an analytic function on a neighbourhood \( U \) of \( \infty \) in \( S \) and for any closed subsector \( S' \subset S \) and any \( N \in \mathbb{N} \), we can find positive constants \( R \) and \( C_{S',R,N} \), such that
\[
(f(s) - \sum_{j=0}^{N-1} a_j s^{-j/p}) \leq C_{S',R,N} |s|^{-N/p}, \forall s \in S' \cap U, |s| > R.
\]

The set of such functions \( f \) with an asymptotic expansion \( \hat{f} \) on \( S \) as above will be denoted by \( \mathcal{A}(S) \).
In accordance with the above, when we write $f(s) \sim 0$, $s \to \infty$ on $S$, we mean that $f(s) = o(s^{-N})$, $s \to \infty$ on $S$, for any $N \in \mathbb{N}$.

Suppose $\hat{u}(s) = \sum_{j=0}^{m} \hat{h}_j(s)(\log s)^j$ with $\hat{h}_j \in C[[s^{-1/p}]]$ for $j = 0, \ldots, m$. If we write $u(s) \sim \hat{u}(s)$, $s \to \infty$ on $S$, we mean that there exist analytic functions $h_{j}$, $j = 0, \ldots, m$, on a neighborhood of $\infty$ in $S$ such that $u(s) = \sum_{j=0}^{m} h_j(s)(\log s)^j$, with $h_j(s) \sim \hat{h}_j(s)$, $s \to \infty$ on $S$.

If $f \in \mathcal{A}(S)$ such that (2.1) holds and if there exist $k > 0$, and $K_{S'}, R, A_{S'} > 0$ such that, for each $N$,

$$C_{S', R, N} \leq K_{S', R} A_{S'}^N \Gamma \left( \frac{N}{pk} \right),$$

then we call $f$ a $k$-Gevrey function on $S$ with respect to the family $\frac{1}{p} \mathbb{N}_0$, and we write $f \in \mathcal{A}_{(1/k)}(S)$. Note that $\frac{1}{p} \mathbb{N}_0$ is an example of a ‘convenient family’, according to the terminology introduced by Malgrange in [Mal95].

Any $f \in \mathcal{A}_{(1/k)}(S)$ has an asymptotic expansion $\hat{f}(s) = \sum_{j=0}^{\infty} a_j s^{-j/p}$ with the $a_j$ satisfying

$$|a_j| \leq K A^j \Gamma \left( \frac{j}{pk} \right), \forall j > 0,$$

for some positive $K$ and $A$. Such a formal series $\hat{f}$ will be called a Gevrey series (in $s^{-1}$) of order $1/k$ with respect to the family $\frac{1}{p} \mathbb{N}_0$, and $C[[s^{-1/p}]] \frac{1}{pk}$ denotes the set of such series.

In the sequel all Gevrey functions and Gevrey series will be with respect to the family $\frac{1}{p} \mathbb{N}_0$ with $p$ as in (1.2), and we omit the references to this family in our notations.

A function $f$ defined on a neighbourhood of $\infty$ in a sector $S$ is exponentially small of order $k > 0$ on $S$ if for any closed subsector $S'$ of $S$ there exists a positive constant $c$ such that $f(s) = O(e^{-c|s|^k})$, $s \to \infty$ on $S'$. If this holds for all positive $c$ then $f$ is said to be supra-exponentially small of order $k$ on $S$. The set of all analytic functions on a neighbourhood of $\infty$ in $S$ which are exponentially or supra-exponentially small of order $k$ on $S$ will be denoted by $\mathcal{A}^{\leq-k}(S)$ and $\mathcal{A}^{<k}(S)$ respectively. If $S = S(\alpha, \beta)$ then we will also write these latter sets as $\mathcal{A}^{\leq-k}(\alpha, \beta)$ and $\mathcal{A}^{<k}(\alpha, \beta)$. Similarly if $S = S(\alpha)$ etc. If $f$ and $g$ both are in $\mathcal{A}_{(1/k)}(S)$, and $f$ and $g$ have the same asymptotic expansion, then it can be shown that their difference $f - g$ is in $\mathcal{A}^{\leq-k}(S)$ (cf. [Mal95]).

Let $I > 0$ and $S$ be an open sector. Let $\{S_i\}_{i \in I}$ be a covering of $S$ consisting of open sectors and let $f^{(i)} \in \mathcal{A}(S_i), i \in I$, such that $f^{(i)} - f^{(j)} \in \mathcal{A}^{\leq-I}(S_i \cap S_j)$ for any $i_1, i_2 \in I$ with $S_{i_1} \cap S_{i_2} \neq \emptyset$. These data determine an $l$-precise quasi-function on $S$. We identify two such sets of data $(\{f^{(i)}\}_{i \in I}; \{S_i\}_{i \in I})$ and $(\{g^{(j)}\}_{j \in J}; \{S_j\}_{j \in J})$ if $f^{(i)} - g^{(j)} \in \mathcal{A}^{\leq-I}(S_i \cap S_j)$ where $i \in I, j \in J$ such that $S_i \cap S_j \neq \emptyset$. They define the same $l$-precise
Definition 3.

Let \( f \) be a representative of \( A/A^{\leq -k}(S) \) on \( S \) and let \( A/A^{\leq -k}(S) \) be a nested sequence of open sectors, where \( S \) has aperture larger than \( \pi/k \). Then \( f \) is \((k, \ldots, k_r)\)-summable on \( S \) if there exist \( f_i \in A/(A/k_i)/A^{\leq -k_i}(S_i) \), \( i = 1, \ldots, r \), such that \( f_i|_{S_{i+1}} \equiv f_{i+1} \mod A^{\leq -k_{i+1}} \), \( i = 0, \ldots, r-1 \), and 

\[
\begin{align*}
\hat{f} &= \frac{1}{p} \sum_{i=0}^{r-1} f_i \\
&= \frac{1}{p} \sum_{i=0}^{r-1} f_i|_{S_{i+1}} \\
&= \frac{1}{p} \sum_{i=0}^{r-1} f_i|_{S_{i+1}} \mod A^{\leq -k_{i+1}}.
\end{align*}
\]

We call \( f \) the \((k_1, \ldots, k_r)\)-sum of \( \hat{f} \) on \( S \), and we have \( f_r(s) \sim \hat{f}(s), s \to \infty \) on \( S \).

According to the \('relative Watson lemma' (MR92, Prop. (2.1))\) \( f_{i+1} \) is completely determined by \( f_i \) and \( S_{i+1}, i = 0, \ldots, r-1 \). Hence the \((k_1, \ldots, k_r)\)-sum of \( \hat{f} \) on \( S \) is uniquely defined. We may extend the definition of multisummability in an obvious way to the case that \( \hat{f} \) is an \( n \)-vector or an \( n \times n \)-matrix with elements in \( \mathbb{C}\).

Definition 4. Let \( f(s) = s^{dx} e^{2\pi ibsd + q(s)s^\gamma} \), with \( d \in \mathbb{Z}/p \), \( b \in \mathbb{C}, q(s) \) identically zero or a polynomial in \( s^{1/p} \) without constant term and of degree at most \( p-1 \), and \( \gamma \in \mathbb{C} \). We will say that \( f(s) \) is of level 1 if \( d \neq 0 \), of level 1 if \( d = 0, b \neq 0 \), of level \( k \) with \( k \in \{\frac{1}{p}, \ldots, \frac{p-1}{p}\} \) if \( d = b = 0, q(s) \neq 0 \), and of level 0 if \( d = b = 0 \) and \( q(s) \equiv 0 \).
Let \( f \) be of level \( k \in (0,1] \), so \( d = 0, 2\pi is b + q(s) \neq 0 \). A closed interval \([\sigma - \pi/k, \sigma]\) will be called a **Stokes interval of level** \( k \) of \( f \) if \( f \in A^{\leq-k}(\sigma - \pi/k, \sigma) \). So if \( k = 1 \) then \( \sigma = \pi - \text{Arg}b \mod 2\pi \) whereas if \( 0 < k < 1 \) and \( q(s) = \omega s^k + o(s^k) \), \( s \to \infty \), \( \omega \neq 0 \), then \( k\sigma = \frac{3}{2}\pi - \text{Arg}\omega \mod 2\pi \).

If \( f \) is of level \( 1^+ \) we will, in Section 7, associate with it a certain **Stokes number**. This number is connected with curves that separate regions of growth from regions of decay. All these curves have the limiting directions \( \frac{\pi}{2} \mod \pi \).

If \( f \) is of level \( 1^+ \), we have \( f(s) = \exp(ds \log s(1 + o(1))) \), \( s \to +\infty \) with \( d \neq 0 \), and so it grows or decays faster than exponentially of order 1 on \( \mathbb{R}^+ \), but slower than any higher exponential order.

**Definition 4.** For any (i.e., not necessarily open) sector \( S \) we will write \( f \in A^{\leq-1^+}(S) \) to express that \( f \) is analytic on a neighbourhood \( U \) of \( \infty \) in \( S \), and that for any closed subsector \( S' \) of \( S \), there exists a positive constant \( c \) (depending on \( S' \)) such that \( f(s) = O(e^{-c|s|\log |s|}) \), uniformly as \( s \to \infty \) on \( S' \cap U \).

We define \( 1^+ \)-precise quasi-functions by replacing \( l \) by \( 1^+ \) in the definition of \( l \)-precise quasi-functions above.

So, for example, \( e^{ds \log s} \in A^{\leq-1^+}(H_0) \) if \( d < 0 \). And if \( p(s) = \sum_{j=0}^p p_j e^{2\pi isj} \) is an analytic 1-periodic function on \( \{ s \in \mathbb{C} | 3s > R \} \) for some \( R > 0 \), then \( p(s) e^{ds \log s} \in A^{\leq-1^+}(H_0 \cap \overline{I_1}) \) if \( d < 0 \).

With Equation (1.1) we associate levels, and with each level certain Stokes intervals or numbers. For this purpose we rewrite the formal fundamental matrix solution (1.2) as follows:

\[
(2.2) \quad \tilde{Y}(s) = \tilde{U}(s) F(s),
\]

where \( \tilde{U}(s) = \hat{H}(s) s^N, N = \bigoplus_{j=1}^n N_j \) and \( F(s) = s^\lambda e^{G(s)} s^C, C = \bigoplus_{j=1}^n c_j I_j \).

The columns \( \tilde{y}_l(s) (l = 1, \ldots, n) \) of \( \tilde{Y}(s) \) form a formal fundamental system of solutions \( \{ \tilde{y}_l \}_{l=1}^n \) and we have

\[
(2.3) \quad \tilde{y}_l(s) = f_l(s) \hat{u}_l(s), \quad f_l(s) = s^d l e^{2\pi isb_l + ql(s)} s^{\gamma_l},
\]

where \( \hat{u}_l(s) \in \mathbb{C}^n[[s^{-1/p}][\log s]] \) is the \( l \)-th column of \( \hat{U}(s) \), and, furthermore, if \( 0 < l - (n_1 + \ldots + n_{j-1}) \leq n_j \), then \( d_l = \lambda_j, b_l \in \mathbb{C} \) and \( q_l(s) \equiv 0 \) or \( q_l(s) \) is a polynomial in \( s^{1/p} \) without constant term and of degree at most \( p - 1 \) such that \( 2\pi isb_l + ql(s) = g_j(s) \) and \( \gamma_l = c_j \). Without loss of generality, we may assume that \( \Re b_l \in [0,1), l = 1, \ldots, n \).

We use the following abbreviations (cf. (2.3)): \( f_{ml} := f_m f_{l-1}^{-1}, d_{ml} := d_m - d_l, b_{ml} := b_m - b_l, q_{ml} := q_m - q_l, \gamma_{ml} := \gamma_m - \gamma_l \). We write \( \kappa_{ml} \) for the level of \( f_{ml} \).
Definition 5. The levels of Equation (1.1) are the levels of the functions $e^{2\pi isj}f_{ml}(s)$, $j \in \mathbb{Z}$, $m, l \in \{1, \ldots , n\}$. Let $k \in \{\frac{1}{p}, \ldots , \frac{p-1}{p}, 1\}$. The Stokes intervals of level $k$ of the equation are the Stokes intervals of level $k$ of the functions $e^{2\pi isj}f_{ml}(s)$, $j \in \mathbb{Z}$, $m, l \in \{1, \ldots , n\}$.

Taking $j = 0$ we see that all the $\kappa_{ml}$ are levels of the equation. Moreover, 0 and 1 always are levels of the equation (take $m = l$ and then $j = 0$ and $j \neq 0$, respectively). By $0 < k_1 < \cdots < k_r = 1$ we denote the increasing sequence of levels of the equation in the interval $(0, 1]$. If $\Im b_{ml} \neq 0$ for some $m$ and $l$ then there are infinitely many Stokes directions (endpoints of Stokes intervals) $-\text{Arg}(b_{ml} + j) \mod \pi, j \in \mathbb{Z}$ which cluster at 0 mod $\pi$.

The following theorem is the counter part in the theory of linear difference equations of the ‘main asymptotic existence’ theorem in the theory of differential equations.

Theorem 6. Let $l \in \{1, \ldots , n\}$ and $\hat{y}_l(s) = f_l(s)\hat{u}_l(s)$ be a formal solution of (1.1) of the form (2.3) with $\hat{u}_l(s) \in \mathbb{C}^n[[s^{-1/p}]]\log s$.

Then for any $\mu \in \mathbb{Z}$ there exists an analytic solution $y_l(s) = f_l(s)u_l(s)$ of the equation such that $u_l(s) \sim \hat{u}_l(s), s \to \infty$ on $H_\mu$.

A proof of this theorem (for the case that no logarithmic terms appear in $\hat{u}_l(s)$) can be found in [vdPS97]. It is based on the so-called quadrant theorem, already stated by Birkhoff and Trjitzinsky in [BT33], but made rigorous by Immink in [Imm91].

3. Two auxiliary lemmas.

The following lemma gives information on the relation between two fundamental systems of solutions of Equation (1.1), which have the same asymptotic behaviour at $\infty$ on some sector.

Lemma 7. Suppose we have two fundamental matrix solutions of Equation (1.1), $Y = UF$ and $Y_1 = U_1F$, such that $U(s) \sim \hat{U}(s)$ and $U_1(s) \sim \hat{U}(s)$ for $s \to \infty$ on an open sector $S$, with $F(s)$ and $\hat{U}(s)$ as in (2.2). Let $u_l$ and $u_{l,1}$ be the $l$-th column of $U$ and $U_1$, respectively.

Then there exist analytic 1-periodic functions $p_{lm}$, $m = 1, \ldots , n$, on a neighbourhood of $\infty$ in $S$, such that

$$u_l - u_{l,1} = \sum_{m=1}^n p_{lm}f_{ml}u_m.$$ 

Moreover,

$$p_{lm}(s)f_{ml}(s) \sim 0, s \to \infty \text{ on } S, \forall m \in \{1, \ldots , n\}.$$ 

If

$$u_l - u_{l,1} \in (A^{\leq -k})^n(S) \text{ for some } k > 0 \text{ (including } k = 1^+),$$
then

\[ p_{lm}f_{ml} \in \mathcal{A}^{\leq -k}(S), \forall m \in \{1, \ldots, n\}. \]

**Proof.** Let \( \tilde{U} := U - U_1 \). Then \( \tilde{U}F = UF\), or, equivalently, \( FPF^{-1} = U^{-1}\tilde{U} \), for some \( 1 \)-periodic analytic matrix function \( P = (P_{ml}) \), on a neighbourhood of \( \infty \) in \( S \). From the diagonal form of the matrix \( F \) it follows that if \( p_{lm} \) is the element in the \( m \)-th row and \( l \)-th column of \( P \), then

\[ p_{lm}(s) f_{ml}(s) \sim (m \text{-th row of } U^{-1})(u_l - u_{l,1}). \]

As \( U(s) \sim \tilde{U}(s) \) we have \( U^{-1}(s) \sim \tilde{U}^{-1}(s) = s^{-N}\tilde{U}^{-1}(s) \). Hence any entry of \( U^{-1} \) is of order \( O(s^\mu (\log s)^\nu) \), \( s \to \infty \) for some \( \mu, \nu \in \mathbb{Z} \). Since \( u_l(s) - u_{l,1}(s) \sim 0 \) as \( s \to \infty \) on \( S \), we thus find that

\[ p_{lm}(s) f_{ml}(s) \sim 0, s \to \infty \text{ on } S, \forall m \in \{1, \ldots, n\}. \]

Similarly, we see that \( p_{lm}f_{ml} \in \mathcal{A}^{\leq -k}(S) \), if \( u_l - u_{l,1} \in (\mathcal{A}^{\leq -k})^n(S) \) for some \( k > 0 \), including \( k = 1^+ \).

The next lemma yields more information on the asymptotic behaviour of the functions \( p_{lm}f_{ml} \) in the previous lemma.

**Lemma 8.** Let \( S := S(\alpha_1, \alpha_2) \) be an open sector with \( 0 < \alpha_2 - \alpha_1 \leq \pi \). Let \( p(s) \) be an analytic, \( 1 \)-periodic function on a neighbourhood of \( \infty \) in \( S \), and let \( f \) be a function of level \( k \in \{0, 1/p, \ldots, 1, 1^+\} \) as in Section 2: \( f(s) = s^{\alpha_1}e^{2\pi is+b+\varphi(s)} \) with \( \Re b \in [0, 1) \). Assume \( g(s) := p(s)f(s) \sim 0, s \to \infty \) on \( S \). Let \( H \) be an upper or lower half plane in \( \mathbb{C}_\infty \) which has a nonempty intersection with \( S \).

Then \( g \in \mathcal{A}^{\leq -k}(S) \) if \( k > 0 \) and \( g \in \mathcal{A}^{\leq -1}(H) \) if \( k = 0 \). If \( \nu \) denotes some integer we have:

1) If \( k = 0 \): If \( \alpha_1 < \nu \pi < \alpha_2 \) then \( p = g = 0 \).

2) If \( 0 < k < 1 \):

Then there exists \( c \in \mathbb{C} \) such that \( g - cf \in \mathcal{A}^{\leq -1}(H) \) and \( cf \in \mathcal{A}^{\leq -k}(S) \). If \( \alpha_1 < \nu \pi < \alpha_2 \) then \( p(s) = c \). If \( g \in \mathcal{A}^{\leq -k}(S) \) then \( g \in \mathcal{A}^{\leq -1}(H) \).

3) If \( k = 1 \):

If \( p \neq 0 \) then with \( H \) corresponds an integer \( N \) such that \( p(s) \sim p_Ne^{2\pi iNs} \) as \( \Im s \to \infty \) on \( H \) where \( p_N \neq 0 \). If \( b \in \mathbb{R}^* \) then \( g \in \mathcal{A}^{\leq -1}(H) \) and if moreover \( \alpha_1 < \nu \pi < \alpha_2 \) then \( p = g = 0 \). If \( \alpha_1 \leq \nu \pi \leq \alpha_2 \) and \( (-1)^\nu \Im b < 0 \) then \( p = g = 0 \).

Next suppose

(I) \( \alpha_1 \leq \nu \pi \leq \alpha_2 \) and \( (-1)^\nu \Im b > 0 \).

(II) \( (\alpha_1, \alpha_2) \subset (\beta_1, \beta_2) \) where \( (\beta_1, \beta_2) \) does not contain a Stokes interval of \( e^{2\pi isj}f(s) \) of level 1 for any \( j \in \mathbb{Z} \).

Then there exist analytic \( 1 \)-periodic functions \( p_+ \) and \( p_- \) such that \( p = p_+ + p_- \) and \( p_+ f \in \mathcal{A}^{\leq -1}(\alpha_1, \beta_2) \) and \( p_- f \in \mathcal{A}^{\leq -1}(\beta_1, \alpha_2) \). If
\[ \alpha_1 = \nu \pi \text{ then } p_+ f \in \mathcal{A}^{\leq 1}[\nu \pi, \beta_2) \text{ and similarly if } \alpha_2 = \nu \pi \text{ then } p_- f \in \mathcal{A}^{\leq 1}(\beta_1, \nu \pi]. \]

4) If \( k = 1^+ \):

If \( \alpha_1 < (\nu + \frac{1}{2}) \pi < \alpha_2 \), then \( p = g = 0 \). If \( \nu \pi \leq \alpha_1 < \alpha_2 \leq (\nu + \frac{1}{2}) \pi \) or \( (\nu - \frac{1}{2}) \pi \leq \alpha_1 < \alpha_2 \leq \nu \pi \) then \( g \in \mathcal{A}^{\leq 1}[\nu \pi, \alpha_2) \) and \( g \in \mathcal{A}^{\leq 1}(\alpha_1, \nu \pi] \) respectively. If \( f \not\in \mathcal{A}^{\leq 1}(S) \) then \( p = g = 0 \).

5) If \( k \leq 1 \) and \( g \in \mathcal{A}^{\leq 1}(S) \) then \( g = 0 \).

**Proof.** We will give the proof for the cases (i) \( 2h \pi < \alpha_1 < \alpha_2 < (2h + 1) \pi \), (ii) \( \alpha_1 = 2h \pi \) and (iii) \( \alpha_1 < 2h \pi < \alpha_2 \) for some \( h \in \mathbb{Z} \). The other cases can be treated similarly.

We may choose \( H = H_{4h+1} \). So \( p(s) \) and \( g(s) \) are analytic on \( H \) for \( \Im s > R \) for some \( R > 0 \). Put \( z = e^{2\pi is} \) and \( P(z) := p(s) \). Then \( |z| = e^{-2\pi \Im s} \) and we have expansions

\[ P(z) = \sum_{j=0}^\infty p_j z^j \text{ if } 0 < |z| < e^{-2\pi R}, \quad p(s) = \sum_{j=-\infty}^\infty p_j e^{2\pi isj} \text{ if } \Im s > R. \]

In case (iii) \( p \) is an entire function. So then (3.1) holds with \( p_j = 0 \) if \( j < 0 \) and \( R \) may be replaced by \( -\infty \). We now treat separately the different cases of the lemma.

**Ad 1**  We have \( f(s) = s^\gamma \). Hence \( p(s) = s^{-\gamma} g(s) \sim 0 \) as \( s \to \infty \) on \( S \). The 1-periodicity of \( p(s) \) then implies that \( p(s) \sim 0 \) as \( \Im s \to \infty \), so \( P(z) \to 0 \) as \( z \to 0 \). Therefore \( p_j = 0 \) if \( j \leq 0 \) and \( p, g \in \mathcal{A}^{\leq 1}(H) \). In case (iii) also \( p(s) \sim 0 \) as \( \Im s \to -\infty \), so \( P(z) \to 0 \) as \( z \to \infty \). Hence \( P = 0 \) and so \( g = p = 0 \).

**Ad 2**  For any closed sector \( S_1 \subset (S \cap H) \) and any \( \rho > R \), there exist positive constants \( K \) and \( a \) such that

\[ |p(s)| = |g(s)f(s)^{-1}| \leq K \exp(a|s|^k), \text{ if } s \in S_1 \text{ and } \Im s \geq \rho. \]

From the 1-periodicity of \( p(s) \) and the fact that \( k \in (0, 1) \) it follows that

\[ |p(s)| \leq K \exp(a'\Im s), \forall \Im s \geq \rho, \text{ for some } a' \in (0, 2\pi), \text{ if we choose } \rho \text{ sufficiently large.} \]

This implies \( P(z) = o(z^{-1}), \text{ as } z \to 0 \). Hence \( p_j = 0 \) if \( j < 0 \) in (3.1). With \( c := p_0 \) we get \( p(s) - c \in \mathcal{A}^{\leq 1}(H) \) and \( g - cf \in \mathcal{A}^{\leq 1}(H) \).

If \( f \not\in \mathcal{A}^{\leq -k}(S \cap H) \) then \( c = 0 \) since otherwise \( g \) is unbounded in a neighbourhood of \( \infty \) in \( S \). In cases (i) and (ii) we have \( cf \in \mathcal{A}^{\leq -k}(S) \). Also \( c = 0 \) if \( g \in \mathcal{A}^{\leq -k}(S) \) and therefore \( g \in \mathcal{A}^{\leq -1}(H) \). In case (iii) we have moreover \( P(z) = o(z) \text{ as } z \to \infty \) and so \( P(z) = p(s) \equiv c \). If \( c \neq 0 \) then \( g = cf \in \mathcal{A}^{\leq -k}(S) \) as \( g \sim 0 \) in \( S \).

**Ad 3**  The fact that \( f \) is of level 1 implies that \( d = 0, b \neq 0 \).

We have \( p(s) = g(s)f(s)^{-1} = O(1) \exp(-2\pi is(b + o(1))) \text{ as } s \to \infty \) on \( S \cap H \) and therefore \( P(z) = O(z^{b+o(1)}) \text{ as } z \to 0 \). So if \( p \neq 0 \) then there exists \( N \in \mathbb{Z} \) such that \( p_j = 0 \) for all \( j < N \) and \( p_N \neq 0 \) in (3.1). Then
\[ p(s) = p_N e^{2\pi i N s} (1 + o(1)) \text{ and } g(s) = p_N e^{2\pi i (b + N + o(1)) s} (1 + o(1)) \text{ as } \Re s \to \infty. \]

As \( g(s) \sim 0 \) as \( s \to \infty \) in \( S \) it follows that \( g \in \mathcal{A}^{\leq 1} (S \cap H) \) where \( S \cap H = S \) in cases (i) and (ii). Moreover, if \( \Im b = 0 \) then \( b + N > 0 \) and so \( g \in \mathcal{A}^{\leq 1} (H) \).

In cases (ii) and (iii) we see that if \( p \neq 0 \) and \( \Im b < 0 \) then \( g(s) \to \infty \) on \( \arg s = 2h\pi + \epsilon \) for \( \epsilon \) sufficiently small positive. Hence \( p = g = 0 \) if \( \Im b < 0 \).

In case (iii) similar reasoning as above leads to \( p_j = 0 \) for all \( j > M \) with some \( M \in \mathbb{Z}, M \geq N \), \( g(s) = e^{2\pi i (b + M + o(1)) s} (p_M + o(1)) \) as \( \Re s \to -\infty \) and \( g \in \mathcal{A}^{\leq 1} (\alpha_1, 2h\pi). \) In particular, if \( \Im b = 0 \) and \( p_N \neq 0 \neq p_M \), then the fact that \( g \sim 0 \) in \( S(2h\pi - \epsilon, 2h\pi + \epsilon) \) for some \( \epsilon > 0 \) implies that \( N + b > 0 > M + b \) in contradiction with \( M \geq N \), and therefore \( p = g = 0 \).

So in case (iii) we have \( \Im b \neq 0 \) if \( p \neq 0 \) and so \( \Im b > 0 \). Consequently \( g \) is exponentially small of order 1 in \( S(2h\pi - \epsilon, 2h\pi + \epsilon) \) for some \( \epsilon > 0 \). Thus \( g \in \mathcal{A}^{\leq 1} (S) \).

Next consider the case that (I) and (II) are satisfied. Now \( \nu = 2h \) and only cases (ii) and (iii) with \( \Im b > 0 \) have to be considered. Let \( \sigma_j := (2h + 1)\pi - \text{Arg}(b + j) \) for all \( j \in \mathbb{Z} \). Then \( S(\sigma_j - \pi, \sigma_j) \) is a maximal sector where \( e^{2\pi i s_j} f(s) \) is exponentially small of order 1 and the behaviour of \( g \) on \( S \cap H \) implies that \( \sigma_N - \pi \leq 2h\pi < \sigma_2 \leq \sigma_N \). If \( \sigma_N \geq \beta_2 \) then we see that \( g \in \mathcal{A}^{\leq 1} (2h\pi, \beta_2) \). Next suppose \( \sigma_N < \beta_2 \). As \( \sigma_j \) increases monotonically from \( 2h\pi \) to \( (2h + 1)\pi \) as \( j \) increases from \( -\infty \) to \( +\infty \), there exists \( A \in \mathbb{Z} \) such that \( \sigma_A < \beta_2 \leq \sigma_{A + 1} \) and \( A \geq N \). The condition on Stokes intervals now implies that \( \sigma_{A + 1} - \pi \leq \beta_1 \).

Let \( p_-(s) := \sum_{j=A}^{N} p_j e^{2\pi i \sigma_j} \). Then \( p_-(s) = O(e^{2\pi i s A}) \) on the lower half plane and \( p_+(s) = O(e^{2\pi i s N}) \) on the upper half plane. So \( p_\pm \) is exponentially small of order 1 for \( arg s \in (\sigma_A - \pi, 2h\pi] \) and for \( arg s \in [2h\pi, \sigma_N). \) Hence \( p_\pm \in \mathcal{A}^{\leq 1} (\beta_1, \alpha_2) \). Furthermore, \( p_+ := p - p_- = \sum_{j=A+1}^{\infty} p_j e^{2\pi i \sigma_j} = O(e^{2\pi i s (A+1)}) \) on the upper half plane and as \( \sigma_{A+1} \geq \beta_2 \) we see that \( p_+ f \in \mathcal{A}^{\leq 1} (2h\pi, \beta_2) \). Moreover, \( p_+ f = g - p_- f \in \mathcal{A}^{\leq 1} (\alpha_1, \alpha_2) \) and we conclude that \( p_+ f \in \mathcal{A}^{\leq 1} (\alpha_1, \beta_2) \).

Ad 4) Now \( f(s) = \exp\{ds(\log s + O(1))\}, s \to \infty \) with \( d \neq 0 \). Since \( \Re(s) \log s = \Re s \log |s| - \Im s \arg s \), we have \( p(s) = g(s) f(s)^{-1} = O(\exp (-d \Re s \log |s| + O(s))) \), as \( s \to \infty \) on \( S \). As \( \Re s / \Im s \) is a nonzero constant on any ray \( \arg s = \psi \notin Z \) we see that if \( S \) contains such a ray on which \( d \Re s > 0 \) then \( p(s) = O(1) \exp (-N|\Im s|) \) for any \( N \in \mathbb{N} \) and therefore \( p = g = 0 \). In particular, if \( f \notin \mathcal{A}^{\leq -1+k} (S) \) then there exists a ray where \( d \Re s > 0 \), so \( p = g = 0 \).

It is now sufficient to consider the case that \( S \) belongs to a right half plane and \( d < 0 \). Choose \( \epsilon > 0 \) with \( 0 < \epsilon < (\alpha_2 - \alpha_1)/2 \) and \( \epsilon < \alpha_2 - \alpha_1 \). Let \( \psi \in (2h\pi, \alpha_2 - 2\epsilon) \). For any \( s \) with \( \arg s = \psi \) there exists \( s_- \in H \) with \( s - s_- \in N \) and \( \arg s_- \in (\alpha_2 - \epsilon, \alpha_2 - \epsilon) \) if \( \Im s \) is sufficiently large. Then \( \Re s_- < \Im s \cot(\alpha_2 - \epsilon) \). Thus

\[ p(s) = p(s_-) = O(1) \exp \left[ -d \cot(\alpha_2 - \epsilon) \Im s + O(s) \right]. \]
From this and $3\delta \log \delta = |s| \sin \psi \log(|s| \sin \psi) = |s|(|s| \log |s| + O(1))$ we conclude that $|p(s)| \leq K_1 \exp\{-d \cot(\alpha_2 - \varepsilon) \sin \psi |s| \log |s| + K_2 |s|\}$ if $\delta$ is sufficiently large where $K_1$ and $K_2$ are some positive constants. Using $g(s) = O(1)p(s)\exp[-|ds| (\cos \psi) \log |s| + O(s)]$ and $\cos \psi - (\sin \psi) \cot(\alpha_2 - \varepsilon) = \sin(\alpha_2 - \varepsilon - \psi)/\sin(\alpha_2 - \varepsilon) > \sin \varepsilon/\sin(\alpha_2 - \varepsilon)$ we see that

$$g(s) = O(1) \exp(-c_\varepsilon |ds| \log |s|)$$

if $\delta$ is sufficiently large and $\arg s \in (2\pi, \alpha_2 - 2\varepsilon)$. Thus we see that $g \in \mathcal{A}^{\leq 1-}[2\pi, \alpha_2)$. In case (iii) we get similarly $g \in \mathcal{A}^{\leq 1-}(\alpha_1, 2\pi]$. Furthermore, the $1$-periodic function $p$ is bounded on any bounded strip parallel to the real axis intersected with $S$. Thus we obtain $g \in \mathcal{A}^{\leq 1-}(S)$ in case (iii).

Ad 5) If $g(s) \in \mathcal{A}^{\leq 1-}(S)$ and $k \leq 1$ we deduce $p(s) = g(s)/f(s) \in \mathcal{A}^{\leq 1-}(S)$. Therefore $p(s) = O(1)e^{-|s|}$ as $s \to \infty$ for all $c > 0$. So $P(z) = O(z^j)$ as $z \to 0$ for all $j$. Hence $P(z) \equiv 0$ and so $p = g = 0$. 

4. A Gevrey property of solutions

Proposition 9. Let $k_1$ be the lowest positive level of (1.1). Then the elements of $\hat{H}(s)$ are Gevrey series of order $1/k_1$. There exist fundamental matrices $Y(\mu)(s) = H(\mu)(s)s^{A_\varepsilon}e^{G(s)s}$ of (1.1) such that $H(\mu)(s)$ is a matrix of $k_1$-Gevrey functions on $H_\mu$ with $H(\mu)(s) \sim \hat{H}(s)$ on $H_\mu$ for all $\mu \in Z$. For any $\mu_0 \in Z$ a representative of $T^{-1}\hat{H}$ (cf. definition of $T^{-1}$ in Section 2) on the covering $\{H_\mu \mid \mu = \mu_0, \ldots, \mu_0 + 4p - 1\}$ of $C_p$ is given by

$$\{H(\mu) \mid \mu = \mu_0, \ldots, \mu_0 + 4p - 1\}.$$

Let $S$ be an open sector of aperture at most $\pi$ and let $U_l$ be given by (2.3) for $l = 1, \ldots, n$. Assume that $f_1 U_l$ is a solution of (1.1) such that $v_l \sim U_l$ on $S$ for $l = 1, \ldots, n$. Then $v_l \in (\mathcal{A}(1/k_1))^n(S)[\log |s|]$. Moreover, $\{f_1 U_l\}_{l=1}^n$ is a fundamental set of solutions of equation (1.1).

Proof. To prove the last statement, let $V$ be the matrix with $v_l$ as $l$-th column. Then $V \sim \hat{U}$ on $S$, where $\hat{U}$ as in (2.2). As $\det \hat{U} \neq 0$, we also have $\det V \neq 0$. Thus $V := VF$ is a matrix solution of Equation (1.1) and $\det V \neq 0$, i.e., it is a fundamental matrix solution.

According to Theorem 6 and the last statement of the proposition under consideration we have fundamental matrices $Y(\mu)(s) = H(\mu)(s)s^{N_F}(s), \mu = \mu_0, \ldots, \mu_0 + 4p - 1$, with $H(\mu)(s) \sim \hat{H}(s), s \to \infty$ on $H_\mu$, $\hat{H}(s)$ as in (1.2). Since $e^{2p\pi i}H_\mu = H_\mu + 4p$ we define $H(\mu_0 + 4p)(s) = H(\mu_0)(se^{-2p\pi i}), s \in H_{\mu_0 + 4p}$. Then $H(\mu_0 + 4p)(s) \sim \hat{H}(se^{-2p\pi i}) = \hat{H}(s), s \to \infty$ on $H_{\mu_0 + 4p}$. If $s \in H_{\mu_0 + 4p}$, $\zeta := se^{-2p\pi i} \in H_{\mu_0}$, then $s + 1 = (\zeta + 1)e^{2p\pi i}$ and

$$Y(\mu_0 + 4p)(s) := H(\mu_0 + 4p)(s)s^{N_F}(s) = H(\mu_0)(\zeta)(\zeta e^{2p\pi i})^{N_F}(\zeta e^{2p\pi i}) = Y(\mu_0)(\zeta)P(\zeta),$$

where $P(\zeta) = \hat{H}(\zeta) = H(\mu_0 + 4p)(\zeta), \zeta = se^{-2p\pi i}$ and $\hat{H}(s) \sim \hat{H}(se^{-2p\pi i})$ as $s \to \infty$ on $H_{\mu_0 + 4p}$.
where \( P(\zeta) = e^{2p\pi i(\Lambda+1)} \), with \( \Lambda \) and \( L \) as in (1.2), is a 1-periodic matrix function, and \( \det P(\zeta) \neq 0 \). Hence \( \Phi^{(\mu_0+4p)}(s) \) is a fundamental matrix.

Next we prove that the entries of \( H^{(\mu)}(s) \) are in \( A_{(1/k_1)}(H_\mu) , \mu = \mu_0, \ldots, \mu_0+4p \). As the half planes \( H_\mu, \mu = \mu_0, \ldots, \mu_0+4p-1 \), cover a neighbourhood of \( \infty \) on the Riemann surface of \( z^{1/p} \), it is, by \([MR92, \text{Theorem 1.6}]\), sufficient to prove that the entries of \( H^{(\mu+1)}(s) - H^{(\mu)}(s) \) are exponentially small of order \( k_1 \) on \( H_\mu \cap H_{\mu+1}, \mu = \mu_0, \ldots, \mu_0+4p-1 \). If we denote by \( u^{(\mu)}_l(s) \) the \( l \)-th column of \( H^{(\mu)}(s) s^N, l = 1, \ldots, n, \mu = \mu_0, \ldots, \mu_0+4p \), this is equivalent to proving that the differences \( u^{(\mu+1)}_l - u^{(\mu)}_l \) are in \( (A^{\leq-k_1})^n(H_\mu \cap H_{\mu+1}) \).

We have

\[
u^{(\mu+1)}_l - u^{(\mu)}_l(s) = \sum_{m=1}^{n} p_{lm} f_{ml} u^{(\mu)}_m(s),
\]

for some 1-periodic functions \( p_{lm} \) on a neighbourhood of \( \infty \) in \( H_\mu \cap H_{\mu+1} \).

Since \( u^{(\mu+1)}_l(s) - u^{(\mu)}_l(s) \sim 0 \), as \( s \to \infty \) on \( H_\mu \cap H_{\mu+1} \), we have \( p_{lm}(s) f_{ml}(s) \sim 0 \), \( s \to \infty \) on \( H_\mu \cap H_{\mu+1} \), for \( m = 1, \ldots, n \), according to Lemma 7. Lemma 8 now yields that

\[ u^{(\mu+1)}_l(s) - u^{(\mu)}_l(s) \in (A^{\leq-k_1})^n(H_\mu \cap H_{\mu+1}) \]

Applying \([MR92, \text{Theorem 1.6}]\), we conclude that

\[ u^{(\mu)}_l(s) \in (A_{(1/k_1)})^n(H_\mu)[\log s], l = 1, \ldots, n, \mu = \mu_0, \ldots, \mu_0 + 4p - 1, \]

\( H^{(\mu)} \) is a \( k_1 \)-Gevrey function on \( H_\mu \) and the elements of \( \hat{H}(s) \) are Gevrey series of order \( 1/k_1 \). Moreover, it follows that \( \{ H^{(\mu)} | \mu = \mu_0, \ldots, \mu_0 + 4p - 1 \} \) is a representative of \( T^{-1}\hat{H} \).

Finally we prove the statement concerning the functions \( v_l(s) \). It is sufficient to consider the case that \( S \subset (H_{\mu_0} \cup H_{\mu_0+1}). \) There exist 1-periodic functions \( \tilde{p}_{lm} \) analytic on a neighbourhood of \( \infty \) in \( S \cap H_{\mu_0}, \) such that

\[ v_l - u^{(\mu_0)}_l = \sum_{m=1}^{n} \tilde{p}_{lm} f_{ml} u^{(\mu_0)}_m. \]

Since \( v_l(s) - u^{(\mu_0)}_l(s) \sim 0 \), as \( s \to \infty \) on \( S \cap H_{\mu_0} \), Lemma 7 and Lemma 8 now tell us that \( \tilde{p}_{lm} f_{ml} \in A^{\leq-k_1}(S \cap H_{\mu_0}) \), hence

\[ v_l - u^{(\mu_0)}_l \in (A^{\leq-k_1})^n(S \cap H_{\mu_0}). \]

The same holds with \( \mu_0 \) replaced by \( \mu_0 + 1 \). It follows that \( v_l \in (A_{(1/k_1)})^n(S)[\log s] \), what had to be proven. \( \square \)
5. Refinement of chains of solutions.

Consider the fundamental matrix $Y^{(\mu)}(s)$ of Proposition 9 for $\mu \in \mathbb{Z}$. Let its columns be denoted by $\{f_l u_{l,0}^{(\mu)}\}$ for $l = 1, \ldots, n$ as in (2.3). Then $u_{l,0}^{(\mu)} \in (\mathcal{A}_{(1/k_1)})^n(H_\mu)[\log s]$, $u_{l,0}^{(\mu)} \sim \hat{u}_l$ on $H_\mu$ and $\{u_{l,0}^{(\mu)} \mid \mu = \mu_0, \ldots, \mu_0 + 4p - 1\}$ represents the $k_1$-precise quasi-function corresponding to $T^{-1}\hat{u}_l$ on the covering $\{H_\mu \mid \mu = \mu_0, \ldots, \mu_0 + 4p - 1\}$ of $C_p$ for any $\mu_0 \in \mathbb{Z}$.

In this section we show how these $k_1$-precise quasi-functions can be refined to $k_2$-precise quasi-functions with representatives $\{u_{l,1}^{(\mu)}\}$ on an open sector $S(\alpha_1, \beta_1)$ with aperture $> \pi/k_1$ such that $(\alpha_1, \beta_1)$ does not contain a Stokes interval of level $k_1$ of the equation and such that $\{f_l u_{l,1}^{(\mu)}\}_{l=1}^n$ again is a fundamental system of (1.1) with the same asymptotic expansion as before. This will be done by expressing the differences $f_l(u_{l,0}^{(\mu)} - u_{l,0}^{(\mu+1)})$ in terms of suitable fundamental systems and distributing the terms that are $f_l$ times an exponentially small factor of order $k_1$, over $u_{l,0}^{(\mu)}$ and $u_{l,0}^{(\mu+1)}$. The same method can be applied to proceed from $k_j$-precise quasi-functions $u_{l,j-1}^{(\mu)}$ to $k_{j+1}$-precise quasi-functions $u_{l,j}^{(\mu)}$ corresponding to solutions $f_l u_{l,j-1}^{(\mu)}$ and $f_l u_{l,j}^{(\mu)}$ of (1.1).

**Proposition 10.** Let $0 < k_1 < \cdots < k_r = 1$ be the levels in $(0,1]$ of (1.1). Let $j \in \{1, \ldots, r\}$, and define $k := k_j$. If $j < r$, then $k' := k_{j+1}$, otherwise $k' := 1^+$. Let $(\alpha, \beta)$ be an open interval of length $> \pi/k$ not containing a Stokes interval of level $k$ of (1.1). Let $M$ and $N$ be the integers, such that $\left(M - 1\right)\frac{\pi}{2} \leq \alpha < M\frac{\pi}{2} < N\frac{\pi}{2} \leq \beta \leq \left(N + 1\right)\frac{\pi}{2}$. Define $\Gamma_\mu := H_\mu \cap S(\alpha, \beta)$ for $\mu = M, \ldots, N$.

Suppose that we have fundamental systems of solutions $\{f_l u_{l,0}^{(\mu)}\}_{l=1}^n$ on $\Gamma_\mu$ for $\mu = M, \ldots, N$ which satisfy for $l = 1, \ldots, n$:

(i) $u_{l,0}^{(\mu)} \in (\mathcal{A}_{(1/k_1)})^n(\Gamma_\mu)[\log s]$, $u_{l,0}^{(\mu)}(s) \sim \hat{u}_l(s)$, $s \to \infty$ on $\Gamma_\mu$,

(ii) $u_{l,0}^{(\mu+1)} - u_{l,0}^{(\mu)} \in (\mathcal{A}^{<k})^n(H_\mu \cap H_{\mu+1})$.

Then there exist fundamental systems of solutions $\{f_l u_{l,1}^{(\mu)}\}_{l=1}^n$ for $\mu = M, \ldots, N$ such that for $l = 1, \ldots, n$:

\[(5.1) \quad \hat{u}_l^{(\mu)} - u_l^{(\mu)} \in (\mathcal{A}^{<k})^n(\Gamma_\mu), \quad \text{if } \mu \in \{M, \ldots, N\}\]

and

\[(5.2) \quad \hat{u}_l^{(\mu+1)} - \hat{u}_l^{(\mu)} \in (\mathcal{A}^{<k'})^n(H_\mu \cap H_{\mu+1}), \quad \text{if } \mu \in \{M, \ldots, N - 1\}\]

Moreover, for each $l \in \{1, \ldots, n\}$ the family of functions $\{\hat{u}_l^{(\mu)}\}_{\mu=M}^N$ defines an element $\hat{u}_l$ in $(\mathcal{A}_{(1/k_1)}/\mathcal{A}^{<k'})^n(\alpha, \beta)[\log s]$, which is uniquely determined by the properties of the $\hat{u}_l^{(\mu)}$ mentioned above.
We prove the proposition subsequently for the cases \( k \in (0, 1) \) and \( k = 1 \).

**Proof for** \( k \in (0, 1) \).

We introduce the following sets:
\[
St^{-}(\mu, l) = \{ m \mid \kappa_{ml} = k, f_{ml} \in \mathcal{A}^{\leq-k}(\alpha, (\mu+1)\pi/2) \} \text{ if } \mu \in \{M-1, \ldots, N-1\};
\]
\[
St^{+}(\mu, l) = \{ m \mid \kappa_{ml} = k, f_{ml} \in \mathcal{A}^{\leq-k}(\mu\pi/2, \beta) \} \text{ if } \mu \in \{M, \ldots, N\}. \text{ Obviously}
\]

\begin{equation}
St^{-}(\mu + 1, l) \subset St^{-}(\mu, l), \text{ and } St^{+}(\mu - 1, l) \subset St^{+}(\mu, l).
\end{equation}

Because of the assumption that \( \beta - \alpha > \pi/k \) the two sets \( St^{-}(\mu, l) \) and \( St^{+}(\mu, l) \) are disjoint. Since \( f_{jl} = f_{jm}f_{ml} \) the following transitivity relation holds:

\begin{equation}
j \in St^{-}(\mu, m) \land m \in St^{-}(\mu, l) \Rightarrow j \in St^{-}(\mu, l).
\end{equation}

Finally, let \( St(\mu, l) := St^{-}(\mu, l) \cup St^{+}(\mu, l), \mu = M, \ldots, N-1 \). If \( \mu \in \{M, \ldots, N-1\} \) then

\begin{equation}
\kappa_{ml} = k \text{ and } f_{ml} \in \mathcal{A}^{\leq-k}(H_{\mu} \cap H_{\mu+1}) \iff m \in St(\mu, l).
\end{equation}

We only give the proof that the left statement implies the right one since the converse is trivial. The left-hand side implies that \( f_{ml} \) has a Stokes interval \([\sigma - \pi/k, \sigma]\) containing \((\mu\pi/2, (\mu + 1)\pi/2)\). Because of the assumptions of the proposition we have either \( \alpha < \sigma - \pi/k < \beta \leq \sigma \) or \( \sigma - \pi/k \leq \alpha < \sigma < \beta \) and therefore \( m \in St(\mu, l) \).

For \( k < 1 \) the first statement of Proposition 10 is an easy consequence of the following two lemmas.

**Lemma 11.** Under the assumptions of Proposition 10 with \( k < 1 \) there exist fundamental systems \( \{f_{m}u_{l,1}^{(\mu)}\}_{l=1}^{n}, \mu = M, \ldots, N, \) satisfying:

1. \( u_{l,1}^{(\mu)} - u_{l}^{(\mu)} \in (\mathcal{A}^{\leq-k})^{n}(\Gamma_{\mu}) \);
2. \( u_{l,1}^{(\mu + 1)} - u_{l,1}^{(\mu)} = \sum_{m \in St^{-}(\mu, l)} c_{m}^{(\mu)} f_{ml}u_{m,1}^{(\mu)} + \psi_{l,1}^{(\mu)}, \text{ where } c_{m}^{(\mu)} \in \mathbb{C} \) and \( \psi_{l,1}^{(\mu)} \in (\mathcal{A}^{\leq-k})^{n}(H_{\mu} \cap H_{\mu+1}) \).

**Proof.** The proof goes by induction on \( \mu \). Define \( u_{m,1}^{(M)} := u_{m}^{(M)}, m = 1, \ldots, n \). Next assume \( u_{m,1}^{(M)}, \ldots, u_{m,1}^{(\mu)} \) have been defined for all \( m = 1, \ldots, n \) and some \( \mu \in \{M, \ldots, N-1\} \). In the remaining part of this section \( m \) will always be understood to be in \( \{1, \ldots, n\} \). Fix \( l \in \{1, \ldots, n\} \).

From condition (i) of the proposition it follows that \( u_{m,1}^{(\mu)} \sim \hat{u}_{m} \) on \( \Gamma_{\mu} \).

Thus the functions \( f_{m}u_{m,1}^{(\mu)}, m \in St^{-}(\mu, l) \), together with the functions \( f_{m}u_{m,1}^{(\mu + 1)} \), \( m \not\in St^{-}(\mu, l) \), form a fundamental system of solutions according
to Proposition 9. Hence, there exist 1-periodic analytic functions \( p_m^{(\mu)} \), \( p_m^{(\mu+1)} \), 1 \( \leq m \leq n \), such that

\[
(5.6) \quad u_l^{(\mu+1)} - u_l^{(\mu)} = \sum_{m \in \text{St}^-(\mu,l)} p_m^{(\mu)} f_{ml} u_{m,1}^{(\mu)} + \sum_{m \in \text{St}^+(\mu,l)} p_m^{(\mu+1)} f_{ml} u_{m}^{(\mu+1)} + \sum_{m \in \text{St}(\mu,l)} p_m^{(\mu+1)} f_{ml} u_{m}^{(\mu+1)}.
\]

We have \( u_l^{(\mu+1)} - u_l^{(\mu)} = u_l^{(\mu+1)} - u_l^{(\mu)} + u_l^{(\mu)} - u_l^{(\mu+1)} \in (A\leq-k)^n(H_\mu \cap H_{\mu+1}) \), and thus (by Lemma 7) we may conclude that each term in the sums on the right-hand side of (5.6) belongs to this set. Next we apply Lemma 8 to these terms, and find:

- If \( m \in \text{St}^-(\mu,l) \) (resp. \( m \in \text{St}^+(\mu,l) \)): Then \( f_{ml} \) is of level \( k < 1 \), and it is an element of \( A\leq-k(H_\mu \cap H_{\mu+1}) \). So there exist complex constants \( c_m^{(\mu)} \) (resp. \( d_{ml}^{(\mu+1)} \)) such that \((p_m^{(\mu)}(s) - c_m^{(\mu)})f_{ml}(s)\) (resp. \((p_{ml}^{(\mu+1)}(s) - d_{ml}^{(\mu+1)})f_{ml}(s)\)) belong to \( A\leq-1(H) \), where \( H \) is the upper or lower half plane containing \( H_\mu \cap H_{\mu+1} \).

- If \( m \notin \text{St}(\mu,l) \): Then \( \kappa_{ml} > k \) with in all cases \( p_m^{(\mu+1)} f_{ml} \in A\leq-k(H_\mu \cap H_{\mu+1}) \). In the last case Lemma 8 implies \( p_m^{(\mu+1)} f_{ml} \in A\leq-k'(H_\mu \cap H_{\mu+1}) \). This also follows in the first two cases from Lemma 8-2 with \( c = 0 \).

So we have

\[
u_l^{(\mu+1)} - u_l^{(\mu)} = \sum_{m \in \text{St}^-(\mu,l)} c_m^{(\mu)} f_{ml} u_{m,1}^{(\mu)} + \sum_{m \in \text{St}^+(\mu,l)} d_{ml}^{(\mu+1)} f_{ml} u_{m}^{(\mu+1)} + \psi_{l,1}^{(\mu)},
\]

with \( \psi_{l,1}^{(\mu)} \) a function in \( (A\leq-k')^n(H_\mu \cap H_{\mu+1}) \). Obviously, if we define

\[
u_{l,1}^{(\mu+1)} := u_l^{(\mu+1)} - \sum_{m \in \text{St}^+(\mu,l)} d_{ml}^{(\mu+1)} f_{ml} u_{m}^{(\mu+1)},
\]

then \( u_{l,1}^{(\mu+1)} \) satisfies the requirements.

So, if we have constructed \( u_{m,1}^{(\lambda)} \) for \( m \in \{1, \ldots, n\} \) and \( \lambda \in \{M, M + 1, \ldots, \mu\} \) \( (\mu \leq N - 1) \), then we can construct \( u_{l,1}^{(\mu+1)} \) for each \( l \in \{1, \ldots, n\} \) and the lemma follows by induction on \( \mu \).

We next refine the solutions of the previous lemma to solutions which satisfy (5.1) and (5.2) in Proposition 10.

**Lemma 12.** Let \( k < 1 \). Suppose the assumptions of Proposition 10 hold, and furthermore, assume that \( u_l^{(\mu+1)} - u_l^{(\mu)} = \sum_{m \in \text{St}^-(\mu,l)} c_m^{(\mu)} f_{ml} u_{m}^{(\mu)} + \psi_{l}^{(\mu)} \), for some constants \( c_m^{(\mu)} \) and a function \( \psi_{l}^{(\mu)} \in (A\leq-k')^n(H_\mu \cap H_{\mu+1}) \).
Then there exist fundamental systems \( \{ f_t \tilde{u}_l^{(\mu)} \}_{l=1}^{n}, \mu = M, \ldots, N \), such that

(i) \( \tilde{u}_l^{(\mu+1)} - u_l^{(\mu)} = \sum_{m \in St^{-}\mu,l} c_{lm}^{(\mu)} f_{ml} u_m^{(\mu)} \) for some constants \( c_{lm}^{(\mu)} \);

(ii) \( \tilde{u}_l^{(\mu+1)} - u_l^{(\mu)} \in (A^{k'} \leq -k')(H_{\mu} \cap H_{\mu+1}) \).

**Proof.** This lemma can also be proven by induction, but this time we start from the other end of the covering \( \{ \Gamma_{\mu} \}_{\mu=M}^{N} \) of \( S(\alpha, \beta) \): define \( \tilde{u}_m^{(N)} := u_m^{(N)} \) for all \( m \in \{1, \ldots, n\} \). Next suppose \( \tilde{u}_m^{(N)}, \ldots, \tilde{u}_m^{(\mu+1)} \) have been defined and possess the properties of the lemma for all \( m \in \{1, \ldots, n\} \) and some \( \mu \in \{M, \ldots, N-1\} \). Let \( l \in \{1, \ldots, n\} \). We have

\[
\tilde{u}_l^{(\mu+1)} - u_l^{(\mu)} = \tilde{u}_l^{(\mu+1)} - u_l^{(\mu+1)} + u_l^{(\mu+1)} - u_l^{(\mu)} = \sum_{m \in St^{-}\mu+1,l} c_{lm}^{(\mu+1)} f_{ml} u_m^{(\mu+1)} + \sum_{m \in St^{-}\mu,l} c_{lm}^{(\mu)} f_{ml} u_m^{(\mu)} + \psi_l^{(\mu)} .
\]

Furthermore,

\[
u_m^{(\mu+1)} - u_m^{(\mu)} = \sum_{j \in St^{-}\mu,m} c_{mj}^{(\mu)} f_{jm} u_j^{(\mu)} + \psi_m^{(\mu)} .
\]

From these two relations and properties (5.3) and (5.4) we obtain

\[
\tilde{u}_l^{(\mu+1)} - u_l^{(\mu)} = \sum_{m \in St^{-}\mu,l} c_{lm}^{(\mu)} f_{ml} u_m^{(\mu)} + \tilde{\psi}_l^{(\mu)} ,
\]

where the \( c_{lm}^{(\mu)} \) are constants in \( C \) and \( \tilde{\psi}_l^{(\mu)} \in (A^{k'} \leq -k')(H_{\mu} \cap H_{\mu+1}) \).

If we define

\[
\tilde{u}_l^{(\mu)} := u_l^{(\mu)} + \sum_{m \in St^{-}\mu,l} c_{lm}^{(\mu)} f_{ml} u_m^{(\mu)} ,
\]

then \( \tilde{u}_l^{(\mu)} \) satisfies the requirements of the lemma. Again the lemma follows by induction on \( \mu \).

The first statement of Proposition 10 follows from the previous lemmas and the last statement is a direct consequence of the relative Watson lemma referred to after Definition 2. We could also prove the uniqueness directly, without reference to this lemma, along the same lines as in [BIS].

**Proof for \( k = 1 \).**

Recall that the Stokes intervals of level 1 of Equation (1.1) with formal solution (2.3) are, by definition, the Stokes intervals of the functions \( e^{2\pi isj} f_{ml}(s), j \in \mathbb{Z}, l, m \in \{1, \ldots, n\} \), that are of level 1. Taking \( l = m \) and...
\[ j \neq 0, \text{ we find that } [(h - 1)\pi, h\pi] \text{ is a Stokes interval of level 1 of the equation for any } h \in \mathbb{Z}. \text{ Hence, due to the assumption that } (\alpha, \beta) \text{ does not contain a Stokes interval of level 1, the sector } S(\alpha, \beta) \text{ does not contain a lift of both the positive and the negative real axis. From this and } \alpha < M\pi/2 < N\pi/2 < \beta \text{ it follows that } N - M \leq 2 \text{ and at least one of the integers } M \text{ and } N \text{ has to be odd.}

We will prove the proposition for the cases } M = -1 \text{ and } N = 0 \text{ or } N = 1, \text{ that is } \alpha \in [-\pi, -\pi/2), \beta \in (0, \pi]. \text{ The other cases can be proven similarly.}

Let } l \in \{1, \ldots, n\}. \text{ There exist 1-periodic analytic functions } p_m^{(-1)}(s), \ m = 1, \ldots, n, \text{ such that}

\[ (5.7) \quad \tilde{u}_l^{(0)} - u_l^{(-1)} = \sum_{m=1}^{n} p_m^{(-1)} f_{ml} u_m^{(-1)}. \]

By assumption (ii) of Proposition 10, we have \( u_l^{(0)} - u_l^{(-1)} \in (A^{\leq-1})^n(-\pi/2, 0). \) So, by Lemma 7, \( p_m^{(-1)} f_{ml} \in A^{\leq-1}(-\pi/2, 0), \) \( m \in \{1, \ldots, n\}. \) From Lemma 8 we conclude that

\[ (5.8) \quad p_m^{(-1)} f_{ml} \in A^{\leq-1}(-\pi, 0) \text{ if } \kappa_{ml} < 1, \]

and

\[ (5.9) \quad p_m^{(-1)} f_{ml} \in A^{\leq-1^+}(-\pi/2, 0) \text{ if } \kappa_{ml} = 1^+. \]

If \( \kappa_{ml} = 1 \) and \( p_m^{(-1)} \neq 0 \) then according to Lemma 8-3 we have \( \Re b_{ml} \geq 0. \) Moreover, this lemma tells us that

\[ (5.10) \quad p_m^{(-1)} f_{ml} \in A^{\leq-1}(-\pi, 0) \text{ if } \kappa_{ml} = 1, \Re b_{ml} = 0. \]

Let \( t_1, \ldots, t_\nu \) denote the numbers \( \Re b_h, \ h \in \{1, \ldots, n\} \) in decreasing order of magnitude. We will use induction on \( \tau \in \{1, \ldots, \nu\}. \)

If \( \Re b_{t_\tau} = t_1, \) then \( \Re b_{ml} \leq 0 \) for all \( m \in \{1, \ldots, n\}. \) So if \( \kappa_{ml} = 1 \) and \( p_m^{(-1)} \neq 0 \) then we know already that \( \Re b_{ml} \geq 0 \) and so \( \Re b_{ml} = 0 \) and (5.10) applies. Therefore if \( \Re b_{t_\tau} = t_1 \) we define

\[ \hat{u}_l^{(-1)} = u_l^{(-1)} + \sum_{\kappa_{ml} \leq 1} p_m^{(-1)} f_{ml} u_m^{(-1)}, \hat{u}_l^{(0)} = u_l^{(0)}. \]

Then from (5.7), (5.8), (5.9) and (5.10) it follows that \( f_{\mu} \hat{u}_l^{(\mu)} \) are solutions of (1.1) with

\[ (5.11) \quad \hat{u}_l^{(\mu)} - u_l^{(\mu)} \in (A^{\leq-1})^n(\Gamma_\mu) \text{ if } \mu \in \{-1, 0\}; \]

\[ \hat{u}_l^{(0)} - \hat{u}_l^{(-1)} \in (A^{\leq-1^+})^n(-\pi/2, 0). \]

Next let \( \tau \in \{2, \ldots, \nu\}, \Re b_{t_\tau} = t_\tau \) and suppose that for all \( m \in I(l) := \{m \in \{1, \ldots, n\}|\kappa_{ml} = 1, \Re b_{ml} > 0\} \) the functions \( \hat{u}_m^{(-1)} \) and \( \hat{u}_m^{(0)} \) have
already been defined such that (5.11) holds with \( l \) replaced by \( m \). We have
\[
   u_t^{(0)} - u_t^{(-1)} = \sum_{m \notin I(l)} \tilde{p}_m^{(-1)} f_{ml} \tilde{u}_m^{(1)} + \sum_{m \in I(l)} \tilde{p}_m^{(-1)} f_{ml} \tilde{u}_m^{(0)},
\]
for some 1-periodic functions \( \tilde{p}_m^{(-1)}(s) \) analytic on a neighbourhood of \( \infty \) in \( H_{-1} \). As before we have
\[
   (5.12) \quad \tilde{p}_m^{(-1)} f_{ml} \in \mathcal{A}^{\leq -1}(-\pi/2,0).
\]
Now we have analogues of (5.8), (5.9) and (5.10), and so the first sum can be written as \( \tilde{u}_l^{(-1)} = \tilde{u}_l^{(0)} - \sum_{m \in I(l)} p_m^{\pm} f_{ml} \tilde{u}_m^{(0)} \),
and therefore
\[
   \tilde{u}_l^{(0)} := \tilde{u}_l^{(-1)} + \tilde{u}_l^{(1)} + \sum_{m \in I(l)} p_m^{\pm} f_{ml} \tilde{u}_m^{(0)},
\]
Next consider the case that \( m \in I(l) \) and \( p_m^{(-1)} \neq 0 \). Then according to Lemma 8-3 there exist analytic 1-periodic functions \( p_m^\pm \) such that \( \tilde{p}_m^{(-1)} = \tilde{p}_m^+ + \tilde{p}_m^- \) and \( p_m^+ f_{ml} \in \mathcal{A}^{\leq -1}(\alpha,0) \) and \( p_m^- f_{ml} \in \mathcal{A}^{\leq -1}(-\pi/2,\beta) \). Now define
\[
   \tilde{u}_l^{(-1)} := u_l^{(1)} - \psi_l^{(-1)} + \sum_{m \in I(l)} p_m^- f_{ml} \tilde{u}_m^{(0)},
\]
and therefore
\[
   \tilde{u}_l^{(0)} := u_l^{(0)} - \sum_{m \in I(l)} p_m^+ f_{ml} \tilde{u}_m^{(0)}.
\]
Then \( \tilde{u}_l^{(0)} - \tilde{u}_l^{(-1)} = \psi_l^{(-1)} + \sum_{m \in I(l)} \tilde{p}_m^{+} f_{ml} (\tilde{u}_m^{(1)} - \tilde{u}_m^{(0)}) \in (\mathcal{A}^{\leq -1})^n(-\pi/2,0) \) and it follows that the functions \( \tilde{u}_l^{(n)} \) satisfy (5.11). By induction (5.11) follows for all \( l \). So in case \( N = 0 \) the proposition has been proved.

Next suppose that \( N = 1 \), so \( \beta > \pi/2 \). If \( p_m^{(1)}(s), m = 1, \ldots, n, \) are the 1-periodic functions analytic on a neighbourhood of \( \infty \) in the upper half plane such that
\[
   u_l^{(1)} - \tilde{u}_l^{(0)} = \sum_{m=1}^{n} p_m^{(1)} f_{ml} u_m^{(1)},
\]
then \( p_m^{(1)} f_{ml} \in \mathcal{A}^{\leq -1}(0,\pi/2) \). If \( \kappa_m < 1 \) then as before \( p_m^{(1)} f_{ml} \in \mathcal{A}^{\leq -1}(H_1) \). Next suppose \( \kappa_m = 1 \) and \( p_m^{(1)} \neq 0 \). Then by Lemma 8-3 there exists an integer \( N \) such that \( p_m^{(1)}(s) = p_N e^{2\pi i s N}(1 + o(1)) \) as \( s \to \infty \) with \( p_N \neq 0 \) and therefore \( e^{2\pi i s N} f_{ml}(s) \in \mathcal{A}^{\leq -1}(0,\pi/2) \). So there exists a Stokes interval \( [\sigma_m-\pi,\sigma_m] \) of \( e^{2\pi i s N} f_{ml} \) which contains \( (0,\pi/2) \). Now \( \sigma_m - \pi \geq -\pi/2 > \alpha \) and therefore \( \sigma_m > \beta \). Hence \( p_m^{(1)} f_{ml} \in \mathcal{A}^{\leq -1}(0,\beta) \). Moreover, if \( \kappa_m = 1^+ \), then \( p_m^{(1)} f_{ml} \in \mathcal{A}^{\leq -1+}(0,\pi/2) \) according to Lemma 8. Thus, if we define
\[
   \tilde{u}_l^{(1)} := u_l^{(1)} - \sum_{m=1}^{\kappa_m \leq 1} p_m^{(1)} f_{ml} u_m^{(1)},
\]
then
\[
   \tilde{u}_l^{(1)} - u_l^{(1)} \in (\mathcal{A}^{\leq -1})^n(0,\beta); \quad \tilde{u}_l^{(1)} - \tilde{u}_l^{(0)} \in (\mathcal{A}^{\leq -1+})^n(0,\pi/2).
\]
The fundamental systems \( \{f_i u_i^{(\mu)}\}_{i=1}^n \), \( \mu = -1, 0, 1 \), thus obtained satisfy (5.1) and (5.2).

The uniqueness property of Proposition 10 is an immediate consequence of a more general form of the relative Watson lemma by Malgrange and Ramis, that can be found in [BIS].  

\[ \square \]

6. Equations without level \( 1^+ \).

Theorem 13 has already been stated and proven in [BF96], but here we present a new proof.

**Theorem 13.** Let \( 0 < k_1 < \cdots < k_r = 1 \) be the levels in \((0, 1)\) of Equation (1.1), and suppose that this equation does not contain a level \( 1^+ \) (i.e., \( d_{ml} = 0, \forall m, l \in \{1, \ldots, n\} \)). Let \( H(s)s^{\Lambda_s}e^{G(s)}s^k \) be a formal fundamental matrix as in (1.2).

Let \( S_i = S(\alpha_i, \beta_i), i = 1, \ldots, r \), be a sequence of open sectors such that \( S_1 \supset \cdots \supset S_r \). \( S_1 \) has aperture less than \( 2\pi \), \( S_i \) has aperture larger than \( \pi/k_i \) and \((\alpha_i, \beta_i)\) does not contain a Stokes interval of level \( k_i, i = 1, \ldots, r \).

Then \( \tilde{H} \) is \((k_1, \ldots, k_r)\)-summable on \((S_1, \ldots, S_r)\) with sum \( \tilde{H} \) such that \( H_r(s)s^{\Lambda_s}e^{G(s)}s^1 \) is an analytic fundamental matrix of (1.1).

**Proof.** Define \( M_j, N_j, j = 1, \ldots, r \), to be the integers such that \((M_j - 1)\pi/2 < \alpha_j < M_j\pi/2 < N_j\pi/2 < \beta_j \leq (N_j + 1)\pi/2 \) and let \( \Gamma_{j,\mu} := S_j \cap H_{\mu}, \mu = M_j, \ldots, N_j, j = 1, \ldots, r \). Also, let \( S_0 := C_p \) be the Riemann surface of \( s^{1/p}, \Gamma_{0,\mu} := H_{\mu}, \mu = M_0, \ldots, N_0 \) where \( M_0 := M_1, N_0 := M_0 + 4p - 1 \).

By Proposition 9 we have a representative \( \{H^{(\mu)}(s)\}_{\mu=M_0}^{N_0} \) of \( T^{-1}\tilde{H}(s) \) on the covering \( \{H_{\mu}\}_{\mu=M_0}^{N_0} \) of \( C_p \) such that \( H^{(\mu)}(s)s^{\Lambda_s}e^{G(s)}s^1 \) is an analytic fundamental matrix of (1.1). To show that the columns \( \tilde{h}_l \) of \( \tilde{H} \) are multi-summable we have to construct \( h_{l,j} \in (A_{(1/k_i)}/A_{<k_{j+1}})^n(S_j), j = 0, \ldots, r \) such that \( h_{l,j}|_{S_{j+1}} \equiv h_{l,j+1} \mod A_{<k_{j+1}}, j = 0, \ldots, r \) if \( k_{r+1} = \infty \).

Let \( U_0^{(\mu)}(s) := H^{(\mu)}(s)s^N \) so that \( U_0^{(\mu)}(s)F(s) \) is a fundamental matrix of (1.1) (cf. (2.2)) and let \( \tilde{u}_{l,0}^{(\mu)} \) denote the \( l \)th column of \( U_0^{(\mu)} \). The construction mentioned above is equivalent to the construction of functions \( \{\tilde{u}_{l,j}^{(\mu)}\}_{\mu=M_j}^{N_j} \) for \( j = 1, \ldots, r \) and \( l = 1, \ldots, n \) such that:

1. \( \{f_i \tilde{u}_{l,j}^{(\mu)}\}_{i=1}^n \) is a fundamental system of Equation (1.1),
2. \( \{\tilde{u}_{l,j}^{(\mu)}\}_{\mu=M_j}^{N_j} \) represents a \( k_{j+1} \)-precise quasi-function \( \tilde{u}_{l,j} \in (A_{(1/k_i)}/A_{<k_{j+1}})^n(S_j)[\log s], j = 1, \ldots, r \),
3. \( \tilde{u}_{l,j-1}|_{S_j} \equiv \tilde{u}_{l,j} \mod (A_{<k_j})^n, j = 1, \ldots, r \).

Suppose we have constructed \( \tilde{u}_{l,i} \) for \( i = 0, \ldots, j - 1 \), for some \( j \in \{1, \ldots, r\} \). Then we can apply Proposition 10 with \( \alpha = \alpha_j, \beta = \beta_j \), and
with \( u^{(\mu)}_t = \hat{u}^{(\mu)}_{t,j-1} |_{\Gamma_{j,\mu}} \), \( \mu = M_j, \ldots, N_j \). Defining \( \hat{u}^{(\mu)}_{t,j} := \hat{u}^{(\mu)}_t , l = 1, \ldots, n \), \( \mu = M_j, \ldots, N_j \) we see that properties (1), (2) and (3) are satisfied for \( i = j \) as well. So they are satisfied for all \( j \in \{1, \ldots, r\} \).

We have \( \hat{u}^{(\mu+1)}_{t,r} - \hat{u}^{(\mu)}_{t,r} \in (\mathcal{A}_{-1}^\pm 1)^n(\Gamma_{r,\mu} \cap \Gamma_{r,\mu+1}) \). We also have \( \hat{u}^{(\mu+1)}_{t,r} - \hat{u}^{(\mu)}_{t,r} = \sum_{m=1}^n p_m f_{ml} \hat{u}^{(\mu)}_{m,r} \) for some 1-periodic analytic functions \( p_m \). Lemma 7 now tells us that each \( p_m f_{ml} \in \mathcal{A}_{-1}^\pm 1(\Gamma_{\mu,r} \cap \Gamma_{\mu+1,r}) \), and then it follows from Lemma 8 and the fact that the equation has no level 1*, that \( p_m = 0 \), for all \( m \). Hence, the functions \( \hat{u}^{(\mu)}_{m,r} \), \( \mu = M_r, \ldots, N_r \), are the restrictions of an analytic function \( \hat{u}_{t,r} \in (\mathcal{A}_1^{(1/k_1)})^n(S_r)[\log s] \).

\[ \square \]

7. Equations with level 1*.

In this section we will consider Equation (1.1) under the assumption that there does exist a pair \((m, l)\) such that \( d_m \neq d_l \) in the notation of (2.3); that is, the equation possesses the level 1*. We will show in this section that we can still assign a uniquely characterizable fundamental system \( \mathcal{Y}(s) \) with asymptotic expansion \( \hat{Y}(s) \) in appropriate regions of the Riemann surface of the logarithm, provided \( \Re b_l \neq \Re b_m \) if \( b_l \neq b_m \) (cf. notation in (2.3)).

Before we state the main result of this paper (Theorem 18), we need to define the Stokes numbers of level 1* of the equation. The Stokes number of a function \( f \) of level 1* of the form

\[ f(s) = \exp(ds \log s + 2\pi ibs + q(s) + \gamma \log s) \],

occuring in formal solutions of equations possessing a level 1*, is associated with curves that separate regions of growth from regions of decay of \( f \). We have

\[ \Re\{ds \log s + 2\pi ibs\} = d\Re s \log |s| - (d \arg s + 2\pi \Re b) \Re s - 2\pi \Im b \Re s \],

and therefore the main contribution to \( |f(s)| \) comes from \( \exp[\Re\{ds(d \log s + 2\pi ibs)\}] \). Let \( h \) be an even integer if \( d < 0 \) and an odd integer if \( d > 0 \). Then \( f \in \mathcal{A}_1^{\pm 1}(H_{2h}) \). Moreover, \( f \) behaves as an exponential function of order 1 in vertical strips, \( f \) becomes exponentially large on any open sector containing \( H_{2h} \) and, if \( S_+(h) := S((h-\frac{1}{2})\pi, (h+\frac{1}{2})\pi) \) and \( S_-(h) := S((h-\frac{1}{2})\pi, (h+\frac{1}{2})\pi) \) then it is easily verified that

\[ f \in \mathcal{A}_{-1}^{\pm 1}(S_\pm) \] iff \( \pm (h + 2\Re b/d) < -1/2 \),

where the upper (lower) signs belong together.

Let \( \{f_{j,u_j}\}_{j=1}^n \) be a fundamental system of (1.1) such that \( u_j \sim \hat{u}_j \) as \( s \to \infty \) on \( S_\pm(h) \). Assume \( d_{ml} = d_m - d_l < 0 \) and \( h \) is even. Then there exists \( N_\pm \in \mathbb{Z} \) such that \( \pm \{h + 2(\Re b_{ml} + N_\pm)/d_{ml}\} < -1/2 \). Then \( e^{2\pi i s N_\pm} f_{ml} \in \mathcal{A}_{-1}^{\pm 1}(S_\pm) \) and therefore the solutions \( f_t u \) and \( f_t u + e^{2\pi i s N_\pm} f_{ml} u_{ml} \) have the same asymptotic behaviour on \( S_\pm(h) \). So in this case it is not possible to
characterize fundamental systems $Y(s)$ by their asymptotic behaviour $\hat{Y}(s)$ on $S_\pm(h)$.

In order for a fundamental system to be in some way uniquely determined by its asymptotic expansion in a sector of $C_\infty$, which contains an open right or left half plane, this sector therefore should contain the closure of this half plane. However, such fundamental systems do not exist in general (see [vdPS97, Chapter 11]).

Hence we have to characterize fundamental systems by their asymptotic behaviour in a more complicated type of region. This region should contain a neighbourhood of $\infty$ in some half plane $H_{2h}$ but not a neighborhood of $\infty$ in $\overline{H}_{2h}$. In the case $d = (-1)^{h+1}$ a suitable region is given by

**Definition 14.** For $\theta \in \mathbb{R}$, $h \in \mathbb{Z}$, we define

$$D(h; \theta) := \{s \in C_\infty \mid (h - 1)\pi < \arg s < (h + 1)\pi; \quad (-1)^h R\{s(\log s + i\theta)\} > 1\}.$$ 

If $h \in \mathbb{Z}$ and $\theta_1, \theta_2 \in \mathbb{R}$ then $D(h; \theta_1, \theta_2) := D(h; \theta_1) \cup D(h; \theta_2)$.

We denote the boundary of $D(h; \theta)$ by $C(h; \theta)$. We have $D(h; \theta) \subset \bigcup_{j=1}^{\frac{h}{2} - 1} H_{2h+j}$ and se$^{h\pi} \in D(h; \theta) \iff s \in D(0; \theta + h\pi)$. Similarly with $D$ replaced by $C$. Details on $C(0; \theta)$ can be found in [Imm84] and [Imm91]. We have $\Re s = O(3s/\log |s|)$ as $s \to \infty$ on $C(h; \theta)$. This implies that $\arg s \to \pm \pi/2 \mod 2\pi$ as $3s \to \pm \infty$, $s \in C(h; \theta)$.

If $h \in \mathbb{Z}$ and $\theta \in \mathbb{R}$, then the following properties are easily established.

- Let $\hat{\theta} < \theta$. As $(-1)^h R\{s(\log s + i\theta)\} = (-1)^h R\{s(\log s + i\theta)\} - \Re (\hat{\theta} - \theta)3s$, it follows that:
  - i) $D(h; \theta) \cap H_{2h-1} \subset D(h; \theta) \cap H_{2h-1}$, and that
  - ii) $D(h; \hat{\theta}) \cap H_{2h+1} \supset D(h; \theta) \cap H_{2h+1}$.

One could say that the regions $D(h; \theta)$ 'rotate' (modulo some deformation) clockwise with increasing $\theta$.

- $D(h; \theta) \cap H_{2h}$ is a neighbourhood of $\infty$ in $H_{2h}$. However, since for any $s_0 \in \overline{H}_{2h}, s_0 + \overline{H}_{2h} \not\subset D(h; \theta)$, the intersection $D(h; \theta) \cap \overline{H}_{2h}$ is not a neighbourhood of $\infty$ in $\overline{H}_{2h}$.

