MONOTONICITY AND SYMMETRY RESULTS FOR
DEGENERATE ELLIPTIC EQUATIONS ON NILPOTENT
LIE GROUPS

I. Birindelli and J. Prajapat

In this paper we prove some monotonicity results for solutions of semilinear equations in nilpotent, stratified groups. We also prove a partial symmetry result for solutions of nonlinear equations on the Heisenberg group.

1. Introduction.

Berestycki and Nirenberg (see e.g., [2]) introduced the so called “sliding method” to prove monotonicity results for semilinear elliptic equations in convex domains of $\mathbb{R}^n$. The idea here is to implement the method in the general setting of nilpotent stratified groups. Let us mention that examples of such groups include the Heisenberg group and, of course, the Euclidean space. Hence, in particular, we obtain monotonicity results for a large class of degenerate elliptic semilinear equations.

More precisely, let $(G, \circ)$ be a nilpotent, stratified Lie group, see Section 2 for definitions and properties. Clearly the notion of “convexity” has to be related to the group action:

**Definition 1.1.** Fix $\eta \in G$. A domain $\Omega \subset G$ is said to be $\eta$-convex (or convex in the direction $\eta$) if for any $\xi_1 \in \Omega$ and any $\xi_2 \in \Omega$ such that $\xi_2 = s\eta \circ \xi_1$ for some $s > 0$, we have $s\eta \circ \xi_1 \in \Omega$ for every $s \in (0, \alpha)$.

Observe that this coincides with the notion of convexity in a given direction for domains in the Euclidean space. Any Koranyi ball in the Heisenberg group $H^n = (\mathbb{R}^{2n+1}, \circ)$ is an example of a domain which is $\eta$-convex for any $\eta \in H^n$.

At the end of the paper we show a “cube” in the Heisenberg group $H^1$, which is obtained by sliding a square of the plane $x_1 = 0$ through the group action in the direction of $(0, 1, 0)$. In the figure, we have shaded the top and bottom surfaces in order to make the cube more visible. Observe that this set is convex in both the directions $e_1 = (1, 0, 0)$ and $e_2 = (0, 1, 0)$.

Let

$$\Delta_G = \sum_{i=1}^{m} X_i^2$$
denote the sub-Laplacian operator defined on $G$ and let $S^Q_2$ denote the Sobolev space for the group $G$ where $Q$ is the homogeneous dimension of $G$; see details in Section 2. For $\eta \in G$ and $u \in S^Q_2$, let $T_\eta u(\xi) := u(\eta \circ \xi)$.

Our main result is the following:

**Theorem 1.1.** Let $(G, \circ)$ be a stratified, nilpotent Lie group and $\Omega$ be an arbitrary bounded domain of $G$ which is $\eta$-convex for some $\eta \in G$. Let $u \in S^Q_2(\Omega) \cap C(\bar{\Omega})$ be a solution of

$$
\begin{cases}
\Delta_G u + f(u) = 0 & \text{in } \Omega \\
u = \phi & \text{on } \partial \Omega
\end{cases}
$$

where $f$ is a Lipschitz continuous function. Assume that for any $\xi_1, \xi_2 \in \partial \Omega$, such that $\xi_2 = \alpha \eta \circ \xi_1$ for some $\alpha > 0$, we have for each $s \in (0, \alpha)$

$$
\phi(\xi_1) < T_{s\eta} u(\xi_1) < \phi(\xi_2) \text{ if } s \eta \circ \xi_1 \in \Omega
$$

and

$$
\phi(\xi_1) < T_{s\eta} \phi(\xi_1) < \phi(\xi_2) \text{ if } s \eta \circ \xi_1 \in \partial \Omega.
$$

Then $u$ satisfies

$$
T_{s_1\eta} u(\xi) < T_{s\eta} u(\xi)
$$

for any $0 < s_1 < s < \alpha$ and for every $\xi \in \Omega$.

Moreover, $u$ is the unique solution of (1.1) in $S^Q_2(\Omega) \cap C(\bar{\Omega})$ satisfying (1.2).

**Remark.** Clearly, (1.4) implies that $u$ is monotone along $\gamma(s) = s \eta \circ \xi$. Observe that the curve $\gamma$ is the integral curve of a right invariant vector field $R_\eta$, even though the operator $\Delta_G$ is left invariant.

An immediate consequence of Theorem 1.1 is:

**Corollary 1.2.** Under the assumptions of Theorem 1.1, if $f$ is $C^1$ and $R_\eta$ commutes with $\Delta_G$ then

$$
R_\eta u > 0 \text{ in } \Omega.
$$

In [1], L. Almeida and Y. Ge have proved monotonicity results in the general setting of manifolds. However, they consider solutions of uniformly elliptic semilinear equations.

An important tool in the proof of Theorem 1.1 is the “Maximum principle in domains with small measure” which is new in the setting of degenerate elliptic equations. On the other hand, it is known for uniformly elliptic and parabolic operators (see [1], [2], [6]) and it has found extensive applications, see for e.g., [3] and [15].
Using the notations of [14], on a bounded domain $\Omega \subset \mathbb{R}^N$, consider the operator
\begin{equation}
Lu(x) = \frac{1}{2} \sigma^{ik}(x)(\sigma^{jk}(x)u_{x^j}(x))_{x^i} + b^i(x)u_{x^i}(x)
\end{equation}
where $\sigma^k = (\sigma^{ik})$, $k = 1, \ldots, n_1$, and $b = (b^i)$ are smooth vector fields given on $\mathbb{R}^N$ and $n_1$ is an integer. We assume that the Lie algebra generated by the family of vector fields $\{b, \sigma^k, k = 1, \ldots, n_1\}$ has dimension $N$ at all points in the closure $\overline{D}$ of a neighborhood $D$ of $\overline{\Omega}$. Equivalently, $L$ is an operator satisfying Hörmander condition.

Similarly to [2], we say that the maximum principle holds for the operator $L + c$ where $c$ is an $L^\infty$ function in $\Omega$ if for $u \in S_2^Q(\Omega)$
\begin{equation}
Lu + c(\xi)u \geq 0 \text{ in } \Omega
\end{equation}
and
\begin{equation}
\lim_{\xi \to \partial \Omega} u(\xi) \leq 0
\end{equation}
implies that $u \leq 0$ in $\Omega$. Note that by embedding theorems (see e.g., [17]), $u \in S_2^Q(\Omega)$ implies that $u$ is continuous in $\Omega$.

The following proposition is the maximum principle for “domains with small measure” of $\mathbb{R}^N$ for the operators $L$:

**Proposition 1.3 (Maximum Principle).** Let $\Omega$ be a bounded domain in $\mathbb{R}^N$ and $L$ be an operator as defined above and $c$ such that $c(\xi) \leq b$ in $\Omega$ for some $b \in \mathbb{R}^+$. There exists $\delta > 0$, depending only on $N$ and $b$, such that the maximum principle holds for $L + c$ in $\Omega$ provided
\begin{equation}
\text{meas } (\Omega) < \delta.
\end{equation}

A weak comparison principle was derived in [1] using a Poincaré type inequality. An anonymous referee raised the question of whether we could similarly use a Poincaré type inequality to give an alternative proof of Proposition 1.3. In the last section, we derive a Poincaré type inequality for subelliptic operators and as a consequence of this inequality, we give an alternative proof of Proposition 1.3. We thank the referee for pointing this out.

Note that the class of operators $L$ defined in (1.6) and the sub-Laplacian $\Delta_G$ associated with a nilpotent Lie group $G$ are examples of subelliptic operators (see (6.1) for the definition of a subelliptic operator). Since it is possible to associate a group structure with the operator $L$ in (1.6) (see [19]), the monotonicity result Theorem 1.1 is in fact true for a more general nilpotent Lie group. We have given the result here for nilpotent, stratified Lie group to avoid technical details. However, it may not be possible to associate a general subelliptic operator with a group structure.
In Section 2 we state the basic definitions concerning nilpotent stratified Lie groups in general and the Heisenberg group in particular, in Section 3 we prove Proposition 1.3 and in Section 4 we prove Theorem 1.1.

Section 5 is a different application of the maximum principle in domains with small measure i.e., Proposition 1.3. We prove a symmetry result for positive “cylindrical” solutions of semilinear equations in a class of bounded symmetric domains in the Heisenberg group under some conditions. The generalization of Gidas, Ni, Nirenberg result (see [10]), to the Heisenberg Laplacian is a difficult open problem. Theorem 5.1 is a step towards the solution of this problem.

Finally in Section 6 we prove a Poincaré type inequality as mentioned above.

2. Preliminaries.

In this section we recall the basic notions of nilpotent, stratified Lie groups from [19]. Let \((G, [\cdot, \cdot])\) be a real finite dimensional Lie algebra, \(G_1 = G\) and \(G_i = [G, G_{i-1}]\) for \(i \geq 2\). Then \(\{G_i\}_{i \geq 2}\) is a decreasing sequence of Lie sub-algebras of \(G\). The Lie algebra \(G\) is said to be nilpotent of rank \(r\) if \(G_{r+1} = 0\). A Lie group \(G\) is said to be nilpotent of rank \(r\) if its Lie algebra is nilpotent of rank \(r\).

A stratified group \(G\) is a simply connected nilpotent group whose Lie algebra admits a direct sum decomposition (as vector space)

\[ G = V_1 \oplus \ldots \oplus V_m \]

with \(\dim V_j = n_j\), \([V_1, V_j] = V_{j+1}\) for \(1 \leq j < m\) and \([V_1, V_m] = 0\). Thus \(V_1\) generates \(G\) as a Lie algebra.

More precisely, given a Lie algebra \((G, [\cdot, \cdot])\) satisfying the above conditions, consider \(\mathbb{R}^N\) where \(N = \sum_{j=1}^m n_j\) with the group operation \(\circ\) given by the Campbell-Hausdorff formula

\[ \eta \circ \xi = \eta + \xi + \frac{1}{2} [\eta, \xi] + \frac{1}{12} [\eta, [\eta, \xi]] + \frac{1}{12} [\xi, [\xi, \eta]] + \ldots. \]

(2.1)

Note that since \(G\) is nilpotent there are only a finite number of nonzero terms in the above sum; precisely those involving commutators of \(\xi\) and \(\eta\) of length less than \(m\). Then \((G, \circ) = (\mathbb{R}^N, \circ)\) is the nilpotent, stratified group whose Lie algebra of left-invariant vector fields coincides with the Lie algebra \((G, [\cdot, \cdot])\).

Consider the standard basis \(e_1, \ldots, e_{n_1}\) of the subspace \(\mathbb{R}^{n_1}\) of \(G\). Let \(X_1, \ldots, X_{n_1}\) denote the corresponding “coordinate vector fields”, i.e.,

\[ X_i(f)(\xi) = \lim_{t \to 0} \frac{f(\xi \circ te_i) - f(\xi)}{t}, \]

for any smooth function \(f\) defined on \(G\) and for \(i = 1, \ldots, n_1\). The family \(\{X_1, \ldots, X_{n_1}\}\) forms a basis for \(V_1\). We define the sub-Laplacian operator
on $G$ as

\begin{equation}
\Delta_G = \sum_{i=1}^{n_1} X_i^2.
\end{equation}

We observe that this operator is subelliptic and satisfies Hörmander’s condition. Hence the Bony’s maximum principle holds (see [5]). Furthermore, the vector fields are invariant with respect to the group action, viz,

\[ X_i \circ T_\eta = T_\eta \circ X_i \]

and clearly so is the operator $\Delta_G$. In fact, this is a fundamental property of the operator which we shall use to prove Theorem 1.1.

Since the vector fields $\{X_1, \ldots, X_{n_1}\}$ generate $\mathcal{G}$ as Lie algebra, we can define recursively for $j = 1, \ldots, m$, and $i = 1, \ldots, n_j$, a basis $\{X_{i,j}\}$ of $V_j$ as

\[ X_{i,1} = X_i (i = 1, \ldots, n_1) \]

\[ X_\alpha = [X_{i_1}, [X_{i_2}, \ldots, [X_{i_{j-1}}, X_{i_j}]] \ldots] \]

with $\alpha = (i_1, \ldots, i_j)$ multi-index of length $j$ and $X_{i_k} \in \{X_1, \ldots, X_{n_1}\}$. With the decomposition $\mathcal{G} = \mathbb{R}^{n_1} \oplus \ldots \oplus \mathbb{R}^{n_m}$, we define a parameter group of dilations $\delta_\lambda$ by setting for

\begin{equation}
\xi = \xi_1 + \ldots + \xi_m, \quad (\xi_i \in \mathbb{R}^{n_i})
\end{equation}

\[ \delta_\lambda(\xi) = \sum_{i=1}^{m} \lambda^i \xi_i. \]

For any $\xi \in G$, the Jacobian of the map $\xi \mapsto \delta_\lambda(\xi)$ is $\lambda^Q$ where

\begin{equation}
Q = \sum_{i=1}^{m} in_i.
\end{equation}

The integer $Q$ is called the homogeneous dimension of $G$. Note that the euclidean dimension of $G$ is $N = \sum_{i=1}^{m} n_i$. We have $Q \geq N$ with equality in the trivial case $m = 1$ and $G = \mathbb{R}^{n_1}$.

Observe that since $G$ is simply connected, the exponential map $\exp : \mathcal{G} \rightarrow G$ is a diffeomorphism and the Lebesgue measure on $\mathcal{G}$, $dx = dx_1 \ldots dx_N$, pulled back to $G$ by the map $\exp^{-1}$, is left and right invariant with respect to the group action.

We recall that the equivalent of the Sobolev spaces, as introduced by Folland and Stein [8, 9], are

\[ S^q_2(\Omega) = \{ f \in L^q(\Omega) \text{ such that } X^I f \in L^q(\Omega) \text{ for } |I| \leq 2 \}, \]
where $I := (\alpha_1, \ldots, \alpha_h)$ (\(\alpha_i \leq n_1\)) denotes a multi-index of length \(|I| = h\) and $X^I = X_{\alpha_1} \ldots X_{\alpha_h}$. The norm in $S^q_2$ is given by:

$$
\|u\|_{S^q_2}^q = \int_\Omega \left( \sum_{|I|=1}^2 |X^I u|^q + |u|^q \right) \, d\xi.
$$

A typical example of a nilpotent, stratified Lie group is the Heisenberg group $H^n = (\mathbb{R}^{2n+1}, \circ)$ endowed with the group action $\circ$ defined by

$$(2.5) \quad \xi_0 \circ \xi = \left( x + x_0, y + y_0, t + t_0 + 2 \sum_{i=1}^n (x_i y_0 - y_i x_0) \right).$$

Here we denote the elements of $H^n$ either by $(z, t) \in \mathbb{C}^n \times \mathbb{R}$ or $(x, y, t) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ where $z = x + iy$, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$.

The Lie algebra of $H^n$ decomposes as $\mathbb{R}^{2n} \oplus \mathbb{R}$. Hence $n_1 = 2n$, $n_2 = 1$ and the anisotropic norm which is homogeneous with respect to the dilation given in (2.3) is defined by

$$
|\xi|_H = \left( (x^2 + y^2)^2 + t^2 \right)^{\frac{1}{4}}.
$$

The so called Koranyi ball is the set: \{\(\xi \in H^n\) such that $|\xi|_H \leq \text{const}$\}.

The generating vector fields are defined by

$$
X_i = \frac{\partial}{\partial x_i} + 2y_i \frac{\partial}{\partial t}, \quad \text{for} \quad i = 1, \ldots, n,
$$

$$
X_{n+i} := Y_i = \frac{\partial}{\partial y_i} - 2x_i \frac{\partial}{\partial t}, \quad \text{for} \quad i = 1, \ldots, n.
$$

Furthermore, we have

$$
X_{(i,j+n)} := [X_i, Y_j] = -4\delta_{ij} T
$$

for $1 \leq i, j \leq n$ and $T := \frac{\partial}{\partial t}$. Also, observe that the homogeneous dimension of $H^n$ is $2n + 2$, which is strictly greater than its linear dimension.

3. **Maximum principle.**

The Proposition 1.3 is a consequence of the following theorem by Krylov:

**Theorem 3.1.** Let $L$ be an operator defined as in (1.6), on a smooth bounded domain $D \subset \mathbb{R}^N$. For a fixed $\epsilon \in (0, 1)$ and $f \in L_p(D)$ with any $p \in (1, \infty)$, let $u := Rf \in W^2_p(D)$ be the unique solution of the equation

$$(L + \epsilon \Delta)u - u = f$$

with zero boundary condition. Here $\Delta$ is the Laplace operator on $\mathbb{R}^N$. Then, there exists a (large) $p_0 \in (1, \infty)$ and a (small) $\alpha \in (0, 1)$ both independent
of $\epsilon$ and such that for any $p \geq p_0$, subdomain $D_1 \subset D \setminus \overline{D}$, and $f \in L_p(D)$ we have

$$\sup_D |Rf| \leq C \|f\|_{L_p(D)},$$

(3.1)

$$|Rf(x) - Rf(y)| \leq C |x - y|^\alpha \|f\|_{L_p(D)},$$

(3.2)

where the constants $C$ are independent of $x$, $y$, $f$, and $\epsilon$.

We refer to [14] for a beautiful proof of this result. Also, it follows from [13] that one does not need the condition that the domain $D$ is smooth in the above theorem.

**Proof of Proposition 1.3.** Define $u^+(x) = \max\{u(x), 0\}$. To prove the proposition, we need to show that $u^+ \equiv 0$. Let $\Omega^+ = \{x \in \Omega : u(x) > 0\}$. Then $u^+$ satisfies the equation

$$Lu^+(x) + c(x)u^+(x) \geq 0$$

(3.3)

for $x \in \Omega^+$ and

$$u^+ = 0$$

on the boundary $\partial \Omega^+$.

Now let $v$ be the solution of the equation

$$Lv - v = -u^+ - bu^+$$

(3.4)

on $\Omega^+$ with zero boundary condition. From Theorem 3.1 we have

$$\sup_{\Omega^+} v \leq C(b + 1)\|u^+\|_{L_p(\Omega^+)}.$$  

But from (3.3), (3.4) and $u^+ - v = 0$ on $\partial \Omega$, the maximum principle implies that $u^+ \leq v$ in $\Omega^+$. Hence it follows that

$$\sup_{\Omega} u^+ \leq C(b + 1)\|u^+\|_{L_p(\Omega^+)}.$$  

Estimating the r.h.s. we have

$$\sup_{\Omega} u^+ \leq C(b + 1)\text{meas}(\Omega)^{1/p} \sup_{\Omega} u^+.$$  

(3.5)

Hence, if we choose $\delta$ such that $C(b + 1)\delta < 1$, then $\text{meas}(\Omega) < \delta$ and (3.5) implies that $u^+ \equiv 0$ i.e., $u \leq 0$ in $\Omega$.

**4. Proof of Theorem 1.1.**

As in previous sections $(G, \circ) = (\mathbb{R}^N, \circ)$ is a nilpotent, stratified Lie group and $\Delta_G$ is the corresponding sub-Laplacian operator. Using the notations of Theorem 1.1, define $u_s(\xi) = T_{s\eta}u(\xi)$ for $s > 0$. The function $u_s$ is defined on the domain $\Omega_s = \{\xi \in G : s\eta \circ \xi \in \Omega\}$, obtained by “translation” of $\Omega$. 
Furthermore, since the sub-Laplacian is invariant under the group action it follows that \( u_s \) satisfies the equation
\[
\Delta_G u_s + f(u_s) = 0 \quad \text{in} \quad \Omega_s.
\]
Since the domain \( \Omega \) is bounded, there exists \( s_0 > 0 \) such that \( \Omega_{s_0} \cap \Omega = \emptyset \) and for \( s < s_0 \) near \( s_0 \), \( \Omega_s \cap \Omega \neq \emptyset \). And as we slide the domain \( \Omega_s \), i.e., we decrease \( s \) to zero, we get \( \Omega_0 = \Omega \). Now for \( s < s_0 \) consider the function \( w_s = u_s - u \) in \( D_s = \Omega_s \cap \Omega \). Clearly, to prove (1.4), we need to show that \( w_s > 0 \) for every \( 0 < s < s_0 \).

Observe that \( w_s \) satisfies the equation
\[
\Delta_G w_s + c_s(\xi) w_s = 0 \quad \text{in} \quad D_s,
\]
where \( c_s \) is a \( L^\infty \) function satisfying \( |c_s(\xi)| \leq C \) for \( \xi \in D_s \) for all \( s \). Furthermore, due to the assumptions (1.2) and (1.3) we have \( w_s \geq 0 \) on the boundary of \( D_s \).

Let \( \delta \) denote the constant appearing in the Proposition 1.3 corresponding to the operator \( \Delta_G \) defined on \( \Omega \). For \( s > s_1 \) and sufficiently near \( s_1 \), \( \text{meas}(D_s) < \delta \). Therefore, by Proposition 1.3 it follows that \( w_s \geq 0 \) in \( D_s \) for \( s < s_1 \). Moreover, (1.2), (1.3) and the strong maximum principle implies that \( w_s > 0 \) in \( D_s \).

Let \( \mu_1 = \min \{ \mu : w_s > 0 \quad \text{for every} \quad s > \mu \} \). We claim that \( \mu_1 = 0 \). Suppose, by contradiction that \( \mu_1 > 0 \). Then \( w_{\mu_1} \geq 0 \). Since \( \mu_1 > 0 \), again the strong maximum principle implies that \( w_{\mu_1} > 0 \) in \( D_{\mu_1} \).

Choose a compact set \( \Sigma \subset D_{\mu_1} \) such that \( \text{meas}(D_{\mu_1} \setminus \Sigma) < \delta/3 \), where \( \delta \) is as fixed above. Since \( \Sigma \) is compact, for \( s < \mu_1 \) with \( \mu_1 - s \) small, we have
\[
\text{meas}(D_{\mu_1} \setminus \Sigma) < \delta.
\]
Further, for \( 0 < \mu < \mu_1 \) and sufficiently close to \( \mu_1 \), we have
\[
\Sigma \subset D_\mu \quad \text{and} \quad \text{meas}(D_{\mu_1} \setminus \Sigma) < \delta.
\]
Fix a \( \mu < \mu_1 \) such that (4.1) and (4.2) hold for all \( s, \mu < s < \mu_1 \). Proposition 1.3 implies \( w_s \geq 0 \) in \( D_s \setminus \Sigma \) for \( \mu < s < \mu_1 \). This, together with (4.1) implies that \( w_s \geq 0 \) on \( D_s \) for all \( s, \mu < s < \mu_1 \). Since \( s > \mu > 0 \) and \( w_s \neq 0 \), we further conclude from the strong maximum principle that \( w_s > 0 \) in \( D_s \) for all \( s, \mu < s < \mu_1 \); which contradicts the definition of \( \mu_1 \). Hence \( \mu_1 = 0 \) and therefore (1.4) holds true.

**Uniqueness.** To prove the uniqueness, suppose \( u, v \in S^Q_\alpha(\Omega) \cap C(\overline{\Omega}) \) are two solutions of (1.1). Consider the function \( w_s(\xi) := v_s(\xi) - u(\xi) \) in \( \Omega_s \cap \Omega \).
where \( v_s(\xi) = v(s\eta \circ \xi) \). We can go through the above proof with this function to conclude that for every \( \xi \in \Omega \),

\[
(4.3) \quad u(\xi) < v(s\eta \circ \xi) \text{ for } s > 0.
\]

Similarly, considering the function \( \tilde{w}_s(\xi) := u_s(\xi) - v(\xi) \) in \( \Omega_s \cap \Omega \) we have for every \( \xi \in \Omega \) that

\[
(4.4) \quad v(\xi) < u(s\eta \circ \xi) \text{ for } s > 0.
\]

Letting \( s \to 0 \) in (4.3) and (4.4), it follows

\[
 u \equiv v \text{ in } \Omega.
\]

\[\square\]

The proof of the Corollary 1.2 is immediate. Observe that since \( f \) is \( C^1 \) and \( R_\eta \) commutes with \( \Delta_G \), then \( R_\eta u \) satisfies the equation

\[\Delta_G R_\eta u + f'(u)R_\eta(u) = 0\]

in \( \Omega \). Furthermore, Theorem 1.1 implies that \( R_\eta u \geq 0 \) in \( \Omega \). But \( R_\eta u \not\equiv 0 \) in \( \Omega \). Hence the maximum principle implies that \( R_\eta u > 0 \) in \( \Omega \).

\[\square\]

5. A symmetry result.

We begin by defining a special class of functions and domains in the Heisenberg group:

**Definition 5.1.** We say that a function \( u \) defined on \( H^n \) is **cylindrical** if there exists \( \xi_0 \in H^n \) such that \( v(\xi) := u(\xi_0 \circ \xi) \) is a function depending only on \( (r, t) \), where \( r = (x^2 + y^2)^{\frac{1}{2}} \). We say that a domain \( C \subset H^n \) is a **cylinder** if there exists a cylindrical function \( \Phi \) such that \( \xi \in C \iff \Phi(\xi) < 0 \).

Observe that a Koranyi ball is a cylinder. Also, the Euclidean ball \( \{(z, t) \in H^n : |z|^2 + t^2 \leq \text{constant}\} \) with center at the origin belongs to this class. However, a Euclidean ball centered at a point other than the origin need not be a cylinder in \( H^n \).

In this section we prove a symmetry result for positive, cylindrical solutions of semilinear equations defined on a “cylinder” (as defined above) in the Heisenberg group \( H^n \). The proof relies on the maximum principle in domains with small measure and the adaptation of the moving plane method to \( H^n \). This method was used for the first time in the setting of the Heisenberg group in [4].

In the rest of the section, without loss of generality we will assume that \( \xi_0 \) occurring in the definition (5.1) is 0.
Theorem 5.1. Let $C$ be a bounded cylinder in $H^n$ defined by a function $\Phi$. Let $u \in C^2(C) \cap C(\overline{C})$ be a positive, cylindrical solution of the equation

\begin{align}
\Delta_H u + f(u) &= 0 \quad \text{in } C \\
u &= 0 \quad \text{on } \partial C
\end{align}

where $f$ is a Lipschitz function. If $\Phi(r, t) = \Phi(r, -t)$ then $u(r, t) = u(r, -t)$ on $C$.

Proof. The proof relies on the adaptation of the moving plane method to $H^n$. Let $T_\lambda = \{ \xi \in H^n : t = \lambda \}$ denote the hyperplane orthogonal to the $t$-direction and let $R_\lambda(x, y, t) = (y, x, 2\lambda - t)$ denote the $H$-reflection (see [4]).

We shift the plane from infinity towards the domain, i.e., we decrease $\lambda$ until it reaches the value $\lambda_0$ such that the plane $T_{\lambda_0}$ "touches" the boundary $\partial C$.

For $\lambda < \lambda_0$, let $D_\lambda = \{ (x, y, t) \in C : t > \lambda \}$ be the subset of $C$ cut off by the plane $T_\lambda$. Define $u_\lambda = u \circ R_\lambda$ on $D_\lambda$. Since $u$ is cylindrical, so is $u_\lambda$ and further $u_\lambda(r, t) = u(r, 2\lambda - t)$. Moreover, since $\Delta_H$ is invariant with respect to the $H$-reflection (see [4]), it follows that $u_\lambda$ satisfies Equation (5.1) in $D_\lambda$.

Now consider the function $w_\lambda = u_\lambda - u$ in $D_\lambda$. It satisfies the equation

\begin{equation}
\Delta_H w_\lambda + c(\xi)w_\lambda \leq 0 \quad \text{in } D_\lambda
\end{equation}

with the boundary conditions

\begin{equation}
w_\lambda \geq 0 \quad \text{on } \partial D_\lambda.
\end{equation}

Let $\delta$ be the constant appearing in Proposition 1.3 corresponding to the operator $L = \Delta_H + c$ of Equation (5.3) on $C$. Observe that here $c(\xi)$ is bounded since $f$ is Lipschitz. For $\lambda < \lambda_0$ and sufficiently close to $\lambda_0$, we have $\text{meas } (D_\lambda) < \delta$. Hence by the maximum principle 1.3, it follows that $w_\lambda \geq 0$ in $D_\lambda$.

We claim that $w_\lambda \geq 0$ in $D_\lambda$ for every $\lambda > 0$. For otherwise, let $\mu = \inf \{ \lambda : w_\lambda \geq 0 \text{ for } \lambda < s < \lambda_0 \}$ and suppose $\mu > 0$. By continuity, $w_\mu \geq 0$. Further since $u$ is positive inside $\Omega$, the maximum principle implies that $w_\mu > 0$ in $D_\mu$.

Let $K \subset D_\mu$ be a compact set such that

\[ \text{meas } (D_\mu \setminus K) < \frac{\delta}{2} \]

where $\delta$ is the constant chosen above. Since $K$ is compact and $w_\mu > 0$ on $K$, there exists $\overline{\lambda}$ near $\mu$ and $0 < \overline{\lambda} < \mu$ such that

\begin{equation}
w_{\overline{\lambda}} > 0 \quad \text{in } K.
\end{equation}

Further we may choose $\overline{\lambda}$ such that

\[ \text{meas } (D_{\overline{\lambda}} \setminus K) < \delta. \]

On $D_{\overline{\lambda}} \setminus K$, $w_{\overline{\lambda}}$ satisfies the differential equation (5.3) with boundary condition $w_{\overline{\lambda}} \geq 0$ on $\partial (D_{\overline{\lambda}} \setminus K)$. Since $\text{meas } (D_{\overline{\lambda}} \setminus K) < \delta$, by Proposition 1.3 it
follows that $w_\lambda \geq 0$ on $D_\lambda \setminus K$. Therefore, $w_\lambda \geq 0$ on $D_\lambda$. From (5.5), the strong maximum principle implies that $w_\lambda > 0$ in $D_\lambda$. This contradicts the definition of $\mu$. Hence $\mu = 0$; which completes the proof. \hfill \Box

**Remark.** It is clear from the proof that we don’t use the fact that the solutions $u$ are cylindrical, we only use that $u(x,y,t) = u(y,x,t)$. Hence the theorem holds true under this weaker condition on the solution. Almeida and Ge ([1]) use a similar condition, precisely if $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ they suppose that $u(x_1, x_2, \ldots, y_1, y_2, \ldots, t) = u(y_1, x_2, \ldots, x_1, y_2, \ldots, t)$.

6. A Maximum principle for locally subelliptic operators.

Here we prove a maximum principle for locally subelliptic operators, using the idea suggested by T. Coulhon. We first recall the definition of subelliptic operator from [18] and [12].

An operator $L$ is said locally subelliptic in $\mathbb{R}^n$ if for an open subset $\Omega$ of $\mathbb{R}^n$, we can write

$$L = \sum_{i,j=1}^{n} \frac{1}{h(x)} \frac{\partial}{\partial x_i} \left( h(x) a_{ij}(x) \frac{\partial}{\partial x_j} \right)$$

(6.1)

where the coefficients $a_{ij}$ and $h$ are $C^\infty$ real valued functions on $\overline{\Omega}$, $h$ is positive and the matrix $A(x) = (a_{ij}(x))$ is symmetric positive semidefinite for every $x \in \overline{\Omega}$.

Further $L$ satisfies a *subelliptic* estimate: There exists a constant $C$ and a number $\varepsilon > 0$ such that all $u \in C^\infty_0(\Omega)$ satisfy

$$\|u\|_2^2 \leq C \left( \int \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} h(x) \, dx + \int |u(x)|^2 \, dx \right)$$

(6.2)

where

$$\|u\|_s = \left( \int |\hat{u}(\xi)|^2 (1 + |\xi|^2)^s \, d\xi \right)^{1/2}$$

denotes the standard Sobolev norm of order $s$.

Clearly, if $A$ is a positive definite matrix, then $L$ is an elliptic operator which satisfies (6.2) with $\varepsilon = 1$. Examples of $L$ include the operators which can be written as sum of vector fields satisfying Hörmander’s condition. In this case, $\varepsilon = 1/2$. See [12] for other examples.

Let $\rho$ denote the distance function canonically associated with $L$ which is continuous and defines a topology on $\mathbb{R}^n$ (see [12] and references therein). We denote this space $\mathcal{M} = (\mathbb{R}^n, \rho)$. 

---

**MONOTONICITY AND SYMMETRY ON LIE GROUPS**
The gradient associated to operator $L$ is defined as
\begin{equation}
\nabla_L(u,v) = \frac{1}{2} L(uv) + uLv + vLu
\end{equation}
see [18]. We denote
\begin{equation}
|\nabla_L u| = (\nabla_L(u,u))^{1/2}.
\end{equation}

Almeida and Ge proved a weak comparison principle (Theorem 2.1) in [1] for $n$-dimensional manifolds $(M,g)$ for the elliptic operator defined locally as
\begin{equation}
L u = -\sum_{i,j=1}^\infty \frac{1}{(\det g)^{1/2}} \frac{\partial}{\partial x^i} \left( a'(|\nabla u|^2)(\det g)^{1/2} g^{ij} \frac{\partial u}{\partial x^j} \right)
\end{equation}
where $a \in W^{2,\infty}((0, \infty)) \cap C^0([0, \infty))$ is such that $a'(t) - 2(a'(t))^{-1}t \geq \alpha > 0$ for some $\alpha > 0$.

Their proof relied on the following Poincaré type inequality: For an open subset $M'$ of $M$, there exists two constants $\gamma, C > 0$ such that if $\text{vol}(M') \leq \gamma$, then
\begin{equation}
(6.4) \quad \int_{M'} |\psi|^2 \, d\text{vol} \leq C \text{vol}(M')^{2/n} \int_{M'} |\nabla \psi|^2 \, d\text{vol}, \text{ for all } \psi \in H^1_0(M').
\end{equation}

We will essentially show that an inequality similar to (6.4) holds for the operator $L$ on $M$.

**Proposition 6.1.** Let $B_L(\xi, R) \subset M$ denote a ball with center $\xi$ and radius $R$ (with respect to the distance $\rho$). Then for every nonempty compact subset $\Omega$ of $B(\xi, R)$, there exists $\nu > 0$ and a constant $C_0$ depending only on $B(\xi, R)$ such that
\begin{equation}
(6.5) \quad \|f\|^2_2 \leq C_0 \text{meas}(\Omega)^\nu \|\nabla_L f\|^2_2 \quad \text{for every } f \in C_0^\infty(\Omega).
\end{equation}
Here, $\nabla_L$ is the gradient associated to $L$.

Observe that when $L$ is an elliptic operator, then (6.5) reduces to (6.4) with $\nu = 2/n$.

**Proof.** Observe that, the distance function $\rho$ satisfies the doubling property (see [12]): There exists a constant $d$ such that
\begin{equation}
|B_L(x, 2R)| \leq d |B_L(x, R)| \quad \text{for all } x \in M, \ R > 0
\end{equation}
where $|B_L(x, R)| = \mu(B_L(x, R))$ is the volume or Lebesgue measure of the ball $B_L(x, R)$.

We also recall the Poincaré inequality proved in [18] (Lemma 2.4): There exists constant $C$ such that for every $f \in C^\infty_0(M)$
\begin{equation}
(6.7) \quad \|f - f_R\| \leq CR\|\nabla f\|_2 \quad \text{for all } R > 0,
\end{equation}
where $f_R$ is the mean of $f$ over the ball $B_L(x, R)$. 

Now as in [11], [18], [7] (see references therein) it can be proved that (6.7) and (6.6) implies the Faber-Krahn type of inequality for \( M \): i.e., there exists constants \( a > 0, \nu > 0 \) such that, for every \( x \in M \), \( R > 0 \) and for every nonempty compact subset \( \Omega \) contained in \( B_L(x, R) \),

\[
\lambda_1(\Omega) \geq \frac{a}{R^2} \left( \frac{|B_L(x, R)|}{|\Omega|} \right)^\nu
\]

where

\[
\lambda_1(\Omega) = \inf \left\{ \frac{\| \nabla G f \|^2_2}{\| f \|^2_2} : f \in C^\infty_0(\Omega) \right\}.
\]

In particular, we can conclude from Faber-Krahn inequality (6.8) that for a fixed ball \( B_L(x, R) \), \( R \geq 1/2 \), for every nonempty subset \( \Omega \subset B_L(x, R) \), we have

\[
\| f \|^2_2 \leq \frac{R^2}{a} \left( \frac{|\Omega|}{|B_L(x, R)|} \right)^\nu \| \nabla G f \|^2_2 \quad \text{for every } f \in C^\infty_0(\Omega)
\]

(6.10)

where \( C_0 = \frac{R^2}{a|B_L(x, R)|^\nu} \) is a fixed constant for \( B_L(x, R) \).

Using the inequality (6.5) we have:

**Proposition 6.2.** Let \( \Omega \) be a bounded domain in \( M \) and \( L \) be a subelliptic operator as defined in (6.1). Assume that \( \| c \|_{L^\infty(\Omega)} \leq b \). For a subset \( \Sigma \subset \Omega \), there exists \( \delta > 0 \) depending only on \( b, \Omega \) and \( C_0 \) (the constant appearing in (6.5)) such that the maximum principle holds for \( L + c \) in \( \Sigma \) provided

\[ \text{meas} (\Sigma) < \delta. \]

**Proof.** First choose a ball \( B_L(x_0, R) \) such that \( \overline{\Omega} \subset B_L(x_0, R) \) and fix it for the following discussion. Note that this \( R \) depends on \( \Omega \). And let \( C_0 \) be the constant defined in the Proposition 6.1 with respect to this ball.

Let \( \Sigma \subset \Omega \) and consider the function \( u \in S^{1,2}(M) \cap L^\infty(M) \) satisfying

\[ Lu + c(x)u \geq 0 \quad \text{in } \Sigma, \quad \lim_{x \to \partial \Sigma} u(x) \leq 0. \]

Here \( S^{1,2}(M) \) is the completion of \( C^1(M) \) under the seminorm

\[ \| f \|_{1,2} = \| \nabla_L f \|_2 + \| f \|_2. \]

Define \( u^+(x) = \max\{u(x), 0\} \) and \( \Sigma^+ = \{ x \in \Sigma : u(x) > 0 \} \). Then \( u^+ \) satisfies the equation

\[
Lu^+(x) + c(x)u^+(x) \geq 0 \quad \text{in } \Sigma^+
\]

(6.11)

\[
u^+ = 0 \quad \text{on } \partial \Sigma^+.
\]
Multiplying (6.11) by $u^+$ and integrating by parts, we have

$$
\int_{\Sigma^+} |\nabla L u^+|^2 \, dx = - \int_{\Sigma^+} c(x)|u^+|^2 \, dx
\leq \|c\|_{L^\infty} \int_{\Sigma^+} |u^+|^2 \, dx \leq b \int_{\Sigma^+} |u^+|^2 \, dx.
$$

Now from (6.13) and (6.5) we obtain

$$
\int_{\Sigma^+} |\nabla L u^+|^2 \, dx \leq b \int_{\Sigma^+} |u^+|^2 \, dx \leq b C_0 \nu \int_{\Sigma^+} |\nabla L u^+|^2 \, dx.
$$

Choose $\delta < (b C_0)^{-1/\nu}$. If $\text{meas}(\Sigma) < \delta$ then (6.14) implies that

$$
\int_{\Sigma^+} |\nabla L u^+|^2 \, dx = 0.
$$

It follows that the inequalities in (6.14) are in fact equalities with each term equal to 0. In particular,

$$
\int_{\Sigma^+} |u^+|^2 \, dx = 0
$$

and hence $u^+ \equiv 0$. \qed

**Acknowledgements.** We would like to thank Xavier Cabré and Arvind Nair for useful conversations.

We also thank Thierry Coulhon for fruitful discussion and suggesting the references related to the inequalities in Section 6.

This work was completed while the second author was visiting the Mathematics Department of Università degli studi di Roma “La Sapienza” with a grant from G.N.A.F.A.-CNR. She thanks the Department of Mathematics for the hospitality.
A “cube” in the Heisenberg group $H^1$.

References


16

I. BIRINDELLI AND J. PRAJAPAT


Università di Roma “La Sapienza”
P.le Aldo Moro 5
00185 Roma
Italia

Indian Statistical Institute
8th Mile, Mysore Road
Bangalore 560 059
India

E-mail address: jyotsna@isibang.ac.in
THE HARMONIC FUNCTIONAL CALCULUS AND HYPERREFLEXIVITY

JOHN B. CONWAY AND MAREK PTAK

A natural $L^\infty$ functional calculus for an absolutely continuous contraction is investigated. It is harmonic in the sense that for such a contraction and any bounded measurable function $\phi$ on the circle, the image can rightly be considered as $\hat{\phi}(T)$, where $\hat{\phi}$ is the solution of the Dirichlet problem for the disk with boundary values $\phi$. The main result shows that if the functional calculus is isometric on $H^\infty$, then it is isometric on all of $L^\infty$. As a consequence we obtain that if the contraction has an isometric $H^\infty$ functional calculus and is in class $C_{00}$, then the range of the harmonic functional calculus is a hyperreflexive subspace of operators. In particular, the space of all Toeplitz operators with a bounded harmonic symbol acting on the Bergman space of the disc is hyperreflexive. Applications of these results to subnormal operators are also presented.

1. Introduction.

Let $\mathcal{H}$ be a complex, separable Hilbert space and let $\mathcal{B}(\mathcal{H})$ be the algebra of all bounded linear operators from $\mathcal{H}$ into itself. Assume $T$ is a contraction in $\mathcal{B}(\mathcal{H})$ that is absolutely continuous. That is, if $T$ has a reducing subspace on which it is unitary, then the spectral measure of this unitary is absolutely continuous. It is a well-known result of Sz.-Nagy [14] that $T$ has a unitary dilation. That is, there is a Hilbert space $\mathcal{K}$ that contains $\mathcal{H}$ and a unitary operator $U$ on $\mathcal{K}$ such that $T^n = P_n U^n | \mathcal{H}$ for all $n \geq 0$. (A proof of this can be found in [10], p. 200. This book will serve as general background for this paper as well [7].) This unitary, moreover, is absolutely continuous. Therefore for any bounded Borel function $\phi$, we can define the operator $\phi(T) \equiv P \phi(U)| \mathcal{H}$. If $L^\infty = L^\infty(\partial \mathbb{D})$ is the $L^\infty$-space of Lebesgue measure on the circle, $\partial \mathbb{D}$, then this defines a map $\xi : L^\infty \to \mathcal{B}(\mathcal{H})$ given by $\xi(\phi) = \phi(T)$. The properties of this map are summarized below. The weak$^*$ topology on $L^\infty$ referred to in this result is the usual one it has as the Banach space dual of $L^1$; the weak$^*$ topology on $\mathcal{B}(\mathcal{H})$ is the one it has as the dual of the trace class, $\mathcal{B}_1$.  