Let $D := D(h; \theta_1, \theta_2)$, for some $h \in \mathbb{Z}$ and $\theta_1 < \theta_2$. A set $U \subset D$ is called a **neighbourhood of** $\infty$ in $D$, if, for any $\theta \in (\theta_1, \theta_2)$, there exists $s_0 \in D$ such that $s_0 + D(h; \theta) \subset U$.

Suppose $g$ is an analytic function on a neighbourhood $U$ of $\infty$ in $D$, and suppose there exist a $k > 0$ and a series $\hat{g}(s) = \sum_{n \geq 0} g_n s^{-n/p}$ such that for any $\theta \in (\theta_1, \theta_2)$ we have

$$\left| g(s) - \sum_{n=0}^{N-1} g_n s^{-n/p} \right| \leq K A |s|^{-N/p}, \quad (7.2)$$
\( \forall s \in D(h; \theta) \cap U, \forall N \in \mathbb{N} \),

for some positive constants \( K \) and \( A \), which only depend on \( \theta \). Then we call \( g \) a \( k \)-Gevrey function on \( D \) (with respect to \( \frac{1}{p} \mathbb{N}_0 \)), and the set of these functions is denoted by \( \mathcal{A}_{(1/k)}(D) \).

Concerning 1-Gevrey functions on \( D \) we have the following theorem by Immink (cf. [Imm96]):

**Theorem 15.** Suppose \( \theta_1, \theta_2 \in \mathbb{R} \), \( \theta_1 < \theta_2 \), and \( h \in \mathbb{Z} \). Let \( D := D(h; \theta_1, \theta_2) \).

If \( g \in \mathcal{A}_{(1)}(D) \), then \( g \) is uniquely determined by its asymptotic expansion \( \hat{g} \).

In [Imm96] it is shown that \( g \) is already uniquely determined by its asymptotic expansion if (7.2) with \( k = 1 \) holds on \( D(h; \theta) \) for one \( \theta \in (\theta_1, \theta_2) \).

Let \( f \) be a function of level \( 1^+ \) given by (7.1) and \( \theta_1 < 2\pi \Re b / d < \theta_2 \).

Let \( h \in \mathbb{Z} \) be odd if \( d > 0 \) and even if \( d < 0 \). Then there exist positive constants \( K \) and \( c \) such that \( |f(s)| \leq Ke^{-c|s|} \) for \( s \in D(h; \theta_1) \cap D(h; \theta_2) \) and \( |s| \) sufficiently large. From this it follows that (7.2) holds on this set with \( g := f, g_n = 0 \), and \( k = 1 \). This is not true if \( \frac{2\pi \Re b}{d} \not\in [\theta_1, \theta_2] \). Therefore we introduce the following definition:

**Definition 16.** Let \( f(s) \) be the function of level \( 1^+ \) given by (7.1). Then we call \( \frac{2\pi \Re b}{d} \) its Stokes number. The Stokes numbers of level \( 1^+ \) of the equation (1.1) are the Stokes numbers of the functions \( e^{2\pi i j s} f_{ml}(s), l, m \in \{1, \ldots, n\}, j \in \mathbb{Z} \), that are of level \( 1^+ \). That is, \( \theta_1 < \theta_2 \) and \( \alpha < \beta \). Let \( D := D(h; \theta_1, \theta_2) \) and \( S := S(\alpha, \beta) \).

Assume \( D \cap S \neq \emptyset \). We define a neighbourhood of \( \infty \) in \( D \cap S \) to be an open set \( U \) in \( D \cap S \) such that for any \( \theta, \gamma, \delta \) satisfying \( \theta_1 < \theta_2 \) and \( \alpha < \gamma < \delta < \beta \), there exists \( s_0 \in D \cap S \) such that \( s_0 + (D(h; \theta) \cap S(\gamma, \delta)) \subset U \). We write \( \mathcal{A}^{(1)}(D \cap S) \) for the set of functions that are analytic on a neighbourhood of \( \infty \) in \( D \cap S \), and exponentially small of order 1, as \( s \to \infty \), uniformly on \( D(h; \theta) \cap S(\gamma, \delta) \), for any \( \theta, \gamma, \delta \) as above.

We extend Definition 2 of multisummability as follows:

**Definition 17.** Let \( 0 < k_1 < \cdots < k_{r-1} < k_r = 1 \), and define \( k_{r+1} = 1^+ \).

Let \( S_1 \supset \ldots \supset S_r \) be a nested sequence of sectors \( S_i \) with aperture \( > \pi/k_i \), \( i = 1, \ldots, r \), aperture of \( S_1 \) at most \( 2p\pi \) and assume \( S_r \supset S((h - \frac{1}{2})\pi - \varepsilon, (h + \frac{1}{2})\pi + \varepsilon) \), for some \( h \in \mathbb{Z} \) and some \( \varepsilon > 0 \). Finally, let \( D := D(h; \theta_1, \theta_2) \) for some \( \theta_1 < \theta_2 \).

A formal power series \( \hat{f} \in C[[s^{-1/p}]]_{1/(p k_1)} \) is called \( (k_1, \ldots, k_r, 1^+) \)-summable on \( (S_1, \ldots, S_r, D) \) with \( (k_1, \ldots, k_r, 1^+) \)-sum \( f \in \mathcal{A}_{(1/k_1)}(D) \),
if there exist quasi-functions \( f_i \in (A_{(1/k_i)}/A_{-k_i}) (S_i) \), \( i = 1, \ldots, r \), satisfying:

\[
f_i|_{S_{i+1}} \equiv f_{i+1} \mod A_{-k_i}, \quad i = 0, \ldots, r - 1 \text{ with } f_0 = T^{-1} \hat{f};
\]

\( f_r \) has a representative \( \{f_{r,\omega}\}_{\omega \in \Omega} \) with respect

to a covering \( \{S_{r,\omega}\}_{\omega \in \Omega} \) of \( S_r \) with open sectors \( S_{r,\omega} \)
such that \( f_{r,\omega} - f \in A_{-1}(D \cap S_{r,\omega}), \forall \omega \in \Omega \).

Let \( g \) be another function such that \( f_{r,\omega} - g \in A_{-1}(D \cap S_{r,\omega}), \forall \omega \in \Omega \). Then \( f - g \in A_{-1}(D \cap S_{r,\omega}), \forall \omega \in \Omega \), hence, \( f - g \in A_{-1}(D \cap S_r) = A_{-1}(D) \). Theorem 15 implies \( f = g \) and it follows that \( f \) is uniquely determined by \( f_r \) and \( D \). By the relative Watson lemma ([MR92, Prop. (2.1)]) \( f_i \) is uniquely determined by \( f_{i-1} \) and \( S_i, i = r, \ldots, 1 \). Hence the sum \( f \) is uniquely determined by \( f \) and \( (S_1, \ldots, S_r, D) \).

Similarly to Definition 2 we extend this definition to the case that \( \hat{f} \) is an \( n \)-vector or \( n \times n \)-matrix with elements in \( \mathbb{C}[[s^{-1/p}]]_{1/(p_1)} \).

The main result of this paper is the following theorem.

**Theorem 18.** Let \( k_1 < \ldots < k_r = 1 \) be the sequence of positive levels \( \leq 1 \) of (1.1). Suppose that \( 1^+ \) is a level of (1.1) and that \( \Re b_i \neq \Re b_m \) if \( b_i \neq b_m \) where \( b_i \) is defined below (2.3). Let \( \tilde{H}(s) s^{\Lambda(s)} G(s) s^{L} \) be a formal fundamental matrix as in (1.2).

Let \( S_i = S(\alpha_i, \beta_i), i = 1, \ldots, r, \) be a sequence of open sectors, such that \( S_1 \supset \ldots \supset S_r \) and \( \beta_i - \alpha_i \leq 2\pi, \beta_i - \alpha_i > \pi/k_i \) and \( (\alpha_i, \beta_i) \) does not contain a Stokes interval of level \( k_i, i = 1, \ldots, r \) of (1.1). Moreover, suppose that \( (h-1)\pi < \alpha_r < (h-\frac{1}{2})\pi \) and \( (h+\frac{1}{2})\pi < \beta_r < (h+1)\pi, \) for some \( h \in \mathbb{Z} \).

Let \( \hat{\theta}, \theta \in \mathbb{R}, \hat{\theta} < \theta, \) such that \( (\hat{\theta}, \theta) \) does not contain a Stokes number of level \( 1^+ \) and define \( D = D(h; \hat{\theta}, \theta) \).

Then \( \tilde{H} \) is \((k_1, \ldots, k_r, 1^+)-\text{summable} \) on \((S_1, \ldots, S_r, D)\) with \((k_1, \ldots, k_r, 1^+)-\text{sum} \) \( H \) such that (1.3) defines an analytic fundamental matrix of (1.1).

For the proof of this theorem we will use the following lemma which extends the results of Lemma 8-4.

**Lemma 19.** Let \( h \in \mathbb{Z} \) and \( Q_\pm \) be the quadrant \( H_{2h} \cap H_{2h \pm 1} \). Here and in the following the upper signs belong together and so do the lower signs. Let \( f \) be given by (7.1). Define \( \theta_j := \frac{2\pi}{d} (\Re b + j), j \in \mathbb{Z}, \) and \( D := D(h; \theta_{N-1}, \theta_N), \) for some \( N \in \mathbb{Z} \).

Suppose that \( p(s) \neq 0 \) is a \( 1 \)-periodic analytic function on a neighbourhood of \( \infty \) in \( H_{2h \pm 1} \) such that \( p(s)f(s) \in A_{-1}^+(Q_\pm) \).

Then there exists a \( 1 \)-periodic function \( p_-(s) \), such that \( p_-(s) \) is analytic on a neighbourhood of \( \infty \) in \( H_{2h \pm 1} \), \( p_+(s) := p(s) - p_-(s) \) is an entire
function, and
\[ p_-(s)f(s) \in \mathcal{A}^{-1}(D \cap H_{2h\pm 1}), \]
\[ p_+(s)f(s) \in \mathcal{A}^{-1}(D \cap H_{2h+1}) \cap \mathcal{A}^{1+}(H_{2h}). \]

Proof. We will give the details of the proof for the lower sign and \( h \) is even. The other cases can be proven in a similar way.

Now \( Q_- \) is a fourth quadrant and \( H := H_{2h-1} \) is a lower half plane. According to Lemma 8-4 we have \( p(s)f(s) \in \mathcal{A}^{-1+}((h-\frac{1}{2})\pi, h\pi], \) and, since \( p(s) \neq 0 \) also \( f(s) \in \mathcal{A}^{-1+}((h-\frac{1}{2})\pi, h\pi). \) Therefore \( d < 0, \) \( f(s) \in \mathcal{A}^{-1+}((h-\frac{1}{2})\pi, (h+\frac{1}{2})\pi) \) and \( \theta_N < \theta_{N-1}. \)

We have an expansion for \( p(s) \) as in (3.1) for \( \Im s < -\rho \) for some \( \rho > 0. \) Let \( p_-(s) := \sum_{j \leq N} p_j e^{2\pi isj}. \) Then \( p_-(s) \) is analytic for \( \Im s < -\rho \) and \( p_-(s) = e^{2\pi is(N-1)}O(1), s \to \infty \) on \( \partial H. \) Moreover, \( p_+(s) = \sum_{j \geq N} p_j e^{2\pi isj} \) is an entire function.

First consider \( p_-(s)f(s). \) The properties of \( p_- \) imply
\[ p_-(s)f(s) \in \mathcal{A}^{1+}((h-\frac{1}{2})\pi, h\pi] \subset \mathcal{A}^{1+}(Q_-). \]

For \( \epsilon > 0, \) let \( S_\epsilon := S((h-\frac{1}{2})\pi - \epsilon, (h-\frac{1}{2})\pi + \epsilon). \) In order to prove that \( p_-(s)f(s) \in \mathcal{A}^{-1}(D \cap H) \) it is sufficient to show, that with any \( \theta \in (\theta_N, \theta_{N-1}) \) we can find positive constants \( K, c \) and \( \epsilon, s \), such that
\[ |p_-(s)f(s)| \leq Ke^{-c|s|}, s \in D(h; \theta) \cap S_\epsilon, \Im s < -\rho. \]

As \( p_-(s)f(s) = O(1) \exp(ds \log s + 2\pi is(b + N - 1) + o(s)), s \to \infty \) on \( H, \) it is sufficient to prove that
\[ \Re\{ds \log s + 2\pi is(b + N - 1)\} \leq -c|s|, \forall s \in D(h; \theta) \cap S_\epsilon, \]
for some \( c > 0. \)

So let \( \theta \in (\theta_N, \theta_{N-1}). \) On \( D(h; \theta) \) we have:
\[ \Re\{ds \log s + 2\pi is(b + N - 1)\} \]
\[ = d\Re\{s(\log s + i\theta) + is(\frac{2\pi}{d}(b + N - 1) - \theta)\} \]
\[ < -d[3s(\theta_{N-1} - \theta) + \frac{2\pi}{d} \Re s \Re b] \]
\[ = -d\sin(\arg s)(\theta_{N-1} - \theta)|s| \left( 1 + \frac{2\pi \Im b}{d(\theta_{N-1} - \theta)} \cot(\arg s) \right). \]

We have \( d\sin(\arg s)(\theta_{N-1} - \theta) > 0 \) on \( H. \) Furthermore, there exists an \( \epsilon > 0 \) such that \( \left| \frac{4\pi \Im b}{d(\theta_{N-1} - \theta)} \cot(\arg s) \right| < 1, \forall s \in S_\epsilon. \) Thus (7.5) and (7.4) follow.

Next consider \( p_+(s)f(s). \) As \( p(s)f(s), p_-(s)f(s) \in \mathcal{A}^{1+}((h-\frac{1}{2})\pi, h\pi], \) also \( p_+(s)f(s) \) belongs to this set. Moreover, \( f(s) \in \mathcal{A}^{1+}(H_{2h}) \) and \( |p_+(s)| \leq Ke^{-2\pi N\Im s}, \Im s \geq -\rho \) for some \( K > 0, \) and therefore \( p_+(s)f(s) \in \mathcal{A}^{1+}(H_{2h}). \) To prove the lemma it now suffices to prove that \( p_+f \in \)
$A^{\leq -1}(D \cap H_{2h+1})$, hence that for any $\theta \in (\theta_N, \theta_{N-1})$ we can find positive constants $K,c$ and $\varepsilon$, such that

$$|p_+(s)f(s)| \leq Ke^{-c|s|}, \forall s \in D(h;\theta) \cap S((h + \frac{1}{2})\pi - \varepsilon, (h + \frac{1}{2})\pi + \varepsilon).$$

A proof of this inequality runs along the same lines as that of (7.4). $\square$

The following proposition extends the results of Proposition 10.

**Proposition 20.** Let $k_1 < \cdots < k_r$ be the positive levels $\leq 1$ of equation (1.1) and assume that (1.1) has a level $1^+$. Let $M$ be odd. Assume $\alpha < M\pi/2$, $\beta > (M + 2)\pi/2$ such that $I = (\alpha, \beta)$ does not contain a Stokes interval of level 1 of (1.1). Define $\Gamma = H_{\mu} \cap S(\alpha, \beta)$ for $\mu = M, M + 1, M + 2$. Assume $\hat{\theta}, \theta \in \mathbb{R}$, $\hat{\theta} < \theta$, such that $(\theta, \theta)$ does not contain any Stokes number of level $1^+$ of (1.1). Define $D := D(\frac{1}{2}(M + 1); \hat{\theta}, \theta)$.

Assume we have fundamental systems $\{f_1 u^{(\mu)}_l\}_{l=1}^n$, $\mu = M, M + 1, M + 2$ such that for all $l \in \{1, \ldots, n\}$:

(i) $u^{(\mu)}_l(\alpha, \beta) \in (A^{1/2})^{n}(\Gamma) [\log s]$, $u^{(\mu)}_l \sim \hat{u}_l$ on $\Gamma$ if $\mu = M, M + 1, M + 2$,

(ii) $u^{(\mu+1)}_l - u^{(\mu)}_l(\alpha, \beta) \in (A^{\leq -1^+})^{n}(\Gamma \cap \Gamma_{\mu+1})$ if $\mu = M, M + 1$.

Then there exists a fundamental system $\{f_1 \hat{u}_l\}_{l=1}^n$ of equation (1.1) such that for all $l \in \{1, \ldots, n\}$:

(7.6) $\hat{u}_l - u^{(\mu)}_l \in (A^{\leq -1})^{n}(D \cap H_{\mu})$ if $\mu = M, M + 2$,

(7.7) $\hat{u}_l - u^{(M+1)}_l \in (A^{\leq -1^+})^{n}(H_{M+1})$.

Moreover, for each $l \in \{1, \ldots, n\}$ the function $\hat{u}_l \in (A^{1/2})^{n}(D) [\log s]$ is uniquely determined by these properties.

**Remark.** We can find $\alpha$ and $\beta$ satisfying the above conditions if and only if $(\nu \pi - \pi/2, \nu \pi + \pi/2)$ is not a Stokes interval of level 1 for any $\nu \in \mathbb{Z}$. This corresponds to the condition that $\Re b_l \neq \Re b_m$ if $b_l \neq b_m$.

**Proof.** Throughout the proof $l, m \in \{1, \ldots, n\}$. Let $h := (M + 1)/2$, so $h$ is an integer. We will write $m \prec l$ if $f_{ml} \in A^{\leq -1^+}(H_{2h})$, and $I(l) := \{m \mid m \prec l\}$. The relation $\prec$ gives a partial ordering on $\{1, \ldots, n\}$. We will prove the proposition by induction with respect to this partial ordering.

Let $p^{(\mu)}_m$, $\mu = M, M + 1$, be the 1-periodic analytic functions such that

(7.8) $u^{(\mu+1)}_l - u^{(\mu)}_l = \sum_{m=1}^n p^{(\mu)}_m f_{ml} u^{(\mu)}_m$, $\mu = M, M + 1$.

As $u^{(\mu+1)}_l - u^{(\mu)}_l \in (A^{\leq -1^+})^{n}(\Gamma \cap \Gamma_{\mu+1})$ by assumption, it follows from Lemma 7 that each of the summands must belong to this set. If $m \notin I(l)$ then $c^{d_m s \log s}$ is unbounded on $H_{2h}$, so $\Re d_m s$ is positive on $H_{2h}$. Therefore
for all \( m \) of Lemmas 7 and 8 we may conclude as above that

\[
A := \Gamma_q \cap \Gamma_{q+1},
\]

so that

\[
\tilde{u}_m := \Gamma_q \cap \Gamma_{q+1} \cap A_{\lambda} := \Gamma_q \cap \Gamma_{q+1} \cap A_{\lambda} = \Gamma_q \cap \Gamma_{q+1} \cap A_{\lambda}
\]

First let \( l \) be such that \( I(l) = \emptyset \). Then \( u_l^{(\mu)} \) is independent of \( \mu \in \{ M, M + 1, M + 2 \} \) and \( u_l = u_l^{(\mu)} \) satisfies (7.6) and (7.7).

Now let \( l \) be such that \( \tilde{u}_m \) have been defined, and satisfy (7.6) and (7.7) for all \( m \in I(l) \). The functions \( f_m \tilde{u}_m, m \in I(l) \), together with the functions \( f_m u_m^{(M+1)}, m \notin I(l) \), form a fundamental system of solutions. With the aid of Lemmas 7 and 8 we may conclude as above that

\[
u_l^{(M+1)} = \sum_{m \in I(l)} f_m \tilde{u}_m,
\]

for some 1-periodic analytic functions \( \tilde{u}_m \) on a neighbourhood of \( \infty \) in \( Q := \Gamma_M \cap \Gamma_M + 1 \), and \( f_m \tilde{u}_m \in A^{1-1}(Q) \). According to Lemma 19 the functions \( \tilde{u}_m = \Gamma_M \cap \Gamma_M + 1 \) and \( \tilde{u}_m \in A^{1-1}(D \cap H_M) \) and \( \tilde{u}_m \in A^{1-1}(D \cap H_{M+2}) \cap A^{1-1}(H_{M+1}) \). We define

\[
u_{l,1} := \nu_l^{(M)} + \sum_{m \in I(l)} \tilde{u}_m
\]

so that \( \nu_{l,1} = \nu_{l,1}^{(M+1)} \in (A^{1-1}_M \cap D \cap H_M) \) and

\[
u_{l,1} = \nu_{l,1}^{(M+1)} - \sum_{m \in I(l)} f_m \tilde{u}_m \in \left( A^{1-1}_M \cap D \cap H_{M+2} \right) \cap \left( A^{1-1}_M \cap D \cap H_{M+1} \right).
\]

From these relations and assumption (ii) it follows that \( \nu_{l,1}^{(M+2)} = \nu_{l,1}^{(M+1)} \in (A^{1-1}_M \cap D \cap H_{M+2}) \) and as above we find that there exist 1-periodic functions \( \tilde{u}_m = \Gamma_M \cap \Gamma_M + 1 \) and \( \tilde{u}_m = \Gamma_M \cap \Gamma_M + 1 \), such that

\[
u_{l,1}^{(M+2)} = \nu_{l,1}^{(M+1)} = \sum_{m \in I(l)} \left( \tilde{u}_m + \tilde{u}_m \right),
\]

where the first term of each summand belongs to \( (A^{1-1}_M \cap D \cap H_{M+2}) \) and the second one to \( (A^{1-1}_M \cap D \cap H_M) \cap A^{1-1}_M \cap D \cap H_{M+1} \). Hence, if we define

\[
u_l := \nu_l^{(M+2)} - \sum_{m \in I(l)} \tilde{u}_m
\]

then

\[
u_l = \nu_{l,1} + \sum_{m \in I(l)} \tilde{u}_m.
\]
and it is easy to verify that $\tilde{u}_l$ satisfies (7.6) and (7.7). The uniqueness of $\tilde{u}_l$ follows from Theorem 15.

Theorem 18 can be proved similarly to Theorem 13, with the aid of Propositions 9, 10 and 20.

References


[Imm96] ---, *A particular type of summability of divergent power series, with an application to linear difference equations*, Preprint, 1996.


Received October 22, 1998.

**University of Groningen**
Department of Mathematics, P.O.Box 800
9700 AV Groningen
The Netherlands
E-mail address: B.L.J.Braaksma@math.rug.nl

**University of Groningen**
Department of Mathematics, P.O.Box 800
9700 AV Groningen
The Netherlands
E-mail address: bernard@math.rug.nl

**University of Groningen**
Department of Econometrics, P.O.Box 800
9700 AV Groningen
The Netherlands
E-mail address: G.K.Immink@eco.rug.nl
Theorem 2 below). The proof of the converse statement crucially depends on the manifold structure, in particular on the fact that two geodesics coincide if they do so on an open interval. Hence, the converse statement fails in our context due to the bifurcation property of geodesics. A counterexample exhibiting this failure is provided in Section 4 below. A geodesic $\gamma$ is called recurrent if there exists a sequence $\{t_n\} \subset \mathbb{R}, t_n \to \infty$ such that $t_n \gamma \to \gamma$ as $t_n \to \infty$. Convergence in this definition is meant to be uniform convergence on compact sets which, in fact, induces the topology on the space $GX$ consisting of all (local) isometries $\mathbb{R} \to X$ when $X$ is (not) simply connected. $\mathbb{R}$ acts on $GX$ by right translations, namely, $(t, g) \to tg$, where $tg : \mathbb{R} \to X$ is the geodesic defined by $tg(s) = g(s + t), s \in \mathbb{R}$. This action is simply the geodesic flow. The notion of convergence in the above definition is analogous to the tangential condition which defines recurrence in the manifold case. We use the notion of approximation given in Definition 6 below which was introduced in [1] in order to characterize recurrent geodesics in hyperbolic manifolds.

$X$ will always denote a locally compact, complete, geodesic metric space with curvature less than or equal to $\chi, \chi < 0$. Recall that a geodesic metric space is said to have curvature less than or equal to $\chi$ if each $x \in X$ has a neighborhood $V_x$ such that every geodesic triangle of perimeter strictly

\begin{align*}
\text{PACIFIC JOURNAL OF MATHEMATICS} \\
\text{Vol. 195, No. 1, 2000} \\
\text{APPROXIMATION OF RECURRENCE IN NEGATIVELY CURVED METRIC SPACES} \\
\text{Charalambos Charitos and Georgios Tsapogas} \\
\text{For metric spaces with curvature less than or equal to $\chi$, $\chi < 0$, it is shown that a recurrent geodesic is approximated by closed geodesics. A counter example is provided for the converse.}
\end{align*}
less than \( \frac{2\pi}{\sqrt{\chi}} \) \((=+\infty \text{ when } \chi \leq 0) \) contained in \( V_x \) satisfies \( \text{CAT}-(\chi) \) inequality (see [11] for definitions and basic properties). We will denote the metric by \( d(\cdot,\cdot) \) and will use the same letter to denote distance when the metric space to which we refer is understood. All curves are assumed to be parametrized by arclength. A \textit{geodesic segment} in \( X \) is an isometry \( c : I \to X \), where \( I \) is a closed interval in \( \mathbb{R} \). A \textit{geodesic} in \( X \) is a map \( c : \mathbb{R} \to X \) such that for each closed interval \( I \subset \mathbb{R} \), the map \( c|_I : I \to X \) is a geodesic segment. A \textit{local geodesic segment} (usually called \textit{geodesic arc}) in \( X \) is a map \( c : I \to X \) such that for each \( t \in I \) there is an \( \varepsilon > 0 \) such that \( c|_{[t-\varepsilon,t+\varepsilon]} \cap I : [t-\varepsilon,t+\varepsilon] \cap I \to X \) is a geodesic segment. Similarly, a \textit{local geodesic} \( \mathbb{R} \to X \) is defined. A \textit{closed geodesic} in \( X \) is a local geodesic \( c : \mathbb{R} \to X \) which is a periodic map.

\textbf{Definition 1.} An oriented geodesic \( g \) in \( X \) is said to be approximated by closed geodesics if, for every \( \varepsilon > 0 \) and every \( x \in \text{Im } g \), there exists a closed oriented geodesic \( c \) such that for some point \( y \in \text{Im } c \),

\[ d(c(t + t_y), g(t + t_x)) < \varepsilon \]

for all \( t \in [0, \text{period } (c)] \), where \( t_x, t_y \in \mathbb{R} \) with \( x = g(t_x) \) and \( y = c(t_y) \).

The following theorem is the main result of this paper.

\textbf{Theorem 2.} Let \( X \) be a locally compact, complete, geodesic metric space which has curvature less than or equal to \( \chi, \chi < 0 \). If a geodesic or geodesic ray in \( X \) is recurrent, then it is approximated by closed geodesics.

The proof of Theorem 2 uses the notion of quasi-geodesic and its stability properties. We will closely follow notation and terminology appearing in [8, Ch. 3] where we refer the reader for first definitions and basic properties of quasi-geodesics. Here we only recall the following definition.

\textbf{Definition 3.} Let \( f : [a,b] \to X \) be a continuous map with \(-\infty \leq a \leq b \leq +\infty \) and \( \lambda, \kappa, L \) real numbers with \( \lambda \geq 1, \kappa \geq 0, L > 0 \).

\( f \) is a \((\lambda, \kappa, L)\)-\textit{quasi-geodesic} if for every subinterval \([a',b']\) of \([a,b]\) satisfying

\[ \text{length } f \left( [a',b'] \right) \leq L, \]

the following inequality holds

\[ \text{length } f \left( [a',b'] \right) \leq \lambda d(f(a'), f(b')) + \kappa. \]

The next proposition is a well know fact for \( \text{CAT}-(\chi) \) spaces. We include a short proof of it, since it is difficult to find exact reference (when \( X \) is a geometric polyhedron this result follows from [3, p. 403]).

\textbf{Proposition 4.} Let \( M \) be a complete geodesic space satisfying \( \text{CAT}-(\chi) \) inequality with \( \chi < 0 \). Every local geodesic segment in \( M \) is a geodesic segment.
Recurrence in Negatively Curved Spaces

Proof. Let \( \delta : [0, L] \to M \) be a local geodesic segment in \( M, L > 0 \). Set 

\[
l = \sup \left\{ t \in [0, L] \mid \delta \mid_{[0,t]} \text{ is a geodesic segment} \right\}.
\]

Apparently, \( l > 0 \) and by completeness of \( M, \delta \mid_{[0,l]} \) is a geodesic segment joining \( \delta (0) \) with \( \delta (l) \). Assuming the conclusion is not true, i.e., \( l < L \), let \( \varepsilon \) be a positive number such that \( \delta \mid_{[l-\varepsilon, l+\varepsilon]} \) is a geodesic segment. Denote by \( \bar{\delta} \mid_{[0, l+\varepsilon]} \) the geodesic segment in \( M \) joining \( \delta (0) \) with \( \delta (l + \varepsilon) \). Since \( \delta \mid_{[0, l+\varepsilon]} \) is not the geodesic segment joining \( \delta (0) \) with \( \delta (l + \varepsilon) \),

\[
(1) \quad d (\delta (0) , \delta (l + \varepsilon)) < d (\delta (0) , \delta (l)) + d (\delta (l) , \delta (l + \varepsilon)) .
\]

The points \( \delta (0), \delta (l) \) and \( \delta (l + \varepsilon) \) define a geodesic triangle in \( M \). Denote by \( \Delta = \left( \bar{\delta} (0), \bar{\delta} (l), \bar{\delta} (l + \varepsilon) \right) \) the corresponding comparison triangle which is non-degenerate by inequality (1). Choose points \( B \) on \( \delta \mid_{[0, l]} \) and \( B' \) on \( \delta \mid_{[l, l+\varepsilon]} \) such that \( d (B, \bar{\delta} (l)) = d (B', \bar{\delta} (l)) = \varepsilon' < \varepsilon \) and denote by \( \overline{BB} \) and \( \overline{BB'} \) the corresponding points on the comparison triangle. Then by (1) the angle of \( \Delta \) at \( \bar{\delta} (l) \) is smaller than \( \pi \) and therefore

\[
d (\overline{BB'}, \overline{BB}) < d (\overline{BB}, \overline{BB'}) + d (\overline{BB'} , \overline{BB}) = 2\varepsilon'.
\]

By comparison, \( d (B, B') \leq d (\overline{BB}, \overline{BB'}) \) so we obtain

\[
d (B, B') < d (B, \delta (l)) + d (\delta (l) , B') .
\]

This contradicts the fact that \( \delta \mid_{[l-\varepsilon', l+\varepsilon']} \) is a geodesic segment. \( \square \)

Let \( \tilde{X} \) be the universal cover of \( X \) and \( p : \tilde{X} \to X \) the projection map. \( \tilde{X} \) becomes a metric space as follows: Given \( \tilde{x}, \tilde{y} \in \tilde{X} \) choose any curve \( \tilde{c} : [a, b] \to \tilde{X} \) with \( \tilde{c} (a) = \tilde{x} \) and \( \tilde{c} (b) = \tilde{y} \) and define the distance from \( \tilde{x} \) to \( \tilde{y} \) to be the length of the unique length minimizing curve in the homotopy class of \( \tilde{p}\tilde{c} \) with endpoints fixed. For the existence of the length minimizing curve see [10]. This distance function is a metric on \( \tilde{X} \) which inherits the properties of \( X \), namely, \( \tilde{X} \) becomes a complete geodesic locally compact (hence, proper) metric space. \( \pi_1 (X) \) acts on \( \tilde{X} \) and the action commutes with \( p \). As the projection \( p \) is a local isometry, it follows that \( \pi_1 (X) \) acts on \( \tilde{X} \) by local isometries. Using the fact that \( \tilde{X} \) is geodesic and Proposition 4, it is routine to show that \( \pi_1 (X) \) acts on \( \tilde{X} \) by isometries. In addition, \( \tilde{X} \) has curvature less than or equal to \( \chi, \chi < 0 \) and, by a theorem of Gromov (see for example [11, p. 325]), \( \tilde{X} \) satisfies \( CAT - (\chi) \) inequality.

\( GX \) is by definition the space of all local geodesics \( \mathbb{R} \to X \) and, by Proposition 4 above, \( G\tilde{X} \) is the space consisting of all global geodesics \( \mathbb{R} \to \tilde{X} \). The topology on these spaces is uniform convergence on compact sets. The boundary \( \partial \tilde{X} \) can be defined using either equivalence classes of sequences or, equivalence classes of geodesic rays. The local compactness assumption on \( X \) implies that \( \tilde{X} \) is proper and hence the two definitions coincide (see
We will be using them interchangeably. For any two distinct points \( \xi, \eta \) in \( \partial \widetilde{X} \) there exists a unique, up to parametrization, (oriented) geodesic \( g \) with \( g(-\infty) = \xi \) and \( g(\infty) = \eta \) (see for example [5, Prop. 2]). We need the following lemma which asserts that the projection of a point onto a geodesic always exists.

**Lemma 5.** Let \( g \) be a geodesic in \( G\widetilde{X} \) (or a geodesic segment) and \( x_0 \) a point in \( \widetilde{X} \). There exists a unique real number \( s \) such that \( g(s) \) realizes the distance of \( x_0 \) from \( \text{Im} \, g \), i.e., \( \text{dist} \,(x_0, \text{Im} \, g) = d \left( x_0, g(s) \right) \).

**Proof.** We may assume that \( x_0 \notin \text{Im} \, g \). Existence is apparent. Assume that \( s \neq s' \) are two such numbers. The points \( g(s), g(s') \) and \( x_0 \) define a non-degenerate geodesic triangle in \( \widetilde{X} \) and denote by \( \Delta = \left( g(s), g(s'), x_0 \right) \) the corresponding comparison triangle. \( \Delta \) is an equilateral triangle in the unique complete simply connected Riemannian 2-manifold of constant sectional curvature \( \chi \). Hence, the angles of \( \Delta \) at \( g(s) \) and \( g(s') \) are each less than \( \pi/2 \). Therefore, there exists a point \( g(t) \) on the side of \( \Delta \) opposite to \( x_0 \) such that \( d \left( x_0, g(t) \right) < d \left( x_0, g(s) \right) = d \left( x_0, g(s') \right) \). By CAT - (\( \chi \)) inequality, \( d \left( x_0, g(t) \right) \leq d \left( x_0, g(s) \right) \), a contradiction. \( \square \)

**Remark 1.** If \( c \in GX \) is a closed geodesic and \( x_0 \in X \), the same argument applied to a lifting \( \widetilde{c} \) of \( c \) shows that there exists a unique point \( B \in \text{Im} \, c \) such that \( d \left( x_0, B \right) = \text{dist} \,(x_0, \text{Im} \, c) \).

**Remark 2.** Set \( \partial^2 \widetilde{X} = \left\{ (\xi, \eta) \in \partial \widetilde{X} \times \partial \widetilde{X} : \xi \neq \eta \right\} \) and let \( \rho : G\widetilde{X} \to \partial^2 \widetilde{X} \) be the fiber bundle given by \( \rho(g) = (g(-\infty), g(+\infty)) \). Since for any two distinct points \( \xi, \eta \) in \( \partial \widetilde{X} \) there exists a unique (oriented) geodesic \( g \) with \( g(-\infty) = \xi \) and \( g(\infty) = \eta \) (see for example [5, Prop. 2]), the fiber of \( \rho \) is \( \mathbb{R} \). Moreover, this bundle is trivial (see for example [4, Th. 4.8]). To define a trivialization, let \( x_0 \) be a base point and let

\[
(2) \quad H : G\widetilde{X} \xrightarrow{\sim} \partial^2 \widetilde{X} \times \mathbb{R}
\]

be the trivialization of \( \rho \) with respect to \( x_0 \) defined by

\[
H(g) = (g(-\infty), g(+\infty), s)
\]

where \( -s \) is the real number provided by Lemma 5. Note that the composite of the geodesic flow \( \mathbb{R} \times G\widetilde{X} \to G\widetilde{X} \) with \( H \) is given by the formula

\[
(\xi_1, \xi_2, s) \to (\xi_1, \xi_2, s + t)
\]

for all \( (\xi_1, \xi_2) \in \partial^2 \widetilde{X} \) and \( s \in \mathbb{R} \).
2. Recurrent geodesics.

**Definition 6.** A geodesic \( \gamma \) in \( X \) is called recurrent if there exists a sequence \( \{t_n\} \subset \mathbb{R} \), \( t_n \to \infty \) such that \( t_n \gamma \to \gamma \) as \( t_n \to \infty \).

For a recurrent geodesic \( \gamma \) in \( X \) there exists a sequence of closed (in fact, piece-wise geodesic) curves \( \{\gamma_n\}_{n \in \mathbb{N}} \), associated to \( \gamma \) as follows: Fix a convex neighborhood \( U \) of \( \gamma(0) \), i.e., a neighborhood which satisfies the following property: For all \( x, y \in U \) there exists a unique geodesic segment with endpoints \( x \) and \( y \) lying entirely in \( U \). Such a neighborhood exists (see for example \([2]\)). If \( \{t_n\} \) is the sequence given by Definition 6 above and \( \varepsilon_n = d(\gamma(0), \gamma(t_n)) \), let \( K \in \mathbb{N} \) such that \( \gamma(t_n) \in U \) for all \( n \geq K \). Define \( \gamma_n, n \geq K \) to be the curve

\[
\gamma_n : [0, t_n + \varepsilon_n] \to X
\]

with \( \gamma_n(t) = \gamma(t) \) \( \forall t \in [0, t_n] \) and \( \gamma_n|_{[t_n, t_n + \varepsilon_n]} \) the unique geodesic segment in \( U \) joining \( \gamma(t_n) \) with \( \gamma(0) \). Note that \( t_n + \varepsilon_n \) is the period of the closed curve \( \gamma_n \). In the sequel, we will refer to these closed curves by writing \( \gamma_n, n \in \mathbb{N} \) but it will always be implicit that \( n \) is large enough so that \( \gamma_n \) are defined.

Using the following lemma, we may assume that given a recurrent geodesic \( \gamma \), the associated closed curves \( \{\gamma_n\}_{n \in \mathbb{N}} \) are not homotopic to a point.

**Lemma 7.** Given a recurrent geodesic \( \gamma \) there exists \( M \in \mathbb{N} \) such that each closed curve \( \gamma_n, n \in \mathbb{N} \) associated to \( \gamma \) is not homotopic to a point, provided \( n \geq M \).

**Proof.** Let \( \widetilde{\gamma} \) be a lift of \( \gamma \) to the universal cover \( \widetilde{X} \) of \( X \) parametrized so that \( \widetilde{\gamma}(0) \) projects to \( \gamma(0) = \gamma_n(0) \). The curve \( \gamma_n|_{[0, t_n]} \) is a local geodesic segment and, by Proposition 4, its lift \( \widetilde{\gamma}_n|_{[0, t_n]} \) to \( \widetilde{X} \) starting at \( \widetilde{\gamma}(0) \) is a geodesic segment. Moreover, \( \gamma_n|_{[t_n, t_n + \varepsilon_n]} \) and its lift \( \widetilde{\gamma}_n|_{[t_n, t_n + \varepsilon_n]} \) to \( \widetilde{X} \) starting at \( \widetilde{\gamma}_n(t_n) \) are both geodesic segments. We have

\[
d(\widetilde{\gamma}_n(t_n + \varepsilon_n), \widetilde{\gamma}_n(0)) \geq d(\widetilde{\gamma}_n(t_n), \widetilde{\gamma}_n(t_n)) = t_n - \varepsilon_n.
\]

Since \( \varepsilon_n \to 0 \) and \( t_n \to \infty \) as \( n \to \infty \), we may choose \( M \in \mathbb{N} \) such that \( \widetilde{\gamma}_n(t_n + \varepsilon_n), \widetilde{\gamma}_n(0) \) are distinct for all \( n \geq M \). Therefore, \( \widetilde{\gamma}_n|_{[t_n, t_n + \varepsilon_n]} \), which is the lift of the closed curve \( \gamma_n \) starting at \( \widetilde{\gamma}(0) = \widetilde{\gamma}_n(0) \), has distinct endpoints and, therefore, \( \gamma_n, n \geq M \) is not homotopic to a point. \( \square \)

The following proposition shows that the lifts (to the universal cover \( \widetilde{X} \)) of the closed curves \( \gamma_n \) associated to a recurrent geodesic \( \gamma \) are, for \( n \) large enough, quasi-geodesics with arbitrarily large \( L \). Recall that a \( CAT - (\chi) \) space is a \( \delta \)-hyperbolic space in the sense of Gromov (see for example \([11]\),
This applies to the universal covering $\tilde{X}$, since it satisfies $\text{CAT}-(\chi)$ inequality globally. Let $\delta$ denote the hyperbolicity constant of the space $\tilde{X}$.

**Proposition 8.** Let $\gamma$ be a recurrent geodesic in $X$ and $\{\gamma_n\}_{n \in \mathbb{N}}$ the associated closed curves. For every $L > 0$, there exists $N \in \mathbb{N}$ such that all lifts $\tilde{\gamma}_n : \mathbb{R} \to \tilde{X}$ of $\gamma_n$ with $n \geq N$ are $(\lambda, \kappa, L)$-quasi-geodesics provided $\kappa > 16\delta$ and $\lambda = 1$, where $\delta$ is the hyperbolicity constant of $X$.

**Proof.** Let $\gamma$ be a recurrent geodesic and $L > 0$ be given. The sequence $\{t_n\}$ given by Definition 6 converges to infinity. Moreover, $\varepsilon_n = d(\gamma(0), \gamma(t_n)) \to 0$ and $t_n + \varepsilon_n = \text{period}(\gamma_n)$ also converges to infinity as $n \to \infty$. Hence, we may choose $N$ such that

$$\begin{align*}
(4) \quad t_n + \varepsilon_n > L \quad \text{and} \quad \varepsilon_n < \frac{1}{2}(\kappa - 16\delta) \quad \text{for all } n \geq N.
\end{align*}$$

Let now $[a, b]$ be any interval with $b - a < L$ (cf. Definition 3). For each $n \geq N$ there exists an integer $k_n$ such that

$$\begin{align*}
(5) \quad \tilde{\gamma}_n([a, b]) \subset \tilde{\gamma}_n\left(\left[\left(k_n - 1\right)(t_n + \varepsilon_n), k_n(t_n + \varepsilon_n) + t_n\right]\right).
\end{align*}$$

Denote by $[\tilde{\gamma}_n(a), \tilde{\gamma}_n(b)]$ the unique geodesic segment in $\tilde{X}$ joining $\tilde{\gamma}_n(a)$ with $\tilde{\gamma}_n(b)$ and set

$$\begin{align*}
A_n & := \tilde{\gamma}_n(k_n(t_n + \varepsilon_n) - \varepsilon_n), \\
y_{k_n} & := \tilde{\gamma}_n(k_n(t_n + \varepsilon_n)).
\end{align*}$$

The distance of any point on $[\tilde{\gamma}_n(a), \tilde{\gamma}_n(b)]$ from $\tilde{\gamma}_n([a, b])$ is bounded by a number which depends on the hyperbolicity constant $\delta$ of the space $\tilde{X}$ and on the number of geodesic segments which constitute $\tilde{\gamma}_n([a, b])$, see [8, Lemma 1.5, p. 25]. In our case here, $\tilde{\gamma}_n([a, b])$ consists of at most three geodesic segments (since the right hand side of inclusion (5) above consists of 3 geodesic segments) and the bound is $8\delta$. Hence we have

$$\begin{align*}
(6) \quad d(y_{k_n}, [\tilde{\gamma}_n(a), \tilde{\gamma}_n(b)]) \leq 8\delta.
\end{align*}$$

By Lemma 5, let $B_n$ be the point on $[\tilde{\gamma}_n(a), \tilde{\gamma}_n(b)]$ which realizes the distance in the left hand side of inequality 6. Assume that neither $\tilde{\gamma}_n(a)$ nor $\tilde{\gamma}_n(b)$ lies on the geodesic segment $[A_n, y_{k_n}]$. Then we have the following triangle inequalities

$$\begin{align*}
&d(\tilde{\gamma}_n(a), A_n) \leq d(\tilde{\gamma}_n(a), B_n) + d(B_n, y_{k_n}) + d(y_{k_n}, A_n) \\
&d(y_{k_n}, \tilde{\gamma}_n(b)) \leq d(y_{k_n}, B_n) + d(B_n, \tilde{\gamma}_n(b))
\end{align*}$$

which, after employing the fact that $d(A_n, y_{k_n}) = \varepsilon_n$, become

$$\begin{align*}
\text{length } \tilde{\gamma}_n([a, b]) & = d(\tilde{\gamma}_n(a), A_n) + d(A_n, y_{k_n}) + d(y_{k_n}, \tilde{\gamma}_n(b)) \\
& \leq 2\varepsilon_n + 2d(y_{k_n}, B_n) + d(\tilde{\gamma}_n(a), \tilde{\gamma}_n(b)) \\
& \leq 2\varepsilon_n + 2 \cdot 8\delta + d(\tilde{\gamma}_n(a), \tilde{\gamma}_n(b))
\end{align*}$$

by inequality (6)

$$\begin{align*}
\text{by inequality (4)} \quad & \leq \kappa + d(\tilde{\gamma}_n(a), \tilde{\gamma}_n(b)).
\end{align*}$$
The case where \( \tilde{\gamma}_n(a) \) and/or \( \tilde{\gamma}_n(b) \) lies on \([A_n,y_{k_n}]\) is treated similarly. \(\square\)

**Corollary 9.** For \( n \in \mathbb{N} \) sufficiently large, the isometry of \( \tilde{X} \) in \( \pi_1(X) \) which corresponds to the homotopy class of the closed curve \( \gamma \) is hyperbolic.

**Proof.** It suffices to show that each \( \tilde{\gamma}_n : \mathbb{R} \to \tilde{X} \) determines exactly two boundary points \( \tilde{\gamma}_n(-\infty), \tilde{\gamma}_n(+\infty) \). By Lemma 8 there exists an \( M \in \mathbb{N} \) such that \( \tilde{\gamma}_n \) is a quasi-isometry for all \( n \geq M \). Each such \( \tilde{\gamma}_n \) induces a map \( \partial \mathbb{R} \to \partial \tilde{X} \) which is a homeomorphism onto its image, see [8, Th. 2.2, p. 35]. As \( \partial \mathbb{R} \) consists of two distinct points, \( \tilde{\gamma}_n(-\infty), \tilde{\gamma}_n(+\infty) \in \partial \tilde{X} \) are also distinct for all \( n \geq M \). \(\square\)

It now follows that a recurrent geodesic \( \gamma \) in \( X \) as well as each of the (oriented) closed curves \( \gamma_n, n \geq M \) (cf. Lemma 7 and Corollary 9 above) determine exactly two boundary points in \( \partial \tilde{X} \) denoted by \( \tilde{\gamma}(-\infty), \tilde{\gamma}(+\infty) \) and \( \tilde{\gamma}_n(-\infty), \tilde{\gamma}_n(+\infty) \) respectively. We need the following lemma concerning these boundary points. Recall that \( \tilde{X} \cup \partial \tilde{X} \) is a compact space which is metrizable (see [8, p. 134]), and we will denote such metric by \( d_{\tilde{X} \cup \partial \tilde{X}} \).

**Lemma 10.** \( \tilde{\gamma}_n(-\infty) \to \tilde{\gamma}(-\infty) \) and \( \tilde{\gamma}_n(+\infty) \to \tilde{\gamma}(+\infty) \) as \( n \to \infty \).

**Proof.** As above, let \( \varepsilon_n = \text{length}(\text{Im} \gamma_n) - t_n \) so that \( t_n + \varepsilon_n \) is the period of \( \gamma_n \). We first show that \( \tilde{\gamma}_n(+\infty) \to \tilde{\gamma}(+\infty) \). Consider the sequence \( \tilde{\gamma}_n(k(t_n + \varepsilon_n)), k \in \mathbb{N} \) which converges to \( \tilde{\gamma}_n(+\infty) \) as \( k \to \infty \). Thus, there exists \( k_n \in \mathbb{N} \) such that

\[
(7) \quad d_{\tilde{X} \cup \partial \tilde{X}}(\tilde{\gamma}_n(k(t_n + \varepsilon_n)), \tilde{\gamma}_n(+\infty)) < 1/n.
\]

Now consider the sequences \( y_n := \tilde{\gamma}_n(k(t_n + \varepsilon_n)) \) and \( x_n := \tilde{\gamma}(t_n), n \in \mathbb{N} \). Since \( x_n \to \tilde{\gamma}(+\infty) \), by inequality (7) above it is enough to show that the sequences \( \{x_n\} \) and \( \{y_n\} \) represent the same element in \( \partial \tilde{X} \) or, in other words, that the hyperbolic product \( (x_n,y_n)_{x_0} \) with respect to the base point \( x_0 := \tilde{\gamma}(0) \) converges to \( +\infty \) as \( n \to +\infty \). For the notion of hyperbolic product of sequences and their equivalence, see [8].

The stability property of quasi-geodesics states (see Corollary 1.10 of [8, p. 31]) that given any two numbers \( \kappa \geq 0 \) and \( \lambda \geq 1 \), there exists a constant \( C \) depending on \( \lambda, \kappa \) and on the hyperbolicity constant \( \delta \) of the space such that if \( L \) is bigger than \( 2C \) then every \( (\lambda, \kappa, L) \)-quasi-geodesic \( f : [a, b] \to \tilde{X} \) lies within a \( C \)-neighborhood of the geodesic segment \([f(a), f(b)]\) · By choosing \( \lambda = 1, \kappa > 16\delta \) where \( \delta \) is the hyperbolicity constant of the space \( \tilde{X} \) and \( L > 2C \) we obtain, by Proposition 8 above, a natural number \( N \) such that all \( \tilde{\gamma}_n : \mathbb{R} \to \tilde{X} \) with \( n \geq N \) are \( (\lambda, \kappa, L) \)-quasi-geodesics. In particular, \( \tilde{\gamma}_n : [0, k_n(t_n + \varepsilon_n)] \to \tilde{X} \) are \( (\lambda, \kappa, L) \)-quasi-geodesics for all \( n \geq N \). Therefore, by Corollary 1.10 of [8, p. 31] as explained above,

\[
d(x_n, x'_n) < C \quad \forall \ n \geq N
\]
where \( x'_n \) denotes the projection of \( x_n \) on the geodesic segment \([\gamma(0), y_n]\) (cf. Lemma 5). Hence,

\[
(x_n, y_n)_{x_0} = \frac{1}{2} \left( d(x_n, x_0) + d(y_n, x_0) - d(x_n, y_n) \right) \\
\geq \frac{1}{2} \left( d(x'_n, x_0) - C + d(y_n, x_0) - d(x'_n, y_n) - C \right) \\
= (x'_n, y_n)_{x_0} - C \\
= d(x_0, x'_n) - C.
\]

Apparently, \( d(x_0, x'_n) \to \infty \) as \( n \to +\infty \) and, hence, \( (x_n, y_n)_{x_0} \to \infty \) as required.

In order to show that \( \hat{\gamma}_n (-\infty) \to \hat{\gamma} (-\infty) \) we work in a similar manner: The sequence \( \hat{\gamma}_n (-k(t_n + \varepsilon_n)), k \in \mathbb{N} \) converges to \( \hat{\gamma}_n (-\infty) \) as \( k \to \infty \). Hence, there exists \( k_n \in \mathbb{N} \) such that \( d_{\hat{X} \cup \hat{X}}(\hat{\gamma}_n (-k_n(t_n + \varepsilon_n))), \hat{\gamma}_n (-\infty) < 1/n \). As before, sequences \( \{y_n\} \) and \( \{x_n\} \) are defined by \( \hat{y}_n := \hat{\gamma}_n (-k_n(t_n + \varepsilon_n)) \) and \( x_n := \hat{\gamma} (-t_n), n \in \mathbb{N} \). Then we use the same arguments to show that the hyperbolic product \((x_n, y_n)_{x_0}\) with respect to the base point \( x_0 := \hat{\gamma}(0) \) converges to \(+\infty\) as \( n \to +\infty \). 

\[\Box\]

3. Proof of main theorem.

Let \( \gamma \) be a recurrent geodesic, \( \varepsilon > 0 \) and \( x \in \text{Im} \gamma \) be given. We may assume that \( x = \gamma(0) \). Let \( \{t_n\} \) be the sequence given by Definition 6 and \( \{\gamma_n\} \) the sequence of the associated closed curves given by formula (3) above. For each \( n \in \mathbb{N} \), there exists a unique closed geodesic \( c_n \) in the free homotopy class of \( \gamma_n \). The number \( t_n + \varepsilon_n \) is the period of \( \gamma_n \) and let \( s_n \) denote the period of \( c_n \) (apparently, \( s_n < t_n + \varepsilon_n \)). Let \( B_n \) be the projection of \( \gamma(0) \) onto \( \text{Im} c_n \), i.e., \( d(\gamma(0), B_n) = d(\gamma(0), \text{Im} c_n) \). Such a point exists and is unique by Remark 1 following Lemma 5. Lift \( \gamma \) to an isometry \( \tilde{\gamma} : \mathbb{R} \to \hat{X} \) with a base point \( \tilde{\gamma}(0) \) satisfying \( p(\tilde{\gamma}(0)) = \gamma(0) \), where \( p : \hat{X} \to X \) is the universal covering map. Lift each \( c_n \) to an isometry \( \tilde{c}_n : \mathbb{R} \to \hat{X} \) and parametrize it so that \( \tilde{c}_n(0) \) is a point \( \tilde{B}_n \) satisfying

\[
d\left( \tilde{B}_n, \tilde{\gamma}(0) \right) = d(B_n, \gamma(0)) \quad \text{and} \quad p\left( \tilde{B}_n \right) = B_n.
\]

For the reader’s convenience, we have gathered all the above notation in Figure 1.

Since

\[
p(\tilde{\gamma}_n(t_n + \varepsilon_n)) = p(\tilde{\gamma}(0)) = \gamma(0)
\]

and \( \gamma_n, c_n \) are homotopic, the isometry \( \phi_n \) of \( \hat{X} \) which translates \( \tilde{c}_n \) (in the positive direction) satisfies

\[
\phi_n(\tilde{\gamma}(0)) = \tilde{\gamma}_n(t_n + \varepsilon_n).
\]
Moreover,

\[ d\left(\tilde{\gamma}_{n}(t_{n} + \varepsilon_{n}), \phi_{n}\left(\tilde{B}_{n}\right)\right) = d\left(\phi_{n}\left(\tilde{\gamma}(0)\right), \phi_{n}\left(\tilde{B}_{n}\right)\right) = d\left(\tilde{\gamma}(0), \tilde{B}_{n}\right). \]  

(8)

We now proceed to show that given \( \varepsilon > 0 \), there exists \( N \in \mathbb{N} \) such that for all \( n \geq N \)

\[ d\left(\tilde{\gamma}(s), \tilde{c}_{n}(s)\right) < \varepsilon \ \forall s \in [0, s_{n}]. \]  

(9)

Recall that \( s_{n} \) is the period of \( c_{n} \) and \( s_{n} < t_{n} + \varepsilon_{n} = \text{period}(\gamma_{n}) \). Using Lemma 10 and the fact that \( \gamma_{n}, c_{n} \) are homotopic for all \( n \) large enough, we have that \( \tilde{c}_{n}(+\infty) \to \tilde{\gamma}(+\infty) \) and \( \tilde{c}_{n}(-\infty) \to \tilde{\gamma}(-\infty) \). Let \( H : G\bar{X} \approx \partial^{2}\bar{X} \times \mathbb{R} \) be the trivialization of the fiber bundle \( G\bar{X} \to \partial^{2}\bar{X} \) with respect to the base point \( x_{0} = \tilde{\gamma}(0) \). This homeomorphism was described in Remark 2 following Lemma 5. By the choice of parametrization for each \( \tilde{c}_{n} \) made above (i.e., \( \tilde{c}_{n}(0) = \tilde{B}_{n} \)), we have that \( H^{-1}(\tilde{c}_{n}(-\infty), \tilde{c}_{n}(+\infty), 0) = \tilde{c}_{n} \). Moreover, \( H^{-1}(\tilde{\gamma}(-\infty), \tilde{\gamma}(+\infty), 0) = \tilde{\gamma} \) and, thus, \( \tilde{c}_{n} \to \tilde{\gamma} \) uniformly on compact sets. Observe that such convergence is weaker than property (9). However, it implies, in particular, that \( \text{dist}(\tilde{\gamma}(0), \text{Im} \tilde{c}_{n}) \to 0 \) as \( n \to \infty \). Hence, we may choose \( N \in \mathbb{N} \) such that

\[ d\left(\tilde{\gamma}(0), \tilde{B}_{n}\right) < \varepsilon/5, \text{ for all } n \geq N. \]  

(10)

Moreover, we may choose \( N \) such that, in addition, the following inequality is satisfied

\[ \varepsilon_{n} = d\left(\tilde{\gamma}(t_{n}), \tilde{\gamma}(t_{n} + \varepsilon_{n})\right) < \varepsilon/5, \text{ for all } n \geq N. \]  

(11)

To show inequality (9), let \( s \in [0, s_{n}] \) be arbitrary and let \( D_{n} \) (resp. \( F_{n} \)) be the point on the geodesic segment \([\tilde{\gamma}(0), \tilde{\gamma}_{n}(t_{n} + \varepsilon_{n})]\)
where $F_n'$ is the point on $[\tilde{B}_n, \phi_n(\tilde{B}_n)]$ satisfying
\[ d\left(F_n, \phi_n(\tilde{B}_n)\right) = d\left(F_n', \phi_n(\tilde{B}_n)\right). \]

By comparison (see for example [12, Prop. 29]) we have
\begin{align*}
    d\left(\tilde{\gamma}(s) , D_n\right) &\leq d\left(\tilde{\gamma}(t_n + \varepsilon_n) , \tilde{\gamma}_n(t_n + \varepsilon_n)\right) \leq 2\varepsilon_n \\
    d\left(D_n, F_n\right) &\leq d\left(\tilde{\gamma}(t_n + \varepsilon_n) , \phi_n(\tilde{B}_n)\right) \\
    d\left(F_n, F_n'\right) &\leq d\left(\tilde{\gamma}(0) , \tilde{B}_n\right) \\
    d\left(F_n', \tilde{c}_n(s)\right) &\leq \left|d\left(\tilde{\gamma}(0) , \phi_n(\tilde{B}_n)\right) - d\left(\tilde{B}_n, \phi_n(\tilde{B}_n)\right)\right| < d\left(\tilde{\gamma}(0) , \tilde{B}_n\right).
\end{align*}

Combining the above inequalities with inequalities (8), (10) and (11), we obtain property (9) which completes the proof of the existence of a sequence of closed geodesics approximating a given recurrent geodesic.

\[ \square \]

\textbf{Remark.} Let $\Gamma$ be a discrete group of isometries of a locally compact, complete geodesic metric space $Y$ satisfying CAT$-(\chi)$ inequality, $\chi < 0$. The notion of controlled concentration points in the limit set of $\Gamma$ can be defined as follows. $\xi \in \partial Y$ is a controlled concentration point if it admits a neighborhood $U$ containing $\xi$ with the following property: For every neighborhood $V$ of $\xi$ there exists an element $\gamma \in \Gamma$ such that $\gamma(U) \subset V$ and $\xi \in \gamma(V)$.

Following [1], one can show that $\xi$ is a controlled concentration point if and only if there exists a sequence of $\{\phi_n\}$ of distinct elements of $\Gamma$ such that $\phi_n(\xi) \to \xi$ and $\phi_n(0) \to \eta$ with $\eta \neq \xi$. The proof in this more general setting is identical with the one provided in [1] except that the convergence property used there, namely, $\phi_n(x) \to \eta$ for all $x \in Y \cup \partial Y$, is provided in our case by Proposition 7.2 in [6, Ch. 1]. The latter property for $\xi$ is equivalent to the existence of a recurrent geodesic $\gamma$ with $\gamma(+ \infty) = \xi$ and $\gamma(- \infty) = \eta$. Hence we obtain the following connection between recurrent geodesics and controlled concentration points which also holds for manifolds (see [1]).

\textbf{Theorem 11.} Let $Y$ be a locally compact, complete geodesic metric space $Y$ satisfying CAT$-(\chi)$ inequality, $\chi < 0$ and $\Gamma$ a discrete group of isometries of $Y$. A limit point $\xi \in \partial Y$ is a controlled concentration point if and only if $\gamma(+ \infty) = \xi$ for some recurrent geodesic $\gamma$ in $Y$. 

As it was mentioned in the introduction, approximation by closed geodesics does not imply recurrence. The following example demonstrates the existence of a geodesic in a $CAT - (\chi), \chi < 0$ space which is not recurrent but can be approximated by closed geodesics in the sense of Definition 1. Let $X$ be the union of two hyperbolic cylinders identified along a (convex) geodesic strip bounded by two geodesic segments (see Figure 2). We may adjust the geometry of $X$ so that the unique simple closed geodesic in each cylinder, denoted by $c_1$ and $c_2$, have a common image in the geodesic strip, namely, the geodesic segment indicated by letters $A$ and $B$ in Figure 2. Using Cor. 5 of [2] and the fact that the geodesic strip is a convex closed subset it follows that $X$ is a $CAT - (\chi)$ space with $\chi < 0$.

![Figure 2.](image)

Let $\omega_1$ and $\omega_2$ be the periods of $c_1$ and $c_2$ respectively and assume that $c_1$ and $c_2$ are parametrized so that $c_1(0) = c_2(0) = B$ and clockwise i.e., $c_1(s) = c_2(s)$ for all $s \in [0, d(A,B)]$. Define $\gamma : \mathbb{R} \rightarrow X$ as follows:

$$
\gamma(t) = c_1(t), \text{ for } t \in [0, \omega_1] \\
\gamma(t) = c_2(t), \text{ for } t \in (-\infty, 0] \cup [\omega_1, +\infty).
$$

It is apparent that $\gamma$ can be approximated by closed geodesics in the sense of Definition 1. We proceed to show that $\gamma$ is not recurrent by showing that, $\gamma$ and $s\gamma$ are not close in the compact open topology for any positive real $s$. For this it suffices to show that there exists $\varepsilon > 0$ and a compact $M \subset \mathbb{R}$.
such that for any positive $s \in \mathbb{R}$,
\begin{equation}
    d\left( s\gamma(t_0), \gamma(t_0) \right) \geq \varepsilon \quad \text{for some } t_0 \in M.
\end{equation}

For simplicity, we may assume that $d(A, B) = \omega_1/2 = \omega_2/4$. Pick $\varepsilon < d(A, B)/2$ and choose a compact $M \subset \mathbb{R}$ containing the real numbers 0 and $3\omega_1/4$. Let $s$ be arbitrary positive real. If
\[ d(\gamma(s), \gamma(0)) = d(s\gamma(0), \gamma(0)) \geq \varepsilon \]
then Equation (12) is satisfied for the number $t_0 = 0$. If $d(\gamma(s), \gamma(0)) < \varepsilon$ then for $t_0 = 3\omega_1/4$ we have
\[ d(s\gamma(t_0), \gamma(t_0)) = d\left( s\gamma \left( \frac{3\omega_1}{4} \right), c_1 \left( \frac{3\omega_1}{4} \right) \right) > \frac{\omega_1}{4} = \frac{d(A, B)}{2} > \varepsilon. \]

This completes the proof that $\gamma$ is not recurrent and, therefore, approximation by closed geodesics does not imply recurrence.

References


Received September 23, 1998 and revised May 27, 1999.

**Agricultural University of Athens**
75 Iera Odos
Athens 11855
Greece
*E-mail address*: bakis@auadec.aua.gr

**University of the Aegean**
Karlovassi
Samos 83200
Greece
*E-mail address*: gtsap@aegean.gr
COHOMOLOGY OF SINGULAR HYPERSURFACES

BERNARD M. DWORK

Professor Dwork passed away on May 9, 1998 after a long illness. The manuscript was completed a few days earlier, and was submitted to the Pacific Journal of Mathematics following his express desire. It is a testimony to his dedication to mathematics even during his last illness - Managing Editor.

Part I.

Our object is to extend earlier work \cite{D1} on singular hypersurfaces defined over an algebraic number field to singular hypersurfaces defined over function fields in characteristic zero.

A key role will be played by the results of Bertolin \cite{B1} which in turn is based upon the Transfer Theorem of André–Baldassarri–Chiarellotto \cite[DGS, Theorem VI 3.2]{DGS}.

Let $h(A, x)$ be the generic form of degree $d$ in $n+1$ variables $x_1, \ldots, x_{n+1}$.

Thus letting $\mathcal{F}_0 = \left\{ u \in \mathbb{N}^{n+1} \ \middle| \sum_{i=1}^{n+1} u_i = d \right\}$,

$$h(A, x) = \sum_{u \in \mathcal{F}_0} A_u x^u$$

where the symbols $\{A_u\}_{u \in \mathcal{F}_0}$ are algebraically independent over $\mathbb{Q}$.

Let $E_i = x_i \frac{\partial}{\partial x_i} (1 \leq i \leq n + 1), h_i = E_i h$. Let $R(A)$ be the resultant of $\{h_1, h_2, \ldots, h_n, h\}$.

Let $V$ be an absolutely irreducible subvariety of the discriminant locus, $R(A) = 0$. Let $k$ be the field of definition of $V$.

Let $\Omega$ be a suitable universal domain in characteristic zero, and let $\mathcal{L}_{\Omega}^*$ be the ring of all formal sums

$$\mathcal{L}_{\Omega}^* = \left\{ \xi^* = \sum_{u \in \mathcal{F}} C_u \frac{1}{x^u} \ \middle| \ C_u \in \pi^{u_0} \Omega \right\}$$

where $\mathcal{F} = \left\{ u = (u_0, u_1, \ldots, u_{n+1}) \ \middle| \ du_0 = u_1 + \cdots + u_{n+1} \right\}$ and where $\pi^{p-1} = -p, p$ a rational prime. (Thus $\pi$ need not be in $\Omega$.)

For $\lambda \in V$, $\lambda$ rational over $\Omega$ we write

$$D_{i, \lambda}^* = \gamma_\ast \circ (E_i + \pi x_0 h_i(\lambda, x)) \quad 1 \leq i \leq n + 1$$
an endomorphism of $L^\ast_{\Omega}$ where $\gamma_-$ is the projection operator

$$
\gamma_- x^v = \begin{cases} 
0 & \text{if any } v_i \geq 1 \\
 x^v & \text{if all } v_i \leq 0.
\end{cases}
$$

For each integer $\ell$, let $\mathcal{K}^{(\ell)}_{\lambda}$ be the set of all $\xi^* \in L^\ast_{\Omega}$ such that $\xi^*$ is annihilated by all monomials of degree $\ell$ in $\{D^*_{i,\lambda}\}_{1 \leq i \leq n+1}$.

In the following, ord refers to a rank one valuation of $\Omega$.

For $b \in \mathbb{R}, b > 0$, let $L^*(b) = \left\{ \sum_{u \in \mathbb{F}} C_u \frac{1}{u^v} \mid \inf u (\text{ord } C_u + u_0 b) > -\infty \right\}$.

Let $\Gamma$ be an indeterminate and consider the polynomial

$$
h(\lambda, x) + \Gamma h(A, x).
$$

Let $R(\lambda, \Gamma, A)$ be the resultant of $E_1(h(\lambda, x) + \Gamma h(A, x)), \ldots, E_{n+1}(h(\lambda, x) + \Gamma h(A, x))$ and write

$$
R(\lambda, \Gamma, A) = \Gamma^c (\rho_0(\lambda, A) + \Gamma \rho_1(\lambda, A) + \Gamma^2 \rho_2(\lambda, A) + \cdots),
$$

where $\rho_0(\lambda, A) \neq 0$. The key result of the research of Bertolin [B1, Theorem 3.11] states that:

**Theorem 1.**

$$
\mathcal{K}^{(\ell)}_{\lambda} \subset L^*(\tau(n, d, e, \ell) \text{ ord } \rho_0(\lambda, A) + \varepsilon)
$$

for all $\varepsilon > 0$. Here $\tau(n, d, e, \ell)$ depends only on $n, d, e$ and $\ell$ and is independent of the coefficients of $h(\lambda, X)$.

**Remark.** Bertolin obtains estimates independent of $\ell$. The estimate given here depends upon $\ell$ but is simpler to state. The slight error in [B1, Theorem 3.11] is corrected in [B2].

**Corollary 1.** If $\lambda \in V$ and $\rho_0(\lambda, A) \neq 0$, then $\mathcal{K}^{(\ell)}_{\lambda} \subset L^*(\varepsilon)$ for all $\ell$ and all $\varepsilon > 0$ and for all but a finite set of valuations (depending on $\lambda$).

**Corollary 2.** For $\lambda \in V$ with $\rho_0(\lambda, A) \neq 0$, dim $\mathcal{K}^{(\ell)}_{\lambda}$ is independent of $\lambda$.

**Proof.** We choose a valuation $v$ of $k(\lambda)$ such that (extending the valuation of $k(\lambda)$ to $k(\lambda, A)$ via the Gauss norm relative to $A$)

$$
|\rho_0(\lambda, A)|_v = 1 \\
|\lambda|_v \leq 1.
$$

By the Lemma of Appendix B, we may choose a generic point $\lambda'$ of $V$ over $k$ so close to $\lambda$ $v$-adically that $|\lambda - \lambda'|_v < 1$ and hence $|\rho_0(\lambda', A)|_v = 1$. Thus $\mathcal{K}^{(\ell)}_{\lambda}$ and $\mathcal{K}^{(\ell)}_{\lambda'}$ lie in $L^*(\varepsilon)$ ($v$-adically) for all $\varepsilon > 0$ and hence $T_{\lambda, \lambda'} = \gamma_- \circ \exp \pi X_0(h(\lambda', x) - h(\lambda, x))$ is an isomorphism between $\mathcal{K}^{(\ell)}_{\lambda}$ and $\mathcal{K}^{(\ell)}_{\lambda'}$ as vector spaces over $\Omega$. \[\square\]
Part II: Koszul complex.

In earlier work [D1, Theorem 19.2] we discussed the (cohomological) Koszul complex of $D^*_1,\lambda(0), \ldots, D^*_{n+1,\lambda(0)}$ operating on $\mathcal{K}^{(\infty)}_{\lambda(0)} = \bigcup_{\ell=1}^{\infty} \mathcal{K}^{(\ell)}_{\lambda(0)}$ where $\lambda(0)$ is algebraic over $\mathbb{Q}$. We denote by $H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda(0)})$ the $s$-th cohomology group of this complex. We showed:

**Theorem 2.**

$$\dim H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda(0)}) < \infty.$$  

We also showed [D1, Theorem 17.1] that this dimension can be bounded in terms of $d$ and $n$ alone.

**Note.** Equation 19.4 of [D1] is stated without proof. This gap will be filled in Appendix A.

**Corollary 3.** For $\lambda' \in V$, $\dim H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda'}) < \infty$ and if $\rho_0(\lambda', A) \neq 0$, then $\dim H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda'})$ is independent of $\lambda'$.

**Proof.** We choose $\lambda(0)$ algebraic over $\mathbb{Q}$ such that $\lambda(0) \in V$ and $\rho_0(\lambda(0), A) \neq 0$. We choose a valuation $v$ such that $|\rho_0(\lambda(0), A)|_v = 1$, $|\lambda(0)|_v \leq 1$ and then a generic point $\lambda$ of $V$ over $k$ in $\Omega$ as in the proof of Corollary 2. Then $T^{(s)}_{\lambda(0),\lambda}$ provides an isomorphism of $\mathcal{K}^{(s)}_{\lambda(0)}$ with $\mathcal{K}^{(s)}_{\lambda}$ which induces an isomorphism of $H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda'})$ with $H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda})$ for all $s$. This shows finiteness for $\lambda$ generic.

If $\rho_0(\lambda', A) \neq 0$, then by the same argument choosing $\lambda$ generic close to $\lambda'$ we conclude that $\dim H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda'}) = \dim H^{(s)}(\mathcal{K}^{(\infty)}_{\lambda})$. If $\rho_0(\lambda', A) = 0$, then $\lambda'$ lies in a proper subvariety of $V$ and we may use induction on the dimension. □

**Notation.** For $B = \{1, 2, \ldots, n+1\}$ and $W$ a vector space over $k(\lambda)$ let $\mathcal{F}_s(W) = \text{Hom}(\wedge^s B, W)$.

**Corollary 4.** For $\ell$ large enough (depending upon $V$) and $\lambda$ a generic point of $V$ over $k$, $H^{s}\left(\mathcal{K}^{(\infty)}_{\lambda}\right) \simeq \ker\left(\delta^*_s,\lambda, \mathcal{F}_s\left(\mathcal{K}^{(\ell)}_{\lambda}\right)\right) / \left(\mathcal{F}_s\left(\mathcal{K}^{(\ell)}_{\lambda}\right) \cap \delta^*_s,\lambda, \mathcal{F}_{s-1}\left(\mathcal{K}^{(\infty)}_{\lambda}\right)\right)$. (For definition of $\delta^*_s$ see [D1].)

**Proof.** There is a natural injection of the right hand space into the left–hand one induced by the inclusion $\mathcal{K}^{(\ell)}_{\lambda} \hookrightarrow \mathcal{K}^{(\infty)}_{\lambda}$. The left–hand space is of finite dimension and so the mapping is surjective. □
We now give $H^{(s)}(K^{(\infty)}_\lambda)$ the structure of a differential module when viewed as a vector space over $k(\lambda)$. Let $\lambda_1, \ldots, \lambda_t$ be a transcendence basis over $k$ of $k(\lambda)$. Viewing $\lambda_{t+1}, \lambda_{t+2}$ etc. as dependent variables we define for $1 \leq i \leq t$

$$\sigma_i^* = \gamma_- \circ \left( \frac{\partial}{\partial \lambda_i} - \pi x_0 \frac{\partial h}{\partial \lambda_i} \right).$$

These operators commute with $\{D^*_j,\lambda\}_{1 \leq j \leq n+1}$ and hence induce a set of commuting operators on $H^{(s)}(K^{(\infty)}_\lambda)$. If $\lambda^{(1)} \in V$, then horizontal elements are obtained by applying $T^{(1)}_{\lambda,\lambda}$ to $H^{(s)}(K^{(\infty)}_\lambda)$.

**Theorem 3.** Let $\lambda^{(1)}$ be a generic point of $V$. We consider all extensions to $k(\lambda^{(1)})$ of valuations of $k$ whose restriction to $k(\lambda^{(1)}_1, \ldots, \lambda^{(1)}_t)$ is given by the Gauss norm of that field relative to $\lambda^{(1)}_1, \ldots, \lambda^{(1)}_t$. For almost all such valuations the horizontal elements converge for $|((\lambda_1, \ldots, \lambda_t) - (\lambda^{(1)}_1, \ldots, \lambda^{(1)}_t))| < 1$.

**Corollary 5.** If $k$ is an algebraic number field, then $H^{(s)}(K^{(\infty)}_\lambda)$ is a $G$-module.

**Appendix A.**

Let $k$ be a field of characteristic zero and let $f(x_1, \ldots, x_{n+1})$ be a form of degree $d$ in $n+1$ variables. If $\Omega$ is an extension of $k$, let us write $L_\Omega$ for the ring of all polynomials in $x_0, x_1, \ldots, x_{n+1}$ of the form

$$\left\{ \sum_{\substack{d\mu_0 = u_1 + \cdots + u_{n+1}}} C_u \pi^{\mu_0} x^u \mid C_u \in \Omega \right\}.$$

We define $D_i = E_i + \pi x_0 f_i$, $E_i = x_i \frac{\partial}{\partial x_i}$, $f_i = E_i f$. The $D_i$ are commuting endomorphisms of $L_\Omega$ and likewise by restricting to $L_k$ we obtain commuting endomorphisms of that ring.

Let $L^*_\Omega$ (resp. $L^*_k$) be the adjoint space of $L_\Omega$ (resp. $L_k$) and $K^{(\ell)}_\Omega$ (resp. $K^{(\ell)}_k$) the set of all $\xi^* \in L^*_\Omega$ (resp. $L^*_k$) annihilated by all forms in $\{D^*_1, D^*_2, \ldots, D^*_{n+1}\}$ of degree $\ell$, where $D^*_i = \gamma_- \circ (-E_i + \pi x_0 f_i)$.

Again let $K^{(\infty)}_\Omega$ (resp. $K^{(\infty)}_k$) be the union $\bigcup_{\ell=1}^{\infty} K^{(\ell)}_\Omega$ (resp. $\bigcup_{\ell=1}^{\infty} K^{(\ell)}_k$).

Finally we define $H^{(s)}(K^{(\infty)}_\Omega)$ (resp. $H^{(s)}(K^{(\infty)}_k)$) to be the $s$-th cohomology group of the (cohomological) Koszul complex of $D^*_1, \ldots, D^*_{n+1}$ operating on $K^{(\infty)}_\Omega$ (resp. $K^{(\infty)}_k$).

**Theorem.**

(i) $K^{(\infty)}_\Omega = K^{(\infty)}_k \otimes \Omega$
COHOMOLOGY OF SINGULAR HYPERSURFACES

(ii) $H^s(\mathcal{K}_\Omega^{(\infty)}) = H^s(\mathcal{K}_k^{(\infty)}) \otimes \Omega$.

This was stated without proof as Equation (19.4) of [D1].

Proof. We first show for $\ell < \infty$

(iii) $\mathcal{K}_\Omega^{(\ell)} = \mathcal{K}_k^{(\ell)} \otimes \Omega$.

We know [D1, Lemma 7.2] that $\dim_k \mathcal{K}_k^{(\ell)} < \infty$, $\dim_\Omega \mathcal{K}_\Omega^{(\ell)} < \infty$. We may view each element of $\mathcal{K}_\Omega^{(\ell)}$ as an $\infty$-tuple $(z_1, z_2, \ldots)$ indexed by a countable set $I$. Indeed $\xi^* \in \mathcal{K}_\Omega^{(\ell)}$ implies $\xi^* = \sum_\mathcal{U} C_u/x^n$. The sum being over all $u$ such that $du_0 = u_1 + \cdots + u_{n+1}$. Here $C_u = \pi^{u_0} \overline{C}_u$ with $\overline{C}_u \in \Omega$. Identifying the $\{\overline{C}_u\}$ with the $\{z_i\}$, the condition that $\xi^* \in \mathcal{K}_\Omega^{(\ell)}$ is equivalent to an infinite set of conditions

$$\sum t_{j,i}z_i = 0 \quad \text{for all } j \in J.$$

Here $t_{j,i} \in k$, $t_{j,i} = 0$ for almost all $i$, for each fixed $j$. For $\xi^* \in \mathcal{K}_k^{(\ell)}$ we have the same set of conditions. Following a suggestion by Wan, by elementary operations on the rows of the matrix $\{t_{j,i}\}$ the finite dimension of the subspace is given by the number of zero columns in the reduced echelon form. The echelon form is the same for the equation over $\Omega$ as over $k$. It follows that indeed $\dim_k \mathcal{K}_k^{(\ell)} < \infty \iff \dim \mathcal{K}_\Omega^{(\ell)} < \infty$ and both are then equal and $\mathcal{K}_\Omega^{(\ell)} = \mathcal{K}_k^{(\ell)} \otimes \Omega$. The first assertion now follows.

For a vector space $W$ we write $F_\ell(W) = \text{Hom}(\wedge^\ell B, W)$ with $B = \{1, 2, \ldots, n+1\}$. Then $\xi^* \in F_\ell(\mathcal{K}_\Omega^{(\infty)})$ implies $\xi^* = \sum \eta_i \xi_i^*$ a finite sum with $\xi_i^* \in F_\ell(\mathcal{K}_k^{(\infty)})$ and $\{\eta_i\}$ a finite set of elements of $\Omega$ linearly independent over $k$.

If $\delta^s_{s+1} \xi^* = 0$ then by linear independence $\delta^s_{s+1} \xi_i^* = 0$ and so $\ker(\delta^s_{s+1}, F_\ell(\mathcal{K}_\Omega^{(\infty)})) = \ker(\delta^s_{s+1}, F_\ell(\mathcal{K}_k^{(\infty)})) \otimes \Omega$.

Also

$$\delta^s_{s} F_{s-1} \left(\mathcal{K}_\Omega^{(\infty)}\right) = \delta^s_{s} F_{s-1} \left(\mathcal{K}_k^{(\infty)}\right) \otimes \Omega.$$

The theorem now follows from the following well known proposition. \qed

Proposition. Let $U$ be a subspace of a linear $k$ space $W$. Then

$$W \otimes \Omega/U \otimes \Omega \simeq (W/U) \otimes \Omega.$$

Appendix B. Approximation by generic points.

Lemma. Let the origin $O$ be on an irreducible affine variety $V$ defined over a field $k$ of characteristic zero. Let $\Omega$ be a universal domain complete under a rank one valuation. Then there exists a generic point of $V$ rational over $\Omega$ which is as close as you please to the origin.

We first show the lemma holds if $V$ is a curve in $\mathbb{A}^n$. 
Proof. Let \( P = (x_1, \ldots, x_n) \) be a generic point of \( V \) over \( k \). Then \( R = k[x_1, \ldots, x_n] \) has a specialization into \( k \) given by \((x_1, \ldots, x_n) \mapsto O \) and hence there exists a place \( \mathfrak{p} \) of \( k(V) \) with center \( O \). Letting \( T \) be a uniformizing parameter of \( \mathfrak{p} \), each coordinate \( x_i \) as element of \( k(V)_{\mathfrak{p}} \), the completion at \( \mathfrak{p} \) of \( k(V) \), is represented as a power series

\[
x_i = a_{i1} T + a_{i2} T^2 + \cdots + \in k'[\![T]\!]
\]

where \( k' \) is the residue class field at \( \mathfrak{p} \) of \( k(V) \), a finite extension of \( k \). This series may have zero radius of convergence in the metric of \( \Omega \), but if we choose (as we shall) the uniformizing parameter, \( T \), in \( k(V) \) then the series represents an algebraic function of \( T \) and hence by Eisenstein’s Theorem (or more elementarily by Clark’s Theorem) the series has a non-trivial radius of convergence.

Since \( P \) is a generic point, these series are not all constant. We think of \( P(T) \) as function of \( T \) for \( T \) restricted to a small disk \( D(0, r^-) \) in \( \Omega \)–space. Trivially \( P(T) \to 0 \) as \( T \to 0 \). We may suppose \( x_1 \) is a non–constant function of \( T \). The theory of Newton polygons shows that the image of \( D(0, r^-) \) under \( x_1 \) contains elements transcendental over \( k \). This completes the proof for \( \dim V = 1 \).

We recall [\( H \), Chapter I, Proposition 7.1]:

\[
\square
\]

**Proposition.** If \( V \) is irreducible of dimension \( s \) in \( k^n \) and \( H \) is a hypersurface not containing \( V \) then each irreducible component of \( H \cap V \) has dimension \( s - 1 \).

**Proof of Lemma.** Letting \( V_0 = V \) we define inductively \( V_1 \supseteq V_2 \supseteq \cdots \) by the condition that \( V_j \) be an irreducible component of \( V_{j-1} \cap \{ x \mid x_j = 0 \} \) which contains the origin. Since

\[
-1 + \dim V_{j-1} \leq \dim V_j \leq \dim V_{j-1}, \; \dim V_n = \{0\}
\]

there exists \( j \) such that \( V_j \) is a curve on \( V \) passing through the origin. \( \square \)

We conclude there exists a curve \( V' \) on \( V \) passing through the origin. Let \( k' \supseteq k \) be a field of definition of \( V' \). By our previous treatment of curves there exists \( P \in V' \), \( P \) as close as you please to \( O \) such that \( k'(P) \) is of transcendence degree unity over \( k' \). Let \( P_1 \) be a coordinate of \( P \) of transcendence degree unity over \( k' \).

Let \( L = k(P_1) \), \( \mathfrak{A} \) be the ideal of all \( f \in k[x_1, \ldots, x_n] \) which are zero everywhere on \( V \). If \( g \in L[x_1, \ldots, x_n] \), \( g = 0 \) on \( V \) then \( g \in \mathfrak{A}L[x] \) and hence for each automorphism \( \tau \) of \( L/k \) we have \( g^\tau = 0 \) on \( V \). In particular \( x_1 - P_1 \) cannot be zero on \( V \) as otherwise \( (x_1 - P_1)^\tau \) would also be zero on \( V \) and hence \( P_1 - P_1^\tau \) would be zero on \( V \) for every \( \tau \) which is impossible as there are nontrivial automorphisms of \( L/k \).

Thus \( V \) does not lie in the hyperplane \( x_1 = P_1 \) and so the intersection has an irreducible component \( W \) passing through \( P \) of dimension \( s-1 \). Let \( k'' \) be
a field of definition of $W, P \in W$. By induction there exists $Q \in W, Q - P$ as small as you please with transc deg $k''(Q)/k'' = s - 1$.

Clearly $Q$ is as close as you please to $O$. It remains to show that $s = \text{trans deg} k(Q)/k$.

Since $Q_1 = P_1, k(Q) \supset k(P_1)$. Hence

$$s \geq \text{trans deg} k(Q)/k = \text{trans deg} k(Q)/k(P_1) + \text{trans deg} k(P_1)/k$$

$$\geq \text{trans deg} k''(Q)/k'' + \text{trans deg} k'(P_1)/k'$$

$$\geq (s - 1) + 1 = s,$$

the two inequalities being based on

if $k' \supset k$ then $\text{trans deg} k(P_1)/k \geq \text{trans deg} k'(P_1)/k'$

if $k'' \supset k(P_1)$ then $\text{trans deg} k(Q)/k(P_1) \geq \text{trans deg} k''(Q)/k''$.

This completes the proof of the lemma.

**Appendix C: (Generalization of Heaviside’s generalized exponential functions).**

In this article we examined the Koszul complex of $\{D^*_1, \lambda, \ldots, D^*_{n+1}, \lambda\}$ operating on $K_\lambda(\infty)$. In this appendix, we replace $L^*$ by

$$L'^* = \left\{\sum_{u \in F'} A_u \frac{1}{x^u} \mid A_u \in \pi u_0 \Omega\right\}$$

and $D^*_{i, \lambda} = \gamma_\lambda \circ (-E_i + \pi x_0 h_i(x_1, x))$ by $D^*_{i, \lambda} = -E_i + \pi x_0 h_i(\lambda, x)$. Here

$$F' = \{(u_0, u_1, \ldots, u_{n+1}) \mid \in Z^{n+2} | du_0 = u_1 + \cdots + u_{n+1}\}.$$

Thus $L'^*$ consists of formal Laurent series in $\{x_i, \frac{1}{x_i}\}_{i = 1, \ldots, n+1}$. We note that $L'^*$ is adjoint to $L'$, the ring of Laurent polynomials with support in

$$du_0 = u_1 + \cdots + u_{n+1}.$$

Let $D^s$ denote the ideal of all forms of degree $s$ in $D_1, \lambda, \ldots, D_{n+1, \lambda}$ with coefficients in $k(\lambda)$. We assert that

$$L' = L + D^s L'.$$

For $s = 1$ this follows by the proof of [D2, Lemma 9.7.1]. Assume the formula valid for some given $s$ then $L' = L + D^s(L + DL') \subset L + D^{s+1}L'$, which completes the proof by induction.

This shows that the natural mapping of $L$ into $L'$ induces a surjection $L/D^sL \rightarrow L'/D^sL'$. We recall that $K^{(s)}$ denotes the annihilator in $L^*$ of $D^sL$. Let $K'^{(s)}$ denote the annihilator in $L'^*$ of $D^sL'$. We now know that the dimension of $L'/D^sL'$ is finite and hence the same holds for $K'^{(s)}$. Thus
by duality the mapping of $\mathcal{K}_{\ell(s)}$ into $\mathcal{K}_{s}^{(s)}$ adjoint to the natural mapping is injective. This adjoint mapping is $\gamma_{-}$.

**Conclusion.** The mapping $\gamma_{-}$ maps $\mathcal{K}_{\ell(s)}$ into $\mathcal{K}_{s}^{(s)}$ injectively.

We now restrict our attention to the case where $h(\lambda, x) \in \mathbb{Q}[\lambda, x]$ and consider $\lambda^{(0)}$ algebraic over $\mathbb{Q}$. For $b > 0$, $c \in \mathbb{R}$, $w$ a finite valuation of $\Omega$, let $L^{s}*(b, c)$ be the set of all formal Laurent series $\xi^{*} = \sum_{u \in \mathcal{F}} B_{u} \frac{1}{x^{u}}$ such that $B_{u} \in \Omega$, and $\text{ord}(B_{u,v}) \geq -b(u_{0} + v_{0}) + c$ for all $u, v, \in \mathcal{F}$. Let $L_{*}^{s}(b) = \bigcup_{c \in \mathbb{R}} L_{*}^{s}(b, c)$, a Banach space. For almost all valuations of $\mathbb{Q}(\lambda^{(0)})$ we have a completely continuous mapping of $L_{*}^{s}(b')$ (giving $\Omega$ a valuation extending that of $\mathbb{Q}(\lambda^{(0)})$) defined by putting $F(x) = \exp \pi(x_{0}b(\lambda^{(0)}, x) - x_{0}^{q}h(\lambda^{(0)}, x^{q}))$ where $q$ is the order of the residue class field of $\mathbb{Q}(\lambda^{(0)})$ and writing

$$\alpha^{*} = F \circ \phi, \quad \phi : x^{u} \rightarrow x^{qu}$$

$$L_{*}^{s}(b') \xrightarrow{\phi} L_{*}^{s}(b'/q) \xrightarrow{F} L_{*}^{s}(b'/q) \xleftarrow{\gamma} L_{*}^{s}(b')$$

where $b'$ is chosen in $[0, \frac{p}{1-p}[$. Here $F$ means multiplication by $F$ and the last map is the inclusion map.

Letting $L_{*}^{s}(b) = \gamma_{-}L_{*}^{s}(b)$ we have the completely continuous endomorphism of $L_{*}^{s}(b')$ (for almost all valuations of $\mathbb{Q}(\lambda^{(0)})$) given by

$$\alpha^{*} = \gamma_{-} \circ F \circ \phi.$$ 

By the trace formula the two mappings have the same Fredholm determinant. Defining

$$W_{z}^{*} = \bigcup_{k} \ker \left( (I - z\alpha^{*})^{k}, L_{*}^{s}(b') \right)$$

$$W_{z}^{\prime s} = \bigcup_{k} \ker \left( (I - z\alpha^{s})^{k}, L_{*}^{s}(b') \right)$$

we conclude equality of dimensions and hence

$$\gamma_{-}W_{z}^{\prime s} = W_{z}^{*}.$$ 

Now $\mathcal{K}_{\lambda^{(0)}}^{(\ell)}$, is covered by a union of spaces $W_{z}^{*}$ and hence by $\gamma_{-}$ (finite union of spaces $W_{z}^{\prime s}$) which lies in $\gamma_{-}\mathcal{K}_{\lambda^{(0)}}^{(\ell)}$ for suitable $\ell'$.

**Conclusion.** $\gamma_{-}$ gives a bijection of $\mathcal{K}_{\lambda^{(0)}}^{(\ell)}$ onto $\mathcal{K}_{\lambda^{(0)}}^{(\ell)}$ provided $\lambda^{(0)}$ is algebraic over $\mathbb{Q}$.

We propose to remove the restriction that $\lambda^{(0)}$ be algebraic. Again let $\lambda^{(0)} \in V$ be an algebraic point, $\rho_{0}(\lambda^{(0)}, A) \neq 0$. Excluding a finite set of primes of $\mathbb{Q}(\lambda^{(0)})$ we choose $\lambda$ generic point of $V$ close to $\lambda^{(0)}$.

If $\xi^{*} \in \mathcal{K}_{\lambda^{(0)}}$, then $\gamma_{-}\xi^{*} \in \mathcal{K}_{\lambda^{(0)}}^{(\ell)}$ for some $\ell$ and hence $\gamma_{-}\xi^{*}$ is a finite sum of elements of spaces $W_{z}^{*}$ and hence is the image under $\gamma_{-}$ of a finite
sum of elements of spaces $W'_{x^*}$. But $\gamma_-$ is injective, and hence $\xi^*$ is a sum of elements of spaces $W'_{x^*}$. Thus for almost all valuations of $\mathbb{Q}(\lambda^{(0)})$, $\xi^* \in L^{b'}(b')$, $b' < \frac{p-1}{p}$. More precisely $\mathcal{K}_{\lambda^{(0)}}^{(\infty)}$ lies in $L^{b'}(b')$ for all $b' > 0$ and almost all primes of $\mathbb{Q}$. Hence for almost all primes multiplication by $\exp \pi x_0 \left( h(\lambda; x) - h(\lambda^{(0)}; x) \right)$ provides an isomorphism $T'_{\lambda^{(0)},\lambda}$ of $\mathcal{K}_{\lambda^{(0)}}^{(\infty)}$ with $\mathcal{K}_{\lambda}^{(\infty)}$. On the other hand, $T_{\lambda^{(0)},\lambda} = \gamma_- \circ T'_{\lambda^{(0)},\lambda}$ gives an isomorphism between $\mathcal{K}_{\lambda^{(0)}}^{(\infty)}$ and $\mathcal{K}_{\lambda}^{(\infty)}$

\[
\begin{array}{ccc}
\mathcal{K}_{\lambda}^{(\infty)} & \xrightarrow{T'_{\lambda^{(0)},\lambda}} & \mathcal{K}_{\lambda}^{(\infty)} \\
\downarrow \gamma_- & & \downarrow \gamma_-
\end{array}
\]

The horizontal arrows of this commutative diagram are isomorphisms. The first vertical arrow is also an isomorphism. It follows that the second vertical arrow is also an isomorphism. This completes the proof.

**Note.** The purpose of the argument involving $\lambda^{(0)}$ is to show that $\gamma_-$ is injective on $\mathcal{K}_{\lambda}^{(\infty)}$.

**References**


Received October 28, 1998.

**Universita degli Studi di Padova**

35131 Padova

**Italy**

E-mail address: baldassa@math.unipd.it
RIEMANNIAN MANIFOLDS ADMITTING ISOMETRIC IMMERSIONS BY THEIR FIRST EIGENFUNCTIONS

AHMAD EL SOUFI AND SAÏD ILIAS

Given a compact manifold $M$, we prove that every critical Riemannian metric $g$ for the functional “first eigenvalue of the Laplacian” is $\lambda_1$-minimal (i.e., $(M,g)$ can be immersed isometrically in a sphere by its first eigenfunctions) and give a sufficient condition for a $\lambda_1$-minimal metric to be critical. In the second part, we consider the case where $M$ is the 2-dimensional torus and prove that the flat metrics corresponding to square and equilateral lattices of $\mathbb{R}^2$ are the only $\lambda_1$-minimal and the only critical ones.

Introduction.

Many recent works concerning the spectrum of compact Riemannian manifolds have pointed out the importance of a particular class of Riemannian metrics which we called in [5] $\lambda_1$-minimal. Recall that a metric $g$ on a compact $m$-dimensional manifold $M$ is $\lambda_1$-minimal if the eigenspace $E_1(g)$ associated to the first nonzero eigenvalue $\lambda_1(g)$ of the Laplacian of $g$ contains a family $f_1, \ldots, f_k$ of functions satisfying: $\sum_{1 \leq i \leq k} df_i \otimes df_i = g$. It follows from a well known result of Takahashi [8] that this last condition is equivalent to the fact that the map $f = (f_1, \ldots, f_k)$ is a minimal isometric immersion from $(M,g)$ into the Euclidean sphere $S_{r}^{m-1}$ of radius $r = \sqrt{\frac{m}{\lambda_1(g)}}$.

The best known examples of $\lambda_1$-minimal metrics are the standard metrics of rank one compact symmetric spaces (i.e., spheres and projective spaces). More generally, any Riemannian irreducible homogeneous space is $\lambda_1$-minimal. Also, Yau [9] conjectured that a minimal embedded hypersurface of a Euclidean sphere, carrying the induced metric, must be $\lambda_1$-minimal.

In [2], Berger showed that the $\lambda_1$-minimality of a metric $g$ is strongly related to the extremality of $g$ for a spectral functional involving the $k$-smallest eigenvalues of the Laplacian (where $k$ is the multiplicity of $\lambda_1(g)$). Recently, Nadirashvili [7] considered the functional $\lambda_1 : g \mapsto \lambda_1(g)$ defined on the set of Riemannian metrics of given area on a compact surface $M$ and showed that the extremal metrics of this functional are $\lambda_1$-minimal (here extremality is defined in a generalized sense because of the non-differentiability of $\lambda_1$).
In the first part of this paper we generalize Nadirashvili’s theorem to higher dimensions (Theorem 1.1). We also give a sufficient condition for a $\lambda_1$-minimal metric to be extremal for $\lambda_1$ (Proposition 1.1).

Using results established by us in [4] about $\lambda_1$-minimal metrics we deduce that (Corollary 1.1), if $g$ is an extremal metric of the $\lambda_1$ functional then:

(i) The multiplicity of $\lambda_1(g)$ is at least equal to $m + 1$ and equality holds only for the standard metric of Euclidean spheres.
(ii) The restriction of the $\lambda_1$ functional to the conformal class of $g$ achieves its maximum at $g$. In particular, the $\lambda_1$ functional has no local minima.
(iii) The metric $g$ is, up to dilatation, the unique extremal metric in its conformal class.
(iv) If $g$ is not isometric to the standard metric of a Euclidean sphere then any conformal diffeomorphism of $(M, g)$ is an isometry.

The second part of this paper deals with the classification of $\lambda_1$-minimal metrics and of the extremal metrics of the $\lambda_1$ functional. The only manifold for which this classification was available is the 2-dimensional sphere. Indeed, on $S^2$ the standard metric is (up to dilatation) the only one to be $\lambda_1$-minimal and the only extremal metric for $\lambda_1$ (this follows from the uniqueness of the conformal class on $S^2$ and property (iii) above).

The main theorem of Section 2 (Theorem 2.1) states that in genus one (i.e., on the torus $T^2$) there exists, up to dilatation, exactly two $\lambda_1$-minimal metrics: The Clifford metric $g_{cl}$ and the equilateral metric $g_{eq}$ induced from the Euclidean metric respectively on $\mathbb{R}^2/\mathbb{Z}^2$ and $\mathbb{R}^2/\Gamma_{eq}$ with $\Gamma_{eq} = \mathbb{Z}(1, 0) \oplus \mathbb{Z}(1/2, \sqrt{3}/2)$. These two metrics are also the only extremal metrics for $\lambda_1$ (Corollary 2.2). Moreover, we prove that for each of them, the standard embedding (in $S^3$ for $g_{cl}$ and $S^5$ for $g_{eq}$) is, up to equivalence, the only full (minimal) isometric immersion by the first eigenfunctions.

Note that a first step towards this classification was achieved by Montiel and Ros [6] who proved that the only minimal torus immersed in $S^3$ by its first eigenfunctions is the Clifford torus. They deduced that if the aforementioned conjecture of Yau is true, then the Clifford torus is the only minimally embedded torus in $S^3$ (Lawson’s conjecture).

1. Extremal metrics for the $\lambda_1$ functional.

Let $M$ be a compact smooth manifold of dimension $m \geq 2$. Denote by $\mathcal{R}_0(M)$ the set of Riemannian metrics of volume 1 on $M$. For any $g \in \mathcal{R}_0(M)$, we denote by $0 < \lambda_1(g) \leq \lambda_2(g) \leq \cdots \leq \lambda_k(g) \leq \cdots$ the increasing sequence of eigenvalues of the Laplacian $\Delta_g$ of $g$. The functional:

$$
\lambda_1 : \mathcal{R}_0(M) \to \mathbb{R} \\
g \mapsto \lambda_1(g)
$$
is continuous but not differentiable in general. However, for any family $(g_t)_t$ of metrics, analytic in $t$, $\lambda_1(g_t)$ has right and left derivatives w.r.t. $t$. Indeed, if $(g_t)_{t \in [-\delta, \delta]}$ is such a family and if $k$ is the multiplicity of $\lambda_1(g_0)$, then there exists $k$ analytic families $\Lambda_{1,t}, \ldots, \Lambda_{k,t}$ of real numbers and $k$ analytic families of smooth functions $u_{1,t}, \ldots, u_{k,t}$ such that: $\forall i \leq k$ and $\forall t$, $\Delta g_t u_{i,t} = \Lambda_{i,t} u_{i,t}$, $\Lambda_{i,0} = \lambda_1(g_0)$ and $\{u_{1,t}, \ldots, u_{k,t}\}$ is $L^2(g_t)$-orthonormal (see [1] and [2] for details). Moreover, Berger [2] gave the following formula for the derivative of $\Lambda_{i,t}$:

$$\frac{d}{dt} \Lambda_{i,t} \bigg|_{t=0} = - \int_M \langle q(u_i), h \rangle \nu_{g_0},$$

where $\nu_{g_0}$ is the Riemannian volume element of $g_0$, $u_i = u_{i,0}$, $h = g_t|_{t=0}$, $\langle , \rangle$ is the inner product induced by $g_0$ on the space $S^2(M)$ of symmetric covariant 2-tensors of $M$ and where for any $u \in C^\infty(M)$,

$$q(u) = du \otimes du + \frac{1}{4} \Delta g_0 (u^2) g_0.$$

From the continuity of $\lambda_i(g_t)$ and $\Lambda_{i,t}$ w.r.t. $t$, we have for $t$ small enough $\{\Lambda_{i,t}\}_{1 \leq i \leq k} = \{\lambda_i(g_t)\}_{1 \leq i \leq k}$ and thus $\lambda_1(g_t) = \min_{1 \leq i \leq k} \{\Lambda_{i,t}\}$. This proves the left and right differentiability of $\lambda_1(g_t)$ and gives:

$$\frac{d}{dt} \lambda_1(g_t) \bigg|_{t=0^+} = \min_{1 \leq i \leq k} \frac{d}{dt} \Lambda_{i,t} \bigg|_{t=0} = - \max_{1 \leq i \leq k} \int_M \langle q(u_i), h \rangle \nu_{g_0},$$

and

$$\frac{d}{dt} \lambda_1(g_t) \bigg|_{t=0^-} = \max_{1 \leq i \leq k} \frac{d}{dt} \Lambda_{i,t} \bigg|_{t=0} = - \min_{1 \leq i \leq k} \int_M \langle q(u_i), h \rangle \nu_{g_0},$$

This suggests the following definition:

**Definition 1.1.** A metric $g \in \mathcal{R}_0(M)$ is said to be extremal for the $\lambda_1$ functional if for any analytic deformation $(g_t)_t \subset \mathcal{R}_0(M)$, with $g_0 = g$, the left and right derivatives of $\lambda_1(g_t)$ at $t = 0$ have opposite signs, i.e.,

$$\frac{d}{dt} \lambda_1(g_t) \bigg|_{t=0^+} \leq 0 \leq \frac{d}{dt} \lambda_1(g_t) \bigg|_{t=0^-}.$$

This last condition is equivalent to:

$$\lambda_1(g_t) \leq \lambda_1(g) + o(t) \quad \text{as} \quad t \to 0.$$

Hence our definition of extremality is an equivalent formulation of Nadi­rashvili’s one [7].

The main result of this section is:

**Theorem 1.1.** If a Riemannian metric $g \in \mathcal{R}_0(M)$ is extremal for $\lambda_1$ then it is $\lambda_1$-minimal.
In the 2-dimensional case this result was proved by Nadirashvili [7]. Some of the arguments in our proof are inspired by his. However, the use of the aforementioned result of Berger makes the proof of this theorem simpler and more transparent.

**Lemma 1.1.** If a metric $g \in \mathcal{R}_0(M)$ is extremal for $\lambda_1$ then for any $h \in S^2_0(M) = \{ h \in S^2(M); \int_M tr_g h \nu_g = 0 \}$ there exists $u \in E_1(g) \setminus \{0\}$ such that:

$$\int_M \langle q(u), h \rangle \nu_g = 0.$$ 

**Proof.** Suppose that $g$ is extremal for $\lambda_1$ and let $h \in S^2_0(M)$. We let, for small $t$, $g_t = g + th \frac{\nu}{V(g+th)^2} \in \mathcal{R}_0(M)$, where $V(g+th)$ is the Riemannian volume of $g + th$. Since $\frac{d}{dt} V(g+th)|_{t=0} = \frac{1}{2} \int_M tr_g h \nu_g = 0$, we find $\frac{d}{dt} g_t|_{t=0} = h$. The extremality condition implies that the quadratic form $u \in E_1(g) \mapsto \int_M \langle q(u), h \rangle \nu_g$ takes on both nonpositive and nonnegative values, and therefore it admits at least one isotropic direction. □

**Proof of Theorem 1.1.** Let $K$ be the convex hull in $S^2(M)$ of $\{q(u), \ u \in E_1(g)\}$. The set $K \cup \{g\}$ is contained in a finite dimensional subspace of $S^2(M)$. We claim that $g \in K$. Indeed, if $g \notin K$ then, since $K$ is a convex cone, the Hahn-Banach theorem implies the existence of $s \in S^2(M)$ such that:

$$\int_M \langle s, g \rangle \nu_g > 0 \quad \text{and for every } l \in K \setminus \{0\}, \int_M \langle l, s \rangle \nu_g < 0.$$ 

The 2-tensor $\tilde{s} = s - \frac{\int_M \langle s, g \rangle \nu_g}{mV(g)} g$ belongs to $S^2_0(M)$ and, for any $u \in E_1(g) \setminus \{0\}$,

$$\int_M \langle q(u), \tilde{s} \rangle \nu_g = \int_M \langle q(u), s \rangle \nu_g - \frac{1}{mV(g)} \left( \int_M \langle s, g \rangle \nu_g \right) \left( \int_M |du|^2 \nu_g \right) < 0.$$ 

By Lemma 1.1, this contradicts the extremality of $g$.

Thus $g \in K$ and there exists $w_1, \ldots w_d \in E_1(g)$ such that:

$$g = \sum_{1 \leq i \leq d} q(w_i) = \sum_{1 \leq i \leq d} dw_i \otimes dw_i + \frac{1}{4} \left( \sum_{1 \leq i \leq d} \Delta w_i^2 \right) g$$

$$= \sum_{1 \leq i \leq d} \left( dw_i \otimes dw_i + \frac{1}{2} \left( \lambda_1(g) w_i^2 - |dw_i|^2 \right) g \right).$$
The traceless part of the last member of this equation must be zero. Therefore,
\[
\sum_{1 \leq i \leq d} \left( dw_i \otimes dw_i - \frac{|dw_i|^2}{m} g \right) = 0,
\]
and then:
\[
\frac{\lambda_1}{2} \sum_{1 \leq i \leq d} w_i^2 = 1 + \left( \frac{m-2}{2m} \right) \sum_{1 \leq i \leq d} |dw_i|^2.
\]
(1)

The \(\lambda_1\)-minimality of \(g\) will follow from the fact that \(\sum_{1 \leq i \leq d} |dw_i|^2\) is constant and equal to \(m\). Indeed, set \(f = \left( \sum_{1 \leq i \leq d} w_i^2 \right) - \frac{m}{\lambda_1(g)}\). From (1) we get:
\[
(m-2) \Delta_g f = 2(m-2) \left( \lambda_1(g) \left( \sum_{1 \leq i \leq d} w_i^2 \right) - \sum_{1 \leq i \leq d} |dw_i|^2 \right) = -4\lambda_1(g) f.
\]
This implies that \(f = 0\) (the Laplacian being a positive operator). Therefore \(\left( \sum_{1 \leq i \leq d} w_i^2 \right) = \frac{m}{\lambda_1(g)}\). Replacing in (1) we obtain \(\sum_{1 \leq i \leq d} |dw_i|^2 = m\). \(\square\)

In [4] we showed that \(\lambda_1\)-minimal metrics satisfy certain remarkable conformal properties. Theorem 1.1 tells us that all these properties are still true for extremal metrics:

**Corollary 1.1.** Let \(g \in R_0(M)\) be an extremal metric for \(\lambda_1\).

(i) The multiplicity of \(\lambda_1(g)\) satisfies: \(\text{mult}(\lambda_1(g)) \geq m+1\), where equality holds if and only if \(g\) is isometric to a standard metric of a Euclidean sphere.

(ii) For any \(g' \in C_0(g) = \{ g' \in R_0(M) \ ; \ g' \ \text{conformal to} \ g \} \) we have \(\lambda_1(g') \leq \lambda_1(g)\), and equality holds if and only if \(g'\) is isometric to \(g\). In particular, the functional \(\lambda_1\) does not admit a local minimum in \(R_0(M)\).

(iii) The metric \(g\) is, up to isometry, the only extremal metric of \(\lambda_1\) in \(C_0(g)\).

(iv) If \((M,g)\) is not isometric to a Euclidean sphere then any conformal diffeomorphism of \((M,g)\) is an isometry.

The following is a converse to Theorem 1.1.

**Proposition 1.1.** Let \(g \in R_0(M)\) and assume there exists an \(L_2(g)\)-orthonormal basis \(\{ \phi_1, \ldots, \phi_k \}\) of \(E_1(g)\) such that the 2-tensor \(\sum_{1 \leq i \leq k} d\phi_i \otimes d\phi_i\) is proportional to \(g\). Then \(g\) is extremal for \(\lambda_1\).
Proof. Let \((g_t)_t \subset \mathcal{R}_0(M)\) be a family of metrics analytic in \(t\) with \(g_0 = g\) and set \(h = \frac{d}{dt} g_t\big|_{t=0}\). With the same notation as above we have for small \(t\):
\[
\sum_{1 \leq i \leq k} \lambda_i(g_t) = \sum_{1 \leq i \leq k} \Lambda_{i,t}.
\]
Therefore, \(\sum_{1 \leq i \leq k} \lambda_i(g_t)\) is differentiable at \(t = 0\) and
\[
\frac{d}{dt} \left. \sum_{1 \leq i \leq k} \lambda_i(g_t) \right|_{t=0} = \frac{d}{dt} \left. \sum_{1 \leq i \leq k} \Lambda_{i,t} \right|_{t=0} = \text{trace } Q_h,
\]
where \(Q_h\) is the quadratic form defined on \(E_1(g)\) by:
\[
Q_h(u) = \int_M \langle q(u), h \rangle \nu_g,
\]
and where the trace of \(Q_h\) is taken w.r.t the \(L_2\) inner product induced by \(g\). Now
\[
\text{trace } Q_h = \sum_{1 \leq i \leq k} Q_h(\phi_i)
\]
\[
= \int_M \left( \sum_{1 \leq i \leq k} d\phi_i \otimes d\phi_i, h \right) \nu_g + \frac{1}{4} \sum_{1 \leq i \leq k} \int_M \langle \Delta \phi_i^2, h \rangle \nu_g.
\]
Since \(\sum_{1 \leq i \leq k} d\phi_i \otimes d\phi_i\) is proportional to \(g\) and \(\int_M \langle g, h \rangle \nu_g = 2 \frac{d}{dt} V(g_t)\big|_{t=0} = 0\) we have \(\int_M \left( \sum_{1 \leq i \leq k} d\phi_i \otimes d\phi_i, h \right) \nu_g = 0\). Moreover, by Takahashi’s theorem \(\sum_{1 \leq i \leq k} \phi_i^2\) is constant. Therefore, \(\text{trace } Q_h = 0\) and \(\frac{d}{dt} \left. \sum_{1 \leq i \leq k} \lambda_i(g_t) \right|_{t=0} = 0\). The extremality of \(g\) then follows from the inequality \(\lambda_1(g_t) \leq \frac{1}{k} \sum_{1 \leq i \leq k} \lambda_i(g_t)\) which is an equality at \(t = 0\). \(\square\)

Remarks.

1) It is known that compact irreducible homogeneous Riemannian spaces satisfy the hypothesis of Proposition 1.1 (see [8]). Thus, their standard metrics are extremal for \(\lambda_1\).

2) We restricted ourselves to \(\lambda_1\). Nevertheless, the results of this paragraph can be carried over to the case of higher eigenvalues.

2. \(\lambda_1\)-minimal and extremal metrics on the torus.

Let \((M, g)\) be an orientable compact surface of genus one endowed with a Riemannian metric \(g\). It is well known that there exists a lattice \(\Gamma\) of \(\mathbb{R}^2\) such that \((M, g)\) is conformally equivalent to the torus \((\mathbb{R}^2/\Gamma, g_\Gamma)\), where \(g_\Gamma\) is the flat metric induced from the Euclidean metric on \(\mathbb{R}^2\). The Clifford torus \((T_\Gamma^c = \mathbb{R}^2/\Gamma_c, g_{\Gamma_c} = g_{\Gamma_c})\) with \(\Gamma_c = \mathbb{Z}(1,0) \oplus \mathbb{Z}(0,1)\), and the equilateral torus \((T_{eq}^2 = \mathbb{R}^2/\Gamma_{eq}, g_{eq} = g_{eq})\) with \(\Gamma_{eq} = \mathbb{Z}(1,0) \oplus \mathbb{Z}(1/2, \sqrt{3}/2)\),
each admit a natural homothetic minimal embedding into a sphere. These embeddings, denoted by $\phi_{cl}$ and $\phi_{eq}$, are those induced on $T^2_{cl}$ and $T^2_{eq}$ from $\tilde{\phi}_{cl} : \mathbb{R}^2 \to \mathbb{S}^3$, where $\tilde{\phi}_{cl}(x,y) = \frac{1}{\sqrt{2}}(\exp 2i\pi x, \exp 2i\pi y)$, and $\tilde{\phi}_{eq} : \mathbb{R}^2 \to \mathbb{S}^5$, where $\tilde{\phi}_{eq}(x,y) = \frac{1}{\sqrt{3}}(\exp 4i\pi y/\sqrt{3}, \exp 2i\pi (x - y/\sqrt{3}), \exp 2i\pi (x + y/\sqrt{3}))$.

**Theorem 2.1.** Let $(M,g)$ be a compact orientable surface of genus one and suppose that there exists a full isometric immersion $\phi = (\phi_1, \ldots, \phi_{n+1})$ from $(M,g)$ in the $n$-dimensional unit sphere $\mathbb{S}^n$ such that $\forall i \leq n+1, \phi_i \in E_1(g)$. Then either:

(i) $(M,g)$ is isometric to the normalized Clifford torus $(T^2_{cl}, 2\pi^2 g_{cl})$, $n = 3$ and $\phi$ is equivalent to $\phi_{cl}$, or

(ii) $(M,g)$ is isometric to the normalized equilateral torus $(T^2_{eq}, \frac{8\pi^2}{3} g_{eq})$, $n = 5$ and $\phi$ is equivalent to $\phi_{eq}$.

Recall that an immersion $\phi$ into $\mathbb{S}^n$ is full if its image is not contained in a great sphere of $\mathbb{S}^n$. Two immersions $\phi$ and $\psi$ into $\mathbb{S}^n$ are called equivalent if there exists an isometry $R$ of $\mathbb{S}^n$ such that $\phi = R \circ \psi$. A direct consequence of Theorem 2.1 is:

**Corollary 2.1.** A compact genus one orientable surface $(M,g)$ is $\lambda_1$-minimal if and only if it is homothetic to $(T^2_{cl}, g_{cl})$ or $(T^2_{eq}, g_{eq})$.

As the metrics $g_{cl}$ and $g_{eq}$ trivially satisfy the hypothesis of Proposition 1.1 we have the following:

**Corollary 2.2.** Let $M$ be a compact orientable surface of genus one. A metric $g$ on $M$ is extremal for $\lambda_1$ if and only if $(M,g)$ is homothetic to $(T^2_{cl}, g_{cl})$ or $(T^2_{eq}, g_{eq})$.

The proof of Theorem 2.1 is based on the following Propositions 2.1 and 2.2 which are valid in a more general setting.

**Proposition 2.1.** Let $(M,g)$ be a $n$-dimensional compact Riemannian homogeneous manifold non homothetic to $\mathbb{S}^n$. If a metric $g = fg_0$, conformal to $g_0$, is $\lambda_1$-minimal, then $f$ is constant on $M$.

*Proof.* As $(M,g)$ is $\lambda_1$-minimal non homothetic to $\mathbb{S}^n$ then any conformal diffeomorphism of $(M,g)$ is an isometry (cf. [4]). It follows that any isometry of $(M,g_0)$ is also an isometry of $(M,g)$. Thus the function $f$ is invariant under the isometry group of $(M,g_0)$. The result follows from the homogeneity of $(M,g_0)$. □

**Proposition 2.2.** Let $\eta_1, \eta_2, \ldots, \eta_N$ be $N$ continuous functions on a domain $\Omega$ of $\mathbb{R}^m$ and assume that the $N^2$ functions: $2\eta_j$ ($1 \leq j \leq N$), $\eta_k + \eta_l$ and $\eta_k - \eta_l$ ($1 \leq k < l \leq N$) are non-constant and mutually distinct modulo $2\pi$. If $\phi = (\phi_1, \ldots, \phi_{n+1})$ is a map from $\Omega$ to $\mathbb{S}^n$ such that all its components $\phi_i$ are in the vector space generated by $\{\cos \eta_j, \sin \eta_j, 1 \leq j \leq N\}$,
then there exists an isometry $R$ of $\mathbb{S}^n$ such that

$$R \circ \phi = (\alpha_1 \exp in j_1, \alpha_2 \exp in j_2, \ldots, \alpha_r \exp in j_r, 0, \ldots, 0),$$

where $r \leq (n + 1)/2$, $j_1, \ldots, j_r \in \{1, \ldots, N\}$ and $\alpha_1, \ldots, \alpha_r$ are positive constants satisfying $\sum_{1 \leq j \leq r} \alpha_j^2 = 1$. In particular, $R(\phi(\Omega)) \subset \mathbb{S}^1(\alpha_1) \times \mathbb{S}^1(\alpha_2) \times \cdots \times \mathbb{S}^1(\alpha_r) \times \{0\}$.  

The proof of this proposition is quite elementary and can be omitted.

Proof of Theorem 2.1. In view of Proposition 2.1, it suffices to consider the case where the metric $g$ is flat. It is well known that there exists $(a, b) \in \mathbb{R}^2$; $0 \leq a < \frac{1}{2}$, $b > 0$ and $a^2 + b^2 \geq 1$, such that $(M, g)$ is homothetic to $\left(T_{a,b}^2 = \mathbb{R}^2/\Gamma(a,b), g_{ab} = g_{\Gamma(a,b)}\right)$ with $\Gamma(a,b) = \mathbb{Z}(1,0) \oplus \mathbb{Z}(a,b)$ (cf. [3]). Now the existence of an isometric immersion from $(M, g)$ into the unit sphere by the first eigenfunctions implies that $\lambda_1(g) = 2$. Since $\lambda_1(g_{ab}) = \frac{4\pi^2}{b^2}$, $(M, g)$ is in fact isometric to $\left(T_{a,b}^2, \frac{2\pi^2}{b^2}g_{ab}\right)$. Let $E_{a,b}$ be the first eigenspace of $g_{ab}$ and $\phi : \left(T_{a,b}^2, \frac{2\pi^2}{b^2}g_{ab}\right) \to \mathbb{S}^n$ a full isometric immersion whose components $\phi_i \in E_{a,b}$.

- If $a^2 + b^2 > 1$ then the dimension of $E_{a,b}$ is 2 and there is no such $\phi$.
- If $a^2 + b^2 = 1$ and $(a, b) \neq (1/2, \sqrt{3}/2)$ then $E_{a,b}$ is generated by $\cos \eta_j$, $\sin \eta_j$, $j \leq 2$, with $\eta_1(x,y) = \frac{2\pi y}{b}$ and $\eta_2(x,y) = 2\pi \left(x - \frac{ay}{b}\right)$. From Proposition 2.2, it follows that $n = 3$ and, up to an isometry of $\mathbb{S}^3$, $\phi$ has the form $\phi = (\alpha_1 \exp(i\eta_1), \alpha_2 \exp(i\eta_2))$ with $\alpha_1 > 0$, $\alpha_2 > 0$ and $\alpha_1^2 + \alpha_2^2 = 1$. As $\phi$ is isometric we deduce that $a = 0$, $b = 1$ and $\alpha_1 = \alpha_2 = \sqrt{2}/2$. Thus $(M, g)$ is isometric to $\left(T_{a,b}^2, 2\pi^2g_{cd}\right)$ and $\phi$ is equivalent to $\phi_{eq}$.
- If $(a, b) = (1/2, \sqrt{3}/2)$ then $E_{a,b}$ is generated by $\cos \eta_j$, $\sin \eta_j$, $j \leq 3$, where $\eta_1(x,y) = 4\pi y/\sqrt{3}$, $\eta_2(x,y) = 2\pi \left(x - \frac{y}{\sqrt{3}}\right)$ and $\eta_3(x,y) = 2\pi \left(x + \frac{y}{\sqrt{3}}\right)$. As before $n \leq 5$ and, up to isometry, $\phi = (\alpha_1 \exp(i\eta_1), \alpha_2 \exp(i\eta_2), \alpha_3 \exp(i\eta_3))$ where $\alpha_1$, $\alpha_2$ and $\alpha_3$ are nonnegative constants such that $\alpha_1^2 + \alpha_2^2 + \alpha_3^2 = 1$. As $\phi$ is isometric we obtain $\alpha_1 = \alpha_2 = \alpha_3 = \sqrt{3}/3$. Thus $\phi$ is equivalent to $\phi_{eq}$.

References


Received October 20, 1998 and revised June 15, 1999.

LABORATOIRE DE MATHEMATIQUES ET PHYSIQUE THEORIQUE
UNIVERSITE DE TOURS
PARC DE GRANDMONT
37200 TOURS
FRANCE
E-mail address: elsoufi@univ-tours.fr

LABORATOIRE DE MATHEMATIQUES ET PHYSIQUE THEORIQUE
UNIVERSITE DE TOURS
PARC DE GRANDMONT
37200 TOURS
FRANCE
E-mail address: ilias@univ-tours.fr
THE RUBINSTEIN–SCHARLEMANN GRAPHIC OF A 3-MANIFOLD AS THE DISCRIMINANT SET OF A STABLE MAP

TSUYOSHI KOBAYASHI AND OSAMU SAEKI

We show that Rubinstein–Scharlemann graphics for 3-manifolds can be regarded as the images of the singular sets (: discriminant set) of stable maps from the 3-manifolds into the plane. As applications of our understanding of the graphic, we give a method for describing Heegaard surfaces in 3-manifolds by using arcs in the plane, and give an orbifold version of Rubinstein–Scharlemann’s setting. Then by using this setting, we show that every genus one 1-bridge position of a non-trivial two bridge knot is obtained from a 2-bridge position in a standard manner.

1. Introduction.

In this paper, we show that Rubinstein-Scharlemann graphics for 3-manifolds can be regarded as the images of the singular sets (: discriminant set) of stable maps from the 3-manifolds into the plane, and as applications, we give a method for describing Heegaard surfaces in 3-manifolds by using arcs in the plane, and give an orbifold version of Rubinstein-Scharlemann’s setting. Then by using this setting, we show that every genus one 1-bridge position of a non-trivial two bridge knot is obtained from a 2-bridge position in a standard manner.

In [18], Rubinstein-Scharlemann introduced a powerful machinery, which is called a graphic, for studying Heegaard splittings of 3-manifolds, and succeeded to obtain deep results on the Reidemeister-Singer distance of two strongly irreducible Heegaard splittings of a 3-manifold. We note that Rubinstein and Scharlemann derived a graphic from two Heegaard splittings of a 3-manifold via Cerf theory [5]. Then the purpose of this paper is to introduce another way for understanding the graphic. That is, we show that we can regard a graphic as the image of the singular set of a “stable map” (for definition, see Sect. 3) from the 3-manifold into the plane \( \mathbb{R}^2 \) (Theorem 4.2).

An immediate consequence of this is that we can regard a Heegaard surface as the preimage of an arc in \( \mathbb{R}^2 \), and as an application of our understanding, we first give a method for instructing a procedure for deforming one Heegaard surface to the other by using the arcs as above (Proposition 5.4),
and describe how stabilization works in this setting (Proposition 5.6). In [19], Rubinstein and Scharlemann give a generalization of results in [18] for 3-manifolds with boundary. As the second application, we will give another formulation for generalizing the idea in [18] for link spaces. In fact, we will introduce an orbifold version of the Rubinstein-Scharlemann type argument (Sect. 6), and, by using this, we show that any genus one 1-bridge position of a 2-bridge knot is obtained from a 2-bridge position in a standard manner (Theorem 8.2).

2. Rubinstein-Scharlemann graphic.

Throughout this paper, we work in the differential category, and for standard terminology in 3-dimensional topology, we refer to [9], and [11].

In this section, we quickly review the setting of Rubinstein-Scharlemann’s paper [18].

Let \( M \) be a closed orientable 3-manifold.

**Definition 2.1.** We say that a decomposition \( M = A \cup_P B \) is a (genus \( g \)) Heegaard splitting of \( M \) if \( A, B \) are 3-dimensional genus \( g \) handlebodies in \( M \) such that \( M = A \cup B, A \cap B = \partial A = \partial B = P \). Then \( P \) is called a (genus \( g \)) Heegaard surface of \( M \).

**Definition 2.2.** A disk \( D \) properly embedded in a handlebody \( H \) is called a meridian disk of \( H \) if \( \partial D \) is an essential simple closed curve in \( \partial H \).

**Definition 2.3.** A Heegaard splitting \( M = A \cup_P B \) is stabilized, if there are meridian disks \( D_A, D_B \) of \( A, B \) respectively such that \( \partial D_A \) and \( \partial D_B \) intersects transversely in a single point.

**Remark 2.4.** We note that a genus \( g \) Heegaard splitting \( M = A \cup_P B \) is stabilized if and only if there exists a genus \( g-1 \) Heegaard splitting \( A' \cup_P B' \) such that \( A \cup_P B \) is obtained from \( A' \cup_P B' \) by adding a “trivial” handle. Then we say that \( M = A \cup_P B \) is obtained from \( A' \cup_P B' \) by a stabilization.

**Definition 2.5.** A Heegaard splitting \( M = A \cup_P B \) is reducible, if there exist meridian disks \( D_A, D_B \) of \( A, B \) respectively such that \( \partial D_A = \partial D_B \).

**Definition 2.6.** A Heegaard splitting \( M = A \cup_P B \) is weakly reducible, if there exist meridian disks \( D_A, D_B \) of \( A, B \) respectively such that \( \partial D_A \cap \partial D_B = \emptyset \).

**Remark 2.7.** It is easy to see that if a Heegaard splitting \( M = A \cup_P B \) is reducible then it is weakly reducible. And it is also easy to see that if \( M = A \cup_P B \) is stabilized and is not a genus one Heegaard splitting of the 3-sphere \( S^3 \), then it is reducible.

**Remark 2.8.** It is known, by Haken [8], that if \( M \) is reducible (that is, if \( M \) is a connected sum of two 3-manifolds which are not \( S^3 \)), then any Heegaard splitting of \( M \) is reducible.
Remark 2.9. It is known, by Casson-Gordon [4], that if a Heegaard splitting \( M = A \cup_P B \) is weakly reducible, then either it is reducible, or \( M \) contains an incompressible surface.

Setting of Rubinstein-Scharlemann graphic.

Let \( A \cup_P B \), \( X \cup_Q Y \) be a pair of Heegaard splittings of a closed 3-manifold \( M \). Let \( \Theta_A, \Theta_B, \Theta_X, \Theta_Y \) be spines of \( A, B, X, Y \) respectively such that (except for genus 0, or 1 Heegaard splittings) each vertex of \( \Theta_A, \Theta_B, \Theta_X, \Theta_Y \) has valency 3 (see Figure 2.1). Note that for a genus \( 0 \) handlebody (:the 3-ball) \( B^3 \), we let the spine of \( B^3 \) be a point in \( \text{Int} B^3 \), and for a genus \( 1 \) handlebody (solid torus), we let the spine be a core circle of the solid torus. Then \( M - (\Theta_A \cup \Theta_B) \) is homeomorphic to \( P \times (0, 1) \), where \( P \times \{ \varepsilon \} \) is close to \( \Theta_A \), and \( P \times \{ 1 - \varepsilon \} \) is close to \( \Theta_B \) for a small \( \varepsilon > 0 \). We let \( P_s \) be the surface in \( M \) corresponding to \( P \times \{ s \} \). Then, by regarding \( P_0 = \Theta_A \), and \( P_1 = \Theta_B \), we obtain a continuous map \( H : P \times I \to M \) such that \( H(P, s) = P_s \), and we call this a sweep-out associated to \( A \cup_P B \). Similarly we obtain a sweep-out \( G : Q \times I \to M \) associated to \( X \cup_Q Y \), and set \( G(Q, t) = Q_t \).

Here we may suppose that \( \Theta_A \cup \Theta_B \) and \( \Theta_X \cup \Theta_Y \), \( \Theta_A \cup \Theta_B \) and \( G \), and \( \Theta_X \cup \Theta_Y \) and \( H \) are in general positions. This implies that \( "H(P \times [0, \varepsilon]) \cup H(P \times [1 - \varepsilon, 1])" \) and \( "G(Q \times [0, \varepsilon]) \cup G(Q \times [1 - \varepsilon, 1])" \) and \( H \) are in a "standard" position. That is:

Regard \( G(Q \times (0, 1)) \xrightarrow{G^{-1}} Q \times (0, 1) \xrightarrow{\text{proj}} (0, 1) \) as a height function. Then except for a neighborhood of the maxima and minima (with respect to the height function) and vertices of \( \Theta_A \), each component of \( H(P \times [0, \varepsilon]) \cap Q_t \) is a meridian disk intersecting \( \Theta_A \) in one point (for a small \( \varepsilon > 0 \)), and in the neighborhoods \( Q_t \) looks as in Figure 2.2. The same picture holds for the other pair.

Figure 2.1.
Then, by Cerf [5], we see that for “generic” sweep-outs $H, G$ we obtain a stratification of $\text{Int}(I \times I)$ which consists of four parts below.

**Regions:** Each region is a component of the subset of $\text{Int}(I \times I)$ consisting of values $(s,t)$ such that $P_s$ and $Q_t$ intersect transversely, and this is an open set.

**Edges:** Each edge is a component of the subset consisting of values $(s,t)$ such that $P_s$ and $Q_t$ intersect transversely except for one non-degenerate tangent point. The tangent point is either a “center” or a “saddle”. Edge is a 1-dimensional subset of $\text{Int}(I \times I)$, which is monotonously increasing or decreasing.

**Crossing vertices:** Each crossing vertex is a component of the subset consisting of points $(s,t)$ such that $P_s$ and $Q_t$ intersect transversely except for two non-degenerate tangent points. Crossing vertex is an isolated point in $\text{Int}(I \times I)$. In a neighborhood of a crossing vertex, four edges are coming in, where one can regard the crossing vertex as
the intersection of two edges $\ell_1, \ell_2$ with the signs of the slopes of $\ell_1$ and $\ell_2$ are either different or the same.

**Figure 2.5.**

**Birth-death vertices:** Each birth-death vertex is a component of the subset consisting of points $(s, t)$ such that $P_s$ and $Q_t$ intersect transversely except for a single degenerate tangent point. In particular, there is a parametrization $(\lambda, \mu)$ of $I \times I$ such that $P_s = \{(x, y, z) | z = 0\}$, and $Q_t = \{(x, y, z) | z = x^2 + \lambda \mu y + y^3\}$. A birth-death vertex is an isolated point in $\text{Int}(I \times I)$, and in a neighborhood of a birth-death vertex, two edges $\ell_1, \ell_2$ are coming in, with one from center tangency, the other from saddle tangency, and the signs of the slopes of $\ell_1$ and $\ell_2$ the same.

**Figure 2.6.**
Let $\Gamma$ be the union of edges and vertices above. By the above, $\Gamma$ is a 1-complex in $\text{Int}(I \times I)$. Since we have assumed that $H$, $G$ are standard in a regular neighborhood of $\Theta_A \cup \Theta_B \cup \Theta_X \cup \Theta_Y$, we see that the 1-complex $\Gamma$ naturally extends to $\partial(I \times I)$. We abuse $\Gamma$ to denote this 1-complex, and we call $\Gamma$ a graphic (obtained from the sweep-outs $H$, $G$).

Example 2.10. We show that there exist infinitely many 3-manifolds, and a pair of Heegaard splittings, say $A \cup_P B$, $X \cup_Q Y$, of each 3-manifold such that the corresponding graphic is as in Figure 2.9.

Note that the picture admits 4-fold ($\mathbb{Z}_2 \oplus \mathbb{Z}_2$) symmetry, and we will give an explicit description of the Heegaard surfaces belonging to the lower-left
quarter of \( I \times I \), which can be naturally extended to the whole picture under the above symmetry.

We may regard the Heegaard surface \( P \) as \( P_{1/2} \), and \( Q \) as \( Q_{1/2} \). Then \( A \) and \( A \cap Q_{\epsilon} (0 \leq \epsilon \leq 1/2) \) look like as follows. (Here we suppose that \( A \) admits a symmetry generated by \( \varphi_1, \varphi_2 \) in Figure 2.10, where \( \varphi_1 \) is an orientation preserving involution which changes the right side and the left side of \( A \) with the fixed point set an arc properly embedded in \( A \), and \( \varphi_2 \) is an orientation reversing involution which changes the right side and the left side of \( A \) with the fixed point set a disk properly embedded in \( A \).)

\( A \cap Q_0 (= A \cap \Theta_X) \) is a 1-complex as in Figure 2.10. Then, when \( \epsilon \) is sufficiently small, \( A \cap Q_{\epsilon} \) is the frontier of a regular neighborhood of the 1-complex (the surface is homeomorphic to a torus with one hole). If we make \( \epsilon \) bigger, then we come to the point (1) of Figure 2.9 and simultaneously four points in the boundary of a torus \( Q_{\epsilon} \) tend to four directions as in (1) of Figure 2.10.

Then, by making \( \epsilon \) bigger further, we come to the point (2) and, then, (3) (: \( \epsilon = 1/2 \)), where the corresponding figures of \( A \cap Q_{\epsilon} \) look as in Figure 2.11.

When we come to the point (2), the boundary component touches itself simultaneously in two places. In the right side, a band is produced, and, in the left side, the surface is boundary compressed when we pass the point (2).

Note that \( A \cap Q_{1/2} \) is a vertical surface (which is a disk with two holes), which is located in the middle of \( A \), that is, \( A \cap Q_{1/2} \) is invariant under \( \varphi_1 \), and \( \varphi_1 \) exchanges the components of \( A - Q_{1/2} \).
Figures 2.12 describes the deformations (4) → (5). When we come to (5) from (1) with passing the edge containing (4), a torus with one hole is boundary compressed and becomes an annulus.

![Figure 2.12.](image)

Figure 2.12.

Figure 2.13 describes the surface of (6). When we come to (6) from (5), two points of a boundary component of an annulus tend to the right side and touch and a band is produced, and simultaneously two points contained in different boundary components of an annulus tend to the back of \( A \) and the surface is boundary compressed.

![Figure 2.13.](image)

Figure 2.13.

Figures 2.14 and 2.15 describe the surfaces of (7), (8), and (9). At (7) \( Q \cap A \) is an inessential disk properly embedded in \( A \). When we come to (8) from (7), two points of the boundary of a disk tend to the right side and touch at a middle part of \( A \), and simultaneously two points in the boundary of a disk tend to the back of \( A \) and the surface is boundary compressed. As a result, an inessential disk becomes a separating essential disk in \( A \).

![Figure 2.14.](image)

Figure 2.14.
Figure 2.15.

Figure 2.16 describes (10). When we come to (10) from (7) with passing an edge, an inessential disk is boundary compressed and becomes two non-separating disks.

Let $A \cap Q_1 - \varepsilon = \varphi_1(A \cap Q_{\varepsilon})$ $(0 \leq \varepsilon \leq 1/2)$, which gives the whole description of $Q_t$ $(0 \leq t \leq 1)$ in $A$.

Let $B$ be a copy of $A$, and $f : A \to B$ the homeomorphism induced by the identification, and $\phi = f|_{\partial A} : \partial A \to \partial B$ the corresponding homeomorphism. Let $\mathcal{F}$ be the pattern on $\partial A$ induced by $Q_t$'s. Let $M$ be the 3-manifold obtained by attaching $A$ to $B$ by the homeomorphism $\phi \circ \varphi_2$, where $\varphi_2 : \partial A \to \partial A$ is isotopic to $\varphi_2$ with $\varphi_2(\mathcal{F}) = \mathcal{F}$, and $\varphi_2(\ell_2) = \ell_1$, $\varphi_2^{-1}(\ell_1) = \ell_2$. (Note that $M$ is actually the connected sum of two $S^2 \times S^1$'s.) We note that $\mathcal{F}$ is invariant under $\varphi_2$, and, hence, the surfaces $Q_t$ $(0 < t < 1)$ in $A$ are matched to the surfaces $f \circ \varphi_2(Q_t)$ $(0 < t < 1)$ in $B$ to make a system of closed surfaces, say $Q_t$ again, in $M$. It is directly observed from the pictures that $Q_t$ gives a sweep out $G : Q \times I \to M$, and, by construction, we immediately see that the corresponding graphic is as in Figure 2.9.

Then let $\ell_1, \ell_2$ be the components of $\partial(A \cap Q_{1/2})$ as in Figure 2.11, and $D_i : \partial A \to \partial A$ $(i = 1, 2)$ the Dehn twist along $\ell_i$. For a pair of integers $(p, q)$, we let $M_{(p,q)}$ be the manifold obtained by attaching $B$ to $A$ by the homeomorphism $\phi \circ D_1^p \circ D_2^q \circ \varphi_2 : \partial A \to \partial B$. Since there is a regular neighborhood $N(\ell_i, \partial A)$ such that $\mathcal{F}$ restricted to $N(\ell_i, \partial A)$ is a foliation by circles parallel to $\ell_i$, we can arrange so that the configuration of the sweep-outs $H$ and $G$ are respected in $M_{(p,q)}$ (and, hence, the corresponding graphic is the same as above). It is easy to see that $M_{(p,q)}$ is a connected
sum of two lens spaces $L(p, 1)$, and $L(q, 1)$, which implies the existence of infinitely many examples.

We note that it is easy to see from Figure 2.11, that the Heegaard surfaces are isotopic in each of the examples.

**Example 2.11.** By using the arguments in Example 2.10, we show that there exist infinitely many 3-manifolds, and a pair of Heegaard splittings, say $A \cup P \cup B, X \cup Q \cup Y$, of each 3-manifold such that the corresponding graphic is as in Figure 2.17.

![Figure 2.17](image1)

**Figure 2.17.**

As in Figure 2.10, $Q_\varepsilon$ is a torus with one hole which is the frontier of a regular neighborhood of a 1-complex for a small $\varepsilon > 0$. Four points in the boundary of the torus with one hole tend to four directions according as the expansion of the regular neighborhood.

![Figure 2.18](image2)

**Figure 2.18.**

When we come to (3a) from (1) with passing (2), a band which goes around the right handle twice is produced, and a slit in the surface which goes around the left handle twice occurs to produce a boundary compression. As a result, a torus with one hole becomes a disk with two holes.
Figure 2.19.

Note that in (3b), $Q_{1/2}$ is invariant with respect to the involution $\varphi_1$.

Figure 2.20a.

When we come to (6) from (3b) with passing (4), two boundary compressions occur, and a disk with two holes becomes a separating essential disk in $A$.

Figure 2.20b.

Figure 2.21.
In the following, we show pictures with turning back to front for the convenience of drawing.

When we come to (5) from (3b) with passing an edge, the right band is boundary compressed and a disk with one hole becomes an annulus.

When we come to (6) from (5) with passing an edge, an annulus is boundary compressed to become a separating essential disk.

When we come to (6) from (1) with passing (7), two bands, one of which goes around the left handle once and the other goes around the left handle twice are produced.

When we come to (8) from (1) with passing an edge, a torus with one hole is boundary compressed and the punctured torus become an annulus. When we come to (8) from (6) with passing an edge, a band which goes around the left handle once is attached and the disk becomes an annulus.
Figure 2.25.

When we come to (6) from (10) with passing (9), two points of the boundary of an inessential disk tend to the right side and touch at a middle part of $A$, and simultaneously two points in the boundary of a disk tend to the back of $A$ and touch to produce a boundary compression. As a result, an inessential disk becomes a separating essential disk in $A$.

Figure 2.26.

When we come to (11) from (10) with passing an edge, an inessential disk is boundary compressed and becomes two non-separating disks.

Figure 2.27.

Note that, by using the arguments in the previous example, we can show that such 3-manifolds are obtained from $A$ and $B$ by pasting their boundaries applying the Dehn twists along $\ell$ in Figure 2.20b. Note that, as a result of this construction, we obtain 3-manifolds each of which is a union of two Seifert fibered spaces with orbit space a disk with two exceptional fibers of index two, and the exterior of a $(2,2n)$-torus link (see [13]). Note that except in one case (case $n = 0$) they are Haken manifolds.
3. Stable maps.

The purpose of this section is to show that any differentiable map from an $n$-manifold into a surface can be deformed to an “excellent” (stable) map, and this assertion is an essential part of this paper. In the following, manifolds have countable basis and all manifolds and maps are assumed to be $C^\infty$.

Let $M$ be a connected $n$-dimensional manifold (possibly with boundary) with $n \geq 2$ and $N$ a surface without boundary. For a smooth map $f: M \to N$, $S(f)$ denotes the singular set of $f$; i.e., $S(f)$ is the set of the points in $M$ where the rank of the differential $df$ is strictly less than 2. The discriminant set is the image of the singular set, $f(S(f))$. We denote by $C^\infty(M, N)$ the space of the smooth maps of $M$ into $N$ endowed with the Whitney $C^\infty$ topology (fine topology) (see [7, 10]).

Definition 3.1. A smooth map $f: M \to N$ is stable if there exists a neighborhood $U$ of $f$ in $C^\infty(M, N)$ such that for each $g \in U$ there exist diffeomorphisms $H: M \to M$ and $h: N \to N$ which make the following diagram commutative.

\[
\begin{array}{ccc}
M & \xrightarrow{f} & N \\
H \downarrow & & \downarrow h \\
M & \xrightarrow{g} & N
\end{array}
\]

Definition 3.2. Let $f: M \to N$ be a proper smooth map of an $n$-dimensional manifold $M$ ($n \geq 2$) into a surface $N$ without boundary. For an open set $U$ of $N$, we say that $f$ is excellent on $U$, if $f^{-1}(U) \cap \partial M = \emptyset$ and the following conditions are satisfied: For all $p \in f^{-1}(U)$, there exist local coordinates $(u, x, y_1, \ldots, y_{n-2})$ centered at $p$ and $(X, Y)$ centered at $f(p)$ such that $f$ has one of the following forms:

$L_0)$ $X \circ f = u$, $Y \circ f = x$ \hspace{1cm} (p: regular point)

$L_1)$ $X \circ f = u$, $Y \circ f = \pm x^2 + \sum_{i=1}^{n-2} \pm y_i^2$ \hspace{1cm} (p: fold point)

$L_2)$ $X \circ f = u$, $Y \circ f = ux - x^3 + \sum_{i=1}^{n-2} \pm y_i^2$ \hspace{1cm} (p: cusp point);

and

$G_1)$ If $p \in f^{-1}(U)$ is a cusp point, then $f^{-1}(f(p)) \cap S(f) = \{p\}$,

$G_2)$ $f|_{(S(f) \cap f^{-1}(U) - \{cusp points\})}$ is an immersion with normal crossings.
Note that it is well known that a proper smooth map \( f : M \to N \) of a manifold \( M \) with \( \partial M = \emptyset \) is stable if and only if \( f \) is excellent on \( N \) (see [7, 14]). The terminology “excellent” comes from [22, §2].

The main purpose of this section is to prove the following.

**Theorem 3.3.** Let \( f : M \to N \) be a proper smooth map. Suppose that \( F \) is a closed 2-dimensional submanifold (possibly with boundary) of \( N \) such that \( f^{-1}(\text{Int } F) \supset \partial M \). Furthermore we suppose that \( f \) is excellent on a neighborhood of \( \partial F \) and that \( f \) is transverse to \( \partial F \). Furthermore, let \( \mathcal{V} \) be an arbitrary open neighborhood of \( f \) in \( C^\infty(M, N) \). Then there exists a smooth map \( g : M \to N \) such that

1. \( g \in \mathcal{V} \),
2. \( g = f \) on \( f^{-1}(F) = g^{-1}(F) \),
3. \( g \) is excellent on a neighborhood of \( N \setminus \text{Int } F \).

**Remark 3.4.** In Theorem 3.3 the condition (2) is essential. In fact, if we drop the condition (2), it has already been well known.

In the following, we use the following notation. For manifolds \( X \) and \( Y \), \( J^k(X, Y) \) denotes the \( k \)-jet bundle over \( X \times Y \), i.e., \( J^k(X, Y) = \{(x, y, j^k(x)) \mid x \in X, y \in Y \exists f : X \to Y : C^\infty \text{ s.t. } f(x) = y \} \). For integers \( s \geq 1 \) and \( k \geq 0 \), \( J^k_s(X, Y) \) denotes the \( s \)-fold \( k \)-jet bundle ([7, p. 57]). We denote by \( X^{(s)} \) the subset of \( X^s = X \times \cdots \times X \) (the \( s \)-fold product space of \( X \)) consisting of the elements \( (x_1, \cdots, x_s) \) such that \( x_i \neq x_j \) for \( i \neq j \). We denote by \( \pi : J^k_s(X, Y) \to X^{(s)} \times Y^s \) the canonical projection and by \( \pi_Y : J^k_s(X, Y) \to Y^s \) the natural projection to the target. We set \( \Delta^k_Y = \{(y, \cdots, y) \in Y^s \} \) and \( d : \Delta^k_Y \to Y \) is the natural identification map. Furthermore, for a smooth map \( f : X \to Y \), \( J^k_{s} f : X^{(s)} \to J^k_s(X, Y) \) denotes the \( s \)-fold \( k \)-jet of \( f \). (For details, see [7, Chapter II, §4].)

In order to prove Theorem 3.3, we need the following.

**Proposition 3.5.** Let \( f : X \to Y \) be a smooth map between manifolds (\( Y \) need not be a surface). Let \( W \) be a submanifold of \( J^k_s(X, Y) \) such that \( \pi_Y(W) \subset \Delta^k_Y \). Suppose that \( U \) is an open subset of \( Y \) and that \( \mathcal{V} \) is an open neighborhood of \( f \) in \( C^\infty(X, Y) \). Then there exists a smooth map \( g : X \to Y \) such that

1. \( g \in \mathcal{V} \),
2. \( g = f \) on \( f^{-1}(U) = g^{-1}(U) \),
3. \( J^k_s g \) is transverse to \( W \) on \( W \cap \pi_Y^{-1}(d^{-1}(Y \setminus U)) \).

**Proof.** Set \( W' = W \cap \pi_Y^{-1}(d^{-1}(Y \setminus U)) \), which is an open submanifold of \( W \). Then there exists a countable family \( \{W_r\}_{r=1}^\infty \) of open sets of \( W' \) with the following properties (a)-(f).

(a) \( \cup_{r=1}^\infty W_r = W' \).
(b) \( \overline{W}_r \subset W'_r \), where \( \overline{W}_r \) denotes the closure of \( W_r \) in \( J^k_s(X,Y) \).
(c) \( \overline{W}_r \) is compact.
(d) There exist coordinate neighborhoods \( U_{r,1}, \ldots, U_{r,s} \) in \( X \) and \( V_{r,1}, \ldots, V_{r,s} \) in \( Y \) such that \( \{U_{r,i}\}_{i=1}^s \) are mutually disjoint and \( \pi(\overline{W}_r) \subset U_{r,1} \times \cdots \times U_{r,s} \times V_{r,1} \times \cdots \times V_{r,s} \).
(e) \( \overline{U}_{r,i} \) is compact for \( 1 \leq i \leq s \).
(f) \( V_{r,i} \cap \overline{U} = \emptyset \) for \( 1 \leq i \leq s \).

Using this family \( \{W_r\}_{r=1}^\infty \) in the argument of [7, proof of Theorem 4.13 (p. 58)] or [15, pp. 311-312], we see that there exists a smooth map \( g_r : X \to Y \) such that \( j^k_s g_r : X^{(s)} \to J^k_s(X,Y) \) is transverse to \( W_r \) on \( \overline{W}_r \) and that \( g_r = f \) on \( f^{-1}(U) = g_r^{-1}(U) \) in an arbitrary neighborhood of \( f \) in \( C^\infty(X,Y) \). Thus, putting

\[ C_{f,U} = \{ g \in C^\infty(X,Y) : g = f \text{ on } f^{-1}(U) = g^{-1}(U) \}, \]

we see that

\[ D_r = C_{f,U} \cap \{ g \in C^\infty(X,Y) : j^k_s g \text{ is transverse to } W \text{ on } \overline{W}_r \} \]

is dense in \( C_{f,U} \). On the other hand, \( D_r \) is open by [7, Lemma 4.14 (p. 57)]. The proposition is proved if we show that \( \cap_{r=1}^\infty D_r \) is dense in \( C_{f,U} \). Thus we have only to show that \( C_{f,U} \) is a Baire space (see [7, Definition 3.2 (p. 44)]). First note that \( C_{f,U} \) is a closed subset of \( C^\infty(X,Y) \). Then by imitating the proof of [7, Proposition 3.3 (p. 44)], we see easily that \( C_{f,U} \) is a Baire space. This completes the proof.

Let \( M \) and \( N \) be as in Theorem 3.3. We consider some submanifolds of the (multi-)jet bundles as follows. For the jet bundle \( J^3(M,N) \), consider the four submanifolds \( \Sigma^{n-1,0}, \Sigma^{n,0}, \Sigma^{n-1,1,0} \) and \( \Sigma^{n-1,1,1,0} \) as defined in [1] (or [7, p. 156, Sect. 5]). Note that their codimensions are equal to \( n-1, 2n, n \) and \( n+1 \) respectively by [1, Theorem (6.2)]. For the multi-jet bundle \( J_2^3(M,N) \), we consider

\[ S_1 = \{ (j^3 f(p), j^3 g(q)) : f(p) = g(q), j^3 f(p) \in \Sigma^{n-1,0}, j^3 g(q) \in \Sigma^{n-1,0} \}, \]

\[ S_2 = \{ (j^3 f(p), j^3 g(q)) : f(p) = g(q), j^3 f(p) \in \Sigma^{n-1,0}, j^3 g(q) \in \Sigma^{n-1,1,0} \}, \]

\[ S_3 = \{ (j^3 f(p), j^3 g(q)) : f(p) = g(q), j^3 f(p) \in \Sigma^{n-1,1,0}, j^3 g(q) \in \Sigma^{n-1,1,0} \}. \]

For the multi-jet bundle \( J_3^3(M,N) \), we consider

\[ S_3 = \{ (j^3 f(p), j^3 g(q), j^3 h(r)) : f(p) = g(q) = h(r), \]

\[ j^3 f(p), j^3 g(q), j^3 h(r) \in \Sigma^{n-1,0} \}. \]

Note that \( S_1, S_2, S_3 \) and \( S_3 \) are easily seen to be submanifolds and that their codimensions are equal to \( 2n, 2n+1, 2n+2 \) and \( 3n+1 \) respectively.

For a smooth map \( f : M \to N \), we have the following facts:
(1) \( j^3 f \) is transverse to \( \Sigma^{n-1,0}, \Sigma^{n,0}, \Sigma^{n-1,1,0} \) and \( \Sigma^{n-1,1,1,0} \) if and only if \( f \) exhibits only fold and cusp points as its singularities.
(2) Suppose \( f \) satisfies (1). Then \( j^3_2f \) is transverse to \( S^3_2 \) if and only if \( f\mid_{(S(f) \setminus \{\text{cusp points}\})} \) has no multiple points of multiplicity greater than two.

(3) Suppose \( f \) satisfies (1) and (2). Then \( j^3_2f \) is transverse to \( S^4_3 \) if and only if \( f\mid_{(S(f) \setminus \{\text{cusp points}\})} \) is an immersion with normal crossings (see [7, Proposition 5.6 (p. 158)]).

(4) Suppose \( f \) satisfies (1). Then \( j^3_2f \) is transverse to \( S^2_3 \) and \( S^3_3 \) if and only if for every cusp point \( p \) of \( f \), we have \( f^{-1}(f(p)) \cap S(f) = \{p\} \).

Using the above facts, we obtain the following.

**Lemma 3.6.** Let \( f : M \to N \) be a proper smooth map of an \( n \)-dimensional manifold \( M \) \((n \geq 2)\) into a surface \( N \). For an open set \( U \) of \( N \), \( f \) is excellent on \( U \) if and only if \( f^{-1}(U) \cap \partial M = \emptyset \) and the jets of \( f \) are transverse to \( \Sigma^{n-1,0}, \Sigma^{n,0}, \Sigma^{n-1,1,0}, \Sigma^{n-1,1,1}, S^1_2, S^3_2, S^3_2 \) and \( S^3_3 \) on the part corresponding to \( f^{-1}(U) \).

**Proof of Theorem 3.3.** Set \( U = \text{Int} F \). By Proposition 3.5 and Lemma 3.6, we see that there exists a smooth map \( g : M \to N \) such that \( g \in \mathcal{V} \), \( g = f \) on \( f^{-1}(U) = g^{-1}(U) \) and that \( g \) is excellent on \( N - F \). Since \( f \) and \( g \) are continuous and \( N \) is Hausdorff, we see that \( g = f \) on the closure of \( f^{-1}(U) \).

However, we do not know if \( g \) is excellent on a neighborhood of \( N - \text{Int} F \). This is because there is a possibility of a point in \( f^{-1}(N - F) \) being mapped into \( \partial F \) by \( g \). In order to exclude this possibility, we modify the argument as follows.

Since \( \partial F \) is a closed submanifold of \( N \), the set of maps of \( M \) into \( N \) transverse to \( \partial F \) forms an open set of \( C^\infty(M, N) \) (see [7, Proposition 4.5 (p. 52)]). Thus we may assume that every map in the open set \( V \) is transverse to \( \partial F \) from the beginning. Furthermore, since the set of the proper maps of \( M \) into \( N \) forms an open set (see [10, Theorem 1.5 (p. 38)]), we may further assume that every element of \( V \) is a proper map. We may further assume that each element of \( V \) maps \( \partial M \) into \( F - V \) by a similar reason, where \( V \) is a closed neighborhood of \( \partial F \) in \( N \). Now suppose that \( g \in V \). Then, since \( g^{-1}(\partial F) \) is a closed regular submanifold of \( \text{Int} M \), we see that the closure of \( g^{-1}(U) \) is equal to \( g^{-1}(F) \). Since \( f = g \) on the closure of \( f^{-1}(U) \) and \( f^{-1}(U) = g^{-1}(U) \), we see that \( f = g \) on \( f^{-1}(F) = g^{-1}(F) \). Combining the facts that \( g \) is excellent on \( N - F \) and that \( f \) is excellent on a neighborhood of \( \partial F \), we see that \( g \) is excellent on a neighborhood of \( N - \text{Int} F \). This completes the proof of Theorem 3.3. \( \Box \)

**Remark 3.7.** Results similar to Theorem 3.3 hold for some other dimension pairs as well.

### 4. Graphic as the discriminant set.

Let \( \Theta_A, \Theta_B, \Theta_X, \Theta_Y, H, G \) be as is Section 2, where \( H, G \) may not be generic. In this section, we first observe that we can obtain a smooth map
$f : M - (\Theta_A \cup \Theta_B \cup \Theta_X \cup \Theta_Y) \to I \times I$ from $H$ and $G$, and we show, by using Theorem 3.3, that $f$ can be deformed to a map $\Phi$ which is excellent in the exterior of a regular neighborhood of $\Theta_A \cup \Theta_B \cup \Theta_X \cup \Theta_Y$ by an arbitrarily small deformation. Then we see that we can obtain sweep-outs $H', G'$ associated to $A \cup B, X \cup Y$ respectively from $\Phi$, which have the feature “generic” in Sect. 2. Finally we observe that the corresponding graphic is actually the discriminant set (for the definition, see Sect. 3) of $\Phi$.