19
Theorem 1.1. $\xi : L^\infty \to B(\mathcal{H})$ is a positive linear contraction that is weak$^*$ continuous.

The proof of this theorem is standard and almost immediate from the definition of $\xi$. Note that $\xi$ is not multiplicative unless $T$ is unitary. When this map $\xi$ is restricted to the bounded analytic functions, $H^\infty$, then it is called the Sz.-Nagy-Foias functional calculus for the operator $T$. It is the functional calculus $\xi : L^\infty \to B(\mathcal{H})$ that is the subject of this paper. Define $T(T)$ to be the range of $\xi$.

This functional calculus was introduced in a more general setting in [11], where many of its properties are deduced and some applications are presented. There is an overlap between [11] and the remainder of this section.

The reason that the functional calculus $\xi$ is called the “harmonic functional calculus” is the following. If $p$ and $q$ are analytic polynomials, then $\xi(p + q) = p(T) + q(T)^*$. Now $p + q$ is the typical trigonometric polynomial. If $\phi \in L^\infty$, then there is a sequence of trigonometric polynomials $\{f_n\}$ such that $\|f_n\|_\infty \leq \|\phi\|_\infty$ and $f_n \to \phi$ weak$^*$ in $L^\infty$. Indeed, one can take the $f_n$ to be the Césaro sums of the Fourier series of $\phi$. It follows that $f_n(T) \to \phi(T)$ weak$^*$ in $B(\mathcal{H})$. But on the open disk the trigonometric polynomials $\{f_n\}$ converge uniformly on compact subsets of $D$ to $\hat{\phi}$, the harmonic extension of $\phi$. This is more than a slight of hand. Indeed, as will be seen below, in the case of many operators such as the Bergman shift, $\phi(T)$ can be equivalently defined in terms of $\hat{\phi}$. A further, more explicit connection with harmonic functions can be seen by a consideration of absolutely continuous contractions that are normal operators.

Let $N$ be a normal operator on $\mathcal{H}$ that is an absolutely continuous contraction and let $N = \int z \, dE(z)$ be its spectral decomposition. For vectors $x$ and $y$ in $\mathcal{H}$, let $\mu_{x,y}$ be the measure defined on $\text{cl} \, D$ by $\mu_{x,y}(\Delta) = \langle E(\Delta)x, y \rangle$. Denote by $\tilde{\mu}_{x,y}$ the sweep of $\mu_{x,y}$ to $\partial D$. (See p. 311 of [9].) It follows that $\mu_{x,y}$ is absolutely continuous on the circle and for every $\phi \in L^\infty$,

$$\int_{\partial D} \phi \, d\tilde{\mu}_{x,y} = \int_{\text{cl} \, D} \phi \, d\mu_{x,y} = \langle \hat{\phi}(N)x, y \rangle.$$

In particular, $P(\Delta) = \hat{\chi}_\Delta(N)$ defines a positive operator-valued measure on the circle with $P(\partial D) = 1$. Call $P$ the sweep of the spectral measure $E$. By the Naimark Dilation Theorem (see p. 197 of [10]), there is a Hilbert space $\mathcal{K}$ containing $\mathcal{H}$ and a spectral measure $F$ on the circle with values in $B(\mathcal{K})$ such that $P(\Delta) = P_{\mathcal{H}}F(\Delta)|\mathcal{H}$ for every Borel subset $\Delta$ of $\partial D$.

Proposition 1.2. Let $N$ be a normal operator that is an absolutely continuous contraction and let $N = \int z \, dE(z)$ be its spectral decomposition. If $P$ is the sweep of the spectral measure $E$, $F$ is the minimal spectral measure with values on $B(\mathcal{K})$ that dilates $P$, $\mathcal{H} \subseteq \mathcal{K}$, then $U = \int z \, dF(z)$ is the minimal unitary dilation of $N$. 
Proof. Using the notation that preceded the statement of the proposition, if $x, y \in H$ and $n \geq 0$, then $\langle U^nx, y \rangle = \int z^n \, d(F(z)x, y) = \int z^n \, d\mu_{x,y} = \langle N^n x, y \rangle$, since $z^n$ is harmonic. Thus $U$ is a unitary dilation. Now to show that it is minimal.

Let $\mathcal{L}$ be the closed linear span of $\{U^nH : n \in \mathbb{Z}\}$. To show the minimality of $U$ it must be shown that $\mathcal{L} = \mathcal{K}$. But $\mathcal{L}$ is clearly a reducing subspace for the unitary operator $U$, so $F(\Delta)\mathcal{L} \subseteq \mathcal{L}$ for every Borel subset $\Delta$ of $\partial \mathbb{D}$. Now the fact that $F$ is the minimal spectral measure that dilates $P$ implies that $\mathcal{L} = \mathcal{K}$.

The next result is a corollary of the preceding proposition, but it can also be proved directly.

**Corollary 1.3.** If $N$ is a normal operator that is an absolutely continuous contraction, then the functional calculus $\xi : L^\infty \to \mathcal{B}(H)$ is given by

$$\xi(\phi) = \hat{\phi}(N) = \int \hat{\phi} \, dE,$$

where $\hat{\phi}$ is the solution of the Dirichlet problem with boundary values $\phi$.

**Corollary 1.4.** If $S$ is a subnormal, absolutely continuous contraction on $H$ with minimal normal extension $N$ on $\mathcal{K}$, then $\xi(\phi) = P_H \hat{\phi}(N)|H$ for every $\phi$ in $L^\infty$.

In the case of a subnormal operator $S$ as described in the preceding corollary, there is a richer linear functional calculus. If $\mu$ is a scalar-valued spectral measure for $N$, then we can define $\phi(S) = P_H \hat{\phi}(N)$ for every $\phi$ in $L^\infty(\mu)$. These operators $\{\phi(S) : \phi \in L^\infty(\mu)\}$ are sometimes called the Toeplitz operators associated with the subnormal operator $S$. Indeed, if $S$ is the unilateral shift, then these are the classical Toeplitz operators. Extending results from the classical case to this more general setting seems almost hopeless. There is evidence, however, that some results will extend from the classical case to the Toeplitz operators with a harmonic symbol.

### 2. The functional calculus.

The key to the proof of the main result of this paper is a result of Tomiyama and Yabuta [17], which we state here for reference. Recall the definition of a uniform algebra and its Shilov boundary.

**Proposition 2.1.** Let $\eta : C(X) \to \mathcal{B}(H)$ be a contractive linear representation such that $\eta(1) = 1$, and let $A$ be a uniform algebra on $X$. If $\eta$ is isometric on $A$ and $X$ is a Shilov boundary of $A$, then $\eta$ is an isometry on $C(X)$.

The main result of the paper is the following:
Theorem 2.2. If $T$ is an absolutely continuous contraction and if the functional calculus $\xi : L^\infty \to B(H)$ is isometric on $H^\infty$, then $\xi$ is isometric on $L^\infty$, $T(T)$ is weak* closed, and $\xi$ is a weak* homeomorphism from $L^\infty$ onto $T(T)$.

Proof. Let $X$ be the maximal ideal space of $L^\infty$ and let $\gamma_L : L^\infty \to C(X)$ be the Gelfand map; so $\gamma_L$ is an isometric isomorphism. We will consider $\xi \circ \gamma_L^{-1} : C(X) \to T(T)$ and show that it is an isometry, which will complete the proof. This is done by applying Proposition 2.1 to $A = \gamma_L(H^\infty)$. Before showing that the assumptions of Proposition 2.1 are fulfilled, we need some notation.

Let $M$ denote the maximal ideal space of $H^\infty$ and let $\gamma_H : H^\infty \to C(M)$ be the Gelfand map. Define $\rho : X \to M$ to be the restriction map, $\rho(\alpha) = \alpha|H^\infty$. By [12], p. 174, $\rho$ is a homeomorphism of $X$ onto $\rho(X)$ and $\rho(X)$ is the Shilov boundary of $H^\infty$.

Note that the diagram

$$
\begin{array}{ccc}
L^\infty & \xrightarrow{\gamma_L} & C(X) \\
\uparrow i & & \uparrow \rho^* \\
H^\infty & \xrightarrow{\gamma_H} & C(M)
\end{array}
$$

is commutative, where $i$ is the inclusion map and $\rho^*(g) = g \circ \rho$ for $g$ in $C(M)$. Indeed, for $\phi$ in $H^\infty$ and $\alpha$ in $X$, $(\rho^* \circ \gamma_H(\phi))(\alpha) = \gamma_H(\phi)(\rho(\alpha)) = \rho(\alpha)(\phi) = \gamma_L(\phi)(\alpha)$.

It is clear that $A$ is norm closed and $1 \in A$. To show that $A$ is a uniform algebra in $C(X)$, it remains to show that $A$ separates the points of $X$. Let $\alpha_1, \alpha_2 \in X$, and assume that $\gamma_L(\phi)(\alpha_1) = \gamma_L(\phi)(\alpha_2)$ for all $\phi$ in $H^\infty$. Since the diagram commutes, this implies that $\rho(\alpha_1)(\phi) = \rho(\alpha_2)(\phi) = \gamma_H(\phi) \circ \rho(\alpha_1) = \gamma_H(\phi) \circ \rho(\alpha_2) = \rho(\alpha_2)(\phi)$. Thus, by the definition of a homomorphism, $\rho(\alpha_1) = \rho(\alpha_2)$. But $\rho$ is injective, so $\alpha_1 = \alpha_2$. Hence $A$ is a uniform algebra.

Note that $X$ is a boundary for $A \subset C(X)$. If $F$ is a proper closed subset of $X$, then $\rho(F)$ is a proper closed subset of the Shilov boundary of $H^\infty$, $\rho(X)$. Hence there is $\phi$ in $H^\infty$ such that $1 = \|\gamma_H(\phi)\| = \|\gamma_H(\phi)\|_{\rho(X)}$ and $\|\gamma_H(\phi)\|_{\rho(F)} < 1$. If $f = \gamma_L(\phi) \in A$, then, for $\alpha$ in $X$,

$$
|f(\alpha)| = |\gamma_L(\phi)(\alpha)| = |(\rho^* \circ \gamma_H(\phi))(\alpha)| = |\gamma_H(\phi)(\rho(\alpha))|.
$$

Hence $\|f\| = 1$, but $|f(\alpha)| < 1$ for $\alpha \in F$. Thus $F$ cannot be a boundary for $A$ and so $X$ is the Shilov boundary for $A$. By Proposition 2.1, $\xi$ is an isometry.

Since $\xi$ is weak* continuous (1.1), the fact that $T(T) = \xi(L^\infty)$ is weak* closed and that $\xi$ is a weak* homeomorphism from $L^\infty$ onto $T(T)$ is a standard consequence of the Krein-Smulian Theorem. (For example, see [8], Proposition 1.2.7.)
Some special cases of this result have appeared in the literature. In [13] this theorem is shown for a class of weighted Bergman operators on the disk (Theorem 10), though the proof there is not correct.

The standard terminology is that $A$ is the set of all absolutely continuous contractions for which the $H^\infty$ functional calculus is isometric. See [5]. So the preceding theorem says that every contraction in class $A$ has an isometric harmonic functional calculus.

This section concludes with an application of this result to all subnormal operators. Let $S$ be a subnormal operator on $H$ and let $N$ be its minimal normal extension acting on $K$. If $\mu$ is a scalar-valued spectral measure for $N$, then a natural multiplicative, functional calculus for $S$ is $\phi \rightarrow \phi(S) = \phi(N)\vert H$ for $\phi$ in $P^\infty(\mu)$, the weak* closure of the polynomials in $L^\infty(\mu)$. By a result of Sarason [16], there is a “special” open set $G$ and a decomposition $\mu = \mu_\infty + \mu_0$, $\mu_0 \perp \mu_\infty$, such that $P^\infty(\mu) = L^\infty(\mu_0) \oplus H^\infty(G,\mu_\infty)$, where $H^\infty(G,\mu_\infty)$ is an isometric, weak* homeomorphic embedding of $H^\infty(G)$ onto $P^\infty(\mu_\infty) \subseteq L^\infty(\mu_\infty)$. (Also see [8], p. 301.) The open set $G$ is called the Sarason hull of $\mu$.

The next result gathers information about this. The reader can consult [8] for details.

**Theorem 2.3.** Let $S$ be a subnormal operator on $H$ with minimal normal extension $N$ acting on $K$ and scalar-valued spectral measure $\mu$. If $G$ is the Sarason hull of $\mu$ and $G_1,G_2,\ldots$ are its components, then there is a decomposition of $\mu$, $\mu = \mu_0 + \mu_1 + \cdots$, where $\mu_n \perp \mu_m$ for $n \neq m$, such that the following hold:

(a) $P^\infty(\mu) = L^\infty(\mu_0) \oplus P^\infty(\mu_1) \oplus \cdots$;

(b) the polynomials are weak* dense in $L^\infty(\mu_0)$;

(c) for $n \geq 1$, the identity map on polynomials extends to an isometric, weak* homeomorphism of $P^\infty(\mu_n)$ onto $H^\infty(G_n)$;

(d) there is a corresponding decomposition of the Hilbert space $H$ as $H = H_0 \oplus H_1 \oplus \cdots$, where each $H_n$ reduces $S$ and, for $n \geq 0$, the weak* closed algebra generated by $S_n \equiv S\vert H_n$ and the identity is precisely $\{f(S)\vert H_n : f \in P^\infty(\mu_n)\} = \{f(S_n) : f \in P^\infty(\mu_n)\}$;

(e) for $n \geq 1$, the set $G_n$ is simply connected and if $\tau : G_n \rightarrow \mathbb{D}$ is a Riemann map, then $\tau$ is a weak* generator of $H^\infty(G_n)$ as an algebra and $\tau^{-1}$ is a weak* generator of $H^\infty(\mathbb{D})$ as an algebra;

(f) for $n \geq 1$, $\mu_n$ is supported on $cG_n$ and $\mu_n\vert \partial G_n$ is absolutely continuous with respect to harmonic measure for $G_n$;

(g) the Riemann map $\tau : G_n \rightarrow \mathbb{D}$ extends to a subset of $\partial G_n$ having full harmonic measure and on this set is a one-to-one, measurable map onto a subset of $\partial \mathbb{D}$ having full measure such that its inverse is also measurable.
Remarks.

1. Condition (b) of this theorem says that $S_0$ is a reductive normal operator.
2. The statement in (e) means that the polynomials in $\tau$ are weak* dense in $H^\infty(\Gamma_n)$ and polynomials in $\tau^{-1}$ are weak* dense in $H^\infty(\Delta)$.
3. In light of (e) no distinction between $P^\infty(\mu_n)$ and $H^\infty(\Gamma_n)$ will be made. That is, when a function $\phi$ in $P^\infty(\mu_n)$ is considered, it will be assumed that $\phi$ is a bounded analytic function on $\Gamma_n$.
4. Using the results from the statement of Theorem 2.3 it can be shown that $G$ is the set of all $\lambda$ in $\mathbb{C}$ such that the map $p \to p(\lambda)$ defined on polynomials extends to a weak* continuous homomorphism from $P^\infty(\mu)$ into $\mathbb{C}$.
5. If $L^\infty(\partial \Gamma_n)$ denotes the $L^\infty$ space of harmonic measure for $\Gamma_n$, then $\rho$ says that the restriction of $\tau$ to $\partial \Gamma_n$ induces an isometric isomorphism and a weak* homeomorphism of $L^\infty(\partial \Delta)$ onto $L^\infty(\partial \Gamma_n)$.

As mentioned after Corollary 1.4, for a subnormal operator $S$ there is an additional linear functional calculus. Using the preceding notation, for every $\phi$ in $L^\infty(\mu_n)$, define $\phi(S) \equiv P_H \phi(N)$. In the proposition below the notation of Theorem 2.3 is used. If $\phi \in L^\infty(\partial \Gamma_n)$ and $\hat{\phi}$ is the solution of the Dirichlet problem for $\Gamma_n$ with boundary values $\phi$, then $\hat{\phi}$ can be considered as an element of $L^\infty(\mu_n)$ by letting $\hat{\phi}$ be itself on $\Gamma_n$ while being $\phi$ on $\partial \Gamma_n$. Note that this makes sense in light of the fact that $\mu_n|\partial \Gamma_n$ is absolutely continuous with respect to harmonic measure. Also note that $\|\hat{\phi}\|_{L^\infty(\mu_n)} = \|\hat{\phi}\|_{\Gamma_n}$.

**Proposition 2.4.** Let $S$ be a subnormal operator and adopt the notation in Theorem 2.3. If $\phi \in L^\infty(\partial \Gamma_n)$, let $\hat{\phi}$ denote the solution of the Dirichlet problem for $\Gamma_n$ with boundary values $\phi$. If $\rho : L^\infty(\mu_0) \oplus L^\infty(\partial \Gamma_1) \oplus L^\infty(\partial \Gamma_2) \oplus \cdots \to B(H)$ is defined by

$$
\rho(\phi_0 \oplus \phi_1 \oplus \phi_2 \oplus \cdots) = \phi_0(S_0) \oplus \hat{\phi}_1(S_1) \oplus \hat{\phi}_2(S_2) \oplus \cdots,
$$

then $\rho$ is a linear isometry, its range is weak* closed, and $\rho$ is a weak* homeomorphism onto its image.

**Proof.** Since $S_0$ is normal, it suffices to show that for $n \geq 1$ and $\phi$ in $L^\infty(\partial \Gamma_n)$, $\|\hat{\phi}(S_n)\| = \|\hat{\phi}\|_{\Gamma_n} \equiv \sup\{|\hat{\phi}(z)| : z \in \Gamma_n\}$. If $\tau : \Gamma_n \to \mathbb{D}$ is a Riemann map, then $\tau(S_n)$ is an absolutely continuous contraction for which the $H^\infty$ functional calculus is isometric. Indeed, for $f$ in $H^\infty$, $f \circ \tau \in H^\infty(\Gamma_n)$. Therefore by Theorem 2.3 we have that $\|f(\tau(S_n))\| = \|f \circ \tau\|_{\Gamma_n} = \|f\|_{\Delta}$. By Theorem 2.2, for every bounded harmonic function $u$ on the disk, $\|u(\tau(S_n))\| = \|u\|_{\Delta}$. Taking $u = \hat{\phi} \circ \tau^{-1}$ implies that $\|\hat{\phi}(S_n)\| = \|(\hat{\phi} \circ \tau^{-1})(\tau(S))\| = \|\hat{\phi} \circ \tau^{-1}\|_{\Delta} = \|\hat{\phi}\|_{\Gamma_n}$.

Again, the rest of the proposition follows from Proposition 2.7 of [8]. □
3. Hyperreflexivity.

Recall that a subspace \( \mathcal{M} \) of \( \mathcal{B}(\mathcal{H}) \) is said to be reflexive if every operator \( B \) in \( \mathcal{B}(\mathcal{H}) \) satisfying \( Bx \in \text{cl} [\mathcal{M}x] \) for all \( x \) in \( \mathcal{H} \) necessarily belongs to \( \mathcal{M} \).

For any \( B \) in \( \mathcal{B}(\mathcal{H}) \), let \( \text{dist} (B, \mathcal{M}) \) denote the usual distance from \( B \) to \( \mathcal{M} \) in \( \mathcal{B}(\mathcal{H}) \), and let

\[
\alpha(B, \mathcal{M}) = \sup\{\|Q^\perp BP\| : P, Q \text{ projections with } Q^\perp MP = (0)\}.
\]

The linear space \( \mathcal{M} \) is called hyperreflexive if there is a constant \( C > 0 \) such that \( \text{dist} (B, \mathcal{M}) \leq C \alpha(B, \mathcal{M}) \) for every \( B \) in \( \mathcal{B}(\mathcal{H}) \). The smallest constant \( C \) is called the hyperreflexive constant and is denoted by \( \kappa(\mathcal{M}) \). (See [10], Chapter 8, for the elementary properties of reflexive and hyperreflexive subspaces.) It is straightforward that hyperreflexive subspaces are reflexive, and reflexive spaces are weakly (WOT) closed.

If \( S \) is the unilateral shift on the Hardy space \( H^2 \), it was shown in [1] (also see [10], Prop. 56.8) that the space \( T(S) \) is far from being reflexive, though it is WOT closed. Indeed, it is transitive. However, \( T(S) \) contains many reflexive subspaces, for example \( \mathcal{A}(S) \), the weakly closed algebra generated by \( S \). All the reflexive subspaces of \( T(S) \) were characterized in [1], where it is also shown that every weak* closed subspace of \( T(S) \) is either reflexive or transitive. Below we will show that \( T(T) \) is hyperreflexive for all \( C_{00} \) contractions \( T \) in the class \( \mathcal{A} \).

Recall that \( \mathcal{B}(\mathcal{H}) \) is the Banach space dual of the trace class \( \mathcal{B}_1 \). If \( \mathcal{M} \) is a linear manifold in \( \mathcal{B}(\mathcal{H}) \), then the preannihilator of \( \mathcal{M} \) is the space of weak* continuous linear functionals \( \mathcal{M} \) \( \equiv \{L \in \mathcal{B}_1 : L(\mathcal{M}) = (0)\} \). Let \( \mathcal{Q}_\mathcal{M} = \mathcal{B}_1 / \mathcal{M} \). So the Banach space dual of \( \mathcal{Q}_\mathcal{M} \) is the weak* closure of \( \mathcal{M} \). For any \( L \) in \( \mathcal{B}_1 \), \( [L] = [L]_\mathcal{M} \) denotes the coset in \( \mathcal{Q}_\mathcal{M} \).

Now consider an absolutely continuous contraction \( T \) and the subspace \( T(T) \). Since \( \xi : L^\infty \rightarrow T(T) \) is weak* continuous, there is a bounded linear map \( \theta : \mathcal{Q}_{T(T)} \rightarrow L^1 \) such that \( \xi = \theta^* \). Thus for every \( L \) in \( \mathcal{B}_1 \) and \( \phi \) in \( L^\infty \), \( \int \phi \theta([L]) \) \( dm = ([L], \xi(\phi)) = \text{tr} (L\xi(\phi)). \) If \( x \) and \( y \) are vectors in \( \mathcal{H} \), \( x \otimes y \) is the rank one operator on \( \mathcal{H} \) given by \( (x \otimes y)(h) = \langle h, y \rangle x \). Using the notation of [6], let \( x \cdot y \equiv \theta(x \otimes y) \). So for every \( \phi \) in \( L^\infty \),

\[
\text{tr} [\xi(\phi)(x \otimes y)] = \int_{\mathcal{B}} \phi \left( x \cdot y \right) dm.
\]

Denote by \( \mathcal{X}_0(\mathcal{M}) \) the set of all \( [L] \) in \( \mathcal{Q}_\mathcal{M} \) such that there exist sequences \( \{x_n\}_{n=1}^\infty \), \( \{y_n\}_{n=1}^\infty \) in \( \mathcal{H} \) with \( \|x_n\| \leq 1 \) and \( \|y_n\| \leq 1 \) for all \( n \) such that

\[
\begin{align*}
(a) \lim_{n \rightarrow \infty} \|[x_n \otimes y_n] - [L]\| &= 0, \\
(b) \lim_{n \rightarrow \infty} \|[x_n \otimes w]\| &= 0 \quad \text{for all } w \in \mathcal{H}, \\
(c) \lim_{n \rightarrow \infty} \|[w \otimes y_n]\| &= 0 \quad \text{for all } w \in \mathcal{H}.
\end{align*}
\]
Say that \( M \) has property \( X_{0,1} \) if the unit ball of \( Q_M \) is contained in the closed convex hull of \( X_0(M) \). (It was shown in [5] that \( X_0(M) \) is in fact absolutely convex and closed.)

**Theorem 3.2.** If \( T \) is a \( C_{00} \) contraction in the class \( \mathbb{A} \), then \( T(T) \) is hyperreflexive with constant at most 3. Moreover, each weak* closed subspace of \( T(T) \) is hyperreflexive.

**Proof.** We will show that \( T(T) \) has property \( X_{0,1} \). By Theorem 3.1 of [4], this proves the hyperreflexivity of \( T(T) \) with the constant 3. The last statement of the theorem follows from hereditary behavior of the property \( X_{0,1} \).

Note also that the weak* and weak operator topologies coincide. (See [2], Theorem 2.)

By [6], Lemma 4.2, the fact that \( T \) is a \( C_{00} \) operator implies that the following is satisfied.

**Condition 3.3.** For any \( f \) in \( L^1 \) with \( \| f \|_1 \leq 1 \) there are sequences \( \{x_n\}, \{y_n\} \) in \( \mathcal{H} \) with \( \|x_n\| \leq 1, \|y_n\| \leq 1 \) such that for every \( w \) in \( \mathcal{H} \),

\[
\begin{align*}
\|f - x_n \cdot y_n\|_1 & \to 0 \\
\|x_n \cdot w\|_1 + \|w \cdot y_n\|_1 & \to 0.
\end{align*}
\]

Let \([L]\) be any weak* continuous linear functional on \( T(T) \) with \( \|[L]\| \leq 1 \). Applying (3.3) to \( \theta([L]) \) and using the hypothesis that \( \xi \), and hence \( \theta \), is an isometry, we get that

\[
\|[L] - [x_n \otimes y_n]\| = \|\theta([L]) - x_n \cdot y_n\|_1 \to 0,
\]

Similarly

\[
\|[x_n \otimes w]\| + \|[w \otimes y_n]\| \to 0 \quad \text{for every } w \text{ in } \mathcal{H}.
\]

This implies that ball \( Q_{T(T)} \subseteq X_{0,1} \), so that \( T(T) \) has property \( X_{0,1} \). \( \square \)

**Remark.** Theorem 3.2 remains true if we assume that \( T \) is an absolutely continuous contraction in the class \( \mathbb{A} \) and Condition (3.3) is fulfilled.

This theorem has application to a large collection of subnormal operators. Once again the notation introduced in connection with Theorem 2.3 is used. Note that if \( S \) is a pure subnormal operator, then the reductive normal summand in (2.3), \( S_0 \), is not present. For an open set \( \Omega \), \( h^\infty(\Omega) \) denotes the bounded harmonic functions on \( \Omega \).

**Theorem 3.4.** Let \( S \) be a pure subnormal operator and adopt the notation of Theorem 2.3 and Proposition 2.4. If \( \mu_n(\partial G_n) = 0 \) for all \( n \geq 1 \), then

\[
T = \{ \tilde{\phi}_1(S_1) \oplus \tilde{\phi}_2(S_2) \oplus \cdots : \phi_n \in L^\infty(\partial G_n) \text{ for } n \geq 1 \}.
\]
is hyperreflexive with constant at most 3. Moreover each weak* closed subspace of \( T \) is hyperreflexive with constant at most 3.

Proof. Let \( \tau_n : G_n \to \mathbb{D} \) be a Riemann map. Recall that \( \tau_n(S_n) \) is a contraction. Put \( T_n = \{ u(S_n) : u \in h^\infty(G_n) \} \) and \( M_n = T(\tau_n(S_n)) \).

Claim. \( T_n = M_n \) for all \( n \geq 1 \).

Indeed, \( M_n = \{ v(\tau_n(S_n)) : v \in h^\infty(\mathbb{D}) \} \). Since \( h^\infty(G_n) = \{ v \circ \tau_n : v \in h^\infty(\mathbb{D}) \} \), the claim is clearly true.

Theorem 2.3 (g) implies that the scalar-valued spectral measure for the subnormal operator \( \tau_n(S_n) \) is \( \mu_n \circ \tau_n^{-1} \). Thus the assumption that \( \mu_n(\partial G_n) = 0 \) implies that \( \tau_n(S_n) \) is a \( C_{00} \) contraction. Hence Theorem 3.2 implies that \( T_n \) is hyperreflexive. More importantly for this proof, Condition (3.3) is satisfied.

Let \( Q_n \) be the predual of \( T_n \). Since \( T = \bigoplus_n T_n \), the predual of \( T \) is \( Q = \bigoplus_n Q_n \), where this direct sum is an \( \ell^1 \) direct sum. That is, \( ||L|| = \sum_n ||[L_n]|| \) for all \( [L] = \bigoplus_n [L_n] \) in \( \bigoplus_n Q_n \). Let \( \xi_n : L^\infty \to T_n \) be the harmonic functional calculus for the contraction \( T_n = \tau_n(S_n) \), \( \xi_n(\phi) = \hat{\phi}(T_n) \), and let \( \theta_n : Q_n \to L^1 \) be its predual.

Let \( [L] \in Q, ||L|| \leq 1 \), let \( X \) be a finite subset of \( \mathcal{H} \), and let \( \epsilon > 0 \); let \( X_n \) be the set of \( n \)-th coordinates of the vectors in \( X \). Using the fact that each \( \theta_n \) is an isometry, (3.3), when applied to \( T_n \), implies there are vectors \( x_n, y_n \) in \( \mathcal{H} \) with \( ||x_n||, ||y_n|| < ||[L_n]||^{1/2} \) satisfying

\[
||[L_n] - [x_n \otimes y_n]|| < 2^{-n} \epsilon
\]

\[
||[x_n \otimes w_n]|| + ||[w_n \otimes y_n]|| < 2^{-n} \epsilon
\]

for all \( w_n \) in \( X_n \). Put \( x = \oplus_n x_n \) and \( y = \oplus_n y_n \). So \( ||x||, ||y|| \leq 1 \). The reader can verify that as elements of \( Q = \bigoplus_n Q_n \), \( [x \otimes y]_T = \bigoplus_n [x_n \otimes y_n]_{T_n} \). Thus

\[
||[L] - [x \otimes y]|| = \sum_n ||[L_n] - [x_n \otimes y_n]|| < \epsilon.
\]

Similarly, for \( w = \oplus w_n \) in \( X \), \( ||[x \otimes w]|| = \sum_n ||[x_n \otimes w_n]|| < \epsilon \) and \( ||[w \otimes y]|| < \epsilon \).

Thus \( T \) has property \( X_{0,1} \). By Theorem 3.1 of [4], \( T \) is hyperreflexive with constant at most 3. As mentioned in the proof of Theorem 3.2, the last statement of the theorem follows from hereditary behavior of the property \( X_{0,1} \).

It is worth singling out the Bergman operators. For a bounded open set \( \Omega \) in the plane, \( L^2_\alpha(\Omega) \) is the Bergman space of all analytic functions that are square integrable with respect to area measure on \( \Omega \). The Bergman operator for \( \Omega \) is the operator \( S \) defined on \( L^2_\alpha(\Omega) \) by \( (Sf)(z) = zf(z) \). This is a subnormal operator, and, as for the general subnormal operator, we can
define $\phi(S)f \equiv P(\phi f)$ for $f$ in $L^2_2(\Omega)$ and $\phi$ in $L^\infty(G)$. The next corollary is a direct consequence of the preceding theorem.

**Corollary 3.5.** If $S$ is the Bergman operator for the bounded open set $\Omega$ and $T$ is the weak* closure in $\mathcal{B}(L^2_2(\Omega))$ of

$$\{p(S) + q(S)^* : p \text{ and } q \text{ are analytic polynomials}\},$$

then every weak* closed subspace of $T$ is hyperreflexive with constant at most 3.

The next result follows from the preceding corollary or directly from Theorem 3.2. Indeed, it was the attempt to investigate this example that led the authors to the results that are contained in this paper.

**Corollary 3.6.** If $S$ is the Bergman operator for the unit disk $\mathbb{D}$, then every weak* closed subspace of

$$T = \{u(S) : u \in h^\infty(\mathbb{D})\}$$

is hyperreflexive with constant at most 3.

These results raise additional questions. The problem of whether the general subnormal operator is hyperreflexive remains open. But if this is settled, especially if it is settled affirmatively, there will be the question of which subnormal operators $S$ have the property that $T(S)$, the weak* closure of $\{p(S) + q(S)^* : p, q \text{ are analytic polynomials}\}$, is reflexive or hyperreflexive. A rich collection of subnormal operators having these properties is shown to exist in this paper. On the other hand, the unilateral shift has neither. An interesting example to explore would be the operator defined as multiplication by the independent variable on $P^2(\mu)$, where $\mu$ equals area measure on the bottom half of the unit disk and arc length measure on the top half of the unit circle.

**References**


Received June 15, 2000 and revised June 26, 2001. This work was done while the second author was visiting the University of Tennessee with partial support from the Tennessee Science Alliance.

\textbf{DEPARTMENT OF MATHEMATICS}  
\textbf{UNIVERSITY OF TENNESSEE}  
\textbf{KNOXVILLE, TN 37996-1300}  
\textit{E-mail address}: conway@math.utk.edu

\textbf{INSTITUTE OF MATHEMATICS}  
\textbf{UNIVERSITY OF AGRICULTURE}  
\textbf{MICKIEWICZA 24/28}  
\textbf{30-059 KRAKÓW}  
\textbf{POLAND}  
\textit{E-mail address}: rmptak@cyf–kr.edu.pl
GROUPS THAT DO NOT ACT BY AUTOMORPHISMS OF CODIMENSION-ONE FOLIATIONS

R. Feres and D. Witte

Let $\Gamma$ be a finitely generated group having the property that any action of any finite-index subgroup of $\Gamma$ by homeomorphisms of the circle must have a finite orbit. (By a theorem of É. Ghys, lattices in simple Lie groups of real rank at least 2 have this property.) Suppose that such a $\Gamma$ acts on a compact manifold $M$ by automorphisms of a codimension-one $C^2$ foliation, $\mathcal{F}$. We show that if $\mathcal{F}$ has a compact leaf, then some finite-index subgroup of $\Gamma$ fixes a compact leaf of $\mathcal{F}$. Furthermore, we give sufficient conditions for some finite-index subgroup of $\Gamma$ to fix each leaf of $\mathcal{F}$.

1. Introduction and statement of results.

The letter $M$ will denote a compact, connected, boundaryless, smooth manifold of dimension $n$. Let $\mathcal{F}$ be a $C^r$ foliation of $M$ by smooth leaves, $r \geq 2$. It will be assumed that $\mathcal{F}$ is a transversely oriented, codimension-one foliation.

Let $D^s(M,\mathcal{F})$ denote the group of $C^s$ automorphisms of $\mathcal{F}$, $s \geq 0$; that is, the group of $C^s$ diffeomorphisms of $M$ that map leaves to leaves. The normal subgroup consisting of automorphisms of $\mathcal{F}$ that send each leaf to itself will be denoted $D^s(M,\mathcal{F})_0$, and will be called the group of inner automorphisms of the foliation. The quotient

$$O^s(M,\mathcal{F}) = D^s(M,\mathcal{F})/D^s(M,\mathcal{F})_0$$

will be called the group of transverse automorphisms (or outer automorphisms) of $\mathcal{F}$. When a group $\Gamma$ acts by automorphisms of $\mathcal{F}$ so as to define a homomorphism into $D^s(M,\mathcal{F})$, the action will be called transversely finite if $\Gamma$ projects to a finite subgroup of $O^s(M,\mathcal{F})$.

The general question that motivates the results of the present paper can be stated thus: Given $(M,\mathcal{F})$, what groups can act by automorphisms of $\mathcal{F}$ so that the action is not transversely finite, and what groups cannot? For example, if $\mathcal{F}$ is the foliation by fibers of a product manifold $M = B \times L$, with leaves $\{b\} \times L$, $b \in B$, then any group that acts nonfinitely on $B$ by $C^s$-diffeomorphisms also acts nonfinitely by outer-automorphisms of $\mathcal{F}$ (e.g., by setting the action on $L$ to be trivial). In this case, $O^s(M,\mathcal{F})$ is the group of $C^s$-diffeomorphisms of $B$. If, on the other hand, $M = \mathbb{T}^n$ (the $n$-dimensional
flat torus) and $\mathcal{F}$ is the foliation by planes parallel to an irrational hyperplane $F \subset \mathbb{R}^n$, then it is an elementary fact that $\mathcal{O}^s(M, \mathcal{F})$ is isomorphic to $H := (\Gamma \ltimes \mathbb{R}^n)/(\Gamma_0 \ltimes F)$, where $\Gamma$ is the stabilizer of $F$ in $GL(n, \mathbb{Z})$ and $\Gamma_0$ is the subgroup of $\Gamma$ that acts trivially on the quotient $\mathbb{R}^n/F$. In this case, only groups that admit homomorphisms with nonfinite image in $H$ can have nonfinite actions by outer-automorphisms of $\mathcal{F}$. (Allowing big codimension, it is quite easy to construct, say, topologically transitive foliated bundles, with large groups of smooth outer-automorphisms.)

A more specific question that will be addressed here is the following: Suppose that no (topological, say) action of a group $\Gamma$ on the circle yields nontrivial dynamics (that is, nonfinite action). Does $\Gamma$ admit nontrivial “transverse dynamics” on some codimension-one foliation of a compact manifold? The two theorems given below provide support for the negative answer.

**Theorem 1.1.** Suppose that $\Gamma$ is a finitely generated discrete group such that every homomorphism from a finite-index subgroup of $\Gamma$ into the group of homeomorphisms of the circle has a finite orbit on the circle. Also suppose that $\Gamma$ acts by $C^0$-automorphisms of $(M, \mathcal{F})$.

If $\mathcal{F}$ has a closed leaf, then some closed leaf of $\mathcal{F}$ is fixed by some subgroup of finite index in $\Gamma$.

By a bounded transverse invariant measure we mean a holonomy invariant transverse measure that is finite on compact transversals.

**Corollary 1.2.** Suppose that $\Gamma$ is a finitely generated discrete group such that every homomorphism from a finite-index subgroup of $\Gamma$ into the group of homeomorphisms of the circle has a finite orbit on the circle. Also suppose that $\Gamma$ acts by $C^0$-automorphisms of $(M, \mathcal{F})$.

If $\mathcal{F}$ admits a bounded transverse invariant measure, then it also admits a bounded transverse invariant measure $\mu$ that is invariant under $\Gamma$, such that every leaf in the support of $\mu$ is sent to itself under the action of a finite-index subgroup $\Gamma'$ of $\Gamma$.

The next theorem provides a class of foliations on which the $\Gamma$-action is transversely finite. A foliation is said to be almost without holonomy if the germinal holonomy groups of all the non-compact leaves are trivial [7, IV-2.11, p. 251]. This is the case, for example, if the non-compact leaves are simply connected.

**Theorem 1.3.** Suppose that $\Gamma$ is a finitely generated discrete group such that every homomorphism from a finite-index subgroup of $\Gamma$ into the group of homeomorphisms of the circle has a finite orbit on the circle.

If $\mathcal{F}$ is almost without holonomy, then every homomorphism of $\Gamma$ into $D^1(M, \mathcal{F})$ yields a transversely finite action.

It is rather likely that the assumption about the holonomy of $\mathcal{F}$ in Theorem 1.3 can be greatly relaxed.
The following theorem of É. Ghys [6] provides examples of groups that satisfy the requirements of Theorems 1.1 and 1.3, that is, groups for which every homomorphism into the group of homeomorphisms of the circle has a finite orbit. Most of these examples were also established by M. Burger and N. Monod [1, 2].

**Theorem 1.4** (Ghys [6]). Suppose that \( \Gamma \) is an irreducible lattice in a connected, semisimple, real Lie group \( G \) of real rank at least 2, and that there is no continuous homomorphism from \( G \) onto \( PSL(2, \mathbb{R}) \). Then every homomorphism from \( \Gamma \) into the group of homeomorphisms of the circle, \( \mathbb{T}^1 \), has a finite orbit. Furthermore, if \( \Gamma \) acts by \( C^1 \) diffeomorphisms, then some finite-index subgroup of \( \Gamma \) acts trivially on \( \mathbb{T}^1 \).

2. General facts about codimension-one foliations.

We use [3] as our main reference for the fundamental concepts and results about codimension-one foliations. These results were established by R. Sacksteder, P. Dippolito, G. Hector, A. Haefliger, J. Cantwell, L. Conlon, and others. Some of the original sources are [5, 9, 12, 14, 4]. The foliation \( \mathcal{F} \) will be said to be without holonomy if the germinal holonomy group of each leaf of \( \mathcal{F} \) is trivial.

Let \( \mathcal{L} \) denote a smooth one-dimensional foliation of \( M \) everywhere transverse to \( \mathcal{F} \). (Cf. [3, 5.1.2].) It is convenient to work with biregular coordinate charts for the pair \((\mathcal{F}, \mathcal{L})\). These are charts that define foliation boxes for both \( \mathcal{L} \) and \( \mathcal{F} \) simultaneously, having local coordinate maps \( \varphi: U \subset M \to V \subset \mathbb{R}^{n-1} \times \mathbb{R}^1, \varphi(p) = (x(p), y(p)) \), such that \( x = \) constant corresponds to plaques of \( \mathcal{L} \) while \( y = \) constant corresponds to plaques of \( \mathcal{F} \). A biregular cover is an atlas comprised of biregular coordinate charts. Such covers exist. (Cf. [3, 5.1.4].) From now on \( \mathcal{L} \) will denote a fixed transverse foliation to \( \mathcal{F} \) and any foliation box will be assumed without mention to be biregular.

An open \( \mathcal{F} \)-saturated set \( U \) is a called a foliated product if it is connected and \( \mathcal{L}|_U \) fibers \( U \) by open intervals over some \((n - 1)\)-dimensional manifold \( B \). Since \( \mathcal{F} \) is orientable, a foliated product is a trivial interval bundle, homeomorphic to \( B \times (0, 1) \) (although the foliation need not be the product foliation). Each leaf of \( \mathcal{F} \) in \( U \) with the restriction to it of the bundle map is a covering space of \( B \). We note, in particular, that each closed transversal that meets \( U \) has to meet every leaf in \( U \). Let \( d \) be the topological metric on \( U \) induced by the restriction to \( U \) of a Riemannian metric on \( M \) and denote by \( \hat{U} \) the completion of \( U \) in the metric \( d \).

An \( \mathcal{F} \)-saturated set \( U \) is a called a foliated bundle if it is connected and \( \mathcal{L}|_U \) fibers \( U \) over some \((n - 1)\)-dimensional manifold \( B \). (This is more general than a foliated product, because there is no restriction on the fibers.) An \( \mathcal{F} \)-saturated set \( U \) is a called a trivially foliated product if there is a connected
1-manifold $F$ (possibly with boundary), a connected $(n - 1)$-dimensional manifold $B$, and a diffeomorphism from $U$ to $B \times F$, that carries $\mathcal{L}|_U$ and $\mathcal{F}|_U$ to the product foliations of $B \times F$.

**Theorem 2.1** (Dippolito [5]). Let $U$ be a connected $\mathcal{F}$-saturated open set. Then $\hat{U}$ is a connected manifold with finitely many boundary components. The interior of $\hat{U}$ is $U$ and the inclusion $i: U \to M$ extends to an immersion $\hat{i}$ of $\hat{U}$ into $M$ that sends the boundary components of $\hat{U}$ onto boundary leaves of $U$. If $L'$ is a boundary leaf of $U$ then $i^{-1}(L')$ consists of one or two components of the boundary of $\hat{U}$, each component being mapped bijectively to $L'$ by $\hat{i}$. Both $\mathcal{F}$ and $\mathcal{L}$ pull-back under $i$ to well-defined foliations on $\hat{U}$. If $U$ is a foliated product, then $\hat{U}$ is a foliated bundle whose fibers are compact intervals.

**Proof.** This is [3, 5.2.10, 5.2.11, 5.2.12] as well as the remarks after 5.2.12 of the same reference. $\square$

The foliation of $\hat{U}$ obtained by the pull-back of $\mathcal{F}$ will be denoted $\hat{\mathcal{F}}$.

**Theorem 2.2** (Sacksteder [14]). Let $\mathcal{F}$ be a transversely orientable foliation of class $C^2$ and codimension one on a compact manifold. Then the following are equivalent:

1) There exists a bounded transverse invariant measure $\mu$.
2) Either $\mathcal{F}$ has a compact leaf or $\mathcal{F}$ is without holonomy.

**Proof.** This is [10, 2.3.8]. $\square$

A foliation $\mathcal{F}$ of a manifold $U$ will be said to fiber over a manifold $B$ if there is a fibration of $U$ with base $B$ having the leaves of $\mathcal{F}$ as fibers.

**Theorem 2.3.** Let $\mathcal{F}$ be a transversely orientable foliation of class $C^2$ and codimension-one of a compact manifold $M$. Let $U$ be a connected $\mathcal{F}$-saturated open set and suppose that $\mathcal{F}|_U$ is without holonomy. Then either every leaf of $\mathcal{F}|_U$ is closed and $\mathcal{F}|_U$ fibers over a connected 1-manifold, or each leaf of $\mathcal{F}|_U$ is dense in $U$. Furthermore, if $\mathcal{F}|_U$ is a fibration over a 1-manifold $B$, but $(\hat{U}, \hat{\mathcal{F}})$ is not a trivially foliated product, then $B \cong S^1$.

**Proof.** This is [3, 9.1.4, 9.1.6]. $\square$

**Theorem 2.4.** Let $\mathcal{F}$ be a transversely orientable foliation of class $C^2$ and codimension-one. Let $U \subset M$ be a connected, nonempty, open, $\mathcal{F}$-saturated set.

1) Suppose that $(U, \mathcal{F}|_U)$ is without holonomy. Then there is a $C^0$ flow $\Phi: \mathbb{R} \times \hat{U} \to \hat{U}$ that fixes the points of $\partial\hat{U}$, carries leaves diffeomorphically to leaves and is transitive on the set of leaves of $\mathcal{F}|_U$. Furthermore, $\mathcal{F}|_U$ admits
a bounded transverse-invariant nonatomic measure $\mu$ of full support that assigns to each transverse arc $\{\Phi_t(p)\mid t \in (a, b)\}$ the measure $b - a$.

2) Conversely, if $(U, \mathcal{F}|_U)$ admits a bounded nonatomic transverse invariant measure $\mu$ of full support, then $\mathcal{F}|_U$ is without holonomy and there exists a continuous flow $\Phi$ that carries leaves diffeomorphically to leaves, transitively on the set of leaves in $\mathcal{F}|_U$, and fixing the leaves on the boundary. The flow $\Phi$ is related to $\mu$ in the way described in Part 1) of this theorem.

Proof. This is a special case of [3, 9.2.1]. The transverse-invariant measure is described in the proof given in the reference. See also [12]. $\square$

Finally, we mention the following basic fact. For simplicity, we state it for a boundaryless $M$ and a $C^1$ foliation, although these conditions can be weakened. (Cf. [3, 6.1.1].)

Theorem 2.5 (Haefliger [8]). Let $\mathcal{F}$ be a codimension-one $C^1$ foliation of a compact connected boundaryless manifold $M$. Then the union of the compact leaves of $\mathcal{F}$ is a compact subset of $M$.

3. Proof of Theorem 1.1.

Suppose that a $\Gamma$-action satisfying the assumptions of Theorem 1.1 has been fixed. We may assume, by passing to a finite-index subgroup of $\Gamma$, that the $\Gamma$-action preserves the transverse orientation for $\mathcal{F}$.

The $\Gamma$-action will be said to be fixing if there exists a finite-index subgroup $\Gamma'$ of $\Gamma$ and a compact leaf $L$ such that every element of $\Gamma'$ sends $L$ to itself.

A nonempty $\mathcal{F}$-saturated subset $\mathcal{P}$ of $M$ will be called $\mathcal{F}$-perfect if for every differentiable curve $\alpha : (-a, a) \to M$, $a > 0$, transverse to $\mathcal{F}$, the intersection of $\mathcal{P}$ with the image of $\alpha$ is a perfect set.

Lemma 3.1. If $\mathcal{F}$ has a compact leaf and the $\Gamma$-action is not fixing, then there exists a compact, $\mathcal{F}$-saturated, $\mathcal{F}$-perfect, $\Gamma$-invariant set $\mathcal{P} \subset M$, that is the union of compact, mutually homeomorphic leaves.

Proof. Let $\mathcal{C}$ denote the union of all compact leaves of $\mathcal{F}$ of a same homeomorphism type. A nonempty set of this kind exists since $\mathcal{F}$ has a compact leaf. It is also clear that $\mathcal{C}$ is invariant under every automorphism of $\mathcal{F}$. Furthermore, by a theorem of Haefliger, [3, 6.1.1], $\mathcal{C}$ is compact. Let

$$\mathcal{A} = \{\mathcal{A} \subset \mathcal{C} \mid \mathcal{A} \text{ is compact, nonempty, } \mathcal{F}\text{-saturated, and } \Gamma\text{-invariant}\}.$$ 

By Zorn’s lemma, $\mathcal{A}$ has an element $A$ that is minimal under inclusion.

We define the derived set $A'$ of $A$ as the subset of $A$ comprised of the union of all leaves $L \subset A$ whose points are limits of sequences in the complement of $L$ in $A$. If $A$ is the union of finitely many leaves, a finite-index subgroup
of \( \Gamma \) would send each of those finitely many leaves to itself, contradicting the assumption that the \( \Gamma \)-action is not fixing. Therefore \( A' \) is nonempty. It is easy to see that \( A' \in A \), so \( A = A' \) by the minimality of \( A \). Therefore, \( P := A \) satisfies the properties required in the lemma.

\[\square\]

**Proposition 3.2.** If \( P = M \), then the foliation fibers over the circle.

**Proof.** This is immediate from Theorem 2.3 and the easy fact that \( \mathcal{F} \) is, in this case, without holonomy. \( \square \)

The connected components of \( M - P \) will be called the **gaps** of \( P \).

**Lemma 3.3.** Suppose that \( P \) is a proper subset of \( M \), and the \( \Gamma \)-action is not fixing. Then

1) each gap is bounded by two leaves of \( P \), and

2) there exists a finite open cover \( \{V_1, \ldots, V_k\} \) of \( M \) by foliated products such that each \( V_i \) is bounded by two leaves in \( P \).

**Proof.** We first remark that each connected component of \( W = M - P \) is a foliated product. In fact, by [3, 5.2.9], only finitely many connected components of \( W \) fail to be foliated products. Suppose that there are connected components of \( W \) which are not foliated products and denote them by \( W_1, \ldots, W_l \). Since any homeomorphism \( \gamma \in \Gamma \) must send each \( W_i \) into some (possibly the same) \( W_j \), one obtains a homomorphism of \( \Gamma \) into the group of permutations of \( l \) symbols, from which it follows that some finite-index subgroup of \( \Gamma \) sends each \( W_i \) to itself. In particular, this subgroup would permute the boundary components of a \( W_i \). By [3, 5.2.5] the boundary of a connected \( \mathcal{F} \)-saturated open set consists of the union of a finite number of leaves, which in this case must be elements of \( P \). But then a finite-index subgroup of \( \Gamma \) would send one leaf in \( P \) to itself, contradicting the assumption that the action is not fixing.

A leaf in \( P \) will be called a **border leaf** of \( P \) if it is a component of the boundary of a gap. If \( P \) is not all of \( M \), there must be a countable infinity of gaps, since otherwise \( P \) would be contained in the union of the finitely many boundary leaves of a finite number of connected \( \mathcal{F} \)-saturated open sets. Each border leaf \( L \) of \( P \) is a boundary component of a gap and on the side of \( L \) opposite the gap a sequence of leaves in \( P \) accumulates on \( L \).

If a leaf \( L \) of \( P \) does not bound a gap, then \( L \) is a limit of sequences of leaves in \( P \) on both of its sides, so that [3, 5.3.4] (due to Dippolito [5]) immediately yields a foliated product neighborhood of \( L \).

We claim that boundary of each gap of \( P \) consists of two (distinct) leaves in \( P \), and that the closure of each gap is contained in a foliated product bounded by two leaves in \( P \). The interiors of these foliated products together with the interiors of the foliated products of the previous paragraph form an open cover for \( M \). Since \( M \) is compact, we can extract a finite open cover.
All that is left is to prove the claim. Let \( W \) denote a gap of \( \mathcal{P} \). We have seen that it is a foliated product. Denote by \( B \) the base manifold. The boundary of \( W \) consists of two (distinct) leaves in \( \mathcal{P} \). (There are not more than two leaves by Theorem 2.1. If a single leaf of \( \mathcal{P} \) bounded \( W \) on both sides, this would be an isolated leaf, which is not the case.) The boundary leaves are homeomorphic to \( B \) (notice that the boundary leaves of \( \widehat{W} \) are homeomorphic to the base of the fibration on \( \widehat{W} \)) and \( \widehat{W} \) maps bijectively onto the closure of \( W \) under the map \( \widehat{\iota} \). Since the boundary leaves \( L_1 \) and \( L_2 \) are compact, it is possible to find neighborhoods \( U_1 \) and \( U_2 \) of \( L_1 \) and \( L_2 \), respectively, such that \( U_i \mid_{L_i} \) is a trivial bundle over \( L_i \), \( i = 1, 2 \), not necessarily \( \mathcal{F} \)-saturated. To obtain \( \mathcal{F} \)-saturated \( U_i \), one applies [3, 5.3.4]. The union of \( W, U_1 \) and \( U_2 \) (for sufficiently small \( U_i \)) gives the desired neighborhood. □

For \( x \in \mathcal{P} \), let \( L_x \) be the leaf of \( \mathcal{F} \) that contains \( x \). Define an equivalence relation \( \sim \) on \( \mathcal{P} \) by specifying that \( x \sim y \) if either \( L_x \cup L_y \) is the boundary of a gap of \( \mathcal{P} \), or \( L_x = L_y \). (In particular, if \( \mathcal{P} = M \), then \( x \sim y \) if and only if \( L_x = L_y \).) Note that each equivalence class is either a leaf or the union of two leaves.

**Lemma 3.4.** If \( \Gamma \) is as in Theorem 1.1 and \( \mathcal{F} \) has a compact leaf, but the \( \Gamma \)-action is not fixing, then \( \mathcal{P}/\sim \) is homeomorphic to \( S^1 \).

**Proof.** We may assume \( \mathcal{P} \) is a proper subset of \( M \); otherwise, the desired conclusion follows from Proposition 3.2. It is immediate from Lemma 3.3 (and the “waterfall construction” described in [3, 3.3.7]) that there exists a closed transversal, \( \alpha \), of \( \mathcal{F} \) that intersects each leaf of \( \mathcal{P} \) exactly once. The intersection \( \alpha \cap \mathcal{P} \) is a perfect set in the embedded circle \( \alpha \), and the saturations of the gaps of this perfect set are the gaps of \( \mathcal{P} \). Thus, the desired conclusion follows from the elementary observation that, by identifying the two endpoints of each of the gaps of \( \alpha \cap \mathcal{P} \) to a single point, we obtain a quotient that is homeomorphic to a circle. □

Suppose the \( \Gamma \)-action is not fixing, and let \( \mathcal{P} \) be as in Lemma 3.1. The action of \( \Gamma \) on \( \mathcal{P} \) factors through to an action of \( \Gamma \) by homeomorphisms of \( \mathcal{P}/\sim \). Now Lemma 3.4 implies that \( \mathcal{P}/\sim \) is homeomorphic to \( S^1 \), so, by assumption, \( \Gamma \) must have a finite orbit on \( \mathcal{P}/\sim \). This finite orbit yields a \( \Gamma \)-invariant, finite collection of compact leaves in \( \mathcal{P} \). Then some finite-index subgroup of \( \Gamma \) fixes each of these compact leaves. This proves Theorem 1.1.

**4. Proof of Corollary 1.2.**

Suppose that \( \mathcal{F} \) admits a transverse invariant measure.

If \( \mathcal{F} \) has a compact leaf, then Theorem 1.1 implies that some finite-index subgroup \( \Gamma' \) of \( \Gamma \) fixes some compact leaf \( L \). Then \( \Gamma' \) fixes the atomic
measure \( \mu \) supported on the single leaf \( L \), so the conclusion of Corollary 1.2 holds.

Thus, we may assume that no leaf of \( F \) is compact. Therefore, from Theorems 2.2 and 2.3, it follows that \( F \) is a minimal foliation (that is, every leaf is dense in \( M \)). Then Corollary 4.2 below completes the proof of Corollary 1.2.

Corollary 4.2 is stated in greater generality than needed for the proof of Corollary 1.2 because it will be used in the given form for the proof of Theorem 1.3.

The next observation follows from the proof of [10, Thm. X.2.3.3, p. 272].

**Proposition 4.1.** Let \( U \) be a connected, \( \Gamma \)-invariant, \( F \)-saturated open set. We suppose that the boundary of \( U \) is either empty or consists of finitely many \( \Gamma \)-invariant compact leaves. We also suppose that each leaf of \( F \) in \( U \) is dense in \( U \). Let \( \mu \) be a bounded transverse invariant measure on \( U \).

If \( \mu' \) is another bounded transverse invariant measure on \( U \), then \( \mu' \) is a scalar multiple of \( \mu \).

**Corollary 4.2.** Let \( U \) be a connected, \( \Gamma \)-invariant, \( F \)-saturated open set. We suppose that the boundary of \( U \) is either empty or consists of finitely many \( \Gamma \)-invariant compact leaves. We also suppose that each leaf of \( F \) in \( U \) is dense in \( U \). Let \( \mu \) be a bounded transverse invariant measure on \( U \). Then

1) \( \mu \) is \( \Gamma \)-invariant; and
2) \([\Gamma, \Gamma]\) is a finite-index subgroup of \( \Gamma \) that fixes each leaf in \( U \).

**Proof.** (1) For each \( \gamma \in \Gamma \), Proposition 4.1 implies there is some \( c(\gamma) \in \mathbb{R}^+ \), such that \( \gamma_\ast \mu = c(\gamma) \mu \). It is easy to see that \( c: \Gamma \to \mathbb{R}^+ \) is a homomorphism, so, because \( \mathbb{R}^+ \) is abelian and has no nontrivial finite subgroups, Lemma 4.3 implies \( c(\Gamma) = 1 \). Thus, \( \mu \) is \( \Gamma \)-invariant.

(2) We use the notations of Theorem 2.4, where the Sacksteder flow \( \Phi_t \) is defined. Integration of \( \mu \) over closed curves representing elements of \( \pi_1(U) \) yields a homomorphism \( \rho: \pi_1(U) \to \mathbb{R} \) whose image group, \( P(\mu) \), is called the group of periods of \( \mu \) [3, 9.3.4]. This is a finitely generated abelian subgroup of \( \mathbb{R} \). The group of periods can be characterized by the following property ([3, 9.3.6]): \( \Phi_t \) sends every leaf to itself exactly when \( t \in P(\mu) \).

Define for each \( p \in U \) and each \( \gamma \in \Gamma \) a class \([t] \in \mathbb{R}/P(\mu) \) where \( t \) is any real number such that \( \Phi_t(p) \) lies in the leaf of \( \gamma(p) \). Then the correspondence \( \gamma \mapsto [t] \) gives a well-defined homomorphism, \( h \) from \( \Gamma \) into \( \mathbb{R}/P(\mu) \). Because \( \mathbb{R}/P(\mu) \) is abelian, we know that \([\Gamma, \Gamma]\) is in the kernel of \( h \), so \([\Gamma, \Gamma]\) sends every leaf to itself. Furthermore, Lemma 4.3 below asserts that \([\Gamma, \Gamma]\) is a finite-index subgroup of \( \Gamma \).

The following well-known observation is easy to prove.
Lemma 4.3. If each homomorphism of a finitely generated group $\Gamma$ into the group of homeomorphisms of the circle has a finite orbit, then $A := \Gamma/[\Gamma,\Gamma]$ is a finite group.

5. Proof of Theorem 1.3.

The following lemma is a slight generalization of the Thurston Stability Theorem [15]:

- We only assume the action of $\Gamma$ is germinal, rather than being globally defined; and
- we do not assume that each element of $\Gamma$ is defined in a neighborhood of 0, but only on a set $X$ that accumulates at 0.

Thurston’s original proof (and many others, such as [11]) can easily be generalized to this setting.

Lemma 5.1 (Thurston, cf. [15]). Suppose $\Gamma$ is a finitely generated group, $X$ is a compact subset of $[0,1]$ that accumulates at 0, and, for each $\gamma \in \Gamma$, we have a $C^1$ diffeomorphism $\phi_\gamma : [0,a_\gamma) \to [0,b_\gamma)$, for positive constants $a_\gamma$ and $b_\gamma$. Assume

1) $\Gamma/[\Gamma,\Gamma]$ is finite; and
2) for each $\gamma_1, \gamma_2 \in \Gamma$, there exists $c > 0$, such that $\phi_{\gamma_1 \gamma_2}|[0,c]\cap X = \phi_{\gamma_1}|[0,c]\cap X$.

Then there exists $a > 0$, such that, for every $\gamma \in \Gamma$ and $x \in [0,a) \cap X$, we have $\phi_\gamma(x) = x$.

All the assumptions of Theorem 1.3 are in force from now on.

Lemma 5.2. Let $\Gamma'$ be a finite-index subgroup of $\Gamma$ and suppose that $W$ is a connected, open, $\Gamma'$-invariant, $\mathcal{F}$-saturated subset of $M$ whose boundary components are $\Gamma'$-invariant compact leaves. If the compact leaves accumulate on a boundary component $L$ of $\hat{W}$, then all the compact leaves in some foliated neighborhood of $L$ in $\hat{W}$ are also $\Gamma'$-invariant.

Proof. Let $U$ be a connected, $\hat{\mathcal{F}}$-saturated neighborhood of $L$ in $\hat{W}$. We may assume $U$ is so small that there is a complete $C^1$ transversal $\alpha : [0,1] \to U$ of $U$, such that each compact leaf in $U$ meets $\alpha$ exactly once. We may assume $\alpha(0) \in L$.

As the $\Gamma$-action and the foliation (hence the local holonomy maps) are $C^1$, we can construct, for each $\gamma \in \Gamma'$, a $C^1$ diffeomorphism $\phi_\gamma : [0,a_1] \to [0,a_2]$, for positive constants $a_i$, such that $\alpha(\phi_\gamma(t))$ is on the same leaf as $\gamma(\alpha(t))$, for each $t \in [0,a_1]$. Let

$$X = \{x \in [0,1] \mid \alpha(x) \text{ is on a compact leaf}\}.$$

For each $\gamma_1, \gamma_2 \in \Gamma$, there exists $c > 0$, such that $\alpha(\phi_{\gamma_1 \gamma_2}(t))$ is on the same leaf as $\alpha(\phi_{\gamma_1}(\phi_{\gamma_2}(t)))$, for each $t \in [0,c)$. Then, for $x \in X \cap [0,c)$,
we must have $\phi_{\gamma_1 \gamma_2}(x) = \phi_{\gamma_1}(\phi_{\gamma_2}(x))$. Also, from Lemma 4.3, we know that $\Gamma'/[\Gamma', \Gamma']$ is finite. Therefore, Lemma 5.1 applies, so we conclude that there is an interval $[0, a)$, such that $\alpha(x)$ is on the same leaf as $\gamma(\alpha(x))$, for all $x \in X \cap [0, a)$. So $\Gamma'$ fixes the compact leaf containing $\alpha(x)$, for each $x \in X \cap [0, a)$.

\begin{proof}
By a standard averaging procedure $H$ can be assumed to act by isometries of a metric on the circle. Therefore the action is conjugate to a homomorphism of $H$ into the group of rotations of the circle, which is an abelian group. It follows that the commutator subgroup must act trivially.
\end{proof}

\begin{lemma}
Let $\Gamma'$ be a finite-index subgroup of $\Gamma$ and suppose that $W \subset M$ is a connected, open, $\Gamma'$-invariant, $\mathcal{F}$-saturated subset, such that the boundary components of $W$ are $\Gamma'$-invariant compact leaves, and every compact leaf in $W$ is $\Gamma'$-invariant. Then all leaves in $W$ are $[\Gamma', \Gamma']$-invariant.
\end{lemma}

\begin{proof}
From Theorem 2.5 and the hypothesis about $W$ it follows that the complement in $W$ of the set of compact leaves is a countable union of connected, $\Gamma'$-invariant, $\mathcal{F}$-saturated sets $U_i$, $i = 1, 2, \ldots$, such that the boundary of each $U_i$ is the union of at most 2 compact, $\Gamma'$-invariant leaves. Since there are no compact leaves in each of the $U_i$, each leaf of $\mathcal{F}|_{U_i}$ is, by assumption, without holonomy. Theorem 2.3 implies that either

1) $\mathcal{F}|_{U_i}$ fibers over the circle; or
2) $(\hat{U}_i, \hat{F})$ fibers over $[0, 1]$; or
3) every leaf of $\mathcal{F}|_{U_i}$ is dense in $U_i$.

In Case 1, we obtain a $C^1$ action of $\Gamma'$ on the circle. This action has to be finite, that is, a finite-index subgroup of $\Gamma'$ must leave invariant each leaf in $U_i$. From Lemma 5.3, we know that this normal subgroup contains the commutator subgroup $[\Gamma', \Gamma']$.

In Case 2, we obtain a $C^1$ action of $\Gamma'$ on $[0, 1]$. Lemma 5.1 implies that this action is trivial.

In Case 3, we can apply Theorem 2.4(1) and Corollary 4.2(2) to conclude that each leaf in $U_i$ is $\Gamma'$-invariant.

Therefore, $[\Gamma', \Gamma']$ fixes every leaf of $W$.
\end{proof}

We now conclude the proof of Theorem 1.3. It suffices to show that there is a finite-index subgroup $\Gamma'$ of $\Gamma$, such that $\Gamma'$ fixes every compact leaf, for then Lemma 5.4 shows that $[\Gamma', \Gamma']$ fixes every leaf (and Lemma 4.3 implies that $[\Gamma', \Gamma']$ has finite index in $\Gamma$).
We may assume that $\mathcal{F}$ has a compact leaf. Then Theorem 1.1 states that there is a finite-index subgroup $\Gamma'$ of $\Gamma$, and a compact leaf $L_0$ of $\mathcal{F}$, such that $\Gamma'$ fixes $L_0$. Let $W_0$ be the complement of the union of all the compact leaves of $\mathcal{F}$. From Theorem 2.1, we know that only finitely many components $U_1, U_2, \ldots, U_k$ of $W_0$ fail to be foliated products, and that each of these components has only finitely many boundary leaves, so, replacing $\Gamma'$ with a finite-index subgroup, we may assume that
\begin{equation}
\Gamma' \text{ fixes each } U_i, \text{ and its boundary components.}
\end{equation}

To complete the proof, we will show that $\Gamma'$ fixes every compact leaf. Let $W$ be a component of the complement of the union of the $\Gamma'$-invariant compact leaves of $\mathcal{F}$. It suffices to show that every compact leaf in $W$ is $\Gamma'$-invariant.

Let $L$ be any leaf on the boundary of $W$. By definition of $W$, we know that $L$ is $\Gamma'$-invariant.

If the compact leaves in $W$ accumulate on $L$, then Lemma 5.2 implies that every compact leaf in $W$ is $\Gamma'$-invariant.

If the compact leaves in $W$ do not accumulate on $L$, then there is a component $U$ of $W_0$ such that $L$ is part of the boundary of $U$, and $U$ is contained in $W$. We will show that $U = W$, so there are no compact leaves in $W$. If $U$ is not a foliated product, then from (5.5), we know that $\Gamma'$ fixes each boundary component of $U$, so $U = W$ as desired. Now suppose that $U$ is a foliated product. Then the boundary of $U$ consists of at most two compact leaves. One of these boundary leaves is $L$, which is known to be $\Gamma'$-invariant. Thus the second boundary leaf (if it exists) must also be $\Gamma'$-invariant, so $U = W$, as desired. \qed

Acknowledgement. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences (Cambridge, U.K.) for providing a congenial environment to carry out this research; and S. Hurder and L. Conlon, for a number of helpful discussions and guidance in the literature of codimension-one foliations. D.W. was partially supported by a grant from the National Science Foundation (DMS-9801136).

References


AN ASYMPTOTIC DIMENSION FOR METRIC SPACES, 
AND THE 0-TH NOVIKOV–SHUBIN INVARIANT

Daniele Guido and Tommaso Isola

A nonnegative number $d_\infty$, called asymptotic dimension, is
associated with any metric space. Such number detects the as-
ymptotic properties of the space (being zero on bounded met-
ric spaces), fulfills the properties of a dimension, and is invari-
ant under rough isometries. It is then shown that for a class
of open manifolds with bounded geometry the asymptotic di-
menion coincides with the 0-th Novikov–Shubin number $\alpha_0$
defined in a previous paper [D. Guido, T. Isola, J. Funct.
Analysis, 176 (2000)]. Thus the dimensional interpretation of
$\alpha_0$ given in the mentioned paper in the framework of noncom-
mutative geometry is established on metrics grounds. Since
the asymptotic dimension of a covering manifold coincides
with the polynomial growth of its covering group, the stated
equality generalises to open manifolds a result by Varopoulos.

0. Introduction.

In a recent paper [14], we extended the notion of Novikov-Shubin numbers
to amenable open manifolds and showed that they have a dimensional inter-
pretation in the framework of noncommutative geometry. Here we introduce
an asymptotic dimension for metric spaces, which is an asymptotic counter-
part of the Kolmogorov dimension [15], and show that for a class of open
manifolds it coincides with the 0-th Novikov-Shubin number.

The dimension introduced by Kolmogorov and Tihomirov, also called
box dimension, “corresponds to the possibility of characterizing the “mas-
siveness” of sets in metric spaces by the help of the order of growth of the
number of elements of their most economical $\varepsilon$-coverings, as $\varepsilon \to 0^+$ [15].
For non-totally bounded sets, denoting by $n(r, R)$ the minimum number of
balls of radius $r$ necessary to cover a ball of radius $R$ (and given center), the
box dimension is the “order of infinite” of $n(r, R)$ when $r \to 0$ (for big $R$, and
often independently of $R$).

We then define the asymptotic dimension $d_\infty$ as the “order of infinite” of
$n(r, R)$ when $R \to \infty$ (for big $r$, and often independently of $r$),

$$d_\infty(X) = \lim_{r \to \infty} \lim_{R \to \infty} \frac{\log n(r, R)}{\log R},$$
show that $d_\infty$ is a dimension, namely $d_\infty(X \cup Y) = \max(d_\infty(X), d_\infty(Y))$ and $d_\infty(X \times Y) \leq d_\infty(X) + d_\infty(Y)$, and prove that it is invariant under rough isometries.

Finally we show that the asymptotic dimension of an open manifold with $C^\infty$-bounded geometry and satisfying an isoperimetric inequality introduced by Grigor’yan \cite{8} coincides with the 0-th Novikov-Shubin number $\alpha_0$ as defined in \cite{14}. On the one hand this strengthens the dimensional interpretation given in \cite{14}, and on the other it shows that the generalised limit procedure used in the definition of $\alpha_0$ does not affect the result. Moreover, the quasi-isometry invariance of the $\alpha_p$ proved in \cite{14} becomes rough isometry invariance for the case of $\alpha_0$.

Since the asymptotic dimension of a manifold with $C^\infty$-bounded geometry may be computed in terms of its volume growth, the equality between $\alpha_0$ and $d_\infty$ may be seen as a generalization of the result of Varopoulos \cite{21} for covering manifolds, namely the equality between $\alpha_0$ and the growth of the covering group.

We recall that the Novikov-Shubin numbers \cite{17} where introduced, after the definition by Atiyah \cite{2} of the $L^2$-Betti numbers in terms of the von Neumann trace of the covering group, as finer invariants of the spectral behaviour of the $p$-Laplacian near zero, and where shown to be homotopy invariant by Gromov and Shubin \cite{10}. It was observed by Roe \cite{18} that, when the covering group is amenable, the von Neumann trace of an operator may be computed as an average of its integral kernel on the manifold w.r.t. a suitable exhaustion. Hence this procedure may be extended to manifolds admitting an amenable exhaustion. We show that, for manifolds satisfying Grigor’yan isoperimetric inequality, an amenable exhaustion exists and is given by concentric balls of increasing radius.

In \cite{13} we showed that, from the operator algebraic point of view, the step from amenable coverings to amenable exhaustions corresponds to passing from a normal semifinite trace on a von Neumann algebra to a semicontinuous semifinite trace on a C*-algebra. The latter does not necessarily contain spectral projections, however, the spectral density function may be still defined, making use of the noncommutative Riemann integration developed in \cite{14}, hence Novikov-Shubin numbers are defined. Moreover, this definition coincides, at least for $\alpha_0$, with the definition given in terms of the trace of the heat kernel, which does not require Riemann integration, and which is used here. More precisely,

$$\alpha_0(M) := 2 \limsup_{t \to \infty} \frac{\log(\tau(e^{-t\Delta}))}{\log 1/t},$$

where $\tau$ is the (semincontinuous semifinite) trace associated with an amenable exhaustion of $M$. 

We now recall the dimensional interpretation of the Novikov-Shubin numbers given in [14] in the framework of noncommutative geometry.

On the one hand these numbers are defined in terms of the low frequency behaviour of the $p$-Laplacians, or the large time behaviour of the $p$-heat kernels, therefore they are the large scale counterpart of the spectral dimension, namely of the dimension as it is recovered by the Weyl asymptotics.

On the other hand, recall that in Alan Connes’ noncommutative geometry [4], a nontrivial singular trace, associated to some power of the resolvent of the Dirac operator, plays the role of integration over the noncommutative space, and such a power is the dimension of the space.

This is analogous of what happens in geometric measure theory where a dimension is the unique exponent of the diameter of a ball which gives, via Hausdorff procedure, a (possibly) nontrivial measure on the space.

Therefore it is in this context that Novikov-Shubin numbers are interpreted as asymptotic spectral dimensions, since it was shown in [14] that the operator $\Delta_p^{-1/2}$, raised to the power $\alpha_p$, is singularly traceable.

The identification of $\alpha_0$ with $d_\infty$ proved here puts on metric grounds the dimensional character of the 0-th Novikov-Shubin number.

Finally we study the relation of $d_\infty$ with the notion of (metric) asymptotic dimension for cylindrical ends given by Davies [6]. Such definition is given in terms of the volume growth of the end, therefore when the end has bounded geometry Davies’ asymptotic dimension coincides with ours. Indeed Davies requires the growth to be exactly polynomial, therefore, in contrast with ours, his dimension is not always defined. Davies also introduced a class of cylindrical ends called standard ends. We show that for standard ends with Davies’ asymptotic dimension $D$ the equality $d_\infty = D$ holds with no further assumptions. Making use of Davies standard ends one observes that $d_\infty$ for a open manifold may take any value in $[1, \infty]$. Also, we discuss examples of standard ends where the growth is not exactly polynomial.

Some of the results contained in the present paper have been announced in several international conferences. In particular we would like to thank the Erwin Schrödinger Institute in Vienna, where the first draft of this paper was completed, and the organisers of the “Spectral Geometry Program” for their kind invitation.

1. An asymptotic dimension for metric spaces.

The main purpose of this section is to introduce an asymptotic dimension for metric spaces. Other notions of asymptotic dimension have been considered by Gromov [9] (see also the papers by Yu [23] and Dranishnikov [7]). Davies [6] proposed a definition in the case of cylindrical ends of a Riemannian manifold. We shall give a definition of asymptotic dimension in the setting of metric dimension theory, based on the (local) Kolmogorov dimension [15]
and state its main properties. We compare our definition with Davies’ in the next Section, and discuss its relations with Gromov’s in Remark 1.18.

In the following, unless otherwise specified, \((X, \delta)\) will denote a metric space, \(B_X(x, R)\) the open ball in \(X\) with centre \(x\) and radius \(R\), \(n_r(\Omega)\) the least number of open balls of radius \(r\) which cover \(\Omega \subset X\), and \(\nu_r(\Omega)\) the largest number of disjoint open balls of radius \(r\) centered in \(\Omega\).

Kolmogorov and Tihomirov [15] defined a dimension for totally bounded metric spaces \(X\) as \(d_0(X) := \limsup_{r \to 0} \frac{\log n_r(X)}{\log(1/r)}\). A natural extension to all metric spaces is given by \(d_0(X) = \lim_{R \to \infty} \limsup_{r \to 0} \frac{\log n_r(B_X(x, R))}{\log(1/r)}\). It can be shown that \(d_0\) is independent of \(x\), is a dimension, namely it satisfies the properties of Theorem 1.8, and is invariant under bi-Lipschitz mappings.

The following definition gives a natural asymptotic counterpart of the dimension of Kolmogorov-Tihomirov.

**Definition 1.1.** Let \((X, \delta)\) be a metric space. We call \(d_\infty(X) := \lim_{R \to \infty} \limsup_{r \to 0} \frac{\log n_r(B_X(x, R))}{\log R}\), the asymptotic dimension of \(X\).

Let us remark that, as \(n_r(B_X(x, R))\) is nonincreasing in \(r\), the function \(r \mapsto \limsup_{R \to \infty} \frac{\log n_r(B_X(x, R))}{\log R}\) is nonincreasing too, so the \(\lim_{r \to \infty}\) exists.

**Proposition 1.2.** \(d_\infty(X)\) does not depend on \(x\).

**Proof.** Let \(x, y \in X\), and set \(\delta := \delta(x, y)\), so that \(B(x, R) \subset B(y, R + \delta) \subset B(x, R + 2\delta)\). This implies,

\[
\frac{\log n_r(B(x, R))}{\log R} \leq \frac{\log n_r(B(y, R + \delta))}{\log(R + \delta)} \leq \frac{\log n_r(B(x, R + 2\delta))}{\log(R + 2\delta)}
\]

so that, taking \(\limsup_{R \to \infty}\) and then \(\lim_{r \to \infty}\) we get the thesis. \(\square\)

The following lemma is proved in [15]. For the sake of completeness, we include a proof.

**Lemma 1.3.** \(n_r(\Omega) \geq \nu_r(\Omega) \geq n_{2r}(\Omega)\).

**Proof.** For the first inequality, let \(B_X(x_i, r)\), \(i = 1, \ldots, \nu_r(\Omega)\), be disjoint balls with centres in \(\Omega\). Then any \(r\)-ball of a covering of \(\Omega\) may contain at most one of the \(x_i\)'s. Indeed, if \(B_X(x, r) \ni x_i, x_j\), then \(B_X(x_i, r) \cap B_X(x_j, r) \ni \{x\} \neq \emptyset\), so that \(x_i = x_j\).
As for the second inequality, we need to prove it only when $\nu_r$ is finite. Let us assume that $\{B(x_i,r)\}_{i=1}^{\nu_r(\Omega)}$ are disjoint balls centered in $\Omega$ and observe that, for any $y \in \Omega$, $\delta(y, \bigcup_{i=1}^{\nu_r(\Omega)} B(x_i,r)) < r$, otherwise $B(y,r)$ would be disjoint from $\bigcup_{i=1}^{\nu_r(\Omega)} B(x_i,r)$, contradicting the maximality of $\nu_r$. So for all $y \in \Omega$ there is $j$ s.t. $\delta(y, B(x_j,r)) < r$, that is $\Omega \subset \bigcup_{i=1}^{\nu_r(\Omega)} B(x_i,2r)$, which implies the thesis.

**Proposition 1.4.**

$$d_\infty(X) = \lim_{r \to \infty} \limsup_{R \to \infty} \frac{\log \nu_r(B_X(x,R))}{\log R}.$$  

**Proof.** Follows easily from Lemma 1.3.  

**Definition 1.5.** Let $X, Y$ be metric spaces, $f : X \to Y$ is said to be a rough isometry if there are $a \geq 1, b, \varepsilon \geq 0$ s.t.

(i) $a^{-1} \delta_X(x_1, x_2) - b \leq \delta_Y(f(x_1), f(x_2)) \leq a \delta_X(x_1, x_2) + b$, for all $x_1, x_2 \in X$,

(ii) $\bigcup_{x \in X} B_Y(f(x), \varepsilon) = Y$.

**Lemma 1.6** ([3], Proposition 4.3). If $f : X \to Y$ is a rough isometry, there is a rough isometry $f^- : Y \to X$, with constants $a^-, \varepsilon^-$, s.t.

(i) $\delta_X(f^- \circ f(x), x) < c_X$, $x \in X$,

(ii) $\delta_Y(f \circ f^-(y), y) < c_Y$, $y \in Y$.

**Theorem 1.7.** Let $X, Y$ be metric spaces, and $f : X \to Y$ a rough isometry. Then $d_\infty(X) = d_\infty(Y)$.

**Proof.** Let $x_0 \in X$, then for all $x \in B_X(x_0, r)$ we have $\delta_Y(f(x), f(x_0)) \leq a \delta_X(x, x_0) + b \leq ar + b$ so that $f(B_X(x_0, r)) \subset B_Y(f(x_0), ar + b)$. Then, with $n := n_r(B_Y(f(x_0), aR + b))$,

$$f(B_X(x_0, R)) \subset \bigcup_{j=1}^{n} B_Y(y_j, r),$$

which implies

$$f^- \circ f(B_X(x_0, R)) \subset \bigcup_{j=1}^{n} f^-(B_Y(y_j, r)) \subset \bigcup_{j=1}^{n} B_X(f^-(y_j), ar + b^-).$$

Let $x \in B_X(x_0, R)$, and $j$ be s.t. $f^- \circ f(x) \in B_X(f^-(y_j), ar + b^-)$, then

$$\delta_X(x, f^-(y_j)) \leq \delta_X(x, f^- \circ f(x)) + \delta_X(f^- \circ f(x), f^-(y_j)) < c_X + ar + b^-,$$
so that
\[ B_X(x_0, R) \subset \bigcup_{j=1}^{n} B_X(f_j(y_j), ar + b - cX), \]
which implies \( n_{ar+b-cX}(B_X(x_0, R)) \leq n_r(B_Y(f(x_0), aR + b)). \)
Finally
\[
d_\infty(X) = \lim_{r \to \infty} \limsup_{R \to \infty} \frac{\log n_r(B_X(x_0, R))}{\log R} = \lim_{r \to \infty} \limsup_{R \to \infty} \frac{\log n_{ar+b-cX}(B_X(x_0, R))}{\log R} \leq \lim_{r \to \infty} \limsup_{R \to \infty} \frac{\log n_r(B_Y(f(x_0), aR + b))}{\log R} = \lim_{r \to \infty} \limsup_{R \to \infty} \frac{\log n_r(B_Y(f(x_0), R))}{\log R} = d_\infty(Y),
\]
and exchanging the roles of \( X \) and \( Y \) we get the thesis. \( \square \)

**Theorem 1.8.** The set function \( d_\infty \) is a dimension, namely it satisfies:

(i) If \( X \subset Y \) then \( d_\infty(X) \leq d_\infty(Y) \).

(ii) If \( X_1, X_2 \subset X \) then \( d_\infty(X_1 \cup X_2) = \max\{d_\infty(X_1), d_\infty(X_2)\} \).

(iii) If \( X \) and \( Y \) are metric spaces, then \( d_\infty(X \times Y) \leq d_\infty(X) + d_\infty(Y) \).

**Proof.** (i) Let \( x \in X \), then \( B_X(x, R) \subset B_Y(x, R) \) and the claim follows easily.

(ii) By part (i), we get \( d_\infty(X_1 \cup X_2) \geq \max\{d_\infty(X_1), d_\infty(X_2)\} \). So we need to prove the converse inequality, only when both \( X_1 \) and \( X_2 \) have finite asymptotic dimension. Let \( x_i \in X_i, \ i = 1, 2 \), and set \( \delta = \delta(x_1, x_2) \), \( a = d_\infty(X_1), \ b = d_\infty(X_2) \), with e.g., \( a \leq b < \infty \). Then \( B_X(x_1, R) \subset B_X(x_1, R) \cup B_X(x_2, R + \delta) \), therefore
\[
n_r(B_{X_1 \cup X_2}(x_1, R)) \leq n_r(B_{X_1}(x_1, R)) + n_r(B_{X_2}(x_2, R + \delta)).
\]

Besides, \( \forall \varepsilon > 0 \ \exists r_0 > 0 \ \text{s.t.} \ \forall r > r_0 \ \exists R_0 = R_0(\varepsilon, r) \ \text{s.t.} \ \forall R > R_0
\]
\[
n_r(B_{X_1}(x_1, R)) \leq R^{a+\varepsilon}
\]
\[
n_r(B_{X_2}(x_2, R + \delta)) \leq R^{b+\varepsilon},
\]
hence, by inequality (1.1),
\[
n_r(B_{X_1 \cup X_2}(x_1, R)) \leq R^{a+\varepsilon} + R^{b+\varepsilon} = R^{b+\varepsilon}(1 + R^{a-b}).
\]

Finally,
\[
\frac{\log n_r(B_{X_1 \cup X_2}(x_1, R))}{\log R} \leq b + \varepsilon + \frac{\log(1 + R^{a-b})}{\log R}.
\]
Taking the \( \limsup_{R \to \infty} \) and then the \( \lim_{r \to \infty} \) we get
\[
d_{\infty}(X_1 \cup X_2) \leq \max\{d_{\infty}(X_1), d_{\infty}(X_2)\} + \varepsilon
\]
and the thesis follows by the arbitrariness of \( \varepsilon \).

(iii) By Theorem 1.7, we may endow \( X \times Y \) with any metric roughly isometric to the product metric, e.g.,
\[
(1.2) \quad \delta_{X \times Y}((x_1, y_1), (x_2, y_2)) = \max\{\delta_X(x_1, x_2), \delta_Y(y_1, y_2)\}.
\]
Then, by \( n_r(B_{X \times Y}((x, y), R)) \leq n_r(B_X(x, R)) n_r(B_Y(y, R)) \), the thesis follows easily. \( \square \)

**Remark 1.9.** (a) In part (ii) of the previous theorem we considered \( X_1 \) and \( X_2 \) as metric subspaces of \( X \). If \( X \) is a Riemannian manifold and we endow the submanifolds \( X_1, X_2 \) with their geodesic metrics this property does not hold in general. A simple example is the following. Let \( f(t) := (t \cos t, t \sin t) \), \( g(t) := (-t \cos t, -t \sin t) \), \( t \geq 0 \) planar curves, and set \( X, Y \) for the closure in \( \mathbb{R}^2 \) of the two connected components of \( \mathbb{R}^2 \setminus (G_f \cup G_g) \), where \( G_f, G_g \) are the graphs of \( f, g \), and endow \( X, Y \) with the geodesic metric. Then \( X \) and \( Y \) are roughly-isometric to \( [0, \infty) \) (see below) so that \( d_{\infty}(X) = d_{\infty}(Y) = 1 \), while \( d_{\infty}(X \cup Y) = 2 \).