Let $M$ be a closed 3-manifold. In this section we consider a smooth map $f$ from $M$ to the Euclidean space $\mathbb{R}^2$. Recall that $S(f)$ denotes the set of singular points (or singular set) of $f$. That is,

$$S(f) = \{q \in M \mid \text{rank } (df_q : T_q M \to T_{f(q)} \mathbb{R}^2) \leq 1\}.$$

Then as a special situation of Definition 3.2, we have:

**Definition 4.1.** Let $f : M \to \mathbb{R}^2$ be a smooth map. For an open set $U$ of $\mathbb{R}^2$, we say that $f$ is excellent on $U$ if $f^{-1}(U) \cap \partial M = \emptyset$, and the following conditions are satisfied.

1. For each point $q \in S(f)$ there exist local coordinates $(u, x, y)$ for $q$, and $(X, Y)$ for $f(q)$ such that:
   1.1. $X \circ f = u, Y \circ f = x^2 + y^2$, or
   1.2. $X \circ f = u, Y \circ f = x^2 - y^2$, or
   1.3. $X \circ f = u, Y \circ f = y^2 + ux - x^3$.
2. For each cusp $q$, $f^{-1}(f(q)) \cap S(f) = \{q\}$ (that is, the fiber which contains $q$ does not contain another singular point).
3. $f|_{S(f) - \{\text{cusps}\}}$ is an immersion (possibly) with normal crossing (that is, an immersion (possibly) with transverse self intersections).

We call a singular point of type (1-1) ((1-2) resp.) a **definite fold** (**indefinite fold** resp.). Recall that a singular point of type (1-3) is called a cusp.

![Figure 4.1](image-url)
Now we describe the relationship between graphic and excellent map. Let $M$, $A \cup P B$, $X \cup Q Y$, $\Theta_A$, $\Theta_B$, $\Theta_X$, and $\Theta_Y$ be as in Section 2. (Here we suppose that $\Theta_A \cup \Theta_B$ and $\Theta_X \cup \Theta_Y$ are in general position.)

Let $H$, $G$ be sweep-outs obtained from the Heegaard splittings $A \cup P B$, $X \cup Q Y$ respectively. We may suppose that $H|_{P \times (0,1)} : P \times (0,1) \to M - (\Theta_A \cup \Theta_B)$ and $G|_{Q \times (0,1)} : Q \times (0,1) \to M - (\Theta_X \cup \Theta_Y)$ are smooth. Let $\Phi : M - (\Theta_A \cup \Theta_B \cup \Theta_X \cup \Theta_Y) \to I \times I$ be the map defined by:

\[
\Phi(p) = (s, t) \iff p \in P_s \cap Q_t.
\]

Since $H|_{P \times (0,1)}$, $G|_{Q \times (0,1)}$ are smooth maps, we see that $\Phi$ is also smooth. Since $\{P_s\}$ and $\Theta_X \cup \Theta_Y$, $\{Q_t\}$ and $\Theta_A \cup \Theta_B$ are generic, we see that:
\((\ast 1)\) \(H\) (\(G\) resp.) is standard (see Sect. 2) in a small regular neighborhood \(N(\Theta_X \cup \Theta_Y)\) of \(\Theta_X \cup \Theta_Y\) \((N(\Theta_A \cup \Theta_B)\) of \(\Theta_A \cup \Theta_B\) resp.), and, hence, \(\Phi\) is transverse to the frontier of a regular neighborhood of \(\partial(I \times I)\) in \(I \times I\).

By Theorem 3.3, we see that we can deform \(\Phi\) in \(M - (N(\Theta_A \cup \Theta_B) \cup N(\Theta_X \cup \Theta_Y))\) by an arbitrarily small deformation, to a map \(\Phi'\) which is excellent on the complement of the regular neighborhood of \(\partial(I \times I)\) in \(I \times I\).

Since \(\Phi'\) is obtained from \(\Phi\) by a small deformation, we may suppose:

\((\ast 2)\) \(pr_1 \circ \Phi', pr_2 \circ \Phi'\) have no critical points, where \(pr_1, pr_2: I \times I \rightarrow I\) are the projections to the first, and second factors respectively.

By condition \((\ast 2)\), we see that, there exist sweep-outs \(H', G'\) such that \(H'(P \times \{s\}) = \Phi'^{-1}(\{s\} \times I)\), and \(G'(Q \times \{t\}) = \Phi'^{-1}(I \times \{t\})\). Note that \(H', G'\) are small deformations of \(H, G\). By the definition of \(H', G'\), we immediately have:

\[(\ast)'\] \(\Phi'(p) = (s, t) \iff p \in H'(P \times \{s\}) \cap G'(Q \times \{t\})\).

Then by Definition 4.1 (and the definition of the graphic in Sect. 2) we see that \(H'\) and \(G'\) are generic in the sense of Rubinstein and Scharlemann (see Sect. 2) and, by comparing Definition 4.1 and the definition of the graphic in Sect. 2, it is easy to see that the corresponding graphic is actually the image of the singular set of \(\Phi\) on \(M - (\Theta_A \cup \Theta_B \cup \Theta_X \cup \Theta_Y)\), where the image of a definite fold corresponds to a center tangency, the image of an indefinite fold corresponds to a saddle tangency, and the image of a cusp corresponds to a birth-death vertex.

Now we summarize the above results to give:

**Theorem 4.2.** Let \(H, G\) be as above. Then, by an arbitrarily small deformation of \(H\) and \(G\), we obtain sweep-outs \(H'\) and \(G'\) such that:

1. The above map \(\Phi'\) (see \((\ast)'\)) is excellent on \(\text{Int}(I \times I)\),
2. The maps \(H'\) and \(G'\) are generic. Hence we can obtain a graphic \(\Gamma\) from \(H'\) and \(G'\), and then \(\Gamma \cap \text{Int}(I \times I)\) is the discriminant set of the excellent map \(\Phi'|_{M - (\Theta_A \cup \Theta_B \cup \Theta_X \cup \Theta_Y)}\).

5. Isotopy, and stabilization of Heegaard surfaces.

Let \(f: M \rightarrow I \times I(\subset \mathbb{R}^2)\) be an excellent map obtained in the previous section (which is denoted by \(\Phi'\) there). Let \(\alpha: I \rightarrow \mathbb{R}^2\) be an embedding of the unit interval.

**Definition 5.1.** We say that \(\alpha\) is transverse to \(f\) if \(\alpha\) satisfies the following two conditions:
(1) \( \alpha(\partial I) \subset \partial(I \times I) \), and \( \alpha \) and \( \partial(I \times I) \) are transverse (i.e., a smooth slight extension of \( \alpha \) is transverse to smooth extensions of \( I \times \{0\} \), \( I \times \{1\} \), \( \{0\} \times I \), and \( \{0\} \times I \)).

(2) for each pair \((t,q) \in (0,1) \times M\) with \( \alpha(t) = f(q) \), we have:
\[
\frac{df_q(T_qM)}{dt} + \frac{d\alpha_t(T_tI)}{dt} = T_{f(q)}R^2.
\]

Then, by Definition 4.1, it is easy to see:

**Lemma 5.2.** Suppose that \( \alpha \) satisfies Definition 5.1 (1). Then \( \alpha \) is transverse to \( f \) if and only if:

- For \( q \in M \) with \( f(q) \in \alpha(I) \), we have either one of:
  1. \( q \) is not a singular point of \( f \),
  2. \( q \) is a fold point which is not a normal crossing, and \( \alpha \) is transverse to the discriminant set at \( f(q) \),
  3. \( q \) is a fold point which is mapped to a normal crossing, and \( \alpha \) is transverse to the two arcs (which are local images of the singular set), or
  4. \( q \) is a cusp. In this case, there are two arcs in a neighborhood of \( f(q) \) (which are local images of the singular set). Then there is a slight smooth extensions of the arcs, say \( \ell_1, \ell_2 \) such that \( \alpha \) is transverse to \( \ell_1 \), and \( \ell_2 \).

![Figure 5.1](image_url)

**Figure 5.1.**

**Lemma 5.3.** Suppose that \( \alpha \) is transverse to \( f \). Then \( f^{-1}(\alpha(I)) \) is a 2-dimensional submanifold of \( M \).

**Proof.** By condition Definition 5.1 (2), we see that \( f^{-1}(\alpha(0,1)) \) is a 2-dimensional proper submanifold in \( M - (\Theta_A \cup \Theta_B \cup \Theta_X \cup \Theta_Y) \). Then by condition Definition 5.1 (1), we see that \( f^{-1}(\alpha(0,\varepsilon)) \) \( f^{-1}(\alpha[1-\varepsilon,1]) \) resp.) is a disk, which cap off \( f^{-1}(\alpha(0,1)) \) to make a closed surface \( f^{-1}(\alpha(I)) \).

**Proposition 5.4.** Let \( \alpha, \beta \) be arcs transverse to \( f \). Suppose that \( \alpha(I) \) is deformed to \( \beta(I) \) through a sequence of moves of the following types.

(0) Ambient isotopy of \( I \times I \) which fixes \( \partial(I \times I) \cup \Gamma \) setwise.

(1) Passing a crossing vertex as \( \alpha_- \to \alpha_0 \to \alpha_+ \) in Figure 5.2.
Figure 5.2.

(2) Passing a cusp as $\alpha_\to \to \alpha_0 \to \alpha_+ \to \alpha_0$ in Figure 5.3.

Figure 5.3.

(3) Passing a vertex in $\partial(I \times I)$ as in Figure 5.4.

Figure 5.4.

(4) Passing a corner of $\partial(I \times I)$ as in Figure 5.5.

Figure 5.5.

Then the surfaces $f^{-1}(\alpha(I))$, and $f^{-1}(\beta(I))$ are isotopic in $M$.

Proof. First, we consider moves (0), (1), and (2). Suppose that $\alpha_0(I)$ is deformed to $\alpha_1(I)$ through a sequence of moves (0), (1), and (2). By Lemma 5.2, we see that there is a 1-parameter family of transverse arcs
\( \alpha_s \) \((0 \leq s \leq 1) \) from \( \alpha_0 \) to \( \alpha_1 \). Then we obtain an isotopy of surfaces \( f^{-1}(\alpha_s) \) in \( M \).

Now we consider move \((3)\). Note that the deformation \((3)\) gives the isotopy as in Figure 5.6.

\[ \text{Figure 5.6.} \]

This shows that the deformation \((3)\) gives mutually isotopic surfaces.

Finally we consider about move \((4)\). Note that in a neighborhood of a corner of \( I \times I \), \( P_s \) and \( Q_t \) are disjoint. Hence the deformation \((4)\) obviously gives equal surfaces.

Combining the above observations, we have the conclusion of the proposition. \( \square \)

As a consequence of Proposition 5.4. we have:

**Corollary 5.5** (cf. Example 2.10). Suppose that the graphic obtained from \( P \) and \( Q \) contains a region as in Figure 5.7. Then \( P \) and \( Q \) are isotopic in \( M \).

\[ \text{Figure 5.7.} \]

**Proof.** Let \( \alpha(t) = (t, \varepsilon) \), and \( \beta(t) = (\varepsilon, t) \) for a small \( \varepsilon > 0 \). It is easy to see that \( \alpha(I) \) is deformed to \( \beta(I) \) within the above region by applying the deformations of Proposition 5.4. \( \square \)

For a stabilization of a Heegaard splitting, we have:

**Proposition 5.6.** Let \( \alpha \) be an arc transverse to \( f \) such that \( f^{-1}(\alpha(I)) \) is a Heegaard surface. Suppose that a transverse arc \( \alpha' \) is obtained by changing
\( \alpha \) locally as in Figure 5.8. Then \( f^{-1}(\alpha'(I)) \) is a Heegaard surface which is a stabilization of \( \Phi^{-1}(\alpha(I)) \).

![Figure 5.8.](image)

**Proof.** We use the following picture (Figure 5.9) for the proof. The picture corresponds to the point of the intersection of \( \alpha(I) \) and the image of an indefinite fold (here \( P_s \)'s are represented by horizontal planes).

![Figure 5.9.](image)

Let \( P_i \) be the subsurface of \( \Phi^{-1}(\alpha'(I)) \) corresponding to \( \Phi^{-1}(\alpha_i) \), where \( \alpha_i \) is as in Figure 5.8. It is directly observed from Figure 5.9 that each \( P_i \) looks as in Figure 5.10.

![Figure 5.10.](image)

By summing up \( P_i \)'s, we see that \( \Phi^{-1}(\alpha'(I)) \) is a stabilization of \( \Phi^{-1}(\alpha(I)) \).

**Corollary 5.7.** Let \( P \) and \( Q \) be the Heegaard surfaces as in Example 2.11. Then \( P \) and \( Q \) become isotopic by applying one stabilization.
Proof. We first take transverse arcs \( \alpha \) and \( \beta \) as in Figure 5.11. Then, by Figure 5.11, we see that \( \alpha(I) \) can be deformed to the arc in Figure 5.12, by one application of the deformation of Proposition 5.6, and deformations in Proposition 5.4. By reflecting the pictures in Figure 5.11 in the line connecting right-bottom corner to left-top corner, we see that \( \beta(I) \) is also deformed to the arc in Figure 5.12, and this gives the conclusion. \( \square \)

6. Orbifold version of Rubinstein-Scharlemann graphic.

In this section, we formulate an orbifold version of the Rubinstein-Scharlemann setting, and show that the local labelling scheme described in \([18]\) holds in this setting.

Let \( M \) be a compact 3-manifold, \( \gamma \) a union of mutually disjoint arcs or simple closed curves properly embedded in \( M \), \( F \) a surface properly embedded in \( M \), which is in general position with respect to \( \gamma \), and \( \ell(\subset F) \) a simple closed curve with \( \ell \cap \gamma = \emptyset \).

**Definition 6.1.** A surface \( D \) is a \( \gamma \)-disk, if \( D \) is a disk intersecting \( \gamma \) in at most one transverse point.

**Definition 6.2.** We say that \( \ell \) is \( \gamma \)-inessential if \( \ell \) bounds a \( \gamma \)-disk in \( F \). We say that \( \ell \) is \( \gamma \)-essential if it is not \( \gamma \)-inessential.
Let \( \ell_1, \ell_2 (\subset F) \) be simple closed curves with \( \ell_i \cap \gamma = \emptyset \) (\( i = 1, 2 \)).

**Definition 6.3.** We say that \( \ell_1 \) and \( \ell_2 \) are \( \gamma \)-parallel if \( \ell_1 \cup \ell_2 \) bounds an annulus \( A \) in \( F \) such that \( A \cap \gamma = \emptyset \).

**Definition 6.4.** We say that \( D \) is a \( \gamma \)-compressing disk for \( F \) if; \( D \) is a \( \gamma \)-disk; and \( D \cap F = \partial D \), and \( \partial D \) is \( \gamma \)-essential in \( F \). The surface \( F \) is \( \gamma \)-compressible if it admits a \( \gamma \)-compressing disk, and it is \( \gamma \)-incompressible if it is not \( \gamma \)-compressible.

Let \( a \) be an arc properly embedded in \( F \) with \( a \cap \gamma = \emptyset \).

**Definition 6.5.** We say that \( a \) is \( \gamma \)-inessential if there is a subarc \( b \) of \( \partial F \) such that \( \partial b = \partial a \), and \( a \cup b \) bounds a disk \( D \) in \( F \) such that \( D \cap \gamma = \emptyset \). We say that \( a \) is weakly \( \gamma \)-inessential if there is a subarc \( b \) of \( \partial F \) such that \( \partial b = \partial a \), and \( a \cup b \) bounds a \( \gamma \)-disk \( D \) in \( F \).

**Definition 6.6.** Let \( F_1, F_2 \) be surfaces embedded in \( M \) such that \( \partial F_1 = \partial F_2 \). We say that \( F_1 \) and \( F_2 \) are \( \gamma \)-parallel, if there is a submanifold \( N \) in \( M \) such that \( (N, F_1 \cap F_2, N \cap \gamma) \) is homeomorphic to \( (F_1 \times I, \partial F_1 \times \{1/2\}, \partial I) \) as a triple, where \( \partial \) is a union of points in \( \text{Int}(F_1) \), and \( F_1 \) (\( F_2 \) resp.) corresponds to the closure of the component of \( \partial(F_1 \times I) - \partial F_1 \times \{1/2\} \) containing \( F_1 \times \{0\} \) (\( F_1 \times \{1\} \) resp.).

The submanifold \( N \) is called a \( \gamma \)-parallelism between \( F_1 \) and \( F_2 \).

We say that \( F \) is \( \gamma \)-boundary parallel if there is a subsurface \( F' \) in \( \partial M \) such that \( F \) and \( F' \) are \( \gamma \)-parallel.

**Definition 6.7.** Let \( F_1, F_2 \) be mutually disjoint surfaces in \( M \) which are in general position with respect to \( \gamma \). We say that \( F_1 \) and \( F_2 \) are \( \gamma \)-isotopic if there is an ambient isotopy \( \phi_t \) (\( 0 \leq t \leq 1 \)) of \( M \) such that; \( \phi_0 = \text{id}_M \); \( \phi_1(F_1) = F_2 \), and; \( \phi_t(\gamma) = \gamma \) for each \( t \).

**Genus \( g \) \( n \)-bridge position.**

Let \( \Gamma = \{ \gamma_1, \ldots, \gamma_n \} \) be a system of mutually disjoint arcs properly embedded in \( M \).

**Definition 6.8.** We say that \( \Gamma \) is trivial if there exists a system of mutually disjoint disks \( \{ D_1, \ldots, D_n \} \) in \( M \) such that (1) \( D_i \cap \Gamma = \partial D_i \cap \gamma_i = \gamma_i \), and (2) \( D_i \cap \partial M \) is an arc, say \( \alpha_i \), such that \( \alpha_i = \text{cl}(\partial D_i - \gamma_i) \).

**Example 6.9.** Let \( \beta \) be a system of trivial two arcs in a 3-ball \( B \). The pair \((B, \beta)\) is often refered as 2-string trivial tangle, or a rational tangle.

Let \( K \) be a link in a closed 3-manifold \( M \). Let \( M = A \cup_{\partial} B \) be a genus \( g \) Heegaard splitting. Then the next definition is borrowed from [6].

**Definition 6.10.** We say that \( K \) is in a \((\text{genus} \ g)\) \( n \)-bridge position (with respect to the Heegaard splitting \( A \cup_{\partial} B \)) if \( K \cap A \) (\( K \cap B \) resp.) is a system of trivial \( n \) arcs in \( A \) (\( B \) resp.).
In this paper, we abbreviate genus 0 \( n \)-bridge position to \( n \)-bridge position.

**Unknotting tunnel.**

Let \( K \) be a knot in a closed 3-manifold \( M \). A tunnel for \( K \) is an embedded arc \( \sigma \) in \( S^3 \) such that \( \sigma \cap K = \partial \sigma \). We say that a tunnel \( \sigma \) for \( K \) is unknotting if \( S^3 - \text{Int} N(K \cup \sigma, S^3) \) is a genus two handlebody.

**Orbifold setting.**

Let \( K \) be a link in a closed 3-manifold \( M \). We regard \( K \) as \( \gamma \) above. Let \( L_1, L_2 \) be a pair of mutually disjoint 1-complexes in \( M \) such that:

1. Each vertex of \( L_i \) has valency zero, one or three,
2. \((L_1 \cup L_2) \cap K\) consists of the union of a (possibly, empty) sublink of \( K \), and a subset of the vertices of \( L_1 \cup L_2 \) with valency one,
3. Let \( N \) be a regular neighborhood of \( L_1 \cup L_2 \), and \( E = \text{cl}(M - N) \).

Then \((E, E \cap K)\) is homeomorphic to \((P \times (0,1), P \times (0,1))\), where \( P \) is a closed surface, and \( P \) is a finite set of (possibly empty) points in \( P \).

Let \( A, B \) be the closures of the components of \( M - (P \times \{1/2\}) \), where \( L_1 \subset A, L_2 \subset B \). We say that \( A \cup B \) is an orbifold Heegaard splitting of \((M,K)\). Then as in Sect. 2, we obtain a sweep-out \( H : P \times I \to M \). Let \( R_1, R_2 \) be another pair of 1-complexes satisfying the above conditions (1), (2), and (3), and \( G : Q \times I \to M \) the corresponding sweep-out. Then, as in Theorem 4.2, we may suppose that we can obtain an excellent map \( f : M - (L_1 \cup L_2 \cup R_1 \cup R_2) \to \mathbb{R}^2 \) from \( H \) and \( G \) such that the graphic obtained from \( H \) and \( G \) is the discriminant set \( f(S(f)) \subset I \times I \). Here we note that we have to slightly generalize the definition of standard position for a neighborhood of a valency one vertex (e.g. Figure 6.1), and it is easy to see the procedures in Sect. 4 work under this situation.

![Figure 6.1](image)

**Example 6.11.** Suppose that \( K \) is in a genus \( g \) \( n \)-bridge position with respect to a Heegaard splitting \( A \cup_P B \). Then, by adding \( n \) edges to each of the appropriate spines of \( A \) and \( B \), we can obtain 1-complexes satisfying the above conditions.
Example 6.12. Let $K$ be a tunnel number one knot with unknotting tunnel $\tau$. Let $L_1 = K \cup \tau$. Let $L_2$ be a spine of the genus two handlebody $c\ell(M - N(K \cup \tau))$ with each vertex having valency three. Then $L_1, L_2$ satisfies the above conditions.

Definition 6.13. Let $H, G$ be sweep-outs as above. We say that $H$ and $G$ are $K$-comparable if $f_{K- (L_1 \cup L_2 \cup R_1 \cup R_2)} : K - (L_1 \cup L_2 \cup R_1 \cup R_2) \to \text{Int}(I \times I)$ is an immersion (possibly) with normal crossing, and $f(K - (L_1 \cup L_2 \cup R_1 \cup R_2))$ and $f(S(f))$ are in general position in $\text{Int}(I \times I)$.

Proposition 6.14. By an arbitrarily small deformation on $K$ rel $(L_1 \cup L_2 \cup R_1 \cup R_2)$ with respect to Whitney topology, we can arrange $H$ and $G$ to be $K$-comparable.

Proof. Note that $f^{-1}(f(S(f)))$ has the structure of a simplicial complex with dimension at most 2 (see Figure 6.3), and, hence, by an arbitrarily small deformation of $K$ with respect to Whitney topology we can arrange so that $K$ and $f^{-1}(f(S(f)))$ are in general position, that is, $K$ and the 1-skeleton are disjoint, and $K$ and $f^{-1}(f(S(f)))$ intersect transversely in a finite number of points. This shows that $f(K)$ and $f(S(f))$ intersect transversely in a finite number of points. Possibly $f(K)$ may contain a crossing vertex of the graphic $f(S(f))$. Then we further apply a small deformation to make $f(K)$ avoid crossing vertices and to make $f_{K- (L_1 \cup L_2 \cup R_1 \cup R_2)}$ an immersion with normal crossing, and this gives the conclusion. \[\square\]
For a $K$-comparable pair $H$ and $G$ we can obtain a graphic $\Gamma$ as in the following manner.

**Regions:** A region is a component of the subset of $\text{Int}(I \times I)$ consisting of values $(s,t)$ such that $P_s$ and $Q_t$ intersect transversely, and $K \cap (P_s \cap Q_t) = \emptyset$.

**Edges:** An edge is a component of the subset consisting of values $(s,t)$ such that either:

1. $P_s$ and $Q_t$ intersect transversely except for one non-degenerate tangent point and $K \cap (P_s \cap Q_t) = \emptyset$, or
2. $P_s$ and $Q_t$ intersect transversely and $K \cap (P_s \cap Q_t)$ consists of one point.

By Definition 6.13, we see that edge is a 1-dimensional subset of $\text{Int}(I \times I)$ which is monotonously increasing or decreasing.

**Crossing vertices:** A crossing vertex is a component of the subset consisting of points $(s,t)$ such that either:

1. $P_s$ and $Q_t$ intersect transversely except for two non-degenerate points of tangency and $K \cap (P_s \cap Q_t) = \emptyset$, or
2. $P_s$ and $Q_t$ intersect transversely except for one non-degenerate tangent point and $K \cap (P_s \cap Q_t)$ consists of one point, or
3. $P_s$ and $Q_t$ intersect transversely and $K \cap (P_s \cap Q_t)$ consists of two points.

Note that in this setting we may (as in Sect. 2) also regard a crossing vertex to be a crossing point of two edges. This follows from the same reason as in Section 2 (Case (1)), or from the condition “$K$ and $f^{-1}(f(S(f)))$ are generic” (Case (2)), or from the condition “$f\lbrack K-(L_1 \cup L_2 \cup R_1 \cup R_2)\rbrack$ is an immersion with normal crossings” (Case (3)).

**Birth-death vertices:** A birth-death vertex is a component of the subset consisting of points $(s,t)$ such that $P_s$ and $Q_t$ intersect transversely except for a single degenerate tangent point and $K \cap (P_s \cap Q_t) = \emptyset$.

**Labelling regions of the graphic.**

Consider a region of the graphic $I \times I - \Gamma$. Then the $K$-isotopy class of $P_s \cap Q_t$ in $P_s$ or $Q_t$ is independent of the choice of $(s,t)$ in each region, and, hence, we often abbreviate $P_s$ by $P$, and $Q_t$ by $Q$.

The purpose of the rest of this section is to claim that the nature of the (local) labelling schemes discussed in [18] holds also in our setting. We assume that the reader is familiar with Sect. 4, 5 of [18].

**Definition 6.15.** We say that an orbifold Heegaard splitting $A \cup_P B$ is weakly $K$-reducible if there exist $K$-compressing disks $D_A, D_B$ for $P$ in $A, B$ respectively such that $\partial D_A \cap \partial D_B = \emptyset$. The orbifold Heegaard splitting $A \cup_P B$ is strongly $K$-irreducible if it is not weakly $K$-reducible.
Definition 6.16. Let \((s, t)\) be a point in a region of \(I \times I - \Gamma\). (Hence, \(P \cap Q\) consists of a system of simple closed curves in \(M\) disjoint from \(K\).) Let \(C^K_P\) (\(C^K_Q\) resp.) be the subset of the simple closed curves which are \(K\)-essential in \(P\) (\(Q\) resp.). Then the subset \(C^K_A\) of \(C^K_P\) is defined by:
\[
C^K_A = \{ c \mid c \text{ bounds a } K \text{-disk } D \text{ in } Q - C^K_P \text{ such that } N(\partial D, D) \subset A \},
\]
where \(N(\partial D, D)\) is a regular neighborhood of \(\partial D\) in \(D\).

Analogously \(C^K_B\) (\(\subset C^K_P\)), and \(C^K_X, C^K_Y\) (\(\subset C^K_Q\)) are defined.

Lemma 6.17 (Lemma 4.3 of [18]). If \(c \in C^K_A\), then \(c\) bounds a \(K\)-disk in \(A\).

Proof. The proof is basically the same as Rubinstein-Scharlemann’s except for the consideration on \(K\). That is:

Let \(D\) be the \(K\)-disk which \(c\) bounds in \(Q\), such that \(N(\partial D, D) \subset A\). If \(\text{Int} D \cap P = \emptyset\), then \(D\) gives a desired \(K\)-disk. Suppose that \(\text{Int} D \cap P \neq \emptyset\). Let \(\Delta(\subset D)\) be an innermost disk. Since \(\text{Int} D \cap C^K_P = \emptyset\), we see that \(\partial \Delta\) bounds a \(K\)-disk \(\Delta'\) in \(P\). For a proof of the next claim, see Appendix A-3.

Claim. \(\Delta' \cap K = \emptyset\) if and only if \(\Delta \cap K = \emptyset\). Furthermore, if \(\Delta' \cap K \neq \emptyset\), then \(\Delta\) and \(\Delta'\) are \(K\)-parallel, i.e., \(\Delta \cup \Delta'\) bounds a 3-ball \(D^3\) such that \(D^3 \cap K\) is an unknotted arc.

By the claim, we see that we can apply cut and paste on \(D\) using \(\Delta\) and \(\Delta'\) to get a new disk \(D'\) with fewer intersections. By applying the argument finitely many times, we obtain the desired disk. \(\square\)

As an immediate consequence of Lemma 6.17, we have:

Corollary 6.18 (Corollary 4.4 of [18]). If there exists a region such that both \(C^K_A\) and \(C^K_B\) are non-empty, then \(A \cup P B\) is weakly \(K\)-reducible.

In the rest of this section, we suppose:

\(M\) admits a 2-fold branched covering space \(p : \tilde{M} \to M\) along \(K\).

Lemma 6.19 (Lemma 4.5 of [18]). Suppose that \(C^K_P = \emptyset, C^K_Q = \emptyset\), and there exists a \(\partial\)-reducing \(K\)-disk in \(A\) which intersects \(Q\) only in \(K\)-inessential simple closed curves. Suppose, moreover, that \(A\) contains a \(K\)-essential curve of \(Q\). Then either \(A \cup P B\) is weakly \(K\)-reducible, or \(M\) is the 3-sphere \(S^3\) and \(K\) is a trivial knot.

Proof. By Appendix A-3, we may suppose, by \(K\)-isotopy, that \(P\) and \(Q\) are disjoint, and that the \(\partial\)-reducing \(K\)-disk \(D\) and \(Q\) are disjoint. Without loss of generality, we may suppose that \(Y\) is contained in \(A\). Now consider the 2-fold branched covers (along \(K\)) \(\tilde{A}, \tilde{B}, \tilde{P}, \tilde{X}, \tilde{Y}, \tilde{D}\) of \(A, B, P, X, Y, D\) respectively. Note that, by the definition of an orbifold Heegaard splitting, \(\tilde{A}, \tilde{B}\) are handlebodies.
Take a maximal compression body $\tilde{C}$ of $\tilde{A} - \text{Int}\tilde{Y}$ for $\partial\tilde{A}$. By the uniqueness of maximal compression body, we may suppose, by applying $\mathbb{Z}_2$-equivariant cut and paste arguments as in the Proof of 10.3 of [9] or [12], that $\tilde{C}$ is invariant under the covering translation $\tau$. Let $P'$ be a component of the inner boundary of $\tilde{C}$.

If $P'$ is a sphere, then $M$ is $S^3$ (see Proof of [18, Lemma 4.5]), and, by $\mathbb{Z}_2$-Smith Conjecture ([21], or [16]), $K$ is a trivial knot in $S^3$.

Suppose that $P'$ is not a sphere. Note that $P'$ is compressible in $\tilde{B} \cup \tilde{C}$ since $P'$ is contained in a handlebody $\tilde{X}$ (see the proof of Lemma 4.5 of [18]). Then note that $\tilde{B} \cup \tilde{P} \tilde{C}$ is a Heegaard splitting in the sense of Casson-Gordon [4]. Hence, by [4], there exists a compressing disk $D' (\subset \tilde{B} \cup \tilde{C})$ for $P'$ such that $D' \cap \tilde{P}$ consists of a circle (hence, $D' \cap \tilde{C}$ is an annulus).

Now we show that we can have such $D'$ which moreover is equivariant with respect to $\tau$. We may suppose, by general position argument, that $D'$ and $\tau(D')$ intersect transversely. Then, by isotopy, we may suppose that each component of $\tilde{C} \cap (D' \cap \tau(D'))$ is an essential arc in the annulus $D' \cap \tilde{C}$, (and $\tau(D') \cap \tilde{C}$).

Then apply cut and paste arguments on $D_B \cap \tilde{C}$ to obtain a compressing disk $\tilde{D}_B$. Since each component of $\tilde{C} \cap (D' \cap \tau(D'))$ is an essential arc of $D' \cap \tilde{C}$, we see that each component of $D_B$ intersects $\tilde{C}$ in an annulus, (hence, intersects $\tilde{P}$ in a circle).

Then apply cut and paste arguments on $D_B \cap \tilde{C}$ and $\tilde{D}$ to obtain a compressing disk $\tilde{D}' (\subset \tilde{C})$ for $\tilde{P}$ such that $\tilde{D}' \cap (D_B \cap \tilde{C}) = \emptyset$. Then, by applying equivariant cut and paste arguments on $\tilde{D}'$, and $\tau(\tilde{D}')$, we obtain equivariant disk(s) $D_A$ for $\tilde{P}$ such that $D_A \cap \tilde{D}_B = \emptyset$. Then $p(D_A)$, and $p(D_B \cap \tilde{B})$ give weak $K$-reducibility of $K$. \qed

**Labelling scheme.**

Now we mimic the procedures in [18, Section 5]. If $C_A^K$ ($C_B^K$, $C_X^K$, $C_Y^K$ resp.) is non-empty, then we label the region $A$ ($B$, $X$, $Y$ resp.). If $C_B^K$ and $C_Q^K$ are both empty and $A$ ($B$ resp.) contains a $K$-essential curve of $Q$, then we label the region $b$ ($a$ resp.), and if $X$ ($Y$ resp.) contains an essential curve of $P$, then we label the region $y$ ($x$ resp.). By Corollary 6.18 we have:

**Rule 1.** If there exists a region with both labels $A$ and $B$ assigned, then $A \cup_P B$ is weakly $K$-reducible.

We obviously have:

**Rule 2.** No region can have both an upper case label and lower case label.

Next, we consider how labels change as one cross an edge of $\Gamma$.

Note that we have the following three possibilities.

1) The edge comes from center tangency.
In this case, the regions have exactly the same label.

2) The edge comes from saddle tangency.

In this case, the effect is banding two components of $P \cap Q$, say $c_0$ and $c_1$, to make a simple closed curve, say $c$, or vice versa.

3) The edge comes from $P \cap Q \cap K$.

In this case the effect is that a component of $P \cap Q$ passes a puncture by $K$ on $P$, (and $Q$).

Note that situation 3) did not appear in Rubinstein-Scharlemann setting. With this fact in mind, it is easy, by tracing the proof of [18, Corollary 5.1], to see:

**Rule 3 ([18, Corollary 5.1]).** If both labels $A$ and $B$ appear in two adjacent regions, then $A \cup_P B$ is weakly $K$-reducible.

Then we have:

**Rule 4 ([18, Corollary 5.2]).** In adjacent regions of $I \times I - \Gamma$, labels $a$ and $b$ ($x$ and $y$ resp.) cannot appear.

*Proof.* Suppose that $a$ and $b$ occur opposite sides of an edge. Then arguments in the proof of [18, Corollary 5.2] show that edge does not come from saddle tangency. Then it is easy to see that this phenomena can occur only in case when $Q$ is a 2-sphere and $Q \cap K$ consists of three points $a_1, a_2, a_3$ and a component of the intersection $P \cap Q$ is changed from a circle separating $a_1$ and $a_2 \cup a_3$ to a circle separating $a_1 \cup a_2$ and $a_3$. However this is impossible, since $Q \cap K$ must consists of even number of points.

With tracing the proof of [18, Lemma 5.3] with consideration on $K$ we easily have:

**Lemma 6.20 ([18, Lemma 5.3]).** Suppose, in $I \times I - \Gamma$, a region labelled $A$ is adjacent to a region labelled with a lower case letter. Then the edge represents either (1) a saddle tangency in which a band which is $K$-essential in $P$ and weakly $K$-inessential in $Q$ is attached to an intersection curve which is $K$-inessential in both $P$ and $Q$, or (2) a passing of $K$ which changes an element of $\mathcal{C}_K^A$ bounding a disk (in $P$) with two punctures by $K$ into a disk with one puncture by $K$.

Then we have:

**Rule 5 ([18, Corollary 5.4]).** Suppose, in $I \times I - \Gamma$, a region labelled $A$ is adjacent to a region labelled $b$. Then either $A \cup_P B$ is weakly $K$-reducible, or $M \cong S^3$ and $K$ is a trivial knot.

*Proof.* We see, by Lemma 6.20, that $A \cup_P B$ satisfies the assumption of Lemma 6.19, and this gives the conclusion.
In the following, the notation \( a \) stands for, as in [18], either \( a \) or \( A \), and similar for \( b \), \( x \), and \( y \). With the above rules, we see that the arguments in the proof of [18, Lemma 5.7] (it is easy to check that [18, Lemma 5.6] holds in our setting since the new phenomenon is the situation 3) in the preceding Rule 3) completely works in our setting to give:

Rule 6 ([18, Lemma 5.7]). If all letters \( A \), \( B \), \( X \), and \( Y \) appear in quadrants of a crossing vertex of \( \Gamma \), then either two opposite quadrants are unlabelled, or one of \( A \cup B \), \( X \cup Y \) is weakly \( K \)-reducible, or \( M \cong S^3 \) and \( K \) is a trivial knot.

By using these rules, the arguments in the proof of [18, Proposition 5.9] show (the difference here is the consideration on \( K \), a possibility that three edges may be joined to a vertex in \( \partial(I \times I) \) (see Figure 6.1)):

Proposition 6.21. Let \( A \cup B \), \( X \cup Y \) be orbifold Heegaard splittings for \( (M,K) \) obtained from two bridge positions as in Example 6.11. Suppose that \( A \cup B \), \( X \cup Y \) are strongly \( K \)-irreducible, and \( K \) is not a trivial knot in \( S^3 \). Then there is an unlabelled region in \( I \times I - \Gamma \).

And, this together with Appendix A-3, and the arguments in the proof of [18, Corollary 6.2] shows:

Corollary 6.22. Let \( A \cup B \), \( X \cup Y \) be as in Proposition 6.21. Then, by applying \( K \)-isotopy, we may suppose that \( P \) and \( Q \) intersect non-empty collection of simple closed curves which are \( K \)-essential in both \( P \) and \( Q \).

7. 2-bridge position of a 2-bridge knot.

Let \( K \) be a non-trivial 2-bridge knot (that is, \( K \) is a non-trivial knot which admits a genus 0 2-bridge position). In this section, we show that the 2-bridge positions of \( K \) are unique up to \( K \)-isotopy, which was originally proved by Schubert [20].

Theorem 7.1. Let \( K \) be a non-trivial 2-bridge knot, and \( P \), \( Q \) are 2-spheres in \( S^3 \) which give 2-bridge positions of \( K \). Then \( P \) is \( K \)-isotopic to \( Q \), i.e., there is an ambient isotopy \( \varphi_t \) \( (0 \leq t \leq 1) \) of \( S^3 \) such that (1) \( \varphi_t(K) = K \) \( (0 \leq t \leq 1) \), (2) \( \varphi_0 = \text{id}_{S^3} \), and (3) \( \varphi_1(P) = Q \).

For the proof of Theorem 7.1, we prepare some lemmas, proofs of which are given in Appendix B. (For the definition of \( \beta \)-essential surface, see Definition 6.2.)

Lemma 7.2 (Appendix B-1). Let \((B,\beta)\) be a 2-string trivial tangle. Let \( F \) be a surface properly embedded in \( B \). Suppose that \( F \) is \( \beta \)-essential. Then \( F \) is a disk which is disjoint from \( \beta \), and \( F \) separates the components of \( \beta \).
Recall that it is often said that \((C, \gamma)\) is a rational tangle if \((C, \gamma)\) is homeomorphic to the 2-string trivial tangle \((B, \beta)\) as a pair.

**Lemma 7.3** (Appendix B-2). Let \((B, \beta)\) be a 2-string trivial tangle, and \(F\) a \(\beta\)-incompressible surface in \(B\).

Then either (0) \(F\) is \(\beta\)-essential (see Lemma 7.2), (1) \(F\) is a \(\beta\)-boundary parallel disk intersecting \(\beta\) in at most one point, (2) \(F\) is a \(\beta\)-boundary parallel disk intersecting \(\beta\) in two points (and, hence, \(F\) separates \((B, \beta)\) into the parallelism and a rational tangle), or (3) \(F\) is \(\beta\) boundary parallel annulus such that \(F \cap \beta = \emptyset\).

**Lemma 7.4** (Appendix B-3). Let \(D\) be a \(\beta\)-compressible disk in \(B\) such that \(\partial D\) is \(\beta\)-essential in \(\partial B\), and \(D \cap \beta\) consists of two points. Then \(D\) separates \((B, \beta)\) into two tangles \((B_1, \beta_1)\), and \((B_2, \beta_2)\), where \((B_1, \beta_1)\) is a rational tangle such that there is a \(\beta\)-essential disk \(D'\) in \((B_1, \beta_1)\) with \(D \cap D' = \emptyset\). Moreover if \((B_2, \beta_2)\) happens to be a rational tangle, then \((B_2, \beta_2)\) is a \(\beta\)-boundary parallelism for \(D\).
Let $A, B$ ($X, Y$ resp.) be the closures of the components of $S^3 - P$ ($S^3 - Q$ resp.).

**Proposition 7.5.** Every genus 0 Heegaard splitting of $S^3$ which gives a 2-bridge position of $K$ is strongly $K$-irreducible.

**Proof.** We give the proof for $A \cup P B$. Assume that $A \cup P B$ is weakly $K$-reducible, and let $D_A, D_B$ be a pair of $K$-essential disks in $A, B$ respectively such that $\partial D_A \cap \partial D_B = \emptyset$. Since $P \cap K$ consists of four points, we see that $\partial D_A$ and $\partial D_B$ are parallel in $P - K$, and this together with Lemma 7.2 implies that $K$ is a 2-component trivial link, a contradiction. \qed

**Proof of Theorem 7.1.** Note that, by Proposition 7.5, $A \cup P B, X \cup Q Y$ are strongly $K$-irreducible. Hence, by Corollary 6.22, we may suppose that $P$ and $Q$ intersect non-empty collection of simple closed curves which are $K$-essential in both $P$ and $Q$.

The proof is carried out by the induction on the number of the components of $P \cap Q$. The following Claims 1 and 2 give the first step of the induction.

**Claim 1.** If $P \cap Q$ consists of one component, then $P$ and $Q$ are $K$-isotopic.

**Proof.** Let $D_A = Q \cap A, D_B = Q \cap B, D_X = P \cap X, and D_Y = P \cap Y$. Then $D_A, D_B, D_X, D_Y$ are disks each of which intersects $K$ in two points. Then we have the following cases.

**Case 1.** The disks $D_A, D_B, D_X, D_Y$ are $K$-incompressible in $A, B, X, Y$ respectively.

In this case, by Lemma 7.3 (2), we have:

1. “$D_A$ and $D_X$ are $K$-parallel in $A$” or “$D_A$ and $D_Y$ are $K$-parallel in $A$”,
2. “$D_B$ and $D_X$ are $K$-parallel in $B$” or “$D_B$ and $D_Y$ are $K$-parallel in $B$”,
3. “$D_X$ and $D_A$ are $K$-parallel in $X$” or “$D_X$ and $D_B$ are $K$-parallel in $X$”, and
4. “$D_Y$ and $D_A$ are $K$-parallel in $Y$” or “$D_Y$ and $D_B$ are $K$-parallel in $Y$”.

It is easy to see that the above 4 conditions imply either one of:

1. “$D_A$ and $D_X$ are $K$-parallel in $A$ (and, $X$), and $D_B$ and $D_Y$ are $K$-parallel in $B$ (and, $Y$)”, or
2. “$D_A$ and $D_Y$ are $K$-parallel in $A$ (and, $Y$), and $D_B$ and $D_X$ are $K$-parallel in $B$ (and, $X$)”.

Since the argument is symmetric, we may suppose that (1) holds. Then, by using the parallelisms, we can move $D_A$ to $D_X$, and $D_B$ to $D_Y$ to give a desired $K$-isotopy.

**Case 2.** One of the disks $D_A, D_B, D_X, or D_Y$ is $K$-compressible in $A, B, X, or Y$. 


Without loss of generality, we may suppose that $D_Y$ is $K$-compressible in $Y$, and the $K$-compressing disk is contained in $B$. Note that this implies that $D_B$ is $K$-compressible in $B$. Note moreover that $D_X$ is $K$-incompressible in $X$, since $D_X$ separates the boundary components of each component of $K \cap X$ in $X$, and similarly $D_A$ is $K$-incompressible in $A$. Then, by Lemma 7.3 (2) and the last half of Lemma 7.4, we see that $D_B$ and $D_X$ are $K$-parallel in $B$ (and, $X$). Similarly we can show that $D_A$ and $D_Y$ are $K$-parallel in $A$ (and, $Y$). Hence we can obtain a desired $K$-isotopy as in Case 1.

This completes the proof of Claim 1.

Claim 2. If $P \cap Q$ consists of two components, then $P$ and $Q$ are $K$-isotopic.

Proof. Let $D_1$, $A_0$, $D_2$ be the closures of the components of $P - (P \cap Q)$ such that $D_1$ and $D_2$ are disks, and $A_0$ is an annulus. Without loss of generality, we may suppose that $D_1 \cup D_2$ is contained in $X$, and $A_0$ is contained in $Y$.

Subclaim. Either $D_1$ or $D_2$ is $K$-boundary parallel in $X$.

Proof. If $D_1$ or $D_2$ is $K$-incompressible, then this immediately follows from Lemma 7.3 (2). Suppose that $D_1$ and $D_2$ are $K$-compressible. Let $B_3^1$ be the closure of the component of $X - D_1$ which corresponds to $B_2$ in Lemma 7.4. By exchanging suffix, if necessary, we may suppose that $\text{Int}(B_3^1) \cap P = \emptyset$. Without loss of generality, we may suppose that $B_3^1$ is contained in $A$. Let $D^1 = B_3^1 \cap Q$. Since $D^1$ separates the boundary components of each component of $K \cap A$ in $A$, we see that $D^1$ is $K$-essential in $A$. Hence, by Lemma 7.3 (2) and the last half of Lemma 7.4, we see that $B_3^1$ is a $K$-parallelism.

Let $B_3^2$ be the parallelism between $D_1$ and $\partial X$ obtained in Subclaim. Then we can push $D_1$ out of $X$ along the parallelism, and we have the conclusion by Claim 1.

This completes the proof of Claim 2.

Completion of Proof. Suppose that $\sharp (P \cap Q) \geq 3$. Note that the components of $P \cap Q$ are mutually $K$-parallel in $P$. Let $D_1$, $A_1$, $\ldots$, $A_m$, $D_2$ be the closures of the components of $P - (P \cap Q)$ such that $D_1$, $D_2$ are disks and $A_1, \ldots, A_m$ are annuli that are located on $P$ in this order. Suppose that $D_1$ or $D_2$ is $K$-boundary parallel in $X$ or $Y$. Then, by using the arguments in the proof of Claim 2, we can reduce $\sharp (P \cap Q)$, to give the conclusion. Suppose that $D_1$ and $D_2$ are not $K$-boundary parallel in $X$ and $Y$. By Lemma 7.3 (2), this implies that $D_1$ and $D_2$ are $K$-compressible.

Claim 3. Both $D_1$ and $D_2$ are contained in the closure of a component of $S^3 - Q$, say $X$. And each component of $P \cap Y$ is a $K$-incompressible annulus.
Proof. Assume that $D_1$ is contained in $X$, and $D_2$ is contained in $Y$. By applying $K$-compression on $D_1$ ($D_2$ resp.) we obtain a $K$-essential disk $E_1$ ($E_2$ resp.) in $X$ ($Y$ resp.) such that $\partial E_i = \partial D_i$. Note that $E_1$ ($E_2$ resp.) separates the components of $K \cap X$ ($K \cap Y$ resp.). This implies that $K$ is a 2-component trivial link, a contradiction. Hence we may suppose that $D_1$ and $D_2$ are contained in $X$. Let $A_i$ be a component of $P \cap Y$. Assume that $A_i$ is $K$-compressible in $Y$. Then, by applying $K$-compression on $A_i$, we obtain two $K$-essential disks in $Y$. Then, by the above argument, we see that $K$ is a 2-component trivial link, a contradiction.

Claim 3 together with Lemma 7.3 (3) implies that each component of $P \cap Y$ is a $K$-boundary parallel annulus in $Y$. Take an outermost one of $P \cap Y$, say $A_j$, and push $A_j$ out of $Y$ along the parallelism. This reduces $\sharp(P \cap Q)$ by two, and we have the conclusion by the assumption of the induction.

This completes the proof of Theorem 7.1. \( \square \)

8. Genus one 1-bridge position of a 2-bridge knot.

For a 2-bridge knot $K$ we can obtain four genus one 1-bridge positions of $K$ as follows.

Let $A \cup_P B$ be the Heegaard splitting which gives the 2-bridge position. Then let $a_1, a_2, b_1, b_2$ be the closures of the components of $K - P$, where $a_1 \cup a_2$ ($b_1 \cup b_2$ resp.) is contained in $A$ ($B$ resp.). Let $T_1 = A \cup N(b_1, B)$, $\alpha_1 = a_1 \cup b_1 \cup a_2$, $T_2 = c(\overline{B - N(b_1, B)})$, and $\alpha_2 = b_2$. Then each $T_i$ is a solid torus and $\alpha_i$ is a trivial arc in $T_i$, and, hence, $T_1 \cup T_2$ gives genus one 1-bridge position of $K$. Moreover, by using $a_1, a_2, b_2$ for $b_1$, we can obtain other three genus one 1-bridge positions of $K$.

![Figure 8.1.](image)

Remark 8.1. In [17], Morimoto-Sakuma study the isotopy classes of the 1-bridge positions above.
We say that these genus one 1-bridge positions are *standard*.

The purpose of this section is to prove:

**Theorem 8.2.** Every genus one 1-bridge positions of a non-trivial 2-bridge knot is standard.

First, we prepare some lemmas for the proof of Theorem 8.2, proofs of which are given in Appendix C. Let \( \alpha \) be a trivial arc in a solid torus \( T \). For \((T, \alpha)\), we have:

**Lemma 8.3** (Appendix C-1). Let \( D \) be an \( \alpha \)-compressing disk for \( \partial T \). Then \( D \) is either:

1. a meridian disk of \( T \) with \( D \cap \alpha = \emptyset \). In this case, we obtain, by cutting \((T, \alpha)\) along \( D \), a 1-string trivial tangle,
2. a meridian disk of \( T \) with \( D \cap \alpha \) consists of one point, and we obtain, by cutting \((T, \alpha)\) along \( D \), a 2-string trivial tangle, or
3. \( \partial \)-parallel disk in \( T \) with \( D \cap \alpha = \emptyset \). In this case, \( D \) cobounds a 1-string trivial tangle in \((T, \alpha)\).

![Figure 8.2](image)

**Lemma 8.4** (Appendix C-2). Let \( D \) be an \( \alpha \)-essential disk in \( T \) such that \( D \cap \alpha \) consists of two points. Then there exists an \( \alpha \)-compressing disk \( D' \) disjoint from \( D \) such that \( D' \cap \alpha \) consists of one point. Moreover, by cutting \((T, \alpha)\) along \( D' \), we obtain 2-string trivial tangle \((B, \beta)\) such that \( D \) is a \( \beta \)-incompressible disk in \((B, \beta)\) (hence \( D \) is \( \beta \)-boundary parallel).

We note that the disk \( D \) in Lemma 8.4 is either separating or non-separating in \( T \).

![Figure 8.3](image)

**Lemma 8.5** (Appendix C-3). Let \( D_1, D_2 \) be mutually disjoint non \( \alpha \)-parallel, \( \alpha \)-essential disks such that \( D_i \cap \alpha \) \((i = 1, 2)\) consists of two points. Then there exists an \( \alpha \)-compressing disk \( D' \) for \( \partial T \) disjoint from \( D_1 \cup D_2 \) such
that $D' \cap \alpha$ consists of one point. Moreover each $D_i$ is non-separating in $T$, and, by cutting $(T, \alpha)$ along $D'$, we obtain 2-string trivial tangle $(B, \beta)$, and $D_1, D_2$ are mutually non $\beta$-parallel, $\beta$-boundary parallel, $\beta$-incompressible disks in $(B, \beta)$.

**Figure 8.4.**

**Lemma 8.6 (Appendix C-4).** Let $D$ be an $\alpha$-compressible disk such that $\partial D$ is $\alpha$-essential in $\partial T$, and $D \cap \alpha$ consists of two points. Then there is a disk $\Delta$ in $T$ such that $\Delta \cap D = \partial \Delta \cap D = \gamma$ an arc, and $\Delta \cap \alpha = c\ell(\partial \Delta - \gamma)$. Particularly, if $D$ is separating in $T$, then $D$ separates $(T, \alpha)$ into $(T', \alpha')$, and $(B', \alpha'')$ such that $\alpha'$ is a trivial arc in a solid torus $T'$. In this case, if $(B', \alpha'')$ happens to be a rational tangle, then $(B', \alpha'')$ is an $\alpha$-boundary parallelism.

**Figure 8.5.**

In the rest of this section, we give a proof of Theorem 8.2. Let $K$ be a non-trivial 2-bridge knot. Let $A \cup_P B$ be a genus 0 Heegaard splitting of $S^3$ which gives a 2-bridge position of $K$, and $X \cup_Q Y$ a genus one Heegaard splitting which gives a genus one 1-bridge position of $K$. Note that $A \cup B$, $X \cup Y$ give orbifold Heegaard splittings of $(S^3, K)$ (see Example 6.11).

**Proposition 8.7.** Exactly one of the following (1) or (2) holds.

1. $X \cup_Q Y$ gives a standard genus one 1-bridge position.
2. $X \cup_Q Y$ is strongly $K$-irreducible.

**Proof.** Suppose that $X \cup_Q Y$ is weakly $K$-reducible. Let $D_X, D_Y$ be a pair of $K$-essential disks in $X, Y$ respectively such that $\partial D_X \cap \partial D_Y = \emptyset$. Since $H_1(S^3) = \{0\}$, we see that either $D_X$ is separating in $X$ or $D_Y$ is separating in $Y$. Without loss of generality, we may suppose that $D_X$ is separating in $X$. 


Claim 1. The disk $D_Y$ is non-separating in $Y$.

Proof. Assume that both $D_X$ and $D_Y$ are separating in $X$ and $Y$ respectively. By Lemma 8.3 (3), we see that the closure of a component of $X - D_X$ (resp.), say $B_X^3$, is a 3-ball such that $K \cap B_X^3$ is a trivial arc. Since $\partial D_X \cap \partial D_Y = \emptyset$, we see that $\partial D_X$ and $\partial D_Y$ are $K$-parallel in $Q$. This implies that $B_X^3 \cup B_Y^3$ is a 3-ball, and $B_X^3 \cap B_Y^3$ is a disk intersecting $K$ in two points. This shows that $K$ is a connected sum of trivial knots, and hence, $K$ is a trivial knot, a contradiction. □

Claim 2. The disk $D_Y$ intersects $K$ in one point.

Proof. Assume this does not hold, i.e., $D_Y \cap K = \emptyset$. Let $N(D_Y)$ be a regular neighborhood of $D_Y$ in $Y$. Let $X' = X \cup N(D_Y)$, and $Y' = \text{cl}(Y - N(D_Y))$. Then, by Lemma 8.3 (3), we see that $X'$ is a 3-ball such that $K \cap X'$ is a trivial arc. Moreover, by Lemma 8.3 (1), we see that $Y'$ is a 3-ball such that $K \cap Y'$ is a trivial arc. Hence we see that $K$ is a trivial knot, a contradiction. □

Let $N'(D_Y)$ be a regular neighborhood of $D_Y$ in $Y$. Let $X'' = X \cup N'(D_Y)$, and $Y'' = \text{cl}(Y - N'(D_Y))$. Then, by Lemma 8.3 (3), we see that $X''$ is a 3-ball such that $K \cap X''$ is a system of 2-string trivial arcs. Moreover, by Lemma 8.3 (2), we see that $Y''$ is a 3-ball such that $K \cap Y''$ is a system of 2-string trivial arcs. Hence $X'' \cup Y''$ gives a 2-bridge position of $K$, and the genus one 1-bridge position $X \cup Y$ is obtained from $X'' \cup Y''$ in a standard manner.

Conversely, suppose that $X \cup_Q Y$ gives a standard genus one 1-bridge position. Then it is clear there exist disks corresponding to $D_X$, $D_Y$ above in $X$, $Y$, and these disks give a weak $K$-irreducibility of $X \cup_Q Y$. This together with the above shows that $X \cup_Q Y$ gives a standard genus one 1-bridge position if and only if $X \cup_Q Y$ is weakly $K$-irreducible.

This completes the proof of Proposition 8.7. □

Then we prove:

Proposition 8.8. Suppose that $P \cap Q$ consists of non-empty collection of simple closed curves which are $K$-essential in both $P$ and $Q$. Then the genus one 1-bridge position $X \cup_Q Y$ of $K$ is obtained from $A \cup_P B$ in a standard manner.

For the proof of Proposition 8.8, we prepare the following lemma, the proof of which is left to the reader.

Lemma 8.9. Let $T$ be a solid torus, and $A$ an annulus properly embedded in $T$ such that each component of $\partial A$ is a longitude of $T$. Then there is a homeomorphism $h : (\text{annulus}) \times I \to T$ such that $h((\text{annulus}) \times \{1/2\}) = A$. 

Proof of Proposition 8.8. The proof is carried out by the induction of the number of the components $P \cap Q$. The following Claims 1 and 2 give the first step of the induction.

Claim 1. If $P \cap Q$ consists of one component, then the genus one 1-bridge position $X \cup_Q Y$ of $K$ is obtained from $A \cup_P B$ in a standard manner.

Proof. Let $D_X = P \cap X$, and $D_Y = P \cap Y$. Then $D_X$, $D_Y$ are disks properly embedded in $X$, $Y$ respectively, each of which intersect $K$ in two points such that $\partial D_X = \partial D_Y$. Since $P$ is separating in $S^3$, $D_X$ is separating in $X$, and this shows that $\partial D_X$ is contractible in $Q$. Let $E$ be the disk in $Q$ bounded by $\partial D_X (= \partial D_Y)$. Then $E$ intersects $K$ in two points. Without loss of generality, we may suppose that $E$ is contained in $A$. Since $E$ separates the boundary components of each component of $K \cap A$, we see that $E$ is $K$-incompressible in $A$. Then, by Lemma 7.3(2) we have either one of the following two cases.

Case 1. $E$ is not $K$-parallel to one of $D_X$ or $D_Y$ in $A$.

In this case, we may suppose without loss of generality that $E$ is not $K$-parallel to $D_X$. Then, by Lemma 7.3 (2), we see that $E$ is $K$-parallel to $D_Y$ in $Y$, and, by Lemma 8.4, we see that there is a $K$-compressing disk $D$ for $\partial X$ in $X$ such that $D$ intersects $K$ in one point and $D \cap D_X = \emptyset$, and these imply that the genus one 1-bridge position $X \cup_Q Y$ of $K$ is $K$-isotopic to a genus one 1-bridge position which is obtained from $A \cup_P B$ in a standard manner by using the arc $K \cap (B \cap X)$.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure86.png}
\caption{Figure 8.6.}
\end{figure}

Case 2. $E$ is $K$-parallel to $D_X$ and $D_Y$ in $A$.

In this case, we consider the genus one surface $F = Q \cap B$. Note that $\partial F$ is a $K$-essential simple closed curve in $P$. Then, by Lemma 7.3, we see that there is a $K$-compressing disk $D$ for $F$ in $B$. Without loss of generality, we may suppose that $D$ is contained in $X$. Then we have the following two cases.

Case 2.1. $D \cap K = \emptyset$.

In this case, we obtain a $K$-compressing disk $D'$ for $\partial B$ in $B$ by compressing $F$ along $D$. By applying a slight isotopy, we may regard $D'$ as a
A $K$-compressing disk for $\partial X$ in $X$ such that $\partial D' = \partial E$. Then, by Lemma 8.6, we see that the arc $K \cap X$ is pushed into by an isotopy rel $\partial$ in $X$ to an arc contained in $E$. Then we further push the arc along the parallelism between $E$ and $D_Y$ to an arc contained in the disk $D_Y$. We denote by $K'$ the image of $K$ under this isotopy. Then $K'$ is contained in $B$ and $K' \cap \partial B$ is an arc. Since this isotopy does not move the component of $K \cap B$ contained in $Y$ ($= c\ell(K' \cap \text{Int}B)$), $c\ell(K' \cap \text{Int}B)$ is a trivial arc in $B$. This implies that $K$ is a trivial knot, a contradiction.

![Figure 8.7](image.png)

**Figure 8.7.**

**Case 2.2.** $D \cap K$ consists of one point.

In this case, we obtain, by applying $K$-compression on $F$ along $D$ and slight isotopy, a disk $D''$ in $B$ with $\partial D'' = \partial E$, and $D''$ intersects $K$ in two points.

**Subclaim.** $D'' \cup E$ bounds a $K$-parallelism between $D''$ and $E$ in $X$.

**Proof.** Assume not. Suppose that $D''$ is $K$-compressible in $X$. Then, by applying $K$-compression and slight isotopy, we obtain a $K$-compressing disk $E'$ in $B$ such that $\partial E' = \partial E$. Then, by the argument of Case 2.1, we see that $K$ is a trivial knot, a contradiction. Suppose that $D''$ is $K$-incompressible in $X$. By the assumption, we see that $E' \cup D_X$ bounds a rational tangle in $X$ which is not a $K$-parallelism between $E'$ and $D_X$. Then, by Lemma 7.3 (2), we see that $E'$ and $D_Y$ must bound a $K$-parallelism in $B$. But this is impossible, since $D_Y$ contains the boundary components of a component of $K \cap B$. □

By the subclaim together with the arguments in Case 1, we see that the genus one 1-bridge position $X \cup_Q Y$ of $K$ is $K$-isotopic to a genus one 1-bridge position which is obtained from $A \cup_P B$ in a standard manner by using the arc $K \cap (B \cap X)$. 
Claim 2. If \( P \cap Q \) consists of two components, then the genus one 1-bridge position \( X \cup Q Y \) of \( K \) is obtained from \( A \cup P B \) in a standard manner.

Proof. Let \( D_1, A_0, D_2 \) be the closures of the components of \( P - (P \cap Q) \) such that \( D_1 \) and \( D_2 \) are disks, and \( A_0 \) is an annulus. Without loss of generality, we may suppose that \( D_1 \cup D_2 \) is contained in \( X \), and \( A_0 \) is contained in \( Y \). Since \( P \) is separating in \( S^3 \), we see that both \( D_1 \) and \( D_2 \) are either separating or non-separating in \( X \).

Case 1. Both \( D_1 \) and \( D_2 \) are separating in \( X \).

Let \( E \) be the disk in \( Q \) bounded by \( \partial D_1 \). By changing subscript, if necessary, we may suppose that \( \partial D_2 \cap E = \emptyset \), i.e., \( D_1 \) is “outer” than \( D_2 \). Without loss of generality, we may suppose that \( E \) is contained in \( A \). Then we have the following cases.

Case 1.1. \( D_1 \) and \( E \) are \( K \)-parallel in \( X \).

In this case, we can push \( D_1 \) along the parallelism out of \( X \) to make \( P \cap X = D_2 \). Hence, we have the conclusion by Claim 1.

Case 1.2. \( D_1 \) and \( E \) are not \( K \)-parallel in \( X \).

In this case, we first claim that \( D_1 \) is \( K \)-incompressible in \( X \). Assume that \( D_1 \) is \( K \)-compressible in \( X \). Then, by Lemma 8.6, the component of \( K - D_1 \) contained in \( X \) is isotopic rel \( \partial \) to an arc in \( D_1 \) joining \( D_1 \cap K \). We denote by \( K' \) the image of \( K \) under this isotopy. Then \( K' \) is contained in \( A \) and \( K' \cap \partial A (= K' \cap D_1) \) is an arc, and \( \text{cl}(K' \cap \text{Int} A) \) is a trivial arc in \( A \) (see Claim 1, Case 2.1). This implies that \( K \) is a trivial knot, a contradiction. Hence \( D_1 \) is \( K \)-incompressible in \( X \).

Let \( B^3 \) be the 3-ball in \( X \) bounded by \( D_1 \cup E \). By Lemma 8.4, we see that \( (B^3, K \cap B^3) \) is a rational tangle. Assume that \( E \) is \( K \)-compressible in \( A \). Then by applying the last half of Lemma 7.4 to \( E \) in \( A \), we see that \( (B^3, K \cap B^3) \) is a \( K \)-parallelism, contradicting the fact that \( D_1 \) and \( E \) are not \( K \)-parallel in \( X \). Hence \( E \) is \( K \)-incompressible in \( A \). Then, by Lemma 7.3 (2), we see that \( E \) and \( \text{cl}(P - D_1) \) bounds a \( K \)-parallelism in \( A \). However
this is impossible, since the boundary components of a component of $K \cap A$ is contained in $\text{cl}(P - D_1)$. This shows that Case 1.2 does not occur.

**Case 2.** Both $D_1$ and $D_2$ are non-separating in $X$.

In this case, we first note that, since $D_i \cap (X \cap K)$ $(i = 1, 2)$ consists of two points, the two points $K \cap Q$ are contained in a component of $Q - (\partial D_1 \cup \partial D_2)$. Then, let $R$ be the closure of the component of $X - (D_1 \cup D_2)$ which does not contain $K \cap Q$. Without loss of generality, we may suppose that $R$ is contained in $A$. We have the following cases.

**Case 2.1.** Both $D_1$ and $D_2$ are $K$-incompressible in $X$.

In this case, we have the following subcases.

**Case 2.1.1.** $D_1$ and $D_2$ are not $K$-parallel in $X$.

By Lemma 8.5, we see that $(R, K \cap R)$ is a rational tangle. By Lemma 8.9, we see that $A_0$ is parallel to the annulus $R \cap Q$ in $Y$, and, hence, $P$ is $K$-isotopic to $\partial R$. Now let $a$ be the component of $K \cap B$ that is contained in $X$. Then, by Lemma 8.5 again, we see that the torus obtained from $\partial R$ by adding a tube along $a$ is $K$-isotopic to $Q$. Hence we have seen that the genus one 1-bridge position $X \cup Q Y$ of $K$ is obtained from $A \cup P B$ in a standard manner by using the arc $a$.

**Case 2.1.2.** $D_1$ and $D_2$ are $K$-parallel in $X$, and there exists a $K$-incompressible disk $D'$ in $X$ such that $D'$ intersects $K$ in two points, $D'$ is non-separating in $X$, $D' \cap (D_1 \cup D_2) = \emptyset$, and $D'$ is not $K$-parallel to $D_1$ (or $D_2$).
Let $R'$ be the closure of the component of $X - (R \cup D')$ which does not contain $K \cap Q$. By exchanging subscript, if necessary, we may suppose that $R' \cap R = D_2$. Let $A' = R' \cap Q$, and $D'' = A' \cup D'$. We note that $D''$ is a disk properly embedded in $B$, which intersects $K$ in two points. Since $D''$ separates the boundary components of each component of $K \cap B$ in $P$, we see that $D''$ is $K$-incompressible. Moreover $D''$ and $D_2$ are not $K$-parallel in $X$, hence in $B$. Hence, by Lemma 7.3 (2), we see that $D''$ and $\text{cl}(P - D_2)$ ($= D_1 \cup A_0$) are $K$-parallel in $B$. This shows that $P$ is $K$-isotopic to $\partial R'$. Then, by the argument of Case 2.1.1, with regarding $R'$ as $R$ we see that the genus one 1-bridge position $X \cup Q Y$ of $K$ is obtained from $A \cup P B$ in a standard manner.

**Case 2.1.3.** $D_1$ and $D_2$ are $K$-parallel in $X$, and there does not exist a $K$-incompressible disk $D'$ as in Case 2.1.2.

![Figure 8.11](image)

In this case, it is easy to see that the argument of Case 2.1.1 works to show that the genus one 1-bridge position $X \cup Q Y$ of $K$ is obtained from $A \cup P B$ in a standard manner.

**Case 2.2.** Either $D_1$ or $D_2$ is $K$-compressible in $X$.

Without loss of generality, we may suppose that there is a $K$-compressing disk $D$ for $D_1 \cup D_2$ such that $\partial D \subset D_2$.

By applying $K$-compression on $D_2$, we obtain a compressing disk $D$ for $\partial X$ in $X$ such that $D \cap K = \emptyset$. By applying a slight isotopy, we may suppose that $D_2 \cap D = \emptyset$. By Lemma 8.9, we see that $A_0$ and $R \cap Q$ are parallel in $Y$. Hence, by $K$-isotopy, we may suppose that $P = \partial R$, (and $A = R$).

Then we have the following subcases.

**Case 2.2.1.** $D$ is not contained in $R$. 
Since $\partial D_1$ and $\partial D_2$ are $K$-parallel, we see that $D_1$ is also $K$-compressible in $X$. Then, by Lemma 8.6, we see that there is a disk $\Delta$ in $X$ such that $\Delta \cap D_1 = \partial \Delta \cap D_1 = \gamma$ an arc, and $\Delta \cap (K \cap X) = c\ell(\partial \Delta - \gamma)$. Hence, we can move $K$ by an isotopy along $\Delta$ such that $\Delta \cap (K \cap X)$ is moved to $\gamma$ (hence, the component of $K \cap B$ which intersects $Y$ is not changed by this isotopy). Since each component of $K \cap B$ is a trivial arc, this shows that $K$ is a trivial knot, a contradiction.

**Case 2.2.2.** $D$ is contained in $R$.

Let $R' = c\ell(X - R)$, and $A^* = R' \cap \partial X$. Then $A^*$ is an annulus properly embedded in $B$ such that each component of $\partial A^*$ is $K$-essential in $\partial B$, and $A^*$ intersects $K$ in two points. By Lemma 7.3, we see that there is a $K$-compressing disk $D'$ for $A^*$ in $B$.

Then we claim that $D'$ is contained in $X$. In fact, assume that $D'$ is contained in $Y$. Since $H_1(S^3) = \{0\}$, we see that $\partial D'$ bounds a disk $E$ in $A^*$ such that $E \cap K$ consists of two points, and $D' \cap K = \emptyset$. Let $B^3$ be the 3-ball in $Y$ bounded by $D' \cup E$, and $B^3 = c\ell(X - N(D))$. By Lemma 8.3 (1), we see that $K \cap B^3(= K \cap X)$ is a trivial arc in $B^3$. By Lemma 8.3 (3), we see that $K \cap B^3$ is a trivial arc in $B^3$. Note that $B^3 \cup B^3$ is a 3-ball, and $B^3 \cap B^3 = E$. This shows that $K$ is a trivial knot, a contradiction.

Hence $D'$ is contained in $X$. Since each component of $\partial A^*$ separates the boundary points of each component of $K \cap B$ in $P$, we see that $D' \cap K \neq \emptyset$, and
and, hence, $D' \cap K$ consists of a point. By applying $K$-compression on $A^*$ along $D'$, we obtain two disks $D'_1$, $D'_2$ in $B$ such that $\partial D'_1 = \partial D_1$, and $\partial D'_2 = \partial D_2$. Since $D'_i$ ($i = 1, 2$) separates the boundary components of a component of $K \cap B$, $D'_i$ is $K$-incompressible in $B$. Then, by Lemma 7.3 (2), $D_i \cup D'_i$ ($i = 1, 2$) bounds a rational tangle in $B$. Let $B^{3m}$ be the 3-ball obtained from $X$ by cutting along $D'$. We regard $D'_1, D'_2$ as contained in $\partial B^{3m}$, and $D_1, D_2$ are properly embedded in $B^{3m}$. By Lemma 8.3 (2), we see that $(B^{3m}, K \cap B^{3m})$ is a 2-string trivial tangle. Then we see that $D_1, D_2$ are $K$-compressible, by the condition of Case 2.2. Hence, by the last half of Lemma 7.4, we see that $D_1$ and $D'_1$ ($D_2$ and $D'_2$ resp.) are $K$-parallel. Let $a$ be the component of $K \cap B$ that is contained in $X$. Then, by the above observations, we see that the genus one 1-bridge position $X \cup_Q Y$ of $K$ is obtained from $A \cup_P B$ in a standard manner by using $a$.

This completes the proof of Claim 2.

Completion of the proof of Proposition 8.8. Suppose that $\sharp(P \cap Q) \geq 3$. Note that the components of $P \cap Q$ are mutually $K$-parallel in $P$. Let $D_1, A_1, \ldots, A_m$, $D_2$ be the closures of the components of $P - (P \cap Q)$ such that $D_1, D_2$ are disks and $A_1, \ldots, A_m$ are annuli that are located on $P$ in this order. Then we have the following cases.

Case 1. Either $D_1$ or $D_2$ is non-separating in the solid torus.

Without loss of generality, we may suppose that $D_1$ is contained in $X$, and is non-separating in $X$. Then $A_1$ is contained in $Y$, and, by Lemma 8.9, there is a homeomorphism from $A_1 \times I$ to $Y$ such that $A_1$ corresponds to $A_1 \times \{1/2\}$. Let $U$ be the closure of the component of $Y - A_1$ which does not contain $K \cap Q$.

Suppose that $(\text{Int}U) \cap P \neq \emptyset$. Then we can push the component of $(\text{Int}U) \cap P$ along the parallelism $U$ to $X$ to reduce $\sharp(P \cap Q)$, yet still have at least two components $\partial A_1$.

Suppose that $(\text{Int}U) \cap P = \emptyset$. Then we can push $A_1$ along the parallelism $U$ to $X$ to reduce $\sharp(P \cap Q)$ by two.

In either case we have the conclusion by the assumption of the induction.

Case 2. Both $D_1$ and $D_2$ are separating in the solid torus.

If $D_1$ or $D_2$ is $K$-boundary parallel, then we can apply the assumption of the induction by the argument as in Case 1. Hence we suppose that $D_1$ and $D_2$ are not $K$-boundary parallel. Let $E_i$ be the disk in $Q$ bounded by $\partial D_i$. Without loss of generality, we may suppose that $E_1 \subset E_2$, and $D_1$ is contained in $X$. Then we have the following subcases.

Case 2.1. $D_1$ is $K$-incompressible in $X$.

Let $B^3$ be the closure of the component of $X - D_1$ such that $\partial B^3 = D_1 \cup E_1$. Then, by Lemma 8.4, $(B^3, K \cap B^3)$ is a rational tangle. Suppose
We note that each component of \( c\ell(\text{Int}B^3) \cap P \) is an annulus whose boundary components are parallel to \( \partial D_1 \) in \( Q \setminus K \). This shows that each annulus is \( K \)-incompressible in \( B^3 \). Hence, by Lemma 7.3 (3), we see that the closure of each component of \( (\text{Int}B^3) \cap P \) is boundary parallel annulus. Hence, we can push them to \( Y \) to reduce \( \sharp(P \cap Q) \), and we have the conclusion by the assumption of the induction. If \( (\text{Int}B^3) \cap P = \emptyset \), then by Lemma 7.3 (2), the last half of Lemma 7.4, and the assumption that \( D_1 \) is not \( K \)-boundary parallel in \( X \), we see that \( c\ell(P - D_1) \) is \( K \)-isotopic to \( E_1 \) rel \( D_1 \). Hence, by the argument in Claim 1, we see that the genus one 1-bridge position \( X \cup_Q Y \) of \( K \) is obtained from \( A \cup_P B \) in a standard manner.

**Case 2.2.** \( D_1 \) is \( K \)-compressible in \( X \).

We moreover have the following subcases.

**Case 2.2.1.** \( D_2 \subset Y \), and \( D_2 \) is \( K \)-incompressible in \( Y \).

Let \( B^3 \) be the closure of the component of \( Y - D_2 \) such that \( \partial B^3 = D_2 \cup E_2 \). In this case, since \( \partial D_1 \subset \text{Int}E_2 \), we see that \( (\text{Int}B^3) \cap P \neq \emptyset \). Hence, by the argument of Case 2.1 (for the case \( (\text{Int}B^3) \cap P \neq \emptyset \) ), we can show that the genus one 1-bridge position \( X \cup_Q Y \) of \( K \) is obtained from \( A \cup_P B \) in a standard manner.

**Case 2.2.2.** \( D_2 \subset Y \), and \( D_2 \) is \( K \)-compressible in \( Y \).

In this case, by \( K \)-compressing \( D_2 \), we obtain a \( K \)-compressing disk \( D'_2 \) for \( \partial Y \) in \( Y \) such that \( \partial D'_2 = \partial D_2 \). Then, by Lemma 8.3 (3), we see that \( K \cap Y \) is rel \( \partial \) isotopic to an arc \( \alpha_Y \) in \( E_1 \). And we also see that \( K \cap X \) is rel \( \partial \) isotopic to an arc \( \alpha_X \) in \( E_1 \) such that \( \alpha_X \cap \alpha_Y = \partial \alpha_X = \partial \alpha_Y \). Hence \( K \) is a trivial knot, a contradiction.

**Case 2.2.3.** \( D_2 \subset X \).

Let \( T' \) and \( B^3 \) be the closures of the component of \( T - D_1 \) such that \( T' \) is a solid torus and \( B^3 \) is a 3-ball. Without loss of generality, we may suppose that \( B^3 \)-side of \( D_1 \) is contained in \( B \). Since \( D_1 \) is \( K \)-compressible, we see that \( K \cap T' \) is rel \( \partial \) isotopic in \( T' \) to an arc \( \alpha \) in \( D_1 \), by an isotopy that does not move \( K \setminus T' \). Note that since \( D_2 \subset X \), \( c\ell(K - T') \) is a component of \( K \cap B \), hence, a trivial arc in \( B \). This shows that \( K \) is a trivial knot, a contradiction.

This completes the proof of Proposition 8.8.

**Proof of Theorem 8.2.** Let \( A \cup_P B \) be a Heegaard splitting which gives a 2-bridge position of \( K \), and \( X \cup_Q Y \) a Heegaard splitting which gives the given genus one 1-bridge position of \( K \). By Proposition 8.7, it is enough to assume that \( X \cup_Q Y \) is strongly \( K \)-irreducible for the proof of Theorem 8.2. Then, by Proposition 7.5, and Corollary 6.22, we may suppose that \( P \) and \( Q \) intersect...
in non-empty collection of simple closed curves which are $K$-essential in both $P$ and $Q$. Then we have the conclusion by Proposition 8.8. □

Appendix A.

Let $\gamma$ be a system of trivial arcs in a handlebody $H$, and $p : \tilde{H} \to H$ the two fold branched cover of $H$ along $\gamma$.