(b) The choice of the \( \limsup \) in Definition 1.1 is the only one compatible with the classical dimensional inequality stated in Theorem 1.8 (iii).

In what follows we show that when \( X \) is equipped with a suitable measure, the asymptotic dimension may be recovered in terms of the volume asymptotics for balls of increasing radius. This is analogous to the fact that the local dimension may be recovered in terms of the volume asymptotics for balls of infinitesimal radius.

**Definition 1.10.** A Borel measure \( \mu \) on \( (X, \delta) \) is said to be uniformly bounded if there are functions \( \beta_1, \beta_2 \), s.t. \( 0 < \beta_1(r) \leq \mu(B(x, r)) \leq \beta_2(r) \), for all \( x \in X \), \( r > 0 \).

That is
\[
\beta_1(r) := \inf_{x \in X} \mu(B(x, r)) > 0, \quad \text{and} \quad \beta_2(r) := \sup_{x \in X} \mu(B(x, r)) < \infty.
\]

**Proposition 1.11.** If \( (X, \delta) \) has a uniformly bounded measure, then every ball in \( X \) is totally bounded (so that if \( X \) is complete it is locally compact).

**Proof.** Indeed, if there is a ball \( B = B(x, R) \) which is not totally bounded, then there is \( r > 0 \) s.t. every \( r \)-net in \( B \) is infinite, so \( n_r(B) \) is infinite, and \( \nu_r(B) \) is infinite too. So that \( \beta_2(R) \geq \mu(B) \geq \sum_{i=1}^{\nu_r(B)} \mu(B(x_i, r)) \geq \beta_1(r) \nu_r(B) = \infty \), which is absurd. \( \square \)

**Proposition 1.12.** If \( \mu \) is a uniformly bounded Borel measure on \( X \) then
\[
d_{\infty}(X) = \limsup_{R \to \infty} \frac{\log \mu(B(x, R))}{\log R}.
\]
Proof. As $\bigcup_{i=1}^{\nu_r(B(x,R))} B(x_i, r) \subset B(x, R+r) \subset \bigcup_{j=1}^{n_r(B(x,R+r))} B(y_j, r)$, we get
\[
\beta_2(r) n_r(B(x, R + r)) \geq \mu(B(x, R + r)) \geq \beta_1(r) \nu_r(B(x, R)) \geq \beta_1(r) n_r(B(x, R)),
\]
by Lemma 1.3. So that
\[
\beta_1(r/2) \leq \frac{\mu(B(x, R + r/2))}{n_r(B(x, R))}, \quad \frac{\mu(B(x, R))}{n_r(B(x, R))} \leq \beta_2(r),
\]
and the thesis follows easily. \qed

Let us now consider the particular case of complete Riemannian manifolds.

Definition 1.13. Let $(M,g)$ be an $n$-dimensional complete Riemannian manifold. We say that $M$ has bounded geometry if it has positive injectivity radius, sectional curvature bounded from above, and Ricci curvature bounded from below.

Lemma 1.14. Let $M$ be an $n$-dimensional complete Riemannian manifold with bounded geometry. Then the Riemannian volume is a uniformly bounded measure.

Proof. We can assume, without loss of generality, that the sectional curvature is bounded from above by some positive constant $c_1$ and the Ricci curvature is bounded from below by $(n-1)c_2g$, with $c_2 < 0$. Then, denoting with $V_\delta(r)$ the volume of a ball of radius $r$ in a manifold of constant sectional curvature equal to $\delta$, we can set $\beta_1(r) := V_{c_1}(\min\{r,r_0\})$, and $\beta_2 := V_{c_2}(r)$, where $r_0 := \min\{\text{inj}(M), \frac{\pi}{\sqrt{c_2}}\}$, and inj$(M)$ is the injectivity radius of $M$. Then the result follows from ([3], p. 119 and 123). \qed

Proposition 1.15. Let $M, N$ be complete Riemannian manifolds.

(i) If $M$ is noncompact, then $d_{\infty}(M) \geq 1$.

(ii) If $M$ has bounded geometry, then
\[
d_{\infty}(M) = \limsup_{R \to \infty} \frac{\log \text{vol}(B_M(x,R))}{\log R}, \quad x \in M.
\]

(iii) If $M, N$ have bounded geometry, and admit asymptotic dimension in a strong sense, that is $d_{\infty}(M) = \lim_{R \to \infty} \frac{\log \text{vol}(B_M(x,R))}{\log R}, \quad x \in M$, and analogously for $N$, then
\[
d_{\infty}(M \times N) = d_{\infty}(M) + d_{\infty}(N).
\]

Proof. (i) Let us fix $x_0 \in M$ and $R > 0$, and consider $x_R \in M$ s.t. $\delta(x_0, x_R) = R$, which exists because $M$ is not compact, and let $\gamma : [0, 1] \to M$ be a minimizing geodesics between $\gamma(0) = x_0, \gamma(1) = x_R$. Clearly $\gamma([0, 1]) \subset$
ASYMPTOTIC DIMENSION

51

$B_M(x_0, R)$, hence, if $x_1, \ldots, x_k$ are the centres of a minimal covering by $r$-balls of $\gamma([0, 1])$, we have $k \leq n_r(B_M(x_0, R))$. Then

$$R = \text{length}(\gamma) \leq \sum_{i=1}^{k} \text{length}(\gamma \cap B_M(x_i, r)) \leq 2rk,$$

namely $n_r(B_M(x_0, R)) \geq \frac{R}{2rk}$, from which the thesis follows.

(ii) The result follows from Lemma 1.14 and Proposition 1.12.

(iii) As in the proof of Theorem 1.8 (iii), we may endow $M \times N$ with the metric (1.2). Then $\text{vol} (B_{M \times N}((x, y), R)) = \text{vol} (B_M(x, R)) \text{vol} (B_N(y, R))$, and we get

$$d_\infty(M \times N) = \lim_{R \to \infty} \frac{\log \text{vol} (B_{M \times N}((x, y), R))}{\log R} = \lim_{R \to \infty} \frac{\log \text{vol} (B_M(x, R))}{\log R} + \lim_{R \to \infty} \frac{\log \text{vol} (B_N(y, R))}{\log R} = d_\infty(M) + d_\infty(N).$$

\[ \square \]

Remark 1.16. (a) Conditions under which the inequality in Theorem 1.8 (iii) becomes an equality are often studied in the case of (local) dimension theory (cf. [1, 19]). The previous Proposition gives such a condition for the asymptotic dimension.

(b) As the asymptotic dimension is invariant under rough isometries, it is natural to substitute the continuous space with a coarse graining, which destroys the local structure, but preserves the large scale structure. Then (cf. [3], Theorem 4.9) if $M$ is a complete Riemannian manifold with Ricci curvature bounded from below, $M$ is roughly isometric to any of its discretizations, endowed with the combinatorial metric. Therefore $M$ has the same asymptotic dimension of any of its discretizations. In particular, when $M$ has a discrete group of isometries $\Gamma$ with a compact quotient, the asymptotic dimension of the manifold coincides with the asymptotic dimension of the group, hence with its growth (cf. [12]). Therefore, by a result of Varopoulos [21], it coincides with the 0-th Novikov-Shubin invariant. We will generalise this result in Section 3.

Let us conclude this Section with some examples. Other examples are contained in the next Section.

Example 1.17.

(i) $\mathbb{R}^n$ has asymptotic dimension $n$.

(ii) Set $X := \cup_{n \in \mathbb{Z}} \{(x, y) \in \mathbb{R}^2 : \delta((x, y), (n, 0)) < \frac{1}{2}\}$, endowed with the Euclidean metric, then $d_0(X) = 2, d_\infty(X) = 1$.

(iii) Set $X = \mathbb{Z}$ with the usual distance, then $d_0(X) = 0$, and $d_\infty(X) = 1$. 

(iv) Let $X$ be the unit ball in an infinite dimensional Banach space. Then $d_0(X) = +\infty$ while $d_\infty(X) = 0$.

(v) Let $X$ be the $\mathbb{Z}^\infty$-lattice determined by an orthonormal base in an infinite dimensional Hilbert space. Then $d_0(X) = 0$ while $d_\infty(X) = \infty$.

**Remark 1.18.** M. Gromov introduced a notion of “large scale dimension” for metric spaces: The asymptotic dimension of $X$ is the smallest integer $n$ such that, for any $r > 0$, there is a cover $\mathcal{U} = \{U_i\}$ of $X$ such that the diameters of the sets $U_i$ are bounded, and no ball of radius $r$ meets more than $n + 1$ of them.

Our asymptotic dimension can be very different from Gromov’s. For example hyperbolic space $H_n$ has finite Gromov dimension, but $d_\infty(H_n) = \infty$. Conversely, one can find a sequence of cylindrical ends with fixed $d_\infty$ and arbitrarily large Gromov dimension (cf. Corollary 2.4).

The two notions however, coincide on some very special spaces, such as cartesian products of $\mathbb{R}^n$ and a compact set with the product metric. Moreover both dimensions are in a sense “coarse”, since they are invariant under rough isometries.

Finally we remark that Gromov dimension is an asymptotic topological dimension, since it is a coarse analogue of the Lebesgue covering dimension, according to Dranishnikov [7]. Ours instead is an asymptotic metric dimension. Indeed it is an asymptotic counterpart of the Kolmogorov-Tihomirov metric dimension, and is a dimension in the context of noncommutative geometric measure theory [12, 14].

2. Asymptotic dimension of some cylindrical ends.

In this Section we want to compare our work with a work of Davies. In [6] he defines the asymptotic dimension of cylindrical ends of a Riemannian manifold $M$ as follows. Let $E \subset M$ be homeomorphic to $(1, \infty) \times A$, where $A$ is a compact Riemannian manifold. Set $\partial E := \{1\} \times A$, $E_r := \{x \in E : \delta(x, \partial E) < r\}$, where $\delta$ is the restriction of the metric in $M$. Then $E$ has asymptotic dimension $D$ if there is a positive constant $c$ s.t.

$$c^{-1}r^D \leq \text{vol}(E_r) \leq cr^D,$$

for all $r \geq 1$. Davies does not assume bounded geometry for $E$. If one does, the two definitions coincide, more precisely if an asymptotic dimension à la Davies exists, it coincides with ours.

**Proposition 2.1.** With the above notation, if the volume form on $E$ is a uniformly bounded measure (as in Definition 1.10), or in particular if $E$ has bounded geometry (as in Definition 1.13), and there is $D$ as in (2.1), then $d_\infty(E) = D$.

**Proof.** Choose $o \in E$, and set $\delta := \delta(o, \partial E)$, $\Delta := \text{diam}(\partial E)$. Then it is easy to prove that $E_{R-\delta-\Delta} \subset B_E(o, R) \subset E_{R+\delta}$. 

Then $c^{-1}(R - \delta - \Delta)^D \leq \operatorname{vol}(B_E(o, R)) \leq c(R + \delta)^D$, and from Proposition 1.12 the thesis follows. \hfill \Box

Motivated by (6), Example 16), let us set the following:

**Definition 2.2.** $E$ is a standard end of local dimension $N$ if it is homeomorphic to $(1, \infty) \times A$, endowed with the metric $ds^2 = dx^2 + f(x)^2d\omega^2$, and with the volume form $d\operatorname{vol} = f(x)^{N-1}dxd\operatorname{vol}_\omega$, where $(A, \omega)$ is an $(N-1)$-dimensional compact Riemannian manifold, and $f$ is an increasing smooth function.

**Proposition 2.3.** The volume form on a standard end $E$ is a uniformly bounded measure. Therefore, if $E$ satisfies equation (2.1), we get $d_\infty(E) = D$.

**Proof.** It is easy to show that, for $(x_0, p_0) \in E$, $r < x_0 - 1$,

$$[x_0 - r/2, x_0 + r/2] \times B_A\left(p_0, \frac{r/2}{f(x_0 + r/2)}\right) \subset B_E((x_0, p_0), r) \subset [x_0 - r, x_0 + r] \times B_A\left(p_0, \frac{r}{f(x_0 - r)}\right).$$

So that, with $V_X(x, r) := \operatorname{vol}(B_X(x, r))$,

$$\int_{x_0 - r/2}^{x_0 + r/2} f(x)^{N-1}dx \ A\left(p_0, \frac{r/2}{f(x_0 + r/2)}\right) \leq V_E((x_0, p_0), r) \leq \int_{x_0 - r}^{x_0 + r} f(x)^{N-1}dx \ A\left(p_0, \frac{r}{f(x_0 - r)}\right),$$

which implies

$$rf(x_0 - r/2)^{N-1} A\left(p_0, \frac{r/2}{f(x_0 + r/2)}\right) \leq V_E((x_0, p_0), r) \leq 2rf(x_0 + r)^{N-1} A\left(p_0, \frac{r}{f(x_0 - r)}\right).$$

As for $x_0 \to \infty$, $V_A\left(p_0, \frac{r}{f(x_0 - r)}\right) \sim c\left(\frac{r}{f(x_0 - r)}\right)^{N-1}$, and the same holds for $V_A\left(p_0, \frac{r/2}{f(x_0 + r/2)}\right)$, we get the thesis. \hfill \Box

**Corollary 2.4.** Let $E$ be the standard end given by $E := (1, \infty) \times S^{N-1}$, endowed with the metric $ds^2 = dr^2 + r^{2(D-1)/(N-1)}d\omega^2$, and with the volume form $d\operatorname{vol} = r^{D-1}drd^{N-1}\omega$ (6), Example 16). Then $d_\infty(E) = D$. 
Remark 2.5. Observe that $d_\infty(M)$ makes sense for any metric space, hence for any cylindrical end, while Davies’ asymptotic dimension does not. Indeed let $E := (1, \infty) \times S^1$, endowed with the metric $ds^2 = dr^2 + f(r)^2d\omega^2$, and with the volume form $d\text{vol} = f(r)d\text{dr}d\omega$, where $f(r) := \frac{d}{dr}(r^2 \log r)$. Then $d_\infty(E) = 2$, but $\text{vol}(E_r)$ does not satisfy one of the inequalities in (2.1).

Before closing this section we observe that the notion of standard end allows us to construct an example which shows that we could obtain quite different results if we used $\lim \inf$ instead of $\lim \sup$ in the definition of the asymptotic dimension. It makes use of the following function

$$f(x) = \begin{cases} \sqrt{x} & x \in [1, a_1] \\ 2 + b_{n-1} + c_{n-1} + (x - a_{2n-1}) & x \in [a_{2n-1}, a_{2n}] \\ 2 + b_{n-1} + c_n + \sqrt{x - a_{2n} + 1} & x \in [a_{2n}, a_{2n+1}] \end{cases}$$

where $a_0 := 0$, $a_n - a_{n-1} := 2^{2^n}$, $b_n := \sum_{k=1}^{n} \sqrt{2^{2k+1} + 1}$, $c_n := \sum_{k=1}^{n} (2^{2k} - 1)$, $n \geq 1$.

Proposition 2.6. Let $M$ be the Riemannian manifold obtained as a $C^\infty$ regularization of $C \cup \varphi E$, where $C := \{(x, y, z) \in \mathbb{R}^3 : (x - 1)^2 + y^2 + z^2 = 1, x \leq 1\}$, with the Euclidean metric, $E := [1, \infty) \times S^1$, endowed with the metric $ds^2 = dx^2 + f(x)^2d\omega^2$, and with the volume form $d\text{vol} = f(x)dxd\omega$, where $\varphi$ is the identification of $\{y^2 + z^2 = 1, x = 1\}$ with $\{1\} \times S^1$. Then the volume form is a uniformly bounded measure, $d_\infty(M) \geq 2$ but $d_\infty(M) \leq 3/2$, where $d_\infty(M) := \lim_{r \to \infty} \lim \inf_{R \to \infty} \frac{\log_2(V_{M}(x, R))}{\log R}$.

Proof. Set $o := (0, 0, 0) \in M$, then it is easy to see that, for $n \to \infty$, $a_n \sim 2^{2^n}$, $b_n \sim c_n \sim 2^{2^{2n}}$, and

$$\text{area}(B_M(o, a_{2n})) \sim \frac{1}{2} a_{2n}^2$$

$$\text{area}(B_M(o, a_{2n-1})) \sim \frac{5}{3} a_{2n-1}^{3/2}$$

so that, calculating the limit of $\frac{\log_2(\text{area}(B_M(o, R)))}{\log R}$ on the sequence $R = a_{2n}$ we get 2, while on the sequence $R = a_{2n-1}$ we get $3/2$. The thesis follows easily, using Proposition 1.12. \qed

3. The asymptotic dimension and the 0-th Novikov-Shubin invariant.

In this Section we show that, for a class of open manifolds of bounded geometry, the asymptotic dimension coincides with the 0-th Novikov-Shubin invariant defined in [14]. In all this Section $M$ denotes a manifold of $C^\infty$-bounded geometry, i.e., $M$ has positive injectivity radius, and the curvature tensor is bounded together with all its covariant derivatives. We assume moreover that $M$ satisfies:
Assumption 3.1. There are $A$, $C$, $C' > 0$ s.t. for all $x \in M$, $r > 0$,

$$V(x, 2r) \leq AV(x, r)$$

(3.1)

$$\frac{C}{V(x, \sqrt{r})} \leq p_t(x, x) \leq \frac{C'}{V(x, \sqrt{r})}$$

(3.2)

where $V(x, r) := \text{vol}(B(x, r))$ and $p_t(x, y)$ is the heat kernel on $M$.

Remark 3.2.

(i) Inequality (3.1) is introduced in [5] and called the volume doubling property.

(ii) A result of Coulhon-Grigor’yan ([5], Corollary 7.3) ([8], Proposition 5.2) states that the assumption above is equivalent to the following isoperimetric inequality introduced in [5]. There are $\alpha$, $\beta > 0$ s.t. for all $x \in M$, $r > 0$, and all regions $U \subset B(x, r)$,

$$\lambda_1(U) \geq \frac{\alpha}{r^2} \left( \frac{V(x, r)}{\text{vol}(U)} \right)^{\beta},$$

where $\lambda_1(U)$ is the first Dirichlet eigenvalue of $\Delta$ in $U$.

(iii) Assumption 3.1 is satisfied by all manifolds with positive Ricci curvature [16], and covering manifolds whose group of deck transformations has polynomial growth [20].

Proposition 3.3. Let $M$ be a complete Riemannian manifold of $C^\infty$-bounded geometry, satisfying Assumption 3.1.

Then $d_\infty(M) = \limsup_{t \to \infty} \frac{-2 \log p_t(x, x)}{\log t}$, for any $x \in M$.

Proof. Follows from Proposition 1.15 and estimates (3.2). □

Remark 3.4. The previous result shows that there are some connections between the asymptotic dimension of a manifold and the notion of dimension at infinity for semigroups (in our case the heat kernel semigroup) considered by Varopoulos (see [22]).

The volume doubling property is a weak form of polynomial growth condition, and still guarantees the finiteness of the asymptotic dimension (for manifolds of bounded geometry).

Proposition 3.5. Let $M$ be a complete Riemannian manifold of $C^\infty$-bounded geometry, and suppose the volume doubling property (3.1) holds. Then $M$ has finite asymptotic dimension.

Proof. Let $R > 1$, and $n \in \mathbb{N}$ be s.t. $2^{n-1} < R \leq 2^n$. Then $V(x, R) \leq V(x, 2^n) \leq A^n V(x, 1)$, so that

$$1 \leq \frac{V(x, R)}{V(x, 1)} \leq A^n \leq AR^{\log_2 A}.$$

Therefore $d_\infty(M) = \limsup_{R \to \infty} \frac{\log V(x, R)}{\log R} \leq \log_2 A$. □
Definition 3.6 ([18]). A regular exhaustion $K$ is an increasing sequence $\{K_n\}$ of compact subsets of $M$, whose union is $M$, and such that, for any $r > 0$

$$\lim_{n \to \infty} \frac{\text{vol}(\text{Pen}^+(K_n, r))}{\text{vol}(\text{Pen}^-(K_n, r))} = 1,$$

where we set $\text{Pen}^+(K, r) := \{x \in M : \delta(x, K) \leq r\}$, and $\text{Pen}^-(K, r) :=$ the closure of $M \setminus \text{Pen}^+(M \setminus K, r)$.

Proposition 3.7. Let $M$ be an open manifolds of $C^\infty$-bounded geometry and satisfying Assumption 3.1.

(i) There is $\gamma > 0$ s.t. for any $x, y \in M, r > 0$, if $B(x, r) \cap B(y, r) \neq \emptyset$, then

$$\gamma^{-1} \leq \frac{V(x, r)}{V(y, r)} \leq \gamma.$$

(ii) There is a sequence $n_k \in \mathbb{N}$ s.t. $\{B(x, n_k)\}$ is a regular exhaustion of $M$.

Proof. (i) The inequality easily follows by a result of Grigor’yan ([8], Proposition 6.2), where it is shown that Assumption 3.1 implies the existence of a constant $\gamma$ such that

$$\gamma^{-1} \left( \frac{R}{r} \right)^{\alpha_1} \leq \frac{V(x, R)}{V(y, r)} \leq \gamma \left( \frac{R}{r} \right)^{\alpha_2}$$

for some positive constants $\alpha_1, \alpha_2$, for any $R \geq r$, and $B(x, R) \cap B(y, r) \neq \emptyset$.

(ii) The statement follows from the fact that the volume doubling property implies subexponential (volume) growth, so that the result is contained in ([18], Proposition 6.2).

Recall from [13] that the $C^*$-algebra $A$ of almost local operators on $M$ is the norm closure of the finite propagation operators on $L^2(M, d\text{vol})$. Then:

Proposition 3.8 ([13]). There is on $A$ a lower semicontinuous semifinite trace $\text{Tr}_K$, which, on the heat semigroup, is given by the following formula,

$$\text{Tr}_K(e^{-t\Delta}) = \text{Lim}_\omega \frac{\int_{K_n} \text{tr}(p_t(x, x))d\text{vol}(x)}{\text{vol}(K_n)},$$

where $\text{Lim}_\omega$ is a generalized limit.

Remark 3.9.

(i) The above formula for the trace was considered by J. Roe in [18]. However, this formula does not describe a semicontinuous trace on the $C^*$-algebra of almost local operators. Therefore we introduced a semicontinuous semifinite regularization in [13].
(ii) $L^2$-Betti numbers for open manifolds have been introduced in [18], where it is shown that the $0$-th $L^2$-Betti number of a noncompact manifold is zero. For this reason it does not appear in the formula for $\alpha_0$ below.

By means of $\text{Tr}_K$ we defined the $0$-th Novikov-Shubin invariant as

$$\alpha_0(M, K) := 2 \limsup_{t \to \infty} \log \frac{\text{Tr}_K(e^{-t\Delta})}{\log 1/t}.$$

**Theorem 3.10.** Let $M$ be an open manifold of $C^\infty$-bounded geometry and satisfying Assumption 3.1, endowed with the regular exhaustion $K$ given by Proposition 3.7 (ii). Then the asymptotic dimension of $M$ coincides with the 0-th Novikov-Shubin invariant, namely $d_\infty(M) = \alpha_0(M, K)$. In particular $\alpha_0$ is independent of the limit procedure $\text{Lim}_\omega$.

**Proof.** First, from Equation (3.2) and Proposition 3.7 (i), we get

$$\frac{C\gamma^{-1}}{V(o, \sqrt{t})} \leq \int_{B(o,r)} \frac{p_t(x, x) \, d\text{vol}(x)}{V(o, r)} \leq \frac{C'}{V(o, r)} \leq \frac{C' \gamma}{V(o, \sqrt{t})}$$

therefore, by Proposition 3.8 we have,

$$\frac{C\gamma^{-1}}{V(o, \sqrt{t})} \leq \text{Tr}_K(e^{-t\Delta}) \leq \frac{C' \gamma}{V(o, \sqrt{t})}$$

hence, finally,

$$d_\infty(M) = 2 \limsup_{t \to \infty} \frac{\log(V(o, t))}{2 \log t} = 2 \limsup_{t \to \infty} \frac{\log(C' \gamma V(o, \sqrt{t})^{-1})}{\log \frac{1}{t}} \leq \alpha_0(M, K) \equiv 2 \limsup_{t \to \infty} \frac{\log \text{Tr}_K(e^{-t\Delta})}{\log \frac{1}{t}} \leq 2 \limsup_{t \to \infty} \frac{\log(C\gamma^{-1} V(o, \sqrt{t})^{-1})}{\log \frac{1}{t}} = 2 \limsup_{t \to \infty} \frac{\log(V(o, t))}{2 \log t} = d_\infty(M).$$

The thesis follows. \qed
Acknowledgement. We would like to thank D. Burghelea, M. Farber and L. Friedlander for conversations. We also thank I. Chavel, E.B. Davies, A. Grigor’yan, P. Li and L. Saloff-Coste for having suggested useful references on heat kernel estimates, and the referee for drawing our attention to the notion of Gromov asymptotic dimension.

References


ASYMPTOTIC DIMENSION


87-136, MR 89a:58102, Zbl 0657.58041.


[22] N.T. Varopoulos, L. Saloff-Coste and T. Couhlon, Analysis and Geometry on Groups,

[23] Guoliang Yu, The Novikov conjecture for groups with finite asymptotic dimension,

Received April 10, 2000 and revised October 16, 2000.

DIPARTIMENTO DI MATEMATICA
UNIVERSITÀ DELLA BASILICATA
I–85100 POTENZA, ITALY

Università di Roma “Tor Vergata”
I–00133 ROMA
ITALY
E-mail address: guido@mat.uniroma2.it

DIPARTIMENTO DI MATEMATICA
UNIVERSITÀ DI ROMA “Tor Vergata”
I–00133 ROMA, ITALY
E-mail address: isola@mat.uniroma2.it
HEEKAARD SPLITTINGS OF HAKEN MANIFOLDS HAVE BOUNDED DISTANCE

KEVIN HARTSHORN

Given a Heegaard splitting and an incompressible surface $S$ and a Heegaard splitting of an irreducible manifold, I shall use a generalization of Haken’s lemma proved by Kobayashi in order to define a pair of simple closed curves on the splitting surface such that each bounds a disc in one of the handlebodies of the splitting. By modifying the proof of Kobayashi’s lemma, I shall show that the sequence of boundary compressions used to isotope $S$ places a bound on the distance between these two simple closed curves in the complex of curves. This will then place a bound on the distance of the Heegaard splitting.

1. Introduction.

Let $\Sigma$ be a closed, orientable surface of genus $g \geq 2$. Associated with $\Sigma$ is a “curve complex” $\mathcal{C}(\Sigma)$ that has been defined by Harvey [4]. A vertex of this complex is an isotopy class of essential simple closed curves on $\Sigma$. Two vertices are joined by an edge if the corresponding isotopy classes have disjoint representatives\(^1\).

Notation 1.1. Throughout this paper, I will use the notation $c$ or $c_i$ to denote a simple closed curve on the surface $\Sigma$, the isotopy class of that curve, or the corresponding vertex in the complex of curves. Generally, the distinction will be unimportant.

On this complex, we define the distance $d$ between two vertices – between two essential simple closed curves on $\Sigma$ – to be the minimum number of edges traversed to get from one vertex to the other. Essentially, the distance is simply the metric on the 1-skeleton of $\mathcal{C}(\Sigma)$ gotten by letting each edge have length 1. Recently, Hempel [6] and Masur and Minsky [9] independently showed that the diameter of $\mathcal{C}(\Sigma)$ is infinite.

Suppose that $\Sigma$ is the splitting surface for a Heegaard decomposition of a 3-manifold $M$. That is, suppose $M$ is decomposed by handlebodies $H_1, H_2$ such that $M = H_1 \cup H_2$ and $H_1 \cap H_2 = \partial H_1 = \partial H_2 = \Sigma$. Let $\mathcal{C}_i(\Sigma) \subset \mathcal{C}(\Sigma)$

\(^1\)In general, $\mathcal{C}(\Sigma)$ is a $(3g-4)$-simplex such that each $k$ cell corresponds to a collection of $k + 1$ disjoint (and distinct) isotopy classes.
denote the subcomplex consisting of essential closed curves that bound a disc in $H_i$. Define the distance of the splitting to be

$$d(H_1, H_2) = \min \{ d(c_1, c_2) \mid c_i \in C_i(\Sigma) \}.$$  

Hempel showed (again in [6]) for any $D$, there is a manifold $M$ that has a Heegaard splitting $(H_1, H_2; \Sigma)$ with $d(H_1, H_2) > D$. Whether there is one manifold $M$ with splittings of arbitrarily large distance is not known (however, they would have to be non-Haken). Several results already known about Heegaard splittings can now be expressed in terms of this distance function:

- **Reducibility implies distance 0:** If $M$ contains an essential sphere, then for any Heegaard splitting of $M$, Haken’s lemma [2, 7] shows that $d(H_1, H_2) = 0$. Specifically, Haken’s lemma states that the sphere can be positioned so that it intersects $\Sigma$ exactly once and bounds a disc in each handlebody. Notice that stabilization creates a splitting of distance 0 regardless of the original splitting’s distance.

- **Strong irreducibility and distance:** Recall that a splitting $(H_1, H_2; \Sigma)$ is called weakly reducible if there is a pair $c_1, c_2 \subset \Sigma$ of essential simple closed curves with empty intersection such that $c_i$ is the boundary of a disc in $H_i$. A splitting that is not weakly reducible is called strongly irreducible. Thus weakly reducible is equivalent to $d(H_1, H_2) \leq 1$ and strongly irreducible is equivalent to $d(H_1, H_2) \geq 2$. A result of Casson and Gordon [1] shows that if $M$ contains a distance 1 splitting, then $M$ is Haken (a more specific description of the genus of the incompressible surface is given in [5]).

- **Conversely,** Hempel [6] proved that if $M$ contains an incompressible torus, then $d(H_1, H_2) \leq 2$ for any splitting $(H_1, H_2; \Sigma)$.

In this paper, I will show the following generalization of the above results:

**Theorem 1.2.** Let $M$ be a Haken 3-manifold containing an orientable incompressible surface of genus $g$. Then any Heegaard splitting of $M$ has distance at most $2g$.

In Section 2, I shall establish the main definitions and lemmas needed to prove the theorem. The theorem itself will be proven in Section 3.

### 2. Euler characteristics and $\partial$-compressions.

A surface $S$ embedded in a 3-manifold $M$ is called *compressible* if there is a simple closed curve $c \subset S$ such that $c$ bounds a disc in $M$, but not in $S$. Otherwise $S$ is called incompressible. A *Haken manifold* is an irreducible manifold containing an incompressible surface of genus $g \geq 1$. If $(S, \partial S) \subset (M, \partial M)$ is a surface in a manifold with boundary, then $S$ is called *boundary compressible* (boundary compressible) if there is an arc $\beta \subset \partial M$ and an
arc $\alpha \subset S$ essential in $S$ such that $\beta \cap S = \beta \cap \alpha = \partial \beta = \partial \alpha$ with the property that $\alpha \cup \beta$ bounds a disc in $M \setminus S$. Otherwise, $S$ is called $\partial$-incompressible (boundary incompressible). Note that if $S$ is $\partial$-comppressible, then we can “compress” along the arc $\alpha$ to simplify the surface.

Let $M$ be a closed, orientable, irreducible 3-manifold. Let $(H_1, H_2; \Sigma)$ be a Heegaard splitting for $M$. Suppose $S$ is a given closed, orientable, incompressible surface in $M$. Note that if $M$ contains an incompressible surface of positive genus, then the statement of Theorem 1.2 allows the distance of the splitting to be at least 2. Thus I shall assume that the splitting $(H_1, H_2; \Sigma)$ is strongly irreducible, and that the genus of $S$ is positive.

Kobayashi [8] proved the following generalization of Haken’s lemma:

**Lemma 2.1.** If $M = (H_1, H_2; \Sigma)$ is a strongly irreducible splitting of an irreducible manifold, then any closed, orientable, incompressible surface $S$ is ambient isotopic to a surface $S'$ such that:

1) $S'$ intersects $\Sigma$ transversely,
2) $S' \cap H_1$ contains exactly one disc component,
3) $S' \cap H_2$ does not contain any disc components.

Kobayashi used what Jaco [7] called an isotopy of type $A$ in order to prove this lemma. Because I will be using a similar isotopy to prove Theorem 1.2, I shall describe it again here.

Suppose that $S \cap H_1$ is $\partial$-comppressible in $H_1$. Let $\alpha \subset S \cap H_1$ be an essential compressing arc and $\beta \subset \partial H_1 = \Sigma$ be the arc such that $\alpha \cup \beta$ bounds a disc $D$ in $H_1$. To perform a $\partial$-comppression of $S$ from $H_1$ or an isotopy of type $A$, we isotope $S$ to $S'$ by “chopping” into $S \cap H_1$ and pushing $\alpha$ through the disc $D$ and across $\Sigma$ (see Figure 1). In a similar way we describe $\partial$-comppressions of $S$ from $H_2$.

**Lemma 2.2.** Suppose that $S'$ is the image of $S$ after a $\partial$-comppression of $S$ from $H_1$. Then $\chi(S' \cap H_1) = \chi(S \cap H_1) + 1$.

**Proof.** To see this, notice that the effect of the $\partial$-comppression on $S \cap H_1$ is removal of a 1-handle. This 1-handle is homotopic to a 1-cell with Euler characteristic -1, and so removal of this handle raises the Euler characteristic by 1. \hfill \Box

Recall that if a surface $S$ is decomposed by $S = X \cup Y$, then the Euler characteristic of $S$ is given by

$$\chi(S) = \chi(X) + \chi(Y) - \chi(X \cap Y).$$

In our current case, we decompose $S$ by $S \cap H_1$ and $S \cap H_2$ and notice that all the intersections are circles (with zero Euler characteristic). Thus we see

---

I find this terminology somewhat unintuitive and so I will usually refer to this isotopy as a boundary compression, or $\partial$-comppression.
that
\[ \chi(S) = \chi(S \cap H_1) + \chi(S \cap H_2). \]

With this in mind, we deduce the following corollary to Lemma 2.2:

**Corollary 2.3.** Suppose that \( S' \) is the image of \( S \) after an \( \partial \)-compression of \( S \) from \( H_1 \). Then \( \chi(S' \cap H_2) = \chi(S \cap H_2) - 1 \).

**Remark 2.4.** On the surface \( \Sigma \), the effect of the \( \partial \)-compression is surgery on a 1-submanifold:

Notice that there are two possibilities for the arc \( \beta \). Either it connects two essential simple closed curves on \( \Sigma \) and forms a single simple closed curve after the surgery; or it connects two points of the same simple closed curve on \( \Sigma \) to form two disjoint simple closed curves. The argument for this is given by Rubinstein and Scharlemann [11] to describe “saddle vertices”.

![Figure 1](image-url)
In order to prove Theorem 1.2, in the next section I will define an *elementary compression* and prove a few facts regarding the definition.

3. Elementary compressions.

When performing $\partial$-compressions on $S \cap H_1$ (or similarly for $S \cap H_2$), there are generally choices for the compression arc $\alpha$ which are non-helpful. Specifically, suppose there is an annular component of $S \cap H_1$ parallel to $\partial H_1$ (such components will be referred to as $\partial$-parallel annuli). Then the $\partial$-compression defined by a meridional arc $\alpha$ of this $\partial$-parallel annulus will leave a component of $S \cap H_1$ which is a disc parallel to $\Sigma$. In particular, the component of $S \cap \Sigma$ formed by this move will be inessential in $\Sigma$. As I am interested only in the components of $S \cap \Sigma$ which are essential in $\Sigma$, I would like to avoid this particular compression. However, the positioning of $S$ in $M$ may be such that the only possible $\partial$-compressions are along $\partial$-parallel annuli.

To deal with this, I define a two step operation. The first is the removal of $\partial$-parallel annular components of $S \cap H_1$. Note that this compression along the meridional arc followed by pushing the resulting disc through $\Sigma$ will constitute the “removal” of a $\partial$-parallel annulus (by pushing it into $H_2$). This operation of “removing” a $\partial$-parallel annular component of $S \cap H_1$ will be referred to as an *annular compression of $S$ from $H_1$*.

The second step is a specific sort of $\partial$-compression called an elementary compression. Define an essential compressing arc $\alpha$ of $S \cap H_1$ to be *strongly essential* if it is not the meridian of a $\partial$-parallel annular component of $S \cap H_1$. An *elementary compression of $S$ from $H_1$* is a $\partial$-compression of $S \cap H_1$ along a strongly essential arc $\alpha \subset S \cap H_1$.

Throughout this section, let $(H_1, H_2; \Sigma)$ be a strongly irreducible Heegaard splitting of $M$ and $S$ an incompressible surface intersecting $\Sigma$ transversely. In the following lemmas, I shall list several helpful properties of both elementary and annular compressions. While all the lemmas in this section will refer to elementary and annular compressions from $H_1$, notice that they are equally true if the roles of $H_1$ and $H_2$ are reversed.

**Lemma 3.1.** Suppose $S \cap H_1$ is incompressible in $H_1$. If $\chi(S \cap H_1) \leq 0$, then there is an elementary or an annular compression of $S$ from $H_1$.

*Proof.* Because $S$ is incompressible in $M$, it must intersect $\Sigma$ nontrivially, and thus $S \cap H_1 \neq \emptyset$. Further, the only incompressible, $\partial$-incompressible surfaces in a handlebody are discs. If all the components of $S \cap H_1$ were discs, then $\chi(S \cap H_1)$ would be positive. Thus, as $S \cap H_1$ is incompressible in $H_1$, there must be at least one component of $S \cap H_1$ which is $\partial$-compressible, and so there is either an elementary compression of $S$ from $H_1$ or an annular compression of $S$ from $H_1$. $\Box$
Remark 3.2. Because any incompressible surface has a nonempty intersection with the Heegaard splitting surface, we can assume that \( S \cap H_1 \neq \emptyset \).

In Lemma 2.2, we saw how an elementary compression (a special case of \( \partial \)-compression) affects the Euler characteristic of \( S \cap H_1 \). The following lemma provides the analogous statement for annular compressions.

Lemma 3.3. Suppose that \( S' \) is the image of an annular compression of \( S \) from \( H_1 \). Then \( \chi(S' \cap H_1) = \chi(S \cap H_1) \), and \( \chi(S' \cap H_2) = \chi(S \cap H_2) \).

Proof. Note that the net effect of an annular compression is simply the removal of an annulus. Removing an annulus has no net effect on \( \chi(S \cap H_1) \), as the Euler characteristic of an annulus is 0. The comments leading to Corollary 2.3 show that there is also no change to \( \chi(S \cap H_2) \). \( \square \)

Note that Lemmas 3.1 and 3.3 provide the following corollary:

Corollary 3.4. If \( S \cap H_1 \) is incompressible in \( H_1 \) and \( \chi(S \cap H_1) \leq 0 \), then (perhaps after some annular compressions) there is an elementary compression of \( S \) from \( H_1 \).

Lemma 3.5. Suppose that \( S \cap H_1 \) is incompressible in \( H_1 \). Then the image \( S' \) of an elementary compression of \( S \) from \( H_1 \) also has incompressible intersection with \( H_1 \).

Proof. Let the elementary compression of \( S \) be defined by the disc \( D \subset H_1 \) with \( \partial D = \alpha \cup \beta \), where \( \alpha \subset S \cap H_1 \) and \( \beta \subset \Sigma \) – as described in Section 2. Consider a small regular neighborhood of \( D \) in \( H_1 \), say \( D \times I \) (where \( I = [0,1] \)). This can be chosen so \((\partial D) \times I = (\alpha \cup \beta) \times I\), where \( \alpha \times I \subset S \cap H_1 \) and \( \beta \times I \subset \Sigma \). The effect of the elementary compression on \( S \cap H_1 \) is to replace the band \( \alpha \times I \) by the pair of discs \( D_0 = D \times \{0\} \) and \( D_1 = D \times \{1\} \). So if \( S' \cap H_1 \) is the image of \( S \cap H_1 \) after the elementary compression, we consider \( D_0 \) and \( D_1 \) as submanifolds of \( S \cap W_1 \).

Let \( c \subset S' \cap H_1 \) be a simple closed curve such that \( c = \partial \Delta \) for some disc \( \Delta \subset H_1 \setminus S' \). We can isotope \( c \) so that \( c \cap D_i = \emptyset \) for \( i = 0,1 \). Further, we can use an innermost disc argument (noting that \( H_1 \) is irreducible) to see that \( \Delta \cap D_i = \emptyset \) for \( i = 0,1 \). Thus by undoing the elementary compression, we can view \( \Delta \) as a compressing disc for \( S \). Since \( S \) is incompressible in \( H_1 \), the curve \( c \subset S \cap H_1 \) must bound a disc in \( S \). Because \( c \cap D_1 = \emptyset \), we see that the disc bounded by \( c \) in \( S \) must be disjoint from the strip \( \alpha \times I \). Thus we can conclude that \( c \) bounds a disc in \( S' \), and so \( S' \) is also incompressible. \( \square \)

Lemma 3.6. Suppose that \( S \cap H_1 \) is incompressible in \( H_1 \), that each component of \( S \cap \Sigma \) is essential in \( \Sigma \) and that \( S' \) is the image of an elementary compression of \( S \) from \( H_1 \). Then each component of \( S' \cap \Sigma \) is essential in \( \Sigma \).

Proof. Assume there is a component $c'$ of $S' \cap \Sigma$ which bounds a disc in $\Sigma$. By the hypothesis of the lemma, $c'$ must be a curve generated by the elementary compression of $S$. Consider the arc $\beta \subset \Sigma$ along which the boundary compression was defined. At this point we divide the proof into two cases, depending on whether $\beta$ joined together one or two components of $S \cap \Sigma$.

**Case 1.** Suppose that $\beta$ joined two points of the same component $c$ of $S \cap \Sigma$. Then $c$ is broken into two components $c', c''$ by the $\partial$-compression and we assume $c'$ bounds a disc $D$ in $\Sigma$. Suppose that $c' = \partial D$ as indicated below.

\[
\begin{array}{c}
\includegraphics[width=0.5\textwidth]{case1.png}
\end{array}
\]

Then we see that $c'' \subset D$. By the Jordan curve theorem, $c''$ must bound a disc in $D$, and thus in $\Sigma$. By looking at the preimage of this disc $D$ in $\Sigma$ before the elementary compression,

\[
\begin{array}{c}
\includegraphics[width=0.5\textwidth]{case1_preimage.png}
\end{array}
\]

we see that the simple closed curve $c$ must bound a disc, contradicting the hypothesis of the lemma.