Let $F$ be a surface properly embedded in $H$, which is in general position with respect to $\gamma$. Then, by using $\mathbb{Z}_2$-equivariant loop theorem [12], we see that:

**Lemma A.1.** $F$ is $\gamma$-incompressible if and only if $\tilde{F} (= p^{-1}(F))$ is incompressible.

Moreover, by using $\mathbb{Z}_2$-equivariant cut and paste argument as in [9, Proof of 10.3], we see that:

**Lemma A.2.** A $\gamma$-incompressible surface $F$ is $\gamma$-boundary compressible if and only if $\tilde{F}$ is boundary compressible.

By using $\mathbb{Z}_2$-Smith conjecture ([21], [16]) together with the $\mathbb{Z}_2$-equivariant cut and paste argument and the irreducibility of $H$, we have:

**Lemma A.3.** A $\gamma$-incompressible surface $F$ is $\gamma$-boundary parallel if and only if $\tilde{F}$ is boundary parallel. In particular, if $F$ is a disk intersecting $\gamma$ in one point, and $\partial F$ bounds a disk $D$ in $\partial H$ such that $D$ intersects $\gamma$ in one point, then $F$ is $\gamma$-boundary parallel (in fact, $F$ and $D$ are $\gamma$-parallel).

Appendix B.

Let $(B, \beta)$ be a 2-string trivial tangle, and $(\tilde{B}, \tilde{\beta})$ the 2-fold branched covering space of $B$ along $\beta$. Then $\tilde{B}$ is a solid torus, $\tilde{\beta}$ a system of two trivial arcs in $\tilde{B}$, and the covering translation $\tau$ is a $\pi$-rotation along $\tilde{\beta}$ (for details, see [3, Chapter 12]).

![Figure B-1.](image)
We leave the proof of the next lemma to the reader.

**Lemma B.0.** Let $F$ be an orientable incompressible surface properly embedded in a solid torus. Then either:

1. $F$ is a meridian disk,
2. $F$ is a boundary parallel disk, or
3. $F$ is a boundary parallel annulus.

Then, we show:

**Lemma B.1.** Let $D$ be a $\beta$-essential surface in $B$. Then $D$ is a disk disjoint from $\beta$, and $D$ separates the components of $\beta$.

*Proof.* Let $\tilde{D}$ be the lift of $D$ in $\tilde{B}$. By Lemmas A-1, A-3, we see that $\tilde{B}$ is an essential surface in the solid torus $\tilde{B}$. By Lemma B-0, we see that $\tilde{D}$ is a meridian disk. Suppose that $\tilde{D} \cap \tilde{\beta} \neq \emptyset$. Then we see that $\tilde{D} \cap \beta$ consists of a point, and hence, $D \cap \beta$ consists of a point. However, this implies that $\partial D$ is $\beta$-inessential in $\partial B$. Hence, by Lemmas A-3, and B-0, we see that $D$ is $\beta$-boundary parallel, a contradiction. Since $\partial D$ is $\beta$-essential in $\partial B$, we see that $\partial D$ separates the points $\partial \beta$ in $\partial B$. This shows that $D$ separates the components of $\beta$. □

**Lemma B.2.** Let $F$ be a $\beta$-incompressible surface in $B$. Then either

0. $F$ is $\beta$-essential,
1. $F$ is a $\beta$-boundary parallel disk intersecting $\beta$ in at most one point,
2. $F$ is a $\beta$-boundary parallel disk intersecting $\beta$ in two points and $F$ separates $(B, \beta)$ into the parallelism and a rational tangle, or
3. $F$ is a $\beta$-boundary parallel annulus such that $F \cap \beta = \emptyset$.

*Proof.* Let $\tilde{F}$ be the lift of $F$ in $\tilde{B}$. By Lemma A-1, we see that $\tilde{F}$ is one of (1), (2), or (3) of Lemma B-0. It is easy to see that (1) ((2) resp.) of Lemma B-0 corresponds to the conclusion (0) ((1) resp.). Suppose that $\tilde{F}$ is an incompressible annulus ((3) of Lemma B-0). Then it is easy to see that we have conclusion (2) if $\tilde{F} \cap \tilde{\beta} \neq \emptyset$, and we have conclusion (3) if $\tilde{F} \cap \tilde{\beta} = \emptyset$. □

**Lemma B.3.** Let $D$ be a $\beta$-compressible disk in $B$ such that $\partial D$ is $\beta$-essential in $\partial B$, and $D \cap \beta$ consists of two points.

Then $D$ separates $(B, \beta)$ into two tangles $(B_1, \beta_1)$, and $(B_2, \beta_2)$, where $(B_1, \beta_1)$ is a rational tangle such that there is a $\beta$-essential disk $D'$ in $(B_1, \beta_1)$ with $D \cap D' = \emptyset$. Moreover, if $(B_2, \beta_2)$ happens to be a rational tangle, then $(B_2, \beta_2)$ is a $\beta$-boundary parallelism for $D$.

*Proof.* Let $\tilde{D}$ be the lift of $D$ in $\tilde{B}$. By Lemma A-1, we see that $\tilde{D}$ is a compressible annulus in $\tilde{B}$. Since $\partial D$ is $\beta$-essential, we see that, by compressing $\tilde{D}$, we obtain two meridian disks, say $D_1$ and $D_2$. Let $B_1^3$, $B_2^3$ be the closures of the components of $\tilde{B} - (D_1 \cup D_2)$. Then $B_1^3$, $B_2^3$ are 3-balls,
and the closure of a component of $\tilde{B} - \tilde{D}$, say $\tilde{T}$, is obtained from one of $B^3_1$ or $B^3_2$, say $B^3_1$, by adding a 1-handle, and hence $\tilde{T}$ is a solid torus, and there is an equivariant compressing disk $\tilde{D}'$ for $\tilde{D}$ in $\tilde{T}$ such that $\tau(\tilde{D}') \cap \tilde{D}' = \emptyset$. Hence $\tilde{T}/\tau$ gives a rational tangle, and $p(\tilde{D}')$ gives $D'$. Then the closure of the other component of $\tilde{B} - \tilde{D}$, say $E$, is obtained from $B^3_2$ by removing a regular neighborhood of an arc properly embedded in $B^3_2$, and taking the closure. This shows that $E$ is homeomorphic to the exterior of a knot in $S^3$ by a homeomorphism such that $E \cap \partial T$ is a regular neighborhood of a meridian loop of the knot in $\partial E$. Suppose that $E$ is a solid torus, i.e., $(E, K \cap E)$ is a rational tangle. Then the knot is a trivial knot. Since $E \cap \partial T$ is a regular neighborhood of a meridian loop of the knot, we see that $E$ is a $\partial$-parallelism for $\tilde{D}$, and by Lemma A-3, we see that $(B_2, \beta_2)$ is a $\beta$-boundary parallelism for $D$.

Appendix C.

Let $\alpha$ be a trivial arc in a solid torus $T$, and $(\tilde{T}, \tilde{\alpha})$ the 2-fold branched covering space of $T$ along $\alpha$. Then $\tilde{T}$ is a genus two handlebody, $\tilde{\alpha}$ a 1-string trivial arc in $\tilde{T}$, and the covering translation is a $\pi$-rotation along $\tilde{\alpha}$.

![Figure C-1.](image)

Lemma C.1. Let $D$ be an $\alpha$-compressing disk for $\partial T$. Then $D$ is either:

1. a meridian disk of $T$ with $D \cap \alpha = \emptyset$. In this case, we obtain, by cutting $(T, \alpha)$ along $D$, a 1-string trivial tangle,
2. a meridian disk of $T$ with $D \cap \alpha$ consists of one point, and we obtain by cutting $(T, \alpha)$ along $D$, a 2-string trivial tangle, or
3. $\partial$-parallel disk in $T$ with $D \cap \alpha = \emptyset$. In this case, $D$ cobounds a 1-string trivial tangle in $(T, \alpha)$. 
Proof. Let $\tilde{D}$ be the lift of $D$ in $\tilde{T}$. Then we have either $\tilde{D}$ is a union of two disks if $D \cap \alpha = \emptyset$ (Case 1), or $\tilde{D}$ is a disk if $D \cap \alpha$ consists of one point (Case 2).

In Case 1, we have either each component of $\tilde{D}$ is non-separating, or separating, which correspond to the conclusions (1), (3) respectively. In Case 2, it is easy, by a homological argument, to see that $\tilde{D}$ is non-separating, and this gives conclusion (2) \(\square\)

Lemma C.2. Let $D$ be an $\alpha$-essential disk in $T$ such that $D \cap \alpha$ consists of two points. Then there exists an $\alpha$-compressing disk $D'$ for $\partial T$ such that $D' \cap D = \emptyset$, and $D' \cap \alpha$ consists of one point. Moreover, by cutting $(T, \alpha)$ along $D'$, we obtain a 2-string trivial tangle $(B, \beta)$ such that $D$ is a $\beta$-incompressible disk in $(B, \beta)$ (hence $D$ is $\beta$-boundary parallel).

Proof. Let $\tilde{D}$ be the lift of $D$ in $\tilde{T}$. By Lemma A-1, we see that $\tilde{D}$ is an essential annulus in $\tilde{T}$. Then it is easy to see that:

The annulus $\tilde{D}$ is obtained from a meridian disk $\Delta$ by attaching a band. (For a proof of this, see, for example, [13, Lemma 3.2].)

Let $\tilde{F} = c\ell(\partial \tilde{T} - N(\partial \tilde{D}))$, where $N(\partial \tilde{D})$ is a regular neighborhood of $\partial \tilde{D}$ in $\partial \tilde{T}$. Note that $\tilde{F}$ is compressible in $\tilde{T}$ (in fact, slightly push off of $\Delta$ gives a compressing disk of $\tilde{F}$). Hence, by $\mathbb{Z}^2$-equivariant loop theorem [12], we have an equivariant compressing disk(s) $G$ for $\tilde{F}$.

Claim 1. $G$ consists of one disk, and, hence, $G \cap \alpha$ consists of one point.

Proof. Assume that $G$ consists of two disks $D_1$, $D_2$. Then we have the following three cases.

Case 1. Each $D_i$ is separating in $\tilde{T}$.

In this case $D_1$ and $D_2$ are parallel and the closures of $\tilde{T} - G$ are two solid tori $T_1$, $T_2$, and a 3-ball $B$, which is a parallelism between $D_1$, and $D_2$. Note that $\tilde{\alpha}$ is contained in $B$, and this shows that $\tilde{D}$ is contained in $B$, contradicting the incompressibility of $\tilde{D}$.

![Figure C-2.](image-url)
Case 2. Each $D_i$ is non-separating, and $G = D_1 \cup D_2$ is non-separating in $\bar{T}$.

Since $G \cap \bar{D} = \emptyset$, this contradicts the incompressibility of $\bar{D}$.

![Figure C-3.](image)

Case 3. Each $D_i$ is non-separating, and $G = D_1 \cup D_2$ is separating in $\bar{T}$.

In this case, we see that $D_i$ ($i = 1, 2$) intersects $\bar{\beta}$, a contradiction.  

![Figure C-4.](image)

Hence, $G$ is a disk, and $G \cap \alpha$ consists of one point. Then:

Claim 2. $G$ is non-separating in $\bar{T}$.

Proof. Assume that $G$ is separating. Then the closures of $\bar{T} - G$ are solid tori $T_1, T_2$ with $\tau(T_i) = T_i$. Moreover the fixed point set of $\tau|_{T_i}$ is an arc in $T_i$. But, by using $\mathbb{Z}_2$-equivariant loop theorem, it is easy to see that such $\tau|_{T_i}$ does not exist. 

By Claim 2, we see that we obtain a solid torus $T'$ by cutting $\bar{T}$ along $G$, and this shows that we obtain a 2-string trivial tangle $(B, \beta)$, and obviously $D$ is $\beta$-incompressible.  

![Figure C-5.](image)
Lemma C.3. Let $D_1$, $D_2$ be mutually disjoint non $\alpha$-parallel, $\alpha$-essential disks such that $D_i \cap \alpha$ ($i = 1, 2$) consists of two points. Then there exists an $\alpha$-compressing disk $D'$ for $\partial \tilde{T}$ disjoint from $D_1 \cup D_2$ such that $D' \cap \alpha$ consists of one point. Moreover each $D_i$ is non-separating in $\tilde{T}$, and by cutting $(T, \alpha)$ along $D'$, we obtain 2-string trivial tangle $(B, \beta)$, and $D_1$, $D_2$ are mutually non $\beta$-parallel, $\beta$-boundary parallel, $\beta$-incompressible disks in $(B, \beta)$.

Proof. Let $\tilde{D}_i$ be the lift of $D_i$ in $\tilde{T}$ ($i = 1, 2$). By Lemma A-1, we see that $\tilde{D}_i$ is an incompressible annulus in $\tilde{T}$. Let $\tilde{F} = cl(\partial \tilde{T} - N(\partial \tilde{D}_1 \cup \partial \tilde{D}_2, \partial \tilde{T}))$. The argument in the proof of Lemma C-2 works in this case to show that there is an equivariant compressing disk $G$ for $\tilde{F}$ such that $G$ is non-separating, and $G$ intersects $\tilde{\alpha}$ in one point. Then let $\tilde{T}'$ be the solid torus obtained by cutting $\tilde{T}$ along $G$, and $G_1$, $G_2$ the copies of $G$ in $\partial \tilde{T}'$. By Lemma B-0 (3), we see that there are annuli $A_1$, $A_2$ in $\partial \tilde{T}'$ such that $A_1$ and $\tilde{D}_2$ are parallel ($i = 1, 2$). Since $\tilde{D}_1$, $\tilde{D}_2$ are essential, we see that $G_1 \subset A_1$, $G_2 \subset A_2$. It is easy to see that this gives the conclusion of Lemma C-3. □

Lemma C.4. Let $D$ be an $\alpha$-compressible disk such that $\partial D$ is $\alpha$-essential in $\partial \tilde{T}$, and $D \cap \alpha$ consists of two points. Then there is a disk $\Delta$ in $T$ such that $\Delta \cap D = \partial \Delta \cap D = \gamma$ an arc, and $\Delta \cap \alpha = cl(\partial \Delta - \gamma)$. Particularly, if $D$ is separating in $T$, then $D$ separates $(T, \alpha)$ into $(T', \alpha')$, and $(B', \alpha'')$ such that $\alpha'$ is a trivial arc in a solid torus $T'$. In this case, if $(B', \alpha'')$ happens to be a rational tangle, then $(B', \alpha'')$ is an $\alpha$-boundary parallelism.

Proof. Let $\tilde{D}$ be the lift of $D$ in $\tilde{T}$. By Lemma A-1, we see that $\tilde{D}$ is a compressible annulus in $\tilde{T}$. It is easy to see that there is a compressing disk $\Delta$ for $\tilde{D}$ such that $\tau(\Delta) = \tilde{\Delta}$, hence $\tilde{\alpha} \subset \tilde{\Delta}$. Then the projection of $\Delta$ gives $\Delta$. Since $\partial D$ is $\beta$-essential, we see that, by compressing $\tilde{D}$ along $\tilde{\Delta}$, we obtain two meridian disks, say $D_1$ and $D_2$, which are mutually parallel in $\tilde{T}$. Suppose that $D$ is separating in $T$. Then $\tilde{D}$ is also separating in $\tilde{T}$, and the closures of the components of $\tilde{B} - (D_1 \cup D_2)$ consist of $B^3$ and $T'$, where $B^3$ is a 3-ball, and $T'$ is a solid torus. Then the closure of a component of $\tilde{T} - \tilde{D}$, say $H$, is obtained from $T'$ by adding a 1-handle, and hence $H$ is a genus two handlebody. Now we consider the closure of the other component of $\tilde{T} - \tilde{D}$, say $E$. Then $E$ is obtained from $B^3$ by removing a regular neighborhood of an arc properly embedded in $B^3$, and taking the closure. This shows that $E$ is homeomorphic to the exterior of a knot in $S^3$ by a homeomorphism such that $E \cap \partial T$ is a regular neighborhood of a meridian loop of the knot in $\partial E$. Suppose that $E$ is a solid torus, i.e., $(B', \alpha'')$ is a rational tangle. Then the knot is a trivial knot. Since $E \cap \partial T$ is a regular neighborhood of a meridian loop of the knot, we see that $E$ is a $\partial$-parallelism for $\tilde{D}$, and by Lemma A-3, we see that $D$ is $\alpha$-boundary parallel. □
References


Received April 1, 1998 and revised June 24, 1999. The authors were supported by Grant-in-Aid for Scientific Research, Ministry of Education, Science and Culture, Japan.

DEPARTMENT OF MATHEMATICS  
NARA WOMEN’S UNIVERSITY  
KITA-UOYA NISHIMACHI  
NARA 630  
JAPAN  
E-mail address: tsuyoshi@cc.nara-wu.ac.jp

DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE  
HIROSHIMA UNIVERSITY  
HIGASHI-HIROSHIMA 739  
JAPAN  
E-mail address: saeki@top2.math.sci.hiroshima-u.ac.jp
GEOMETRIC REALIZATIONS OF FORDY–KULISH NONLINEAR SCHRÖDINGER SYSTEMS

JOEL LANGER AND RON PERLINE

A method of Sym and Pohlmeyer, which produces geometric realizations of many integrable systems, is applied to the Fordy–Kulish generalized non-linear Schrödinger systems associated with Hermitian symmetric spaces. The resulting geometric equations correspond to distinguished arclength-parametrized curves evolving in a Lie algebra, generalizing the localized induction model of vortex filament motion. A natural Frenet theory for such curves is formulated, and the general correspondence between curve evolution and natural curvature evolution is analyzed by means of a geometric recursion operator. An appropriate specialization in the context of the symmetric space $SO(p+2)/SO(p) \times SO(2)$ yields evolution equations for curves in $\mathbb{R}^{p+1}$ and $S^p$, with natural curvatures satisfying a generalized mKdV system. This example is related to recent constructions of Doliwa and Santini and illuminates certain features of the latter.

1. Introduction.

Shortly after it was discovered that the Korteweg-deVries equation could be linearized via the spectral transform method [G-G-K-M], Shabat and Zakharov [S-Z] showed that the method could also be applied to the (cubic) non-linear Schrödinger equation,

\[ -i \psi_t = \psi_{ss} + \frac{1}{2} |\psi|^2 \psi. \]

Almost concurrently, Hasimoto [Has] discovered the connection between NLS and the localized induction equation (LIE), an idealized model of the evolution of the curved centerline of a thin vortex tube in a three-dimensional ideal fluid. (See [Bat] for a derivation, and [Ric] for the history of this equation, also known as the Betchov-Da Rios equation.) Denoting this evolving centerline by $\gamma(s,t)$ (where $s$ is arclength along the curve and $t$ is time), the curve evolution in this model is described by

\[ \gamma_t = \gamma_s \times \gamma_{ss} = \kappa B, \]

where $\kappa(s)$ is the curvature and $B$ the binormal. Recall that along a space curve, the Frenet frame $\{T,N,B\}$ satisfies the equations $T_s = \kappa N$, $N_s =$
−κT + τB, \( B_s = −τN \). The LIE-NLS connection is this: If a curve \( γ \) with curvature \( κ \) and torsion \( τ \) evolves according to LIE, then the associated complex curvature function, \( ψ = κe^{i\int^u τ(u)du} \), evolves according to NLS.

In view of the LIE-NLS correspondence, it is not surprising that LIE manifests familiar integrability characteristics, but in geometric form: Soliton solutions, a hierarchy of conserved Hamiltonians in involution, a recursion operator generating the corresponding infinite sequence of commuting Hamiltonian vectorfields – the localized induction hierarchy.

The Hamiltonian nature of LIE itself was introduced by Marsden and Weinstein in [M-W]; the equation’s Poisson geometry was further elucidated in [L-P1]; Yasui and Sasaki developed the structure of LIE in the setting of hereditary operator, Hamiltonian pairs, and master symmetries [S-Y].

Other recent papers have addressed a variety of closely related geometric topics of a more concrete nature, including: Knotted soliton curves of constant torsion [C-I]; planar, spherical, and constant torsion-preserving curve evolution [L], [L-P3, L-P4]; integrable variational problems for curves [Lan-S1, Lan-S2, Lan-S3]; pseudospherical surfaces and Weingarten systems [Per1, Per2], evolution of immersed Riemann surfaces in \( \mathbb{R}^3 \) preserving the Willmore integral [G-L]. It is by now clear that the localized induction hierarchy is a rich source of examples and structure in the classical differential geometry of curves and surfaces.

Here we consider natural generalizations of the LIE hierarchy in higher dimensional spaces. Our starting point is the Fordy-Kulish [F-K] construction of a generalized nonlinear Schrödinger equation (gNLS) (with spectral problem) associated to a Hermitian symmetric Lie algebra \( g \). We apply a technique due to Sym [Sym] and Pohlmeyer [Pohl], differentiation with respect to the spectral parameter, which produces geometric realizations of many integrable systems. By this route, we arrive at a generalized LIE hierarchy for distinguished arclength-parametrized curves evolving in \( g \), the first three terms of which are:

\[
\begin{align*}
γ_t &= γ_s, \\
γ_t &= −[γ_s, γ_{ss}], \\
γ_t &= −\left(γ_{sss} + \frac{3}{2}[γ_{ss}, [γ_s, γ_{ss}]]\right).
\end{align*}
\]

Here, \([\ , \]\) is the Lie bracket in \( g \), and the subscript \( s \) denotes derivative by a curve parameter which is unit speed with respect to the Cartan-Killing form on \( g \). In this setting, a direct generalization of Hasimoto’s result is proved (Theorem 3), establishing the correspondence between the above curve evolution equations and evolution of natural curvatures by equations in the gNLS hierarchy; in particular, a curve evolving by the second order flow, gLIE, has curvatures satisfying gNLS.
Interestingly, as in the three-dimensional case, the odd-order flows are more amenable to geometrically meaningful reductions. In fact, by an ad hoc reduction, in the class of symmetric spaces $SO(p+2)/SO(p) \times SO(2)$, we were able to fully realize our original goal; namely, we obtain geometric evolution equations applicable to arbitrary smooth curves in $E^n$ and $S^n$. For the third order flow, our equations take the form

$$\gamma_t = -\left(\gamma_{sss} + \frac{3}{2}\|\gamma_{ss}\|^2\gamma_s\right) = -\left(\frac{1}{2}k^2T + \sum_i (u_i)_sU_i\right).$$

Here, $u_1, \ldots, u_{n-1}$ are curvatures belonging to a natural frame $T, U_1, \ldots, U_{n-1}$. We show (Theorem 5) that the corresponding natural curvature vector, $u = (u_1, \ldots, u_{n-1})$, satisfies the vector modified Korteweg-deVries equation:

$$(mKdV) \quad u_t = -\left(u_{sss} + \frac{3}{2}|u|^2u_s\right).$$

Note that these simple equations for $\gamma$ and $u$ are given, finally, without reference to a Lie algebra.

We now describe the contents of the paper. Section 2 is a review of the Fordy-Kulish construction of generalized NLS equations. In Section 3, we apply the Sym-Pohlmeyer geometrization procedure in the Fordy-Kulish setting, and develop a natural Frenet theory for the resulting curves. In Section 4, we introduce the geometric recursion operator for the generalized LIE hierarchy, and derive key variation formulas. Section 5 treats the special class of Hermitian symmetric spaces mentioned above, and the reduction yielding curve evolutions in Euclidean spaces and spheres. We note that our constructions in the latter case are related to recent work of Doliwa and Santini [D-S]; in fact, our investigation developed out of an effort to better understand their equations. Since the completion of our paper, we have learned from Chuu-Lian Terng of her own work (with K. Uhlenbeck) on generalizations of LIE [T-U1, T-U2].

2. The Fordy-Kulish generalizations of NLS.

Following a standard framework in the theory of integrable systems, the nonlinear soliton equations arise as compatibility conditions for an overdetermined linear system

$$(LS) \quad \phi_s = (\lambda A + Q)\phi, \quad \phi_t = V\phi.$$  

This system involves two independent variables, $s$ ("position") and $t$ ("time"), and a scalar $\lambda$, the spectral parameter. The eigenfunction $\phi(s, t; \lambda)$ has values in a Lie group $G$, while $U(s, t; \lambda) = \lambda A + Q(s, t)$ and $V(s, t; \lambda)$ have values in the Lie algebra $g$ of $G$. Here $Q$ is the potential, which is meant to evolve isospectrally, hence the lack of $\lambda$-dependence. For the
Fordy-Kulish generalized NLS equations, \( g \) is taken to be the compact real form of a complex semi-simple Lie algebra \( g_C \); in fact, \( g \) is required to be a *Hermitian symmetric Lie algebra*, and \( A, Q \), are specific to the structure of \( g \). To recall briefly some of the relevant features of this structure, \( g \) has a decomposition as a vector space sum, \( g = k \oplus m \), of a compact subalgebra \( k \) and complement \( m \), satisfying the bracket conditions \([k, k] \subset k, [m, m] \subset k, \) and \([k, m] \subset m \). Also, \( k \) is associated with a special element \( A \) in \( h \), a Cartan subalgebra of \( g \); namely, \( k \) is the commutator algebra of \( A : k = \ker(ad_A) = \{ B \in g : [B, A] = 0 \} \). Further, \( J = ad_A \) satisfies \( J^2|_m = -Id \), i.e., \( J \) is a complex structure on \( m \). Such an element \( A \) is fixed to form (LS) and \( Q \) is required to be an \( m \)-potential, that is, \( Q(s, t) \in m \) for all \( t \). The set of \( m \)-potentials \( Q(s) \) is clearly a vector space; we will refer to a tangent vectorfield \( W \) as a \( m \)-field. Some of the above will be explained more explicitly, as required for specializations, below. (Also, see [F-K], [Hel] for more details.)

Cross-differentiating the equations in (LS) gives the zero curvature condition \( U_t - V_s + [U, V] = 0 \) or

\[
Q_t = V_s - [\lambda A + Q, V].
\]

With the aim of finding \( V \) in terms of \( Q \), such that the compatibility condition (ZC1) is satisfied, a *polynomial ansatz* is invoked: \( V = \sum_{j=0}^n P^{(j)}(s, t)\lambda^{n-j} \). (Our indexing convention reverses the order of \([F-K]\).) The strategy here is to substitute this expression for \( V \) into (ZC1), set the coefficients of powers of \( \lambda \) equal to zero, then solve recursively for the \( P^{(j)} \) and, finally, obtain a nonlinear PDE for the \( m \)-field \( Q \) (from the \( \lambda^0 \) term). To carry this out requires the decomposition of \( g \) given above. Namely, each \( P^{(j)} \) is decomposed as \( P^{(j)} = P^{(j)}_m + P^{(j)}_k \), with \( P^{(j)}_m \in m \) and \( P^{(j)}_k \in k \). Then, using the above bracket conditions and \( J^2 = -Id \), one obtains the equations:

\[
P^{(0)}_m = 0,
\]

\[
P^{(j)}_m = -J(\partial_s P^{(j-1)}_m - [Q, P^{(j-1)}_k]), \quad j = 1, \ldots, n;
\]

\[
\partial_s P^{(j)}_k = [Q, P^{(j)}_m], \quad j = 0, \ldots, n,
\]

\[
Q_t = \partial_s P^{(n)}_m - [Q, P^{(n)}_k].
\]

Note the first and third of these equations imply \( P^{(0)} \) is necessarily a constant in \( k \). – we will take the “obvious choice” \( P^{(0)}_k = P^{(0)} = A \). Also, a choice of “constant of integration” is made, at each stage, as we are apparently required to compute an antiderivative to obtain \( P^{(j)}_k \). An essential (and remarkable) feature of the recursion scheme is that the antiderivative is explicit, and is polynomial in \( Q \) and its derivatives; here, we simply illustrate this point with the important case \( n = 2 \). The required terms are readily generated in the order: \( P^{(1)}_m = Q, P^{(1)}_k = 0, P^{(2)}_m = -[A, Q], P^{(2)}_k = \ldots \).
\[ Q_t = -JQ_{ss} - \frac{1}{2} ad_Q A, \]

which we will refer to as the Fordy-Kulish generalized NLS equation. (In [F-K], the NLS equations are displayed componentwise, rather than in vector notation.)

The above recursion scheme may be more compactly described by introducing a recursion operator \( \tilde{R} = \tilde{R}_Q \) which takes \( m \)-fields \( \tilde{X} \) to \( n \)-fields:

\[ \tilde{R} \tilde{X} = - (\partial_s - ad_Q \partial_s^{-1} ad_Q) J \tilde{X}. \]

(Adjusted specification of the antiderivative \( \partial_s^{-1} \) depends on the context.)

Defining \( \tilde{X}^{(j)} = J P_m^{(j)} \), we can now write the recursion scheme and nonlinear equations as

\[ \tilde{X}^{(1)} = JQ, \quad \tilde{X}^{(2)} = \tilde{R} \tilde{X} = Q_s, \]

\[ \tilde{X}^{(j+1)} = \tilde{R} \tilde{X}^{(j)}, \quad j = 1, 2, 3, \ldots, \]

\[ Q_t = \tilde{X}^{(j+1)}, \quad j = 0, 1, 2, \ldots. \]

The last of these equations defines the \((j+1)\)rst term in the Fordy-Kulish NLS hierarchy, the above NLS equation being the third term.

It will be useful to have a concrete (and particularly simple) example at hand for illustrating the main ideas in the next few sections; thus, we begin our:

**Running example.**

For the “classical NLS”, we take \( g = su(2) \). We use the basis \( A = -\frac{i}{2} \sigma^3, \quad B = \frac{i}{2} \sigma^1, \quad C = \frac{i}{2} \sigma^2 \), where \( \sigma^1, \sigma^2, \sigma^3 \) are the Pauli matrices

\[ \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \]

The bracket relations \([A, B] = C, \quad [B, C] = A, \quad [C, A] = B\) imply that \( k = \text{span}(A) \) and \( m = \text{span}(B, C) \) define a Hermitian symmetric Lie algebra structure on \( g = su(2) \). Writing \( Q = bB + cC \), and plugging into the generalized NLS equation yields:

\[ Q_t = -J \left( Q_{ss} + \frac{1}{2} |Q|^2 Q \right), \]

where \( |Q|^2 = (b^2 + c^2) \). In this case, we can identify \( m \) with the complex numbers (\( Q \) with \( \psi = b + ic \)), and then \( J \) coincides with multiplication by \( i \). Using this identification (and a time reversal), we obtain exactly (NLS), given in the introduction.
We conclude this section by recording some useful general identities, to be used in later sections. First of all, we note that the above treatment of the overdetermined linear system $LS$ did not fully reflect the dependence of $\phi$ on the three variables, $s$, $t$, and $\lambda$. Introducing $W(s, t, \lambda) = \phi^{-1}\phi \lambda$, we write down the augmented linear system

$\phi_s = U\phi$, $\phi_t = V\phi$, $\phi_\lambda = W\phi$,

with corresponding compatibility conditions

\[(ZC2) \quad U_t - V_s + [U, V] = V_\lambda - W_t + [V, W] = W_s - U_\lambda + [W, U] = 0.\]

Secondly, the geometric objects related to $LS$ will be expressed in terms of conjugates of $U$, $V$, and $W$, for which we will use the following notational shorthand: For $B, C \in g$, we write $\{B\} = \phi^{-1}B\phi$, and $\{B, C\} = \{[B, C]\}$. The following lemma (whose proof involves straightforward differentiation) states analogues of a standard principle of rigid body mechanics (with “time” being $s$, $t$, or $\lambda$): absolute velocity = relative velocity + transferred velocity ([Arn], p. 128).

**Proposition 1.** For any $g$-field $B(s, t, \lambda)$,

1. $\{B\}_s = \{B_s\} + \{B, U\}$,
2. $\{B\}_t = \{B_t\} + \{B, V\}$, and
3. $\{B\}_\lambda = \{B_\lambda\} + \{B, W\}$.

Finally, combining Proposition 1 with (ZC2) yields at once six simple identities; three of these we will use, so we collect them in:

**Proposition 2.** For $U$, $V$, and $W$ as above, we have

$\{V\}_s = \{U_t\}$, $\{W\}_t = \{V_\lambda\}$, and $\{W\}_s = \{U_\lambda\}$.

### 3. Sym-Pohlmeyer curves.

Throughout this section, we “freeze time” in the definitions of the previous section. In other words, we consider a time-independent potential $Q(s)$, and suppose $\phi = \phi(s; \lambda)$ satisfies the linear system $\phi_s = U\phi = (\lambda A + Q)\phi$, for each value of the “parameter” $\lambda$. Setting $W(s, \lambda) = \phi\lambda \phi^{-1}$, we consider the $g$-valued function

$\gamma(s, \lambda) = \{W\} = \phi^{-1}\phi_\lambda$.

By Proposition 2, we have $\gamma_s = \{W\}_s = \{U_\lambda\} = \{A\}$. If $K$ is the Cartan-Killing form on $g$, $K(B, C) = \text{tr}(ad_B ad_C)$, then by $Ad$-invariance of $K$, $K(\gamma_s, \gamma_s) = K(A, A) = \text{constant}$. In fact, $K(A, A) = \text{tr}((ad_A)^2) = -d$, where $d = \dim(m)$. Therefore, $\gamma$ will be an arclength-parameterized curve in $g$ with respect to the rescaled form $\langle.,.\rangle = -\frac{1}{d}K$. Henceforth, we refer to any curve in the one parameter family $\gamma(s, \lambda)$ as a Sym-Pohlmeyer curve, and denote by $T$ the unit tangent vector $T = \gamma_s = \{A\}$.
To develop a Frenet theory for such curves, we first use Proposition 1 to obtain an expression for the curvature normal $\kappa N$ of a Sym-Pohlmeyer curve,

$$\kappa N = T_s = \{ A \}_s = \{ A_s \}_s + \{ A, \lambda A + Q \} = \{ A, Q \} = \{ \dot{Q} \}.$$  

We refer to the $m$-field $Q$ itself as a curvature coefficients vector. Next, we fix a basis for $g$ of the form $A_1 = A, A_2, \ldots, A_c, B_1, \ldots, B_d$, where the $A_i$ span $k$ and the $B_j$ span $m$. Since the Killing form $K$ is definite ($g$ is compact), we can further specify the basis to be orthonormal with respect to $\langle , \rangle$. The curvature normal vector is now expressible as

$$\kappa N = \left\{ \sum_{j=1}^{d} \kappa_j B_j \right\}_s = \sum_{j=1}^{d} \kappa_j N_j$$

where $N_j = \{ B_j \}$.

Next, we write the derivatives of the $N_j$,

$$(N_j)_s = \{ B_j \}_s = \{ B_j, U \} = \{ B_j, \lambda A + Q \},$$

as a linear combination of themselves and the vectors $T_i = \{ A_i \}, i = 1, \ldots, c$. Finally, we write the derivatives

$$(T_i)_s = \{ A_i \}_s = \{ A_i, Q \}, i = 2, \ldots, c.$$  

At this point, we have a closed system of Frenet equations for the $c + d = \dim(g)$ frame vectors $T_i, N_j$, involving only the curvature functions $\kappa_j$, the spectral parameter $\lambda$, and the structure constants of $g$.

**Running example.**

For Sym-Pohlmeyer curves in $(su(2); -\frac{1}{2}K) \cong (R^3; \langle , \rangle)$ with curvature vector $Q = bB + cC$, the curvature normal is given by $T_s = \kappa N = -c\{ B \} + b\{ C \} = \kappa_1 N_1 + \kappa_2 N_2$, with $\kappa_1 = -c, \kappa_2 = b, N_1 = \{ B \}, N_2 = \{ C \}$. Our Frenet system is completed by the two equations $(N_1)_s = -\kappa_1 T - \lambda N_2$, and $(N_2)_s = -\kappa_2 T + \lambda N_1$. For $\lambda = 0$, this is none other than the natural Frenet system for curves in $R^3$ (see, e.g., [Bis]). For a general value of the constant $\lambda$, such a system may be thought of as inertial, in that the rigid body defined by $\{ T, N_1, N_2 \}$ (identifying $s$ with time) has constant tangential component of angular velocity. The relationship to the classical Frenet system can be written $\kappa_1 + i\kappa_2 = \kappa e^{i\theta}$, and $N_1 + iN_2 = (N + iB)e^{i\theta}$, where $\theta = \int^s \tau(u) + \lambda du$; also, $\kappa^2 = \kappa_1^2 + \kappa_2^2$ and $\tau = \kappa^{-2}(\kappa_1(\kappa_2)_s - \kappa_2(\kappa_1)_s) - \lambda$. While $\kappa, \tau$ and $\{ T, N, B \}$ are uniquely defined along a regular space curve $\gamma$ (with $\kappa \neq 0$), the curvatures $\kappa_1, \kappa_2$ and frame vectors $N_1, N_2$ are determined (given $\lambda$) only up to multiplication by a complex unit, $e^{i\alpha}$ – this freedom corresponds to the choice of antiderivative in the above formulas. Aside from this difference, the natural Frenet theory resulting from these definitions is similar to the classical Fundamental Theorem for space curves. In particular, the set of unit speed curves $\Gamma = \{ \gamma : R \rightarrow R^3 \}$ can be parametrized.
by the following data: *initial position* $\gamma(0)$, *initial frame* $T(0)$, $N_1(0)$, $N_2(0)$, and *shape* $\kappa_1(s), \kappa_2(s)$. For a given curve $\gamma$, this data is unique up to choice of arclength parameter $s$, real parameter $\lambda$, and $S^1$-parameter $e^{i\theta}$.

How does the Sym-Pohlmeyer construction fit together with the above parametrization of $\Gamma$? Since $Q = \kappa_2 B - \kappa_1 C$ and $\lambda$ are explicitly part of the construction, it suffices to discuss the initial data $\gamma(0)$ and $T(0), N_1(0), N_2(0)$. Writing $T = \text{Ad}_{\phi^{-1}} A$, $N_1 = \text{Ad}_{\phi^{-1}} B$, $N_2 = \text{Ad}_{\phi^{-1}} C$, we see that the initial frame is determined by the initial condition on $\phi$, via the adjoint representation of $SU(2)$.

In fact, the two-to-one homomorphism $\text{Ad} : SU(2) \rightarrow SO(3)$ implies all initial frames are achieved (twice) as $\phi(0)$ varies over $SU(2)$. Next, allow the initial condition on $\phi$ to depend on $\lambda$, and regard $\phi(0, \lambda)$ as an arbitrary curve in $SU(2)$. Since $\phi^{-1} \phi_\lambda$ describes the usual trivialization of the tangent bundle $TSU(2)$, it follows that $\gamma(0, \lambda_0) = \phi^{-1} \phi_\lambda|_{(0, \lambda_0)}$ is an arbitrary point in $su(2) \cong R^3$. In conclusion, the Sym-Pohlmeyer curves are precisely the unit speed curves in $R^3$, and the correspondence between the Sym-Pohlmeyer construction and $\Gamma$ is fully described.

Remarks.

1) The above example is prototypical in some, but not all respects. In general, Sym-Pohlmeyer curves constitute a very special subclass of the unit speed curves in $g \cong R^{c+d} -$ the latter cannot all be described by only $d$ curvatures. In fact, the tangent indicatrix $T(s)$ of a regular curve in $R^{c+d}$ can be any smooth curve in the unit sphere $S^{c+d-1} \subset R^{c+d}$, whereas a Sym-Pohlmeyer curve has tangent of the form $T = \{ A \} = \text{Ad}_{\phi^{-1}} A$. Now the $Ad$-orbit of $A$ can be identified with the Hermitian symmetric space $G/K$ ($K$ having Lie algebra $k$). Thus, the tangent indicatrix of a Sym-Pohlmeyer curve lies in $G/K \subset S^{c+d-1} \subset g$. In special cases, the above procedure may produce a closed system with fewer than $(c+d)$ frame vectors - this will be true of our main construction of Section 5 - and the situation may resemble the example more closely.

2) In the general case, it is reasonable to refer to $k_1, \ldots, k_d$ as natural curvatures (though this term will have a more special meaning in Section 5). Note that the non-uniqueness of natural curvatures is described by the group $K$ ($SU(1) = \{ e^{i\theta} \}$ in the example). Specifically, suppose $\phi_s = (\lambda A + Q)\phi$, and consider the Sym-Pohlmeyer curve $\gamma = \phi^{-1} \phi_\lambda$. Now let $\varphi = \phi_0 \phi$, where $\phi_0 \in K$ is any constant element. Then $\gamma = \varphi^{-1} \varphi_\lambda$, and $\varphi$ satisfies the linear system $\varphi_s = (\lambda A + \text{Ad}_{\phi_0} Q) \varphi$, as is easily checked. So the “gauge transformation” $\phi \mapsto \varphi = \phi_0 \phi$ leaves the curve $\gamma$ unchanged, while transforming the natural curvatures according to $Q \mapsto \text{Ad}_{\phi_0} Q$. 
4. The recursion operator and variation formulas.

Next we “un-freeze” time, and apply the above constructions to time-dependent potentials $Q(s,t)$, obtaining two-parameter families of unit speed curves $\gamma(s,t,\lambda)$. The $t$-derivatives of these will be called Sym-Pohlmeyer (variation) fields. Note that Proposition 2 gives a formula for such vectorfields:

$$\gamma_t = \{W\}_t = \{V\}_\lambda.$$

Also, $V$ satisfies the zero curvature equation, $V_s = U_t + [U, V] = Q_t + [\lambda A + Q, V]$. Comparing $k$-components gives $(V_k)_s = [Q, (V_m)_m]$, i.e., the $k$-component of $V$ is determined by the $m$-component of $V$, according to $V_k = \partial_s^{-1}[Q, V_m]$. Differentiation of this equation by $\lambda$ shows that, similarly, $$(V_\lambda)_k = \partial_s^{-1}[Q, (V_\lambda)_m].$$

It is convenient to introduce an operator $K$ which takes $m$-fields to $k$-fields:

$$K(B_m) = \partial_s^{-1}[Q, B_m].$$

Thus any Sym-Pohlmeyer field is of the form $Y = \{K(B_m) + B_m\}$, for some $m$-field $B_m$ (modulo integration constant in $K$). Now the above definitions easily imply the following formula for the arclength derivative of such a Sym-Pohlmeyer field:

$$Y_s = \{(B_m)_s\} + \{K(B_m), Q\} + \lambda\{B_m, A\} = \{C_m\},$$

where $C_m$ is an $m$-field.

**Remark.** The result just obtained has the following (partial) interpretation in the context of curve geometry (in a Riemannian manifold). Suppose $\gamma(s,t)$ is any one-parameter family of arclength parametrized curves, and let $X$ be the vectorfield $X = \partial_t \gamma$. Then $X_s$ (the covariant derivative of $X$ with respect to the unit tangent $T$) has no tangential component; in fact, the condition for a vectorfield $X$ to be locally arclength preserving ([L-P2]) is $\langle X_s, T \rangle = 0$. Of course, a Sym-Pohlmeyer field satisfies this condition: $\langle T, Y_s \rangle = \langle \{A\}, \{C_m\} \rangle = \langle A, C_m \rangle = 0$, since $k$ and $m$ are orthogonal with respect to the Killing form. In the special case $c = 1$, the Sym-Pohlmeyer fields are exactly the locally arclength preserving vectorfields, while for $c > 1$, the Sym-Pohlmeyer vectorfields form a strictly smaller class of vectorfields.

For a Sym-Pohlmeyer curve $\gamma$ in a Hermitian symmetric Lie algebra $g$, we now define three operators on vectorfields $Y = \{B\}$ along $\gamma$.

(i) renormalization operator:

$$\mathcal{P}\{B\} = \{K(B_m) + B_m\} = \partial_s^{-1}[Q, B_m] + B_m;$$

(ii) geometric recursion operator:

$$\mathcal{R}Y = -\mathcal{P}(\{T, \partial_s Y\});$$
(iii) intertwining operator:
\[ Z(Y) = ad_A(\text{Ad}_\phi Y). \]
The next lemma explains the nomenclature for \( Z \):

**Proposition 3.** For \( Y \) a Sym-Pohlmeyer field,
\[ Z \mathcal{R} Y = (\tilde{\mathcal{R}} - \lambda) ZY. \]

**Proof.** Using the above computation of \( Y_s \), we have
\[
Z \mathcal{R} Y = -ZP([T, Y_s])
\]
\[
= -Z\{A, (B_m)_s + [\mathcal{K}(B_m), Q] + \lambda [B_m, A]\}
\]
\[
= -Z\{J((B_m)_s + [\mathcal{K}(B_m), Q] + \lambda B_m)\}
\]
\[
= -J(J((B_m)_s - ad_Q\partial_s^{-1}ad_QB_m) + \lambda B_m)
\]
\[
= (\partial_s - ad_Q\partial_s^{-1}ad_Q)B_m - \lambda J B_m
\]
\[
= -(\partial_s - ad_Q\partial_s^{-1}ad_Q)J^2 B_m - J \lambda B_m
\]
\[
= (\tilde{\mathcal{R}} - \lambda) J B_m = (\tilde{\mathcal{R}} - \lambda) ZY.
\]

Next, we consider a Sym-Pohlmeyer variation, \( \gamma(s, t, \lambda) = \phi^{-1}_s \phi_\lambda = \{W\} \), and the corresponding Sym-Pohlmeyer field (infinitesimal variation) \( X = \gamma_t = \{V\} \).

**Proposition 4.** \( \mathcal{R} X = \{V + \tilde{A}\} \), for some constant \( \tilde{A} \in \mathbb{k} \).

**Proof.**
\[
\mathcal{R} X = -\mathcal{P}\left(\left[ T, \frac{\partial}{\partial s} \frac{\partial}{\partial t} \gamma \right]\right) = -\mathcal{P}([T, T_t])
\]
\[
= -\mathcal{P}([T, \{A\}_t]) = -\mathcal{P}([T, \{A, V\}])
\]
\[
= -\mathcal{P}(\{A, [A, V]\}) = -\mathcal{P}(\{J^2 V\})
\]
\[
= \mathcal{P}(\{V_m\}) = \{\mathcal{K}(V_m) + V_m\} = \{V + \tilde{A}\}.
\]

This last step uses \( \mathcal{K}(V_m) = V_m + \tilde{A} \), as observed above, with the arbitrary “integration constant” \( \tilde{A} \in \mathbb{k} \) explicitly displayed here.

**Theorem 1.** Variation of curvatures formula:
The time variation of the “curvature coefficients vector” \( Q \) induced by a Sym-Pohlmeyer field \( X = \gamma_t = \{V\} \) is given by
\[ Q_t = Z \mathcal{R}^2 X + [Q, \tilde{A}]. \]

In the “gauge term”, \( [Q, \tilde{A}] \), \( \tilde{A} \in \mathbb{k} \) is a constant.

**Proof.** Combining Propositions 2 and 4, we have
\[
Z \mathcal{R}^2 (\gamma_t) = Z \mathcal{R} (\{V + \tilde{A}\}) = -Z \mathcal{P}([T, \{V + \tilde{A}\}_s])
\]
\[
= -Z \mathcal{P}([T, \{Q_t + \tilde{A}, Q\}]) = -Z \mathcal{P}(\{A, Q_t + [\tilde{A}, Q]\})
\]
\[-(ad_{\hat{A}})^2(Q_t + [\hat{A}, Q]) = Q_t + [\hat{A}, Q].\]

The term $[Q, \hat{A}]$ can be interpreted as follows. As explained in Remark 2 of the previous section, the non-uniqueness of natural curvatures for a given curve corresponds to the set of transformations $Q \mapsto \text{Ad}_{\phi_0}Q$, where $\phi_0 \in K$ is a constant. For a curve evolving in time $t$, $\phi_0$ should be treated as a function of $t$ as well (with initial value $\phi_0|_{t=0} = \text{Id}$). Differentiation of $\phi_0$ with respect to $t$ results in the term $[(\phi_0)_t|_{t=0}, Q]$ in the infinitesimal variation of $\text{Ad}_{\phi_0}Q$.

The appearance of the square of the recursion operator in this formula suggests that between the curve $\gamma$ and curvature coefficients vector $Q$, there is an intermediate object whose variation ought to be considered in this context. The appropriate intermediate object is a Sym-Pohlmeyer frame $T_i = \{K_i\}$, $i = 1, \ldots, c$, $N_j = \{M_j\}$, $j = 1, \ldots, d$, as considered above.

**Theorem 2. Variation of Frames formula:**

If $B \in \mathfrak{g}$ is constant, then the time variation of $F = \{B\}$ induced by a Sym-Pohlmeyer field $X = \gamma_t$ is given by

$$F_t = [F, R X] + \{\hat{A}, B\},$$

for some constant $\hat{A} \in \mathfrak{k}$; i.e., $R X$ is essentially the “Darboux vector” for any Sym-Pohlmeyer frame along $\gamma$.

**Proof.** Using Propositions 1 and 4, we compute

$$F_t = \{B, V\} = [(B), \{V\}] = [F, R X] + \{\hat{A}, B\}.$$

We are now in a position to geometrize the Fordy-Kulish NLS hierarchy, the first few terms of which we list here for convenience:

$$\begin{align*}
\hat{X}^{(1)} &= JQ, \\
\hat{X}^{(2)} &= Q_s, \\
\hat{X}^{(3)} &= -JQ_{ss} - \frac{1}{2} (ad_{Q_s})^2 A, \\
\hat{X}^{(n+1)} &= R X^{(n)}.
\end{align*}$$

For a Sym-Pohlmeyer curve $\gamma$ with curvature vector $Q$ and with $\lambda = 0$, let vector fields $X^n$ be defined along $\gamma$ according to:

$$\begin{align*}
X^{(0)} &= \{A\} = T, \\
X^{(1)} &= \{Q\} = -[\gamma_s, \gamma_{ss}], \\
X^{(2)} &= -\left\{ -\frac{1}{2} [Q, [Q, A]] + [A, Q_s] \right\} = -\left( \gamma_{sss} + \frac{3}{2} [\gamma_{ss}, [\gamma_s, \gamma_{ss}]] \right), \\
X^{(n+1)} &= R X^{(n)}.
\end{align*}$$
Now consider the curve evolution equation $\gamma_t = X^{(n)}$. By Theorem 1 and Proposition 3, we can write the corresponding curvature evolution as

$$Q_t = Z\mathcal{R}^2 X^{(n)} + [Q, \tilde{A}] = \mathcal{Z}\mathcal{R}^{(n+1)}X^{(1)} + [Q, \tilde{A}]$$

$$= \tilde{\mathcal{R}}^{(n+1)} Z X^{(1)} + [Q, \tilde{A}] = \tilde{\mathcal{R}}^{(n+1)} \tilde{X}^{(1)} + [Q, \tilde{A}]$$

$$= \tilde{X}^{(n+2)} + [Q, \tilde{A}].$$

We summarize this result (suppressing the gauge term, $[Q, \tilde{A}]$) as:

**Theorem 3.** Evolution of a Sym-Pohlmeyer curve (with $\lambda = 0$) by $\gamma_t = X^{(n)}$ corresponds to curvature evolution by $Q_t = \tilde{X}^{(n+2)}$. In particular, the generalized LIE, $(g\text{LIE})$

$$\gamma_t = -[\gamma_s, \gamma_{ss}],$$

corresponds to the curvature evolution by $g\text{NLS}$, $Q_t = -JQ_{ss} - \frac{1}{2}ad_Q^3 A$ (the analogue of Hasimoto’s result).

Running example.

The geometric recursion operator for curves in $\mathbb{R}^3$ can be written: $\mathcal{R}X = -\mathcal{P}(T \times \frac{\partial}{\partial s} X)$. Here, $\times$ is the cross product in $\mathbb{R}^3$, and the reparameterization operator $\mathcal{P}$ turns an arbitrary vectorfield along $\gamma$, $Y = fT + gU + hV$, into a locally arclength preserving vectorfield, $\mathcal{P}Y = \int (\kappa_1 g + \kappa_2 h) ds T + gU + hV$, by changing only the tangential component. Using the identification of $\mathfrak{m}$ with the complex plane, the operator $\mathcal{Z}$ may be regarded as a simple isomorphism between normal vectorfields $Y = gN_1 + hN_2$ and complex functions $\mathcal{Z}(Y) = i(g + ih)$. On the other hand, $Q$ has already been identified with the complex function $\psi = b + ic = \kappa_2 - i\kappa_1$. Using these definitions, the infinitesimal variation of $\psi$ induced by the vectorfield $X = \gamma_t$ may be written as: $\frac{\partial}{\partial t} \psi = Z\mathcal{R}^2 X + ir\psi$, $r$ a real constant. This differential formula for the Hasimoto transformation easily implies, e.g., that if $\gamma(s, t)$ evolves by LIE, $\frac{\partial}{\partial t} \gamma = \kappa B = -\kappa_2 N_1 + \kappa_1 N_2$, then $\psi$ (with $\lambda = 0$) evolves according to NLS. [There is a minor difference between the formulas discussed here and those of [L-P1, L-P2, L-P3]. In those references, $\psi$ was the “complex curvature function” $\kappa_1 + i\kappa_2 = i(b + ic)$ mentioned in the introduction. The differential formula was written in terms of this $\psi$ (with a minor difference in the definition of $\mathcal{Z}$); if $\frac{\partial}{\partial \tau} \gamma = \kappa B$, then NLS is satisfied by the latter $\psi$ as well (NLS being $i$-equivariant).]

5. Evolution of curves in $\mathbb{R}^{p+1}$ and $S^p$.

We have seen that the geometric realization of the Fordy-Kulish NLS hierarchy is a sequence of evolution equations on the space of Sym-Pohlmeyer curves in a (real compact) Lie algebra $\mathfrak{g}$. The Sym-Pohlmeyer curves have curvature vectors $Q$ which are $\mathfrak{m}$-valued. As stated in Section 3, Sym-Pohlmeyer curves in general form a proper subset of the set of all arc-length
parameterized curves in the Lie algebra $g$. In this section we describe a specific instance of our constructions which allows for a more complete analysis and full geometric interpretation. We will consider the Lie algebra $g = \text{so}(p + 2)$ with subalgebra $k = \text{so}(p) \oplus \text{so}(2)$ corresponding to the Hermitian symmetric space $BDI$. After describing the relevant structure and commutation relations in appropriate detail, we give an explicit formula for the the operator $\tilde{R}^2$ restricted to a distinguished subspace of $m$-fields. We then show that the Sym-Pollmann curves associated with appropriately restricted curvature functions can be naturally considered as corresponding to all curves in the Euclidean space $\mathbb{R}^{p+1}$, and the geometric realizations of the terms in an associated mKdV hierarchy appear as quite natural evolution equations on curves. We explicitly compute the first non-trivial term, and show that it induces curvature evolution corresponding to a particularly simple coupled mKdV system. This mKdV system is a rather special reduction of a system which fits into the general framework of $[F-K], [A-F]$ (though $BDI$ is an exceptional case in the framework of $[A-F]$).

We consider $\text{so}(p+2)$ lying in $\text{gl}(p+2, R)$. We have $\text{gl}(p+2, R)$ commutation relations $[e_{j,k}, e_{l,m}] = \delta_{k,l} e_{j,m} - \delta_{m,j} e_{l,k}, 1 \leq j, k, l, m \leq p + 2$, where $e_{j,k}$ is the matrix with 1 in the $j$th row, $k$th column, zero otherwise. Setting $f_{i,j} = e_{i,j} - e_{j,i}$, we can express $\text{so}(p+2)$ commutation relations in the form $[f_{j,k}, f_{l,m}] = \delta_{j,m} f_{k,l} + \delta_{k,l} f_{j,m} - \delta_{j,l} f_{k,m} - \delta_{k,m} f_{j,l}$.

As it turns out, in addition to the natural notation for the $\text{so}(p+2)$ basis, $\{f_{i,j}\}, 1 \leq i < j \leq p + 2$, it will be convenient to have a notation adapted to a particular decomposition of $\text{so}(p+2)$; thus we define

$$A = f_{1,2},$$
$$X_j = f_{1,j+2}, \; j = 1, \ldots, p,$$
$$Y_k = f_{k+2,2}, \; k = 1, \ldots, p,$$
$$K_{m,n} = f_{m+2,n+2}, \; m, n = 1, \ldots, p.$$ 

The $\text{so}(p+2)$ commutation relations now take the form:

$$[A, X_j] = Y_j, \; [A, Y_j] = -X_j, \; [A, K_{m,n}] = 0,$$
$$[X_j, Y_k] = \delta_{j,k} A,$$
$$[X_j, K_{m,n}] = \delta_{j,m} X_n - \delta_{j,n} X_m,$$
$$[Y_j, K_{m,n}] = \delta_{j,m} Y_n - \delta_{j,n} Y_m,$$
$$[K_{j,k}, K_{l,m}] = \delta_{j,m} K_{k,l} + \delta_{k,l} K_{j,m} - \delta_{j,l} K_{k,m} - \delta_{k,m} K_{j,l}.$$ 

Now consider the following subspaces of $g = \text{so}(p+2)$:

$$k = \text{span}\{A\} \oplus \text{span}\{K_{m,n}\},$$
$$m_x = \text{span}\{X_j\}, \; m_y = \text{span}\{Y_k\},$$
$$m_z = \text{span}\{\}.$$
and

\[ m = m_x \oplus m_y. \]

Part of the structure implicit in these definitions is summarized in

**Proposition 5.**

i) \( g = k \oplus m, \) and \( k = \text{so}(2) \oplus \text{so}(p); \)

ii) \( k \) is the commutator subalgebra of \( A \) in \( g; \)

iii) \( J = ad_A|m \) satisfies \( J^2 = -I. \)

In particular, \( g \) admits a Hermitian symmetric Lie algebra structure (as defined in Section 2).

**Remark.** The proposition does not fully capture all the relevant structure of \( g \) for the geometric considerations to follow. In this connection it should be noted that the same Hermitian symmetric Lie algebra \( g = k \oplus m \) arises as a byproduct of the standard construction of \( \text{so}(p+2) \) as compact real form of \( \text{so}(p+2, C) \). However, a different \( X,Y \)-decomposition of \( m \) appears, which lacks the required properties; specifically, the \( X,Y \)-bracket relations are not as simple as above.

Next, recall the recursion operator \( \hat{R} = \hat{R}_Q \) (first introduced in Section 2) which takes an \( m \)-field \( \hat{X} \) to an \( m \)-field \( \hat{R}\hat{X} \) (and which depends on the \( m \)-potential \( Q \)). Henceforth, we adopt the following:

**Specialization.** \( Q \) is an \( m_x \)-valued potential (more briefly, an \( m_x \)-potential) and \( \hat{X} \) is an \( m_x \)-field.

The fact that this specialization is preserved by \( \hat{R}^2 \) is an immediate consequence of the following:

**Proposition 6.** Let \( Q = \sum_k u_k(s)X_k, \) and \( \hat{X} = \sum_m x_m(s)X_m. \) Then

i) \( \hat{R}\hat{X} = -\sum_k \left( \partial_s x_k + u_k \sum_l \partial_s^{-1}(u_l x_l) \right) Y_k; \)

ii) \( \hat{R}^2\hat{X} = -\sum_k \left( \partial_s^2 x_k + \sum_l (\partial_s(u_k \partial_s^{-1}(u_l x_l)) \right. \right.

\[ + u_l \partial_s^{-1}(u_l \partial_s x_k - u_k \partial_s x_l)) \right) X_k. \]

**Proof.** The proof is by straightforward computation; however, we include it, since (i) depends on the nice bracket formula \([X_j,Y_k] = \delta_{j,k}A,\) and (ii) involves a noteworthy cancellation. Consider \( \hat{R}\hat{X} = (-\partial_s J + ad_Q \partial_s^{-1}ad_Q J)\hat{X}. \) The first term, \(-\partial_s J\hat{X},\) can immediately be written as \(-\sum_k \partial_s x_k Y_k.\) The second term can be rewritten as

\[ ad_Q \partial_s^{-1}ad_Q J \hat{X} = \sum_{k,l,m} u_k \partial_s^{-1}(u_l x_m)[X_k,[X_l,Y_m]] \]

\[ = \sum_{k,l,m} u_k \delta_{l,m}A \]
\[
\begin{align*}
&= -\sum_{k,l,m} u_k \partial_s^{-1}(u_l x_m) \delta_{l,m} Y_k \\
&= -\sum_k u_k \sum_l \partial_s^{-1}(u_l x_l) Y_k .
\end{align*}
\]
Summing these two terms gives the desired formula (i).

To prove (ii), write \( \tilde{R} \tilde{R} \tilde{X} = (-\partial_s J + ad_Q \partial_s^{-1} ad_Q J) \tilde{R} \tilde{X} \), and note that the first term is \(-\partial_s J \tilde{R} \tilde{X} = -\partial_s (\sum_k (\partial_s x_k + u_k \sum_l \partial_s^{-1}(u_l x_l)) X_k) \); this accounts for the first two terms of (ii). Next, using \([X_i, [X_j, X_k]] = [X_i, -K_{j,k}] = \delta_{i,k} X_j - \delta_{i,j} X_k \), one computes
\[
\begin{align*}
ad_Q \partial_s^{-1} ad_Q J \tilde{R} \tilde{X} \\
&= \sum_{i,j,k} u_i \partial_s^{-1} \left( u_j \left( \partial_s x_k + u_k \sum_r \partial_s^{-1}(u_r x_r) \right) \right) (\delta_{i,k} X_j - \delta_{i,j} X_k) \\
&= \sum_{i,j} u_i \partial_s^{-1} \left( u_j \left( \partial_s x_l + u_l \sum_r \partial_s^{-1}(u_r x_r) \right) \right) X_j \\
&\quad - \sum_{i,k} u_i \partial_s^{-1} \left( u_k \left( \partial_s x_k + u_k \sum_r \partial_s^{-1}(u_r x_r) \right) \right) X_k \\
&= \sum_{i,k} u_i \partial_s^{-1} \left( u_k \left( \partial_s x_l + u_l \sum_r \partial_s^{-1}(u_r x_r) \right) \right) - u_l \left( \partial_s x_k + u_k \sum_r \partial_s^{-1}(u_r x_r) \right) X_k \\
&= -\sum_{i,k} u_i \partial_s^{-1}(u_i \partial_s x_k - u_k \partial_s x_l) X_k ,
\end{align*}
\]
which is the last term of formula (ii).

We return now to the Fordy-Kulish NLS hierarchy. The second evolution equation, \( Q_t = \tilde{X}^{(2)} = Q_s \) evidently preserves the space of \( m_x \)-potentials. In fact, writing \( Q = \sum_i u_i(s) X_i \), the resulting evolution for the components \( u_i \) is given by \( (u_i)_t = (u_i)_s \).

From the proposition, it now follows that the evolution \( Q_t = \tilde{X}^{(4)} = \tilde{R}^2 \tilde{X}^{(2)} \) also preserves the space of \( m_x \)-potentials, and the same is true for all of the even evolution equations \( Q_t = \tilde{X}^{(2n)} \). In particular, we can apply formula (ii) to the \( m_x \)-field \( \tilde{X}^{(2)} = \sum_i(u_i)_s X_i \), obtaining
\[
\tilde{X}^{(4)} = -\sum_k \left( \partial_s^2 u_k + \sum_l (\partial_s(u_k \partial_s^{-1}(u_l \partial_s u_l))) \right)
\]
\[ + u_l \partial_s^{-1} (u_l \partial_s^2 u_k - u_k \partial_s^2 u_l)) \right) X_k. \]

Substituting \( \frac{1}{2} u_l^2 \) for \( \partial_s^{-1} (u_l \partial_s u_l) \), and \( u_l \partial_s u_k - u_k \partial_s u_l \) for \( \partial_s^{-1} (u_l \partial_s^2 u_k - u_k \partial_s^2 u_l) \), we obtain a simpler expression for \( \tilde{X}^{(4)} \):

\[
\tilde{X}^{(4)} = - \sum_k \left( \partial_s^3 u_k + \frac{3}{2} \sum_l u_l^2 \partial_s u_k \right) X_k.
\]

It follows that the evolution equation \( Q_t = \tilde{X}^{(4)} \) in terms of the components \( u_i \) is a modified Korteweg-deVries system:

\[(mKdVS) \quad (u_i)_t = - \left( \partial_s^3 u_i + \frac{3}{2} \sum_l u_l^2 \partial_s u_i \right), \quad i = 1, \ldots, p.\]

**Remark.** So far, we have seen that the space of \( m_x \)-fields (along \( m_x \)-potentials) is preserved by the operator \( \tilde{R}^2 \); as a consequence, the even terms in the NLS hierarchy induce evolution equations on the space of \( m_x \)-potentials. It turns out that a corresponding result holds in the general setting of Hermitian symmetric Lie algebras, with the \( X, Y \)-decomposition of \( m \) mentioned in the previous remark. For reasons already given, one cannot generally expect such simple formulas and equations corresponding to those just presented. But from our point of view, the most important difference with the present case shows up in the construction of Sym-Pohlmeyer curves from curvature data \( Q \). Note the key role of the \( X, Y \)-bracket relations in the following

**Proposition 7.** Consider an \( m_x \)-potential \( Q = \sum_i u_i X_i \) with associated Sym-Pohlmeyer curve \( \gamma \) (with \( \lambda = 0 \)). The unit tangent vector field \( T = \{A\} \) and vector fields \( U_i = \{Y_i\}, \quad i = 1, \ldots, p \), defined along \( \gamma \), satisfy the following closed, linear system:

\[
T_s = \sum_i u_i U_i, \quad (U_i)_s = -u_i T.
\]

**Proof.** We have

\[
T_s = \{A\}_s = \{A_s\} + \{[A, Q]\} = \left\{ \left[ A, \sum_i u_i X_i \right] \right\} = \sum_i u_i \{Y_i\} = \sum_i u_i U_i; \quad (U_i)_s = \{Y_i\}_s = \{(Y_i)_s\} + \{[Y_i, Q]\}
\]
\[
\begin{align*}
&= \left\{ Y_i, \sum_j u_j X_j \right\} \\
&= \sum_j u_j \{[Y_i, X_j]\} \\
&= -\sum_j u_j \delta_{i,j} \{A\} = -u_i \{A\} = -u_i T.
\end{align*}
\]

Let \( \Psi \) be a \((p+1) \times (p+1)\) matrix which is the fundamental solution to the matrix differential equation

\[
\Psi_s = \begin{pmatrix}
0 & u_1 & \cdots & u_p \\
-u_1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
-u_p & 0 & \cdots & 0
\end{pmatrix} \Psi,
\]

\[\Psi(s_0) = I_{p+1} = (p+1) \times (p+1) \text{ identity matrix}.\]

Then by the fundamental theorem of differential equations, we can express the moving frame \(T(s), U_1(s), \ldots, U_p(s)\) as

\[
\begin{pmatrix}
T(s) \\
U_1(s) \\
\vdots \\
U_p(s)
\end{pmatrix} = \Psi
\begin{pmatrix}
T(s_0) \\
U_1(s_0) \\
\vdots \\
U_p(s_0)
\end{pmatrix}.
\]

It follows that the Sym-Pohlmeyer curve \(\gamma \subset g\) actually lies in the affine space \(\gamma(s_0) + R^{p+1}\), where \(R^{p+1}\) is here identified as the span of \(T(s_0), U_i(s_0), i = 1, \ldots, p\). Moreover, the equations given in the proposition are the natural Frenet equations (see Running example in Section 3) for a curve in \(R^{p+1}\) with curvatures \(u_i(s), i = 1, \ldots, p\) and natural frame \(T, U_i, i = 1, \ldots, p\). Thus, we conclude that the Sym-Pohlmeyer curve \(\gamma\) may be regarded as a general space curve in \(R^{p+1}\).

Since the natural Frenet equations are not so well-known, we take a moment to indicate some of their geometric significance. To begin with, the conclusion just reached depends on an analogue of the classical Fundamental Theorem of Curve Theory (for curves in \(n\)-dimensional Euclidean space). In particular, every smooth curve \(\gamma\) in \(R^{p+1}\) satisfies the above system for some choice of curvature functions \(u_i(s), i = 1, \ldots, p\), and is uniquely determined, up to congruence, by these functions. (The converse statement differs a bit from that of the classical theorem, in that the natural curvatures \(u_i\) are uniquely determined by a curve \(\gamma\) only after the frame \(T(s_0), U_i(s_0)\) has been specified at some initial point \(\gamma(s_0)\).)

Note that \(T_s = \kappa N = \sum_i u_i U_i\),
implies $k^2 = \sum_i (u_i)^2$, so one can recover the standard (first) curvature from natural curvatures.

What’s more important, the curvatures $u_i$ measure the sphericity of a curve in $R^{p+1}$. In particular, suppose that, for some $j$, $u_j = c_j = \text{constant} \neq 0$. Then $(\gamma(s) + (1/c_j)U_j)_s = T + (1/c_j)c_j(-T) = 0$, so $\gamma$ lies on a $p$-dimensional sphere of radius $1/c_j$. Further, $U_j$ is the inward pointing unit normal to the sphere along $\gamma$, and the remaining frame vectors determine a natural Frenet system along the spherical curve $\gamma$ in the sense of covariant differentiation in the sphere. Namely,

$$\nabla_T T = \sum_{i \neq j} u_i U_i,$$

$$\nabla_T U_i = -u_i T, \ i \neq j.$$

More generally, if $u_{i_1} = c_1, u_{i_2} = c_2, \ldots, u_{i_l} = c_l$, then $\gamma$ lies on a $(p+1-l)$-dimensional sphere of radius $(c_1^2 + c_2^2 \ldots c_l^2)^{-1/2}$. In the exceptional case, $u_{i_1} = u_{i_2} = \ldots u_{i_l} = 0$, $\gamma(s)$ lies on a $(p+1-l)$-plane. This corresponds to the case in classical Frenet theory in which the last $l$ curvatures vanish – here the order of the curvatures matters.

As an application of the above discussion, we are now in a position to give a purely geometric version of our earlier variation formula, in the context of Euclidean and spherical curves.

**Theorem 4. Variation of curvatures formula:**

Let $M$ be Euclidean space of dimension $d = (p+1)$, or a round sphere of dimension $d = p$. Denote by $G$ the (constant) scalar curvature of $M$ (so $G = 0$ or $G = \frac{1}{r^2}$ in the case of a sphere of radius $r$). Let $\gamma_t = X = \alpha T + \sum_{i=1}^{d-1} x_i U_i$ describe a variation of a curve in $M$ through unit speed curves, where $U_i, \ i = 1, \ldots, (d-1)$, is a natural frame along $\gamma(s,t)$. Then the induced variation of the associated natural curvatures $u_i$ is given by

$$(u_i)_t = (\partial_s^2 + G)x_i + \partial_s (\alpha u_i) + \sum_{l=1}^{d-1} (u_l \partial_s^{-1}(u_l \partial_s x_i - u_i \partial_s x_l))$$

$$+ \sum_j c_{i,j} u_j, \ i = 1, \ldots, (d-1).$$

In the gauge term $\sum_j c_{i,j} u_j$, the $c_{i,j}$ are constants with $c_{i,j} = -c_{j,i}$.

**Proof.** Combine Theorem 1 with Propositions 3 and 6, after taking account of the above discussion.

We now merge the last few topics and consider the geometric evolution of curves in $R^{p+1}$ and $S^p$ corresponding to the above mKdV system. According to Theorem 3, the curve evolution $\gamma_t = X^{(2)}$ corresponds to the curvature evolution $Q_t = \tilde{X}^{(4)}$. In the present context, a fully geometric interpretation
of this result is possible; note that the following theorem (like the previous one) is formulated entirely in terms of curve geometry – no Lie algebras!

**Theorem 5.** Motion of a curve \( \gamma(s,t) \) in \( \mathbb{R}^{p+1} \) by the geometric evolution equation

\[
\gamma_t = -\left( \frac{3}{2} \kappa^2 T + T_{ss} \right) = -\left( \frac{1}{2} \kappa^2 T + \sum_i (u_i)_s U_i \right),
\]

corresponds to curvature evolution by the mKdV system

\[
(u_i)_t = -\left( \partial^3_s u_i + \frac{3}{2} |u|^2 \partial_s u_i \right), \quad i = 1, \ldots, p.
\]

Here, the functions \( u_i \) are natural curvatures and \( k^2 = |u|^2 = \sum_i (u_i)^2 \) the squared curvature of \( \gamma \). In particular, if one of the natural curvatures \( u_j \) is initially constant along \( \gamma \), then this condition is preserved, and \( \gamma \) evolves on a sphere.

**Proof.** The first statement is easily obtained from the previous theorem, by direct computation. Alternatively, in view of Theorem 3, it suffices to note that in the present context,

\[
X^{(2)} = -\left\{-\frac{1}{2} [Q, [Q, A]] + [A, Q_s] \right\} = -\left\{-\frac{1}{2} \sum_{i,j} u_i u_j [X_i, [X_j, A]] + \sum_i (u_i)_s [A, X_i] \right\} = -\left\{ \frac{1}{2} \sum_i (u_i)^2 A + \sum_i (u_i)_s Y_i \right\} = -\left( \frac{3}{2} \kappa^2 T + T_{ss} \right).
\]

It is also evident from the form of the coupled mKdV equations that the condition \( u_j = \text{constant} \) is preserved in time, so the last statement follows.

**Remark.** In the case \( k \neq 0 \), the vectorfield \( X^{(2)} \) is readily expressed in terms of the standard Frenet frame:

\[
X^{(2)} = -\left( \frac{3}{2} \kappa^2 T + T_{ss} \right) = -\left( \frac{1}{2} \kappa^2 T + \kappa \tau B \right).
\]

Here, \( \tau \) and \( B \) are, respectively, the second curvature and second normal (in three dimensions, the torsion and binormal).

Expressed in this form, this vectorfield appeared in \([L-P1]\) as the “next” term (above \( X^1 = kB \)) in the localized induction hierarchy, and the connection to the (complex) mKdV equation was discussed. The planar version of
this curve evolution and the connection to mKdV were considered in [G-P] and in [L-P3]. In [L-P4], the authors showed that the even terms of the $R^3$ localized induction hierarchy preserve curves lying on the sphere $S^2$, expressing these vectorfields in terms of the natural curvatures and frames. In [D-S], Doliwa and Santini show that $X = \frac{1}{2}\kappa^2 T + \kappa s N + \kappa T B$ describes an evolution of curves on $S^3$, and discuss corresponding curve evolution equations in spheres of arbitrary dimension. Our curve evolution equations (of this section) apparently coincide with those of [D-S] (though our approach is considerably different). On the other hand, the corresponding curvature evolutions of [D-S] are increasing complicated as dimension of the sphere increases, and do not generally bear a close resemblance to the familiar (scalar) mKdV equation; we attribute this to their use of standard Frenet systems, rather than the natural Frenet systems we have employed here. Finally, our general formalism suggests a new perspective on the main conclusions of [D-S], regarding the characterization of integrable curve dynamics (as will be discussed in a future paper).

In this work we have described a Lie-theoretic construction of generalizations of the localized induction hierarchy and, in a special case, have shown how curvature evolution equations related to (but simpler in form than) those of Doliwa and Santini may be extracted as a subhierarchy. These geometric realizations of the Fordy-Kulish NLS systems have a structure remarkably similar to the $R^3$ LIE equations studied previously. Given the known relations between the localized induction hierarchy and classical geometric constructions, it is not unreasonable to expect that the geometric realizations of the Fordy-Kulish NLS systems will have similar interesting relations to geometry.

References


Received September 2, 1998 and revised March 16, 1999.

DEPT. OF MATHEMATICS
CASE WESTERN RESERVE UNIVERSITY
CLEVELAND OH 44106
E-mail address: jxl6@po.cwru.edu

DEPT. OF MATHEMATICS AND COMP. SCI.
DREXEL UNIVERSITY
PHILADELPHIA PA 19104
E-mail address: rperline@mcs.drexel.edu
SEIBERG–WITTEN INVARIANTS FOR 3-MANIFOLDS IN THE CASE $b_1 = 0$ OR $1$

Yuhan Lim

In this note we give a detailed exposition of the Seiberg-Witten invariants for closed oriented 3-manifolds paying particular attention to the case of $b_1 = 0$ and $b_1 = 1$. These are extracted from the moduli space of solutions to the Seiberg-Witten equations which depend on choices of a Riemannian metric on the underlying manifold as well as certain perturbation terms in the equations. In favourable circumstances this moduli space is finite and naturally oriented and we may form the algebraic sum of the points. Given any two sets of choices of metric and perturbation which are connected by a 1-parameter family, we analyse in detail the singularities which may develop in the interpolating moduli space. This leads then to an understanding of how the algebraic sum changes. In the case $b_1 = 0$ a topological invariant can be extracted with the addition of a suitable counter-term, which we identify (this idea is attributed to Donaldson). In the case $b_1 = 1$ a topological invariant is defined which depends only on cohomological information related to the perturbation term. We prove a ‘wall-crossing’ formula which tells us how the invariant changes with different choices of this perturbation. Throughout we pay careful attention to genericity statements and the issue of orientations and signs in all the relations. The equivalence of this invariant in the case of an integral homology sphere with the Casson invariant is treated in Lim, 1999 (see also works of Nicolescu, preprint). The equivalence with Reidemeister Torsion in the case $b_1 > 0$ is a result of Meng & Taubes, 1996. Some related material is in Marcolli, 1996, Froyshov, 1996 and in the survey Donaldson, 1996. Taubes, 1990 contains the originating construction in this article in the context of flat $SU(2)$-connections.

1. The Seiberg-Witten Invariants.

We denote by $Y$ an oriented 3-manifold. Let $g$ be a Riemannian metric on $Y$ and $P \to Y$ a spin-c structure (see for example [LM]). Denote by $S \to Y$ the associated positive spinor bundle, i.e., $S = P \times_\rho \mathbb{C}^2$ where $\rho : \text{spin}_c(3) \to \text{End}_\mathbb{C}(\mathbb{C}^2)$ is the irreducible representation of the complex Clifford algebra.
$\text{Cl}(\mathbb{R}^3)$ with $g(dy) = +1$, where $dy$ is the oriented volume form on $Y$. Let $\xi : \text{spin}_c(3) = \text{spin}(3) \times_{\{\pm 1\}} U(1) \to U(1)$ be the map which takes the square of the second factor. Then the bundle $L = LP = P \times_\xi C \to Y$ is the determinant line bundle of $P$; it is clearly a $U(1)$-bundle. We remark that a classical fact is $w_2(Y) = 0$ and this gives the existence of a spin-c structure $P \to Y$ with $c_1(L)$ any given class in $2H^2(Y; \mathbb{Z}) \subset H^1(Y; \mathbb{Z})$.