On the other hand, suppose that $c'$ bounds the disc as indicated.

\[
\begin{array}{c}
\includegraphics[width=0.5\textwidth]{case2.png}
\end{array}
\]

Then I claim that the $\partial$-compressing took place along an inessential arc $\alpha \subset S \cap H_1$. To see this, notice that on the preimage we have the picture:

\[
\begin{array}{c}
\includegraphics[width=0.5\textwidth]{case2_preimage.png}
\end{array}
\]

The curve $\beta \cup \gamma$ is isotopic to $c'$ and thus bounds a disc in $\Sigma$. If $\alpha \subset S \cap H_1$ is the compressing arc for the elementary compression, then $\alpha \cup \beta$ bounds a disc in $H_1$. This implies that $\alpha \cup \gamma$ must also bound a disc in $H_1$. The simple closed curve $\alpha \cup \gamma$ can be homotoped to a simple closed curve in the interior of $S \cap H_1$. The incompressibility of $S$ in $H_1$ shows that this simple
closed curve is not essential in $S \cap H_1$, and thus $\alpha$ is not an essential arc, contradicting the definition of an elementary compression.

**Case 2.** Suppose that $\beta$ joined two different components of $S \cap \Sigma$. In this case, both “components” of the image (near the arc $\beta$)

![Diagram](image)

belong to the same curve $c'$. There are then two ways in which $c'$ can bound a disc:

![Diagram](image)

In the first case, the two curves comprising the preimage (before the elementary compression)

![Diagram](image)

must also have bounded discs in $\Sigma$, contradicting the initial hypothesis. In the second case, we see that the preimage

![Diagram](image)

must have bounded an annulus in $\Sigma$. In addition, we can push $c'$ into $H_1$ so it defines a compressing disc for $S' \cap H_1$. By Lemma 3.5, we see that $c'$ must be inessential in $S' \cap H_1$. Thus we see that rather than an elementary compression, this must have been the first step of an annular compression.

Thus we see that $S' \cap \Sigma$ consists only of simple closed curves which are essential in $\Sigma$. □

**Lemma 3.7.** Suppose that $S \cap H_1$ is incompressible in $H_1$ and each component of $S \cap \Sigma$ is essential in $\Sigma$. If $S'$ is the image of $S$ by an annular compression from $H_1$, then $S' \cap H_1$ is incompressible in $H_1$ and each component of $S' \cap \Sigma$ is essential in $\Sigma$. 
Proof. Because an annular compression simply deletes a component of $S \cap H_1$ (by moving it into $H_2$), all the remaining components will still be incompressible in $H_1$. Similarly, the effect of the annular compression on the splitting surface $\Sigma$ is to delete two components of $S \cap \Sigma$, and thus all the remaining components of $S \cap \Sigma$ will still be essential in $\Sigma$. □

Finally, we close the section with the lemmas that tie together the notion of distance and these elementary compressions:

**Lemma 3.8.** Suppose $S'$ is the image of an elementary compression of $S$. Let $c \subset S \cap \Sigma$ and $c' \subset S' \cap \Sigma$ be any choice of components that are essential in $\Sigma$. Then $d(c, c') \leq 1$.

Proof. Suppose $c$ is not a component of $S \cap \Sigma$ affected by the elementary compression. Then $d(c, c') \leq 1$, as either $c \simeq c'$ or $c \cap c' = \emptyset$. Similarly, if $c'$ is not in the image of the affected component(s) of the elementary compression, then $d(c, c') \leq 1$. So suppose $c$ is a component of $S \cap \Sigma$ affected by the elementary compression and $c'$ is in the image of this $\partial$-compression.

By considering the arc $\beta \subset \Sigma$ along which the elementary compression is defined,

```
     β
```

we can divide the proof into two cases, depending on whether $\beta$ connects one or two components of $S \cap \Sigma$.

**Case 1.** If $\beta$ connects two components $c_a, c_b$ of $S \cap \Sigma$, then we can choose normal directions on $c_a$ and $c_b$ in $\Sigma$ in such a way that near $\beta$ we have the picture below.

```
        c_a
          β
        c_b
```

The image of this 1-surgery will be a single curve $c'$ and this choice of a normal direction on $c_a, c_b$ in $\Sigma$ ensures that $c' \cap c_a = c' \cap c_b = \emptyset$. 

```
Case 2. Suppose that $\beta$ connects two points of the same component $c$ of $S \cap \Sigma$. Consider the compressing arc $\alpha \subset S \cap H_1$ for the boundary compression as well as the compressing disc $D$ with $\partial D = \alpha \cup \beta$. Because $S$ is orientable, we can choose a nonzero normal vector field $N(S) \subset M$.

If we then let $S_\epsilon$ be the image of $S$ after flowing in the direction $N(S)$ for a very short time, then $S_\epsilon \cap S = \emptyset$. Further, this normal direction can be chosen so that $S_\epsilon \cap D = \emptyset$. Note that on $\Sigma$, we then get the following picture.

$$ S_\epsilon \cap \Sigma \quad \beta \quad S \cap \Sigma $$

Let $c_\epsilon \subset S_\epsilon \cap \Sigma$ be the isotope of $c \subset S \cap \Sigma$. Then after the elementary compression of $S$, we see the intersections of $S'$ and $S_\epsilon$ with $\Sigma$ as below.

$$ c_\epsilon \quad \beta \quad c' $$

From this we see that even if $c \cap c' \neq \emptyset$, the curve $c$ is isotopic to a curve $c_\epsilon \subset \Sigma$ which is disjoint from $c'$. Thus $d(c, c') \leq 1$. □

**Lemma 3.9.** Suppose that $S'$ is the image of $S$ by an annular compression from $H_1$. Then the collection $S' \cap \Sigma$ of simple closed curves is (up to isotopy) a proper subset of the collection $S \cap \Sigma$.

**Proof.** As in Lemma 3.7, this is due to the fact that annular compression simply removes two components of $S \cap \Sigma$, leaving the rest in place. □

Together, Lemmas 3.8 and 3.9 will provide the means to place a bound on the distance $d(H_1, H_2)$. The idea in Section 4 will be to use a series of elementary compressions in order to develop a chain of essential simple closed curves on $\Sigma$, each distance 1 from the next. Because there may be $\partial$-parallel annuli in the way of this plan, some annular compressions may be needed at each stage. Lemma 3.9 ensures that this action will not affect the chain of essential curves.

**4. Proof of Theorem 1.2.**

Let $M$ be a closed, orientable, irreducible 3-manifold and suppose $S \subset M$ is a closed, orientable, incompressible surface of positive genus. Using an ambient isotopy, assume that $S$ meets $\Sigma$ transversely and minimally. That is, among all surfaces isotopic to $S$ in $M$, the number of components of $S \cap \Sigma$
is minimal. In most of the lemmas from Section 3, the embedded surface $S$ was required to have incompressible intersection with $H_1$ and/or have all components of $S \cap \Sigma$ essential in $\Sigma$. The following pair of lemmas show that these conditions are satisfied by placing $S$ in this minimal position with respect to $\Sigma$.

**Lemma 4.1.** Suppose that $S$ intersects $\Sigma$ minimally. Then for $i = 1, 2$, the intersection $S \cap H_i$ is incompressible in $H_i$.

*Proof.* Suppose that $S \cap H_i$ has a compressible component. Then there is a disc $(D, \partial D) \subset (H_i, S \cap H_i)$ such that $\partial D$ does not bound a disc in $S \cap H_i$. Because $S$ is incompressible in $M$, there is a disc $D' \subset S$ with $\partial D' = \partial D$. Note that $D' \cap \Sigma \neq \emptyset$. Because the manifold $M$ is irreducible, the sphere $D \cup D'$ bounds a ball in $M$. Thus the surface $S' = (S \setminus D') \cup D$ is isotopic to $S$ and $S' \cap \Sigma$ has strictly fewer components than $S \cap \Sigma$. □

**Lemma 4.2.** If the number of components of $S \cap \Sigma$ is minimal, then each component of $S \cap \Sigma$ is a simple closed curve which is essential in $\Sigma$.

*Proof.* Suppose $c \subset S \cap \Sigma$ is non-essential on $\Sigma$. Then let $D \subset \Sigma$ be the disc with $\partial D = C$. The incompressibility of $S$ in $M$ provides a disc $D' \subset S$ with $\partial D' = c$. Then the surface $S' = (S \setminus D') \cup D$ is isotopic to $S$ as $M$ is irreducible. We can then push the closed disc $D \subset S'$ slightly off $\Sigma$ so that $S' \cap \Sigma$ has strictly fewer components than $S \cap \Sigma$. □

**Lemma 4.3.** If the number of components of $S \cap \Sigma$ is minimal, then there are no $\partial$-parallel annular components of either $S \cap H_1$ or $S \cap H_2$.

*Proof.* This follows as a direct consequence of Lemma 3.9. □

The crux of the main theorem lies in the following pair of lemmas. The idea for the first lemma is to assume that $S \cap \Sigma$ is in minimal position and $S \cap H_2$ already has disc components of intersection. I want to move $S$ across $\Sigma$ until there is only one disc component. From that point, I will count the number of elementary compressions it takes to move $S$ further across $\Sigma$ to where $S \cap H_1$ contains a disc component.

**Lemma 4.4.** Let $M = (H_1, H_2; \Sigma)$ describe a strongly irreducible splitting of an irreducible manifold $M$. Suppose that $S \subset M$ is a closed, incompressible surface such that each component of $S \cap \Sigma$ is essential in $\Sigma$ and each component of $S \cap H_1$ is incompressible in $H_1$. If $S \cap H_2$ contains essential discs, then there is a sequence of isotopies

$$S \simeq S_0 \simeq S_1 \simeq \cdots \simeq S_k \simeq S_{k+1} \simeq \cdots \simeq S_n$$

of $S$ having the following properties:

* Each component of $S_i \cap \Sigma$ is essential in $\Sigma$ (and each $S_i \cap H_1$ is incompressible in $H_1$).
For any choice of components $c_i$ of $S_i \cap \Sigma$, $d(c_i, c_{i+1}) \leq 1$ for $0 \leq i < n$.
For $0 \leq i \leq k$, $S_i \cap H_2$ contains disc components.

$k \leq n - 2$, and for $k \leq i \leq n - 1$, neither $S_i \cap H_1$ nor $S_i \cap H_2$ contains any disc components.

$S_n \cap H_1$ contains a single disc component.

**Proof.** If there are any $\partial$-parallel annular components of $S \cap H_1$, use annular compressions to remove them and form $S_0$. Form $\widehat{S}_1$ by performing an elementary compression on $S_0$ from $H_1$. If $\widehat{S}_1 \cap H_1$ has any $\partial$-parallel annuli, perform annular compressions on $\widehat{S}_1$ from $H_1$ to form $S_1$. Otherwise, let $S_1 = \widehat{S}_1$. Continue in this way recursively to form $\widehat{S}_i$ and $S_i$ for $1 \leq i \leq n$. The first point then follows inductively from Lemma 3.6 and Lemma 3.7.

Suppose that $c_i$ is a component of $S_i \cap \Sigma$. Then by Lemma 3.8 we know $d(c_i, \widehat{c}_{i+1}) \leq 1$ for any component $\widehat{c}_{i+1}$ of $\widehat{S}_{i+1} \cap \Sigma$. Further, by Lemma 3.9, we know that the collection of components of $S_{i+1} \cap \Sigma$ is a subset of the collection of components of $\widehat{S}_{i+1} \cap \Sigma$. Thus for any choice $c_{i+1}$ of component of $S_{i+1} \cap \Sigma$, we get the bound $d(c_i, c_{i+1}) \leq 1$.

Let $k$ be the greatest number such that $S_k$ contains a disc component of intersection with $H_2$. By noting Lemmas 2.2 and 3.3, we see that the Euler characteristic changes by exactly 1 through each stage. Thus if $k$ is the greatest number such that $S_k \cap H_2$ has a disc component, then $S_k \cap H_2$ must have exactly one disc component. Otherwise, $\chi(S_{k+1} \cap H_1) \geq \chi(S_k \cap H_1) + 2$. Similarly, $n$ is chosen to be the least number such that $S_n$ has disc intersection with $H_1$ and we see that $S_n$ has exactly one disc component of intersection with $H_1$. Again, from Lemma 2.2, we know that such an $n$ exists, as after a finite number of compressions, the Euler characteristic will be positive, forcing there to be a disc component.

Note that because the splitting is strongly irreducible, we cannot have the case where $S \cap H_1$ and $S \cap H_2$ both have disc intersections. Recall from the remarks in the introduction that if $k = n$, then the splitting is reducible. If $k = n - 1$, then by applying Lemma 3.8, we would find that the splitting is weakly reducible. Thus the inequality $k \leq n - 2$ is due to the fact that $(H_1, H_2; \Sigma)$ is strongly irreducible. \qed

If $S \cap H_1$ has disc intersections, then we can apply the above lemma with the roles of $H_1$ and $H_2$ reversed. If, however, neither $S \cap H_1$ nor $S \cap H_2$ have disc intersections when $S \cap \Sigma$ is minimal, then we need to restate the lemma slightly. The idea here is to boundary compress $S$ in both directions, until we reach a disc intersection on either side.

**Lemma 4.5.** Let $M = (H_1, H_2; \Sigma)$ describe a strongly irreducible splitting of an irreducible manifold $M$. Suppose that $S \subset M$ is a closed, incompressible surface such that each component of $S \cap \Sigma$ is essential in $\Sigma$ and each component of $S \cap H_i$ is essential in $H_i$ (for $i = 1, 2$). If neither $S \cap H_2$ nor
$S \cap H_1$ contains essential disc components, then there is a sequence

$$S_{-m} \simeq S_{-m+1} \simeq \cdots \simeq S = S_0 \simeq \cdots \simeq S_n$$

of isotopic copies of $S$ having the following properties:

- Each component of $S_i \cap \Sigma$ is essential in $\Sigma$.
- $S_n \cap H_1$ contains a disc component, as does $S_{-m} \cap H_2$.
- For $-m + 1 \leq i \leq n - 1$, neither $S_i \cap H_1$ nor $S_i \cap H_2$ contains any disc components.
- For any choice of components $c_i$ of $S_i \cap \Sigma$, $d(c_i, c_{i+1}) \leq 1$, where $-m \leq i < n$.

Proof. For $i > 0$, define $\tilde{S}_i$ and $S_i$ exactly as in Lemma 4.4, again noting that there is a least $n$ such that $S_n \cap H_1$ has a disc component. To define $\tilde{S}_i$ and $S_i$ for $i < 0$, we use the same method, but with the roles of $H_1$ and $H_2$ reversed. Yet again, note that there is a least $m$ such that $S_{-m} \cap H_2$ has a disc component.

Now the proof for each of the points of the lemma is completely analogous to the proof of Lemma 4.4. □

Remark 4.6. If the number of components of $S \cap \Sigma$ is minimal, then the hypotheses of either Lemma 4.4 or Lemma 4.5 must be satisfied, depending on whether $S \cap H_1$, $S \cap H_2$, or neither contains disc components.

In either case, the idea is actually to start the incompressible surface in the position described in the conclusion to Kobayashi's lemma (Lemma 2.1). From there, we use a sequence of elementary compressions (together with annular compressions as necessary) to move the incompressible surface across $\Sigma$ until it again satisfies Lemma 2.1, but with the roles of $H_1$ and $H_2$ reversed. We then merely need to count the number of elementary compressions needed.

Main Theorem. If $M$ is a Haken 3-manifold containing an orientable incompressible surface of genus $g$, then any Heegaard splitting of $M$ has distance at most $2g$.

Proof. Recall that we have an irreducible manifold $M$ decomposed by a strongly irreducible splitting $(H_1, H_2; \Sigma)$. A closed, incompressible surface $S$ lies in $M$ such that the intersection $S \cap \Sigma$ is transverse and has a minimal number of components. If $g$ is the genus of $S$, we wish to show that there are a pair of essential simple closed curves $c_1$ and $c_2$ in $\Sigma$ such that $c_i$ bounds a disc in $H_i$ and $d(c_1, c_2) \leq 2g$.

First, suppose that $S$ in this starting minimal position satisfies the hypothesis of Lemma 4.4. That is, the intersection of $S$ with $H_2$ already contains disc components. Then the conclusion of the lemma provides (up to a relabeling of the indices) a sequence $S_0, \cdots, S_\ell$ such that $S_0$ has a single
disc component of intersection with $H_2$, $S_\ell$ has a single disc component of intersection with $H_1$, and all components of $S_i \cap \Sigma$ are essential in $\Sigma$ for $0 \leq i \leq \ell$. Notice by Lemma 2.2 that $\chi(S_i \cap H_1) = \chi(S_{i-1} \cap H_1) + 1$ for each $0 < i \leq \ell$.

Note that by relabeling indices, an identical sequence can be constructed if $S \cap H_1$ (instead of $S \cap H_2$) contained disc components.

On the other hand, suppose $S$ in this starting minimal position does not have disc intersections with either $H_1$ or $H_2$. Then by relabeling the indexing of the sequence formed in Lemma 4.5, we again have a sequence $S_0, \cdots, S_\ell$ such that $S_0 \cap H_2$ and $S_\ell \cap H_1$ each contain a single disc component and each component of $S_i \cap \Sigma$ is essential in $\Sigma$ for all $i$. Using both Lemma 2.2 and Corollary 2.3, we again see that $\chi(S_i \cap H_1) = \chi(S_{i-1} \cap H_1) + 1$ for each $0 < i \leq \ell$.

In any case, let $c_2$ be the boundary of the disc component of $S_0 \cap H_2$ and let $c_1$ be the boundary of the disc component of $S_\ell \cap H_1$. We know that if we choose any components $\gamma_i$ of $S_i \cap \Sigma$, then $d(\gamma_i, \gamma_{i+1}) \leq 1$. By letting $\gamma_0 = c_2$ and $\gamma_\ell = c_1$, we get the bound $d(c_1, c_2) \leq \ell$. Now it is left to show that this number $\ell$ is at most $2g$.

Recall the remarks leading to Corollary 2.3 that the decomposition of $S$ into $S \cap H_1$ and $S \cap H_2$ gives the equation

$$\chi(S) = \chi(S \cap H_1) + \chi(S \cap H_2).$$

Because each $S_i$ is isotopic to $S$, it is also true that

$$\chi(S) = \chi(S_0 \cap H_1) + \chi(S_0 \cap H_2).$$

Since $S_0 \cap H_2$ contains exactly one disc component, note that $\chi(S_0 \cap H_2) \leq 1$. Substituting the well-known formula for $\chi(S)$, we see

$$\chi(S) \leq \chi(S_0 \cap H_1) + 1$$

$$1 - 2g \leq \chi(S_0 \cap H_1).$$

Using the inductive equation $\chi(S_i \cap H_1) = \chi(S_{i-1} \cap H_1) + 1$, it is now clear that $S_{2g} \cap H_1$ is forced to have at least one disc component of intersection. Because we chose $\ell$ to be the least integer with this property, we finally conclude that

$$d(H_1, H_2) \leq d(c_1, c_2) \leq \ell \leq 2g.$$

□
References


Received August 3, 2000 and revised January 16, 2001.

Department of Mathematics
University of California, Davis
Davis, CA 95616
E-mail address: khartsho@math.ucdavis.edu
NORMAL HOLONOMY AND WRITHING NUMBER OF POLYGONAL KNOTS

JAMES J. HEBDA AND CHICHEN M. TSAU

The normal holonomy of a polygonal knot is a geometrical invariant which is closely related to the writhing number. We show that normal holonomy fibers the space of knots over the circle and deduce that the writhing number fibers the space of knots over the real line. Consequently, two isotopic knots which have the same writhing number are isotopic through a family of knots having the same writhing number. In a similar vein, two isotopic knots having zero holonomy are isotopic through a family of such knots if and only if they have the same autoparallel linking number.

More generally, the definition of normal holonomy makes sense for immersed polygonal knots. This time normal holonomy fibers the space of immersed knots over the circle, but now there are only two isotopy classes of immersed knots of zero holonomy.

1. Introduction.

The writhing number of a smooth knot in Euclidean 3-space is a real number which measures the extent to which the knot coils around itself. Although named by F.B. Fuller [8], the writhing number was originally discovered by G. Călugăreanu [4] and [5], in the form of an integral which he obtained while investigating the behavior in the limit of the Gauss formula for the linking number of a pair of space curves as one of the curves is allowed to approach the other. In this way, he established the formula

\[ LK(K, U) = WR(K) + TW(K, U) \]

where \( U \) is a unit vector field along the knot \( K \), \( LK(K, U) \) is the linking number of \( K \) with a nearby knot that approaches \( K \) from a direction along \( U \), \( WR(K) \) is the writhing number, and \( TW(K, U) \) is the twist of \( U \) along \( K \). Recently, the writhing number of a smooth knot has been related to the helicity of vector fields [6] and [17], and applied in the discussion of models of bacterial fibers [19].

The normal holonomy of the knot is a geometric invariant closely related to the writhing number, but more easily computed. Consider parallel translation with respect to the induced connection in the normal bundle of the
knot. Parallel translating a vector that is perpendicular to the knot at some point, once around the knot, has the effect of rotating the vector through an angle in the normal plane. Every such vector is rotated through the same angle. This common “phase” is the normal holonomy of the knot and may be regarded variously as an angle, a real number modulo $2\pi$, or a point on the unit circle. Suppose the knot $K$ is parametrized on the interval $[0,\ell]$. Let $X$ be a parallel unit normal vector field along $K$. The angle in the normal plane from $X_0$ to $X_\ell$ equals the normal holonomy $\text{HOL}(K)$. Suppose $U$ is a smooth unit normal vector field along $K$, and let $\theta$ be a continuous choice of angle between $X$ and $U$ along $K$. Because $U_0 = U_\ell$, $\theta_0 - \theta_\ell = \text{HOL}(K)$. On the other hand, $\theta_\ell - \theta_0 = 2\pi \text{TW}(K,U)$. Since $LK(K,U)$ is an integer, the formula (1.1) implies

$$2\pi WR(K) \equiv \text{HOL}(K) \mod 2\pi.$$  
(1.2)

Thus the normal holonomy may be identified with the “fractional part” of the writting number. The connection of writhe to the normal holonomy or Berry’s phase has appeared in the literature [15]. Earlier Banchoff and White [3] and [20], observed that the fractional part of the twist depends only on the knot. A closely related formula to (1.2) was observed by F.B. Fuller [9, Equation 6.1]. (See also [1, Equation 16].)

In this paper we study the normal holonomy and writhing number of polygonal, rather than smooth, knots. For us, polygonal knots are piecewise linear embeddings of the circle into 3-space, and so come already parametrized. This differs with the usual definition of polygonal knots as the image of such embeddings in $\mathbb{R}^3$, but our definition simplifies the discussion of the topology on the space of knots. This is important because our point of view is to regard the normal holonomy and writhing number as functions on the space of polygonal knots. Our main result is that normal holonomy fibers the space of knots over the circle and that writhing number fibers it over the real line.

As a corollary, we prove that two isotopic polygonal knots which have the same writhing number are isotopic through a family of knots having the same writhing number. This corollary was proved for $C^1$-smooth knots by Miller and Benham [16] by considering individual knots and adjusting an isotopy between them. For knots of zero holonomy, the writhe is an integer which is equal to the linking number of the knot with a nearby, geometrically determined, parallel curve. We call this integer the autoparallel linking number of the knot of zero holonomy and prove that two isotopic knots having zero holonomy are isotopic through a family of such knots if and only if they have the same autoparallel linking number.

These results may be viewed as the three dimensional version of the fact that two oriented isotopic knot diagrams are regularly isotopic if and only if they have the same writhing number and rotation number. (For a special
case, see [14, pp. 172-173].) They also bear comparison to the result that two isotopic smooth knots of non-vanishing curvature are isotopic through a family of such knots if and only if they have the same self-linking number [10] and [11]. Smooth knots of non-vanishing curvature form an open dense subset of the space of smooth knots, while the polygonal knots of zero holonomy are nowhere dense in the space of polygonal knots.

Finally we prove that holonomy fibers the space of immersed polygonal knots over the circle. Consequently, we obtain a result concerning immersed polygonal knots of zero holonomy which is analogous to Feldman’s results for smoothly immersed space curves of non-vanishing curvature [7].

2. The normal holonomy of polygonal knots.

A polygonal knot is a piecewise linear embedding of the circle $S^1$ into Euclidean 3-space $\mathbf{R}^3$. Thus, given a polygonal knot $K : S^1 \to \mathbf{R}^3$, there is a finite subset $\sigma$ of $S^1$ which subdivides $S^1$ into a finite number of intervals, such that the mapping $K$ restricted to each of these intervals is a linear affine mapping into $\mathbf{R}^3$ as a function of the usual angle parameter $t \mod 2\pi$ in $S^1$. The elements of $\sigma$ are denoted $t_0 < t_1 < \cdots < t_{n-1} < t_n = t_0$ and are to be regarded as cyclically ordered. The point $K(t_i)$ is the $i$-th vertex of $K$, and the restriction of $K$ to $[t_{i-1}, t_i]$ is the $i$-th edge. Because of cyclic ordering, the $n$-th edge is the restriction to the interval $[t_{n-1}, t_0]$.

\begin{equation}
T_i = \frac{K(t_i) - K(t_{i-1})}{|K(t_i) - K(t_{i-1})|}
\end{equation}

denote the oriented unit tangent vector along the $i$-th edge. The exterior angle $\alpha_i$ at the $i$-th vertex is the angle between $T_i$ and $T_{i+1}$ with $0 \leq \alpha_i < \pi$. Since $K$ is an embedding, $\alpha_i \neq \pi$. The definition allows for the possibility that $\alpha_i = 0$. When $\alpha_i \neq 0$, the unit normal vector $\xi_i$ at the $i$-th vertex is defined by the cross product

\begin{equation}
\xi_i = \frac{T_i \times T_{i+1}}{|T_i \times T_{i+1}|}.
\end{equation}

When $\alpha_i = 0$, any unit vector $\xi_i$ perpendicular to $T_{i+1} = T_i$ may be chosen and used in the following formulas.

Let $T_i^\perp$ denote the two-dimensional space of normal vectors to the $i$-th edge, oriented so that $X, Y$ is an oriented basis if and only if $T_i, X, Y$ is an oriented triad in $\mathbf{R}^3$. Rotation about the vector $\xi_i$ through the angle $\alpha_i$ is a linear isometry that carries $T_i^\perp$ onto $T_{i+1}^\perp$. It is the identity when $\alpha_i = 0$.

Let $R_i$ denote this isometry. The composition of all these rotations

\begin{equation}
R_n \circ \cdots \circ R_1 : T_1^\perp \to T_1^\perp
\end{equation}

is an orientation preserving linear isometry of $T_1^\perp$, hence is a rotation of $T_1^\perp$ through some angle HOL ($K$) (regarded as a number modulo $2\pi$) that we
will call the normal holonomy of the knot $K$. Clearly, the angle $\text{HOL}(K)$ does not depend upon which edge is considered to be the first. This will be obvious from the formula we derive for $\text{HOL}(K)$.

A simple formula for the normal holonomy can be obtained by the following considerations. There are two canonical, oriented orthonormal ordered bases for the two-dimensional vector space $T_i^\perp$, the alpha frame $A_i = \{\xi_{i-1}, T_i \times \xi_{i-1}\}$ and the omega frame $\Omega_i = \{\xi_i, T_i \times \xi_i\}$. One can regard these frames as the linear isometries $A_i, \Omega_i : \mathbb{R}^2 \to T_i^\perp$ which carry the standard basis to the respective bases. By construction the matrix representing $R_i$ relative to the omega basis of $T_i^\perp$ and the alpha basis of $T_{i+1}^\perp$ is the identity matrix, that is,

$$A_{i+1}^{-1} \circ R_i \circ \Omega_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. $$

The change of basis matrix $\Phi_i$ between the alpha and omega frames of $T_i^\perp$ is given by rotation through some angle $\phi_i$ as follows:

$$\Phi_i = \Omega_i^{-1} \circ A_i = \begin{bmatrix} \cos(\phi_i) & -\sin(\phi_i) \\ \sin(\phi_i) & \cos(\phi_i) \end{bmatrix}. $$

Since the columns of a change of basis matrix are the coordinates of the old basis in terms of the new, we have

$$\xi_{i-1} = \cos(\phi_i)\xi_i + \sin(\phi_i)T_i \times \xi_i. $$

Therefore

$$\cos(\phi_i) = \xi_i \cdot \xi_{i-1} \quad \text{(2.5a)}$$

and

$$\sin(\phi_i) = T_i \times \xi_i \cdot \xi_{i-1} = T_i \cdot \xi_i \times \xi_{i-1}. \quad \text{(2.5b)}$$

This pair of formulas is sufficient to determine $\phi_i$ modulo $2\pi$.

Now if we express the rotation $R_i$ relative to the alpha bases of $T_i^\perp$ and $T_{i+1}^\perp$ we obtain

$$A_{i+1}^{-1} \circ R_i \circ A_i = A_{i+1}^{-1} \circ R_i \circ \Omega_i \circ \Omega_i^{-1} \circ A_i = \Phi_i. $$

Since the holonomy is the product of the $R_i$, we obtain the formula

$$\text{HOL}(K) = \sum_{i=1}^{n} \phi_i \mod 2\pi. \quad \text{(2.6)}$$

**Remark.** When the vertices of $K$ are in general position, $\sum_{i=1}^{n} \phi_i$ is equal to the negative of the total torsion $TT(K)$ defined in [2], because $-\phi_i$ is the signed angle from $\xi_{i-1}$ to $\xi_i$ by (2.5). Formula (2.6) shows that $\text{HOL}(K)$ depends continuously on the vertices of $K$. In contrast, the total torsion is discontinuous at polygonal knots whose vertices are not in general position.

**Example 2.1.** Planar curves have zero holonomy mod $2\pi$. 

Example 2.2. Polygonal knot diagrams can be made into knots by replacing each crossing by a small polygonal bridge as in Figure 2.1. Knots constructed in this way clearly have zero holonomy mod $2\pi$.

Figure 2.1

Example 2.3. Consider the 1-parameter family of curves

$$K_\theta : S^1 \to \mathbb{R}^3$$

defined so that the vertices are the sequence of points $v_0 = (0,0,0)$, $v_1 = (0,1,0)$, $v_2 = (1,2,0)$, $v_3 = (0,3,0)$, $v_4 = (0,2,0)$, $v_5 = (-\cos \theta,1,\sin \theta)$. Note that the first and fourth edges lie along the $y$-axis. The parameter $\theta$ is the dihedral angle between the $xy$-plane and the plane containing the triangle $v_0v_4v_5$. Imagine the triangle $v_0v_4v_5$ rotating about the $y$-axis. See Figure 2.2.

Figure 2.2
A computation with the formulas \((2.4)\) and \((2.5)\) gives:

\[
\Phi_1 = \Phi_4 = \begin{bmatrix} - \cos \theta & \sin \theta \\ - \sin \theta & - \cos \theta \end{bmatrix}, \\
\Phi_2 = \Phi_5 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \\
\Phi_3 = \Phi_6 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
\]

Thus

\[
\Phi_6 \circ \cdots \circ \Phi_1 = \begin{bmatrix} \cos 2\theta & - \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{bmatrix}.
\]

Therefore \(\text{HOL}(K_\theta) = 2\theta\) modulo \(2\pi\).

**Example 2.4.** In this example we examine the effect on holonomy of putting a twist in an edge. Let \(E^\delta\) be the 2-parameter family of polygonal arcs whose vertices are: 
\[v_0 = (0, -3, 0), \quad v_1 = (0, -\frac{8\delta}{3}, 0),\]
and \(v_6 = (0, 3, 0),\) where \(0 \leq \epsilon < \frac{1}{5}\) and \(0 < \delta \leq 1.\) See Figure 2.3. Then the corresponding unit tangent vectors are found to be: 
\[T_1 = T_5 = (0, 1, 0), \quad T_2 = (-3\epsilon, \sqrt{1 - 25\epsilon^2}, 4\epsilon), \quad T_3 = T_4 = (5\epsilon, \sqrt{1 - 25\epsilon^2}, 0)\text{ and } T_5 = (-3\epsilon, \sqrt{1 - 25\epsilon^2}, -4\epsilon).\]

The unit normal vectors at the 0th, 3rd and 6th vertex are chosen for convenience while the others are computed from \((2.2)\):

\[
\xi_0 = \xi_3 = \xi_6 = (0, 0, 1)
\]

\[
\xi_1 = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \quad \xi_2 = \begin{pmatrix} \sqrt{1 - 25\epsilon^2} \\ \sqrt{5(1 - 20\epsilon^2)} \end{pmatrix}, \quad \xi_4 = \begin{pmatrix} \sqrt{1 - 25\epsilon^2} \\ \sqrt{5(1 - 20\epsilon^2)} \end{pmatrix}
\]

\[
\xi_5 = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \quad \xi_6 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}.
\]
The change of base matrices are computed from (2.4) and (2.5):

\[ \Phi_1 = -\Phi_6 = \Psi^* \equiv \begin{bmatrix} 3 & 4 \\ -1 & 5 \end{bmatrix} \]

(2.7) \[ \Phi_2 = \Phi_3 = -\Phi_4 = \Phi_5 = \Psi_\epsilon \equiv \begin{bmatrix} -\frac{2\sqrt{1-2\epsilon^2}}{\sqrt{5(1-20\epsilon^2)}} & \frac{1}{\sqrt{5(1-20\epsilon^2)}} \\ -\frac{1}{\sqrt{5(1-20\epsilon^2)}} & -\frac{2\sqrt{1-2\epsilon^2}}{\sqrt{5(1-20\epsilon^2)}} \end{bmatrix} . \]

(2.8)

Since the two negative signs in the matrix product will cancel, we have

\[ \Phi_6 \circ \cdots \circ \Phi_1 = \Psi^* \circ \Psi_\epsilon^4 \circ \Psi^* . \]

Therefore parallel translation from \( T_{1\perp} \) to \( T_{6\perp} \) is realized as rotation through the angle \( 2\psi^* + 4\psi_\epsilon \). Observe that this angle increases monotonically from 0 to \( \pi \) as \( \epsilon \) goes from 0 to \( \sqrt{7/180} \). Indeed, by (2.7), \( \psi^* = \arcsin(-4/5) \), and, by (2.8), as \( \epsilon \) goes from 0 to \( \frac{1}{5} \), the angle \( \psi_\epsilon \) increases from \( -\arccos(-\frac{2}{\sqrt{7}}) \) to \( -\pi/2 \), which is a change in angle exceeding \( \pi/4 \). A simple calculation shows the angle change is equal to \( \pi/4 \) when \( \epsilon = \sqrt{7/180} \).

**Lemma 2.5.** Given any polygonal knot \( K \), there is a family of polygonal knots \( K_\epsilon \), \( 0 \leq \epsilon \leq \sqrt{7/180} \), such that

1. \( K_0 = K \), and
2. \( \text{HOL}(K_\epsilon) \) continuously increases monotonically through an angle equal to \( \pi \) as \( \epsilon \) runs from 0 to \( \sqrt{7/180} \).

**Proof.** Pick a unit vector \( U \) which is perpendicular to the first edge of \( K \). There is a unique orientation preserving similarity transformation \( S \) of \( \mathbb{R}^3 \) into itself that carries the oriented line segment from \((0, -3, 0)\) to \((0, 3, 0)\) onto the first edge of \( K \) and the \( z \)-axis onto a positive multiple of \( U \). Define \( K_\epsilon \) to be \( K \) with the first edge replaced by \( S(E_\epsilon^\delta) \), where \( E_\epsilon^\delta \) is as in
Example 2.4, and where $\delta$ is chosen small enough so that $S(E^\delta)$ does not intersect any remaining edges of $K$, except for the endpoints of the first edge. This ensures that $K_\epsilon$ is a family of embeddings. Since the normal holonomy is unaffected under similarity transformations, the computation of Example 2.4 shows that $K_\epsilon$ has the desired properties.

**Remark.** Since reflection through a plane changes the sign of holonomy, the construction in Lemma 2.5 produces an isotopy $K_\epsilon$ of $K$ such that the holonomy decreases monotonically through an angle of $-\pi$ as $\epsilon$ runs from 0 to $\sqrt{7/180}$ when an edge of $K$ is replaced with an appropriately scaled reflection of $E^\delta$.

### 3. The writhing number of polygonal knots.

The same double integral used to define the writhing number of a smooth knot is used to define the writhing number of a polygonal knot $K$.

\[
WR(K) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{2\pi} \frac{K'(s) \times K'(t) \cdot (K(s) - K(t))}{|K(s) - K(t)|^3} ds dt.
\]

The integral is convergent and can be given by a simple formula involving the vertices. To see this, let $t_0 < t_1 < \cdots < t_n = t_0$ be the subdivision of $S^1$ into subintervals on which $K$ is linear. Set

\[
W_{ij} = \frac{1}{4\pi} \int_{t_{i-1}}^{t_i} \int_{t_{j-1}}^{t_j} \frac{K'(s) \times K'(t) \cdot (K(s) - K(t))}{|K(s) - K(t)|^3} ds dt.
\]

When $i = j$, the vectors $K'(s), K'(t), (K(s) - K(t))$ are collinear, and when $|i - j| = 1$, they are coplanar. Thus, if $|i - j| \leq 1$, $W_{ij} = 0$, because the integrand is zero almost everywhere. Therefore,

\[
WR(K) = \sum_{|i-j| \geq 2} W_{ij}.
\]

For a polygonal knot $K : S^1 \to \mathbb{R}^3$, consider the Gauss mapping

\[
\gamma : S^1 \times S^1 - \Delta \to S^2
\]

defined by

\[
\gamma(s,t) = \frac{K(s) - K(t)}{|K(s) - K(t)|}
\]

where $\Delta$ is the diagonal of $S^1 \times S^1$. When $|i-j| \geq 2$, the rectangle $[t_{i-1}, t_i] \times [t_{j-1}, t_j]$ does not meet $\Delta$, and its image under $\gamma$ is a convex, geodesic quadrilateral $Q_{ij}$ in $S^2$, which is possibly degenerate since the $i$–th and $j$–th edges of $K$ could be coplanar. When non-degenerate, $\gamma$ carries $[t_{i-1}, t_i] \times [t_{j-1}, t_j]$ diffeomorphically onto $Q_{ij}$.
Lemma 3.1. Suppose $|i - j| \geq 2$, then

$$4\pi W_{ij} = \pm \text{Area}(Q_{ij})$$

where $\text{Area}(Q_{ij})$ is the area of $Q_{ij}$. The plus sign is taken when $\gamma$ is orientation reversing on $[t_{i-1}, t_i] \times [t_{j-1}, t_j]$.

Proof. Because

$$4\pi W_{ij} = -\int_{[t_{i-1}, t_i] \times [t_{j-1}, t_j]} \gamma^*(\omega_{S^2}),$$

where $\omega_{S^2}$ is the area form on $S^2$ oriented by the outward normal, we have $4\pi W_{ij} = \pm \text{Area}(Q_{ij})$. (See [18, p. 132].) The sign is positive if and only if $\gamma$ is orientation reversing on the rectangle $[t_{i-1}, t_i] \times [t_{j-1}, t_j]$. This occurs precisely when the scalar triple product $T_i \times T_j \cdot \gamma(t_i, t_j) > 0$.

Suppose $|i - j| \geq 2$. Let $A = K(t_{j-1})$, $B = K(t_j)$, $C = K(t_{i-1})$, and $D = K(t_i)$. Let $\beta_{AD}, \beta_{AC}, \beta_{BD},$ and $\beta_{BC}$ denote the interior dihedral angles of the tetrahedron $ABCD$ at the respective edges $AD, AC, BD,$ and $BC$. Each of these four dihedral angles corresponds under $\gamma$ to one of the four interior angles of $Q_{ij}$ having the same measure. Thus the area of $Q_{ij}$ is given by the excess formula:

$$\text{Area}(Q_{ij}) = \beta_{AD} + \beta_{AC} + \beta_{BD} + \beta_{BC} - 2\pi.$$ 

Since the dihedral angles can be computed in terms of the vertices $A, B, C,$ and $D$, this results in a formula for $W_{ij}$ in terms of the vertices (cf. [4, pp. 15-17] and [2, p. 1177]).

Proposition 3.2. For any polygonal knot $K$,

$$2\pi WR(K) \equiv \text{HOL}(K) \mod 2\pi.$$ 

Remark. As noted in the introduction, the same formula (1.2) holds for smooth knots.

Proof. If the vertices of $K$ are in general position, then $WR(K) + (1/2\pi)TT(K)$ is an integer by [2, Theorem 4] where $TT(K)$ is the total torsion. Combining this with the previous remark, $\text{HOL}(K) \equiv -TT(K) \mod 2\pi$, shows that (3.4) holds when the vertices of $K$ are in general position.

If the vertices of $K$ are not in general position, (3.4) still holds by continuity of $WR$ and $\text{HOL}$ because $K$ can be approximated by polygonal knots whose vertices are in general position.

For sufficiently small $\epsilon$, the $\epsilon$-tube about a polygonal knot $K$ is a surface homeomorphic to the torus built out of pieces of cylinders which are pasted together along elliptical boundary curves. Each cylindrical piece of the tube is associated to an edge of $K$ which forms part of the axis of the cylinder. Moreover, each cylindrical piece is striated by a family of line segments which
are parallel to the associated edge of \( K \). There is an obvious correspondence between the set of stria and the set of directions in the plane perpendicular to the associated edge. Each stria in the \( i \)-th cylindrical piece connects to a unique stria in the adjoining \((i + 1)\)-st cylindrical piece in such a way that the corresponding directions \( U_i \in T_{i}^{\perp} \) and \( U_{i+1} \in T_{i+1}^{\perp} \) are related by \( R_i(U_i) = U_{i+1} \) where \( R_i \) is the rotation about the \( i \)-th vertex of \( K \) described in §2. Therefore, if \( K \) has zero holonomy modulo \( 2\pi \), one of these stria, followed around the knot, will return to itself to produce a closed polygonal curve \( J \) disjoint from \( K \). Any two such \( J \) are isotopic with one another. Thus, the linking number of \( K \) with \( J \) is independent of the choice of a sufficiently small \( \epsilon \) and the choice of the stria. We define the autoparallel linking number of the polygonal knot \( K \) with zero holonomy to be the linking number of \( K \) with any such \( J \).

**Remark.** The autoparallel linking number is different from the self-linking number defined in [2] for polygonal knots in general position.

**Proposition 3.3.** For polygonal knots \( K \) of zero holonomy, \( \text{WR}(K) \) is equal to the autoparallel linking number of \( K \).

**Proof.** By rounding out the corners of \( K \) we may replace \( K \) by a smooth knot, still denoted \( K \), with zero holonomy and the same writhe. Let \( U \) be a smooth parallel normal vector field along \( K \). Then as in §1, the twist of \( U \) along \( K \) is zero. Moreover, \( \text{LK}(K, U) \) equals the autoparallel linking number of \( K \). Therefore \( \text{WR}(K) \) equals the autoparallel linking number by applying Călugăreanu’s formula (1.1).

**Remark.** This shows that for oriented knots constructed from oriented knot diagrams as in Example 2.2, the writhing number is computed as the linking number of the oriented knot diagram and its blackboard framing, and therefore is given by the well-known combinatorial formula, see for example [14, p. 163].

### 4. The space of polygonal knots.

Given a finite subset \( \sigma = \{ t_0 < t_1 < \cdots < t_n = t_0 \} \) of \( S^1 \), let \( \mathcal{K}(\sigma) \) denote the collection of all polygonal knots \( K : S^1 \to \mathbb{R}^3 \) with vertices \( K(t_i), t_i \in \sigma \). Because a piecewise linear map \( K : S^1 \to \mathbb{R}^3 \) is determined by its values at the vertices, \( \mathcal{K}(\sigma) \) is in one-to-one correspondence with an open subset of \( \mathbb{R}^{3n} \). The correspondence takes a knot \( K \) in \( \mathcal{K}(\sigma) \) to the ordered \( n \)-tuple of vectors \( (K(t_1), \ldots, K(t_n)) \). Topologize \( \mathcal{K}(\sigma) \) so that the correspondence is a homeomorphism.

The family \( \mathcal{F} \) of all finite subsets of \( S^1 \) is partially ordered by set inclusion. The collection of the \( \mathcal{K}(\sigma) \) for \( \sigma \in \mathcal{F} \) form a directed system of topological spaces. For all \( \sigma_1, \sigma_2 \) in \( \mathcal{F} \), \( \mathcal{K}(\sigma_1) \) is a closed subspace of \( \mathcal{K}(\sigma_2) \) if and only
if \( \sigma_1 \subset \sigma_2 \), and always \( \mathcal{K}(\sigma_1) \cap \mathcal{K}(\sigma_2) = \mathcal{K}(\sigma_1 \cap \sigma_2) \) and \( \mathcal{K}(\sigma_1) \cup \mathcal{K}(\sigma_2) \subset \mathcal{K}(\sigma_1 \cup \sigma_2) \).

Define the space of polygonal knots to be the union
\[
\mathcal{K} = \bigcup_{\sigma \in \mathcal{F}} \mathcal{K}(\sigma)
\]
topologized with the direct limit topology. This means that \( \mathcal{K} \) has the weak topology, that is, a subset is closed in \( \mathcal{K} \) if and only if its intersection with every \( \mathcal{K}(\sigma) \) is closed. See [12, Chapter 15] or [21, pp. 27-28].

**Lemma 4.1.** \( \mathcal{K} \) is Hausdorff.

**Proof.** Given two distinct polygonal knots \( K_1 \) and \( K_2 \) in \( \mathcal{K} \), there exists a \( t^* \in S^1 \) such that \( K_1(t^*) \neq K_2(t^*) \). But the evaluation map
\[
eq \]
\( ev_{t^*} : \mathcal{K} \to \mathbb{R}^3 \)
defined by \( ev_{t^*}(K) = K(t^*) \) is continuous on \( \mathcal{K}(\sigma) \) for every \( \sigma \in \mathcal{F} \). Therefore \( ev_{t^*} \) is continuous on \( \mathcal{K} \), since \( \mathcal{K} \) has the limit topology. If \( U_1 \) and \( U_2 \) are disjoint open sets in \( \mathbb{R}^3 \) containing respectively \( K_1(t^*) \) and \( K_2(t^*) \), then \( ev_{t^*}^{-1}(U_1) \) and \( ev_{t^*}^{-1}(U_2) \) are disjoint open sets in \( \mathcal{K} \) containing respectively \( K_1 \) and \( K_2 \).

By applying Lemma 15.10 in [12] we have the following result.

**Lemma 4.2.** Every compact subset of \( \mathcal{K} \) is contained in some \( \mathcal{K}(\sigma) \).

**Remark.** In view of Lemma 4.2, every path in \( \mathcal{K} \) is a path in some \( \mathcal{K}(\sigma) \). Thus paths in \( \mathcal{K} \) correspond to isotopies of polygonal knots, and vice versa.

## 5. Serre fibrations over circles.

Recall that a continuous map \( f : X \to B \) is a Serre fibration if \( f \) has the homotopy lifting property with respect to simplices \( \Delta^n \) for all \( n \). (See [12, p. 79] and [21, p. 29].) There is a long exact homotopy sequence associated to a Serre fibration [12, p. 84]. This section is devoted to proving a characterization of Serre fibrations over the circle.

**Proposition 5.1.** Let \( f : X \to S^1 \) be a continuous map from a topological space \( X \) to the circle. Then \( f \) is a Serre fibration if and only if for every continuous simplex \( s : \Delta^n \to X \) there exists a pair of homotopies
\[
h, \tilde{h} : \Delta^n \times [0, 1] \to X
\]
such that, for every \( x \in \Delta^n \),

1. \( h(x, 0) = \tilde{h}(x, 0) = s(x) \),
2. \( f(h(x,t)) \) continuously increases monotonically through an angle equal to \( \pi \) as \( t \) runs from 0 to 1, and
(3) $f(h(x, t))$ continuously decreases monotonically through an angle equal to $-\pi$ as $t$ runs from 0 to 1.

The proof is based on several lemmas.

The exponential map $\exp : \mathbb{R} \to S^1$ is a covering map. Consider the pullback $\tilde{X}$ of this covering space by the continuous map $f : X \to S^1$. Then there is a covering map $\text{EXP} : \tilde{X} \to X$ and a bundle map $\tilde{f} : \tilde{X} \to \mathbb{R}$ such that the following diagram commutes.

\[
\begin{array}{ccc}
\tilde{X} & \xrightarrow{\tilde{f}} & \mathbb{R} \\
\text{EXP} \downarrow & & \downarrow \exp \\
X & \xrightarrow{f} & S^1.
\end{array}
\]

The following lemma is obvious (cf. [21, pp. 36-37]).

**Lemma 5.2.** $f : X \to S^1$ is a Serre fibration if and only if $\tilde{f} : \tilde{X} \to \mathbb{R}$ is a Serre fibration.

**Lemma 5.3.** The map $f : X \to S^1$ satisfies the hypothesis of Proposition 5.1, if and only if for every positive integer $m$ and every continuous simplex $s : \Delta^n \to \tilde{X}$, there exists a homotopy $h_m : \Delta^n \times [-m, m] \to \tilde{X}$ such that, for every $x \in \Delta^n$,

1. $h_m(x, 0) = s(x)$,
2. $t \mapsto f(h_m(x, t))$ is monotone increasing, and
3. $\tilde{f}(h_m(x, m)) - \tilde{f}(h_m(x, 0)) \geq m\pi$ and $\tilde{f}(h_m(x, 0)) - \tilde{f}(h_m(x, -m)) \geq m\pi$.

**Proof.** Assuming $f$ satisfies the hypothesis of Proposition 5.1, this is proved by induction on $m$. To construct $h_1$, patch together the lifts from $X$ to $\tilde{X}$ of the homotopies $h$ and $\tilde{h}$ obtained from the simplex $\text{EXP} \circ s : \Delta^n \to X$, which exist according to the hypothesis of Proposition 5.1. Then

$$\text{EXP} (h_1(x, t)) = \begin{cases} 
\tilde{h}(x, -t) & \text{if } -1 \leq t \leq 0 \\
h(x, t) & \text{if } 0 \leq t \leq 1.
\end{cases}$$

Once $h_m$ has been constructed, let $\tilde{h}$ be the homotopy in $X$ obtained from the simplex $x \mapsto \text{EXP} (h_m(x, -m))$ by hypothesis, and let $h$ be the homotopy obtained from the simplex $x \mapsto \text{EXP} (h_m(x, m))$. Then $h_{m+1}$ is constructed as an extension of $h_m$ by patching the homotopy $\tilde{h}$ on the left of $h_m$ and patching the homotopy $h$ on the right of $h_m$ after suitable reparametrizations of the interval domains of $\tilde{h}$ and $h$.

The converse is obvious.

**Lemma 5.4.** A continuous map $\tilde{f} : \tilde{X} \to \mathbb{R}$ is a Serre fibration if and only if for every positive integer $m$ and every continuous simplex $s : \Delta^n \to \tilde{X}$ there exists a homotopy $h_m : \Delta^n \times [-m, m] \to \tilde{X}$ such that, for every $x \in \Delta^n$,
(1) \( h_m(x, 0) = s(x) \),

(2) \( t \mapsto \tilde{f}(h_m(x, t)) \) is monotone increasing, and

(3) \( \tilde{f}(h_m(x, m)) - \tilde{f}(h_m(x, 0)) \geq m\pi \) and \( \tilde{f}(h_m(x, 0)) - \tilde{f}(h_m(x, 0)) \geq m\pi \).

**Proof.** Let \( s : \Delta^n \rightarrow \tilde{X} \) be a continuous simplex in \( \tilde{X} \), and let \( k : \Delta^n \times [0, 1] \rightarrow \mathbb{R} \) be a homotopy of \( \tilde{f} \circ s \). Thus we have the following commutative diagram where \( i_0(x) = (x, 0) \) for all \( x \) in \( \Delta^n \).

\[
\begin{array}{ccc}
\Delta^n & \xrightarrow{s} & \tilde{X} \\
\downarrow{i_0} & & \downarrow{\tilde{f}} \\
\Delta^n \times [0, 1] & \xrightarrow{k} & \mathbb{R}
\end{array}
\]

Pick an integer \( m > 0 \) so that

\[ |k(x, t) - k(x, 0)| < m\pi \]

for all \( x \) in \( \Delta^n \) and \( t \) in \( [0, 1] \). Consider the map

\[ (id, \tilde{f} \circ h_m) : \Delta^n \times [-m, m] \rightarrow \Delta^n \times \mathbb{R}. \]

By (2), this is one-to-one, and hence a homeomorphism onto its image. Consider also the map

\[ (id, k) : \Delta^n \times [0, 1] \rightarrow \Delta^n \times \mathbb{R}. \]

By (3), the image of the map \( (id, k) \) is contained in the image of the map \( (id, \tilde{f} \circ h_m) \).

It is straightforward to verify that the map

\[ h_m \circ (id, \tilde{f} \circ h_m)^{-1} \circ (id, k) : \Delta^n \times [0, 1] \rightarrow \tilde{X} \]

is a lift of \( k \). Here are the details of the calculation:

\[
\begin{align*}
h_m \circ (id, \tilde{f} \circ h_m)^{-1} \circ (id, k)(x, 0) &= h_m \circ (id, \tilde{f} \circ h_m)^{-1}(x, k(x, 0)) \\
&= h_m \circ (id, \tilde{f} \circ h_m)^{-1}(x, \tilde{f}(s(x))) \\
&= h_m \circ (id, \tilde{f} \circ h_m)^{-1}(x, \tilde{f}(h_m(x, 0))) \\
&= h_m(x, 0) = s(x)
\end{align*}
\]

and

\[
\begin{align*}
\tilde{f} \circ h_m \circ (id, \tilde{f} \circ h_m)^{-1} \circ (id, k)(x, t) &= \tilde{f} \circ h_m \circ (id, \tilde{f} \circ h_m)^{-1}(x, k(x, t)) \\
&= \tilde{f} \circ h_m(x, t') \text{ where } \tilde{f}(h_m(x, t')) = k(x, t) \\
&= k(x, t).
\end{align*}
\]
This proves that $\tilde{f}$ is a Serre fibration.

Conversely, suppose $\tilde{f}$ is a Serre fibration. Given a positive integer $m$ and a continuous simplex $s : \Delta^n \to \tilde{X}$, $h_m$ can be constructed from the homotopy lifting property by lifting the homotopy $k : \Delta^n \times [-m, m] \to \mathbb{R}$ defined by the formula $k(x, t) = \tilde{f}(s(x)) + t\pi$ to $\tilde{X}$.

6. HOL and WR are Serre fibrations.

The formula in §2 shows that the normal holonomy of a polygonal knot depends continuously on the vertices. Thus $\text{HOL} : \mathcal{K} \to S^1$ is continuous because $\mathcal{K}$ has the limit topology.

**Theorem 6.1.** $\text{HOL} : \mathcal{K} \to S^1$ is a Serre fibration.

**Proof.** Let $s : \Delta^n \to \mathcal{K}$ be a continuous simplex in $\mathcal{K}$. Since $\Delta^n$ is compact, Lemma 4.2 implies there exists a finite set $\sigma \subset S^1$ such that $s(\Delta^n) \subset \mathcal{K}(\sigma)$. Moreover, there exists a continuous unit vector field $U : \Delta^n \to \mathbb{R}^3$ such that $U(x)$ is perpendicular to the first edge of the knot $s(x)$ for every $x$ in $\Delta^n$. To prove this, consider the vector bundle over $\Delta^n$ such that the fiber over $x \in \Delta^n$ is the vector subspace of $\mathbb{R}^3$ consisting of all vectors perpendicular to the first edge of the knot $s(x)$. This bundle must be trivial because the base space $\Delta^n$ is contractible [13, p. 29]. Hence this bundle admits a continuous nonzero section, which after normalizing gives the unit vector field $U$.

For each $x \in \Delta^n$, consider the construction in Lemma 2.5 of the family of knots starting from $K = s(x)$ with $U = U(x)$. For any fixed $\delta$, the construction depends continuously on $K$ and the choice of $U$. Because $\Delta^n$ is compact, there is a value of $\delta$ that can be used uniformly for all $x \in \Delta^n$. Changing the parameter interval to $[0, 1]$ produces a homotopy $h : \Delta^n \times [0, 1] \to \mathcal{K}$ such that, for all $x \in \Delta^n$, $h(x, 0) = s(x)$ and $t \mapsto \text{HOL}(h(x, t))$ is strictly monotone increasing through an angle equal to $\pi$ as $t$ runs from 0 to 1. Likewise, in view of the remark after Lemma 4.2, there is also a homotopy $\tilde{h} : \Delta^n \times [0, 1] \to \mathcal{K}$ which decreases holonomy monotonically through an angle $-\pi$. Proposition 5.1 implies that $\text{HOL}$ is a Serre fibration.