For a pair $(A, \Phi)$ consisting of a $U(1)$-connection on $L$ and a section of $S$ the $\pi$-perturbed Seiberg-Witten equations (SW$_\pi$) read:

$$F_A = \frac{1}{4} \sigma(\Phi, \Phi) + \omega, \quad D_{A+\alpha} \Phi = 0.$$ 

Here $F_A$ denotes the curvature 2-form of $A$ and $D_A$ the Dirac operator coupled to $A$. The term $\sigma(\cdot, \cdot)$ is a certain symmetric bilinear form $S \to \Lambda^2(i \mathbb{R})$ (see Section 2 for details). The perturbation $\pi$ consists of a pair $(\alpha, \omega)$ where $\alpha \in \Omega^1(i \mathbb{R})$ and $\omega \in \Omega^2(i \mathbb{R})$, $d\omega = 0$.

$U(1)$ embeds into $\text{spin}_c(3)$ by the map $\iota(z) = [\text{Id}, z]$. Given a smooth map $g : Y \to U(1) \subset \mathbb{C}$, we have an automorphism of $P$ given by the rule $p \mapsto p \iota(g)$. This induces, by pulling back, the action $(A, \Phi) \to g(A, \Phi) = (A + 2g^{-1}dg, g^{-1}\Phi)$. If $(A, \Phi)$ is a Seiberg-Witten solution, then $g(A, \Phi)$ is also a solution. Thus the set of Seiberg-Witten solutions is invariant under the above automorphisms of $S$. The automorphism $g$ called a gauge transformation and the group of all gauge transformations is the gauge group denoted $G$.

We will consider the set of $L^2$-SW$_\pi$-solutions modulo gauge equivalence (details in Sec. 2). The set of such solutions will be denoted by $Z_\pi(\mathbb{P}; g)$ or $Z_\pi(\mathbb{P}; g)$. When the underlying metric is understood we shall omit it and simply write $Z_\pi(\mathbb{P})$. Fixing a value of $k \geq 1$, denote by $\mathcal{P}_k$ the space of all perturbations $\pi$ of class $C^k$.

**Theorem 1.** For $\pi$ from an open dense subset of $\mathcal{P}_k$ the irreducible part $Z^*_\pi(\mathbb{P}; g)$ of $Z_\pi(\mathbb{P}; g)$ is a finite set of points and these are naturally oriented. Let $\#Z^*_\pi(\mathbb{P}; g)$ denote the algebraic sum, assuming $\pi$ as above. Then:

(i) if $b_1(Y) > 1$, $\#Z^*_\pi(\mathbb{P}; g)$ is independent of $g$ and $\pi$

(ii) if $b_1(Y) = 1$, $\#Z^*_\pi(\mathbb{P}; g)$ depends only on the component of $H^2(Y; \mathbb{R}) \setminus \{c_1(L)\}_{\mathbb{R}}$ in which $[\frac{1}{2\pi} \omega]$ lies in, $\omega$ being the 2-form component of $\pi$

(iii) if $b_1(Y) = 0$, $\#Z^*_\pi(\mathbb{P}; g)$ is independent of $\pi$ and $g$ after the addition of a counter-term $\zeta(\pi, g)$ which is a combination of the spectral invariants of Atiyah-Patodi-Singer. $\#Z^*_\pi(\mathbb{P}; g) + \zeta(\pi; g)$ takes values in $\mathbb{Z}$ if $H_1(Y; \mathbb{Z}) = \{0\}$ and $\mathbb{Z}\left[\frac{1}{8[H_1(Y; \mathbb{Z})]}\right]$ if $H_1(Y; \mathbb{Z}) \neq \{0\}$.

The exact expression for the counter-term $\zeta(\pi, g)$ is in Proposition 17. For $b_1(Y) = 1$ the formula for the change in $\#Z_\pi(\mathbb{P}; g)$ when we cross the ‘wall’ in $H^2(Y; \mathbb{R})$ defined by $\{c_1(L)\}_{\mathbb{R}}$ is given in Corollary 20.
Let $\text{spin}_c(Y)$ denote the equivalence classes of spin-c structures on $Y$. There is a well-defined map $\text{spin}_c(Y) \to 2H^2(Y;\mathbb{Z})$ which sends a representative $P$ to $c_1(L_P)$.

**Corollary 2.** Let $Y$ be connected. The Seiberg-Witten equations define an oriented diffeomorphism invariant $\tau$ of $Y$ in the following form:

(i) if $b_1(Y) > 0$, $\tau : \text{spin}_c(Y) \to \mathbb{Z}$

(ii) if $b_1(Y) = 1$, let $\text{spin}_c^*(Y)$ be the set of pairs $([P],\mathcal{U})$ where $\mathcal{U}$ a component of $H^2(Y;\mathbb{R}) \setminus \{c_1(L_P)\mathbb{R}\}$. Then $\tau : \text{spin}_c(Y) \to \mathbb{Z}$

(iii) if $b_1(Y) = 0$ and $H_1(Y;\mathbb{Z}) = \{0\}$ there is a unique spin-c structure and so $\tau \in \mathbb{Z}$

(iv) if $b_1(Y) = 0$ and $H_1(Y;\mathbb{Z}) \neq \{0\}$, then $\tau : \text{spin}_c(Y) \to \mathbb{Z} \left[ \frac{1}{8|H_1(Y;\mathbb{Z})|} \right]$.

In the subsequent sections we work up to a proof of Theorem 1. Section 2 discusses the framework for defining the moduli space and their first properties. Section 3 looks at generic properties. The details of the proof of Theorem 1 are in Section 4.

**Addendum.** The reviewer has brought to the attention of the author of an alternative exposition of some of the material in this article, in [Ch].

## 2. The Moduli Space.

Throughout this section $Y$ denotes a closed oriented 3-manifold with Riemannian metric $g$ and $P \to Y$ a fixed spin-c structure.

### 2.1. The Basic Set-up.

As in the usual gauge theory set-up we work with the following spaces. (For more details see, for instance, [M].) Let $C(P)$ denote the space of pairs $(A,\Phi)$ consisting of a $L^2_S$ connection $A$ on $L$ and $\Phi$ a $L^2_2$-section of $S \to Y$. This forms a Hilbert manifold. Let $\mathcal{G}$ denote the space of $L^2_S$ gauge transformations of $S$ i.e., $L^2_S$ maps $g : Y \to S^3 \subset \mathbb{C}$. This forms a Hilbert Lie group. $\mathcal{G}$ acts on $C(P)$ by $g(A,\Phi) = (A + 2g^{-1}dg, g^{-1}\Phi)$. This action is smooth with Hausdorff quotient $\mathcal{B}(P)$.

A pair $(A,\Phi)$ is *irreducible* if $\Phi$ is not identically 0. Otherwise it is called *reducible*. $\mathcal{G}$ acts freely on $C^*(P)$, the open set of irreducibles and its quotient is denoted by $\mathcal{B}^*(P)$. The projection map $C^*(P) \to \mathcal{B}^*(P)$ forms a principle $\mathcal{G}$-bundle. At a reducible $(A,0)$, for which we simply write $A$, the stabilizer of $\mathcal{G}$ is exactly those gauge transformations $g$ for which $dg = 0$; thus the stabilizer is identified with $U(1) \subset \mathbb{C}$, the constant gauge transformations.

Let $\Omega^0_k(\mathbb{R})$ denotes the $p$-forms on $Y$ of class $L^2_k$, and $\Gamma_k(S)$ the sections of $S$ of class $L^2_k$. Since $C(P)$ is an affine space modelled on the vector space $\Omega^0_2(\mathbb{R}) \times \Gamma_2(S)$ the tangent space at any point is canonically identified with the vector space itself. On the other hand the tangent space to the identity of $\mathcal{G}$ is identified with $\Omega^0_2(\mathbb{R})$. 

The derivative of the map $g \mapsto g(A, \Phi)$ at the identity is given by $\gamma \mapsto (2d\gamma, -\gamma \Phi)$. The tangent bundle of $C(P)$ carries a natural Riemannian metric which is the $L^2$-inner product on $\Omega^2_2(i|R) \times \Gamma_2(S)$. This inner product is invariant under complex multiplication in $\Gamma_2(S)$. By taking the $L^2$-orthogonal to the image of the derivative of the gauge group action we obtain a slice $(A, \Phi) + X_{A, \Phi}$ for the action at $(A, \Phi)$. $X_{A, \Phi}$ is defined as

$$\{(a, \phi) \mid 2d^* a = i \langle \Phi, \phi \rangle \} \subset \Omega^1_2(i|R) \times \Gamma_2(S).$$

If $\Phi = 0$ then $X_{A, \Phi}$ reduces to $\ker d^* \times \Gamma_2(S)$. The stabilizer of $A$ preserves $X_A$ and acts by $z(a, \phi) = (a, z^{-1}\phi)$; therefore the stabilizer acts as the opposite complex structure on $\Gamma_2(S)$. If $\Phi \neq 0$ then a sufficiently small neighbourhood $N$ of zero in $X_{A, \Phi}$ models an open set for the gauge equivalence class $[(A, \Phi)]$ in $B(P)$. If $\Phi = 0$ then the same is true except that we should take $N/U(1)$ instead.

The symmetric bilinear form $\sigma(\Phi, \Psi) \in \Lambda^2(i|R)$ for us will be defined as the adjoint of Clifford multiplication, that is defined by the condition that for all $\omega \in \Lambda^2(i|R)$,

$$\langle \omega \cdot \Psi, \Phi \rangle = \langle \omega, \sigma(\Phi, \Psi) \rangle.$$

The representation $c : \Lambda^2(i|R) \to \text{End}_C(S)$ given by Clifford multiplication is an isomorphism onto its image which is the trace-free Hermitian symmetric endomorphisms of $S$. If we identify $\Lambda^2(i|R)$ with its image under $c$ then

$$\sigma(\Phi, \Psi) = \frac{1}{2} (\Phi \otimes \Psi^* + \Psi \otimes \Phi^* - \langle \Phi, \Psi \rangle R \text{Id}).$$

In the formula expressions of the form $v \otimes w^*$ mean the endomorphism $v \otimes w^*(u) = v(u, w)_C$. We remark that this formula assumes the convention that if $\tau$ is unit length in $\Lambda^2(R)$ then $c(i\tau)$ is to be unit length in $\text{End}_C(S)$.

Fix a perturbation term $\pi = (\alpha, \omega)$ of class $C^k$, $k \geq 1$ (we assume this from now on). To set up the moduli space we define the $\text{SW}_\pi$-section $s = s_{\pi, g} : \Omega^2_2(i|R) \times C(P) \to \Omega^1_2(i|R) \oplus \Gamma_1(S)$ by

$$s(\eta, A, \Phi) = \left( F_A - \frac{1}{4} \sigma(\Phi, \Phi) - \omega \right) + 2d\eta, D_{A+\alpha} \Phi - \eta \Phi \right).$$

Since we will want to vary the perturbation term later, we introduce the the Banach space $Q_k$ of $C^k$-sections of $\Lambda^1(Y) \otimes i|R$ and the Banach space $\Omega_k$ of closed $C^k$-sections of $\Lambda^2(Y) \otimes i|R$. Then our perturbations $\pi$ are from the space $\mathcal{P}_k$ which is $Q_k \times \Omega_k$.

From the definition of the $\text{SW}_\pi$-section, it would seem that the zeros might capture a much larger set than the $\text{SW}_\pi$-solutions themselves; but as the following Lemma shows this is only so in a minor way.

**Lemma 3.** Let $s(\eta, A, \Phi) = 0$. If $\Phi \neq 0$ then $\eta = 0$ and $(A, \Phi)$ is a $\text{SW}_\pi$-solution. If $\Phi = 0$ then $\eta =$ constant and $(A, 0)$ is a $\text{SW}_\pi$-solution.
Proof. We claim that the vector \((F_A - \frac{1}{2}\sigma(\Phi, \Phi) - \omega), D_{A+\alpha}\Phi)\) is \(L^2\)-orthogonal to \(2d\eta, -\eta\Phi\). Before we show this we recall some useful identities: (i) if \(\eta \in \Omega_{A}^k(i\mathbb{R})\) then \(D_{A+\alpha}(\eta\Phi) = d\eta \cdot \Phi + \eta D_{A+\alpha}\Phi\) (ii) the Clifford action of \(a \in \Omega_{A}^1(i\mathbb{R})\) is equal to the action of \(-a\) (iii) if \(\eta \in \Omega_{A}^k(i\mathbb{R})\) then \(\langle \eta\Phi, \Psi \rangle = -\langle \Phi, \eta\Psi \rangle\). To prove the claim we compute:

\[
\langle 2d\eta, *(F_A - \frac{1}{2}\sigma(\Phi, \Phi) - \omega) \rangle_{L^2} - \langle \eta\Phi, D_{A+\alpha}\Phi \rangle_{L^2} \\
= \langle 2d\eta, d^* (F_A - \omega) \rangle_{L^2} - \langle 2d\eta, \frac{1}{2}\sigma(\Phi, \Phi) \rangle_{L^2} - \langle \eta\Phi, D_{A+\alpha}\Phi \rangle_{L^2} \\
= -\frac{1}{2} \langle d\eta \cdot \Phi, \Phi \rangle_{L^2} - \langle \eta\Phi, D_{A+\alpha}\Phi \rangle_{L^2} \\
= \frac{1}{2} \langle D_{A+\alpha}(\eta\Phi), \Phi \rangle_{L^2} - \frac{1}{2} \langle \eta D_{A+\alpha}\Phi, \Phi \rangle_{L^2} - \langle \eta\Phi, D_{A+\alpha}\Phi \rangle_{L^2} \\
= \frac{1}{2} \langle D_{A+\alpha}(\eta\Phi), \Phi \rangle_{L^2} + \frac{1}{2} \langle \eta D_{A+\alpha}\Phi, \eta\Phi \rangle_{L^2} - \langle \eta\Phi, D_{A+\alpha}\Phi \rangle_{L^2} = 0.
\]

The last step follows from \(D_{A+\alpha}\Phi\) being self-adjoint. Therefore \(s(\eta, A, \Phi) = 0\) if and only if \((A, \Phi)\) is a SW\(\pi\)-solution and \((2d\eta, -\eta\Phi) = 0\). \(\square\)

Thus we identify the space of Seiberg-Witten solutions with

\[s^{-1}(0) \cap \{0\} \times \mathcal{C}(P)\]

and the Seiberg-Witten moduli space \(Z_{\pi}(P)\) is the quotient by \(\mathcal{G}\) of this, where \(\mathcal{G}\) acts only on the \(\mathcal{C}(P)\) factor. The local structure of the moduli space near a solution \((\eta, A, \Phi) \in s^{-1}(0)\) is determined by the elliptic complex associated to the map \(s\):

\[
\Omega_{A}^0(i\mathbb{R}) \xrightarrow{\delta_{0, A, \Phi}} \Omega_{A}^2(i\mathbb{R}) \oplus \Omega_{A}^1(i\mathbb{R}) \oplus \Gamma_2(S) \xrightarrow{\delta_{A+\alpha}} \Omega_{A}^1(i\mathbb{R}) \oplus \Gamma_1(S)
\]

where

\[
\delta_{0, A, \Phi}(\gamma) = (0, 2d\gamma, -\gamma\Phi) \\
\delta_{A, A, \Phi}(\xi, a, \phi) = \ast(da - \frac{1}{2}\sigma(\Phi, \phi)) + 2d\xi, D_{A+\alpha}\phi + \frac{1}{2}a \cdot \Phi - \xi\Phi - \eta\phi.
\]

Since we will be interested only in the case where \(\eta = 0\) we will in subsequent notation omit it when that is understood. Thus we let \(H_{A, \Phi}^i\), \((i = 0, 1, 2)\) denote the cohomology of the complex when \(\eta = 0\).

Lemma 4. Let \((A, \Phi)\) be a SW\(\pi\)-solution. 

(i) If \(\Phi \neq 0\) then \(H_{A, \Phi}^0 = 0, H_{A, \Phi}^1 = H_{A, \Phi}^2\)

(ii) If \(\Phi = 0\) then \(H_{A, \Phi}^0 = H_{A, \Phi}^1 = H_{A, \Phi}^2 = H_{A+\alpha, \Phi}^0 = H_{A+\alpha, \Phi}^1 = H_{A+\alpha, \Phi}^2\).

Note. In this article \(H_{A}^k(i\mathbb{R})\) always denotes the pure imaginary harmonic forms of degree \(k\) and \(\mathcal{H}_A\) the kernel of \(D_A\).
Proof. A direct computation shows that the $L^2$-adjoint of $\delta^1_{A,\Phi}$ is given by $\delta^1_{A,\Phi}^*(b,\psi) = (2d^*b - i\langle i\Phi, \psi \rangle, \delta^1_{A,\Phi}(a,\phi))$. The adjoint of $\delta^0_{A,\Phi}$ on the other hand, is $\delta^0_{A,\Phi}(\eta,a,\phi) = 2d^*a - i\langle i\Phi, \phi \rangle$. Therefore

\[ H^1_{A,\Phi} = \ker \delta^0_{A,\Phi} \cap \ker \delta^1_{A,\Phi} \]
\[ = \{ (\xi,a,\phi) \mid 2d^*a - i\langle i\Phi, \phi \rangle = 0, D_{A+\alpha}\phi + \frac{1}{2}a \cdot \Phi = 0, d\xi = 0, \xi \Phi = 0 \} \]

\[ H^2_{A,\Phi} = \ker \delta^1_{A,\Phi} \]
\[ = \{ (b,\psi) \mid 2d^*b - i\langle i\Phi, \psi \rangle = 0, db - \frac{1}{2}\sigma(\Phi,\psi) = 0, \]
\[ D_{A+\alpha}\psi + \frac{1}{2}b \cdot \Phi = 0 \}. \]

The remainder of the proof follows easily. □

The local structure of the moduli space may now be deduced by the Kuranishi argument. Let $(0, A, \Phi) \in s^{-1}(0)$. If $\Phi \neq 0$, then a neighbourhood of $[(0, A, \Phi)]$ in $Z_\pi(P)$ is modelled on the zeros of a map (the obstruction map) $\Xi : H^1_{A,\Phi} \to H^2_{A,\Phi}$. If $\Phi = 0$ then the same is true except that $\Xi$ is $S^1$-equivariant and we should take $\Xi^{-1}(0)/S^1$.

If $(A, \Phi)$ is a regular solution, i.e., $H^2_{A,\Phi} = \{0\}$, and $\Phi \neq 0$ then $\Xi^{-1}(0)$ is exactly one point. Thus a regular irreducible solution is isolated. Therefore we have:

**Proposition 5.** If $Z_*^+(P)$ consists solely of gauge equivalence classes of regular solutions then $Z_*^+(P)$ is a discrete set, i.e., every point is isolated.

2.2. Compactness and Regularity.

**Proposition 6.** Fix $\pi \in \mathcal{P}_k, k \geq 1$.

(i) If $(A, \Phi)$ is a SW$_\pi$-solution then $(A, \Phi)$ is gauge equivalent to a SW$_\pi$-solution of class $L^p_{k+1}, p \geq 2$.

(ii) Let $\{(A_i, \Phi_i)\}_{i=1}^\infty$ be a sequence of SW$_\pi$-solutions.

Then there is a subsequence $\{i'\} \subset \{i\}$ and gauge transformations $\{g_{i'}\}$ such that $\{g_{i'}(A_{i'}, \Phi_{i'})\}$ converges in $L^2_k$ to a $L^2_k$ SW$_\pi$-solution. In particular, this converges in $\mathcal{C}(P)$, and therefore $\mathcal{B}(P)$.

Proof. The proof is due to [KM]. We include it here for completeness. The Dirac operator for the Dirac operator reads (see for instance [LM])

\[ D^*_A + \kappa \Phi = \nabla^*_A + \frac{1}{4}\kappa \Phi + \frac{1}{2}F_A \cdot \Phi, \]
\[ \kappa \text{ being scalar curvature.} \]

We also have Kato’s inequality

\[ \frac{1}{2} \Delta |\Phi|^2 \leq \langle \nabla_A \Phi, \nabla_A \Phi \rangle. \]
If \((A, \Phi)\) is a SW\(_{\alpha,\omega}\)-solution then \(D_{A+\alpha}\Phi = 0\) and
\[
\langle F_{A+\alpha} \cdot \Phi, \Phi \rangle = \frac{1}{4} |\sigma(\Phi, \Phi)|^2 + \langle (\omega + d\alpha) \cdot \Phi, \Phi \rangle
\]
\[
= \frac{1}{8} |\Phi|^4 + \langle (\omega + d\alpha) \cdot \Phi, \Phi \rangle.
\]
Applying Kato’s inequality to the Bochner formula we obtain
\[
\frac{1}{2} \Delta |\Phi|^2 \leq -\frac{1}{4} \kappa |\Phi|^2 - \frac{1}{8} |\Phi|^4 + |\omega + d\alpha| |\Phi|^2.
\]
At a maximum for \(\Phi\), \(\Delta |\Phi|^2 \geq 0\). If this is non-zero we obtain
\[
|\Phi|^2 \leq \max_Y (-2\kappa + 8|\omega + d\alpha|, 0).
\]

Since \(\omega\) and \(\alpha\) are in \(C^1\), we obtain a uniform pointwise bound on the spinor component of any SW\(_{\alpha,\omega}\)-solution.

Let us prove (ii). Suppose that \((A_i, \Phi_i)\) is a given sequence of SW\(_{\alpha,\omega}\)-solutions. Choose a fixed reference smooth connection \(A\), and write \(A_i = A + a_i\). Then after a gauge transformation we may assume that \(da_i = 0\) and the harmonic component of \(a_i\) is uniformly bounded, since the component group of maps \(Y \to U(1)\) is \(H_1(Y; \mathbb{Z})\), and thus \(H^1(Y; i\mathbb{R})/H^1(Y; i\mathbb{Z})\) is compact. Let \(\hat{a}_i\) be the \(L^2\)-component of \(a_i\) which is \(L^2\)-perpendicular to the harmonic forms. Since the harmonic forms are \(C^\infty\), \(a_i - \hat{a}_i\) must lie in \(C^k\) for every \(k\). The SW\(_{\alpha,\omega}\) equations together with the uniform pointwise bound on \(\Phi_i\) gives (by ellipticity) a uniform \(L^p_1\) bound for the \(\hat{a}_i\). Applying this to the equation for \(\Phi_i\) this gives again by ellipticity a uniform \(L^p_1\) bound on the \(\Phi_i\). Circulating inductively we terminate with uniform \(L^p_{k+1}\) bounds on both \(\hat{a}_i\) and \(\Phi_i\), since \(\alpha\) and \(\omega\) are assumed to be in \(C^k\). The uniform bound holds for all \(p \geq 2\). Therefore the sequence \((a_i, \Phi_i)\) is uniformly bounded in \(L^p_{k+1}\). By Rellich’s theorem a subsequence of \((a_i, \Phi_i)\) converges in \(L^p_k\). For \(p\) sufficiently large, \(L^p_k \subset L^2_{k+1}\); thus the sequence converges in \(C(P)\), since the underlying topology in \(L^2_2\).

The proof of (i) follows from the preceding by applying it to the constant sequence. \(\square\)

### 2.3. Reducible Solutions.

When \(\Phi = 0\), the Seiberg-Witten equations reduce to a single equation for the connection \(A\): \(F_A = \omega\). If \(\omega = 0\) then the reducible (up to gauge equivalence) is identified with the moduli space of flat \(U(1)\)-connections on \(L\). Thus a necessary condition is that \(c_1(L)_\mathbb{R} = 0\). If this is so, then by a well-known fact in differential geometry, the gauge equivalence classes of flat connections is completely determined by the holonomy representation of \(\pi_1(Y)\) and is therefore topologically a product \(U(1) \times \cdots \times U(1)\) where the number of factors equals \(b_1(Y)\). In particular if \(b_1(Y) = 0\) then the
reducible is exactly one point. (Note: When $b_1(Y) = 0$, $L$ admits only one flat connection, up to gauge.)

**Lemma 7.** The equation $F_A = \omega$ has a solution if and only if the real cohomology classes $[F_A] = [\omega]$ or equivalently $[\frac{i}{2\pi}\omega] = c_1(L)_R$. If the latter holds, then the space of equivalence classes of reducible solutions is topologically $U(1) \times \cdots \times U(1)$ where the number of factors equals $b_1(Y)$, and in the case $b_1(Y) = 0$, a single point.

**Proof.** As explained above the condition $[F_A] = [\omega]$ is necessary. For sufficiency, let $A_0$ be such that $[F_{A_0}] = [\omega]$. Then we only need to solve for $a$ in $F_{A_0} + a = \omega$ which is equivalent to $da = \omega - F_{A_0}$. Since $\omega - F_{A_0}$ is exact such an $a$ can be found. Assuming solutions exist, let $F$ denote the space of all $A$’s such that $F_A = 0$. Then $F + a$ describes all the solutions to $F_A = \omega$. Therefore the space of reducible solutions up to gauge are topologically the same as in the case $\omega = 0$. □

2.4. Orientation.

Suppose $Z^*_\pi(P)$ consists only of regular points. It is clear that $Z^*_\pi(P)$ is orientable. We want to produce a procedure for inducing a global orientation. The fundamental elliptic complex can be combined into a single operator $L_{\eta,A,\Phi} : \Omega^0_2(iR) \oplus \Omega^1_2(iR) \oplus \Gamma_2(S) \rightarrow \Omega^0_1(iR) \oplus \Omega^1_1(iR) \oplus \Gamma_1(S)$, $L_{\eta,A,\Phi} = \delta^1_{\eta,A,\Phi} + \delta^0_{\eta,A,\Phi}$. A direct computation verifies that this operator is formally self-adjoint.

Let $\Lambda = \det \text{Ind} \{L_{\eta,A,\Phi}\}$. Let $g \in G$ and $(a,\phi) \in \ker L_{\eta,A,\Phi}$. Then $(a,g^{-1}\phi) \in \ker L_{\eta,A,\Phi}$. Therefore the action of $G$ lifts to an action on $\Lambda$. Note that if $\Phi = 0$ then the stabilizer $U(1)$ maps the fibre of $\Lambda$ at $A = (A,0)$ back to itself by the identity. Hence $\Lambda$ descends to a line bundle $\hat{\Lambda}$ over $\Omega^0_2(iR) \times \mathcal{B}(P)$.

**Proposition 8.** The real line bundle $\hat{\Lambda}$ is trivial.

**Proof.** We need to show that $\Lambda$ possesses a $G$-equivariant trivialization. The substitution of $(1 - \varepsilon)\Phi$, $0 \leq \varepsilon \leq 1$, for $\Phi$ and $(1 - \varepsilon)\eta$ for $\eta$ in the definition of $\delta^1_{\eta,A,\Phi}$ and $\delta^0_{\eta,A,\Phi}$ defines a homotopy of $L_{\eta,A,\Phi}$ to an operator $L'_{\eta,A,\Phi}$ given by $L'_{\eta,A,\Phi}(\xi,a,\phi) = (da + 2d\xi,D_A\phi)$. This homotopy is $G$-equivariant. We have $\ker L'_{\eta,A,\Phi} = H^0(iR) \oplus H^1(iR) \oplus \mathcal{H}_A = \ker L''_{\eta,A,\Phi}$. This family has a trivial determinant, and this proves $\Lambda$ is $G$-equivariantly trivial. □

Notice that the homotopy given in the proof is the identity over $\{0\} \times \mathcal{C}^{\text{Red}}(P)$. Thus over this set $\det \text{Ind} \{L_{\eta,A,\Phi}\}$ is the determinant of the index of a constant family $\{L_{\text{dRham}}\}$ tensored with the complex family $\{D_A\}$. Since a complex family is canonically oriented, we may ignore it. The kernel and cokernel of $L_{\text{dRham}}$ are $H^0(iR) \oplus H^1(iR)$ and by identifying them with each other we obtain a trivialization of $\det \text{Ind} \{L_{\eta,A,\Phi}\}$ over $\{0\} \times \mathcal{C}^{\text{Red}}(P)$. 

This orients $\text{det} \text{Ind}\{L_{\eta,A,\Phi}\}$ over all of $\Omega^0_2(i\mathbb{R}) \times \mathcal{C}(P)$. This is the natural orientation of $\hat{\Lambda}$.

Consider the trivial real line bundle $\mathbb{R}$ over $\Omega^0_2(i\mathbb{R}) \times \mathcal{B}^*(P)$. Then $\hat{\Lambda}$ has the property that over the open set $\mathcal{O}$ of $\Omega^0_2(i\mathbb{R}) \times \mathcal{B}^*(P)$ defined by the condition that $\ker L_{\eta,A,\Phi} = 0$, there is a canonical isomorphism $h: \hat{\Lambda}|_{\mathcal{O}} \cong \mathbb{R}|_{\mathcal{O}}$.

**Proposition 9.** Let $\hat{\Lambda}$ have the natural orientation described above. Assume $Z_\pi^*(P)$ consists only of regular points. Then the following rule defines an orientation $\varepsilon: Z_\pi^*(P) \to \{\pm 1\}$. Let $x \in Z_\pi^*(P)$. Denote by $o(x, \mathbb{R})$ the canonical orientation of $\mathbb{R}|_x$ and $o(x, \hat{\Lambda})$ the orientation induced by $\hat{\Lambda}$ via the isomorphism $h$ above. Then

$$
\varepsilon(x) = \begin{cases} 
1 & \text{if } o(x, \mathbb{R}) = o(x, \hat{\Lambda}) \\
-1 & \text{otherwise.}
\end{cases}
$$


Let $\{g(t)\}, t \in I_\varepsilon = (-\varepsilon, 1 + \varepsilon)$ be a 1-parameter family of metrics on $Y$. In this section we examine the parameters $(\pi, t)$ in $\mathcal{P} \times I_\varepsilon$ for which the moduli space $Z_\pi^*(P; g(t))$ consists solely of regular points.

In order to understand how the geometric structures and operators changes with the 1-parameter family of metrics it will be useful to be able to work with a single reference underlying metric, spin-c structure and model for the spinors. Fix an underlying metric which we take to be $g$, let $P_{SO}$ be the corresponding oriented orthonormal frame bundle and $h$ an automorphism of $TY$. $h$ induces an automorphism $h_*$ of $P_{GL+}$, the component of positively oriented frames of the frame bundle of $Y$. The image of $P_{SO}$ in $P_{GL+}$ under $h$ describes the orthonormal frame bundle of another metric. Conversely, the positive orthonormal frame bundle of any other metric can be recovered on this way. Call the second metric $g'$.

$h$ can be lifted to an isomorphism between $P$, the spin-c structure for $g$ and $P'$, the spin-c structure for $g'$. This, in turn, induces a fibrewise isometry $\hat{h}$ between the corresponding spinor bundles $S$ and $S'$. By changing $\hat{h}$ to $e^u\hat{h}$ where $u$ is a smooth function on $Y$ we can arrange it so that $e^u\hat{h}$ gives an isometry between $\Gamma(S)$ and $\Gamma(S')$ with respect to their $L^2$-norms.

Given the 1-parameter family $g(t)$ the construction of $e^u\hat{h}$ above can be carried out smoothly in the parameter $t$, taking for instance $g = g(0)$ to be the reference metric. Therefore using these isomorphisms as identifications we may assume $(A, \Phi)$ etc. for every $t$ is defined on a fixed reference bundle. The Dirac operator now depends also on $t$, and we denote this as $D_A^{g(t)}$. It is always self-adjoint with respect to the reference spinor bundle. Further information regarding the relation between the Dirac operator for different metrics can be found in [B], [BG] and [H].
3.1. Singular Locus of Dirac Operators.

In this section we discuss the singular locus for certain families of Dirac operators, i.e., the parameters for which the Dirac operator is singular. This will be crucial for us later.

Suppose \( b_1(Y) = 0 \). Since \( H^2(Y; i\mathbb{R}) = 0 \), we have a bounded right inverse \( d^{-1} : \Omega_k \to \Omega^1_2(i\mathbb{R}) \) for the operator \( d \). Let \( \theta \) be a fixed \( C^\infty \) flat connection on the determinant \( L \). Then for any given \( \omega, A = \theta + d^{-1}(\omega) \) solves \( F_A = \omega \). Define \( \{ D(\alpha, \omega, t) \} \) to be the family of Dirac operators

\[
D : \mathcal{P} \times I_\varepsilon \to \text{Fred}^0(\Gamma(S)), \quad D(\alpha, \omega, t) = D_{\theta + \alpha + d^{-1}(\omega)}^0.
\]

Here \( \text{Fred}^0 \) denotes the Banach space of Fredholm operators of index zero.

In the case \( b_1(Y) = 1 \), we shall also define a family as follows: This time we keep the metric fixed, so we drop it from the notation. Let \( A_0 \) be a fixed \( C^\infty \) connection on \( L \) and denote by \( \omega_0 \) its curvature. Fix a choice of non-zero \( a_0 \in H^1(i\mathbb{R}) \) such that \( \frac{i}{4\pi} a_0 \) defines a generator for \( H^1(\mathbb{Z}) \). (The choice of constants here is so that \( a_0 \) is the class of a gauge change \( 2g^{-1}dg \).) Then the set \( \{ A_0 + ta_0 \mid t \in [0, 1] \} \), parametrizes all the reducible SW \( \alpha, \omega_0 \) solutions up to gauge equivalence. Let \( \mathcal{E}_k, k \geq 2 \), denote the exact forms in \( \Omega_k \). Then on \( \mathcal{E}_k \) we can as before define a bounded inverse \( d^{-1} : \mathcal{E}_k \to \Omega^1_2(i\mathbb{R}) \). Given \( \omega \in \mathcal{E}_k \) then \( A = A_0 + ta_0 + d^{-1}(\omega) \) solves \( F_A = \omega_0 \). Thus the set \( \{ A_0 + ta_0 + d^{-1}(\omega_0) \mid t \in [0, 1] \} \) parameterizes up to gauge equivalence all the reducible SW \( \alpha, \omega_0 \)-solutions, \( \omega \in \mathcal{E}_k \). Define the family \( \{ \overline{D}(\alpha, \omega, t) \} \) by

\[
\overline{D} : Q \times \mathcal{E}_k \times I_\varepsilon \to \text{Fred}^0(\Gamma(S)), \quad \overline{D}(\alpha, \omega, t) = D_{A_0 + ta_0 + \alpha + d^{-1}(\omega)}.
\]

**Proposition 10.** Let \( \mathcal{N} \) be the subset of \( \mathcal{P} \times I_\varepsilon \) consisting of all \( (\alpha, \omega, t) \) for which \( D(\alpha, \omega, t) \) is singular. Similarly define the subset \( \mathcal{K} \) of \( Q \times \mathcal{E}_k \times I_\varepsilon \) for \( \overline{D}(\alpha, \omega, t) \). Then \( \mathcal{N} \) and \( \mathcal{K} \) are nowhere dense closed subspaces.

**Proof.** We prove only the case for \( \mathcal{N} \). The other is done similarly. Let \( B \) be the unit \( L^2 \)-ball in \( \Gamma_2(S) \). Let \( V \to \mathcal{P} \times I_\varepsilon \times B \) be the vector bundle whose fibre at \( (\pi, \phi) \) is the real \( L^2 \)-orthogonal to \( \phi \) in \( \Gamma_1(S) \). By evaluating \( D(\alpha, \omega, t) \) on \( \phi \in B \) we obtain a section, call it \( D \), of \( V \). We claim this section is transverse to the zero section. Let \( D(\alpha_0, \omega_0, t_0)\phi_0 = 0 \). Then \( \psi \in V_{\alpha_0, \omega_0, t_0, \phi_0} \) be \( L^2 \)-orthogonal to the derivative \( ddD \) at \( (\alpha_0, \omega_0, t_0, \phi_0) \). By varying \( \phi_0 \) in the tangent direction \( \delta \phi \) we find \( ddD(\delta \phi) = D(\alpha_0, \omega_0, t_0)\delta \phi \); thus \( \psi \) must also satisfy \( D(\alpha_0, \omega_0, t_0)\psi = 0 \) (since \( D(\alpha, \omega, t) \) is self-adjoint). On the other hand, by varying \( \alpha_0 \), \( dD(\delta \alpha) = \delta \alpha \cdot \gamma_{t_0} \phi_0 \) and if \( \psi \) is \( L^2 \)-perpendicular to this then \( \psi = if \phi_0 \) for some real function \( f \). The condition \( D(\alpha_0, \omega_0, t_0)\psi = 0 \) then leads to \( df \cdot \gamma_{t_0} \phi = 0 \); but since \( \phi_0(x) \neq 0 \) on an open set it must be that \( df = 0 \), and so \( f \) is a constant. Finally \( \psi \) being in \( V \) is necessarily \( L^2 \)-orthogonal to \( \phi_0 \); thus \( f = 0 \). Hence \( \psi = 0 \) and transversality holds and the zeros of \( D \) defines a smooth infinite dimensional submanifold \( \mathcal{M} \) of \( \mathcal{P} \times I_\varepsilon \times B \). The projection map \( p : \mathcal{M} \to \mathcal{P} \times I_\varepsilon \) is proper since the kernel
of the Dirac operator is always finite dimensional. Applying the Sard-Smale theorem we can conclude that there exists an open dense set $\mathcal{O}$ in $\mathcal{P} \times I_\varepsilon$ with the property that $\mathcal{D}(\alpha, \omega, t), (\alpha, \omega, t) \in \mathcal{O}$ has nullity $\leq 1$ over the reals. But since this is a complex linear operator and self-adjoint, $\mathcal{D}(\alpha, \omega, t)$ must be non-singular. The proposition now follows. □

Let $\pi_i = (\alpha_i, \omega_i), i = 0, 1$, be given perturbations. Denote by $\{\pi(t)\} = \{(\alpha(t), \omega(t))\}, t \in I_\varepsilon$ the 1-parameter family of perturbations defined by $\pi(t) = (1 - t)\pi_0 + t\pi_1$. For a fixed value of $\pi$, $\mathcal{D}(\pi(t) + \pi, t)$ defines a 1-parameter family of Dirac operators. We use the notation $\{\mathcal{D}_\pi(t)\}$ for this 1-parameter family. We call this family transverse (for the choice of $\pi$) if $\mathcal{D}_\pi(0)$ and $\mathcal{D}_\pi(1)$ are non-singular and the family has transverse spectral flow, as $t$ varies over $[0, 1]$. Transverse spectral also includes the condition that multiple zero-eigenvalues do not occur as $t$ varies. For the case $b_1(Y) = 1$, fix a value of $\alpha_0$. In a similar way we have a 1-parameter family $\{\mathcal{D}_\pi(t)\}$ obtained by considering $\mathcal{D}((\alpha_0, \omega_0) + \pi, t)$ for a fixed $\pi$. Transversality is defined in the same way as before.

**Proposition 11.** Suppose $\{\mathcal{D}_0(t)\}$ has the property that $\mathcal{D}_0(t)$ is non-singular for $t = 0, 1$. Then there are arbitrarily small $\pi$ such that $\{\mathcal{D}_\pi(t)\}$ is a transverse family. A similar statement holds for $\{\mathcal{D}_\pi(t)\}$.

**Proof.** We shall only prove the case of $\{\mathcal{D}_0(t)\}$. The other case is handled by essentially the same argument. Let $x = (\pi_0, t_0) \in \mathcal{P} \times I, I = (0, 1)$. Consider the map

$$G : \Gamma_2(S) \times \mathcal{P} \times I \to \Gamma_2(S), \quad G(\phi, \pi, t) = \mathcal{D}(\pi, t)\phi.$$  

The differential of $G$ at $(0, \pi_0, t_0)$ is given by

$$dG(\delta\phi, \delta\omega, \delta t) = \mathcal{D}(\pi_0, t_0)\delta\phi.$$ 

Then $\ker dG = \mathcal{H}_{t_0} \oplus \mathcal{P} \oplus \mathbb{R}$ and $\coker dG = \mathcal{H}_{t_0}$. Here $\mathcal{H}_{t_0}$ denotes the kernel of $\mathcal{D}(\pi(t_0) + \pi_0, t_0)$ (acting on $\Gamma_2(S)$). By the implicit function theorem there is a neighbourhood $V$ of $(0, x) \in \mathcal{H}_{t_0} \times \mathcal{P} \times I$ and a unique smooth map $f : V \to \mathcal{H}_{t_0}$ such that for $(\phi, \pi, t) \in V$,

$$\begin{align*}
(1) \quad (I - \Pi)G(\phi + f(\phi, \pi, t), \pi, t) = 0, \quad \Pi = L^2\text{-projection onto } \mathcal{H}_{t_0}.
\end{align*}$$

Note that the linear extension of $f$ in the $\phi$ variable continues to satisfy (1) so we may take $V$ to be of the form $\mathcal{H}_{t_0} \times W, W$ a neighbourhood of $x$. Because of (1) the injective/surjective properties of $\mathcal{D}(\pi, t), (\pi, t) \in W$, are completely determined by the finite-dimensional operator finite dimensional operator $\mathcal{H}_{t_0} \to \mathcal{H}_{t_0}$,

$$T(\pi, t)\phi = \Pi G(\phi + f(\phi, \pi, t), \pi, t).$$

We claim $T(\pi, t)$ is self-adjoint with respect to the (real $\mathbb{C}$-invariant) $L^2$-inner product on $\mathcal{H}_{t_0}$. We introduce the notation $\langle \cdot, \cdot \rangle_{\mathcal{H}_{t_0}}, \psi, \phi \in \mathcal{H}_{t_0},$ to
Herm(\mathcal{H}_t) denote the (real) vector space of Hermitian transformations on \mathcal{H}_t. Let Herm(\mathcal{H}_t) denote the (real) vector space of Hermitian transformations on \mathcal{H}_t. The determinant function \det: Herm(\mathcal{H}_t) \rightarrow \mathbb{R} \subset \mathbb{C} and \det^{-1}(0) is a closed subvariety of codimension 1 in Herm(\mathcal{H}_t). Introduce the notation

\[ N^{(k)} = \{ l \in \text{Herm}(\mathcal{H}_t) \mid \dim \ker(l) \geq k \}. \]

Lemma 12. Suppose \dim_{\mathbb{C}}(\mathcal{H}_t) > 0. The derivative \((dT)_{\pi_0,t_0}\mid_{\mathcal{Q}}\) of \(T\) restricted to \(\mathcal{Q}\) has non-trivial image in \(\text{Herm}(\mathcal{H}_t)\). If \(\dim_{\mathbb{C}}(\mathcal{H}_t) \geq 2\) then this image is of dimension \(\geq 2\).

First let us show that the image of \((dT)_{\pi_0,t_0}\) is non-trivial. The derivative at \((\pi_0, t_0)\) is computed to be

\[ dT_{\pi_0,t_0}(\delta \alpha, \delta t) \phi = \Pi((\delta \alpha + d^{-1} \delta \omega) \cdot_{t_0} \phi), \quad \pi = (\alpha, \omega). \]

Suppose that \(\langle dT_{\pi_0,t_0}(\delta \alpha) \phi, \phi \rangle_{L^2} = 0\) for all \(\delta \alpha \in \mathcal{Q}\). Since

\[ \langle \delta \alpha \cdot_{t_0} \phi, \phi \rangle_{L^2} = \int_Y \langle \delta \alpha, \sigma_{t_0}(\phi, \phi) \rangle. \]

This implies \(\sigma_{t_0}(\phi(y), \phi(y)) = 0\) for all \(y \in Y\) and thus \(\phi = 0\). Thus the image of \(dT_{\pi_0,t_0}\) is non-trivial.

Assume now that \(\dim_{\mathbb{C}}(\mathcal{H}_t) \geq 2\). Let \(\delta \alpha\) be such that \(dT_{\pi_0,t_0}(\delta \alpha) \neq 0\). Let \(\phi_1, \ldots, \phi_n\) be an complex orthonormal basis for \(\text{Herm}(\mathcal{H}_t)\) such that \(dT_{\pi_0,t_0}(\delta \alpha)\) is diagonal with respect to this basis. Thus \(\langle \delta \alpha \cdot \phi_i, \phi_j \rangle_{L^2,\mathbb{C}} = 0\) for \(i \neq j\). Since the \(\phi_k\) are harmonic spinors, unique continuation implies that there is a open set in \(Y\) on which \(\phi_i \neq \phi_j, i \neq j\) on this open set, in particular say at the point \(y \in Y\). We can find a \(\delta \alpha'\) with support in an arbitrarily small neighbourhood of \(y\) such that \(\int_Y \langle \delta \alpha' \cdot \phi_i, \phi_j \rangle_{\mathbb{C}} = 0, i \neq j\). Thus \(dT_{\pi_0,t_0}(\delta \alpha')\) is independent of \(dT_{\pi_0,t_0}(\delta \alpha)\). This shows that the image is at least 2-dimensional. This proves the Lemma.
Consider for each \((\alpha, \omega) \in W\),
\[
\tau_{\alpha,\omega}(t) = T(\alpha, \omega, t), \quad |t - t_0| < \varepsilon
\]
where \(\varepsilon > 0\) is chosen so that \((\alpha, \omega, t) \in W\). Clearly \(\tau_{\alpha,\omega}\) defines a path in \(\text{Herm}(\mathcal{H}_{t_0})\).

By an open cover argument the following exists: (1) a finite set \([t_1, \ldots, t_n]\) \(\subset [0, 1]\) together with open neighbourhoods \(V_i\) of \(t_i\) in \(\mathbb{R}\) such that \([V_i]_i\) covers \([0, 1]\) (2) an open neighbourhood \(W' \subset W\) of \(0 \in \mathcal{P}\) (3) maps \(T^i : W' \times I_i \rightarrow \text{Herm}(\mathcal{H}_{t_0})\) as in the preceding which preserves the injectivity/surjectivity properties of \(D(\alpha, \omega, t)\), \((\alpha, \omega, t) \in W' \times I_i\). We denote the corresponding paths by \(\tau^i_{\alpha,\omega}(t)\).

Let \(N = \max_i\{\dim_C(\mathcal{H}_{t_j})\}\). Suppose \(N > 1\). Let \(j\) be such that \(\dim(\mathcal{H}_{t_j}) = N\). Note that \(\mathcal{N}^{(N)} = \{0\} \subset \text{Herm}(\mathcal{H}_{t_j})\). Thus any non-zero element in \(\text{Herm}(\mathcal{H}_{t_0})\) lies in \(\mathcal{N}^{(k)}\), \(0 \leq k < N\). Then by the above Lemma we can find a sufficiently small perturbation \((\alpha, 0) \in W'\) so that \(\tau^i_{\alpha,0}(t) \in \mathcal{N}^{(k)}\), \(0 \leq k' < N\), \(t \in I_i\) and for \(i \neq j\), \(\tau^i_{\alpha,0}(t) \in \text{Herm}(\mathcal{H}_{t_i})\) for all \(t \in I_i\). Thus we establish that there is an arbitrarily small \(\alpha\) so that \(\tau^i_{\alpha,0}(t) \in \mathcal{N}^{(k)}\) with \(0 \leq k < N\) for every \(i\). Repeating the above construction over but with \(W'\) taken to be an open neighbourhood of \((\alpha, 0)\) instead, we inductively prove that we can find an arbitrarily small \(\alpha'\) so that we have \(N = 1\). The perturbation argument in this case makes each path \(\tau^i_{\alpha',0}\) transverse to \(\mathcal{N}^{(1)} = \{0\} \subset \text{Herm}(\mathcal{H}_{t_0}) \cong \mathbb{R}\).

Let us show that \(\{D_{\alpha',0}(t)\}\) is a transverse family. Suppose at \(s\), \(\tau^i_{\alpha',0}(s) = 0\). Let \(\phi \in \mathcal{H}_{t_i}\) be unit length. Let \(\lambda(t)\) be the 1-parameter family of eigenvalues satisfying \(D_{\alpha',0}(t)\phi = \lambda(t)\phi\) for \(t\) close to \(s\). Thus we have
\[
\langle D_{\alpha',0}(t)\phi, \phi \rangle_{L^2} = \lambda(t)\langle \phi, \phi \rangle_{L^2}.
\]
Differentiating this equation with respect to \(t\) and evaluating at \(t = s\) gives
\[
\langle dT_{\pi_0, t_0}(0, 0, 1)\phi, \phi \rangle_{L^2} = \lambda'(s).
\]
The left hand term is simply the velocity of \(\tau^i_{\alpha',0}\) at \(t = s\) and transversality means this is non-zero. Thus \(\lambda'(s) \neq 0\) and we have transverse spectral flow at \(t = s\).

\[\square\]

### 3.2. The Parameterized Moduli Space.

As before we assume the 1-parameter families \(\{g(t)\}, t \in I_\varepsilon\). Define the parameterized Seiberg-Witten section to be the map
\[
\tilde{s} : \Omega^0_2(i\mathbb{R}) \times \mathcal{C}(P) \times \mathcal{P} \times I_\varepsilon \rightarrow \Omega^1_1(i\mathbb{R}) \oplus \Gamma_1(S),
\]
\[
\tilde{s}(\eta, A, \Phi, \pi, t) = s_{\pi, g(t)}(\eta, A, \Phi).
\]
Then the parameterized moduli space is \(\mathcal{Z}(P) = \tilde{s}^{-1}(0)/\mathcal{G}\) (with \(\mathcal{G}\) acting only on the \(\mathcal{C}(P)\) factor). There is the projection map \(p : \mathcal{Z}(P) \rightarrow \mathcal{P} \times I_\varepsilon\) and clearly \(p^{-1}(\pi, t) = \mathcal{Z}_\pi(P; g(t))\).
**Proposition 13.** The irreducible part $Z^*(P)$ of the parameterized moduli space is smooth Hilbert space manifold and the projection map $p|_{Z^*(P)} : Z^*(P) \to \mathcal{P} \times I_\varepsilon$ is a smooth Fredhom map of index zero.

If $b_1(Y) = 0$ we determined in Prop. 10 that $p^{-1}(\pi)$ contains no reducibles if we are off the set $\mathcal{N} \subset \mathcal{P} \times I_\varepsilon$. If $b_1(Y) > 0$ we observed in Lemma 7 that there are no reducibles in $p^{-1}(\alpha, \omega)$ if and only if $[\frac{c}{b} \omega]$ is not an integral class in cohomology. Thus:

**Lemma 14.** Let $q : \mathcal{P}_k \times I_\varepsilon \to H^2(Y; \mathbb{R})$, $q(\alpha, \omega, t) = [\frac{c}{b} \omega]$ and set $\mathcal{W} = q^{-1}(c_1(L_R))$. Then $p^{-1}(\pi, t)$ has reducibles if and only if $(\pi, t) \in \mathcal{W}$. $\mathcal{W}$ is a closed nowhere dense subset of of codimension equal to $b_1(Y)$.

We remark that in the case $b_1(Y) > 0$ the condition of regularity (i.e., $H^2_{A, \Phi} = \{0\}$) for reducible solutions can never be satisfied. This is because when $\Phi = 0$, $H^2_{A, \Phi}$ reduces to $H^2(\mathbb{R}) \oplus \mathcal{H}_A$. So regularity implies the absence of reducibles in this case.

**Corollary 15.** $p|_{Z^*(P)} : Z^*(P) \to \mathcal{P} \times I_\varepsilon$ is proper over $\mathcal{P} \times I_\varepsilon \setminus \mathcal{W}$ where (i) $\mathcal{W} = \mathcal{N}$ if $b_1(Y) = 0$ (ii) $\mathcal{W} = \mathcal{W}$ if $b_1(Y) > 0$. Therefore for an open dense set $\mathcal{O} \subset \mathcal{P} \times I_\varepsilon$, $p^{-1}(z)$, $z \in \mathcal{O}$, is a finite set of regular points.

**Proof.** The properness assertion is the content of Proposition 6 and the fact that a regular reducible point in the case $b_1(Y) = 0$ is necessarily isolated, by the Kuranishi local model. The Sard-Smale theorem then gives the ‘open dense set’ statement since regularity of an irreducible solution $(A, \Phi)$ is equivalent to the derivative $d\tilde{s}$ at $(0, A, \Phi)$ being surjective. (Note: without the properness assertion we can only conclude regularity on a Baire set.)

**Proof of Proposition 13.** We have to show that the derivative $d\tilde{s}$ is surjective at every point $(0, A_0, \Phi_0, \pi_0, t_0) \in \tilde{s}^{-1}(0)$ for which $\Phi_0 \neq 0$. Let $(b, \psi)$ lie in the cokernel of $d\tilde{s}_0(A_0, \Phi_0)$, i.e., $(b, \psi) \in H^1_{A, \Phi}$, thus

(2) \hspace{1cm} (i) $db = \frac{1}{2} \sigma(\Phi, \psi)$, \hspace{0.5cm} (ii) $D_A \psi + \frac{b}{2} \cdot \Phi = 0$, \hspace{0.5cm} (iii) $2d^* b = i(\Phi, \psi)$.

Suppose $(b, \psi)$ is $L^2$-orthogonal to the image of $d\tilde{s}$. The Proposition is proven as soon as we can show $(b, \psi) = 0$. If $\delta \omega \in \Omega_k$, then $d\tilde{s}(\delta \omega) = (* \delta \omega, 0)$. Thus $b$ must be $L^2$ orthogonal to all the co-closed forms; this implies that $b$ must be closed. Then from (i) we obtain the condition $\sigma(t_0(\Phi_0, \psi) = 0$. Working at a point, the kernel of the transformation $v \mapsto \sigma(t_0(w, v)$ is of dimension 1 and it is easy to check that $\sigma(t_0(w, iv) = 0$. Therefore $\psi = i f \Phi_0$ for some real valued function $f$. Putting this into (ii) of (2) we obtain

$$0 = D_{A_0}^g(t_0) (i f \Phi) + \frac{b}{2} \cdot t_0 \cdot \Phi$$

$$= (idf + \frac{b}{2}) \cdot \Phi \quad \text{(since } D_{A_0}^g(t_0) \Phi_0 = 0\text{).}$$
Hence we obtain the pointwise condition $idf + \frac{b}{2}f = 0$ on the open dense set $O$ where $\Phi_0 \neq 0$. By continuity it holds on all of $Y$. Substituting into (iii) of (2) we get the equation

$$4\Delta f = -|\Phi|^2 f.$$ 

Taking the product with $f$ and integrating we obtain:

$$\int_Y 4|df|^2 + |\Phi|^2 |f|^2 = 0.$$ 

Thus $f = 0$ on $O$ and therefore $Y$ and we finally obtain $(b, \psi) = 0$. Finally the index zero assertion follows directly from Lemma 4.

\[\square\]

4. Proof of Theorem 1.

$Y$ is assumed to be a closed oriented 3-manifold with Riemannian metric $g$ and spin-c structure $P \to Y$. According to Corollary 15 applied to the constant family $\{g(t) = g\}$, we may choose a perturbation $\pi$ from an open dense set in $\mathcal{P}_k$ ($k \geq 3$) such that $Z_{\pi, g}(P)$ consists of a finite set of regular points, i.e., the cohomology $H^2_{\mathcal{A}, \phi}$ is trivial at these points. For this $\pi$, if $b_1(Y) = 0$ there is a unique isolated reducible (up to gauge equivalence) and if $b_1(Y) > 0$ there are no reducibles. $Z^*_{\pi, g}(P)$ is then naturally oriented by our conventions (Proposition 9) and we can form the algebraic sum

$$\# Z_{\pi, g}(P).$$

To prove the claimed invariance properties of $\# Z_{\pi, g}(P)$, let $g_0$, $g_1$ be two metrics on $Y$ and let $\pi_i$ be two perturbations which satisfy the above with respect to $g_i$. We want to relate $\# Z_{\pi_0, g_0}(P)$ and $\# Z_{\pi_1, g_1}(P)$. Consider the 1-parameter family of metrics $\{g(t)\} = \{(1-t)g_0 + tg_1\}$ defined for $t \in I_\varepsilon = (-\varepsilon, 1 + \varepsilon)$. Thus as in Section 3 we have a parameterized moduli space $\mathcal{Z}(P)$ and projection map $p : \mathcal{Z}(P) \to \mathcal{P}_k \times I_\varepsilon$.

We consider a smooth path $\sigma : [0, 1] \to \mathcal{P}_k \times I_\varepsilon$, $\sigma(0) = (\pi_0, g_0)$, $\sigma(1) = (\pi_1, g_1)$. We introduce the notation $Z_\sigma(P)$ for the $\sigma$-parametrized moduli space $\{(x, t) \mid x \in p^{-1}(\sigma(t))\}$. If a portion of $\sigma$ misses the 'singular' sets $N$ or $W$ of Sec.3.1 and is transverse $p$, then that portion of $Z^*_\sigma(P)$ consists purely of regular points and therefore is a smooth arc. This is oriented in the following way. The local deformation theory of $Z_\sigma(P)$ is described by an elliptic complex of the form

$$\Omega^0_3(i\mathbb{R}) \to \Omega^0_2(i\mathbb{R}) \oplus \Omega^1_2(i\mathbb{R}) \oplus \Gamma_2(S) \oplus \mathbb{R} \to \Omega^1_1(i\mathbb{R}) \oplus \Gamma_1(S).$$

Therefore the orientation is determined by looking at the ‘wrapped up’ operator

$$L_{\eta, A, \Phi, t} : \Omega^0_2(i\mathbb{R}) \oplus \Omega^1_2(i\mathbb{R}) \oplus \Gamma_2(S) \oplus \mathbb{R} \to \Omega^0_1(i\mathbb{R}) \oplus \Omega^1_1(i\mathbb{R}) \oplus \Gamma_1(S).$$
The orientation of the regular irreducible points of $Z_{\sigma}(P)$ is determined by an orientation of the determinant of the index of the family $\{L_{\eta,A,\Phi,t}\}$. We have the short exact sequence of 2-step complexes:

\[ 0 \rightarrow \Omega^{3}_{0}(i\mathbb{R}) \oplus \Omega^{2}_{1}(i\mathbb{R}) \oplus \Gamma_{2}(S) \xrightarrow{L_{\eta,A,\Phi}} \Omega^{3}_{1}(i\mathbb{R}) \oplus \Omega^{2}_{1}(i\mathbb{R}) \oplus \Gamma_{1}(S) \rightarrow 0 \]

\[ 0 \rightarrow \Omega^{2}_{0}(i\mathbb{R}) \oplus \Omega^{1}_{1}(i\mathbb{R}) \oplus \Gamma_{2}(S) \oplus \mathbb{R} \xrightarrow{L_{\eta,A,\Phi,t}} \Omega^{3}_{1}(i\mathbb{R}) \oplus \Omega^{2}_{1}(i\mathbb{R}) \oplus \Gamma_{1}(S) \rightarrow 0 \]

\[ 0 \rightarrow \mathbb{R} \rightarrow 0 \]

This gives rise to a canonical isomorphism

(3) \quad h : \ker L_{\eta,A,\Phi,t} \oplus \text{coker} L_{\eta,A,\Phi} \rightarrow \ker L_{\eta,A,\Phi} \oplus \mathbb{R} \oplus \text{coker} L_{\eta,A,\Phi,t}.

An orientation for $\text{det} \text{Ind}\{L_{\eta,A,\Phi}\}$ defines an orientation for $\text{det} \text{Ind}\{L_{\eta,A,\Phi,t}\}$ according to this rule: Choose an orientation for $\text{coker} L_{\eta,A,\Phi}$. Then an orientation of $\ker L_{\eta,A,\Phi}$ is determined, since $\text{det} \text{Ind}\{L_{\eta,A,\Phi}\}$ is oriented. Now given an orientation of $\text{coker} L_{\eta,A,\Phi,t}$, then $\ker L_{\eta,A,\Phi,t}$ is oriented so that $h$ is an orientation-preserving isomorphism, where the domain and range spaces are given the product orientation in the order written in (3). With this orientation convention, if $Z_{\sigma}$ consists entirely of regular irreducible points and if compact then its boundary is precisely $Z_{\sigma(1)}(P) - Z_{\sigma(0)}(P)$, as oriented spaces.

The proof of Theorem 1 in the case $b_{1}(Y) > 1$ can now be easily established. By Lemma 14 $\sigma$ may be chosen to be disjoint from the subset of $W$ for which $Z_{\sigma}(P)$ has reducibles. Furthermore $\sigma$ can be assume to be transverse to the projection $p : Z(P) \rightarrow \mathcal{P}_{k} \times I_{\varepsilon}$. Thus $Z_{\sigma}(P)$ defines a smooth compact oriented cobordism between $Z_{\pi_{0},g_{0}}(P)$ and $Z_{\pi_{1},g_{1}}(P)$. This proves the invariance of $\#Z_{\pi,g}(P)$ in this case.

This argument extends to the cases $b_{1}(Y) = 0,1$ provided $(\pi_{0},0)$ and $(\pi_{1},1)$ can be connected by a path which missed the ‘bad’ sets $\mathcal{N}$, $W$ of Sect. 3.1, Lemma 14 respectively. However this is not generally true, as we shall describe below.

4.1. The case $b_{1}(Y) = 0$.

The argument in the case $b_{1}(Y) > 1$ may fail here due to the presence of a reducible (unique up to gauge) solution in each $Z_{\sigma(t)}(P)$. The reducible stratum of $Z_{\sigma}(P)$ is an arc which under $p$ projects diffeomorphically onto $I_{\varepsilon}$. The path $\sigma$ may meet the subset $\mathcal{N}$ of Prop. 10 and singularities may occur in $Z_{\sigma}$. Choose $\sigma$ to be the path defined by the family $\{g(t)\}$ and the family of perturbations $\{\pi(t)\} = \{(t-1)\pi_{0} + t\pi_{1}\}; t \in I_{\varepsilon}$. Fix $\theta$ a flat connection on $L$. Writing $\pi(t) = (\alpha(t), \omega(t))$, in the notation of Sec. 3.1, the reducible solution up to equivalence in $Z_{\sigma(t)}(P)$ is given by $\theta(t) = \theta + d^{-1}(\omega(t))$. Furthermore the associated family of Dirac operators $\{D_{0}(t)\}$ determine the cohomology.
Proposition 16. Assume \( \{D(t)\} \) is a transverse family and \( \sigma \) is transverse to the projection \( p: Z_\sigma(P) \to P_k \times I_\varepsilon \) away from \( \mathcal{N} \). Let \( \sigma \cap \mathcal{N} = \{\sigma(t_i)\}^n_{i=1} \). Then for each \( t_i \) there is an open neighbourhood \( N_i \) of \( \theta(t_i) \) such that:

1. \( Z_\sigma(P) \setminus N_i \) is a smooth compact 1-manifold with boundary
2. \( Z_\sigma(P) \cap N_i \) is diffeomorphic to the zeros of the map \( \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}, (t, \xi) \mapsto t\xi \)
3. \( Z^*_{\sigma+\pi}(P) \cap N_i = 0 \times \mathbb{R}^+ \) and the orientation of \( 0 \times \mathbb{R}^+ \) is \( -\varepsilon_i \frac{\partial}{\partial \xi} \) where \( \varepsilon_i \) is the sign of the spectral flow of \( \{D(t)\} \) at \( t_i \).

An immediate consequence of this is the formula

\[
\#Z_{\sigma(1)}(P) - \#Z_{\sigma(0)}(P) = -\text{SF}\{D(t)\}
\]

where ‘SF’ on the right denotes the total spectral flow as \( t \) varies from 0 to 1. Notice that the left-hand term is actually independent of choice of path, therefore the spectral-flow term only depends on the end-points of the path. (In fact, it is possible to verify this directly as well.)

To define an invariant in this case it is necessary to introduce a counter-term which should be a function of \( \omega \) and \( g \) which has the same change as \( \#Z_{\pi,g}(P) \) as we cross from one connected component of \( P_k \times I_\varepsilon \setminus \mathcal{N} \) to another. Such a function can be obtained from the spectral invariants of [APS].

Proposition 17. Assume \( b_1(Y) = 0 \). Let \( (\alpha, \omega, g) \) be given. Let \( \theta \) be the unique (up to gauge) flat connection on \( L \) and let \( a \) be defined by the condition \( d^*a = 0, da = \omega \). Define

\[
\zeta(\alpha, \omega, g) = \frac{1}{8} \eta \left( d \ast - \ast d \right)_{\text{even}}, g \]

\[
+ \frac{1}{2} \left( \dim \text{ker } D_{\theta+\alpha}^0 + \eta \left( D_{\theta+\alpha}^0 \right) \right) + \frac{1}{32 \pi^2} \int_Y (a + \alpha) \wedge (a + \alpha)
\]

where \( \eta \) denotes the Atiyah-Patodi-Singer spectral invariant of the associated operator. Then:

1. \( \zeta(\omega, g) \) lies in \( \mathbb{Z} \left[ \frac{1}{8 b_1(Y)} \right] \); if in addition \( H_1(Y, \mathbb{Z}) = 0 \) it lies in \( \mathbb{Z} \)
2. given the path \( \sigma \) as Prop. 16, we have \( \zeta(\sigma(1)) - \zeta(\sigma(0)) = \text{SF}\{D(t)\} \).
Thus we see that the combination

$$\#Z_{\pi,g}(P) + \zeta(\pi,g)$$

defines a topological invariant in the case $b_1(Y) = 0$.

**Proof of Proposition 17.** Every spin-c structure on $Y$ is obtained by tensoring a spin structure on $Y$ with a complex line bundle. By a Theorem of Milnor every spin $Y$ is the oriented spin boundary of an oriented spin $4$-manifold $X$ with $b_1(X) = 0$. Every complex line bundle over $Y$ can be extended over $X$; therefore we may assume a spin-c structure $P' \to X$ which induces the given $P \to Y$. We may also assume $X$ to have a metric which near the boundary which is a product $Y \times [0, \varepsilon)$ of an interval and the metric on $Y$ and with orientation $dy \wedge dt$.

We may extend the connection $\theta$ over $L(P') = \det(P')$ as the connection $\Theta$, $a$ as $\hat{a}$ and $\alpha$ over $X$. These extensions can be taken to be products over $Y \times [0, \varepsilon)$. The index theorem of [APS] applied to the Dirac operator $D_{\Theta+\hat{a}+\hat{\alpha}}$ over $X$ associated to $P'$ gives:

$$\text{Index } D^g_{\Theta+\hat{a}+\hat{\alpha}} = \int_X \exp \left( \frac{1}{2} c_1(\Theta + \hat{a} + \hat{\alpha}) \right) \hat{A} - \frac{1}{2} \left( \dim \ker D^g_{\theta+a+\alpha} + \eta(D^g_{\theta+a+\alpha}) \right).$$

Here

$$c_1(\Theta + \hat{a} + \hat{\alpha}) = \frac{i}{2\pi} F_{\Theta+\hat{a}+\hat{\alpha}}$$

and $\hat{A}$ is the $\hat{A}$-polynomial in the Pontrjagin classes. On the other hand consider the signature operator on $X$. This has index

$$\text{sig}(X) = \int_X L - \eta(d^* - d|_{\text{even}}, g).$$

$L$ is the Hirzebruch $L$-polynomial in the Pontrjagin classes. Since

$$\exp \left( \frac{1}{2} c_1(B) \right) = 1 + \frac{1}{2} c_1(B) + \frac{1}{8} c_1(B) \wedge c_1(B) + \ldots,$$

$$\hat{A} = 1 - \frac{1}{24} p_1 + \ldots,$$

$$L = 1 + \frac{1}{3} p_1 + \ldots,$$

the above index formulas give

$$\frac{1}{8} \eta(d^* - d|_{\text{even}}, g) + \frac{1}{2} \left( \dim \ker D^g_{\theta+a} + \eta(D^g_{\theta+a}) \right) + \frac{1}{32\pi^2} \int_Y (a + \alpha) \wedge d(a + \alpha).$$
\[
\frac{1}{8} \int_X c_1(\Theta) \wedge c_1(\Theta) - \frac{1}{8} \text{sig}(X) - \text{Index } D_{\Theta + \bar{a} + \bar{\alpha}}^g.
\]

If \( Y \) is an integral homology sphere then \( L \) is trivial and we may choose its extension over \( X \) as the trivial bundle; therefore \( \Theta \) in this case may be assumed trivial. Furthermore the intersection form on \( X \) is then unimodular so \( \text{sig}(X) \) is divisible by 8. Thus \( \zeta \) is an integer. When \( Y \) is not an integral homology sphere then the term \( \int_X c_1(\Theta) \wedge c_1(\Theta) \) depends only on the topological type of the extension of \( L \) over \( X \). It can be identified with the \( \mathbb{Z} \left[ \frac{1}{g_1(Y; \mathbb{Z})} \right] \)-intersection product of the class \([c_1(\Theta)] \in H^2(X; \mathbb{Z})/\text{torsion}\) with itself. Therefore \( \zeta \) takes values in \( \mathbb{Z} \left[ \frac{1}{8g_1(Y; \mathbb{Z})} \right] \).

The term \( \eta(d*\sigma + d[\text{even},g]) \) depends continuously on \( g \) whereas according to \[ \text{APS} \] \( \frac{1}{27}\eta(D_{\theta + a + \alpha}^g) \) jumps by the spectral flow. Thus \( \zeta \) has the correct behaviour as we cross components of \( P_k \times I_{1c} \setminus N \), as claimed. \( \square \)

**Proof of Proposition 16.** The proof of the proposition relies on a detailed understanding of the Kuranishi local model at the singular points \([\theta(t_i)]\). Without loss of generality we assume that there is only one value \( t = t_1 \) where \( \{ D(t) \} \) is singular. The Seiberg-Witten section which gives \( Z_\sigma(P) \) is of the form \( \hat{s} : \Omega^0(\mathbb{R}) \times \mathcal{C}(P) \times \mathbb{R} \to \Omega^1(\mathbb{R}) \oplus \Gamma_1(S) \),

\[
\hat{s}(\eta, A, \Phi, t) = \left( \ast_t(F_A - \frac{1}{4}\sigma_t(\Phi, \Phi) - \omega(t)), D_{A + \alpha(t)}^{\theta(t)} \Phi - \eta \Phi \right)
\]

where the ‘\( t \)’ in the notation denotes a dependence on \( t \). (Note: just as in Sec. 3 we work with a fixed \( P \to Y \) with respect to a basepoint metric.) The linearization of \( \hat{s} \) at \( \eta = 0, A = \theta(t_1), \Phi = 0, t = t_1 \) is given by

\[
d\hat{s}(\delta \eta, \delta a, \delta \phi, \delta t) = \left( \ast_{t_1}(d(\delta a) + \omega'(t_1)\delta t), D(t_1)\delta \phi \right).
\]

Let \( X_{\theta(t_1),0} \) be the slice of the gauge group action on \( \mathcal{C}(P) \) at \((0, \theta(t_1), 0)\) (Sec. 2). Then

\[
\ker (d\hat{s}) \cap (\Omega^0(\mathbb{R}) \times X_{\theta(t_1),0} \times \mathbb{R}) = \mathbb{H}^0(\mathbb{R}) \oplus \mathcal{H}_{\theta(t_1)} \oplus \mathbb{R}^t
\]

\[
coker (d\hat{s}) = \mathcal{H}_{\theta(t_1)}.
\]

Here \( \mathbb{R}_t = \text{span}\{ (d^{-1}(\omega'(t_1)), 1) \} \) in the \( \Omega^1(\mathbb{R}) \oplus \mathbb{R} \) factor. The Kuranishi obstruction map then takes the form

\[
\Xi : i\mathbb{R} \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}_t \supset U \to \mathcal{H}_{\theta(t_1)}.
\]

This gives \( \hat{s}^{-1}(0) / \mathbb{G} \) near \((0, \theta(t_1), 0, t_1)\) as \( \Xi^{-1}(0) / S^1 \). A direct verification shows that \( \Xi^{-1}(0) \supset (i\mathbb{R} \times 0 \times \mathbb{R}_t) \cap U \). The subset \((i\mathbb{R} \times 0 \times 0) \cap U \) consists of ‘virtual’ Seiberg-Witten solutions and thus should be ignored to get the Seiberg-Witten moduli space proper (Lemma 3). The subset \((0 \times 0 \times \mathbb{R}_t) \cap U \) are the reducible solutions near \((0, \theta(t_1), 0, t_1)\). Our assumption on \( \sigma \) being transverse to \( p \) away from \( W \) means that the closure of the irreducible part
of $\Xi^{-1}(0)/S^1$, is a compact 1-manifold with boundary except possibly at $(0,0,0)$. Furthermore $\Xi^{-1}(0) \cap U \cap (\mathbb{R} \times 0 \times \mathbb{R}_t - (0,0,0)) = \emptyset$.

By construction the derivative of $\Xi$ at $(0,0,0)$ is the zero map. We aim to compute the second derivative: This will give us the quadratic approximation to $\Xi$ which will be sufficient for our purposes. In the following, we identify $\mathbb{R}_t$ with $\mathbb{R}$ via $t \mapsto (\omega(t_1) - 1, 1)$.

**Claim 18.** The second derivative of $\Xi$ at $(0,0,0)$ is given by

$$D^2\Xi(\delta \eta, \delta \phi, \delta t) = c \delta t \delta \phi - \delta \eta \delta \phi$$

where $c$ is a non-zero real constant and has the same sign as that of the spectral flow of $\{D(t)\}$ at $t = t_1$.

To prove the claim: the obstruction map $\Xi$ is constructed as a composition of the form $x \mapsto \Pi \circ \hat{s}(x + f(x))$ where $x \in \mathcal{O} \subset i\mathbb{R} \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}_t$ and $f: \mathcal{O} \to (i\mathbb{R} \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}_t)^\perp$ is given by the implicit function theorem. As such its derivative at 0 is the zero map. It is then seen that $D^2\Xi$ is given by $\Pi \circ D^2\hat{s}$. This is given by the expression

$$D^2\Xi(\delta \eta, \delta \phi, \delta t) = \Pi (D'(t_1)(\delta t \delta \phi)) - \delta \eta \delta \phi.$$  

The map $\delta \phi \mapsto \Pi(D'(t_1)\delta \phi)$ defines a Hermitian transformation on $\mathcal{H}_{\theta(t_1)} \cong \mathbb{C}$ with respect to the complex $L^2$-inner product. Thus it is multiplication by a real constant $c = \langle D'(t_1)v, v \rangle$, $v$ being of unit length. Our assumption on $\sigma$ was that at $t = t_1$ a single eigenvalue $\lambda(t)$, $|t - t_1| < \delta$, for $D(t)$ changed from negative to positive or vice-versa. In the first case the spectral flow is $+1$ and in the latter $-1$. We have a 1-parameter family of unit eigenvectors $v(t)$, $|t - t_1| < \delta$, such that

$$D(t)v(t) = \lambda(t)v(t).$$

Differentiating this equation at $t = t_1$ and taking the inner product with $v(t_1)$ we obtain using self-adjointness of $D(t)$,

$$\langle D'(t_1)v(t_1), v(t_1) \rangle = \lambda'(t_1).$$

Thus the sign of the spectral flow is seen to be same as that of $\lambda'(t_1)$. This proves the Claim.

To continue the proof of Proposition 16: by the Claim, $\Xi(\eta, \phi, t)$ is approximated up to second order by

$$(ct - \eta)\phi.$$  

The zeros of the quadratic approximation fall into two branches: $0 \times 0 \times \mathbb{R}_t$ and $0 \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}$. We claim that these two branches gives a complete picture of the zeros of $\Xi$ near $(0,0,0)$. To get a clearer picture, consider restricting the $\phi$ variable to the real span of a fixed non-zero vector in $\mathcal{H}_{\theta_0}$. Call this map $\tilde{\Xi}$. Then $\Xi^{-1}(0)/S^1 = \tilde{\Xi}^{-1}(0)/\pm 1$. Since $\tilde{\Xi}(\eta,0,t) = 0$ we may factor out the branch $\{\phi = 0\}$ by setting $\tilde{\Xi}(\eta, \phi, t) = \phi.\Theta(\eta, \phi, t)$. Using
Claim 18 the linearization of $\Theta$ is seen to be $d\Theta(\delta\eta, \delta t, \delta \phi) = (c\delta t - \delta \eta)$. We have $\ker d\Theta = \{\delta t = \delta \eta = 0\}$. Invoking the Implicit Function Theorem we see that near $(0,0,0)$, $\Theta^{-1}(0)$ is a smooth arc tangent to $\{\phi = 0\}$ at $(0,0,0)$. This demonstrates the claimed local structure near the singular point. For later we note the following: the implicit function theorem gives a map $G : \mathcal{H}_{\theta(t_1)} \to \mathbb{R}_t = \{0\} \times \mathbb{R}_t \subset i\mathbb{R} \times \mathbb{R}_t$ such that the closure of the irreducible part of $\Xi^{-1}(0)$ is given by $\text{graph}(G) = \{(0,\phi,G(\phi)) \mid \phi \in \mathcal{H}_{\theta(t_1)}\}$. What remains is to determine the orientation of the above arc of irreducible solutions. We have an orientation of $\det \text{Ind}(L_{\eta,A,\phi,t})$ determined by that of $\det \text{Ind}(L_{\eta,A,\phi})$ according to the map $h$ of (3). (See discussion following there.) Since we are working at a point where $\Phi = 0$, the orientation of $\det \text{Ind}(L_{\eta,A,\phi})$ is determined by the kernel and cokernel of $L_{\theta(t_1),0}$; namely $\mathbb{H}^0(i\mathbb{R}) \oplus \mathcal{H}_{\theta(t_1)}$. The long exact sequence inducing $h$ takes the form

$$0 \to \mathbb{H}^0(i\mathbb{R}) \oplus \mathcal{H}_{\theta(t_1)} \to \ker L_{\theta(t_1),0,t_1} \xrightarrow{\kappa} \mathbb{R} \to \mathbb{H}^0(i\mathbb{R}) \oplus \mathcal{H}_{\theta(t_1)} \xrightarrow{\beta} \text{coker } L_{\theta(t_1),0,t_1} \to 0.$$

The map $\kappa$ in the sequence sends the subspace $\mathbb{R}_t$ isomorphically onto the target space. This isomorphism sends $(1,d^{-1}(\omega(t_1))) r \mapsto r, \ r \in \mathbb{R}$. Choose an orientation of $\text{coker } L_{\theta(t_1),0} = \mathbb{H}^0(i\mathbb{R}) \oplus \mathcal{H}_{\theta(t_1)}$; then since $\beta$ in the sequence is an isomorphism, our orientation convention dictates that $\text{coker } L_{\theta(t_1),0,t_1}$ has the orientation induced by $\beta$, and the orientation on $\ker L_{\theta(t_1),0,t_1}$ is the product orientation $\ker L_{\theta(t_1),0} \oplus \mathbb{R}_t$, where $\mathbb{R}_t$ is oriented via $\kappa$.