**Theorem 6.2.** $WR : \mathcal{K} \to \mathbb{R}$ is a Serre fibration.

**Proof.** By Proposition 3.2, $\text{HOL}$ factors as the composition of the map $2\pi WR : \mathcal{K} \to \mathbb{R}$ and the covering map $\exp : \mathbb{R} \to S^1$. Thus, because $\text{HOL}$ is a Serre fibration, so will be the map $2\pi WR : \mathcal{K} \to S^1$. Therefore $WR$ is a Serre fibration.

**Corollary 6.3.** Two isotopic polygonal knots which have the same writhing number are isotopic through a family of polygonal knots with the same writhing number.

**Proof.** For each real number $w$, the long exact homotopy sequence for a fibration, together with the fact that $\mathbb{R}$ is contractible, implies that the
inclusion of the fiber $WR^{-1}(w)$ into $\mathcal{K}$ is a weak homotopy equivalence. In particular, 

$$WR_* : \pi_0(WR^{-1}(w)) \to \pi_0(\mathcal{K})$$

is an isomorphism of sets. Thus the path connected components of $WR^{-1}(w)$ are in one-to-one correspondence with the path connected components of $\mathcal{K}$.

**Corollary 6.4.** A polygonal knot with zero writhe is isotopic to its reflection if and only if it is isotopic to its reflection through polygonal knots with zero writhe.

*Proof.* This follows from Corollary 6.3 because the writhe of the reflection of a polygonal knot is the negative of the writhe of the knot.

**Remark.** This is a three dimensional version of the Mirror Theorem [14, p. 173].

Reversing the orientation of a polygonal knot does not change the writhing number. Therefore we have the following:

**Corollary 6.5.** A polygonal knot is isotopic to its inverse if and only if it is isotopic to its inverse through a family of polygonal knots of the same writhe.

The class of a polygonal knot with zero holonomy will be denoted $\mathcal{K}_0$. This class is a fiber of HOL. Consider part of the long exact homotopy sequence for HOL:

$$\pi_1(\mathcal{K}) \xrightarrow{\text{HOL}^*} \pi_1(S^1) \xrightarrow{\partial} \pi_0(\mathcal{K}_0) \to \pi_0(\mathcal{K}).$$

By Proposition 3.2, HOL factors through $\mathbb{R}$. Thus $\text{HOL}^*$ is the trivial group homomorphism, and the above long exact sequence gives the short exact sequence

$$1 \to \pi_1(S^1) \xrightarrow{\partial} \pi_0(\mathcal{K}_0) \to \pi_0(\mathcal{K}).$$

This shows that the path connected components of $\mathcal{K}_0$ are indexed by the set of integers in the form of $\pi_1(S^1)$ and the path connected components of $\mathcal{K}$. By Corollary 6.3, WR is constant on each path connected component of $\mathcal{K}_0$. By Proposition 3.3, this constant is an integer equal to the autoparallel linking number of any knot in that component. Therefore the following holds:

**Corollary 6.6.** Two isotopic knots having zero holonomy are isotopic through a family of such knots if and only if they have the same autoparallel linking number.

Recall that the fundamental group of the base of a Serre fibration acts on the set of path connected components of the fiber (cf. [21, p. 186]). Thus, by Theorem 6.2, $\pi_1(S^1) \approx \mathbb{Z}$ acts on the set $\pi_0(\mathcal{K}_0)$. Let $n \cdot [K]$ denote the action of $n \in \mathbb{Z}$ on $[K]$ in $\pi_0(\mathcal{K}_0)$. In view of Corollary 6.5, we can define writhe on $\pi_0(\mathcal{K}_0)$ by $WR([K]) = WR(K)$. 
Theorem 6.7. For every $[K]$ in $\pi_0(K_0)$ and $n$ in $\mathbb{Z}$,
$$\text{WR}(n \cdot [K]) = \text{WR}([K]) + n.$$  
If $K$ is a knot diagram, then $1 \cdot [K]$ has a knot diagram representative that results from $K$ by a type I Reidemeister move.

Proof. Let $[K] \in \pi_0(K_0)$ and $n \in \mathbb{Z}$. Consider the commutative diagram:

$$\begin{array}{ccc}
\mathbb{K} & \xrightarrow{2\pi\text{WR}} & \mathbb{R} \\
\downarrow{\text{HOL}} & & \downarrow{\exp S^1} \\
S^1 & \rightarrow & S^1
\end{array}$$

The loop in $S^1$ representing $n$ lifts to a curve $\alpha$ in $\mathbb{K}$ starting at $K$ and ending in $n \cdot [K]$. This lift can be achieved in two stages. First the loop lifts to a path in $\mathbb{R}$ starting at $2\pi\text{WR}(K)$ and ending at $2\pi(\text{WR}(K) + n)$, which then lifts to the curve $\alpha$. This proves the first statement.

If $K$ is a knot diagram, it defines a knot class $[K]$ in $\pi_0(K_0)$ by Example 2.2. Rearrange the diagram $K$ to $K_0$ as shown in Figure 6.1. By rotating the triangle $ABC$ around the line $AC$, one produces a family of knots $K_\theta$, $0 \leq \theta \leq \pi$, such that $\text{HOL}(K_\theta) = 2\theta$ just as in Example 2.3. Since $\text{HOL}(K_\theta)$ wraps once counterclockwise around the circle, $[K_\pi] = (+1) \cdot [K]$. Clearly, the diagram for $K_\pi$ results from that of $K$ by a type I Reidemeister move. See Figure 6.1. (For a description of the three types of Reidemeister moves see [14, p. 9].)

![Figure 6.1](image)

7. Immersed polygonal knots.

A piecewise linear immersion of $S^1$ into $\mathbb{R}^3$ is a piecewise linear map which is locally an embedding. We will call such maps immersed polygonal knots. The notation in §2 carries over to immersed polygonal knots. It is still the case that the exterior angle at every vertex must be strictly less than $\pi$, even though an immersed polygonal knot may have self-intersections. Thus normal holonomy is defined for immersed polygonal knots. However, the
The writhing number cannot be defined, and there is no analogue for Proposition 3.2.

On the other hand, there is an analogue for the autoparallel linking number of a polygonal knot of zero holonomy. Let $K$ be an immersed polygonal knot of zero holonomy. Recalling the notation of §2, pick a unit vector $U_1 \in T^1$, and define $U_i \in T^i$ inductively by $U_{i+1} = R_i(U_i)$. Since $K$ has zero holonomy, $R_n(U_n) = U_1$ by (2.3). Let $M_i$ be the orthogonal matrix in $SO(3)$ whose columns are the vectors $T_i, U_i, T_i \times U_i$. Recalling Formula (2.2), define $\Gamma^\theta_i$ to be rotation about $\xi_i$ through an angle $\theta$, where $0 \leq \theta \leq \alpha_i$. Then $\theta \mapsto \Gamma^\theta_i \circ M_i$ is a path in $SO(3)$ joining $M_i$ to $M_{i+1}$. The collection of these paths forms a loop in $SO(3)$ whose free homotopy class is independent of the choice of $U_1$. The free homotopy class of loops in $SO(3)$ associated to $K$ in this way will be called the rotation class of $K$. There are two possible rotation classes because $\pi_1(SO(3)) \approx \mathbb{Z}_2$ (see, for example, [21, pp. 198-199]). Both rotation classes are realizable. For example, the rotation class of any regular planar polygon is the nontrivial homotopy class in $SO(3)$, while the rotation class of a polygonal planar figure eight is the trivial homotopy class.

The space $I$ of immersed polygonal knots can be topologized as the direct limit of finite dimensional approximations $I(\sigma)$ just as in §4. The proof of Theorem 6.1 goes over without modification. In fact it is a little easier since one need not worry about self-intersections introduced by the homotopies.

**Theorem 7.1.** $HOL : I \to S^1$ is a Serre fibration.

Let $I_0$ denote the space of immersed polygonal knots of zero holonomy. Consider the long exact homotopy sequence of $HOL$:

$$\pi_0(I) \to \pi_1(I) \xrightarrow{HOL^*} \pi_1(S^1) \xrightarrow{\partial} \pi_0(I_0) \to \pi_0(I).$$

$\pi_0(I)$ is a set with one element since $I$ is path connected. The loop $K_\theta$, $0 \leq \theta \leq 2\pi$, of Example 2.3 shows that the image of $HOL_\ast$ in $\mathbb{Z}$ contains $2\mathbb{Z}$. Thus $\pi_0(I_0)$ has two elements corresponding to the two rotation classes. This proves the following analogue to Corollary 6.3. There is a similar result for smooth space curves of non-vanishing curvature [7].

**Corollary 7.2.** A pair of immersed polygonal knots of zero holonomy are isotopic through a family of immersed polygonal knots of zero holonomy if and only if they have the same rotation class.

Since $K_0 \subset I_0$, every $K \in K_0$ has a rotation class $ROT(K) \in \mathbb{Z}_2 \approx \pi_1(SO(3))$ as well as an autoparallel linking number, which equals $WR(K) \in \mathbb{Z}$ by Proposition 3.3.

**Corollary 7.3.** Let $K$ be a polygonal knot of zero holonomy. Then

$$WR(K) + 1 \equiv ROT(K) \mod 2.$$
Remark. This result should be compared with Fuller’s formula [9, Equation 6.1] or [1, Equation 16].

Proof. If $K$ is a knot diagram, $\text{ROT}(K)$ is clearly equal to the rotation number [22] of the diagram modulo 2. (See [14, p. 170] for a combinatorial definition of rotation number.) Furthermore, the equivalence

$$WR(K) + 1 \equiv \text{ROT}(K) \mod 2$$

holds for knot diagrams $K$ because it holds for the trivial knot diagram and the equivalence is unchanged under Reidermeister moves and crossing changes. Now observe every connected component of $K_0$ contains a knot diagram as in Example 2.2 and Theorem 6.7. Because HOL and $\text{ROT}$ are constant on the connected components of $K_0$ by Corollaries 6.6 and 7.2, the equivalence must hold for every $K \in K_0$.

References


NORMAL HOLONY AND WRITING NUMBER


Received July 11, 2000 and revised June 13, 2001.

Department of Mathematics
Saint Louis University
St. Louis, MO 63103
E-mail address: hebdajj@slu.edu

Department of Mathematics
Saint Louis University
St. Louis, MO 63103
E-mail address: tsauem@slu.edu
A DESCRIPTION OF CERTAIN AFFINE OPEN SUBSCHEMES THAT FORM AN OPEN COVERING OF \( \text{Hilb}_{A_k^2}^n \)

MARK E. HUIBREGTSE

In this paper, we study certain affine open subschemes of the Hilbert scheme of \( n \) points of the affine plane. We express the coordinate rings of these subschemes explicitly as quotients of polynomial rings; as an application, we give sufficient conditions for these subschemes to be isomorphic to \( 2n \)-dimensional affine space.

1. Introduction.

Let \( k \) be an algebraically closed field of any characteristic, \( A_k^2 = \text{Spec}(k[x, y]) \) the affine plane over \( k \), and \( \text{Hilb}_{A_k^2}^n = \text{H}^n \) the Hilbert scheme parameterizing 0-dimensional closed subschemes \( \text{Spec}(k[x, y]/I) \subseteq A_k^2 \) having length \( n \), that is,

\[
\text{dim}_k(k[x, y]/I) = \text{colength of } I = n.
\]

In particular, the \( k \)-points of \( \text{H}^n \) are in natural bijective correspondence with the ideals \( I \subseteq k[x, y] \) of colength \( n \); we often identify the ideal \( I \) with its associated point \( I \in \text{H}^n \). The correspondence is defined by the universal closed subscheme \( Z_n \subseteq \text{H}^n \times A_k^2 \), which is finite and flat of degree \( n \) over \( \text{H}^n \) via the first projection: The \( k \)-point \( t \in \text{H}^n \), given as a map

\[
t: \text{Spec}(k) \to \text{H}^n,
\]

corresponds to the closed subscheme

\[
Z_t = \text{Spec}(k) \times_{\text{H}^n} Z_n \subseteq \text{Spec}(k) \times_{\text{H}^n} (\text{H}^n \times A_k^2) \approx A_k^2.
\]

In a recent paper [7], Haiman defines, for each partition

\[
\mu = (p_1, p_2, \ldots, p_\ell) \text{ of } n, \text{ with } p_1 \geq p_2 \geq \cdots \geq p_\ell > 0,
\]

an open affine subscheme \( U_\mu \subseteq \text{H}^n \), as follows: We first encode the Ferrers’ diagram of \( \mu \) as an arrangement of monomials in \( x \) and \( y \), with the parts
corresponding to the rows. For example, if \( \mu = (4, 3, 1) \), the diagram is

\[
\begin{array}{ccc}
  x^2 & x & y^2 \\
  x & xy & xy^2 \\
  1 & y & y^3 \\
\end{array}
\]

we call the monomials in the diagram the **partition monomials**, and we write \((h, k) \in \mu\) to denote that \(x^h y^k\) is a partition monomial. (The display can be extended in the obvious way to comprise all monomials in \(x\) and \(y\); when this is done, the monomial \(x^r y^s\) resides in row \(r\) and column \(s\).) We then define \(U_\mu\) (as a point set) to be

\[
U_\mu = \{ I \in H^n | k[x, y]/I \text{ is spanned by the partition monomials} \}
\]

\[
= \{ I \in H^n | \text{the partition monomials are a k-basis of } k[x, y]/I \}.
\]

Haiman proves that \(U_\mu\) is an open affine subscheme of \(H^n\), and describes an infinite family of regular functions on \(U_\mu\) that generates the coordinate ring \(O_{U_\mu} \) [7, Proposition 2.1]. These functions arise as follows: For any \(I \in U_\mu\), and monomial \(x^r y^s\), we have a unique expansion

\[
x^r y^s \equiv \sum_{(h, k) \in \mu} c_{hk}^r s(I) x^h y^k \pmod{I};
\]

as \(I\) varies over \(U_\mu\), the coefficients in this expansion define functions \(c_{hk}^r s\) on \(U_\mu\). Haiman shows that the set

\[
\{ c_{hk}^r s | (h, k) \in \mu, \text{ and all } (r, s) \}
\]

generates \(O_{U_\mu}\) as a k-algebra; he also identifies a 2n-member subset of the generating functions that is a set of local parameters at the point

\[
I_\mu = \{ x^r y^s | (r, s) \notin \mu \} \in U_\mu,
\]

a monomial ideal that is the “origin” of \(U_\mu\) in the sense that all the (nontrivial) functions \(c_{hk}^r s\) vanish there. Haiman gives a new proof of the well-known fact that \(H^n\) is nonsingular (and irreducible) by reducing the question to the nonsingularity of \(H^n\) at monomial ideals (each of which is \(I_\mu\) for some \(\mu\)), for which his explicit local parameters provide an affirmative result [7, Proposition 2.4, Corollary 2.5].

The purpose of this paper is to give a fuller description of \(U_\mu\) and of the restriction

\[
Z_{U_\mu} = \text{Spec}(O_{U_\mu}[x, y]/I_\mu)
\]

of the universal closed subscheme \(Z_n\) to \(U_\mu\). We begin by identifying a finite subset \(c_\mu\) of the functions \(c_{hk}^r s\) that generates the k-algebra \(O_{U_\mu}\). More precisely, Proposition 3.1.2 states that the set

\[
c_\mu = \{ c_{hk}^r s | x^r y^s \text{ is a leading monomial and } (h, k) \in \mu \}
\]
generates \( \mathcal{O}_{U_\mu} \), where a leading monomial is a monomial that lies either immediately to the right of a row or immediately above a column in the diagram of \( \mu \) (see Figure 1). We then prove in Proposition 3.2.1 that the polynomials

\[
g_{rs} = x^r y^s - \sum_{(h,k) \in \mu} c_{hk}^r x^h y^k \in \mathcal{O}_{U_\mu}[x,y]
\]

(1)

corresponding to the leading monomials \( x^r y^s \) generate the ideal \( \mathcal{J}_\mu \) cutting out the universal closed subscheme over \( U_\mu \). A simple algorithm (Algorithm 4.2.1) yields a basis of the first syzygy module of the generators \( g_{rs} \) (Corollary 4.4.1), and hence a free resolution of the \( \mathcal{O}_{U_\mu}[x,y] \)-algebra \( \mathcal{O}_{U_\mu}[x,y] / \mathcal{J}_\mu \) (Corollary 4.4.2); moreover, the \( g_{rs} \) can be recovered (up to sign) as the maximal minors of the matrix whose rows are the members of the syzygy basis (Theorem 4.4.4).

To express the ring \( \mathcal{O}_{U_\mu} \) concretely as a quotient, we introduce the set of indeterminates

\[ \mathcal{C}_\mu = \{ c_{hk}^{rs} \mid c_{hk}^{rs} \in \mathcal{C}_\mu \}, \]

and define the surjection

\[
u^*_\mu : k[\mathcal{C}_\mu] \to \mathcal{O}_{U_\mu}, \quad c_{hk}^{rs} \mapsto c_{hk}^{rs},
\]

(2)

which is the comorphism of a closed immersion

\[ u_\mu : U_\mu \to \text{Spec}(k[\mathcal{C}_\mu]). \]

We give two different (computable) sets of generators for the kernel \( \mathfrak{R}_\mu \) of the map \( u^*_\mu \) (Theorem 5.1.1 and Theorem 6.1.1); the first of these is obtained as a byproduct of the (generalized) syzygy algorithm, and the second emerges from the recovery of the \( g_{rs} \) as subdeterminants of the syzygy matrix. As an application, we obtain sufficient conditions on \( \mu \) for \( U_\mu \) to be an affine cell in \( \mathbb{H}^n \), that is, an open subscheme isomorphic to \( \mathbb{A}^{2n}_k \) (Corollary 7.3.2 and Corollary 7.5.1); these conditions subsume some important special cases,
including \( \mu = (1, 1, \ldots, 1) \), which is discussed by Haiman [7, Corollary 2.8], and \( \mu = (r, r-1, r-2, \ldots, 1) \), for which

\[
I_\mu = (x^r, x^{r-1}y, \ldots, y^r) = (x, y)^r
\]

is a “fat point” ideal (see Remarks 7.3.3).

We end the introduction with the following brief table of contents that summarizes the organization and contents of the paper.

**Section 1:** Introduction.

**Section 2:** \( \text{Hilb}^n_{\mathbb{A}^2} \) and the affine open subschemes \( U_\mu \). Summarizes the definitions and results from Haiman’s paper [7] needed in the sequel, including certain relations among the functions \( c_{hk}^{rs} \), and Haiman’s local parameters at the point \( I_\mu \). We also show in Section 2.5 that Haiman’s result can be used to obtain local parameters at each point of \( \mathbb{H}^n \).

**Section 3:** A finite set of generators of \( \mathcal{O}_{U_\mu} \). Obtains the finite generating set \( \mathcal{C}_\mu \) of \( \mathcal{O}_{U_\mu} \), and proves that the polynomials \( g_{rs} \) (1) generate the ideal \( \mathcal{I}_\mu \).

**Section 4:** A free resolution of \( \mathcal{O}_{U_\mu}[x, y]/\mathcal{I}_\mu \). Presents the algorithm for computing a basis of the first syzygy module of the polynomials \( g_{rs} \). (To prepare for subsequent sections, we study this algorithm in greater generality than is needed for the immediate application.) Consequences of the syzygy basis include, as previously noted, a free resolution of the \( \mathcal{O}_{U_\mu}[x, y] \)-module \( \mathcal{O}_{U_\mu}[x, y]/\mathcal{I}_\mu \), and the recovery of the \( g_{rs} \) as the signed maximal minors of the syzygy matrix.

**Section 5:** An explicit representation of \( \mathcal{O}_{U_\mu} \) as a quotient ring. Obtains the first set of generators of the kernel \( \mathcal{R}_\mu \) of the map \( u_\mu^* \) (2).

**Section 6:** A second set of generators of the ideal \( \mathcal{R}_\mu \). Obtains the second set of generators of the kernel of the map \( u_\mu^* \).

**Section 7:** Smaller generating sets for \( \mathcal{O}_{U_\mu} \) and affine cell criteria. Identifies a “small” subset of \( \mathcal{C}_\mu \) that generates \( \mathcal{O}_{U_\mu} \) as a \( k \)-algebra, and presents sufficient conditions on \( \mu \) for \( U_\mu \) to be an affine cell in \( \mathbb{H}^n \).

2. \( \text{Hilb}^n_{\mathbb{A}^2} \) and the affine open subschemes \( U_\mu \).

In this section of the paper, we briefly recall the definition and some properties of the Hilbert scheme \( \text{Hilb}^n_{\mathbb{A}^2} = \mathbb{H}^n \), and summarize the necessary background from [7] regarding the open subschemes \( U_\mu \subseteq \mathbb{H}^n \) (defined for every partition \( \mu \) of \( n \)) which are the main focus of this paper. Recall that the ground field \( k \) is algebraically closed and of arbitrary characteristic.

2.1. Definition of \( \mathbb{H}^n \). The variety \( \mathbb{H}^n \) can be defined most naively as the set of ideals \( I \subseteq k[x, y] \) having colength \( n \), that is, \( \dim_k(k[x, y]/I) = n \).
Less naively, the variety $H^n$ parameterizes in a natural way the set of 0-dimensional closed subschemes $Z = \text{Spec}(k[x, y]/I) \subseteq \mathbb{A}^2_k = \text{Spec}(k[x, y])$ that have length $n$. It is defined as the open subscheme of $\text{Hilb}^n_{\mathbb{P}^2_k}$ (the existence of which is a consequence of Grothendieck’s general construction given in [5]) arising from the inclusion of $\mathbb{A}^2_k$ into $\mathbb{P}^2_k$ as a standard affine.

By pullback, $H^n$ inherits a universal closed subscheme $Z_n \subseteq H^n \times \mathbb{A}^2_k$, which is finite and flat of degree $n$ over $H^n$ via the first projection, and satisfies the following property:

Let $T$ be a separated scheme of finite type over $k$. Then the set of maps $f: T \to H^n$ is in natural bijective correspondence with the set of closed subschemes $Z_f \subseteq T \times \mathbb{A}^2_k$ that are finite and flat of degree $n$ over $T$; the bijection $f \mapsto Z_f$ is defined by $Z_f = T \times H^n Z_n$.

In particular, the inclusion of the $k$-point $t \in H^n$ corresponds to a unique closed subscheme $Z_t \subseteq t \times \mathbb{A}^2_k \approx \mathbb{A}^2_k$; the map $t \mapsto Z_t$ defines a bijection from the set of $k$-points of $H^n$ to the set of 0-dimensional closed subschemes of length $n$ (or, equivalently, to the set of ideals $I \subseteq k[x, y]$ of colength $n$). We often identify $I$ with its associated point in $H^n$, allowing us to write, for example, $I \in H^n$.

### 2.2. The affine open subschemes $U_\mu$. In [7], M. Haiman obtains a finite covering of $H^n$ by affine open subschemes $U_\mu$, where $\mu$ runs through the partitions of $n$. Given the partition $\mu = (p_1, p_2, \ldots, p_\ell)$, with (positive) parts listed in decreasing order, we define the set of $n$ monomials $B_\mu = \{x^h y^k | 0 \leq h < \ell, 0 \leq k < p_{h+1}\}$.

When displayed as an array with rows indexed by $x$-degree and columns indexed by $y$-degree (starting at 0 in each case), $B_\mu$ yields a diagram of the partition with the rows representing the parts. An example should suffice to make the idea clear; witness the set $B_{(5,3,2,2)}$:

\[
\begin{array}{ccccccc}
x^3 & x^3 y & & & & & \\
x^2 & x^2 y & & & & & \\
x & x y & x y^2 & & & & \\
& 1 & y & y^2 & y^3 & y^4.
\end{array}
\]

We write $(h, k) \in \mu$ to indicate that the inequalities in (4) are satisfied, that is, that $x^h y^k \in B_\mu$. We call the elements of $B_\mu$ partition monomials. Following [7, p. 206],

[w]e now define

$$U_\mu = \{ I \in H^n | B_\mu \text{ spans } k[x, y]/I \}.$$
Here we really mean that the image of $\mathfrak{B}_\mu$ modulo $I$ spans $k[x,y]/I$. Of course this makes $\mathfrak{B}_\mu$ a basis modulo $I$, since $\dim_k(k[x,y]/I) = n$. Since $\mathfrak{B}_\mu$ is a basis, for each monomial $x^r y^s$ and ideal $I \in U_\mu$ there is a unique expansion

$$x^r y^s = \sum_{(h,k) \in \mu} c_{hk}^{rs}(I)x^h y^k \pmod{I},$$

whose coefficients depend on $I$ and thus define a collection of functions $c_{hk}^{rs}$ on $U_\mu$.

We have:

**Proposition 2.2.1** (Haiman [7, Proposition 2.1]). The sets $U_\mu$ are open affine subvarieties which cover $H^n$. The affine coordinate ring $\mathcal{O}_{U_\mu}$ is generated by the functions $c_{hk}^{rs}$, for $(h,k) \in \mu$ and all $(r,s)$.

[We do not quote the proof.]

### 2.3. Universal property of $U_\mu$.

For the proof of Theorem 5.1.1, we need to construct a map with $U_\mu \subseteq H^n$ as target. By the universal property of $H^n$, such a map corresponds to a suitable family of subschemes over the source of the map; we make precise one formulation of “suitable family” in Proposition 2.3.2.

**Lemma 2.3.1.** Let $T$ be a separated scheme of finite type over $k$, and $f: T \to H^n$ a map such that for every $k$-rational point $t \in T$ we have that $f(t) \in U_\mu$. Then $f$ factors through the inclusion $U_\mu \hookrightarrow H^n$.

**Proof.** It suffices to show that if $x \in T$ is an arbitrary scheme-theoretic point, then $f(x) \in U_\mu$. If not, then $f(\overline{x}) \subseteq H^n - U_\mu$; however, since $k$ is algebraically closed, the $k$-rational points of $T$ are very dense [6, Corollaire 6.5.3, p. 309], whence $\overline{x}$ contains a $k$-rational point that perforce maps to the complement of $U_\mu$ under $f$, a contradiction. \qed

$U_\mu$ now inherits the following universal property from $H^n$:

**Proposition 2.3.2.** Let $T$ be as in the lemma. Then the set of maps $f: T \to U_\mu$ is in natural bijective correspondence with the set of closed subschemes $Z_f \subseteq T \times \mathbb{A}_k^2$ that are finite and flat of degree $n$ over $T$, and such that the fiber over every $k$-point $t \in T$ is cut out by an ideal $I_t \subseteq k[x,y]$ having the set of partition monomials $\mathfrak{B}_\mu$ as a $k$-basis of the quotient $k[x,y]/I_t$ (briefly, $I_t \in U_\mu$).

**Proof.** A closed subscheme $Z_f$ as described in the proposition is uniquely associated with a map $f: T \to H^n$ by the universal property of $H^n$ (3), and this map must factor through $U_\mu$ by the lemma. It is clear that the correspondence $Z_f \leftrightarrow f$ is bijective, and that $Z_f$ is obtained as the pullback
of the restriction to $U_\mu \times A^2_k$ of the universal closed subscheme $Z_n \subseteq H^n \times A^2_k$. 

2.4. Nonsingularity of $H^n$. It is well-known that $\text{Hilb}_X^n$ is irreducible and nonsingular whenever $X$ is an irreducible and nonsingular surface [4]. Haiman gives a delightful proof that these facts hold for $H^n$ [7, Proposition 2.4, pp. 208-211]. The proof begins with the observation that

[t]he two-dimensional torus group

$$T^2 = \{(t, q) \mid t, q \in k^*\}$$

acts algebraically on $A^2_k$ by $(t, q) \cdot (\xi, \zeta) = (t\xi, q\zeta)$, or equivalently on $k[x, y]$ by $(t, q) \cdot x = tx, (t, q) \cdot y = qy$. There is an induced action on $H^n$ which, since (7) must remain invariant, is given by $(t, q) \cdot c_{hk}^{rs} = tr^{-h}q^{s-k}c_{hk}^{rs}$. One must take care in computing $(t, q) \cdot I$ for $I \in H^n$ to remember that this means the pullback of $I$ via the homomorphism $(t, q) : k[x, y] \to k[x, y]$, given by $(t, q) \cdot I = \{p(t^{-1}x, q^{-1}y) \mid p(x, y) \in I\}$.

Haiman observes that the $T^2$ fixed points of $H^n$ are exactly the points corresponding to monomial ideals. Every monomial ideal has the form

$$I_\mu = \{x^p y^q \mid (p, q) \notin \mu\}$$

for some partition $\mu$ of $n$, and every partition $\mu$ gives rise to a monomial ideal in this way. Note that the subscheme

$$Z_\mu = \text{Spec}(k[x, y]/I_\mu)$$

is concentrated at the origin of $A^2_k$, and that $I_\mu \in U_\mu$. Haiman proves [7, Lemma 2.3, p. 209] that every ideal $I \in H^n$ has a torus fixed point in the closure of its orbit (in fact, the initial ideal of $I$ for the lexicographic monomial ordering with $y > x$ is a monomial ideal with this property). Since the singular locus of $H^n$ is closed and $T^2$-stable, it must either be empty or contain a monomial ideal. Therefore, to prove that $H^n$ is nonsingular, it suffices to show that each monomial ideal in $H^n$ is a nonsingular point. Haiman does this by explicitly constructing local parameters at each such point $I_\mu$. In his words [7, p. 210],

[t]he maximal ideal $m$ of $I_\mu$ in $0_{U_\mu}$ is given by

$$m = \{c_{hk}^{rs} \mid (h, k) \in \mu, (r, s) \notin \mu\}.$$  

(For $(r, s) \in \mu$, we have $c_{hk}^{rs} = 0$ identically for $(h, k) \neq (r, s)$, and $c_{rs}^{rs} = 1$, so we omit these $c_{hk}^{rs}$ from the ideal.) ... [Consider now the diagram of the monomials in $B_\mu$, as in (5).] We single out two special coordinate functions $c_{hk}^{rs}$ for each
Let \((f, k)\) be the top [entry] in column \(k\) and let \((h, g)\) be the last [entry] in row \(h\). This given, let

\[
\begin{align*}
u_{h,k} &= c_{f,k}^{h,g+1}, \\
d_{h,k} &= c_{f+1,k}^{h,g}.
\end{align*}
\tag{10}
\]

These will be our spanning parameters for \(m/m^2\).

For example, when \(\mu = (5, 3, 2, 2)\), as in (5), and \((h, k) = (0, 0)\), we have

\[
u_{0,0} = c_{3,0}^{0,5} \text{ and } d_{0,0} = c_{0,4}^{4,0},
\]

it is a good exercise to list the remaining 22 spanning parameters in this case.

Haiman’s proof that the functions (10) span \(m/m^2\) over \(k\) proceeds as follows [7, pp. 210-211]:

- Multiplying (7) through by \(x\), then expanding each term on the right by (7) again and comparing coefficients yields the identity

\[
c^{r+1,s}_{h,k} = \sum_{(h',k') \in \mu} c^{rs}_{h'k'} c^{h'+1,k'}_{h,k}
\tag{11}
\]

for all \((h, k) \in \mu\) and all \((r, s)\). Proceeding similarly with \(y\) in place of \(x\) yields

\[
c^{r,s+1}_{h,k} = \sum_{(h',k') \in \mu} c^{rs}_{h'k'} c^{h',k'+1}_{h,k}
\tag{12}
\]

Modulo \(m^2\), the terms \(c^{rs}_{h'k'} c^{h'+1,k'}_{h,k}\) on the right-hand side of (11) reduce to zero for \((h' + 1, k') \notin \mu\) and for \((h' + 1, k') \in \mu\), \((h' + 1, k') \neq (h, k)\) [here we are assuming that \((r, s) \notin \mu\), so that \(c^{rs}_{h,k} \in \mathfrak{m}\)]. The remaining term is \(c^{rs}_{h-1,k}\), or zero if \(h = 0\). Corresponding reductions apply to the right-hand side of (12). Thus in \(m/m^2\) we have

\[
\begin{align*}
c^{r+1,s}_{h,k} &= c^{rs}_{h-1,k}, \quad \text{or if } h = 0; \\
c^{r,s+1}_{h,k} &= c^{rs}_{h,k-1}, \quad \text{or if } k = 0.
\end{align*}
\tag{13}
\]

It is convenient to depict each \(c^{rs}_{h,k}\) by an arrow from \((r, s)\) to \((h, k)\), as shown [in Figure 2]: Equations (13) say that we may move these arrows horizontally or vertically without changing their values modulo \(m^2\), provided we keep the head inside \(\mu\) and the tail outside. More generally, as long as we keep the tail in the first quadrant and outside \(\mu\), we may even move the head across the \(x\)- or \(y\)-axis. When this is
Figure 2. $c_{hk}^{rs}$ represented as an arrow.

possible, the value of the arrow is zero. [The passage goes on to verify that any arrow representing one of the generators of $m$ (9) can be translated horizontally and/or vertically (mod $m^2$) until it has been shown either to be zero or to coincide with one of the functions (10); whence, the latter span $m/m^2$, and therefore constitute a set of local parameters at $I_{\mu}$.]

2.5. Local parameters at every point of $H^n$. Haiman’s construction of local parameters at monomial ideals in fact yields explicit local parameters at every ideal $I \in H^n$. Given $I$, we begin by computing the initial ideal $\text{in}(I) = I_{\mu}$ for the lexicographic monomial ordering with $y > x$. We write

$$p_{\mu} = \{u_{h,k}, d_{h,k} \mid (h, k) \in \mu\}$$

for the set of local parameters (10) at $I_{\mu}$, which is accordingly a $k$-algebraically independent set; the inclusion $k[p_{\mu}] \subseteq \mathcal{O}_{U_{\mu}}$ yields the morphism

$$\varepsilon_{\mu}: U_{\mu} \to \text{Spec}(k[p_{\mu}]) = A_{k}^{2n}.$$ We have the following:

**Lemma 2.5.1.** The map $\varepsilon_{\mu}$ is scheme-theoretically dominant [6, 5.4, p. 283] and étale at $I_{\mu} \in U_{\mu}$.

*Proof.* The first assertion follows at once from the definition and the injectivity of the map $\varepsilon_{\mu}^*: k[p_{\mu}] \to \mathcal{O}_{U_{\mu}}$. The second assertion follows from [1, Corollary 4.5, p. 116], since the induced map on completions $\hat{\varepsilon}_{\mu}^*: \hat{\mathcal{O}}_0 \to \hat{\mathcal{O}}_{I_{\mu}}$ is the identity $k[[p_{\mu}]] \to k[[p_{\mu}]]$. \qed

We claim that $\varepsilon_{\mu}$ is étale at $I$. To see this, observe first of all that $\text{Spec}(k[p_{\mu}]) = A_{k}^{2n}$ inherits a $T^2$-action from $H^n$; for $(t, q) \in T^2$, the comorphism of the map

$$(t, q): \text{Spec}(k[p_{\mu}]) \to \text{Spec}(k[p_{\mu}])$$

is defined by $(t, q) \cdot c_{hk}^{rs} = t^{-h}q^{-k}c_{hk}^{rs}$ for $c_{hk}^{rs} \in p_{\mu}$. In other words, the action is defined by restricting the comorphism of $(t, q): U_{\mu} \to U_{\mu}$ to $k[p_{\mu}] \subseteq \mathcal{O}_{U_{\mu}}$; it follows at once that the left-hand diagram in Figure 3 is commutative for all $(t, q) \in T^2$. 

![Diagram](https://via.placeholder.com/150)
Figure 3. Commutative diagrams associated with $\varepsilon_\mu$.

The right-hand square of tangent spaces and $k$-linear maps in Figure 3 is induced by the left-hand square; to show that $\varepsilon_\mu$ is étale at $I$, we must show that the map $d\varepsilon_\mu(I)$ is an isomorphism. However, since $\varepsilon_\mu$ is étale at $I_\mu$, and therefore in a neighborhood thereof (see, e.g., [1, Proposition 4.6, p. 116]), and since $I_\mu$ lies in the closure of the $T^2$-orbit of $I$, as stated in Section 2.4, we can choose $(t, q) \in T^2$ so that $d\varepsilon_\mu((t, q) \cdot I)$ is an isomorphism. Since the horizontal arrows are clearly isomorphisms, we conclude that $d\varepsilon_\mu(I)$ is an isomorphism, as desired. We restate our conclusion as:

**Proposition 2.5.2.** Let $I \in H^n$ and let $I_\mu$ be the initial ideal of $I$ for the lexicographic monomial order with $y > x$. Then the set of functions
\[
\{ c_{rhk}^{rs} - c_{rhk}^{rs}(I) \mid c_{rhk}^{rs} \in p_\mu \}
\]
is a set of local parameters at $I \in H^n$.

In the next section, we exhibit a certain finite subset $c_\mu$ of the (infinite set of) $c_{rhk}^{rs}$ that generates $\mathcal{O}_{U_\mu}$ as a $k$-algebra; the set $c_\mu$ contains the set $p_\mu$ of local parameters at $I_\mu$. (Please note that the subset $c_\mu$ is chosen for convenience; it is typically far from a minimal generating set.) In order to prove that $c_\mu$ does in fact generate $\mathcal{O}_{U_\mu}$, we need the relations (11), (12). Later, in Section 7, we show that $\mathcal{O}_{U_\mu}$ is in fact generated by a subset $e_\mu \subseteq c_\mu$ that also contains $p_\mu$; the subset $e_\mu$ is typically much smaller than $c_\mu$ (see (58) and the following example), but again may not be a minimal generating set. In certain special cases, one has either that $p_\mu = e_\mu$ (Section 7.3) or that $e_\mu \subseteq k[p_\mu]$ (Section 7.4, where the relations (11), (12) again play a key role), from which follows that $\mathcal{O}_{U_\mu}$ is the polynomial ring $k[p_\mu]$, or, equivalently, that the map $\varepsilon_\mu$ (15) is an isomorphism. Of course, whenever $p_\mu$ generates $\mathcal{O}_{U_\mu}$, it is a minimal generating set (since it has cardinality $2n$ equal to the dimension of $U_\mu$), but there are partitions $\mu$ for which $p_\mu$ fails to generate $\mathcal{O}_{U_\mu}$. I do not know how to find a minimal generating set in all cases.

**3. A finite set of generators of $\mathcal{O}_{U_\mu}$.**

We have two objectives in this section: The first is to demonstrate that a certain finite subset $c_\mu$ of the functions $c_{rhk}^{rs}$ generates the affine coordinate
ring $\mathcal{O}_{U_\mu}$ as a $k$-algebra. The second is to show that certain polynomials associated to the set $\mathfrak{c}_\mu$ form a basis for the ideal

$$\mathfrak{I}_\mu \subseteq \mathcal{O}_{U_\mu}[x,y]$$

that cuts out the universal closed subscheme over $U_\mu$.

3.1. Leading monomials and generators of $\mathcal{O}_{U_\mu}$. We begin with a partition $\mu = (p_1, p_2, \ldots, p_\ell)$ of $n$ (with parts listed in decreasing order) which we view as the set of partition monomials $\mathfrak{B}_\mu$ (4) arranged in rows and columns as in (5). We write

$$\begin{align*}
\ell &= \text{the number of parts of } \mu \\
&= \text{the number of rows of } \mathfrak{B}_\mu; \\
d_\mu &= \text{the number of distinct parts of } \mu; \\
\mu(i) &= \text{the number of times the integer } i \text{ occurs in } \mu; \\
p_1 &= \text{the largest part in } \mu \\
&= \text{the number of columns of } \mathfrak{B}_\mu; \\
p_\ell &= \text{the smallest part in } \mu.
\end{align*}$$

We say that a monomial $x^r y^s$ is a leading monomial of $\mu$ if it lies on the “boundary” of $\mathfrak{B}_\mu$ — either immediately above a column or immediately to the right of a row. For example, when $\mu = (5,3,2,2)$, the leading monomials are shown in bold in Figure 4. We call the leading monomials situated

Figure 4. The partition monomials for $\mu = (5,3,2,2)$ enclosed in a box, with the leading monomials shown in boldface. In this case, $\ell = 4$, $d_\mu = 3$, $\mu(2) = 2$, $\mu(3) = 1$, $\mu(5) = 1$, $p_1 = 5$, and $p_\ell = p_4 = 2$.

above the columns (resp. to the right of the rows) of $\mathfrak{B}_\mu$ the top (resp. side) monomials associated to $\mu$. Because $d_\mu - 1$ of the leading monomials (those in the “notch” positions — $x^2 y^2$ and $x = y^3$ in Figure 4) are both top and side monomials, we have in general that

$$\begin{align*}
\Lambda_\mu &= \text{the number of leading monomials} \\
&= \text{(# top monomials)} + \text{( # side monomials)} - (d_\mu - 1) \\
&= p_1 + \ell - d_\mu + 1.
\end{align*}$$
We index the leading monomials $x^r y^s$ from $j = 1$ to $\Lambda_{\mu}$ by starting at the upper left of the diagram of $\mu$ and traversing the boundary in a clockwise fashion. In the example, this yields the sequence

$$x^4, x^4 y, x^3 y^2, x^2 y^2, xy^3, xy^4, y^5.$$  

**Remark 3.1.1.** Observe that the first (resp. last) leading monomial in the sequence is $x^{r_1} = x^1$ (resp. $y^{s_\Lambda} = y^{p_1}$).

For $I \in U$, form the expansion (7) for each of the leading monomials $x^r y^s$, $1 \leq j \leq \Lambda_{\mu}$, to obtain

$$x^r y^s = \sum_{(h,k) \in \mu} c_{hk}^r y^s (I) x^h y^k \pmod{I};$$

recall that the coefficients can be viewed as functions on $U_{\mu}$. Our generating set $c_{\mu}$ consists of all the coefficient functions appearing in (18), that is,  

$$c_{\mu} = \{c_{hk}^r y^s \mid 1 \leq j \leq \Lambda_{\mu}, \ (h,k) \in \mu\};$$

note that the cardinality of $c_{\mu}$ is

$$|c_{\mu}| = n \cdot \Lambda_{\mu},$$

and that $p_{\mu} \subseteq c_{\mu}$ (immediate from the definition of $p_{\mu}$ (14)). Of course, we should not refer to $c_{\mu}$ as a generating set until we have proven it so:

**Proposition 3.1.2.** The set $c_{\mu}$ generates the affine coordinate ring $O_{U_{\mu}}$ as an algebra over $k$.

**Proof.** By Proposition 2.2.1, the ring $O_{U_{\mu}}$ is generated by the set of all $c_{hk}^r y^s$, so it suffices to show that the subset $c_{\mu}$ generates every $c_{hk}^r y^s$. In fact, we only need to do this for $(r,s) \notin \mu$, since, as was observed following (9), $c_{hk}^r y^s$ is identically either 0 or 1 if $(r,s) \in \mu$. We say that the pair $(r,s) \notin \mu$ is **covered** if the function $c_{hk}^r y^s$ is in the subring of $O_{U_{\mu}}$, generated by $c_{\mu}$ for all $(h,k) \in \mu$. We must show that every pair $(r,s) \notin \mu$ is covered; it is clear at the outset that this is so for each of the pairs $(r_j, s_j)$ associated to the leading monomials.

The relation (11) implies that the monomial $x^r y^s$ is covered provided that

1) its “downstairs” neighbor $x^{r-1} y^s$ is covered, and  
2) every pair $(h+1, k) \notin \mu$, with $(h,k) \in \mu$, is covered.

However, the pairs $(h+1, k)$ in item 2 are exactly the pairs $(r_j, s_j)$ associated to the top leading monomials (see Figure 4), and we know that all such pairs are covered. Therefore, whenever $(r,s)$ is covered, we can proceed inductively to conclude that $(r+1,s)$, $(r+2,s)$, \ldots, are covered as well. Similarly, using relation (12) and working horizontally rather than vertically, we see that whenever $(r,s)$ is covered, so too are $(r, s+1)$, $(r, s+2)$, \ldots. Since it is clear that every pair $(r,s) \notin \mu$ can be reached by traversing a path
beginning at a leading monomial and consisting of vertical and/or horizontal segments, it follows that every such pair \((r,s)\) is covered, as desired. □

**Remark 3.1.3.** Recall that, as stated at the end of Section 2, the generating set given by Proposition 3.1.2 is typically far from minimal. For example, the proof of the proposition can be modified to show that the subset

\[ \{ c_{hk}^{r_j,s_j} \mid x^{r_j} y^{s_j} \text{ is either a top monomial or } y^{p_1}, \ (h,k) \in \mu \} \subseteq c_{\mu} \]

also generates \( O_{U_{\mu}} \); this subset is a proper subset of \( c_{\mu} \) whenever there is at least one side monomial \( x^{r_j} y^{s_j} \) with \( j < \Lambda_{\mu} \) that is not also a top monomial (such as \( x^3 y^2 \) in Figure 4). At the moment, we can assert that a set of \( k \)-algebra generators of \( O_{U_{\mu}} \) of minimal cardinality must have at least \( 2n \) members (for reasons of dimension), and can have at most \( \Lambda_{\mu} n \) members (by (20)); we will tighten the upper bound in Section 7.

**Remark 3.1.4.** Recalling (9), it is clear that the monomial ideal \( I_{\mu} \subseteq k[x,y] \) (8) is the point of \( U_{\mu} \) at which all the functions in our generating set \( c_{\mu} \) vanish.

### 3.2. The ideal of the universal closed subscheme.

We now consider the restriction of the universal closed subscheme \( Z_n \subseteq \text{Hilb}^n A^2_k \) (see Section 2.1 and the proof of Proposition 2.3.2) to \( U_{\mu} \); we denote the restricted subscheme by \( Z_{U_{\mu}} \), and write

\[ Z_{U_{\mu}} = \text{Spec}(O_{U_{\mu}}[x,y]/\mathcal{I}_{\mu}) \tag{21} \]

By definition, the ring \( O_{U_{\mu}}[x,y]/\mathcal{I}_{\mu} \) is finite and flat of degree \( n \) over \( O_{U_{\mu}} \). Moreover, for any point \( I \in U_{\mu} \), given as a map \( i: \text{Spec}(k) \to U_{\mu} \) with comorphism \( i^*: O_{U_{\mu}} \to k \), we have that

\[ k \otimes_{O_{U_{\mu}}} O_{U_{\mu}}[x,y]/\mathcal{I}_{\mu} = k[x,y]/I; \tag{22} \]

that is, \( \mathcal{I}_{\mu} \) extends to \( I \) under the map \( O_{U_{\mu}}[x,y] \to k[x,y] \) induced by \( i^* \). We seek to exhibit a basis of the ideal \( \mathcal{I}_{\mu} \); to this end, we define, for each leading monomial \( x^{r_j} y^{s_j} \), the polynomial

\[ g_j = x^{r_j} y^{s_j} - \sum_{(h,k) \in \mu} c_{hk}^{r_j,s_j} x^h y^k \in O_{U_{\mu}}[x,y]. \tag{23} \]

If the leading monomial \( x^{r_j} y^{s_j} \) is a top (resp. side) monomial of \( \mu \), then by extension we refer to \( g_j \) as a top (resp. side) **polynomial of** \( \mu \).

**Proposition 3.2.1.** The ideal \( \mathcal{I}_{\mu} \) is generated by the \( g_j \), \( 1 \leq j \leq \Lambda_{\mu} \); in symbols, \( \mathcal{I}_{\mu} = (g_1, g_2, \ldots, g_{\Lambda_{\mu}}) \).

For the proof, we need two lemmas.

**Lemma 3.2.2.** The quotient \( O_{U_{\mu}}[x,y]/\mathcal{I}_{\mu} \) is free of rank \( n \) over \( O_{U_{\mu}} \), with the (image of the) set \( \mathfrak{B}_{\mu} \) of partition monomials constituting a basis.
Proof. We have already noted that \( \mathcal{O}_{U_\mu}[x,y]/\mathfrak{J}_\mu = Q \) is finite and flat of degree \( n \) over \( \mathcal{O}_{U_\mu} = A \); whence, the sheaf \( \tilde{Q} \) is locally free of degree \( n \) on \( U_\mu = \text{Spec}(A) \). By definition of \( U_\mu \) (6), (the image of) \( \mathfrak{B}_\mu \) yields a basis of \( k[x,y]/I \) for every \( I \in U_\mu \); it follows from Nakayama’s lemma that \( \mathfrak{B}_\mu \) generates (and therefore gives a basis of) \( \tilde{Q} \) in a local neighborhood of every point of \( U_\mu \). Consequently, the map \( A^n \to Q \) induced by the \( n \)-element set \( \mathfrak{B}_\mu \) localizes to an isomorphism everywhere on \( U_\mu \), and is therefore itself an isomorphism. \( \Box \)

Lemma 3.2.3. The quotient \( \mathcal{O}_{U_\mu}[x,y]/(g_1, g_2, \ldots, g_\Lambda_\mu) \) is generated as an \( \mathcal{O}_{U_\mu} \)-module by the (images of the) partition monomials.

Proof. We write \( (g_1, g_2, \ldots, g_\Lambda_\mu) = \mathfrak{G} \) and, as before, \( \mathcal{O}_{U_\mu} = A \). It suffices to prove that every monomial in \( x \) and \( y \) is congruent (mod \( \mathfrak{G} \)) to an \( A \)-linear combination of partition monomials. This is immediate for the partition monomials themselves, and is also clearly true for the leading monomials, by definition of the \( g_j \). We may therefore proceed by induction on the total degree of a monomial, the base case having already been checked since \( x^0y^0 = 1 \) is always a partition monomial.

Suppose therefore that every monomial of total degree \( < r+s \) is congruent (mod \( \mathfrak{G} \)) to an \( A \)-linear combination of partition monomials, and consider the monomial \( x^r y^s \). If \( r > 0 \), we have that \( x^r y^s = x \cdot x^{r-1} y^s \); whence, by the induction hypothesis,

\[
x^r y^s = x \cdot x^{r-1} y^s \equiv x \cdot \left( \sum_{(h,k) \in \mu} a_{hk}^{r-1,s} x^h y^k \right) \equiv \sum_{(h,k) \in \mu} a_{hk}^{r-1,s} x^{(h+1)} y^k \pmod{\mathfrak{G}},
\]

where the coefficients \( a_{hk}^{r-1,s} \in A \). If any of the monomials \( x^{h+1} y^k \) in the last sum are not partition monomials, then they are top leading monomials, and can accordingly be expanded (mod \( \mathfrak{G} \)) as \( A \)-linear combinations of partition monomials, showing that such an expansion also obtains for \( x^r y^s \), as desired. If \( r = 0 \), then \( s > 0 \), and the proof is similar. \( \Box \)

Proof of Proposition 3.2.1. We first show that \( g_j \in \mathfrak{J}_\mu \) for \( 1 \leq j \leq \Lambda_\mu \). By Lemma 3.2.2, we have that each leading monomial \( x^{r_j} y^{s_j} \) is congruent (mod \( \mathfrak{J}_\mu \)) to a unique \( \mathcal{O}_{U_\mu} \)-linear combination of partition monomials. Put another way, \( \mathfrak{J}_\mu \) contains \( \Lambda_\mu \) uniquely determined polynomials of the form

\[
(24) \quad x^{r_j} y^{s_j} - \sum_{(h,k) \in \mu} d_{hk}^{r_j,s_j} x^h y^k, \quad 1 \leq j \leq \Lambda_\mu, \quad d_{hk}^{r_j,s_j} \in A.
\]

Specializing to any point \( I \in U_\mu \), we see that \( c_{hk}^{r_j,s_j}(I) = d_{hk}^{r_j,s_j}(I) \) must hold; in other words, the functions \( c_{hk}^{r_j,s_j} \) and \( d_{hk}^{r_j,s_j} \) have the same \( k \)-values at every point. Since \( \mathcal{O}_{U_\mu} \) is an integral domain (in particular, reduced), we conclude
that
\[ c_{hk}^{r_j,s_j} = d_{hk}^{r_j,s_j}, \quad 1 \leq j \leq \mu, \quad (h,k) \in \mu, \]
which implies that the two sets of polynomials (24) and (23) are the same.
It follows that each \( g_j \) is in \( \mathcal{I} \), as desired; whence,
\[ (g_1, g_2, \ldots, g_{\mu}) = \mathcal{G} \subseteq \mathcal{I}_\mu. \]

Now set \( Q' = \mathcal{O}_{U_\mu}[x,y]/\mathcal{G} \), and, as before, \( A = \mathcal{O}_{U_\mu} \) and \( Q = A[x,y]/\mathcal{I}_\mu \).
Since \( \mathcal{G} \) maps to 0 under the quotient map \( A[x,y] \to Q \), we obtain an \( A \)-linear surjection \( \alpha: Q' \to Q \) which maps the coset \( x^h y^k + \mathcal{G} \) to the coset \( x^h y^k + \mathcal{I}_\mu \) for all \((h,k) \in \mu\). By Lemma 3.2.2, we obtain an \( A \)-linear map \( \beta: Q \to Q' \) by sending \( x^h y^k + \mathcal{I}_\mu \) to \( x^h y^k + \mathcal{G} \) and extending linearly. Since the cosets \( x^h y^k + \mathcal{G} \) generate \( Q' \) as an \( A \)-module, by Lemma 3.2.3, and are mapped to themselves by the composition \( \beta \circ \alpha \), we have that \( \beta \circ \alpha \) is the identity map; whence \( \alpha \) is injective (as well as surjective), and therefore an isomorphism. It follows at once that \( \mathcal{G} = \mathcal{I}_\mu \), which is the desired conclusion.

**Remark 3.2.4.** The generating set of \( \mathcal{I}_\mu \) given by Proposition 3.2.1 is not in general minimal. Indeed, it can be shown that the subset
\[ \{ g_j \mid g_j \text{ is a top polynomial or } g_{\lambda_\mu} \} \subseteq \{ g_1, \ldots, g_{\lambda_\mu} \} \]
is a generating set. For example, consider the partition shown in Figure 4. The displayed subset of \( \{ g_1, g_2, \ldots, g_{\lambda_\mu} \} \) omits only the polynomial with leading term \( x^3 y^2 \); let \( \mathcal{G}' \subseteq \mathcal{I}_\mu \) denote the ideal generated by this subset. By multiplying the polynomial with leading term \( x^3 y^2 \) by \( x \) and adding appropriate multiples of the \( g \)'s with top leading monomials, we see that the ideal \( \mathcal{G}' \) contains a polynomial of the form