In order to determined orientations in the local Kuranishi picture correctly we shall need to combine the obstruction map $\Xi$ with a local slice condition coming from the $S^1$-action, which is the inverse of complex multiplication on the $\mathcal{H}_{\theta(t_1)}$ factor. The set $\text{graph}(G) \subset \Xi^{-1}(0)$ represents the closure of the $S^1$-orbits of the irreducible solutions near $(0,0,0)$. Let $0 \neq v \in \mathcal{H}_{\theta(t_1)}$. Then the linearization of the $S^1$-action at $(0,v,G(v))$ is a map $i\mathbb{R} \to i\mathbb{R} \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}_t$, $\delta \gamma \mapsto (0,(-\delta \gamma)v,0)$. The adjoint of this map sends $\delta \eta, \delta \phi, \delta t \mapsto -i(iv, \delta \phi)$. Therefore a further local description for $\Xi^{-1}(0)/S^1$ near $(0,v,G(v))$ is the zeros of the map

$$\chi : i\mathbb{R} \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}_t \to i\mathbb{R} \times \mathcal{H}_{\theta(t_1)}, \quad \chi(\eta,\phi,t) = (\eta,\phi,\chi(\eta,\phi,t)).$$

In what follows, we may for simplicity assume $G = 0$; the result for general $G$ is obtained by working sufficiently close to $(0,0,0)$ where $\text{graph}(G)$ is approximated to arbitrarily high order by $0 \times \mathcal{H}_{\theta(t_1)} \times 0$. With this assumed, the irreducible zeros of $\chi$ is the set of positive multiples of $(0,v,0)$. The normal bundle to $0 \times \mathcal{H}_{\theta(t_1)} \times 0$ at $(0,0,0)$ is $i\mathbb{R} \times 0 \times \mathbb{R}_t$. This is mapped via $d\Theta$ isomorphically onto $\mathcal{H}_{\theta(t_1)}$. If we pull-back the complex orientation by $d\Theta$, then the induced orientation on $i\mathbb{R} \times 0 \times \mathbb{R}_t$ is $c\delta \eta \wedge \delta t$. By continuity this is carried to the point $v$ as the same orientation. The last remaining direction
to the normal bundle of $\chi^{-1}(0)$ at $v$ is given by $(0, iv, 0)$. This is mapped by $d\chi$ to $(-i(iv, iv), 0) = (-i, 0)$. Let us take the product orientation on $i\mathbb{R} \times \mathcal{H}_{\theta(t_1)}$ for the range space of $\chi$, in this order (the final answer is independent of this choice); then our orientation convention dictates that the domain space of $\chi$ is oriented in the order $i\mathbb{R} \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}_t$. Then the pull-back of the orientation on $i\mathbb{R} \times \mathcal{H}_{\theta(t_1)}$ to the normal bundle of $\chi^{-1}(0)$ at $v$ is $-\delta \theta \wedge c \delta \eta \wedge \delta t$, where $\delta \theta$ is the $(0, iv, 0)$ direction. Let $\delta r$ be the direction given by $v$, and $\varepsilon \delta r$ the induced orientation of $\chi^{-1}(0)$ near $v$. Then we require that the orientation on $\chi^{-1}(0)$ followed by the orientation in the normal direction equals the orientation on $i\mathbb{R} \times \mathcal{H}_{\theta(t_1)} \times \mathbb{R}_t$, that is

$$
\varepsilon \delta r \wedge (-\delta \theta \wedge c \delta \eta \wedge \delta t) = \delta \eta \wedge \delta r \wedge \delta \theta \wedge \delta t.
$$

This shows the induced orientation on $\chi^{-1}(0)$ as $-c \delta r$, as claimed.

4.2. The case $b_1(Y) = 1$.

This case is similar but with some slight differences to $b_1(Y) = 0$. Any path $\sigma$ which connects $(\pi_0, g_0)$ to $(\pi_1, g_1)$ in $\mathcal{P}_k \times I_\varepsilon$ may cross the codimension 1 subset $\mathcal{W}$. At the points where $\sigma$ meets $\mathcal{W}$ the corresponding $Z_{\sigma(t)}(P)$ will admit an $S^1$'s worth of reducibles, otherwise $Z_{\sigma(t)}(P)$ contains no reducibles. As following our notation conventions, $\sigma(t)$ in components is $(\pi(t), g(t))$ or $(\alpha(t), \omega(t), g(t))$.

We may by general transversality arguments assume that $\sigma$ meets $\mathcal{W}$ transversely and orthogonally and transverse to the projection $p : Z^*(P) \to \mathcal{P}_k \times I_\varepsilon$ away from $\mathcal{W}$. Let $\{t_i\}$ be the finite set of values for which $\sigma(t_i) \in \mathcal{W}$. To simplify matters even more, since $\mathcal{W}$ the preimage of a set in $\mathcal{P}_k$ we may assume near $\mathcal{W}$ that $\sigma$ lies in the subset $\mathcal{P}_k \times \{t_i\}$. Hence for values near $t_i$, the metric represented by $\sigma$ is unchanging.

We can always find connections $A_i$ such that $F_{A_i} = \omega(t_i)$. Using the value $\alpha(t_i)$, we can as in Sec. 3.1 form the 1-parameter family of operators $\{D_{\alpha_i}(s)\}$. By Prop. 11 we can make this family transverse by an arbitrarily small perturbation $\pi_i$. This perturbation can be achieved by a perturbation of $\sigma$, supported for values of $t$ near $t_i$, and maintaining the original properties of $\sigma$. Thus we can assume $\{D_{\alpha_0}(s)\}$ is a transverse family and we drop the '0' subscript notation. Finally let $s_{i,j}$ be the values of $s$ for which $\{D_i(s)\}$ has spectral flow. Denote by $A_{i,j}$ the connection $A_i + s_{i,j}a_0$.

A technical issue which will be significant is the orientation of the family $\{D_i(s)\}$. Looking back at the definition in Sec. 3.1 we see that this involved a certain choice of a non-zero element $a_0$ in $H^1(i\mathbb{R})$. We shall make a specific choice for each $i$. The assumption that $\sigma$ meets $\mathcal{W}$ orthogonally means in particular that the derivative $\omega'(t_i)$ is $L^2$-orthogonal to the exacts. Thus $d^*(\omega'(t_i)) = 0$ so $\ast \omega'(t_i)$ is closed. For a chosen $i$ we now make the convention that the $a_0$ should be a positive multiple of $[- \ast \omega'(t_i)]$. 

200 YUHAN LIM
Proposition 19. For each \((i,j)\) there is a open neighbourhood \(N_{i,j}\) of \([A_{i,j}]\) such that (i) \(\overline{Z_\sigma(P)} \setminus N_{i,j}\) is a smooth compact 1-manifold with boundary (ii) \(N_{i,j}\) is diffeomorphic to the zeros of the map \(\mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}\), \((s, \xi) \mapsto s\xi\) with \(Z_\sigma^2(P) \cap N_{i,j} = 0 \times \mathbb{R}^+\) (iii) the orientation of \(0 \times \mathbb{R}^+\) is \(\varepsilon_{i,j} \partial_\mathbb{R}\) where \(\varepsilon_{i,j}\) is the sign of the spectral flow of \(\overline{\mathcal{D}}(s)\) at \(s_{i,j}\), as \(s\) varies from 0 to 1.

Proof. This largely proceeds in the manner of the case \(b_1(Y) = 0\). We continue to use notation introduced there. Again, without loss, we may assume \(\sigma\) meets \(W\) exactly once, at \(t = t_1\). We consider again the map \(\bar{s}\) of the case \(b_1(Y) = 0\). Computed at \(\eta = 0\), \(A = A_{1,j}, \Phi = 0\), this time we find

\[
\ker (d\bar{s}) \cap (\Omega^0_\mathbb{R}(i\mathbb{R}) \times X_{A_{1,j}} \times \mathbb{R}) = i\mathbb{R} \oplus H^1(i\mathbb{R}) \oplus \mathcal{H}_{A_{1,j}},
\]

\[
coker (d\bar{s}) = \mathcal{H}_{A_{1,j}}.
\]

We have the obstruction map \(\Xi : i\mathbb{R} \times H^1(i\mathbb{R}) \times \mathcal{H}_{A_{1,j}} \to \mathcal{H}_{A_{1,j}}\) whose second derivative at \((0,0,0)\) is

\[
D^2\Xi(\delta \eta, \delta a, \delta \phi) = \frac{1}{2}\Pi(\delta a \cdot \delta \phi) - \delta \eta \delta \phi.
\]

Then if we let \(\delta a = -* \omega(t_1) \delta t\), the term

\[
\frac{1}{2}\Pi(\delta a \cdot \delta \phi) = \frac{1}{2}\delta t \Pi(-* \omega(t_1) \cdot \delta \phi) = c \delta t \delta \phi
\]

where \(c\) is a non-zero real constant with the same sign of the spectral flow \(\{\overline{\mathcal{D}}(s)\}\) at \(s_{1,j}\). As in the \(b_1 = 0\) case, the irreducible zeros of \(\Xi\) are modelled by the subset \(0 \times 0 \times \mathcal{H}_{A_{1,j}}\) and the reducible zeros by \(0 \times H^1(i\mathbb{R}) \times 0\).

Let us now deal with the orientations in this case. Looking at the long exact sequence inducing \(h\) of (3) we see

\[
0 \to H^0(i\mathbb{R}) \oplus H^1(i\mathbb{R}) \oplus \mathcal{H}_{A_{1,j}} \xrightarrow{\Xi} \ker L_{A_{1,j},0,t_1} \to \mathbb{R} \xrightarrow{\kappa} H^0(i\mathbb{R}) \oplus H^1(i\mathbb{R}) \oplus \mathcal{H}_{A_{1,j}} \xrightarrow{\beta} \text{coker } L_{A_{1,j},0,t_1} \to 0.
\]

We note that \(\kappa(t) = -\omega(t_1) t\) maps isomorphically onto the \(H^1(i\mathbb{R})\) factor, and \(\text{coker } L_{A_{1,j},0,t_1}\) is \(H^0(i\mathbb{R}) \oplus H^1(i\mathbb{R})\). \(\beta\) is the obvious projection. Let us assume the canonical orientations on \(H^0(i\mathbb{R}), \mathcal{H}_{A_{1,j}}\), and the orientation on \(H^1(i\mathbb{R})\) induced by \(\kappa\), which is given by \(-* \omega(t_1)\). Finally choose the product orientation (in the order indicated) on \(H^0(i\mathbb{R}) \oplus H^1(i\mathbb{R}) \oplus \mathcal{H}_{A_{1,j}}\).

Then \(\ker L_{A_{1,j},0,t_1}\) is identically oriented and \(\text{coker } L_{A_{1,j},0,t_1}\) is oriented according to the order \(H^0(i\mathbb{R}) \oplus \mathcal{H}_{A_{1,j}}\).

Let \(v \in \mathcal{H}_{A_{1,j}}\). Then combining the slice condition with \(\Xi\) gives the moduli space near \((0,0,v)\) (as before we may assume \(G = 0\)) as the zeros of the map

\[
\chi : i\mathbb{R} \times H^1(i\mathbb{R}) \times \mathcal{H}_{A_{1,j}} \to i\mathbb{R} \times \mathcal{H}_{A_{1,j}}\],
\]

\[
\chi(\eta, a, \phi) = (-i\langle iv, \phi\rangle, \Xi(\eta, a, \phi)).
\]
As before the zeros of $\chi$ are the positive multiples of $(0, 0, v)$. The pull-back of the orientation on the target space onto the normal bundle of $\chi^{-1}(0)$ is given by $-\delta \theta \wedge c \delta \eta \wedge \delta a$ where $\delta \theta$ is the angular coordinate on $H_{A_{1,j}}$. Letting $r$ be the direction determined by $v$ and $\varepsilon r$ the induced orientation, then we require

$$\varepsilon r \wedge (-c \delta \theta \wedge \delta \eta \wedge \delta a) = \delta \eta \wedge \delta a \wedge \delta r \wedge \delta \theta$$

which gives the induced orientation on $\chi^{-1}(0)$ as $-c \delta r$. \hfill $\Box$

As mentioned before, $Z_{\sigma(t)}(P)$ admits reducible solutions exactly $\sigma(t) \in \mathcal{W}$. This corresponds to when $\left[ \frac{i}{2\pi} \omega(t) \right]$ coincides with the class $c_1(L)_{\mathbb{R}}$. Let $\mathcal{U}$ denote a connected component of $H^1(Y; \mathbb{R}) - \{ c_1(L)_{\mathbb{R}} \}$. Then if our path $\sigma$ has the property that $\left[ \frac{i}{2\pi} \omega(t) \right] \in \mathcal{U}$ for all $t$, then $\# Z_{\sigma(0)}(P) = \# Z_{\sigma(0)}$. Therefore $\# Z_{a, \omega, g}(P)$ is an integer-valued function depending only on the choice of $\mathcal{U}$. Denote this function as $\tau(\mathcal{U})$.

We think of $\{ c_1(L)_{\mathbb{R}} \}$ as a ‘wall’ in $H^2(Y; \mathbb{R})$. Then as we cross this wall $\tau$ changes. This change can be determined from the previous proposition to give a ‘wall-crossing’ formula.

**Corollary 20.** Let $a \in H^2(Y; \mathbb{Z})/\text{torsion}$ be an indivisible class and let $c_1(L)_{\mathbb{R}} = 2na$. Let $\mathcal{U}_\pm$ be the component of $H^2(Y; \mathbb{R}) - \{ c_1(L)_{\mathbb{R}} \}$ containing $(2n \pm 1/2)a$. Then

$$\tau(\mathcal{U}_+) - \tau(\mathcal{U}_-) = n.$$

**Proof.** Take $(\pi_0, g_0)$ and $(\pi_1, g_1)$ which define the values $\tau(\mathcal{U}_+)$ and $\tau(\mathcal{U}_-)$ respectively. Choose our connecting path $\sigma$ with properties as used for as for Prop. 19. Without loss, We may suppose that $\sigma$ crosses $W$ exactly once, say at $t = t_1$. We now follow the notation and ideas in the proof of Prop. 19. According to Prop. 19 We need then to compute the total spectral flow of the family $\{ \overline{D}^1(s) \}$ as $s$ varies from 0 to 1. The orientation of this family is determined by $-\ast \omega'(t_1)$. We shall choose $a$ to be consistent with this orientation, but the statement of the corollary is actually independent of this choice. Take a positive multiple $\omega$ of $\omega'(t_1)$ such that with $\left[ \frac{i}{2\pi} \omega \right] = 2a$. Thus $A_1 - s\omega$, $0 \leq s < 1$ parameterizes all the reducibles in $Z_{\sigma(t_1)}(P)$.

We may deform the family $\{ \overline{D}^1(s) \}$ preserving self-adjointness to the family $\{ D_{A_{1,s\omega}}^{(t_1)} \}$, $s \in [0, 1]$. Thus it suffices to compute the spectral flow for this family. Notice that there is a gauge transformation $g$ such that $g(A_1) = A_1 - s\omega$, or equivalently $g^{-1}dg = -s\omega$. A theorem of [APS] says that the spectral flow of the Dirac operators $\{ D_{A_{1,s\omega}}^{(t_1)} \}$ is equivalent to computing the index of a Dirac operator $D_{A_1}^{(4)}$ on $Y \times S^1$ with a spin-$c$ structure obtained by taking the product $P \times [0, 1]$ over $Y \times [0, 1]$ and identifying via $g : P \times \{ 1 \} \to P \times \{ 0 \}$. $A$ is a connection which is in temporal gauge and coincides with $A - s\omega$ on $L \times \{ s \}$. (Remark: We follow the orientation conventions of [APS] closely.
In particular $Y \times S^1$ has the product orientation $dy \wedge ds$ where $dy$ is the orientation form on $Y$ and $s$ the real coordinate on $S^1$ thinking of it as $\mathbb{R}/\mathbb{Z}$.) Denote the resulting determinant on $Y \times S^1$ by $L'$. The index of $D^{(4)}_A$ is given by

$$\frac{1}{8} \left< c_2(L'), [Y \times S^1] \right> + \frac{1}{8} \text{sig}(Y \times S^1).$$

To compute the first term, we notice $F_A = d(A_1 - \ast \omega s) = F_{A_1} - \ast \omega ds$. Then

$$\left< c_1^2(L'), [Y \times S^1] \right> = \int_{Y \times [0,1]} \frac{i}{2\pi} F_A \wedge \frac{i}{2\pi} F_A$$

$$= -\frac{1}{4\pi^2} \int_{Y \times [0,1]} (F_{A_1} - \ast \omega ds) \wedge (F_{A_1} - \ast \omega ds)$$

$$= -\frac{1}{4\pi^2} (-2) \int_{Y \times [0,1]} F_{A_1} \wedge \ast \omega ds$$

$$= -\frac{1}{4\pi^2} (-2) \int_{Y \times [0,1]} n \omega \wedge \ast \omega ds$$

$$= -\frac{1}{4\pi^2} (-2n) \int_Y \omega \wedge \ast \omega \int_0^1 ds$$

$$= -\frac{1}{4\pi^2} (-2n) 4\pi ia \cdot PD(4\pi ia)$$

$$= 8n.$$  

Here ‘PD’ denotes Poincare Duality. Since $\text{sig}(Y \times S^1) = 0$, the index of $D^{(4)}_A$ is $n$ and the corollary follows. $\square$

References


Received February 1, 1998 and revised May 14, 1999.

University of California
Santa Barbara, CA 93106
E-mail address: ylim@gauss.math.ucsb.edu
VIRTUAL HOMOLOGY OF SURGERED TORUS BUNDLES

JOSEPH D. MASTERS

Let M be a once-punctured torus bundle over $S^1$ with monodromy $h$. We show that, under certain hypotheses on $h$, “most” Dehn-fillings of $M$ (in some cases all but finitely many) are virtually $\mathbb{Z}$-representable. We apply our results to show that even surgeries on the figure eight knot are virtually $\mathbb{Z}$-representable.

1. Introduction.

Embedded incompressible surfaces are fundamental in the study of 3-manifolds. Accordingly, the following conjecture of Waldhausen and Thurston has attracted much attention:

**Conjecture 1.1.** Let $M$ be a closed, irreducible 3-manifold with infinite $\pi_1$. Then $M$ has a finite cover which is Haken.

The focus of this paper is the following, stronger, conjecture:

**Conjecture 1.2.** Let $M$ be as above. Then $M$ has a finite cover $\tilde{M}$ with $H_1(\tilde{M}, \mathbb{Z})$ infinite.

If $M$ is a compact 3-manifold, we say that $M$ is $\mathbb{Z}$-representable if $H_1(M, \mathbb{Z})$ is infinite. If $M$ satisfies the conclusion of Conjecture 1.2, we say that $M$ is virtually $\mathbb{Z}$-representable.

We shall give what appear to be the first examples of 3-manifolds with torus boundary for which all but finitely many fillings are virtually $\mathbb{Z}$-representable, but not $\mathbb{Z}$-representable (in fact non-Haken). Boyer and Zhang have independently given examples of knot complements for which all but finitely many fillings are virtually Haken, but non-Haken [BZ].

Before we can state our results, we must establish some notation. Let $F$ be a once-punctured torus with $\pi_1(F) = \langle [x], [y] \rangle$, and basepoint $x_0 \in \partial F$ (see Fig. 1).

Any orientation-preserving homeomorphism $h : F \to F$ is isotopic to one of the form $h = D_x^{r_1}D_y^{s_1} \cdots D_x^{r_k}D_y^{s_k}$. Here $D_x$ and $D_y$ are Dehn twists along simple closed curves homologous to $x$ and $y$, respectively. The twists $D_x$
and $D_y$ induce the following actions on $\pi_1(F)$:

\[
\begin{align*}
D_{x^\sharp}(x) &= x \\
D_{x^\sharp}(y) &= yx \\
D_{y^\sharp}(x) &= yx \\
D_{y^\sharp}(y) &= y.
\end{align*}
\]

We may assume $h$ fixes $\partial F$. Let $M_h = (F \times I)/h$ be the once-punctured torus bundle with monodromy $h$. We specify a framing for $H_1(\partial M_h, \mathbb{Z})$ by setting the longitude $\beta = \partial F$ oriented counter-clockwise, and the meridian $\alpha = (x_0 \times I)/h$, where $x_0$ is some point on $\partial F$, and $\alpha$ is oriented as in Fig. 1. Then, for coprime integers $(\mu, \lambda)$, $M_h(\mu, \lambda)$ denotes the manifold obtained by gluing a solid torus to $M_h$ in such a way that the curve $\alpha^\mu \beta^\lambda$ becomes homotopically trivial.

We shall prove:

**Theorem 1.3.** Let $M_h$ be a once-punctured torus bundle over $S^1$, with monodromy $h = D_{x^1}^r D_{y^1}^s \cdots D_{x^k}^r D_{y^k}^s$, and let $n = \gcd\{s_1, \ldots, s_k\}$, $R = r_1 + \cdots + r_k$. 

(i) If $n$ is divisible by some $m$ such that $m \geq 6$ and $m$ is even or $m = 7$, and if $|\lambda| > 1$, then all but finitely many Dehn-fillings $M_h(\mu, \lambda)$ are virtually $\mathbb{Z}$-representable.

(ii) If $n$ is divisible by some $m$ such that $m \geq 5$, $m$ is odd, and $m \neq 7$, and if $1/|R\mu - \lambda| + 1/|R\mu - 2\lambda| + 1/|\lambda| < 1$, then $M_h(\mu, \lambda)$ is virtually $\mathbb{Z}$-representable.

(iii) If $n$ is divisible by 4, and if $2/|R\mu - 2\lambda| + 1/|\lambda| < 1$, then $M_h(\mu, \lambda)$ is virtually $\mathbb{Z}$-representable.
Remarks. 1. Analogous results hold if we replace $n$ by $\gcd\{r_1, \ldots, r_k\}$ and $R$ by $s_1 + \cdots + s_k$.

2. It was shown in [B1] that if $m \geq 2$, $n \geq 2$ and $mn \geq 8$ but $mn \neq 9$, then all non-integral surgeries are virtually $\mathbb{Z}$-representable. In [B2] it was shown that if $4|n$, then for each $\mu$, $M_h(\mu, \lambda)$ is virtually $\mathbb{Z}$-representable for all but finitely many $\lambda$ coprime to $\mu$.

3. From [CJR] and [FH], all but finitely many surgeries on a once-punctured torus bundle over $S^1$ yield non-Haken manifolds.

Theorem 1.3 may be used to show that, for certain choices of $f$, all but finitely many surgeries on $M_f$ are virtually $\mathbb{Z}$-representable. For example:

**Theorem 1.4.** Let $f = (D_xD_y)^{18}$. Then every surgery on $M_f$ is virtually $\mathbb{Z}$-representable.

The proof of Theorem 1.4 appears in Section 3.

In order to state the next theorem, we require some notation. Let $-1 = (D_xD_y^{-1}D_x)^2$, the central involution on the punctured torus. If $h$ is a homeomorphism of the punctured torus, $-h$ stands for $(-1)h$.

**Theorem 1.5.** Let $N = M_{-D_xD_y}$ (also known as “the figure eight knot’s sister”). Then if $1/|\mu - \lambda| + 1/|\mu - 2\lambda| + 1/|\lambda| < 1$, $N(\mu, \lambda)$ is virtually $\mathbb{Z}$-representable.

**Theorem 1.6.** Let $K$ denote the figure-eight knot and let $M$ denote $S^3 - K$. Then, with respect to the canonical framing of knots in $S^3$, any surgery of the form $M(2\mu, \lambda)$ is virtually $\mathbb{Z}$-representable.

Other results on virtually $\mathbb{Z}$-representable figure-eight knot surgeries may be found in [M], [KL], [H], [N] and [B3]. In particular, it was shown in [KL] and [B3] that surgeries of the form $M(4\mu, \lambda)$ are virtually $\mathbb{Z}$-representable. It was also shown in [B3] that surgeries of the form $M(2\mu, \lambda)$ are virtually $\mathbb{Z}$-representable if $\lambda = \pm 7\mu \pmod{15}$. Finally, it was shown in [Bart] that every non-trivial surgery of $M$ contains an immersed incompressible surface.

Our techniques are extensions of Baker’s. The main new ingredient is the use of group theory to encode the combinatorics of cutting and pasting.

I would like to thank Professor Alan Reid for his help and patience.

2. Construction of covers.

We begin by recalling Baker’s construction of covering spaces of $M_h(\mu, \lambda)$ (see [B1], [B2]). Let $n$ be as in the statement of Theorem 1.3, and let $\hat{F}$ be the $kn$-fold cover of $F$ associated to the kernel of the map $\phi: \pi_1(F) \to \mathbb{Z}_k \times \mathbb{Z}_n$, with $\phi([x]) = (1, 0)$ and $\phi([y]) = (0, 1)$ (see Fig. 2).

Now create a new cover, $\tilde{F}$, of $F$ by making vertical cuts in each row of $\hat{F}$, and gluing the left side of each cut to the right side of another cut in the
same row. An example is pictured in Figure 3, where the numbers in each row indicate how the edges are glued.

If \( h \) lifts to a map \( \tilde{h} : \tilde{F} \to \hat{F} \), then the mapping cylinder \( \tilde{M}_h = \tilde{F}/\tilde{h} \) is a cover of \( M_h \). Furthermore, if the loop \( \alpha^\mu \beta^\lambda \) lifts to loops in \( \tilde{M}_h \), then the cover extends to a cover \( \tilde{M}_h(\mu,\lambda) \) of \( M_h(\mu,\lambda) \).

If the cover \( \tilde{M}_h \) exists, then we may compute its first Betti number with the formula

\[
b_1(\tilde{M}_h) = \text{rank} (\text{fix}(\tilde{h}^*)),
\]

where \( \tilde{h}^* \) is the map on \( H_1(\tilde{M},\mathbb{Z}) \) induced by \( \tilde{h} \), and \( \text{fix}(\tilde{h}^*) \) is the subgroup of \( H_1(\tilde{M},\mathbb{Z}) \) fixed by \( \tilde{h}^* \) (see [H] for a proof). We shall use this formula to prove that, in some cases, \( b_1(\tilde{M}) \) is greater than the number of boundary components of \( \tilde{M} \), which ensures that \( b_1(\tilde{M}(\mu,\lambda)) > 0 \).

We now introduce some notation to describe the cuts of \( \tilde{F} \) (see Fig. 3). \( \tilde{F} \) is naturally divided into rows, which we label 1, \ldots, n. The cuts divide each row into pieces, each of which is a square minus two half-disks; we number them 1, \ldots, k. If we slide a point in the top half of the \( i^{th} \) row through the cut to its right, we induce a permutation on \( \{1, \ldots, k\} \), which we denote

---

**Figure 2.** The cover \( \hat{F} \) of \( F \).
σ_i. Thus the cuts on \( \tilde{F} \) may be encoded by elements \( \sigma_1, \ldots, \sigma_n \in S_k \), the permutation group on \( k \) letters.

Next, we find algebraic conditions on the \( \sigma_i \)'s which will guarantee that the cover of \( F \) extends to a cover of \( M(\mu, \lambda) \). We first must pick \( k, n, \) and \( \{\sigma_1, \ldots, \sigma_n\} \) so that \( h \) lifts to \( \tilde{F} \).

**Lemma 2.1.** If

I. \( [\sigma_i, \sigma_1 \sigma_2 \cdots \sigma_{i-1}] = 1 \) for all \( i \) and

II. \( \sigma_1 \sigma_2 \cdots \sigma_n = 1 \)

then \( h \) lifts to \( \tilde{F} \).

**Proof.** Note that \( D^n_x \) lifts to Dehn twists on \( \tilde{F} \). Therefore, we need only ensure that \( D_x \) lifts. We shall attempt to lift \( D_x \) to a sequence of “fractional Dehn twists” along the rows of \( \tilde{F} \). Let \( \tilde{x}_i \) denote the disjoint union of the lifts of \( x \) to the \( i^{th} \) row of \( \tilde{F} \). We first attempt to lift \( D_x \) to row 1, twisting \( 1/k^{th} \) of the way along \( \tilde{x}_1 \). Considering the action on the bottom half of row 1, we find that the cuts are now matched up according to the permutation \( \sigma_1^{-1} \sigma_2 \sigma_1 \). Thus, for \( D_x \) to lift to row 1 we assume \( \sigma_1 \) and \( \sigma_2 \) commute. We now twist along \( \tilde{x}_2 \). The top halves of the squares in row 2 are moved according to the permutation \( \sigma_1 \sigma_2 \), and the lift will extend to all of row 2 if and only if \( \sigma_3 \) commutes with \( \sigma_1 \sigma_2 \). We continue in this manner, obtaining the conditions in I. After we twist through \( \tilde{x}_n \), we need to be back where we started in row 1, so we require the additional condition \( \sigma_1 \sigma_2 \cdots \sigma_n = 1 \). □

Note that the loop \( \alpha^\mu \) lifts homeomorphically to loops in \( \tilde{M}_h \) if \( \tilde{h}^\mu = \text{id} \), and that the loop \( \beta^\lambda \) lifts to loops in \( \tilde{M}_h \) if \( (\sigma_{i+1} \sigma_i^{-1})^\lambda = \text{id} \) for all \( i = 1, \ldots, n \). Then, by considering the action of \( \tilde{h} \) on \( \tilde{M}_h \), the following condition for a loop in \( \partial M_h \) to lift to \( \tilde{M}_h \) is easily verified:
Lemma 2.2. The loop $\alpha^\mu \beta^\lambda \subset \partial M_h$ lifts homeomorphicly to loops in $\tilde{M}_h$ if and only if

III. $(\sigma_1 \cdots \sigma_i)^{R^\mu}(\sigma_{i+1}^{-1}\sigma_i^{-1})^\lambda = 1$, for $i = 1, \ldots, n$.

Therefore we may construct covers of $M_h(\mu, \lambda)$ simply by finding permutations satisfying conditions I-III.

Proof of Theorem 1.3.

Case 1. $m = 4$.

Construction of the cover of $M_h(\mu, \lambda)$.

To construct a cover of $M_h(\mu, \lambda)$, we must first construct a cover of $F$. It was shown in the discussion prior to Lemma 2.1 that there is a unique such cover associated to any four permutations $\sigma_1, \sigma_2, \sigma_3$ and $\sigma_4$ in any permutation group $S_k$.

To ensure that the cover of $F$ extends to a cover of $M_h$, we shall set $\sigma_2 = \sigma_1^{-1}$ and $\sigma_4 = \sigma_3^{-1}$ (see Fig. 4a). Then conditions I and II of Lemma 2.1 are satisfied automatically, so that any choice of $\sigma_1$ and $\sigma_3$ will determine a cover of $M_h$.

To ensure that the cover extends to $M_h(\mu, \lambda)$, we must arrange for the surgery curve $\alpha^\mu \beta^\lambda$ to lift to $\tilde{M}_h$. By Lemma 2.2, $\alpha^\mu \beta^\lambda$ will lift provided that $\sigma_1, \ldots, \sigma_4$ satisfy condition III, which reduces to:

\begin{align}
(1) & \quad \sigma_1^{R^\mu - 2\lambda} = 1 \\
(2) & \quad (\sigma_3\sigma_1)^\lambda = 1 \\
(3) & \quad \sigma_3^{R^\mu - 2\lambda} = 1 \\
(4) & \quad (\sigma_1\sigma_3)^\lambda = 1.
\end{align}

Any pair of permutations $\sigma_1$ and $\sigma_3$ satisfying Equations (1)-(4) determines a unique cover of $M_h(\mu, \lambda)$. We now turn our attention to the construction of such permutations.

Consider the abstract group $G$ generated by the symbols $\bar{\sigma}_1$ and $\bar{\sigma}_3$, satisfying relations (1)-(4). $G$ is a $(|R^\mu - 2\lambda|, |R^\mu - 2\lambda|, |\lambda|)$-triangle group. It is well-known that if $1/|R^\mu - 2\lambda| + 1/|R^\mu - 2\lambda| + 1/|\lambda| < 1$, then $G$ is residually finite, and hence surjects a finite group $H$ such that the images of $\bar{\sigma}_1, \bar{\sigma}_3$, and $\bar{\sigma}_3\bar{\sigma}_1$ have order $|R^\mu - 2\lambda|$. By taking the permutation representation of $H$, we then obtain permutations $\sigma_1$ and $\sigma_3$ satisfying conditions (1)-(4). Note that the permutations act on $|H|$ letters, so $\tilde{M}$ is a $4|H|$-fold cover of $M_h$.

Associated with the permutations $\sigma_1$ and $\sigma_3$ we have covers $\tilde{M}_h$ and $\tilde{F}$ of $M_h$ and $F$, and a cover $M_h(\mu, \lambda)$ of $M_h(\mu, \lambda)$.

Claim. $b_1(M_h(\mu, \lambda)) > 0$. 
Proof of claim. It suffices to show that $\tilde{h}_n$ has a non-peripheral class $[\delta] \in H_1(\tilde{F})$ with $\tilde{h}_n([\delta]) = [\delta]$. To construct this element, we shall first find a non-peripheral class $[\delta_2]$ in row 2, as follows.

Consider the sub-surface $\tilde{F}_2$ obtained by deleting rows 1, 3 and 4 from $\tilde{F}$ (see Fig. 5). The punctures of $\tilde{F}_2$ are in 1-1 correspondence with the cycles
of \(\sigma_1, \sigma_3\) and \(\sigma_3\sigma_1\). Any permutation \(\tau\) coming from the permutation representation of \(H\) decomposes as a product of \(|H|/\text{order}(\tau)\) disjoint \(\text{order}(\tau)\)-cycles. Therefore \(\tilde{F}_2\) has \(|H|(1/\text{order}(\sigma_1) + 1/\text{order}(\sigma_3) + 1/\text{order}(\sigma_3\sigma_1)) < |H|\) punctures. Since \(\chi(\tilde{F}_2) = -|H|\), we deduce that \(\tilde{F}_2\) contains a non-peripheral class \([\delta_2]\). The class \(\delta_2\) also represents a non-peripheral class in \(\tilde{F}\), since it has non-zero intersection number with a class of \(\tilde{F}\) in row 2.

We may find a corresponding non-peripheral loop \(\delta_4\) in row 4, such that \(I([\delta_2 + \delta_4], [\tilde{y}_i]) = 0\) for all \(i\) (see Fig. 7 for the notation and the idea of the proof). Let \([\delta] = [\delta_2 + \delta_4]\). Then, since \([\delta]\) has non-zero intersection number with classes in row 2 and row 4, it is a non-peripheral class. We have \(I[\delta, \tilde{y}_i] = 0\) for all \(i\) (where \(I(.,.)\) denotes oriented intersection number); therefore \([\delta]\) is fixed by \(\tilde{D}_{y_i}\), and since \(\tilde{D}_{x}\) fixes rows 2 and 4, it is fixed by \(\tilde{D}_{x_i}\). Therefore it is fixed by \(\tilde{h}_s\), concluding the proof of the claim, and of Case 1.

**Case 2.** \(m \geq 5\) and \(m\) is odd.

**Case 2a.** \(m = 5\).

The construction proceeds analogously to the case \(m = 4\). We require permutations \(\sigma_1, \ldots, \sigma_5\) satisfying conditions I-III. Again, to simplify matters, we shall impose some extra conditions: \(\sigma_2 = \text{id}, \sigma_3 = \sigma_1^{-1}\), and \(\sigma_5 = \sigma_4^{-1}\) (see Fig. 4). Then conditions I-III reduce to:

\[
\begin{align*}
\sigma_1^{R\mu - \lambda} &= 1 \\
(\sigma_1 \sigma_4)^\lambda &= (\sigma_4 \sigma_1)^\lambda = 1 \\
\sigma_4^{R\mu - 2\lambda} &= 1.
\end{align*}
\]

Again, these relations determine a triangle group, which, under the hypotheses on \(\mu\) and \(\lambda\), is hyperbolic. The rest of the proof is identical to Case 1, except that now the fixed class is in rows 3 and 5.

**Case 2b.** \(m \geq 9\) and \(m\) is odd.

Consider the cover obtained by setting \(\sigma_2 = \sigma_1^{-1}, \sigma_4 = \text{id}, \sigma_5 = \sigma_3^{-1}, \sigma_6 = \sigma_1, \sigma_7 = \sigma_1^{-1}\), and for \(i = 4, \ldots, k, \sigma_{2i+1} = \sigma_{2i}^{-1}\) (see Fig. 7a). □

The corresponding relations are:

\[
\begin{align*}
(5) & \quad \sigma_1^{R\mu - 2\lambda} = 1 \\
(6) & \quad \sigma_4^{R\mu - \lambda} = 1 \\
(7) & \quad (\sigma_3 \sigma_1)^\lambda = (\sigma_1 \sigma_3)^\lambda = 1 \\
(8) & \quad (\sigma_8 \sigma_1)^\lambda = 1 \\
(9) & \quad \sigma_{2i}^{R\mu - 2\lambda} = 1 \text{ for } i = 4, \ldots, k \\
(10) & \quad (\sigma_{2i+2} \sigma_{2i})^\lambda = 1 \text{ for } i = 4, \ldots, k - 1 \\
(11) & \quad (\sigma_1 \sigma_{2k})^\lambda = 1.
\end{align*}
\]
These relations again determine a Coxeter group. It is well-known (see [V]) that any such group surjects a finite group “without collapsing” – i.e., such that the orders of the images of the $\sigma_i$’s and $\sigma_i\sigma_j$’s are as given in (5)-(11). Then, arguing as in Case 1, we may find a non-peripheral fixed class in rows 2 and 5.
Figure 7. a. The cover for $n = 2k + 1 \geq 9$. b. The cover for $n = 2k \geq 8$.

Case 3. $n = 6$

Case 3a. $2/|R\mu - \lambda| + 1/|\lambda| < 1$.

Again, we need permutations $\sigma_1, \ldots, \sigma_6$ satisfying I-III. In this case we impose the additional conditions $\sigma_2 = id, \sigma_3 = \sigma_1^{-1}, \sigma_5 = id$, and $\sigma_6 = \sigma_4^{-1}$.
Then conditions I-III reduce to:

\[ \sigma_1^{R_\mu - \lambda} = 1 \]
\[ (\sigma_1 \sigma_4)^\lambda = (\sigma_4 \sigma_1)^\lambda = 1 \]
\[ \sigma_4^{R_\mu - \lambda} = 1. \]

These relations determine a triangle group, and we find a fixed class in rows 3 and 6.

Case 3b. \(|\lambda| > 2 \) and \(|R_\mu - 3\lambda| \geq |\lambda|\), or \(\lambda\) is even (non-zero) and \(|R_\mu - 3\lambda| \geq 4\).

When \(n = 3\), conditions I-III may be abelianized to obtain a cyclic group of order \(|R_\mu - 3\lambda|\). Specifically, they are satisfied by setting \(\sigma_1 = (1, 2, \ldots, R_\mu - 3\lambda), \sigma_2 = \sigma_1^{-2}, \sigma_3 = \sigma_1\). For \(n = 6\), we may “double” this cover: That is take \(\sigma_1, \sigma_2, \sigma_3\) as above, and then set \(\sigma_4 = \sigma_1, \sigma_5 = \sigma_2, \sigma_6 = \sigma_3\). Then we modify the corresponding cover \(\tilde{M}^{\mu, \lambda}\) of \(M^{\mu, \lambda}\) by making horizontal cuts in adjacent squares of row 3 and gluing the flaps back together as indicated by Fig. 8. If \(\lambda\) is even, we make two non-adjacent cuts and glue the top of one to the bottom of the other. If \(\lambda\) is odd, we make \((|\lambda| - 1)/2\) pairs of adjacent cuts and glue the top of the one cut to the bottom of the other cut in its pair. Now make the same cuts in row 6, with the same identifications. Since rows 3 and 6 are fixed by \(\tilde{D}_x, D_x\) still lifts to the modified \(\tilde{M}_h(\mu, \lambda)\), and since the \(\tilde{y}\)’s still project 6 to 1 onto \(y, D_y\) lifts also; so \(h\) lifts. Also, one may check that \(\alpha^\mu \beta^\lambda\) still lifts, so \(\tilde{M}_h(\mu, \lambda)\) remains a cover of \(M_h(\mu, \lambda)\).

To see that \(b_1(\tilde{M}_h(\mu, \lambda)) > 0\), note that \(\tilde{D}_x\) fixes rows 3 and 6, so it is enough to find a non-peripheral loop in row 3 and add it to the corresponding loop in row 6 with opposite orientation. As in Case 1, the existence of such a non-peripheral loop follows from an Euler characteristic argument (or see Fig. 8).

Note that Case 3a or 3b applies to all but finitely many \((\mu, \lambda)\) with \(|\lambda| > 1\).

Case 4. \(n = 2k \geq 8\).

Case 4a. \(2/|R_\mu - 2\lambda| + 1/|\lambda| < 1\). Set \(\sigma_2 = \sigma_1^{-1}, \sigma_4 = \sigma_3^{-1}, \sigma_5 = \sigma_1, \sigma_6 = \sigma_1^{-1}\), and \(\sigma_{2i} = \sigma_{2i-1}^{-1}\) for \(i = 4, \ldots, k\) (see Fig. 7b). Then, as in Case 2, these relations determine a Coxeter group. We may find a non-peripheral fixed class in rows 2 and 4.

Case 4b. \(|R_\mu - \lambda| \leq 2\) We cannot guarantee, in this case, that there will always be a cover with \(b_1 > 0\), but we shall show that there are at most finitely many exceptions.

We argue as in Case 3b. Take permutations \(\sigma_1, \ldots, \sigma_k\), and consider the relations obtained by abelianizing conditions I-III. We claim that they can be satisfied by setting \(\sigma_1 = (1, 2, 3, \ldots, N)\), for some \(N\), and setting each
\[ \sigma_i \] to an appropriate power of \( \sigma_1 \). We have already seen that this may be done when \( k = 3 \).

The \( \sigma_i \)'s must satisfy the following conditions:

1. \[
\sigma_1^{R\mu - \lambda} \sigma_2^\lambda = 1
\]
2. \[
\sigma_1^{R\mu} \sigma_2^{R\mu - \lambda} \sigma_3^\lambda = 1
\]
3. \[\vdots\]
4. \[
\sigma_1^{R\mu} \sigma_2^{R\mu} \cdots \sigma_{k-2}^{R\mu} \sigma_{k-1}^{R\mu - \lambda} \sigma_k^\lambda = 1
\]
5. \[
\sigma_1^{R\mu + \lambda} \sigma_2^{R\mu} \cdots \sigma_{k-1}^{R\mu} \sigma_k^{R\mu - \lambda} = 1
\]
6. \[
\sigma_1 \sigma_2 \cdots \sigma_k = 1.
\]

**Figure 8.** a. The cover and fixed class for \( n = 6, R\mu - 3\lambda = 4, \lambda = 3 \). b. The cover and fixed class for \( n = 6, R\mu - 3\lambda = 4, \lambda = 2 \).
We shall assume that this system has a cyclic solution, so we may substitute \( \sigma_i = \sigma_1^i \). Then, Equations (12)-(18) are equivalent to the following conditions on the exponents (all of the following equations in this case are taken mod \( N \)):

\[
R\mu - \lambda + \lambda e_2 = 0 \quad (19)
\]

\[
R\mu + (R\mu - \lambda)e_2 + \lambda e_3 = 0 \quad (20)
\]

\[
R\mu + (R\mu - \lambda)e_2 + \lambda e_3 = 0 \quad (21)
\]

\[
R\mu + (R\mu - \lambda)e_2 + \lambda e_3 = 0 \quad (22)
\]

\[
1 + e_2 + \cdots + e_k = 0 \quad (23)
\]

(22) and (23) imply that \( \lambda = \lambda e_k \). Let us set \( e_k = 1 \), eliminating Equation (22). Then, using (23), we may pair off (19) and (21) to deduce that \( \lambda e_2 = \lambda e_{k-1} \), and we set \( e_2 = e_{k-1} \) to eliminate (21). Similarly, we set \( e_3 = e_{k-2} \), and so on. If \( k \) is even, we are left with equations:

\[
R\mu - \lambda + \lambda e_2 = 0 \quad (24)
\]

\[
R\mu + (R\mu - \lambda)e_2 + \lambda e_3 = 0 \quad (25)
\]

\[
R\mu + (R\mu - \lambda)e_2 + \lambda e_3 = 0 \quad (26)
\]

\[
R\mu + (R\mu - \lambda)e_2 + \lambda e_3 = 0 \quad (27)
\]

\[
2 + 2e_2 + \cdots + 2e_{k/2} = 0 \quad (28)
\]

If we replace (28) by

\[
1 + e_2 + \cdots + e_{k/2} = 0 \quad (29)
\]

then we may eliminate (27). Then solve for \( \lambda e_2, \lambda^2 e_3, \ldots, \lambda^{k/2-1} e_{k/2} \). By (29), we have:

\[
\lambda^{k/2-1} + \lambda^{k/2-2}(\lambda e_2) + \lambda^{k/2-3}(\lambda^2 e_3) + \cdots + \lambda^{k/2-1} e_{k/2} = 0 \quad (30)
\]

Substituting our solutions for \( \lambda e_2, \lambda^2 e_3 \) and so on, we get the equation \( N = 0 \) for some integer \( N \); the system has a solution in \( \mathbb{Z}/N\mathbb{Z} \).

If \( k \) is odd, then our reduced system looks like:

\[
R\mu - \lambda + \lambda e_2 = 0 \quad (31)
\]

\[
R\mu + (R\mu - \lambda)e_2 + \lambda e_3 = 0 \quad (32)
\]
Case 5.

4b. Case 5a. Let \( \mu, \lambda \) respectively prime pairs \((\mu, \lambda)\) with \(|R\mu - \lambda| \leq 2\). Then we may eliminate (34). By (35), we have:

\[
2\lambda^{k-1/2} + 2\lambda^{k-3/2}(\lambda e_2) + 2\lambda^{k-5/2}(\lambda^2 e_3) + \cdots
\]
\[
+ 2\lambda(\lambda^{k-3/2} e_{(k-1)/2}) + \lambda^{k-1/2} e_{(k+1)/2} = 0.
\]

Adding (33) and (34) gives a multiple of (35), so we may eliminate (34).

We are supposing that \(k\) is odd. By (35), we have:

\[
2 + 2e_2 + \cdots + 2e_{(k-1)/2} + e_{(k+1)/2} = 0.
\]

Then, as in Case 3b, \(M(\mu, \lambda)\) will have a cover with \(b_1 > 0\), provided that \(|N| \geq |\lambda|\) and \(|\lambda| > 2\). Solving for \(N\), if \(k\) is even, gives:

\[
N = \lambda^{k/2-1} + \lambda^{k/2-3}(\lambda - R\mu) + \lambda^{k/2-3}[(\lambda - R\mu)^2 - R\mu \lambda]
\]
\[
+ \lambda^{k/2-4}[(\lambda - R\mu)((\lambda - R\mu)^2 - R\mu \lambda) - R\mu \lambda(\lambda - R\mu) - R\mu \lambda^2] + \cdots
\]

and if \(k\) is odd:

\[
N = 2\lambda^{k-1/2} + 2\lambda^{k-3/2}(\lambda - R\mu) + 2\lambda^{k-5/2}[(\lambda - R\mu)^2 - R\mu \lambda]
\]
\[
+ 2\lambda^{k-7/2}[(\lambda - R\mu)((\lambda - R\mu)^2 - R\mu \lambda) - R\mu \lambda(\lambda - R\mu) - R\mu \lambda^2]
\]
\[
+ \cdots + 1[\ldots].
\]

We are supposing that \(|R\mu - \lambda| \leq 2\). So for large \(\mu\) or \(\lambda\), \(R\mu/\lambda \to 1\), and for \(k\) even,

\[
N = o[\lambda^{k/2-1} + \lambda^{k/2-3}(\lambda^2) + \lambda^{k/2-4}(\lambda^3) + \cdots]
\]
\[
= o[\lambda^{k/2-1} - \lambda^{k/2-3} - \lambda^{k/2-4} - \cdots].
\]

So if \(k\) is even and \(k \geq 8\), then for all but finitely many \(\mu\) and \(\lambda\), \(|N| > |\lambda|\), and we are done. Similarly, if \(k\) is odd and \(k \geq 7\), then we are done. In the remaining cases, \(|N|\) is given by:

\[
k = 4, \quad |N| = |R\mu - 2\lambda|
\]
\[
k = 5, \quad |N| = |(R\mu)^2 - 5R\mu \lambda + 5\lambda^2|
\]
\[
k = 6, \quad |N| = |(R\mu)^2 - 4R\mu \lambda + 3\lambda^2|.
\]

One may check that each condition is satisfied by only finitely many relatively prime pairs \((\mu, \lambda)\) with \(|R\mu - \lambda| \leq 2\). This concludes the proof in Case 4b.

Case 5. Let \(n = 7\), and \(|\lambda| > 1\).

Case 5a. Let \(1/|R\mu - \lambda| + 1/|\lambda| < 2/3\) and \(|(R\mu - 2\lambda)^2 - 2\lambda^2| > 2|R|\),
We shall consider covers with $\sigma_2 = id$, $\sigma_3 = \sigma_1^{-1}$, $\sigma_6 = \sigma_5^{-1}$, and $\sigma_7 = \sigma_4^{-1}$ (see Fig. 9a).

Figure 9. Two covers for $n = 7$.

We obtain conditions:

\begin{align*}
(38) & \quad [\sigma_4, \sigma_5] = 1 \\
(39) & \quad \sigma_1^{R \mu - \lambda} = 1 \\
(40) & \quad (\sigma_4 \sigma_1)^\lambda = 1
\end{align*}
\[ \sigma_4 R_\mu (\sigma_5 \sigma_4^{-1})^\lambda = 1 \]
\[ (\sigma_4 \sigma_5) R_\mu \sigma_5^{-2\lambda} = 1 \]
\[ (\sigma_1 \sigma_4)^\lambda = 1. \]

Let us also assume for simplicity that \( \sigma_5 \) commutes with \( \sigma_1 \). Equations (38), (41) and (42) determine an abelian group \( A \) of order \(|(R_\mu - 2\lambda)^2 - 2\lambda^2|\); we must show that \( \sigma_4^2 \) is non-trivial in \( A \). The elements \( \sigma_4^2 \) and \( \sigma_5 \) generate a subgroup \( H \) of \( A \) of index at most 2. If \( \sigma_4^2 = id \), then \( H \) is cyclic of order \( gcd(|\lambda|, |R_\mu - 2\lambda|) \). Then \(|(R_\mu - 2\lambda)^2 - 2\lambda^2| = |A| \leq 2|H| = 2gcd(|\lambda|, |R_\mu - 2\lambda|) = 2gcd(|\lambda|, |R|) \leq 2|R| \). So if
\[ |(R_\mu - 2\lambda)^2 - 2\lambda^2| > 2|R| \]
then \( \sigma_4^2 \neq id \). Therefore, under our hypotheses, the relations generate a group which is isomorphic to the direct sum of a cyclic group with a hyperbolic triangle group. As in the previous cases, we may then find a non-peripheral fixed class (in rows 3 and 7), and we are done.

However, note that if \( R = 1 \), then Equation (44) is false for all \((\mu, \lambda)\) satisfying
\[ (\mu + 2\lambda)^2 - 2\lambda^2 = 1. \]
This is an example of Pell’s equation, which has infinitely many solutions, and hence (44) may be false infinitely often.

Case 5b. \( 1/|R_\mu - 2\lambda| + 1/|\lambda| < 2/3 \) and \(|(R_\mu - \lambda)^2 - 2\lambda^2| > 2|R| \).

Let \( \sigma_3 = id, \sigma_4 = \sigma_2^{-1}, \sigma_5 = \sigma_1^{-1}, \) and \( \sigma_7 = \sigma_6^{-1} \) (see Fig. 9b). The conditions for a cover are:
\[ [\sigma_1, \sigma_2] = 1 \]
\[ \sigma_1 R_\mu (\sigma_2 \sigma_1^{-1})^\lambda = 1 \]
\[ (\sigma_1 \sigma_2) R_\mu \sigma_2^{-\lambda} = 1 \]
\[ (\sigma_6 \sigma_1)^\lambda = 1 \]
\[ \sigma_6 R_\mu^{-2\lambda} = 1 \]
\[ (\sigma_1 \sigma_6)^\lambda = 1. \]

For simplicity, suppose also that \( \sigma_2 \) commutes with \( \sigma_6 \). Then (45), (46), (47) determine an abelian group \( B \) of order \(|(R_\mu - \lambda)^2 - R_\mu \lambda| \). If \( \sigma_1^2 = 1 \), then \(|B| \leq 2gcd(|\lambda|, |R_\mu - \lambda|) = 2gcd(|\lambda|, |R|) \leq 2|R| \). Therefore, in this case, the group determined by conditions (45)-(50) is again the direct product
of an abelian group with a hyperbolic triangle group, and we may find a non-peripheral fixed class in rows 5 and 7.

Note that Case 5a or 5b applies to all but finitely many surgeries where $|\lambda| > 1$.

This concludes the proof of Theorem 1.3.

\section{Examples.}

We begin with the proof of Theorems 1.5 and 1.6 (see Section 1 for notation).

\textbf{Lemma 3.1.} Let $g = D_y^5 D_x^{-1}$ and $h = D_x D_y$. Then $M_{-h}(\mu, \lambda) \cong M_g(\mu, \lambda - \mu)$, and $M_{h^2}(\mu, \lambda) \cong M_{(-h)^2}(\mu, \lambda + \mu) \cong M_{g^2}(\mu, \lambda - \mu)$.

\textbf{Proof.} Recall that the mapping class group of the once-punctured torus is isomorphic to $SL_2(\mathbb{Z})$, under the identifications $D_x \rightarrow R = [1 1] \text{ and } D_y \rightarrow L = [1 0]$. Under these identifications, we compute that $h$ has monodromy matrix $[2 1]$, $(-1)$ has monodromy matrix $[0 1]$, and $g$ has monodromy matrix $[1 -1]$. The homeomorphisms $h^2$ and $(-h)^2$ have the same monodromy matrix, and hence are isotopic. Therefore $M_{h^2} \cong M_{(-h)^2}$. Also, $[2 -1]^{-1}(-RL[2 1])^{-1} = L^5 R^{-1}$, so $g$ and $-h$ have conjugate monodromy matrices. It follows that $M_{-h} \cong M_g$, and $M_{(-h)^2} \cong M_{g^2}$.

It remains to determine the effect of these homeomorphisms on the framings. Computing the maps on $\pi_1(F)$ gives:

\[ (-h)^2 = (x^{-1} y x y^{-1}) (h^2) (x^{-1} y x y^{-1})^{-1}. \]

Therefore the isotopy which takes $h^2$ to $(-h)^2$ twists $\partial F$ once in a counterclockwise manner, so the induced bundle homeomorphism sends $M_{h^2}(\mu, \lambda)$ to $M_{(-h)^2}(\mu, \lambda + \mu)$.

Let $f = D_y^2 D_x^{-1}$. The bundle homeomorphism induced by conjugation preserves the framing, so $M_{-h}(\mu, \lambda) \cong M_{f(-h)f^{-1}}(\mu, \lambda)$. The homeomorphisms $f(-h)f^{-1}$ and $g$ have identical monodromy matrices, and hence are isotopic. We compute $g_2 = (yx^{-1} y^{-1} x)f(-h)f_2^{-1}(yx^{-1} y^{-1} x)^{-1}$ so the isotopy from $f(-h)f^{-1}$ to $g$ twists $\partial F$ once in a clockwise manner. The induced bundle homeomorphism sends $M_{f(-h)f^{-1}}(\mu, \lambda)$ to $M_g(\mu, \lambda - \mu)$. So $M_{-h}(\mu, \lambda) \cong M_g(\mu, \lambda - \mu)$.

Likewise, $M_{f(-h)^2f^{-1}}(\mu, \lambda) \cong M_{g^2}(\mu, \mu - 2\mu)$. Thus

\[ M_{h^2}(\mu, \lambda) \cong M_{(-h)^2}(\mu, \lambda + \mu) \cong M_{f(-h)^2f^{-1}}(\mu, \lambda + \mu) \cong M_{g^2}(\mu, \lambda - \mu). \]

\[ \square \]

\textbf{Proof of Theorem 1.5.} This is an immediate consequence of Lemma 3.1 and Theorem 1.3.

\[ \square \]
Proof of Theorem 1.6. We have $M(2\mu, \lambda) \cong M_{g_2}(2\mu, \lambda)$, which is double covered by $M_{g_2}(\mu, \lambda) \cong M_{g_2}(\mu, \lambda - \mu)$. So it is enough to show that $M_{g_2}(\mu, \lambda - \mu)$ is virtually $\mathbb{Z}$-representable. By Theorem 1.3, we are done unless

$$\frac{1}{|2\mu - (\lambda - \mu)|} + \frac{1}{|2\mu - 2(\lambda - \mu)|} + \frac{1}{|\lambda - \mu|} \geq 1$$

or, simplifying:

$$\frac{1}{|\mu + \lambda|} + \frac{1}{|2\lambda|} + \frac{1}{|\mu - \lambda|} \geq 1. \tag{51}$$

By $[B3]$, $M(2\mu, \lambda)$ is virtually $\mathbb{Z}$-representable if $2\mu$ is divisible by 4; hence we may assume $\mu$ is odd. Also, since $gcd(2\mu, \lambda) = 1$, we may assume $\lambda$ is odd, and, assuming $(\mu, \lambda) \neq (\pm 1, 1)$, $|\lambda| \neq |\mu|$. It follows that

$$|\mu - \lambda| \geq 2 \tag{52}$$

$$|\mu + \lambda| \geq 2. \tag{53}$$

The only simultaneous solutions to inequalities (51), (52) and (53) with $\mu$ and $\lambda$ odd are: $(\mu, \lambda) = \pm (-3, 1)$ and $\pm (3, 1)$. So the only possible exceptions to Theorem 1.6 are $M(-6, 1) \cong M(6, 1) \cong M(-2, 1)$. The virtual $\mathbb{Z}$-representability of these manifolds may be verified with either of the computer programs GAP or Snappea. \hfill $\square$

We now turn to the proof of Theorem 1.4.

Proof of Theorem 1.4. Let $g$ and $h$ be as in the statement of Lemma 3.1, let $f = h^{18}$, and let $i = D_{2y}^2 D_{y}^{-4} D_{x}^{-4} D_{x}$. Both $h^3$ and $i$ have monodromy matrix $[13 \ 8; 5]$. Hence $h^3$ and $i$ are isotopic. By arguments similar to those used in the proof of Lemma 3.1, we compute that $M_{h^3}(\mu, \lambda) \cong M_i(\mu, \mu + \lambda)$. Therefore $M_f(\mu, \lambda) \cong M_{g_2}(\mu, \lambda + 6\mu)$. Hence by Theorem 1.3 iii, $M_f(\mu, \lambda)$ is virtually $\mathbb{Z}$-representable if

$$\frac{1}{|6\mu - \lambda|} + \frac{1}{|6\mu + \lambda|} < 1. \tag{54}$$

By Lemma 3.1 we have $M_f(\mu, \lambda) \cong M_{g_2}(\mu, \lambda - 9\mu)$. Hence by Theorem 1.3 ii, $M_f(\mu, \lambda)$ is virtually $\mathbb{Z}$-representable if

$$\frac{1}{|9\mu + \lambda|} + \frac{1}{|2\lambda|} + \frac{1}{|9\mu - \lambda|} < 1. \tag{55}$$

The only simultaneous solutions to the inequalities 54 and 55 have $\mu = 0$. The proof is completed by noting that $M(0, 1)$ has positive first Betti number, as it is a torus bundle over $S^1$. \hfill $\square$

We remark that the same methods may be applied to many other examples of once-punctured torus bundles, to show that all but finitely many surgeries are virtually $\mathbb{Z}$-representable. The idea is to start with a monodromy $f$ to which Theorem 1.3 i or ii applies. Since $L^1$ and $R$ generate a finite-index subgroup of $SL_2(\mathbb{Z})$, there exists an integer $\ell$ such that $f^\ell$ is
isotopic to a $g$ satisfying the hypotheses of Theorem 1.3 iii. Usually Theorem 1.3 will then imply that all but finitely many surgeries on $\mathcal{N}_f \ell$ are virtually $\mathbb{Z}$-representable.

References


Received June 15, 1998.

University of Texas at Austin
Austin, Texas 78712

Rice University
Houston, Texas 77005-1892

E-mail address: mastersj@rice.edu
GROUP ACTIONS ON POLYNOMIAL AND POWER SERIES RINGS

Peter Symonds

When a finite group $G$ acts faithfully on a graded integral domain $S$ which is an algebra over a field $k$, such as a polynomial ring, we consider $S$ as a $kG$-module. We show that $S$ is asymptotically mostly projective in each degree, and also that it is in fact mostly free in an appropriate sense. Similar results also hold for filtered algebras, such as power series rings.

1. Introduction.

Let $S = \bigoplus_{n=0}^{\infty} S_n$ be a graded algebra over a field $k$. We suppose that $S$ is finitely generated over $k$ as a $k$-algebra and that the homogeneous components $S_n$ are finite dimensional vector spaces over $k$. Let $G$ be a finite group of grading preserving automorphisms of $S$ (so $G$ acts faithfully). We are concerned with the structure of $S$ as a $kG$-module.

The classical theory of Hilbert and Serre asserts that for large $n$, $\dim_k S_n$ is given by a function

$$\phi_S(n) = c_{d-1}(n)n^{d-1} + c_{d-2}(n)n^{d-2} + \cdots + c_0(n),$$

where the $c_i(n)$ are rational valued functions periodic in $n$, i.e., $\phi_i(n + p) = \phi_i(n)$ for some integer $p$ (see Section 2). If $c_{d-1}$ is assumed not to be identically zero then $d$ is equal to the dimension of the ring in various senses. If $S$ is a polynomial ring then $d$ is equal to the number of variables.

From now on, we assume that $S$ is an integral domain. Let $P_n$ denote the maximal projective summand of $S_n$ (defined up to isomorphism).

**Theorem 1.1.** $\dim_k (S_n/P_n)$ is bounded by a polynomial in $n$ of degree $d-2$.

Thus $S_n$ is mostly projective, and if $S$ is a polynomial ring then the non-projective part grows like a polynomial ring in one fewer variables.

In fact $S$ is mostly free, although the individual $S_n$ do not have to contain a free module at all; the different projectives can occur in different degrees. To explain this let $R = S^G$, the ring of invariants.
Theorem 1.2. \( S \) contains a free \( kG \)-submodule \( F \) of rank 1, a sum of homogeneous pieces, such that the product map \( R \otimes_k F \to S \) is injective. Denote its image by \( RF = \bigoplus_n (RF)_n \). Then \( RF \) is a free summand of \( S \) and \( \dim_k(S_n/(RF)_n) \) is bounded by a polynomial of degree \( d-2 \).

Of course, the first theorem is a corollary to the second. Versions of these theorems were proven by Howe [4] in characteristic 0 and by Bryant [2, 3] for polynomial rings.

Section 2 contains the main proof, except for some technical details which appear in Section 3. Section 4 proves similar results for filtered algebras.

2. Main Proof.

Proof. We can assume that \( k \) is a splitting field for \( G \), since if a \( kG \)-module contains a free or projective summand after extension of scalars then it did so before. Let \( Q_S \) (resp. \( Q_R \)) denote the fields of fractions of \( S \) (resp. \( R \)), so \( Q_S \cong Q_R \otimes_R S \). By the Normal Basis Theorem, \( Q_S \) is a free module of rank 1 over \( Q_R \). Let \( e \) be a generator. Then, over \( kG \), \( e \) generates a free submodule \( E \) of rank 1 such that \( Q_S \cong Q_R \otimes_k E \). Now there is an \( r \in R \) such that \( re \in S \). Let \( F \) be the \( kG \)-module generated by \( re \), so \( F \subset S \) and \( F = rE \cong kG \). Also the product map \( R \otimes_k F \to RF \subset S \) is injective.

We claim that \( F \) can be assumed to be a sum of homogeneous pieces, \( F = \bigoplus_i F_n_i \). The proof of this plausible statement is somewhat delicate, and we postpone it to the next section.

Let \( x_1, \ldots, x_s \) be homogeneous generators for \( S \) as a \( k \)-algebra. Then \( x_i = \sum_j a_{i,j} b_{i,j} e_j \), where \( a_{i,j}, b_{i,j} \in R \) and the \( e_j \) form a homogeneous \( k \)-basis for \( F \). By writing \( b_{i,j} x_i = \sum_j a_{i,j} e_j \) and taking the homogeneous component of this equation in some degree where \( b_{i,j} x_i \) is non-zero, we see that we may assume that the \( b_{i,j} \) are homogeneous. Let \( \alpha \in R_a \) be the product of all the \( b_{i,j} \). Then each \( x_i \in \alpha^{-1} RF \), so \( S \subset \alpha^{-1} RF \).

Thus
\[(RF)_n \subset S_n \subset \alpha^{-1}(RF)_{n+a},\]
and so, identifying \( RF \) with \( R \otimes_k F \), we have
\[\bigoplus_i R_{n-n_i} \otimes F_{n_i} \subset S_n \subset \alpha^{-1} \bigoplus_i R_{n+a-n_i} \otimes F_{n_i} .\]

In particular, the dimension of \( S_n/(RF)_n \) is bounded by the difference in the dimensions of the two sides, i.e., by
\[ \sum_i (\phi_R(n + a - n_i) - \phi_R(n - n_i)) \dim_k F_{n_i}.\]
But
\[ \phi_R(n + a - n_i) - \phi_R(n - n_i) = c_{d-1}(n + a - n_i)(n + a - n_i)^{d-1} - c_{d-1}(n - n_i)(n - n_i)^{d-1} + \text{lower degree terms}, \]
and \( c_{d-1} \) is periodic, with period dividing \( a \) (see 3.1), so the \( n^{d-1} \) term cancels and we are done. \( \square \)

3. Technical Details.

The form of \( \phi_S(n) \) given above is not quite the standard one, although it is quoted in [4]. The usual references deal with a module over a polynomial ring which has all the variables in degree 1, and then all the coefficients of \( \phi \) are constants. To deduce the version given in the introduction, note that if \( S \) is generated by \( x_1, \ldots, x_s \) then it is a finitely generated module over \( k[x_1, \ldots, x_s] \). By taking suitable powers \( y_i \) of the \( x_i \) we can get all the \( y_i \) in the same degree \( b \), and \( S \) will still be finitely generated over \( k[y_1, \ldots, y_s] \). For \( 0 \leq j \leq b-1 \), set \( T_j = \bigoplus_{i=0}^{\infty} S_{j+b} \). Then \( R \cong \bigoplus_j T_j \) as a \( k[y_1, \ldots, y_s] \)-module, and after regrading each \( T_j \) so that each \( y_i \) can have degree 1, we can apply the usual theory ([1] 11.2, [5] VII Theorem 41) to each \( T_j \) and sum the results. It is the summation that leads to the periodic coefficients.

**Lemma 3.1 ([4]).** If \( R \) is an integral domain (as it always is for us), \( c = \gcd\{r \in \mathbb{Z} | R_r \neq 0 \} \) and \( \phi_R(n) = c_{d-1}(n)n^{d-1} + \cdots + c_0(n) \), then there is a constant \( b \) such that
\[ c_{d-1}(n) = \begin{cases} b, & \text{if } n|c, \\ 0, & \text{otherwise}. \end{cases} \]

**Proof.** If \( 0 \neq \alpha \in R_n \), then multiplication by \( \alpha \) embeds \( R_n \) in \( R_{n+a} \), so for large \( n \), \( \phi_R(n) \leq \phi_R(n+a) \). Now consider the limit of \( \phi_R(n)/n^{d-1} \) as \( n \to \infty \) through elements of the same residue class modulo the period of \( c_{d-1} \) to see that \( c_{d-1}(n) \leq c_{d-1}(n+a) \). This, together with the periodicity, implies the result. \( \square \)

Now we prove the claim made in Section 2.

**Proposition 3.2.** The free module \( F \subset S \) can be assumed to be a sum of homogeneous pieces in such a way that the product map \( Q_R \otimes_k F \to Q_S \) is still an isomorphism.

**Proof.** For each simple \( kG \)-module \( V \), let \( T_V = \text{Hom}_{kG}(V, S) \), a graded \( R \)-module. Now \( socF \) is a direct sum of simples. Let \( soc_V(F) \) denote the sum of those isomorphic to \( V \), so \( soc_V(F) = V_1 \oplus \cdots \oplus V^s \), where \( V^i \cong V \), and let \( P_{V^i} \) be a projective summand of \( F \) with \( soc(P_{V^i}) = V^i \). The inclusions of the \( V^i \) in \( S \) give us \( s \) homomorphisms \( f^i \in T_V \), which are linearly independent over \( R \).
Lemma 3.3. Let $f^1, \ldots, f^s$ be elements of a graded $R$-module $T$ which are linearly independent over $R$. Write each $f^j$ as a sum of its homogeneous components: $f^j = \sum_k f^j_k$, $f^j_k \in T_k$. Then for each $j$ there is an integer $k_j$ such that $f^j_{k_1}, \ldots, f^j_{k_s}$ are linearly independent over $R$.

Proof. For each $0 \leq t \leq s$, let $P_t$ be the claim that there exist integers $k_1, \ldots, k_t$ such that $f^{j_1}_1, \ldots, f^{j_t}_{k_t}, f^{j_{t+1}}, \ldots, f^s$ are linearly independent over $R$. $P_0$ is true by hypothesis and we want $P_s$. We give a proof by induction on $t$, so assume $P_t$.

If $P_{t+1}$ is false, then for each $k \in \mathbb{Z}$ we can find $u_k, r^t_k \in R$, $u_k \neq 0$, such that $$u_k f^{t+1}_k = r^t_k f^1_k + \cdots + r^t_k f^t_k + r^{t+2}_k f^{t+2} + \cdots + r^s_k f^s.$$ Let $u$ be the product of the $u_k$ for which $f^{t+1}_k \neq 0$. Then $uf^{t+1} = \sum_k (\frac{u_k}{u}) u_k f^{t+1}_k$, contradicting $P_t$. \hfill $\square$

Applying this to the $\{f^i\} \subset T_V$ we obtain homogeneous $\{\tilde{f}^i\} \subset T_V$, $\tilde{f}^i \in T_{a_i}$, say, linearly independent over $R$.

Lemma 3.4. The evaluation map $ev : T_V \otimes_k V \to S$ is injective.

Proof. In fact $ev : \text{Hom}_{kG}(V, M) \otimes_k V \to M$ is injective for any $kG$-module $M$. This is because it factors through $\text{soc}_V(M)$, which is a direct sum of $V$’s, so we are reduced to proving the case $M = V$. But then $ev$ is an isomorphism, since $\text{Hom}_{kG}(V, V) \cong k$, by the assumption that $k$ is a splitting field. \hfill $\square$

Corollary. The product map $R \otimes_k (\bigoplus_i \tilde{f}^i(V)) \to S$ is injective.

Now let $\tilde{P}_V$ be the image of the projection of $P_{V_i}$ to $S_{a_i}$. The projection map is injective on $\text{soc}(P_{V_i})$, by the construction of $a_i$, so $\tilde{P}_V \cong P_{V_i}$ and $\text{soc}(\tilde{P}_V) = \tilde{f}(V)$. Let $\tilde{P}_V = \bigoplus_i \tilde{P}_{V_i}$ and consider the product map $R \otimes_k \tilde{P}_V \to S$. Since $\text{soc}(R \otimes_k \tilde{P}_V) = R \otimes \bigoplus_i \tilde{f}(V)$, it is injective on the socle, so is injective.

Finally, we sum the $\tilde{P}_V$ over the simples $V$ to obtain $\tilde{F}$, a free $kG$-module of rank 1, which is a sum of homogeneous pieces, as required. \hfill $\square$

Remark. If $G$ is a $p$-group, where $p$ is the characteristic of $k$, then the proof is much simpler because $\text{soc}(F) \cong k$. Under at least one of the projections of $F$ onto its homogeneous components the image of $\text{soc}(F)$ must be non-zero. Let $F$ be the image of $F$ under this projection. Then $\tilde{F} \cong kG$ and $Q_R \otimes_k F \to Q_S$ is an isomorphism because it is injective on the socle, and both sides have the same dimension over $Q_R$.

This is enough to prove 1.1 for general $G$. For if $P = \text{Syl}_p(G)$ then $S$ is a direct summand of $\text{Ind}_G^R \text{Res}_P^G S$. 

228 PETER SYMONDS
Remark. It is not hard to see that, given any degree $m$, the summands of $\bar{F}$ can be moved by multiplication by a scalar to lie in $T_{m+lc} = S_{m+lc} \oplus \cdots \oplus S_{m+(l+1)c-1}$, for some $l$. The argument of the proof of 1.2 now shows that the non-free part of $T_n$ has dimension bounded by a polynomial of degree $d-2$ (cf. [2, 3]).

4. Filtered Rings.

The case of filtered rings is slightly different. Consider the power series ring $k[[x]]$ in characteristic 2 and let the group of order 2 act by $x \mapsto x/(x+1) = x + x^2 + x^3 + \cdots$. The action on the associated graded ring is trivial, yet the action on $k[[x]]$ certainly contains free summands (the only alternative is trivial).

We consider finitely generated $k$-algebras $S$ which are integral domains and have a filtration $S = I_0 \subset I_1 \subset I_2 \subset \cdots$. Each $S/I_n$ is assumed to be finite dimensional over $k$, and $\cap I_n = \{0\}$. There is a finite group $G$ of automorphisms of $S$, which preserves the filtration. The invariants are $R = S^G$ with the induced filtration $J_n = R \cap I_n$. Again there is a function $\chi_S(n) = c_d(n)n^d + c_{d-1}n^{d-1} + \cdots + c_0(n)$, where the $c_i$ are periodic, such that $\dim_k(S/I_n) = \chi_S(n)$ for large $n$. If $S$ is a power series ring, then $d$ is equal to the number of variables.

As before there is a free $kG$-module of rank 1 in $S$, and the product map $R \otimes_k F \to S$ is injective. Since $F$ is finite dimensional there is some integer $f$ such that $F \cap I_f = 0$, so $F$ injects into $S/I_f$.

For each $n$, let $K_n$ be a vector space complement to $J_n$ in $R$. Then the product map $K_n \otimes F \to S/I_{f+n}$ is injective, so its image, $K_nF$, is a free summand of $S/I_{f+n}$.

Proceeding in the same way as before we can prove:

**Theorem 4.1.** $\dim_k((S/I_{f+n})/K_nF)$ is bounded by a polynomial of degree $d-1$.

So $S$ is mostly free. Again, for a power series ring, the non-free part grows like a power series ring in one fewer variables.

**References**


Received October 13, 1998.

DEPARTMENT OF MATHEMATICS
U.M.I.S.T.
MANCHESTER M60 1QD
ENGLAND
E-mail address: psymonds@umist.ac.uk
THE MODULI OF FLAT PU(2,1) STRUCTURES ON RIEMANN SURFACES

Eugene Z. Xia

For a compact Riemann surface $X$ of genus $g > 1$, $\text{Hom}(\pi_1(X), \text{PU}(p, q))/\text{PU}(p, q)$ is the moduli space of flat $\text{PU}(p, q)$-connections on $X$. There are two integer invariants, $d_P, d_Q$, associated with each $\sigma \in \text{Hom}(\pi_1(X), \text{PU}(p, q))/\text{PU}(p, q)$. These invariants are related to the Toledo invariant $\tau$ by $\tau = 2d_Q - d_P$. This paper shows, via the theory of Higgs bundles, that if $q = 1$, then $-2(g - 1) \leq \tau \leq 2(g - 1)$. Moreover, $\text{Hom}(\pi_1(X), \text{PU}(2, 1))/\text{PU}(2, 1)$ has one connected component corresponding to each $\tau \in \frac{2}{3}\mathbb{Z}$ with $-2(g - 1) \leq \tau \leq 2(g - 1)$. Therefore the total number of connected components is $6(g - 1) + 1$.

1. Introduction.

Let $X$ be a smooth projective curve over $\mathbb{C}$ with genus $g > 1$. The deformation space

$$\mathcal{C}_B = \text{Hom}^+(\pi_1(X), \text{PGL}(n, \mathbb{C}))/\text{PGL}(n, \mathbb{C})$$

is the space of equivalence classes of semi-simple $\text{PGL}(n, \mathbb{C})$-representations of the fundamental group $\pi_1(X)$. This is the $\text{PGL}(n, \mathbb{C})$-Betti moduli space on $X$ [22, 23, 24]. A theorem of Corlette, Donaldson, Hitchin and Simpson relates $\mathcal{C}_B$ to two other moduli spaces, $\mathcal{C}_{\text{DR}}$ and $\mathcal{C}_{\text{Dol}}$—the $\text{PGL}(n, \mathbb{C})$-de Rham and the $\text{PGL}(n, \mathbb{C})$-Dolbeault moduli spaces, respectively [3, 5, 11, 21]. The Dolbeault moduli space consists of holomorphic objects (Higgs bundles) over $X$; therefore, the classical results of analytic and algebraic geometry can be applied to the study of the Dolbeault moduli space.

Since $\text{PU}(p, q) \subset \text{PGL}(n, \mathbb{C})$, $\mathcal{C}_B$ contains the space

$$\mathcal{N}_B = \text{Hom}^+(\pi_1(X), \text{PU}(p, q))/\text{PU}(p, q).$$

The space $\mathcal{N}_B$ will be referred to as the $\text{PU}(p, q)$-Betti moduli space which similarly corresponds to some subspaces $\mathcal{N}_{\text{DR}}$ and $\mathcal{N}_{\text{Dol}}$ of $\mathcal{C}_{\text{DR}}$ and $\mathcal{C}_{\text{Dol}}$, respectively. We shall refer to $\mathcal{N}_{\text{DR}}$ and $\mathcal{N}_{\text{Dol}}$ as the $\text{PU}(p, q)$-de Rham and the $\text{PU}(p, q)$-Dolbeault moduli spaces.

The Betti moduli spaces are of great interest in the field of geometric topology and uniformization. In the case of $p = q = 1$, Goldman analyzed
\[ N_{Dol} \] and determined the number of its connected components to be \( 4g - 3 \) [6]. Hitchin subsequently considered \( N_{Dol} \) in the case of \( p = q = 1 \) and determined its topology [11].

In this paper, we analyze \( N_{Dol} \) for the case of \( p = 2, q = 1 \) and determine its number of connected components. In addition, we produce a new algebraic proof, via the Higgs-bundle theory, of a theorem by Toledo on the bounds of the Toledo invariant [26, 27].

An element \( \sigma \in \text{Hom}^+(\pi_1(X), \text{PU}(p, q)) \) defines a flat principal \( \text{PU}(p, q) \)-bundle \( P \) over \( X \). Such a flat bundle may be lifted to a principal \( \text{U}(p, q) \)-bundle \( \hat{P} \) with a Yang-Mills connection \( D \) [2, 3, 5, 11, 21]. Let \( E \) be the rank-\((p + q)\) vector bundle associated with \( (\hat{P}, D) \). The second cohomology \( H^2(X, \mathbb{Z}) \) is isomorphic to \( \mathbb{Z} \), so one may identify the Chern class \( c_1(E) \in H^2(X, \mathbb{Z}) \) with an integer, the degree of \( E \). Suppose we impose the additional condition

\[
0 \leq \text{deg}(E) < n.
\]

Then the above construction gives rise to a unique obstruction class \( o_2(E) \in H^2(X, \pi_1(\text{U}(p, q))) \) [25]. The obstruction class is invariant under the conjugation action of \( \text{PU}(p, q) \); therefore, one obtains the obstruction map:

\[
o_2 : \text{Hom}^+(\pi_1(X), \text{PU}(p, q))/\text{PU}(p, q) \longrightarrow H^2(X, \pi_1(\text{U}(p, q))) \cong \mathbb{Z} \times \mathbb{Z}.
\]

The maximum compact subgroup of \( \text{U}(p, q) \) is \( \text{U}(p) \times \text{U}(q) \). Hence topologically \( E \) is a direct sum \( E_P \oplus E_Q \) with

\[
\text{deg}(E) = \text{deg}(E_P) + \text{deg}(E_Q).
\]

The obstruction class \( o_2(E) \) is then \( (\text{deg}(E_P), \text{deg}(E_Q)) \in \mathbb{Z} \times \mathbb{Z} \). Associated with \( \sigma \) is the Toledo invariant \( \tau \) which relates to \( d_P = \text{deg}(E_P) \) and \( d_Q = \text{deg}(E_Q) \) by the formula [7, 26, 27]

\[
\tau = 2 \frac{\text{deg}(E_P \otimes E_Q^*)}{p + q} = 2 \frac{qd_P - pd_Q}{p + q}.
\]

This explains why the Toledo invariant of a \( \text{PU}(2, 1) \) representation cannot be an odd integer [7]. The main result presented here is the following:

**Theorem 1.1.** \( \text{Hom}^+(\pi_1(X), \text{PU}(2, 1))/\text{PU}(2, 1) \) has one connected component for each \( \tau \in \frac{2}{3} \mathbb{Z} \) with \( -2(g - 1) \leq \tau \leq 2(g - 1) \). Therefore the total number of connected components is \( 6(g - 1) + 1 \).

We shall also provide a new proof en route to the following theorem:

**Theorem 1.2 (Toledo).** Suppose \( \sigma \in \text{Hom}^+(\pi_1(X), \text{PU}(p, 1)) \) and \( \tau \) is the Toledo class of \( \sigma \). Then

\[
-2(g - 1) \leq \tau \leq 2(g - 1).
\]

Moreover \( \tau = \pm 2(g - 1) \) implies \( \sigma \) is reducible.
These results are related to the results of Domic and Toledo \[4, 26, 27\] and, as being pointed out to the author recently, are also related to the work of Gothen \[8\] which computed the Poincaré polynomials for the components of \(\text{Hom}(\pi_1(X), \text{PSL}(3, \mathbb{C}))/\text{PSL}(3, \mathbb{C})\), where \(\deg(E)\) is coprime to 3.

**Acknowledgments.**

Most of this research was carried out while the author was at the University of Maryland at College Park. I thank J. Adams, K. Coombes, P. Green, K. Joshi, S. Kudla, P. Newstead, J. Poritz and especially W. Goldman and C. Simpson for insightful discussions over the course of the research. I thank the referee for helpful suggestions.

2. **Backgrounds and Preliminaries.**

In this section, we briefly outline the constructions of the Betti, de Rham and Dolbeault moduli spaces. For details, see \[2, 3, 5, 11, 12, 18, 21, 22, 23, 24\].

2.1. **The Betti Moduli Space.** The fundamental group \(\pi_1(X)\) is generated by \(S = \{A_i, B_i\}_{i=1}^g\), subject to the relation

\[
\prod_{i=1}^g A_i B_i A_i^{-1} B_i^{-1} = e.
\]

Denote by \(I\) and \([I]\) the identities of \(\text{GL}(n, \mathbb{C})\) and \(\text{PGL}(n, \mathbb{C})\), respectively. Define

\[
R : \text{PGL}(n, \mathbb{C})^{2g} \rightarrow \text{PGL}(n, \mathbb{C})
\]

\[
\mathcal{R} : \text{GL}(n, \mathbb{C})^{2g} \rightarrow \text{GL}(n, \mathbb{C})
\]

to be the commutator maps:

\[
(X_1, Y_1, \ldots, X_g, Y_g) \xrightarrow{R, \mathcal{R}} \prod_{i=1}^g X_i Y_i X_i^{-1} Y_i^{-1}.
\]

The group

\[
\{\zeta I : \zeta \in \mathbb{C}, \zeta^n = 1\}
\]

is isomorphic to \(\mathbb{Z}_n\). The space \(\mathcal{R}^{-1}(\mathbb{Z}_n)\) is identified with the representation space \(\text{Hom}(\Gamma, \text{GL}(n, \mathbb{C}))\), where \(\Gamma\) is the central extension \[2, 11\];

\[
0 \rightarrow \mathbb{Z}_n \rightarrow \Gamma \rightarrow \pi_1(X) \rightarrow 0.
\]

Each element \(\rho \in \mathcal{R}^{-1}(\mathbb{Z}_n)\) acts on \(\mathbb{C}^n\) via the standard representation of \(\text{GL}(n, \mathbb{C})\). The representation \(\rho\) is called reducible (irreducible) if its action on \(\mathbb{C}^n\) is reducible (irreducible). A representation \(\rho\) is called semi-simple if it is a direct sum of irreducible representations. Let \(\zeta_1 = e^{2\pi i/n}\) and define

\[
\mathcal{CM}_B(e) = \{\sigma \in \mathcal{R}^{-1}(\zeta_1 I) : \sigma\ \text{is semi-simple}\}/\text{GL}(n, \mathbb{C}),
\]
\[ \mathbb{C}M_B = \bigcup_{c=0}^{n-1} \mathbb{C}M_B(c), \]

\[ \mathbb{C}N_B(c) = \mathbb{C}M_B(c)/\text{Hom}(\pi_1(X), \mathbb{C}^*) \]

\[ = \text{Hom}^+(\pi_1(X), \text{PGL}(n, \mathbb{C}))/\text{PGL}(n, \mathbb{C}). \]

Fix \( p, q \) such that \( p + q = n \). Denote by \( \mathcal{R}_U \) the restriction of \( \mathcal{R} \) to the subgroup \( U(p, q) \).

Define

\[ \mathcal{M}_B(c) = \{ \sigma \in \mathcal{R}_U^{-1}(\zeta I) : \sigma \text{ is semi-simple} \}/U(p, q), \]

\[ \mathcal{M}_B = \bigcup_{c=0}^{n-1} \mathcal{M}_B(c). \]

Note the center of \( U(p, q) \) is \( U(1) \) and is contained in the center of \( \text{GL}(n, \mathbb{C}) \). It follows that \( \mathcal{M}_B(c) \subset \mathbb{C}M_B(c) \). Define

\[ \mathcal{N}_B(c) = \mathcal{M}_B(c)/\text{Hom}(\pi_1(X), U(1)) \]

\[ \mathcal{N}_B = \mathcal{M}_B/\text{Hom}(\pi_1(X), U(1)) = \text{Hom}^+(\pi_1(X), U(p, q))/U(p, q). \]

All the spaces constructed here that contain the symbols \( \mathcal{M}_B \) or \( \mathcal{N}_B \) will be loosely referred to as Betti moduli spaces. The subspace of irreducible elements of a Betti moduli space will be denoted by an \( s \) superscript. For example, \( \mathbb{C}M_B^s \) denotes the subspace of irreducible elements of \( \mathbb{C}M_B \).

### 2.2. The de Rham Moduli Space

Suppose \( P \) is a principal \( \text{GL}(n, \mathbb{C}) \)-bundle on \( X \), \( E \) its associated vector bundle of rank \( n \) and \( \mathcal{G}_C(E) \) the group of \( \text{GL}(n, \mathbb{C}) \)-gauge transformations on \( E \). A connection is called Yang-Mills (or central) if its curvature is central \[ 2 \]

\[ \pi \]

This action corresponds to the action of \( \text{Hom}(\pi_1(X), \mathbb{C}^*) \) on \( \mathbb{C}M_B(c) \) and the quotient \( \mathbb{C}N_{DR}(c) \) corresponds to \( \mathbb{C}N_B(c) \). Similarly, the space of \( \text{U}(1) \)-gauge equivalence classes of \( \text{U}(1) \)-connections on \( X \) is \( H^1(X, \mathbb{C}^*) \) which acts on \( \mathbb{C}M_{DR}(c) \) and the quotient is denoted by \( \mathbb{C}N_{DR}(c) \). Define

\[ \mathbb{C}M_{DR} = \bigcup_{c=-\infty}^{\infty} \mathbb{C}M_{DR}(c), \quad \mathbb{C}N_{DR} = \bigcup_{c=-\infty}^{\infty} \mathbb{C}N_{DR}(c) \]
$\mathcal{M}_{DR} = \bigcup_{c=-\infty}^{\infty} \mathcal{M}_{DR}(c), \quad \mathcal{N}_{DR} = \bigcup_{c=-\infty}^{\infty} \mathcal{N}_{DR}(c).$

All the spaces constructed here that contain the symbols $\mathcal{M}_{DR}$ or $\mathcal{N}_{DR}$ will be loosely referred to as de Rham moduli spaces. A central connection is irreducible if $(E,D) = (E_1 \oplus E_2, D_1 \oplus D_2)$ implies $\text{rank}(E_1) = 0$ or $\text{rank}(E_2) = 0.$ The subspace of irreducible elements of a de Rham moduli space will be denoted by an $s$ superscript.

**Theorem 2.1.** The moduli space $\mathcal{C}M_B(c)$ is homeomorphic to $\mathcal{C}M_{DR}(c).$

*Proof.* See [3, 5, 11]. □

Consider all the objects we have defined so far with subscripts $B$ or $DR.$ With Theorem 2.1, one can verify the following: Suppose two objects have subscripts $B$ or $DR.$ Then the two objects are homeomorphic if they only differ in subscripts. For example, $\mathcal{N}_B(c)$ is homeomorphic to $\mathcal{N}_{DR}(c)$.

Since the maximum compact subgroup of $U(p,q)$ is $U(p) \times U(q),$ $(E,D) \in \mathcal{M}_{DR}$ implies $E$ is a direct sum of a $U(p)$ and a $U(q)$-bundle:

$$E = E_p \oplus E_q,$$

where the ranks of $E_p$ and $E_q$ are $p$ and $q,$ respectively. Therefore, associated to each $(E,D)$ are the invariants $d_P = \text{deg}(E_P)$ and $d_Q = \text{deg}(E_Q),$

with $d_P + d_Q = \text{deg}(E) = c.$

The Toledo invariant $\tau$ is $[7, 26, 27]$

$$\tau = 2 \frac{\text{deg}(E_P \otimes E_Q^*)}{n} = 2 \frac{qd_P - pd_Q}{n}.$$

The subspace of $\mathcal{M}_{DR}(c)$ with a fixed Toledo invariant $\tau$ is denoted by $\mathcal{M}_{DR}^\tau.$ By the equivalence of Betti and de Rham moduli spaces, one may define the Toledo invariant on $\mathcal{M}_B(c).$ Denote by $\mathcal{M}_B^\tau$ the subspace of $\mathcal{M}_B(c)$ with a fixed Toledo invariant $\tau.$ The $H^1(X, U(1))$ action on $\mathcal{M}_{DR}(c)$ preserves $\mathcal{M}_{DR}^\tau$ and the quotient is denoted by $\mathcal{N}_{DR}^\tau.$ In the Betti moduli space, the $\text{Hom}(\pi_1(X), U(1))$ action on $\mathcal{M}_B$ preserves $\mathcal{M}_B^\tau,$ and the quotient is denoted by $\mathcal{N}_B^\tau.$

**2.3. The Dolbeault Moduli Space.** Let $E$ be a rank $n$ complex vector bundle over $X$ with $\text{deg}(E) = c.$ Denote by $\Omega$ the canonical bundle on $X.$ A holomorphic structure $\overline{\partial}$ on $E$ induces holomorphic structures on the bundles End($E$) and End($E$) $\otimes \Omega.$ A Higgs bundle is a pair $(E_{\overline{\partial}}, \Phi),$ where $\overline{\partial}$ is a holomorphic structure on $E$ and $\Phi \in H^0(X, \text{End}(E_{\overline{\partial}}) \otimes \Omega).$ Such a $\Phi$ is called a Higgs field. We denote the holomorphic bundle $E_{\overline{\partial}}$ by $V.$
Define the slope of a Higgs bundle \((V, \Phi)\) to be
\[
s(V) = \deg(V) / \text{rank}(V).
\]
For a fixed \(\Phi\), a holomorphic subbundle \(W \subset V\) is said to be \(\Phi\)-invariant if \(\Phi(W) \subset W \otimes \Omega\). A pair \((V, \Phi)\) is stable (semi-stable) if \(W \subset V\) is \(\Phi\)-invariant implies
\[
s(W) < (\leq) s(V).
\]
A Higgs bundle is called poly-stable if it is a direct sum of stable Higgs bundles of the same slope \([11, 22]\).

The gauge group \(G_C(E)\) acts on holomorphic structures by pull-back and on Higgs fields by conjugation. Moreover the \(G_C(E)\) action preserves stability, poly-stability and semi-stability. The Dolbeault moduli space \(\mathcal{CM}_{\text{Dol}}(c)\) on \(E\) (with \(\deg(E) = c\)), is the \(G_C(E)\)-equivalence classes of poly-stable (or \(S\)-equivalence classes of semi-stable \([18]\)) Higgs bundles \((V, \Phi)\) on \(X\) \([11, 12, 18, 22]\). A Higgs bundle is called reducible if it is poly-stable but not stable. Let
\[
\mathcal{CM}_{\text{Dol}} = \bigcup_{c = -\infty}^{\infty} \mathcal{CM}_{\text{Dol}}(c).
\]

If \(D \in \mathcal{CM}_{DR}(c)\), then for any Hermitian metric \(h\) on \(E\), there is a decomposition,
\[
D = D_A + \Psi,
\]
where \(D_A\) is compatible with \(h\) and \(\Psi\) is a 1-form with coefficients in \(p\). The \((0, 1)\) part of \(D_A\) determines a holomorphic structure \(\tilde{\partial}_A\) on \(E\) while the \((1, 0)\) part of \(\Psi\) is a section of the bundle \(\text{End}(E) \otimes \Omega\). There exists a metric \(h\) such that the pair
\[
(V, \Phi) = (E_{\tilde{\partial}_A}, \Psi^{1,0})
\]
so constructed is a poly-stable Higgs bundle \([11, 21, 22]\). Therefore this construction gives a map
\[
f : \mathcal{CM}_{DR}(c) \rightarrow \mathcal{CM}_{\text{Dol}}(c).
\]

**Theorem 2.2** (Corlette, Donaldson, Hitchin, Simpson). The map \(f\) is a homeomorphism.

**Proof.** See \([3, 5, 11, 21]\) \(\square\)

3. The \(U(p, q)\)-Yang-Mills Connections.

Assume \(p \geq q\) and \(p + q = n\). From the previous section, we know that \(\mathcal{M}_{DR} \subset \mathcal{CM}_{DR}\). Let \(D \in \mathcal{CM}_{DR}(c)\) be a \(\text{GL}(n, \mathbb{C})\)-Yang-Mills connection on a rank \(n\) vector bundle
\[
E \rightarrow X.
\]
Proposition 3.1. $D$ is a $U(p,q)$-Yang-Mills connection if and only if its corresponding Higgs bundle $(V, \Phi) \in \mathbb{C}M_{\text{Dol}}(c)$ satisfies the following two conditions:

1) $V$ is decomposable into a direct sum:

$$V = V_P \oplus V_Q,$$

where $V_P, V_Q$ are of rank $p, q$, respectively.

2) The Higgs field decomposes into two maps:

$$\Phi_1 : V_P \longrightarrow V_Q \otimes \Omega,$$

$$\Phi_2 : V_Q \longrightarrow V_P \otimes \Omega.$$

Proof. Suppose $D$ is a $U(p,q)$-Yang-Mills connection. Denote by $h$ the Hermitian-Yang-Mills metric on $(E, D)$. Then $D$ decomposes as

$$D = D_A + \Psi,$$

where $D_A$ is the part compatible with $h$. The Cartan decomposition ($g = \mathfrak{k} \oplus \mathfrak{p}$) for $u(p,q)$ is

$$u(p,q) = (u(p) \oplus u(q)) \oplus \mathfrak{p}.$$

If we take the standard representation of $u(p,q)$, then elements in $\mathfrak{k}$ are of the form

$$\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix},$$

where $a \in u(p), b \in u(q)$, respectively. The elements in $\mathfrak{p}$ are then of the form

$$\begin{pmatrix} 0 & b \\ c & 0 \end{pmatrix},$$

where $b \in \text{Hom}(\mathbb{C}^q, \mathbb{C}^p), c \in \text{Hom}(\mathbb{C}^p, \mathbb{C}^q)$, respectively. Hence on local charts, $D_A$ and $\Psi$ have coefficients in $\mathfrak{k}$ and $\mathfrak{p}$, respectively. In particular, the connection $D_A$ is reducible.

The Higgs bundle corresponding to $D$ is $(E_D, \Phi)$ where $D_A$ is the $(0,1)$-part of $D_A$ and $\Phi$, the $(1,0)$-part of $\Psi$, is considered as a holomorphic bundle map:

$$\Phi : V \longrightarrow V \otimes \Omega.$$}

Since $D_A$ has coefficient in $\mathfrak{k}$, the holomorphic structure on $V$ defined by $D_A^{0,1}$ is a direct sum:

$$V = V_P \oplus V_Q.$$}

Since $\Psi$ is block off-diagonal, $\Phi$ is also block off-diagonal implying $\Phi$ can be decomposed into two maps:

$$\Phi_1 : V_P \longrightarrow V_Q \otimes \Omega,$$

$$\Phi_2 : V_Q \longrightarrow V_P \otimes \Omega.$$}

This proves the only if part of the proposition.
Suppose \((V, \Phi)\) is a Higgs bundle that satisfies the two conditions of Proposition 3.1. Let \(\alpha\) be the constant gauge
\[
\alpha = \begin{pmatrix}
I_p & 0 \\
0 & -I_q
\end{pmatrix},
\]
where \(I_p, I_q\) are \(p \times p, q \times q\) identity matrices, respectively. Then \(\alpha\) acts on the space of holomorphic structures on \(E\) and fixes \(V\). Moreover,
\[
\alpha \Phi \alpha^{-1} = -\Phi
\]
since \(\Phi\) is of the form
\[
\Phi = \begin{pmatrix}
0 & \Phi_1 \\
\Phi_2 & 0
\end{pmatrix}.
\]
Hence by a theorem of Simpson, the corresponding Hermitian-Yang-Mills metric \(h\) is invariant under the action of \(\alpha\) \([21]\). In other words, on local charts, \(h\) is a Hermitian matrix of the form
\[
h = \begin{pmatrix}
a & 0 \\
0 & d
\end{pmatrix},
\]
where \(a, d\) are Hermitian matrices of dimension \(p \times p, q \times q\), respectively. Hence the corresponding Yang-Mills connection is
\[
D = D_A + \Phi + \Phi^\dagger,
\]
where \(\Phi^\dagger\) is the adjoint of \(\Phi\) with respect to \(h\). In local coordinates, \(D_A\) has coefficient of the form
\[
\begin{pmatrix}
a & 0 \\
0 & d
\end{pmatrix}
\]
and \(\Phi + \Phi^\dagger\) is of the form
\[
\begin{pmatrix}
0 & b \\
b^\dagger & 0
\end{pmatrix}.
\]
Hence \(D_A\) and \(\Phi + \Phi^\dagger\) have coefficients in \(u(p) \oplus u(q)\) and \(p, q\), respectively. This implies \(D\) is a \(U(p, q)\)-Yang-Mills connection. \(\square\)

Denote by \(\mathcal{M}_{\text{Dol}}(c)\) the subspace of \(\mathbb{C}\mathcal{M}_{\text{Dol}}(c)\) satisfying the hypothesis of Proposition 3.1. Then \(\mathcal{M}_{\text{Dol}}(c)\) is homeomorphic to \(\mathcal{M}_{\text{DR}}(c)\).

The invariants \(d_P, d_Q\) and \(\tau\) on \((E, D)\) translate to invariants on the corresponding \(U(p, q)\)-Higgs bundles \((V_P \oplus V_Q, \Phi)\):
\[
d_P = \deg(V_P), \quad d_Q = \deg(V_Q), \quad \tau = 2\frac{qd_P - pd_Q}{n}.
\]
The subspace of \(\mathcal{M}_{\text{Dol}}(c)\) with a fixed Toledo invariant \(\tau\) is denoted by \(\mathcal{M}^\tau_{\text{Dol}}\).
4. Group Actions and Kähler Structures on $\mathcal{CM}_{\text{Dol}}$.

4.1. The Action of line bundles. The space of holomorphic line bundles, $\text{H}^1(X, \mathcal{O}^*)$, acts freely on $\mathcal{CM}_{\text{Dol}}$ as follows:

\[
\text{H}^1(X, \mathcal{O}^*) \times \mathcal{CM}_{\text{Dol}} \rightarrow \mathcal{CM}_{\text{Dol}},
\]

\[
(L, (V, \Phi)) \mapsto (V \otimes L, \Phi \otimes 1),
\]

where 1 is the identity map on $L$. An immediate consequence is:

**Proposition 4.1.** If $c_1 \equiv c_2 \mod n$, then $\mathcal{CM}_{\text{Dol}}(c_1)$ is homeomorphic to $\mathcal{CM}_{\text{Dol}}(c_2)$.

4.2. The Action of $\text{H}^0(X, \Omega)$. The vector space $\text{H}^0(X, \Omega)$ acts freely on $\mathcal{CM}_{\text{Dol}}$ as follows:

\[
\text{H}^0(X, \Omega) \times \mathcal{CM}_{\text{Dol}} \rightarrow \mathcal{CM}_{\text{Dol}},
\]

\[
(\phi, (V, \Phi)) \mapsto (V, \Phi + \phi 1).
\]

The actions of $\text{H}^1(X, \mathcal{O}^*)$ and $\text{H}^0(X, \Omega)$ commute and the quotient is defined to be

\[
\mathcal{CM}^\tau_{\text{Dol}} = \mathcal{CM}_{\text{Dol}}/(\text{H}^1(X, \mathcal{O}^*) \times \text{H}^0(X, \Omega)).
\]

The $\text{H}^1(X, \mathcal{O}^*)$ action preserves the subspaces $\mathcal{M}_{\text{Dol}}(c)$ and $\mathcal{M}^\tau_{\text{Dol}}$. The quotients are defined to be

\[
\mathcal{N}_{\text{Dol}}(c) = \mathcal{M}_{\text{Dol}}(c)/\text{H}^1(X, \mathcal{O}^*),
\]

\[
\mathcal{N}^\tau_{\text{Dol}} = \mathcal{M}^\tau_{\text{Dol}}/\text{H}^1(X, \mathcal{O}^*).
\]

All the spaces constructed so far that contain the symbols $\mathcal{M}_{\text{Dol}}$ or $\mathcal{N}_{\text{Dol}}$ will be loosely referred to as the Dolbeault moduli spaces. The subspace of stable Higgs bundles of a Dolbeault moduli space will be denoted by an $s$ superscript. For example, $\mathcal{CM}^s_{\text{Dol}}$ will denote the subspace of irreducible elements of $\mathcal{CM}_{\text{Dol}}$.

**Remark 1.** The Betti, de Rham and Dolbeault moduli spaces $\mathcal{CM}_B$, $\mathcal{CM}_{\text{Dol}}$ and $\mathcal{CM}_{\text{Dol}}$ constructed here are variations of those of Simpson’s [22, 23, 24].

With Theorems 2.1 and 2.2, one can obtain the following equivalence relations between the various Betti, de Rham and Dolbeault moduli spaces.

**Corollary 4.2.** Suppose $\mathcal{M}^\tau_{DR} \subset \mathcal{M}_{DR}(c)$. Then one obtains the following commutative diagram:

\[
\begin{array}{ccc}
\mathcal{M}^\tau_B & \rightarrow & \mathcal{M}_B(c) & \rightarrow & \mathcal{CM}_B(c) \\
\downarrow & & \downarrow & & \downarrow \\
\mathcal{M}^\tau_{DR} & \rightarrow & \mathcal{M}_{DR}(c) & \rightarrow & \mathcal{CM}_{DR}(c) \\
\downarrow & & \downarrow & & \downarrow \\
\mathcal{M}^\tau_{\text{Dol}} & \rightarrow & \mathcal{M}_{\text{Dol}}(c) & \rightarrow & \mathcal{CM}_{\text{Dol}}(c).
\end{array}
\]
Moreover the horizontal maps are continuous injections and vertical maps are homeomorphisms. One obtains three additional commutative diagrams by respectively replacing the symbol \( M \) by \( M^s \), \( N \) and \( N^s \) in the above diagram. In the case of \( M^s \), the maps in the commutative diagram are smooth.

4.3. The Dual Higgs Bundles. There is a \( \mathbb{Z}_2 \) action on \( \mathcal{C}M_{Dol} \). Let 
\[(V, \Phi) \in \mathcal{C}M_{Dol} \text{ where } \Phi \text{ is a holomorphic map:}\]
\[\Phi : V \rightarrow V \otimes \Omega.\]

This induces a map on the dual bundles
\[\Phi^* : V^* \otimes \Omega^* \rightarrow V^*.\]

Tensoring with \( \Omega \),
\[\Phi^* \otimes 1 : V^* \rightarrow V^* \otimes \Omega,\]
where 1 denotes the identity map on \( \Omega \). This produces the dual Higgs bundle \((V^*, \Phi^* \otimes 1)\). We shall abbreviate it as \((V^*, \Phi^*)\).

**Proposition 4.3.** If \((V, \Phi) \in \mathcal{C}M_{Dol}(c)\), then \((V^*, \Phi^*) \in \mathcal{C}M_{Dol}(-c)\).

**Proof.** One must show that \((V, \Phi)\) is stable (semi-stable) implies \((V^*, \Phi^*)\) is stable (semi-stable). Suppose \( W_1 \subset V^* \) is \( \Phi^* \)-invariant. Then we have the following commutative diagram
\[
\begin{array}{cccccc}
0 & \rightarrow & W_1 & \rightarrow & V^* & \rightarrow & W_2 & \rightarrow & 0 \\
\downarrow \Phi^* & & \downarrow \Phi^* & & \downarrow \Phi^* & & \\
0 & \rightarrow & W_1 \otimes \Omega & \rightarrow & V^* \otimes \Omega & \rightarrow & W_2 \otimes \Omega & \rightarrow & 0
\end{array}
\]
where \( W_2 = V^*/W_1 \). The proposition follows by dualizing the diagram. \( \square \)

In light of Propositions 4.1 and 4.3 we have:

**Corollary 4.4.** If \( c_2 = \pm c_1 \mod n \), then \( \mathcal{C}M_{Dol}(c_1) \) is homeomorphic to \( \mathcal{C}M_{Dol}(c_2) \).

4.4. The \( U(1) \) and \( \mathbb{C}^* \)-Actions on the Complex Moduli Spaces. If \((V, \Phi) \in \mathcal{C}M_{Dol}(c)\), then for \( t \in \mathbb{C}^* \), \((V, t\Phi) \in \mathcal{C}M_{Dol}(c)\). This defines an analytic action \([11, 12, 22]\)
\[\mathbb{C}^* \times \mathcal{C}M_{Dol}(c) \rightarrow \mathcal{C}M_{Dol}(c).\]

Since \( U(1) \subset \mathbb{C}^* \), this also induces a \( U(1) \)-action on \( \mathcal{C}M_{Dol}(c) \).
4.5. The Moment Map. The moduli space $\mathcal{CM}_{\text{Dol}}(c)^s$ is Kähler [11, 12]. Denote by $i, \omega$ the corresponding complex and symplectic structures, respectively. Define the Morse function $[11, 12]

\begin{align*}
m : \mathcal{CM}_{\text{Dol}}(c)^s & \longrightarrow \mathbb{R}, \\
m(V, \Phi) & = 2i \int_X \text{tr}(\Phi \Phi^\dagger),
\end{align*}

where $\Phi^\dagger$ is the adjoint of $\Phi$ with respect to the Hermitian-Yang-Mills metric on $(E, D)$. Denote by $X$ the vector field on $\mathcal{CM}_{\text{Dol}}(c)^s$ such that

$$\text{grad} \ m = iX.$$ 

**Theorem 4.5.**

1) The map $m$ is proper.
2) The $U(1)$-action generates $X$.
3) The $\mathbb{C}^*$ action is analytic with respect to $i$; therefore, the orbit of $\mathbb{C}^*$ is locally an analytic subvariety with respect to $i$.

**Proof.** See [11, 12, 22]. \qed

**Corollary 4.6.** Each component of $\mathcal{CM}_{\text{Dol}}(c)$ contains a point that is a local minimum of $m$.

**Corollary 4.7.** If the $\mathbb{C}^*$ action preserves $M \subset \mathcal{CM}_{\text{Dol}}(c)^s$, then the gradient flow $\text{grad} \ m$ preserves $M$.

Let $m_r$ be the restriction of $m$ to the subspace $\mathcal{M}_{\text{Dol}}^r \subset \mathcal{CM}_{\text{Dol}}(c)$.

**Corollary 4.8.** Every component of $\mathcal{M}_{\text{Dol}}^r$ contains a point that is a local minimum of $m_r$. If $(V, \Phi)$ is stable and is a local minimum of $m_r$, then $(V, \Phi)$ is a critical point of $m$.

**Proof.** Consider $\mathcal{M}_B^r \subset \mathcal{M}_B(c) \subset \mathcal{CM}_B(c)$. Since $U(p, q)$ is closed in $\text{GL}(n, \mathbb{C})$, $\mathcal{M}_B(c)$ is a closed subspace of $\mathcal{CM}_B(c)$. Since the obstruction map $o_2$ is continuous, $\mathcal{M}_B^r$ is a closed subspace of $\mathcal{M}_B(c)$. Hence $\mathcal{M}_B^r$ is closed in $\mathcal{CM}_B(c)$. Hence by Theorem 4.5, $m_r$ is proper. Thus each component of $\mathcal{M}_{\text{Dol}}^r$ contains a local minimum of $m_r$.

The points in $(\mathcal{M}_{\text{Dol}}^r)^s$ are smooth. Suppose $(V, \Phi) \in (\mathcal{M}_{\text{Dol}}^r)^s$. Then $(V, \Phi)$ is of the form described in Proposition 3.1. Hence the $\mathbb{C}^*$ action preserves the subspace $(\mathcal{M}_{\text{Dol}}^r)^s \subset \mathcal{CM}_{\text{Dol}}^s$. By Corollary 4.7, the gradient flow of $m$ preserves $(\mathcal{M}_{\text{Dol}}^r)^s$. Hence

$$\text{grad} \ m_r = \text{grad} \ m = iX.$$

If $m_r$ is a local minimum at $(V, \Phi)$, then

$$\text{grad} \ m(V, \Phi) = \text{grad} \ m_r(V, \Phi) = 0.$$

Hence $(V, \Phi)$ is a critical point of $m$. \qed
5. Bounds on Invariants.

In this section, we assume \( q = 1 \) and let \( n = p + q = p + 1 \). In light of Proposition 4.3 and Corollary 4.4, one may further assume that \( \tau \geq 0 \) and \( 0 \leq c < n \), or equivalently,

\[
s(V_Q) \leq s(V) \leq s(V_P), 0 \leq c < n.
\]

**Proposition 5.1.** If \((V, \Phi) = (V_P \oplus V_Q, (\Phi_1, \Phi_2)) \in \mathcal{M}_{Dol}(c)^a (\mathcal{M}_{Dol}(c))\), then

\[
d_P < \left(\leq\right) \frac{c(n-1)}{n} + (g-1)
\]

\[
d_Q > \left(\geq\right) \frac{c}{n} - (g-1).
\]

**Proof.** Suppose \((V_P \oplus V_Q, \Phi) \in \mathcal{M}_{Dol}(c)^a \) with \( \Phi = (\Phi_1, \Phi_2) \) in the notation of Proposition 3.1. Since \( s(V_P) \geq s(V) \),

\[
\Phi_1 : V_P \longrightarrow V_Q \otimes \Omega
\]

is non-zero.

Construct the canonical factorization for \( \Phi_1 \) [20]: There exist holomorphic bundles \( V_1, V_2 \) and \( W_1, W_2 \) such that the following diagram

\[
\begin{array}{cccccc}
0 & \longrightarrow & V_1 & \xrightarrow{f_1} & V_P & \xrightarrow{f_2} & V_2 & \longrightarrow & 0 \\
& & \Phi_1 \downarrow & & \varphi \downarrow & & \\
0 & \longleftarrow & W_2 & \xleftarrow{g_2} & V_Q \otimes \Omega & \xleftarrow{g_1} & W_1 & \longleftarrow & 0
\end{array}
\]

commutes, and the rows are exact, \( \text{rank}(V_2) = \text{rank}(W_1) \) and \( \varphi \) has full rank at a generic point of \( X \). This implies

\[
\begin{cases}
\deg(V_1) + \deg(V_2) = d_P \\
\deg(W_1) + \deg(W_2) = d_Q + 2(g-1).
\end{cases}
\]

Since \( \Phi_1 \neq 0 \), we have \( \varphi \neq 0 \), \( \text{rank}(W_2) = 0 \) and \( W_1 = V_Q \otimes \Omega \).

The case of \( p = 1 \) has been dealt with by Hitchin [11], so we assume \( p > 1 \). Then \( V_1 \) is a \( \Phi \)-invariant subbundle of positive rank. Stability implies

\[
s(V_1) < s(V) = (d_P + d_Q)/n = c/n.
\]

Since the map

\[
V_2 \xrightarrow{\varphi} W_1 = (V_Q \otimes \Omega)
\]

is not trivial,

\[
\deg(V_2) \leq \deg(W_1) = \deg(V_Q \otimes \Omega).
\]
So one has
\[
\begin{aligned}
&\begin{cases}
  s(V_1) &< s(V) \\
  d_P &\quad = \quad \text{deg}(V_1) + \text{deg}(V_2) \\
  \text{deg}(V_2) &\quad \leq \quad d_Q + 2(g - 1).
\end{cases}
\end{aligned}
\]
This implies
\[
d_P < \frac{(n-2)c}{n} + d_Q + 2(g - 1).
\]
Since \(d_P + d_Q = c\),
\[
d_P < \frac{c(n-1)}{n} + (g - 1)
\]
and
\[
d_Q > \frac{c}{n} - (g - 1).
\]
When \((V, \Phi)\) is semi-stable, one has either \(\Phi \not\equiv 0\) or \(\Phi \equiv 0\). In the former case, one has \(s(V_1) \leq s(V)\) implying
\[
\begin{aligned}
  d_P &\quad \leq \quad \frac{c(n-1)}{n} + (g - 1) \\
  d_Q &\quad \geq \quad \frac{c}{n} - (g - 1).
\end{aligned}
\]
In the latter case, \(V_p\) is \(\Phi\)-invariant. By the assumption \(s(V_Q) \leq s(V_P)\), \(d_P = d_Q = 0\) and \(\tau = 0\). \(\square\)

By definition,
\[
\tau = 2 \frac{d_P - pdQ}{n}
\]
\[
\leq 2 \frac{c(n-1)}{n} + (g - 1) - (n-1)\frac{c}{n} + (n-1)(g - 1)
\]
\[
= 2(g - 1).
\]
Equality holds only when \((V, \Phi)\) is semi-stable but not stable, in which case, the associated flat connection is reducible. This proves Theorem 1.2.

6. Reducible Higgs Bundles.

Let \(p = 2\) and \(q = 1\) and assume \(\tau \geq 0\) and \(0 \leq c < 3\). By definition, a reducible poly-stable Higgs bundle is a direct sum of stable Higgs bundles of the same slope. These Higgs bundles correspond to the reducible representations in \(\mathcal{M}_B\). A direct computation shows that if \((V, \Phi)\) is reducible, then
\[
\text{deg}(V) = d_P + d_Q = 0
\]
and the associated Toledo invariant \(\tau\) is an even integer. Hence one has:
Proposition 6.1. If $c = \deg(V) \neq 0$ and $(V, \Phi) \in \mathcal{M}_{\text{Dol}}(c)$, then $(V, \Phi)$ is stable. In particular, $\mathcal{M}_{\text{Dol}}(c)$ is smooth.

An example of a reducible Higgs bundle is $(\mathcal{O} \oplus \Omega^1 \oplus \Omega^{-1}, \Phi)$, where

$$\Phi : \Omega^1 \longrightarrow \Omega^{-1} \oplus \Omega$$

is a holomorphic bundle isomorphism. That is, $\Phi$ is of the form

$$\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 1 & 0
\end{pmatrix}.$$

The Toledo invariant in this case is $2(g - 1)$. All the flat $\text{U}(2, 1)$-connections with $\tau = 2(g - 1)$ are reducible by Proposition 5.1. The fact that there is no irreducible deformation for the $\text{U}(2, 1)$-connections with $\tau = 2(g - 1)$ was first demonstrated by Toledo [26]. In particular, this component is connected [6, 11].

7. Hodge Bundles and Deformation.

Let $p = 2$ and $q = 1$ and assume $\tau \geq 0$ and $0 \leq c < 3$. A Hodge bundle on $X$ is a direct sum of holomorphic bundles [22]

$$V = \bigoplus_{s,t} V^{s,t}$$

together with holomorphic maps (Higgs field)

$$\Phi_1 : V^{s,t} \longrightarrow V^{s-1,t+1} \otimes \Omega.$$ 

An immediate consequence of Proposition 3.1 is:

Corollary 7.1. Suppose $(V_P \oplus V_Q, (\Phi_1, \Phi_2)) \in \mathcal{M}_{\text{Dol}}(c)$ (in the notations of Proposition 3.1). Then $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$ is a Hodge bundle if and only if $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$ is either binary or ternary in the following sense:

1) Binary: $\Phi_2 \equiv 0$.

2) Ternary: $V_P = V_1 \oplus V_2$ and the Higgs field consists of two maps:

$$\Phi_1 : V_2 \longrightarrow V_Q \otimes \Omega,$$

$$\Phi_2 : V_Q \longrightarrow V_1 \otimes \Omega.$$ 

Denote by $B(d_P, d_Q)$ the space of all poly-stable (or $S$-equivalence classes of semi-stable) binary Hodge bundles $(V_P \oplus V_Q, (\Phi_1, 0))$ with $\deg(V_P) = d_P$ and $\deg(V_Q) = d_Q$. Denote by $T(d_1, d_2, d_Q)$ the space of all poly-stable (or $S$-equivalence classes of semi-stable) ternary Hodge bundles $(V_1 \oplus V_2 \oplus V_Q, (\Phi_1, \Phi_2))$ with $\deg(V_1) = d_1$, $\deg(V_2) = d_2$ and $\deg(V_Q) = d_Q$. Denote the subspaces of stable Hodge bundles by $B(d_P, d_Q)^s, T(d_1, d_2, d_Q)^s$. When $\tau$ is not an integer, these are the type $(2,1)$ and $(1,1,1)$ spaces in [8]. Note the $(1,2)$ types give $\tau < 0$ and therefore need not be considered here.
Proposition 7.2. Every stable binary Hodge bundle in \((\mathcal{M}_{\text{Dol}}^\tau)^s\) may be deformed to a stable ternary Hodge bundle within \(\mathcal{M}_{\text{Dol}}^\tau\).

A family (or flat family) of Higgs pairs \((V_Y, \Phi_Y)\) is a variety \(Y\) such that there is a vector bundle \(V_Y\) on \(X \times Y\) together with a section \(\Phi_Y \in \Gamma(Y, (\pi_Y)_* (\pi_X^* \Omega \otimes \text{End}(V_Y)))\) [18]. \(\mathcal{CM}_{\text{Dol}}\) being a moduli space implies that if \(Y\) is a family of stable (poly-stable or \(S\)-equivalence classes of semi-stable) Higgs bundles, then there is a natural morphism \([15, 17]\)

\[ t : Y \longrightarrow \mathbb{C}\mathcal{M}_{\text{Dol}}. \]

Moreover \(t\) takes every point \(y \in Y\) to the point of \(\mathbb{C}\mathcal{M}_{\text{Dol}}\) that corresponds to the Higgs bundle in the family over \(y\) [15, 17, 18].

The space \(\mathcal{M}_{\text{Dol}}(c)\) is a subvariety of \(\mathbb{C}\mathcal{M}_{\text{Dol}}(c)\); hence, to show that two stable (poly-stable or \(S\)-equivalence classes of semi-stable) Higgs bundles \((V_1, \Phi_1)\) and \((V_2, \Phi_2)\) belong to the same component of \(\mathcal{M}_{\text{Dol}}(c)\), it suffices to exhibit a connected family \(Y\) (within \(\mathcal{M}_{\text{Dol}}(c)\)) of stable (poly-stable or \(S\)-equivalence classes of semi-stable) Higgs bundles containing both \((V_1, \Phi_1)\) and \((V_2, \Phi_2)\).

**Proof.** Suppose \((V, \Phi) = (V_P \oplus V_Q, (\Phi_1, 0)) \in B(d_P, d_Q)^s \subset (\mathcal{M}_{\text{Dol}}^\tau)^s\). Since \(s(V_P) \geq s(V)\) (This is due to the assumption \(\tau \geq 0\), and \(0 \leq c < 3\), \(\Phi_1 \not\equiv 0\). Construct the canonical factorization for \(\Phi_1:\)

\[
\begin{array}{cccccc}
0 & \longrightarrow & V_1 & \xrightarrow{f_1} & V_P & \xrightarrow{f_2} & V_2 & \longrightarrow & 0 \\
& & \Phi_1 & \downarrow & \varphi & & \\
0 & \longleftarrow & W_2 & \xleftarrow{g_2} & V_Q \otimes \Omega & \xleftarrow{g_1} & W_1 & \longleftarrow & 0.
\end{array}
\]

\(V_1\) being \(\Phi_1\) invariant implies

\[\text{deg}(V_1) = s(V_1) < s(V) \leq s(V_P) \leq s(V_2) = \text{deg}(V_2).\]

The space \(\text{Pic}^0(X)\) of line bundles of degree 0 over \(X\) is identified with the Jacobi variety \(J_0(X)\). Construct the universal bundle [2, 19]

\[ U \longrightarrow X \times J_0(X) \]

such that \(U\) restricts to the bundle \(L \otimes V_1 \otimes V_2^{-1}\) on \((X, L)\). Let \(\pi\) be the projection

\[ \pi : X \times J_0(X) \longrightarrow J_0(X). \]

Applying the right derived functor \(R^1\) to \(\pi\) gives the sheaf \(\mathcal{F} = R^1 \pi_* (U)\) [10] such that

\[\mathcal{F}|_L = H^1(X, L \otimes V_1 \otimes V_2^{-1}).\]

Since

\[\text{deg}(L \otimes V_1 \otimes V_2^{-1}) = \text{deg}(V_1) - \text{deg}(V_2) < 0,\]
by Riemann-Roch,
\[ h^1(L \otimes V_1 \otimes V_2^{-1}) = h^0(L \otimes V_1 \otimes V_2^{-1}) - \deg(L \otimes V_1 \otimes V_2^{-1}) + (g - 1) \]
\[ = \deg(V_2) - \deg(V_1) + (g - 1) \]
is a constant. By Grauert’s theorem, \( F \) is locally free, hence, is associated with a vector bundle
\[ F \hookrightarrow J_0(X) \]
of rank \( \deg(V_2) - \deg(V_1) + (g - 1) \). In particular the total space \( F \) is smooth and parameterizes extensions [9, 10]:
\[ 0 \to L \otimes V_1 \xrightarrow{f_3} W_P \xrightarrow{f_4} V_2 \to 0 \]
for fixed \( V_1, V_2 \). Tensoring the above sequence with \( \Omega \) gives:
\[ 0 \to L \otimes V_1 \otimes \Omega \xrightarrow{g_3} W_P \otimes \Omega \xrightarrow{g_4} V_2 \otimes \Omega \to 0. \]
Fix \( \varphi \). Then \( F \) also parameterizes a family of Higgs bundles \((W_P, \Phi_1')\) that fit into the factorization
\[ 0 \to L \otimes V_1 \xrightarrow{f_3} W_P \xrightarrow{f_4} V_2 \to 0 \]
\[ \Phi_1' \downarrow \] \[ \Phi \downarrow \]
\[ 0 \leftarrow W_2 \xrightarrow{g_2} V_Q \otimes \Omega \xrightarrow{g_1} W_1 \leftarrow 0. \]

Let \( V \subset F \) be the subset of stable extensions (i.e., \( W_P \in V \) implies \( W_P \) is a stable holomorphic bundle [19]).

**Lemma 7.3.** \( V \cap H^1(L \otimes V_1 \otimes V_2^{-1}) \) and \( V \) are open and dense in \( H^1(L \otimes V_1 \otimes V_2^{-1}) \) and \( F \), respectively. Moreover if \( W_P \in V \), then \((W_P \oplus V_Q, (\Phi_1', 0))\) is stable.

**Proof.** Since \( \deg(L \otimes V_1) < \deg(V_2) \) for each \( L \in J_0(X) \), by a theorem of Lange and Narasimhan [13], there always exists a stable extension \( W_P \in H^1(L \otimes V_1 \otimes V_2^{-1}) \). In addition, a theorem of Maruyama states that being stable is an open property [14]. The open dense property follows from the smoothness of \( F \) and \( H^1(L \otimes V_1 \otimes V_2^{-1}) \).

Let \( p_P, p_Q \) be the projections of \( W_P \oplus V_Q \) onto its \( W_P \) and \( V_Q \) factors, respectively. Suppose \( W \) is \((\Phi_1', 0)\)-invariant. Suppose \( W \) has rank 1. If \( P_Q(W) = 0 \), then \( W = L \otimes V_1 \); otherwise, \( \deg(W) \leq \deg(V_Q) \). In either case, \( s(W) < s(V) \). Suppose \( W \) has rank 2. If \( p_Q(W) = 0 \), then \( W = W_P \) and \( s(W) < s(V) \). Suppose \( P_Q(W) \neq 0 \). Then there exists a line bundle \( L_1 \) such that
\[ 0 \to L_1 \to W \xrightarrow{p_Q} p_Q(W) \to 0. \]

Now let \( L_P = p_P(L_1) \subset W_P \). Then
\[ \deg(W) = \deg(L_1) + \deg(p_Q(W)) \leq \deg(L_P) + \deg(V_Q). \]
Since \( W_P \) is stable, \( s(L_P) < s(W_P) \). By the assumptions \( \tau \geq 0 \) and \( 0 \leq c < 3 \), one has \( s(V_Q) \leq 0 \) and \( s(W_P) \geq 0 \). Therefore,

\[
s(W) \leq s(L_P \oplus V_Q) = \frac{s(L_P) + s(V_Q)}{2} < \frac{s(W_P) + s(V_Q)}{2} = s(V).
\]

Thus \( (W_P \oplus V_Q, (\Phi'_1, 0)) \) is stable. \( \square \)

Since \( \Phi_1 \neq 0 \), \( \deg(V_2) \leq d_Q + 2(g - 1) \) and

\[
\deg(V_1) = d_P \neq \deg(V_2) \geq d_P - d_Q - 2(g - 1).
\]

Hence

\[
\deg(V_Q^{-1} \otimes V_1 \otimes \Omega) = -d_Q + \deg(V_1) + 2(g - 1) \geq d_P - 2d_Q > 0.
\]

Hence there exists \( L' \in J(X) \) such that

\[
h^0(V_Q^{-1} \otimes L' \otimes V_1 \otimes \Omega) > 0
\]

implying there exists a non-trivial holomorphic map

\[
\phi : V_Q \longrightarrow L' \otimes V_1 \otimes \Omega.
\]

Fix \( \phi \neq 0 \). By Lemma 7.3, the family parameterized by \( \mathcal{V} \) contains both \( (V_P \oplus V_Q, (\Phi_1, 0)) \) and \( (W_P \oplus V_Q, (\Phi'_1, 0)) \) implying there is deformation between the two.

Set \( L = L' \) and \( \Phi'_2 = g_3 \circ \phi \). Then the family of stable Higgs bundles parameterized by \( H^0(X, V_Q^{-1} \otimes L' \otimes V_1 \otimes \Omega) \) contains \( (W_P \oplus V_Q, (\Phi'_1, 0)) \) and \( (W_P \oplus V_Q, (\Phi'_1, \Phi'_2)). \)

Now the family of bundle extensions of \( V_2 \) by \( L' \otimes V_1 \) is \( H^1(L' \otimes V_1 \otimes V_2^{-1}) \).

With a fixed \( \phi \) and the canonical factorization with \( \varphi \) fixed, \( H^1(L' \otimes V_1 \otimes V_2^{-1}) \) parameterizes a family of Higgs bundles. This family contains \( (W_P \oplus V_Q, (\Phi'_1, \Phi'_2)). \) The zero element in \( H^1(L' \otimes V_1 \otimes V_2^{-1}) \) corresponds to the bundle extension

\[
0 \longrightarrow L' \otimes V_1 \xrightarrow{f_5} (L' \otimes V_1) \oplus V_2 \xrightarrow{f_6} V_2 \longrightarrow 0.
\]

Tensoring with \( \Omega \) gives

\[
0 \longrightarrow L' \otimes V_1 \otimes \Omega \xrightarrow{g_5} ((L' \otimes V_1) \oplus V_2) \otimes \Omega \xrightarrow{g_6} V_2 \otimes \Omega \longrightarrow 0.
\]

**Lemma 7.4.** If \( (W_P \oplus V_Q, (\Phi'_1, \Phi'_2)) \) is stable (semi-stable), then \( H^1(L' \otimes V_1 \otimes V_2^{-1}) \) parameterizes a stable (semi-stable) family.

**Proof.** Suppose \( (U_P \oplus V_Q, (\Psi_1, \Psi_2)) \in H^1(L' \otimes V_1 \otimes V_2^{-1}) \) and \( W \subset U_P \oplus V_Q \) is \( (\Psi_1, \Psi_2) \)-invariant. Since \( \varphi, \phi \neq 0 \), one has \( W = V_1 \) or \( W = V_Q \oplus V_1 \). A direct computation shows \( s(W) < s(U_P \oplus V_Q) \) \( (s(W) \leq s(U_P \oplus V_Q)). \) \( \square \)
Proposition 7.2 follows from Lemma 7.4 because the family of Higgs bundles parameterized by $H^1(L' \otimes V_1 \otimes V_2^{-1})$ contains $(W_P \oplus V_Q, (\Phi'_1, \Phi'_2))$ and $((L' \otimes V_1) \oplus V_2 \otimes V_Q, (g_1 \circ \varphi \circ f_6, g_5 \circ \phi))$.

To summarize, a stable binary Hodge bundle $(V_P \oplus V_Q, (\Phi_1, 0))$ is first deformed to $(W_P \oplus V_Q, (\Phi'_1, 0))$ such that non-trivial holomorphic maps exist between $V_Q$ and $(L' \otimes V_1) \otimes \Omega \subset W_P \otimes \Omega$. Such a non-trivial map $\Phi'_2$ is then chosen and attached to the existing Higgs field $\Phi'_1$. Finally $W_P$ is deformed to a direct sum making the resulting stable Higgs bundle a ternary Hodge bundle. \[ \Box \]

Let $B = B(0, 0) \backslash (B(0, 0)^s \cup T(0, 0, 0))$.

**Proposition 7.5.** $B$ is connected and can be deformed to a stable ternary Hodge bundle in $M^0_{\text{Dol}}$.

**Proof.** Consider the space $U \times J_0(X)$, where $J_0(X)$ is the Jacobi variety identified with the set of holomorphic line bundles of degree zero on $X$ and $U$ is the moduli space of rank-2 poly-stable holomorphic bundles of degree 0 on $X$. The space $U$ is connected \[ 2, 19 \]. Hence $U \times J_0(X)$ is connected. Each poly-stable Higgs bundle in $B$ is contained in the family of Higgs bundles parameterized by $U \times J_0(X)$. Hence the natural morphism

$$ t : U \times J_0(X) \longrightarrow B $$

is surjective. This proves that the set $B$ is connected.

Choose holomorphic line bundles $V_1, V_2, V_Q$ of degrees $-1, 1, 0$, respectively such that

$$ h^0(X, V_2^{-1} \otimes V_Q \otimes \Omega) > 0, $$

$$ h^0(X, V_Q^{-1} \otimes V_1 \otimes \Omega) > 0. $$

Choose

$$ 0 \neq \psi_1 \in H^0(X, V_2^{-1} \otimes V_Q \otimes \Omega) $$

$$ 0 \neq \psi_2 \in H^0(X, V_Q^{-1} \otimes V_1 \otimes \Omega). $$

The space of extension of $V_2$ by $V_1$,

$$ 0 \longrightarrow V_1 \xrightarrow{f_1} V_P \xrightarrow{f_2} V_2 \longrightarrow 0, $$

is $H^1(X, V_1 \otimes V_2^{-1})$. Tensoring the exact sequence with $\Omega$ gives

$$ 0 \longrightarrow V_1 \otimes \Omega \xrightarrow{g_1} V_P \otimes \Omega \xrightarrow{g_2} V_2 \otimes \Omega \longrightarrow 0. $$

Since deg($V_1$) < deg($V_2$), by the theorem of Lange and Narasimhan \[ 13 \], stable extensions always exist. Fix a stable extension $V_P$ and set

$$ \Phi_1 = \psi_1 \circ f_2, $$

$$ \Phi_2 = g_1 \circ \psi_2. $$

Note $(V_P \oplus V_Q, 0) \in B$. The connected family

$$ FC = H^0(X, V_2^{-1} \otimes V_Q \otimes \Omega) \times H^0(X, V_Q^{-1} \otimes V_1 \otimes \Omega) $$

is connected.
of Higgs bundles contains $(V_P \oplus V_Q, 0)$ and $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$. Note the family $FC$ contains semi-stable Higgs bundles. This is allowed since the points in the moduli space $M_{\text{Dol}}$ are also interpreted as $S$-equivalence classes of semi-stable Higgs bundles. However one may choose $FC$ to be a strictly poly-stable family:

$$FC = (H^0(X, V_{-1}^2 \otimes V_Q \otimes \Omega) \times H^0(X, V_{-1}^1 \otimes V_1 \otimes \Omega)) \setminus \{(\emptyset) \times H^0(X, V_{-1}^1 \otimes V_1 \otimes \Omega) \cup (H^0(X, V_{-1}^1 \otimes V_Q \otimes \Omega) \times \emptyset)\}.$$ 

Since $V_P$ is stable, by Lemma 7.3, any element in $FC$ is semi-stable. Hence the family $FC$ provides a deformation between $(V_P \oplus V_Q, 0)$ and $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$. The cohomology $H^1(X, V_1 \otimes V_2^{-1})$ parameterizes bundle extensions of $V_2$ by $V_1$ and also parameterizes a family of Higgs bundles with fixed $\psi_1, \psi_2$. By Lemma 7.4, this is a stable family which contains $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$ and $(V_1 \oplus V_2 \oplus V_Q, (\psi_1 \circ f_4, g_3 \circ \psi_2))$ where $f_3, f_4, g_3, g_4$ come from the trivial extensions

$$0 \longrightarrow V_1 \xrightarrow{f_3} V_1 \oplus V_2 \xrightarrow{f_4} V_2 \longrightarrow 0,$$

$$0 \longrightarrow V_1 \otimes \Omega \xrightarrow{g_3} (V_1 \oplus V_2) \otimes \Omega \xrightarrow{g_4} V_2 \otimes \Omega \longrightarrow 0.$$

Hence $H^1(X, V_1 \otimes V_2^{-1})$ provides a deformation between $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$ and $(V_1 \oplus V_2 \oplus V_Q, (\psi_1 \circ f_4, g_3 \circ \psi_2)) \in T(-1, 1, 0)$.

To summarize, one first shows that the space $B$ is connected. Then choose a specific element $(V_P \oplus V_Q, 0) \in B$ with $V_P$ a stable extension of $V_2$ by $V_1$ and that there exists non-trivial holomorphic maps

$$\psi_1 : V_2 \longrightarrow V_Q \otimes \Omega$$

$$\psi_2 : V_Q \longrightarrow V_1 \otimes \Omega.$$

This provides a deformation from $(V_P \oplus V_Q, 0)$ to $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$. Finally, since $V_P$ is an extension of $V_2$ by $V_1$, $(V_P \oplus V_Q, (\Phi_1, \Phi_2))$ is deformed to $(V_1 \oplus V_2 \oplus V_Q, (\psi_1 \circ f_4, g_3 \circ \psi_2))$ in $H^1(X, V_1 \otimes V_2^{-1})$.

**Corollary 7.6.** Every Binary Hodge bundle can be deformed to a ternary Hodge bundle.

**Proof.** Every poly-stable reducible Hodge bundle is either ternary or in $B$. The result then follows from Proposition 7.2 and 7.5.  

**Lemma 7.7.** For fixed integers $d_1, d_2, d_3$, $T(d_1, d_2, d_3)$ is connected.

**Proof.** We first consider the stable bundles. Stability implies the Higgs fields $\Phi_1, \Phi_2$ are not identically zero. Denote by $J_d(X)$ the Jacobi variety identified with the set of holomorphic line bundles of degree $d$. For each $L_1 \in J_{d_1}(X)$, the set of all $(L_3, \Phi_2)$ such that $L_3 \in J_{d_3}(X)$ and

$$0 \neq \Phi_2 \in H^0(X, L_3^{-1} \otimes L_1 \otimes \Omega)$$


is $\mathbb{C}^* \times \text{Sym}^{d_1+2(g-1)-d_3}X$, where $\text{Sym}^dX$ is the $d$-th symmetric product of $X$. Hence the set of all triples $(L_3, L_1, \Phi_2)$ such that

$$L_3 \xrightarrow{\Phi_2} L_1 \otimes \Omega$$

with $\Phi_2 \neq 0$ is the space $(\mathbb{C}^* \times \text{Sym}^{d_1+2(g-1)-d_3}X) \times J_{d_1}(X)$.

Similarly, for each $L_3 \in J_{d_3}(X)$, the space of all triples $(L_2, L_3, \Phi_1)$ such that

$$L_2 \xrightarrow{\Phi_1} L_3 \otimes \Omega$$

with $\Phi_1 \neq 0$ is $\mathbb{C}^* \times \text{Sym}^{d_3+2(g-1)-d_2}X$. The set of Higgs bundles parameterized by the total space

$$S = (\mathbb{C}^* \times \text{Sym}^{d_3+2(g-1)-d_2}X) \times (\mathbb{C}^* \times \text{Sym}^{d_1+2(g-1)-d_3}X) \times J_{d_2}(X)$$

contains every Higgs bundle in $T(d_1, d_2, d_3)$. Hence the natural morphism

$$t : S \longrightarrow T(d_1, d_2, d_3)$$

is surjective. Since $S$ is connected, $T(d_1, d_2, d_3)$ is connected.

The reducible bundles consist of $T(0, 0, 0)$ and $T(0, d_2, -d_2)$. All polystable Higgs bundles associated with the points in $T(0, 0, 0)$ and $T(0, d_2, -d_2)$ are contained in the families parameterized by

$$S_1 = J_0(X) \times J_0(X) \times J_0(X)$$

and

$$S_2 = (\mathbb{C}^* \times \text{Sym}^{2(g-1)-2d_2}X) \times J_{-d_2}(X) \times J_0(X),$$

respectively. Both $S_1, S_2$ are connected. Since the natural morphisms

$$t_1 : S_1 \longrightarrow T(0, 0, 0)$$

$$t_2 : S_2 \longrightarrow T(0, d_2, -d_2)$$

are surjective, both $T(0, 0, 0)$ and $T(0, d_2, -d_2)$ are connected. \qed

**Proposition 7.8.** Every component of $\mathcal{M}_{\text{Dol}}^\tau$ contains a Hodge bundle.

**Proof.** By Corollary 4.8, every component of $\mathcal{M}_{\text{Dol}}^\tau$ contains a local minimum $(V, \Phi)$ of $m_r$. If $(V, \Phi)$ is a smooth point, then $(V, \Phi)$ is a critical point of $m$. A theorem of Hitchin and Simpson implies that $(V, \Phi)$ is a Hodge bundle [12, 22]. Singular points of $\mathcal{M}_{\text{Dol}}^\tau$ correspond to reducible Higgs bundles. The space of all reducible Higgs bundles correspond to either the space of $U(2) \times U(1)$ representations or the space of $U(1) \times U(1, 1)$ representations. Each component of $U(2) \times U(1)$ and $U(1) \times U(1, 1)$ representations contains points that correspond to Hodge bundles [11]. In fact, these points are exactly the ones corresponding to the points in $B$ and $T(0, d_2, -d_2)$. \qed

Let $K$ be a divisor of $\Omega$ and let

$$w : X \longrightarrow |K| \cong \mathbb{CP}^{g-1}$$

be the canonical map [10].
Lemma 7.9. \( \Omega \) has a section with simple zeros.

Proof. The linear system \(|K|\) is base point free \([10]\). If \( X \) is hyperelliptic, then the map \( w \) is a 2-1 branch map into \( \mathbb{CP}^{g-1} \) and an embedding otherwise. In both cases, by Bertini’s theorem, there exists a hyperplane \( H \in \mathbb{CP}^{g-1} \) such that \( H \cap X \) is regular. Then \( w^{-1}(H) \) is an effective divisor equivalent to \( K \) and with simple zeros. \( \square \)

Choose 
\[ K = \{ x_1, x_2, \ldots, x_{2(g-1)} \}, \]
such that the \( x_i \)'s are all distinct.

Proposition 7.10. Let \( 0 \leq \tau < 2(g-1) \). Suppose 
\[ T(d_1 - 1, d_2 + 1, d_Q), T(d_1, d_2, d_Q) \subset M^\tau_{\text{Dol}}. \]
Then there is deformation between \( T(d_1, d_2, d_Q) \) and \( T(d_1 - 1, d_2 + 1, d_Q) \) within \( M^\tau_{\text{Dol}}. \)

Proof. Suppose 
\[ (V_1 \oplus V_2 \oplus V_Q, (\Phi_1, \Phi_2)) \in T(d_1 - 1, d_2 + 1, d_Q), \]
\[ (U_1 \oplus U_2 \oplus U_Q, (\Psi_1, \Psi_2)) \in T(d_1, d_2, d_Q). \]
By the semi-stability of \( (U_1 \oplus U_2 \oplus U_Q, (\Psi_1, \Psi_2)) \) and the assumptions \( \tau \geq 0, 0 \leq c < 3 \), one has \( d_Q \leq 0 \) and
\[ d_1 - 1 < d_1 \leq \frac{d_P + d_Q}{3} < 1; \]
hence,
\[ d_1 - 1 < d_1 \leq 0 \text{ and } d_2 + 1 > 0. \]
This implies \( (V_1 \oplus V_2 \oplus V_Q, (\Phi_1, \Phi_2)) \) is stable. Hence \( \Phi_1 \neq 0 \) and
\[ -\deg(V_2) + d_Q + 2(g-1) \geq 0. \]
On the other hand, \( \deg(V_1) + \deg(V_2) = d_P \), so
\[ d_P - \deg(V_1) - d_Q \leq 2(g-1), \]
\[ -d_1 < 1 - d_1 = -\deg(V_1) \leq -d_P + d_Q + 2(g-1) \leq 2(g-1). \]
In light of Lemma 7.7, it suffices to demonstrate the existence of \( (U_1 \oplus U_2 \oplus U_Q, (\Psi_1, \Psi_2)) \in T(d_1, d_2, d_Q) \) and \( (V_1 \oplus V_2 \oplus U_Q, (\Phi_1, \Phi_2)) \in T(d_1 - 1, d_2 + 1, d_Q) \) and a deformation between the two.
Since \(|K|\) is base point free, there exists \( K' \in |K| \) such that
\[ K' = \{ y_1, y_2, \ldots, y_{2(g-1)} \} \]
with $y_i \neq x_{2(g-1)}$ for all $1 \leq i \leq 2g$. The bounds on the degrees of the various bundles allow us to construct the following divisors:

\[
\begin{align*}
D_1 &= \{-x_1, \ldots, -x_{-\deg(U_1)}\} \\
D_2 &= \{y_1, \ldots, y_{d_P-\deg(V_1)}, -x_{2(g-1)}\} \\
D_Q &= \{-y_{d_P-\deg(V_1)+1}, \ldots, -y_{d_P-\deg(V_1)-d_Q}\}.
\end{align*}
\]

Let $u$ be the basic epimorphism [1]

\[u : \text{Div}(X) \rightarrow H^1(X, \mathcal{O}^*)\]

and set

\[
\begin{align*}
U_1 &= u(D_1) \\
U_2 &= u(D_2) \\
U_Q &= u(D_Q) \\
U_P &= U_1 \oplus U_2.
\end{align*}
\]

Let $\psi_1, \psi_2$ be meromorphic sections associated with the divisors $D_1, D_2$. Then the meromorphic section $\psi_1 \oplus \psi_2$ of $U_P$ is associated with the divisor

\[D'_1 = \{-x_1, \ldots, -x_{-\deg(U_1)}, -x_{2(g-1)}\}.
\]

Hence there exists $V_1 \subset U_P$ [9] such that

\[V_1 = u(D'_1).
\]

Let

\[V_2 = U_P/V_1.
\]

Since

\[V_1 \otimes V_2 = \text{det}(U_P) = U_1 \otimes U_2,
\]

\[V_2 = u(D'_2),
\]

where

\[D'_2 = \{y_1, \ldots, y_{d_P-\deg(V_1)}\}.
\]

In short, the bundle $U_P$ is constructed in such a way that it is the trivial extension of $U_2$ by $U_1$, and is also an extension of $V_2$ by $V_1$:

\[0 \rightarrow U_1 \xrightarrow{f_1} U_P \xrightarrow{f_2} U_2 \rightarrow 0
\]

\[0 \rightarrow V_1 \xrightarrow{f_3} U_P \xrightarrow{f_4} V_2 \rightarrow 0.
\]

Tensoring with $\Omega$ gives

\[0 \rightarrow U_1 \otimes \Omega \xrightarrow{g_1} U_P \otimes \Omega \xrightarrow{g_2} U_2 \otimes \Omega \rightarrow 0
\]

\[0 \rightarrow V_1 \otimes \Omega \xrightarrow{g_3} U_P \otimes \Omega \xrightarrow{g_4} V_2 \otimes \Omega \rightarrow 0.
\]
Since
\[
\begin{align*}
-D_2 + D_Q + K' &= \left\{x_2(g-1), y_{d_P - \deg(V_1) - d_Q + 1}, \ldots, y_{2(g-1)} \right\} \\
-D_Q + D_1 + K &= \left\{y_{d_P - \deg(V_1) + 1}, \ldots, y_{d_P - \deg(V_1) - d_Q}, \right. \\
&\left. x_{-\deg(U_1) + 1}, \ldots, x_{2(g-1)} \right\}
\end{align*}
\]
are effective divisors, there exists
\[
0 \not\equiv \psi_1 \in H^0(X, U_2^{-1} \otimes U_Q \otimes \Omega) \\
0 \not\equiv \psi_2 \in H^0(X, U_Q^{-1} \otimes U_1 \otimes \Omega).
\]
Set
\[
\Psi_1 = \psi_1 \circ f_2 \quad \text{and} \quad \Psi_2 = g_1 \circ \psi_2.
\]
Then \((U_1 \oplus U_2 \oplus U_Q, (\Psi_1, \Psi_2))\) is a semi-stable ternary Hodge bundle.

The divisors
\[
\begin{align*}
-D'_2 + D_Q + K' &= \left\{y_{d_P - \deg(V_1) - d_Q + 1}, \ldots, y_{2(g-1)} \right\} \\
-D_Q + D'_1 + K &= \left\{x_{-\deg(U_1) + 1}, \ldots, x_{2(g-1) - 1}, \\
&\left. y_{d_P - \deg(V_1) + 1}, \ldots, y_{d_P - \deg(V_1) - d_Q} \right\}
\end{align*}
\]
are effective. Hence there exist
\[
0 \not\equiv \phi_1 \in H^0(X, V_2^{-1} \otimes U_Q \otimes \Omega) \\
0 \not\equiv \phi_2 \in H^0(X, U_Q^{-1} \otimes V_1 \otimes \Omega).
\]

**Remark 2.** This is the critical step where the assumption \(\tau < 2(g - 1)\) is needed. In the case of \(\tau = 2(g - 1)\), the degree of \(V_2^{-1} \otimes U_Q \otimes \Omega\) equals \(-1\) thus rendering it impossible to find a non-zero global section \(\phi_1\). This reflects the fact that every representation with \(\tau = 2(g - 1)\) is reducible. (See Section 6.)

Set
\[
\Psi'_1 = \phi_1 \circ f_4 \quad \text{and} \quad \Psi'_2 = g_3 \circ \phi_2.
\]
Then \((U_P \oplus U_Q, (\Psi'_1, \Psi'_2))\) is a semi-stable Higgs bundle. Since
\[
\begin{align*}
&h^0(X, U_2^{-1} \otimes U_Q \otimes \Omega) > 0 \\
&h^0(X, U_Q^{-1} \otimes U_1 \otimes \Omega) > 0,
\end{align*}
\]
\(H^0(X, U_1^{-1} \otimes U_Q \otimes \Omega)\) and \(H^0(X, U_1^{-1} \otimes U_2 \otimes \Omega)\) are proper subspaces of \(H^0(X, U_P^{-1} \otimes U_Q \otimes \Omega)\) and \(H^0(X, U_Q^{-1} \otimes U_P \otimes \Omega)\), respectively. Hence
\[
FC = \left( H^0(X, U_2^{-1} \otimes U_Q \otimes \Omega) \setminus H^0(X, U_1^{-1} \otimes U_Q \otimes \Omega) \right) \times \\
\left( H^0(X, U_Q^{-1} \otimes U_P \otimes \Omega) \setminus H^0(X, U_2^{-1} \otimes U_Q \otimes \Omega) \right)
\]
is connected and parameterizes a family of semi-stable Higgs bundles that contains both \((U_P \oplus U_Q, (\Psi_1, \Psi_2))\) and \((U_P \oplus U_Q, (\Psi'_1, \Psi'_2))\). Hence there is deformation between the two.

The space of bundle extensions of \(V_2\) by \(V_1\),

\[
0 \to V_1 \overset{f_5}{\to} V \overset{f_6}{\to} V_2 \to 0,
\]

is parameterized by the vector space \(H^1(V_1 \otimes V_2^{-1})\) containing both \(U_P\) and \(V_1 \oplus V_2\) (the zero element in \(H^1(V_1 \otimes V_2^{-1})\)). Again tensoring with \(\Omega\) gives

\[
0 \to V_1 \otimes \Omega \overset{g_5}{\to} V \otimes \Omega \overset{g_6}{\to} V_2 \otimes \Omega \to 0.
\]

Let

\[
\Phi_1 = \phi_1 \circ f'_6 \quad \text{and} \quad \Phi_2 = g'_6 \circ \phi_2,
\]

where

\[
0 \to V_1 \overset{f'_5}{\to} V_1 \oplus V_2 \overset{f'_6}{\to} V_2 \to 0
\]

\[
0 \to V_1 \otimes \Omega \overset{g'_5}{\to} (V_1 \oplus V_2) \otimes \Omega \overset{g'_6}{\to} V_2 \otimes \Omega \to 0
\]

correspond to the trivial extensions. By Lemma 7.4, \(H^1(V_1 \otimes V_2^{-1})\) parameterizes a family of semi-stable Higgs bundles that contains both \((U_P \oplus U_Q, (\Psi'_1, \Psi'_2))\) and \((V_1 \oplus V_2 \oplus U_Q, (\Phi_1, \Phi_2))\).

To summarize, the first step consists of fixing \(U_P = U_1 \oplus U_2\) and deform the Higgs field \((\Psi_1, \Psi_2)\) to \((\Psi'_1, \Psi'_2)\). In the second step, fix \(\phi_1, \phi_2\) and deform \(U_P\) to \(V_1 \oplus V_2\).

\[\square\]

Consider the space \(T(0, d_2, -d_2)\). By Proposition 7.5, one may assume \(d_2 > 0\). To deform points in \(T(0, d_2, -d_2)\), the family \(FC\) constructed in the above proof contains semi-stable Higgs bundles. However, one may also opt to construct the deformation family of poly-stable Higgs bundles by setting:

\[
FC = \left(\left(\left(H^0(X, U_P^{-1} \otimes U_Q \otimes \Omega) \right) \setminus \left(\left(H^0(X, U_1^{-1} \otimes U_Q \otimes \Omega) \cup H^0(X, U_2^{-1} \otimes U_Q \otimes \Omega)\right) \times \right) \left(H^0(X, U_Q^{-1} \otimes U_P \otimes \Omega) \setminus \left(H^0(X, U_Q^{-1} \otimes U_2 \otimes \Omega) \cup H^0(X, U_1^{-1} \otimes U_1 \otimes \Omega)\right) \right) \cup \left(\left(H^0(X, U_Q^{-1} \otimes U_Q \otimes \Omega \times \{0\}\right) \right)\right).
\]

The case with \(\tau = 2(g - 1)\) has been covered in Section 6 and \(\mathcal{M}_{\text{Dol}}^{2(g-1)}\) is connected. Suppose \(\tau < 2(g - 1)\). By Proposition 7.8, every component of \(\mathcal{M}_{\text{Dol}}^{\tau}\) contains a Hodge bundle. By Corollary 7.6, every component of \(\mathcal{M}_{\text{Dol}}^{\tau}\) contains a ternary Hodge bundle. It follows from Proposition 7.10 and induction that \(\mathcal{M}_{\text{Dol}}^{\tau}\) is connected. Since

\[
\mathcal{N}_{\text{Dol}}^{\tau} = \mathcal{M}_{\text{Dol}}^{\tau}/H^1(X, \mathcal{O}^*)
\]

Theorem 1.1 then follows from Corollary 4.2.
References

[24] ______, Moduli of representations of the fundamental group of a smooth projective
[26] D. Toledo, Harmonic maps from surfaces to certain Kähler manifolds, Math. Scand.,
[27] ______, Representations of surface groups in complex hyperbolic space, J. Differential

Received September 2, 1998 and revised March 16, 1999.

University of Massachusetts
Amherst, MA 01003-4515
E-mail address: xia@math.umass.edu
Guidelines for Authors

Authors may submit manuscripts at pjm.math.berkeley.edu/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use L\LaTeX, but papers in other varieties of \LaTeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as L\LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of Bib\LaTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text (“the curve looks like this:”). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as “Place Figure 1 here”. The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal’s preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.
Motion of hypersurfaces by Gauss curvature

BEN ANDREWS

Summation of formal solutions of a class of linear difference equations

B.L.J. BRAAKSMA, B.F. FABER AND G.K. IMMINK

Approximation of recurrence in negatively curved metric spaces

CHARALAMBOS CHARITOS AND GEORGIOS TSAPOGAS

Cohomology of singular hypersurfaces

BERNARD M. DWORK

Riemannian manifolds admitting isometric immersions by their first eigenfunctions

AHMAD EL SOUF AND SAID ILIAS

The Rubinstein–Scharlemann graphic of a 3-manifold as the discriminant set of a stable map

TSUYOSHI KOBAYASHI AND OSAMU SAEKI

Geometric realizations of Fordy–Kulish nonlinear Schrödinger systems

JOEL LANGER AND RON PERLINE

Seiberg–Witten invariants for 3-manifolds in the case $b_1 = 0$ or 1

YUHAN LIM

Virtual homology of surgered torus bundles

JOSEPH D. MASTERS

Group actions on polynomial and power series rings

PETER SYMONDS

The moduli of flat $PU(2,1)$ structures on Riemann surfaces

EUGENE Z. XIA