\[ x^3 y^2 - \sum_{(h,k) \in \mu} d_{hk} x^h y^k, \quad d_{hk} \in \mathcal{O}_{U_\mu}, \]
which, lying in \( \mathcal{I}_\mu \), must in fact equal the omitted \( g \) polynomial with leading monomial \( x^3 y^2 \) (the unique polynomial of its form in \( \mathcal{I}_\mu \)). It follows that \( \mathcal{I}_\mu = \mathcal{G}' \).

In view of (22), Proposition 3.2.1 yields:

**Corollary 3.2.5.** For \( I \in U_\mu \), we have that
\[ I = (g_1(I), g_2(I), \ldots, g_{\lambda_\mu}(I)) \subseteq k[x,y], \]
where \( g_j(I) \) denotes the polynomial obtained from \( g_j \) by replacing each coefficient function \( c_{hk}^{r_j,s_j} \) by its value \( c_{hk}^{r_j,s_j}(I) \) at the point \( I \).

In the next section, we present a basis of the first syzygy module of \( \{ g_1, g_2, \ldots, g_{\lambda_\mu} \} \), which leads to a free resolution of the \( \mathcal{O}_{U_\mu} \)-module \( \mathcal{O}_{U_\mu}[x,y]/\mathcal{I}_\mu \).
4. A free resolution of \( \mathcal{O}_{U_\mu}[x,y]/\mathfrak{I}_\mu \).

In this section we study the relations among the generators \( g_j \) (23) of the ideal \( \mathfrak{I}_\mu \) (21). We present an algorithm that yields a basis for the first syzygy module of these generators. (In preparation for subsequent sections, we study the syzygy algorithm in greater generality than is required for the purposes of this section.) As a corollary, we obtain a free resolution of the \( \mathcal{O}_{U_\mu} \)-module \( \mathcal{O}_{U_\mu}[x,y]/\mathfrak{I}_\mu \). We then show that the \( g_j \) are recovered (up to sign) as the maximal minors of the matrix whose rows are the elements of the syzygy basis.

4.1. Syzygies of \( (g_1, \ldots, g_{\Lambda_\mu}) \). Recall that a syzygy of \( (g_1, \ldots, g_{\Lambda_\mu}) \) is a \( (\Lambda_\mu) \)-tuple \( (f_1, \ldots, f_{\Lambda_\mu}) \) of elements of \( \mathcal{O}_{U_\mu}[x,y] \) such that \( \sum_{j=1}^{\Lambda_\mu} (f_j \cdot g_j) = 0 \); the set of all such syzygies is a submodule \( \text{Syz}_\mu \subseteq (\mathcal{O}_{U_\mu}[x,y])^{\Lambda_\mu} \), the first syzygy module of \( (g_j) \). We will show that this syzygy module is free of rank \( \Lambda_\mu - 1 \), with an easily-obtained basis. The key observation for finding syzygies is the following:

**Lemma 4.1.1.** If \( (f_1, \ldots, f_{\Lambda_\mu}) \in (\mathcal{O}_{U_\mu}[x,y])^{\Lambda_\mu} \) is such that the polynomial \( \sum_{j=1}^{\Lambda_\mu} (f_j \cdot g_j) \) is an \( \mathcal{O}_{U_\mu} \)-linear combination of partition monomials, then in fact \( (f_1, \ldots, f_{\Lambda_\mu}) = (f_j) \) is a syzygy of \( (g_j) \).

**Proof.** The polynomial \( L = \sum_{j=1}^{\Lambda_\mu} (f_j \cdot g_j) \) lies in the ideal \( \mathfrak{I}_\mu \) generated by the \( g_j \), and so \( L \equiv 0 \) (mod \( \mathfrak{I}_\mu \)). But by Lemma 3.2.2, \( \mathcal{O}_{U_\mu}[x,y]/\mathfrak{I}_\mu \) is free over \( \mathcal{O}_{U_\mu} \) with the partition monomials constituting a basis. Therefore, the only \( \mathcal{O}_{U_\mu} \)-linear combination of the partition monomials that \( L \) could equal is the trivial one, which implies that \( (f_j) \) is a syzygy of \( (g_j) \). \( \square \)

The basic idea is therefore to find linear combinations of the \( g_j \) that involve only partition monomials. For example, consider again the special case \( \mu = (5, 3, 2, 2) \) shown in Figure 4. We will build a syzygy of \( (g_1, \ldots, g_7) \); we begin by multiplying \( g_1 \) by \( -y \):

\[
(25) \quad -y \cdot g_1 = -y \cdot \left( x^4 - c_{3,0}^4 \cdot x^3 - c_{3,1}^4 \cdot x^3 y - c_{2,0}^4 \cdot x^2 - c_{2,1}^4 \cdot x^2 y - c_{1,0}^4 \cdot x - c_{1,1}^4 \cdot xy - c_{0,2}^4 \cdot y^2 - c_{0,3}^4 \cdot y^3 - c_{0,4}^4 \cdot y^4 \right).
\]

With reference to Figure 4, the multiplication by \( -y \) shifts each term in \( g_1 \) one place to the right (and changes its sign). One easily lists the terms in (25) that are not scalar multiples of pattern monomials; we shall call such terms exposed terms, and their coefficient functions exposed coefficients:

\[
-x^4 y, \quad c_{3,1}^4 \cdot x^3 y^2, \quad c_{2,1}^4 \cdot x^2 y^2, \quad c_{1,2}^4 \cdot xy^3, \quad c_{0,4}^4 \cdot y^5.
\]
Since these terms involve, respectively, the second, third, fourth, fifth, and seventh leading monomials, one sees easily that the dot product
\[
(-y, 1, -c_{3,1}^4, -c_{2,1}^4, -c_{1,2}^4, 0, -c_{0,4}^4) \cdot (g_1, g_2, \ldots, g_7)
\]
is an \(\mathcal{O}_{U_\mu}\)-linear combination of partition monomials. (Each exposed term in \(-y \cdot g_1\) is cancelled by the addition of the appropriate multiple of the corresponding \(g_j\).) Therefore, by Lemma 4.1.1, the tuple
\[
h_1 = (-y, 1, -c_{3,1}^4, -c_{2,1}^4, -c_{1,2}^4, 0, -c_{0,4}^4)
\]
is a syzygy of \((g_1, \ldots, g_7)\).

To build a second syzygy, we multiply the polynomial \(g_2\), with leading term \(x^4y\), by \(-y\). In this case, one of the exposed terms, namely \(-x^4y^2\), is not a scalar multiple of a leading monomial. To cancel this exposed term, we add \(xg_3\) to the linear combination. Doing this introduces additional exposed terms, all of which are scalar multiples of leading monomials. Therefore, we can proceed as before to add appropriate multiples of the \(g_j\) to cancel all the remaining exposed terms, and thereby obtain the following syzygy:
\[
h_2 = (c_{3,1}^2 - y + c_{3,1}^2, x - c_{3,1}^4, -c_{2,1}^4 + c_{1,2}^4, -c_{1,2}^4 + c_{0,3}^4, c_{0,4}^4, c_{0,4}^4, -c_{1,2}^4).
\]

Continuing in this way, we construct six syzygies; these we gather together as the rows of the following matrix \(m(5,3,2,2)\):

\[
\begin{bmatrix}
-\frac{y}{1} & -c_{3,1}^4 & -c_{2,1}^4 & -c_{1,2}^4 & 0 & -c_{0,4}^4 \\
c_{3,1}^4 & -\frac{y + c_{3,1}^4}{x - c_{3,1}^4} & -c_{2,1}^4 & -c_{1,2}^4 & c_{0,4}^4 & -c_{0,4}^4 \\
\frac{c_{3,1}^4}{x - c_{3,1}^4} & -\frac{y + c_{3,1}^4}{x - c_{3,1}^4} & -c_{2,1}^4 & -c_{1,2}^4 & c_{0,4}^4 & -c_{0,4}^4 \\
-\frac{c_{3,1}^4}{c_{3,1}^4} & -\frac{y - c_{3,1}^4 + c_{1,2}^4 - c_{0,4}^4}{x - c_{3,1}^4} & -\frac{y - c_{3,1}^4 + c_{1,2}^4 - c_{0,4}^4}{c_{3,1}^4} & -\frac{y - c_{3,1}^4 + c_{1,2}^4 - c_{0,4}^4}{c_{0,4}^4} & -\frac{y - c_{3,1}^4 + c_{1,2}^4 - c_{0,4}^4}{c_{0,4}^4} & -\frac{y - c_{3,1}^4 + c_{1,2}^4 - c_{0,4}^4}{c_{0,4}^4} \\
0 & 0 & -c_{3,1}^4 & -c_{2,1}^4 & c_{0,4}^4 & 0 \\
\frac{c_{3,1}^4}{c_{3,1}^4} & -\frac{c_{3,1}^4}{c_{3,1}^4} & -\frac{c_{3,1}^4}{c_{3,1}^4} & -\frac{c_{3,1}^4}{c_{3,1}^4} & -\frac{c_{3,1}^4}{c_{3,1}^4} & -\frac{c_{3,1}^4}{c_{3,1}^4}
\end{bmatrix}
\]

We claim that the six syzygies \(h_1, \ldots, h_6\) in \((26)\) compose an \(\mathcal{O}_{U_\mu}\)-basis of the full syzygy module of \((g_1, \ldots, g_7)\); to prove this, and to prepare the ground for later sections, we turn to a more general treatment of the syzygy-making process.

4.2. \(\mu\)-Pseudosyzygies. Within the confines of this section and the next, we abuse our previous notation by using it in a more general context. Let \(A\) be any commutative ring with identity (in place of \(\mathcal{O}_{U_\mu}\)), \(B = A[x,y]\), and (as before) \(\mathfrak{B}_\mu\) the set of partition monomials \((4)\) associated to a partition \(\mu\) of \(n\). Let \(g_j \in A[x,y], 1 \leq j \leq \Lambda_\mu,\) be polynomials of the form \((23)\), with the coefficients \(c_{nk,\mathfrak{B}_\mu}\) arbitrarily chosen in \(A\). We define a \(\mu\)-pseudosyzygy of \((g_1, \ldots, g_{\Lambda_\mu}) = (g_k)\) to be a \((\Lambda_\mu)\)-tuple \(f = (f_1, \ldots, f_{\Lambda_\mu})\) of elements of \(B\)
that satisfies

\[(g_1, \ldots, g_{\Lambda_\mu}) \cdot f = \sum_{k=1}^{\Lambda_\mu} (g_k \cdot f_k) = \left( \begin{array}{c} \text{an A-linear combination of} \\ \text{partition monomials} \end{array} \right),\]

We can construct \(\mu\)-pseudosyzygies of \((g_k)\) by repeating the process used in Section 4.1. More precisely, we execute:

**Algorithm 4.2.1.** Begin with \(g_1\) and iterate through the \(g_k\), \(1 \leq k < \Lambda_\mu\). At the \(j\)-th iteration, carry out the unique enumerated step that applies to \(g_j\) to produce the \(\mu\)-pseudosyzygy \(h_j\). In all, \(\Lambda_\mu - 1\) \(\mu\)-pseudosyzygies will be produced; note that the \(h_j\) are indexed by the position of the \(-y\) (for Types 1 and 2) or the \(-1\) (for Type 3).

1) If \(g_j\) is a top polynomial and \(x^{r_j}y^{s_j+1}\) is a (top) leading monomial (equal to \(x^{r_j+1}y^{s_j+1}\)) multiply \(g_j\) by \(-y\), add \(1\cdot g_{j+1}\) to cancel \(-x^{r_j}y^{s_j+1}\), and then add appropriate multiples of the \(g_k\) to cancel the remaining exposed monomials, all of which are \(A\)-multiples of (side) leading monomials. We say that the resulting \(\mu\)-pseudosyzygy is of **Type 1**; it has the value

\[h_j = (a_{j,1}, a_{j,2}, \ldots, a_{j,j-1}, -y + a_{j,j}, 1, a_{j,j+2}, \ldots, a_{j,\Lambda_\mu}),\]

where \(a_{j,i} \in A\) is equal to \(-c_{r_i,s_i}^{r_j,s_j}\) when \(g_i\) is a side polynomial, and 0 otherwise.

2) If \(g_j\) is a top polynomial and \(x^{r_j}y^{s_j+1}\) is not a leading monomial, then \(g_{j+1}\) has (side) leading monomial \(x^{r_j+1}y^{s_j+1} = x^{r_j-1}y^{s_j+1}\). Multiply \(g_j\) by \(-y\), add \(x \cdot g_{j+1}\) to cancel \(-x^{r_j}y^{s_j+1}\), and then add appropriate multiples of the \(g_k\) to cancel the remaining exposed terms, all of which are \(A\)-multiples of (top and side) leading monomials. We say that the resulting \(\mu\)-pseudosyzygy is of **Type 2**; it has the value

\[h_j = (a_{j,1}, a_{j,2}, \ldots, a_{j,j-1}, -y + a_{j,j}, x + a_{j,j+1}, a_{j,j+2}, \ldots, a_{j,\Lambda_\mu}),\]

where \(a_{j,i} \in A\) is equal to \((-c_{r_i,s_i}^{r_j,s_j} + c_{r_{j+1},s_i}^{r_j+1,s_j+1})\) when \(g_i\) is both a top and a side polynomial, \(-c_{r_i,s_i}^{r_j,s_j}\) when \(g_i\) is only a side polynomial, and \(c_{r_i-1,s_i}^{r_{j+1},s_j+1}\) when \(g_i\) is only a top polynomial.

3) If \(g_j\) is a side polynomial and \(x^{r_j-1}y^{s_j}\) is a (side) leading monomial (equal to \(x^{r_j+1}y^{s_j+1}\)), multiply \(g_j\) by \(-1\), add \(x \cdot g_{j+1}\) to cancel \(x^{r_j}y^{s_j}\), and then add appropriate multiples of the \(g_k\) to cancel the remaining exposed monomials, all of which are \(A\)-multiples of (top) leading monomials. We say that the resulting \(\mu\)-pseudosyzygy is of **Type 3**; it has the value

\[h_j = (a_{j,1}, a_{j,2}, \ldots, a_{j,j-1}, -1, x + a_{j,j+1}, a_{j,j+2}, \ldots, a_{j,\Lambda_\mu}),\]
where \( a_{j,i} \in A \) is equal to \( c_{r_j+1, s_j+1} \) when \( g_i \) is a top polynomial, and 0 otherwise.

**Remark 4.2.2.** For future reference, we highlight that the \( \mu \)-pseudosyzygies \( h_j \) of Type 1 in Algorithm 4.2.1 have nonzero coefficients \( a_{j,i} \) only in positions corresponding to side monomials; likewise, the \( h_j \) of Type 3 have nonzero coefficients \( a_{j,i} \) only in positions corresponding to top monomials. The matrix of syzygies (26) provides an example.

### 4.3. Main theorem on \( \mu \)-Pseudosyzygies

We retain the notation of the previous section: \( A \) is a commutative ring, \( B = A[x, y] \), \( g_j \in B \), \( 1 \leq j \leq \Lambda_\mu \), are polynomials of the form (23), with the coefficients \( c_{r_j, s_j} \) arbitrarily chosen in \( A \), and \( h_1, \ldots, h_{\Lambda_\mu - 1} \) are the \( \mu \)-pseudosyzygies of \( (g_j) \) produced by Algorithm 4.2.1. We write \( H \subseteq B^{\Lambda_\mu} \) for the \( B \)-linear span of the \( h_k \). We shall soon show that every \( H \)-coset has a unique representative \( f' = (f'_1, \ldots, f'_{\Lambda_\mu}) \) with the following properties:

1) \( f'_k = 0 \) for all \( k < \Lambda_\mu \) such that \( g_k \) is not a top polynomial, and
2) each nonzero \( f'_k \) with \( k < \Lambda_\mu \) is an element of \( A[x] \).

We say that a \( B^{\Lambda_\mu} \)-tuple \( f' \) with these properties is **side-minimized**, because the non-top entries \( f'_k \) for \( k < \Lambda_\mu \) are all 0, and because the other entries (except for the \( \Lambda_\mu \)-th), being elements of \( A[x] \), cause only upward, not sideways, motion of monomials when they multiply their corresponding \( g_k \) (see Figure 4).

Our main result on \( \mu \)-pseudosyzygies is Theorem 4.3.6; since the proof is lengthy, we break it up into a series of lemmas.

**Lemma 4.3.1.** A nontrivial \( B \)-linear combination of the \( \mu \)-pseudosyzygies \( h_1, \ldots, h_{\Lambda_\mu - 1} \) can never be side-minimized. In particular, the \( h_k \) form a \( B \)-linearly independent set.

**Proof.** Suppose given a nontrivial \( B \)-linear combination

\[
\gamma = \gamma_1 \cdot h_1 + \gamma_2 \cdot h_2 + \cdots + \gamma_{\Lambda_\mu - 1} \cdot h_{\Lambda_\mu - 1}
\]

that is side-minimized. Let \( \gamma_q \) have maximal \( y \)-degree among the nonzero \( \gamma_k \in B \). If \( h_q \) is of Type 1 or Type 2 (see Algorithm 4.2.1), so that the \( q \)-th component of \( h_q \) has the form \( -y + a_{q,q} \), we have that the \( q \)-th component of \( \gamma \) has the form

\[
\gamma_q \cdot (-y + a_{q,q}) + \sum_{k \neq q} \gamma_k \cdot (a_{k,q} \text{ or } x + a_{k,q}), \quad a_{q,q}, a_{k,q} \in A.
\]

A moment’s reflection shows that the terms of maximum \( y \)-degree in \( -y \cdot \gamma_q \) cannot cancel out of this sum, which implies that the \( q \)-th component of \( \gamma \) does not lie in \( A[x] \); this yields a contradiction, since \( q < \Lambda_\mu \), and \( \gamma \) is by hypothesis side-minimized. From this we deduce that a nonzero coefficient \( \gamma_k \) of maximal \( y \)-degree cannot multiply an \( h_k \) of Type 1 or 2; whence, \( h_q \)
must be of Type 3 (see Algorithm 4.2.1), with $q$-th component equal to -1. Therefore, the $q$-th component of $\gamma$ has the form

$$
\gamma_q \cdot (-1) + \gamma_{q-1} \cdot (x + a_{q-1,q}) + \sum_{k \neq q, q-1} \gamma_k \cdot a_{k,q}, \ a_{q-1,q}, a_{k,q} \in A. \tag{29}
$$

Since $g_q$ is side polynomial that is not also a top polynomial, we have from Remark 4.2.2 that the only nonzero $a_{k,q}$’s in (29) are those for which $h_k$ is of Type 1 or 2, which we recently saw implies that $\gamma_k$ has $y$-degree less than the (maximal) $y$-degree of $\gamma_q$. Therefore the terms of maximal $y$-degree in (29) can only cancel — as they must, since the $q$-th component of the side-minimized linear combination $\gamma$ is 0 — provided $\gamma_{q-1}$ has $y$-degree equal to the $y$-degree of $\gamma_q$, which again forces $h_{q-1}$ to be of Type 3. Iterating this argument, we see that $h_{q-2}, h_{q-3}, \ldots, h_1$ are all of Type 3. Since $h_1$ is always of Type 1 or 2, we eventually achieve a contradiction. It follows that the only side-minimized $B$-linear combination of the $h_k$ is the trivial one. The $B$-linear independence of the $h_k$ is an immediate consequence. □

**Remark 4.3.2.** The basic idea of the preceding proof comes from [9, proof of Proposition 4.1, page 51].

**Lemma 4.3.3.** Let $f = (f_1, \ldots, f_{\Lambda_\mu})$ be an arbitrary element of $B^\Lambda_\mu$. Then we can express $f$ uniquely in the form

$$
f = \left( \sum_{k=1}^{\Lambda_\mu - 1} b_k \cdot h_k \right) + f', \ b_k \in B,
$$

where $f' \in B^\Lambda_\mu$ is side-minimized. In brief, every $H$-coset has a unique side-minimized representative.

**Proof.** We organize the operations needed to express $f$ in the desired form into a repeated alternation of two sub-procedures that we call **column clearing** and **reduction of $y$-degree**. In this context, a **column** refers to a maximal subset of the leading monomials having the same $y$-degree; for example, in Figure 4, we have six columns, five of which consist of one leading monomial, and one of which consists of two leading monomials (namely, $x^2y^2$ and $x^3y^2$).

**Column clearing.** Let $x^{r_j}y^{s_j}$ be a side leading monomial that is not of lowest $x$-degree in its column (for example, $x^3y^2$ in Figure 4). Then the syzygy $h_j$ is of Type 3 (see Algorithm 4.2.1), with -1 in the $j$-th component, $x + a_{j,j+1}$ in the $j + 1$st component, and otherwise nonzero $a_{j,k}$’s only in components associated to top polynomials $g_k$. Therefore we can write

$$
f = -f_j \cdot h_j + (f + f_j \cdot h_j)
= -f_j \cdot h_j + \tilde{f},
$$
in which the first summand lies in $H$ and the second is a $\Lambda_\mu$-tuple that is 0 in the $j$-th component and otherwise differs from $f$ only in components associated to top monomials and the $(j + 1)$-st component. If $x^{r_j - 1}y^{s_j}$ (the leading monomial immediately below $x^ry^s$) is not of lowest $x$-degree in the column, then we may apply this procedure to $\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_{\Lambda_\mu})$ to obtain

$$f = (-f_j \cdot h_j - \tilde{f}_{j+1} \cdot h_{j+1}) + (\tilde{f} - \tilde{f}_{j+1} \cdot h_{j+1})$$

in which $\tilde{f}$ is 0 in the $(j+1)$-st component, and otherwise differs from $\tilde{f}$ only in components associated to top monomials as well as the $(j+2)$-nd component (in particular, $\tilde{f}$ is still 0 in the $j$-th component). It is now clear that by working column-by-column and starting at the top of each column, we can write $f$ in the form

$$f = \left(\sum_{k=1}^{\Lambda_\mu-1} b_k \cdot h_k\right) + f_{cc}, \ b_k \in B,$$

where $f_{cc}$ is zero in all components except for those associated to top monomials and the last component (associated to the leading monomial $y^{p_1}$). At this point we have completed the column clearing operation on $f$.

**Remark 4.3.4.** Following a column clearing operation on $f$, the maximum $y$-degree among the components of $f_{cc}$ is no greater than the maximum $y$-degree among the components of $f$. Indeed, the column clearing operation can only modify the components of $f$ in $y$-degrees $\leq$ the maximum $y$-degree among the $f_j$ corresponding to the positions that are set to zero by the process.

**Reduction of $y$-degree.** Let $f = (f_1, \ldots, f_{\Lambda_\mu}) \in B^{\Lambda_\mu}$ and let $s$ be the maximal $y$-degree (achieved at $f_j$) among the components of $f$ associated to top monomials. We write

$$f_j = q_s(x)y^s + \text{(lower degree terms in } y)$$

(we here view the elements of $B = A[x, y]$ as polynomials in $y$ with coefficients in $A[x]$). If $s > 0$, we may use the syzygy $h_j$, which is of Type 1 or 2 with $-y + a_{j,j}$ in the $j$-th component, to rewrite $f$ as follows:

$$f = -q_s(x)y^{s-1} \cdot h_j + (f + q_s(x)y^{s-1} \cdot h_j)$$

$$= -q_s(x)y^{s-1} \cdot h_j + \tilde{f},$$

in which the $y$-degree of the $j$-th component of $\tilde{f}$ is $\leq s - 1$, and the other components of $\tilde{f}$ differ from the corresponding components of $f$ only in $y$-degree $s - 1$. If the maximal $y$-degree among the components of $\tilde{f}$ associated
to top monomials is still \( s \), the process can be iterated to yield

\[
(31) \quad f = \left( \sum_{k=1}^{\Lambda_{\mu}} b_k \cdot h_k \right) + f_{y\text{-red}}, \ b_k \in B,
\]

where \( f_{y\text{-red}} \) has \( y \)-degree \( \leq s - 1 \) in all components associated to top monomials, and otherwise differs from \( f \) only in \( y \)-degree \( s - 1 \). This completes the process of reduction of \( y \)-degree.

To complete the proof of the lemma, we begin with an arbitrary \( f = (f_1, \ldots, f_{\Lambda_{\mu}}) \), and apply column clearing to write \( f \) in the form (30). Let \( s \) be the maximum \( y \)-degree among the components of \( f_{cc} \) corresponding to top monomials. By applying reduction of \( y \)-degree to \( f_{cc} \), we write \( f \) in the form (31), where \( f_{y\text{-red}} \) has \( y \)-degree \( \leq s - 1 \) in each top position and in each position that is zeroed out by column clearing (because the latter positions are 0 in \( f_{cc} \), and reduction of \( y \)-degree can change them only in \( y \)-degree \( s - 1 \)).

In light of Remark 4.3.4, we see that column clearing applied to \( f_{y\text{-red}} \) will express \( f \) again in the form (30), where now the maximum \( y \)-degree among the components of \( f_{cc} \) corresponding to top monomials is \( \leq s - 1 \). By iterating the alternation of reduction of \( y \)-degree and column clearing, we eventually obtain an expression for \( f \) as specified in the statement of the lemma; that is, we produce a side-minimized representative \( f' \) for the coset \( f + H \). Given a second such side-minimized representative \( f'' \), we have that \( f' - f'' \) is a side-minimized element of \( H \); Lemma 4.3.1 now yields \( f' - f'' = 0 \), which proves the uniqueness of the side-minimized representative and completes the proof of the lemma. \( \square \)

**Lemma 4.3.5.** If \( f = (f_1, \ldots, f_{\Lambda_{\mu}}) \) is a \( \mu \)-pseudosyzygy of \( (g_k) \) that is side-minimized, then \( f = 0 \).

**Proof.** We first suppose that \( f_{\Lambda_{\mu}} \neq 0 \). Viewing \( f_{\Lambda_{\mu}} \) as a polynomial in \( y \) with coefficients in \( A[x] \), we let \( q(x) y^a \) be the term of maximal \( y \)-degree. Recall from (16) that we write \( p_1 \) for the largest part in the partition \( \mu \); whence, the term of maximal \( y \)-degree in \( f_{\Lambda_{\mu}} \cdot g_{\Lambda_{\mu}} \) is \( q(x) y^{p_1 + s} \) (as an example, see Figure 4). Since the other nonzero products \( f_k \cdot g_k, 1 \leq k < \Lambda_{\mu} \), have maximal \( y \)-degree \( < p_1 \) (\( f_k \) is either 0 or in \( A[x] \), by hypothesis, and the maximal \( y \)-degree of every top \( g_k \) is \( < p_1 \)), we have that the term \( q(x) y^{p_1 + s} \) cannot cancel out of the expression \( \sum_{k=1}^{\Lambda_{\mu}} (g_k \cdot f_k) \) in Equation (27), a contradiction. It follows that \( f_{\Lambda_{\mu}} = 0 \).

Assume now that not all of the remaining components of \( f \) are 0. Among the nonzero \( f_k, 1 \leq k < \Lambda_{\mu} \), let \( f_j \) have maximal \( x \)-degree \( d \), and let \( a x^d \) be the corresponding term (\( a \in A \)). Recall that the (top) leading monomial of \( g_j \) is \( x^{r_j} y^{s_j} \). The product \( f_j \cdot g_j \) contributes the term \( a x^{d+r_j} y^{s_j} \) to the expression \( \sum_{k=1}^{\Lambda_{\mu}} (g_k \cdot f_k) \) in (27); if this term is to cancel out, as it
must, then for at least one \( j' \neq j \) with \( g_{j'} \) a top polynomial, the product \( f_{j'} \cdot g_{j'} \) must contain a term of the form \( a'x_{r_j}y_{s_j} \). But \( f_{j'} \) lies in \( A[x] \), by hypothesis; moreover, the (top) leading monomial of \( g_{j'} \) must have \( y \)-degree \( \neq s_j \), since distinct top monomials must differ in \( y \)-degree. Therefore, among the terms of \( g_{j'} \) having \( y \)-degree \( s_j \), the largest possible \( x \)-degree is \( r_j - 1 \). This means that the \( x \)-degree of \( f_{j'} \) must be at least \( d + 1 \), which contradicts the maximality of \( d \). We conclude that \( f_k = 0 \) for all \( k < \Lambda_{\mu} \), and the lemma is proved. \( \square \)

We are now ready to state and prove our main result on \( \mu \)-pseudosyzygies.

Let \( \mathfrak{G} \subseteq B = A[x, y] \) denote the ideal generated by the polynomials \( g_k \).

Observe that the proof of Lemma 3.2.3 carries over to show that \( B/\mathfrak{G} \) is generated as an \( A \)-module by the partition monomials.

**Theorem 4.3.6.** If \( B/\mathfrak{G} \) is free over \( A \) with the partition monomials constituting a basis, then any \( \mu \)-pseudosyzygy of \( (g_k) \) is in fact a syzygy of \( (g_k) \); in this case, \( \{h_1, \ldots, h_{\Lambda_{\mu} - 1}\} \) is a \( B \)-basis of the first syzygy module of \( (g_k) \). Conversely, if the \( \mu \)-pseudosyzygies \( h_j \) are all syzygies of \( (g_k) \), then the \( A \)-module \( B/\mathfrak{G} \) is free with basis given by the partition monomials.

**Proof.** First suppose that \( B/\mathfrak{G} \) is \( A \)-free with basis consisting of the partition monomials, and let \( f = (f_1, \ldots, f_{\Lambda_{\mu}}) \) be a \( \mu \)-pseudosyzygy of the \( (g_k) \). To prove that \( f \) is a syzygy of \( (g_k) \), we repeat the argument of Lemma 4.1.1:

We have that

\[
L = \sum_{k=1}^{\Lambda_{\mu}} (f_k \cdot g_k) \equiv 0 \pmod{\mathfrak{G}}
\]

and

\[
L = (\text{a linear combination of partition monomials}).
\]

But by the freeness hypothesis, the only \( A \)-linear combination of partition monomials that is \( \equiv 0 \pmod{\mathfrak{G}} \) is the trivial one; that is, \( L = 0 \), which implies that \( f \) is a syzygy of \( (g_k) \).

In particular, the \( \mu \)-pseudosyzygies \( h_1, \ldots, h_{\Lambda_{\mu} - 1} \) given by Algorithm 4.2.1 are all syzygies of \( (g_k) \); we must show that the \( h_j \) form a \( B \)-basis of the syzygy module. Since the \( h_j \) form a \( B \)-linearly independent set, by Lemma 4.3.1, it remains to show that any given syzygy \( f = (f_1, \ldots, f_{\Lambda_{\mu}}) \) lies in \( H \), the \( B \)-linear span of the \( h_j \). However, by Lemma 4.3.3, we know that the coset \( f + H \) has a unique side-minimized representative \( f' \), and it is clear that \( f' \) is also a syzygy (and therefore a \( \mu \)-pseudosyzygy) of \( (g_k) \). Lemma 4.3.5 now yields \( f' = 0 \); whence, \( f \in H \), as desired.

We turn now to the last assertion of the theorem. Suppose that the \( h_j \) are all syzygies (not just \( \mu \)-pseudosyzygies) of \( (g_k) \). As noted prior to the statement of the theorem, the \( A \)-module \( B/\mathfrak{G} \) is generated by the partition
monomials; it remains to prove that the latter are $A$-linearly independent (mod $G$).

To this end, we assume given an $A$-linear combination of partition monomials $L$ such that $L \equiv 0 \pmod{G}$. Since $L \in G$, there exists a $\mu$-pseudosyzygy $f = (f_1, \ldots, f_{\Lambda_{\mu}})$ of $(g_k)$ such that

$$L = f \cdot (g_1, \ldots, g_{\Lambda_{\mu}}) = \sum_{k=1}^{\Lambda_{\mu}} (f_k \cdot g_k).$$

Let $f'$ be the unique side-minimized representative of $f + H$ given by Lemma 4.3.3. Since the $h_j$ are assumed to be syzygies of $(g_k)$, we see that

$$L = f \cdot (g_k) = f' \cdot (g_k);$$

that is, $f'$ is a side-minimized $\mu$-pseudosyzygy of $(g_k)$, and is perforce 0 by Lemma 4.3.5. We conclude that $L = 0$, which implies that the partition monomials are $A$-linearly independent (mod $G$). This completes the proof of the theorem. □

4.4. Corollaries in the main case. We now return to the case of main interest, in which $A = \mathcal{O}_{U_{\mu}}, B = A[x, y]$, and the polynomials $g_j \in B$ are those given by (23). In light of Proposition 3.2.1 and Lemma 3.2.2, the preceding theorem specializes to yield:

**Corollary 4.4.1.** The $\mu$-pseudosyzygies $h_1, \ldots, h_{\Lambda_{\mu}-1}$ of $(g_j)$ given by Algorithm 4.2.1 are in fact syzygies that constitute an $\mathcal{O}_{U_{\mu}}[x, y]$-basis of the first syzygy module of $(g_j)$.

Before stating the remaining corollaries, we introduce some notation. We denote the $(\Lambda_{\mu}-1) \times \Lambda_{\mu}$ matrix with rows given by the syzygies $h_j$ as follows:

$$m_{\mu} = (h_1, h_2, \ldots, h_{\Lambda_{\mu}-1}).$$

We also define the maps

$$\alpha_{\mu}: B^{\Lambda_{\mu}-1} \to B^{\Lambda_{\mu}}, \quad (b_1, \ldots, b_{\Lambda_{\mu}-1}) \mapsto (b_1, \ldots, b_{\Lambda_{\mu}-1}) \cdot m_{\mu},$$

$$\beta_{\mu}: B^{\Lambda_{\mu}} \to B, \quad (b_1, \ldots, b_{\Lambda_{\mu}}) \mapsto \sum_{k=1}^{\Lambda_{\mu}} b_k \cdot g_k.$$

**Corollary 4.4.2.** The composition

$$0 \to B^{\Lambda_{\mu}-1} \xrightarrow{\alpha_{\mu}} B^{\Lambda_{\mu}} \xrightarrow{\beta_{\mu}} B \to B/\mathcal{I}_{\mu} \to 0$$

is a free resolution of the $B$-module $B/\mathcal{I}_{\mu}$ of length 2.

**Proof.** The exactness of the sequence at $B$ follows from Proposition 3.2.1, which states that $\mathcal{I}_{\mu}$ is generated by the $g_j$, $1 \leq j \leq \Lambda_{\mu}$. Since

$$\alpha_{\mu}((b_1, \ldots, b_{\Lambda_{\mu}-1})) = \sum_{k=1}^{\Lambda_{\mu}-1} (b_k \cdot h_k),$$

...
the exactness at $B^{\Lambda\mu}$ restates that the syzygies $h_k$ span the full syzygy module of $(g_j)$, which is the kernel of $\beta_\mu$. Finally, the injectivity of $\alpha_\mu$ is equivalent to the linear independence of the $h_k$. □

**Corollary 4.4.3.** For any $A$-module $N$, one obtains an exact sequence by applying the functor $(\cdot) \otimes_A N$ to the free resolution of $B/\mathcal{I}_\mu$ given in Corollary 4.4.2.

**Proof.** Since $B = A[x, y]$, we have that the free resolution of the preceding corollary is a free resolution of $B/\mathcal{I}_\mu$ as an $A$-module. Therefore, the modules $\text{Tor}_q^A(N, B/\mathcal{I}_\mu)$ are the homology modules of the tensored sequence. But since $B/\mathcal{I}_\mu$ is flat (in fact, free, by Lemma 3.2.2) over $A$, we have that

$$\text{Tor}_q^A(N, B/\mathcal{I}_\mu) \approx \text{Tor}_q^A(B/\mathcal{I}_\mu, N) = 0;$$

whence, the corollary. □

We end this section with:

**Theorem 4.4.4.** The generators $g_j$ of the ideal $\mathcal{I}_\mu$ may be recovered (up to sign) as the maximal minors of the matrix $m_\mu$ as follows:

$$g_j = \det(e_j, h_1, \ldots, h_{\Lambda\mu-1}), \quad 1 \leq j \leq \Lambda_\mu,$$

where $e_j$ is the standard unit $\Lambda_\mu$-tuple with 1 in the $j$-th component.

**Proof.** Recall that $A$ is an integral domain. Since $B/\mathcal{I}_\mu$ is finite and flat of degree $n$ over $A$, we have that $\mathcal{I}_\mu$ is unmixed of codimension 2; we also have the free resolution of Corollary 4.4.2. In this situation, Schaps’s proof of [11, Theorem 1, pp. 671-673] applies, and shows that if we put

$$g'_j = \det(e_j, h_1, \ldots, h_{\Lambda\mu-1}), \quad 1 \leq j \leq \Lambda_\mu, \quad (32)$$

then there is a nonzero scalar $c \in k$ such that $g'_j = c \cdot g_j$ for all $j$. This result is a version of a theorem of Burch [2] (see also [3, Theorem 20.15, p. 502]), for which Schaps references [10, Ex. 8, p. 148]. It remains to show that $c = 1$.

To this end, let $\lambda: A \rightarrow k$ be the $k$-algebra map corresponding to the inclusion of the point $I_\mu \in U_\mu$, and $\tilde{\lambda}: A[x, y] \rightarrow k[x, y]$ the map induced by $\lambda$. Recall from Remark 3.1.4 that $I_\mu$ is the point of $U_\mu$ at which all of the functions $e_{h_k}^{r_j s_j} \in e_\mu$ vanish; whence,

$$\tilde{\lambda}(g_j) = \text{the leading monomial } x^{r_j} y^{s_j}, \quad 1 \leq j \leq \Lambda_\mu,$$

and therefore, since $g'_j = c \cdot g_j$,

$$\tilde{\lambda}(g'_j) = \tilde{\lambda}(c \cdot g_j) = c \cdot x^{r_j} y^{s_j}, \quad 1 \leq j \leq \Lambda_\mu. \quad (33)$$

On the other hand, applying $\tilde{\lambda}$ to both sides of (32), we obtain

$$\tilde{\lambda}(g'_j) = \det(e_j, \tilde{\lambda}(h_1), \ldots, \tilde{\lambda}(h_{\Lambda\mu-1})), \quad 1 \leq j \leq \Lambda_\mu,$$
where we apply $\tilde{\lambda}$ componentwise to each row of the matrix on the right-hand side. Referring (for example) to (26), and reading all the $e_{hk}^{r_j,s_j}$ therein as 0, we see that

\begin{equation}
\tilde{\lambda}(g'_1) = \det(e_1, \tilde{\lambda}(h_1), \ldots, \tilde{\lambda}(h_{\Lambda_\mu-1}))
\end{equation}

is the determinant of a lower-triangular matrix whose main diagonal comprises only 1’s and $x$’s. Moreover, there is one $x$ for each side monomial associated to $\mu$, and since the number of side monomials is equal to the number of parts $\ell$ of $\mu$, we may expand the determinant in (34) to obtain

$\tilde{\lambda}(g'_1) = x^\ell = x^{r_1}y^{s_1}$

(recall Remark 3.1.1). Confronting this with (33), we find that $c = 1$, as desired. □

5. An explicit representation of $O_{U_\mu}$ as a quotient ring.

According to Proposition 3.1.2, the set $c_\mu$ (19) generates the affine coordinate ring $O_{U_\mu}$ as a $k$-algebra. To represent this ring as a quotient, we introduce the set of indeterminates

\begin{equation}
C_\mu = \{ C_{r_j,s_j}^{h,k} | c_{r_j,s_j}^{h,k} \in c_\mu \} = \{ C_{r_j,s_j}^{h,k} | 1 \leq j \leq \Lambda_\mu, (h,k) \in \mu \},
\end{equation}

and define the surjection

\begin{equation}
u_\mu^* : k[C_\mu] \to O_{U_\mu}, \quad C_{r_j,s_j}^{h,k} \mapsto c_{r_j,s_j}^{h,k},
\end{equation}

which is the comorphism of a closed immersion $u_\mu : U_\mu \to \text{Spec}(k[C_\mu])$.

Our main goal in this section is to describe the kernel $R_\mu$ of $u_\mu^*$. We show that explicit generators of $R_\mu$ can be obtained from the $\mu$-pseudosyzygies studied in Sections 4.2 and 4.3. As an example, we compute these generators in case $\mu = (2,1)$, for which the associated monomial ideal is the “fat point” ideal

$I_\mu = (x^2, xy, y^2) = (x, y)^2 \subseteq k[x, y]$.

5.1. Generators of the kernel of $u_\mu^*$. We begin by defining the polynomials

\[ G_j = x^{r_j}y^{s_j} - \sum_{(h,k) \in \mu} C_{r_j,s_j}^{h,k} x^h y^k \in k[C_\mu][x, y], \quad 1 \leq j \leq \Lambda_\mu; \]

it is clear that each $G_j$ is a preimage of $g_j$ (23) under the map

\begin{equation}\overline{u}_\mu^* : k[C_\mu][x, y] \to O_{U_\mu}[x, y]\end{equation}

induced by $u_\mu^*$ (36). We next execute Algorithm 4.2.1, with $A = k[C_\mu]$ and $B = A[x, y]$, to generate the $\mu$-pseudosyzygies $H_1, \ldots, H_{\Lambda_\mu-1}$ of $(G_k)$. In
particular, for $1 \leq j \leq \Lambda_{\mu} - 1$, we have that

$$\left( G_1, \ldots, G_{\Lambda_{\mu}} \right) \cdot H_j = \sum_{(h,k) \in \mu} \rho_{hk}^j x^h y^k$$

is an $A$-linear combination of partition monomials. Applying $\overline{u}_{\mu}^*$ componentwise to the $(\Lambda_{\mu})$-tuples in (38), we see that $\overline{u}_{\mu}^*(H_j) = h_j$ is the $j$-th $\mu$-pseudosyzygy of $(g_k)$ generated by Algorithm 4.2.1, which is in fact a syzygy of $(g_k)$ by Corollary 4.4.1. It follows that

$$\overline{u}_{\mu}^* \left( \sum_{(h,k) \in \mu} \rho_{hk}^j x^h y^k \right) = 0 \in \mathcal{O}_{U_{\mu}}[x,y];$$

whence,

$$u_{\mu}^*(\rho_{hk}^j) = 0 \in \mathcal{O}_{U_{\mu}}.$$ 

In other words, the ideal $R' = \left\{ \rho_{hk}^j \mid 1 \leq j \leq \Lambda_{\mu} - 1, \ (h,k) \in \mu \right\} \subseteq R_{\mu}.$

Theorem 5.1.1. The functions $\rho_{hk}^j \in k[\mathcal{C}_\mu]$ generate the kernel of the surjection $u_{\mu}^*: k[\mathcal{C}_\mu] \to \mathcal{O}_{U_{\mu}}$; that is, $R' = R_{\mu}.$

Proof. We write

$$V = \text{Spec}(k[\mathcal{C}_\mu]/R'), \quad \mathcal{O}_V = k[\mathcal{C}_\mu]/R',$$

and observe that, by (39), the map (36) induces a surjection

$$v^*: \mathcal{O}_V \to \mathcal{O}_{U_{\mu}}$$

that is the comorphism of a closed immersion

$$v: U_{\mu} \to V.$$ 

Since the desired conclusion is equivalent to these maps being isomorphisms, we seek an inverse map

$$\omega: V \to U_{\mu}.$$ 

By the universal property of $U_{\mu}$ given in Proposition 2.3.2, $\omega$ corresponds to a closed subscheme $Z \subseteq V \times \mathbb{A}_k^2$ that is finite and flat of degree $n$ over $V$, and has fibers over $k$-points $v \in V$ that are cut out by ideals $I_v \in U_{\mu}$; we therefore seek such a family of subschemes.

Let

$$u': k[\mathcal{C}_\mu] \to \mathcal{O}_V, \quad \overline{u}': k[\mathcal{C}_\mu][x,y] \to \mathcal{O}_V[x,y]$$

denote, respectively, the quotient map and the map induced thereby on polynomials in $x$ and $y$, and set

$$g_j' = \overline{u}'(G_j), \quad 1 \leq j \leq \Lambda_{\mu},$$

$$G' = (g_1', \ldots, g_{\Lambda_{\mu}}') \subseteq \mathcal{O}_V[x,y].$$
Applying \( \tilde{u}' \) to both sides of Equation (38) (componentwise to the \((\Lambda_\mu)\)-tuples), and recalling that all the \( \rho_{h_k}^j \) vanish under \( u' \), we see that each \( h'_j = \tilde{u}'(H_j) \), the \( j \)-th \( \mu \)-pseudosyzygy of \((g'_k)\) produced by Algorithm 4.2.1, is in fact a syzygy of \((g'_k)\). It now follows from the last statement of Theorem 4.3.6 that \( \mathcal{O}_V[x,y]/\mathcal{G}' \) is free of dimension \( n \) as an \( \mathcal{O}_V \)-module, with the partition monomials constituting a basis. Therefore, the map

\[
\text{Spec}(\mathcal{O}_V[x,y]/\mathcal{G}') \xrightarrow{F} \text{Spec}(\mathcal{O}_V) = V,
\]

induced by the natural map \( \mathcal{O}_V \to \mathcal{O}_V[x,y]/\mathcal{G}' \), is finite and flat of degree \( n \), and has fibers over \( k \)-points cut out by ideals \( I \in U_\mu \). We define \( \omega : V \to U_\mu \) to be the map corresponding to the family of subschemes \( F \).

We claim that the composition of comorphisms

\[
\mathcal{O}_V \xrightarrow{\upsilon^*} \mathcal{O}_{U_\mu} \xrightarrow{\omega^*} \mathcal{O}_V
\]

is the identity map. This being granted, the theorem follows at once, since \( \upsilon^* \), already known to be surjective, is now seen to be injective as well, and hence an isomorphism, as desired. It remains to prove the claim.

By definition, we have that the map \( \upsilon^* \) behaves as follows on the images of the \( C_{h_k}^{r_j,s_j} \):

\[
\upsilon^*(C_{h_k}^{r_j,s_j}) \mapsto \upsilon^*_\mu(C_{h_k}^{r_j,s_j}) = c_{h_k}^{r_j,s_j}.
\]

Furthermore, from the universal property of \( U_\mu \) (Proposition 2.3.2), the ideal \( \mathcal{G}' \subseteq \mathcal{O}_V[x,y] \) that cuts out the family of subschemes inducing the map \( \omega \) is the extension of the ideal \((g_1, \ldots, g_{\Lambda_\mu}) = I_\mu \subseteq \mathcal{O}_{U_\mu}[x,y] \) that cuts out the universal closed subscheme over \( U_\mu \) (Proposition 3.2.1), under the map

\[
\tilde{\omega}^* : \mathcal{O}_{U_\mu}[x,y] \to \mathcal{O}_V[x,y]
\]

induced by \( \omega^* \). In other words, the polynomials

\[
\tilde{\omega}^*(g_j) = x^{r_j}y^{s_j} - \sum_{(h,k) \in \mu} \omega^*(c_{h_k}^{r_j,s_j})x^hy^k \in \mathcal{O}_V[x,y], \quad 1 \leq j \leq \Lambda_\mu,
\]

generate the ideal \( \mathcal{G}' \). By construction, however, \( \mathcal{G}' \) is generated by the polynomials

\[
g_j = \tilde{u}'(G_j) = x^{r_j}y^{s_j} - \sum_{(h,k) \in \mu} u'(C_{h_k}^{r_j,s_j})x^hy^k \in \mathcal{O}_V[x,y], \quad 1 \leq j \leq \Lambda_\mu;
\]

moreover, since the partition monomials give an \( \mathcal{O}_V \)-basis of the quotient \( \mathcal{O}_V[x,y]/\mathcal{G}' \), we have that \( \mathcal{G}' \) contains a unique polynomial of the form

\[
(\text{leading term } x^{r_j}y^{s_j}) + \text{ an } \mathcal{O}_V \text{-linear combination of partition monomials}
\]

for each \( j \). From this it follows at once that

\[
c_{h_k}^{r_j,s_j} \xrightarrow{\omega^*} u'(C_{h_k}^{r_j,s_j}),
\]
which, in light of (41), shows that the composition (40) maps each element \(u'(C_{h,k}^{r_j,s_j})\) to itself; since these elements generate the source as a \(k\)-algebra, we conclude that the composition is the identity, as claimed. This completes the proof of the theorem. \(\square\)

5.2. An example: \(\mu = (2,1)\). To illustrate Theorem 5.1.1, we compute the \(H_j\) and the \(\rho_{hk}^{j}\) in case \(\mu = (2,1)\); the monomial ideal associated to \(\mu\) is \((x^2, xy, y^2) = (x, y)^2\) (see Figure 5). Our set \(\mathfrak{C}_{\mu=(2,1)}\) of indeterminates (35)

\[
\begin{array}{c}
\text{x}^2 \\
x \\
\text{xy} \\
1 \\
y \\
y^2
\end{array}
\]

Figure 5. Diagram of \(\mu = (2,1)\) with the partition monomials boxed in and the leading monomials shown in boldface.

consists of the nine coefficients \(C_{h,k}^{r_j,s_j}\) in the polynomials

\[
\begin{align*}
G_0 &= x^2 - C_{0,0}^{2,0} - C_{1,0}^{2,0}x - C_{0,1}^{2,0}y, \\
G_1 &= xy - C_{0,0}^{1,1} - C_{1,0}^{1,1}x - C_{0,1}^{1,1}y, \\
G_2 &= y^2 - C_{0,0}^{0,2} - C_{1,0}^{0,2}x - C_{0,1}^{0,2}y.
\end{align*}
\]

It is easy to check that Algorithm 4.2.1 yields the following two pseudodosyzygies:

\[
\begin{align*}
H_1 &= (-y + C_{1,0}^{1,1}, x - C_{1,0}^{2,0} + C_{0,1}^{1,1}, -C_{0,1}^{2,0}), \\
H_2 &= (C_{1,0}^{0,2}, -y - C_{1,0}^{1,1} + C_{0,1}^{0,2}, x - C_{1,0}^{1,1}).
\end{align*}
\]

Computing the dot product (38) for \(j = 1\) and \(2\), and reading off the coefficients \(\rho_{hk}^{j}\) on the right-hand side, we obtain our set of generators for \(\mathfrak{R}_{\mu=(2,1)}\):

\[
\begin{align*}
\rho_{0,0}^{1} &= -C_{0,0}^{1,1}C_{0,0}^{2,0} + C_{0,0}^{0,2}C_{0,1}^{2,0} + C_{0,0}^{1,1}(-C_{0,1}^{1,1} + C_{0,1}^{2,0}), \\
\rho_{1,0}^{1} &= -C_{0,0}^{1,1} - C_{0,1}^{1,1}C_{1,0}^{1,1} + C_{1,0}^{0,2}C_{0,1}^{2,0}, \\
\rho_{0,1}^{1} &= -(C_{0,1}^{1,1})^2 + C_{0,0}^{2,0} + (C_{0,1}^{0,2} - C_{1,0}^{1,1})C_{0,1}^{2,0} + C_{0,1}^{1,1}C_{1,0}^{2,0}, \\
\rho_{0,0}^{2} &= -C_{0,0}^{0,2}C_{0,0}^{1,1} + C_{0,0}^{0,2}C_{0,1}^{1,1} + C_{0,0}^{1,1}C_{1,0}^{1,0} - C_{1,0}^{0,2}C_{0,0}^{2,0}, \\
\rho_{1,0}^{2} &= -(C_{0,0}^{2,0} + C_{0,1}^{1,1}C_{1,0}^{1,1} - C_{0,1}^{0,2}C_{1,0}^{1,1} + (C_{1,0}^{1,1})^2 - C_{1,0}^{0,2}C_{1,0}^{2,0}, \\
\rho_{0,1}^{2} &= C_{0,0}^{1,1} + C_{0,1}^{1,1}C_{1,0}^{1,1} - C_{1,0}^{0,2}C_{0,1}^{2,0}.
\end{align*}
\]

There is considerable redundancy among these generators; indeed, we have that

\[
\mathfrak{R}_{(2,1)} = (\rho_{1,0}^{1}, \rho_{0,1}^{1}, \rho_{1,0}^{2}).
\]
which follows from the relations
\begin{align}
\rho_{0,0}^1 &= (C_{0,1}^{1,1} - C_{1,0}^{2,0})\rho_{1,0}^1 - C_{1,0}^{1,1}\rho_{0,1}^1 - C_{0,1}^{2,0}\rho_{1,0}^2, \\
\rho_{0,0}^2 &= (C_{0,1}^{0,2} - C_{1,1}^{1,1})\rho_{1,0}^1 - C_{1,0}^{0,2}\rho_{0,1}^1 - C_{0,1}^{1,1}\rho_{1,0}^2, \\
\rho_{0,1}^2 &= -\rho_{1,0}^1.
\end{align}
(We derive these relations in the next section.) Note that the generators in (44) express $C_{0,0}^{2,0}$, $C_{0,0}^{1,1}$, and $C_{0,0}^{0,2}$ as polynomials in the remaining $C$’s (mod $\mathcal{R}_{(2,1)}$); it follows from this that
\[ \mathfrak{O}_{U_{(2,1)}} \approx k[\mathcal{C}_{(2,1)}]/\mathcal{R}_{(2,1)} \approx k[C_{1,0}^{2,0}, C_{0,1}^{2,0}, C_{1,1}^{1,1}, C_{0,1}^{1,1}, C_{1,0}^{0,2}, C_{0,1}^{0,2}]. \]
In other words,
\begin{equation}
U_{(2,1)} \text{ is a six-dimensional affine cell in } H^3; \text{ that is, } U_{(2,1)} \text{ is an open neighborhood in } H^3 \text{ that is isomorphic to } \mathbb{A}^6_k.
\end{equation}

We reconfirm this fact in Section 7 (second of Remarks 7.3.3) as one example of a more general sufficient condition for $U_\mu$ to be an affine cell (Corollary 7.3.2). In particular, this condition holds for all $\mu$ whose associated monomial ideal $I_\mu$ is a “fat point” ideal $(x,y)^r$, $r = 1, 2, 3, \ldots$. The proof relies on a second set of generators of the ideal $\mathcal{R}_\mu$ that we present in the next section.

6. A second set of generators of the ideal $\mathcal{R}_\mu$.

In this section we present a second set of generators of the ideal $\mathcal{R}_\mu \subseteq k[\mathcal{C}_\mu]$ that is more convenient for certain purposes than the set of generators given by Theorem 5.1.1. The key ingredient is provided by Theorem 4.4.4, which states that we can recover the polynomials $g_k \in \mathfrak{O}_{U_\mu}[x,y]$ (up to sign) as the maximal minors of the matrix $\mathbf{m}_\mu$ whose rows are the basic syzygies $h_j$ of $(g_k)$. As an example, we continue our study of the case $\mu = (2,1)$ that we began in Section 5.2.

6.1. The second set of generators. Retaining all previous notation, we begin by defining the $((\Lambda_\mu - 1) \times \Lambda_\mu)$-matrix
\[ \mathbf{M}_\mu = (H_1, H_2, \ldots, H_{\Lambda_\mu - 1}), \]
with rows $H_j$ the $\mu$-pseudosyzygies of $(G_k)$ introduced in Section 5.1. By analogy with the determinantal expression for $g_j$ given by Theorem 4.4.4, we define the polynomials
\begin{equation}
\mathcal{D}_j = \det(e_j, H_1, \ldots, H_{\Lambda_\mu - 1}) \subseteq k[\mathcal{C}_\mu][x,y], \quad 1 \leq j \leq \Lambda_\mu,
\end{equation}
where $e_j$ is the $j$-th standard unit vector. We noted in the discussion preceding Theorem 5.1.1 that the map $\mathcal{U}_\mu^\ast$ (37) maps each $G_k$ to $g_k$ and (acting
componentwise) each $H_k$ to $h_k$ (and therefore $M_\mu$ to $m_\mu$). It follows from Theorem 4.4.4 that

$$\tilde{u}_\mu^*(D_j) = g_j, \quad 1 \leq j \leq \Lambda_\mu;$$

since $g_j$ has the form (23), we are led to write each $D_j$ in the form

$$D_j = x^{r_j}y^{s_j} - \sum_{(h,k) \in \mu} D_{rk}^{r_j,s_j} x^h y^k + N_j,$$

where $D_{rk}^{r_j,s_j} \in k[\mathcal{C}_\mu]$ and $N_j \in \ker(\tilde{u}_\mu^*)$ involves only non-partition monomials (in $x$ and $y$), possibly including the leading monomials $x^{r_k}y^{s_k}$. Since $G_j$ and $D_j$ both map to $g_j$ under $\tilde{u}_\mu^*$, we have that $C_{hk}^{r_j,s_j}$ and $D_{hk}^{r_j,s_j}$ both map to $c_{hk}^{r_j,s_j}$ under $u_\mu^*$; whence,

$$u_\mu^*(C_{hk}^{r_j,s_j} - D_{hk}^{r_j,s_j}) = 0, \quad 1 \leq j \leq \Lambda_\mu, \ (h,k) \in \mu.$$

We set

$$\delta_{hk}^{r_j,s_j} = C_{hk}^{r_j,s_j} - D_{hk}^{r_j,s_j} = C_{hk}^{r_j,s_j} + \left( \text{the coefficient of } x^h y^k \text{ in the polynomial } D_j \right),$$

and define the ideal

$$\mathfrak{R}' = \{ \delta_{hk}^{r_j,s_j} \mid 1 \leq j \leq \Lambda_\mu, \ (h,k) \in \mu \} \subseteq k[\mathcal{C}_\mu].$$

**Theorem 6.1.1.** The functions $\delta_{hk}^{r_j,s_j} \in k[\mathcal{C}_\mu]$ generate the kernel of the surjection $u_\mu^*; k[\mathcal{C}_\mu] \to \mathcal{O}_{U_\mu}$; that is, $\mathfrak{R}' = \mathfrak{R}_\mu$.

We need the following:

**Lemma 6.1.2.** The $H_q$ are syzygies of $(D_k)$, that is,

$$H_q \cdot (D_1, \ldots, D_\Lambda_\mu) = 0, \quad 1 \leq q \leq \Lambda_\mu - 1.$$

**Proof.** If one replaces the first row $e_j$ of the determinant in (47) with $H_q$, then the result must be 0, since the matrix now has two equal rows. On the other hand, by linearity of the determinant in the first row, the value of this determinant is given by $H_q \cdot (D_1, \ldots, D_\Lambda_\mu)$; whence, the lemma. \qed

**Proof of Theorem 6.1.1.** By (49), we have that $\mathfrak{R}' \subseteq \mathfrak{R}_\mu$; it remains to establish the reverse inclusion. Recall from (38) that, for $1 \leq q \leq \Lambda_\mu - 1$,

$$H_q \cdot (G_1, \ldots, G_\Lambda_\mu) = \sum_{(h,k) \in \mu} \rho_{hk}^q x^h y^k;$$

since $H_q$ is a syzygy of $(D_k)$, by the lemma, we may write

$$H_q \cdot (G_1 - D_1, \ldots, G_\Lambda_\mu - D_\Lambda_\mu) = \sum_{(h,k) \in \mu} \rho_{hk}^q x^h y^k.$$
By (48) and (50), we have that

\begin{equation}
G_j - D_j = \left( \sum_{(h,k) \in \mu} \delta_{h,k}^{r_i,s_j} x^h y^k \right) - N_j.
\end{equation}

Now, since the term \( \rho_{h,k}^q x^h y^k \) on the right-hand side of (51) can involve only those terms of the factors on the left-hand side of the form

\[ ax^r y^s, \quad \text{with } r \leq h, \ s \leq k, \]

we see that (51) expresses \( \rho_{h,k}^q \) as a \( k[\mathbf{C}] \)-linear combination of the \( \{ \delta_{h,k}^{r_i,s_j} \} \).

Since this holds for all \( 1 \leq q \leq \Lambda_\mu - 1 \), \( (h, k) \in \mu \), we may invoke Theorem 5.1.1 to conclude that

\[ \mathfrak{K}_\mu = \{ \{ \rho_{h,k}^q \} \} \subseteq \mathfrak{K}', \]

which completes the proof. \( \square \)

6.2. An example continued: \( \mu = (2, 1) \). For an example, we return to the case \( \mu = (2, 1) \) considered in Section 5.2. Recalling (42), we see that

\[
\begin{align*}
\mathcal{D}_1 &= \det \begin{pmatrix} 1 & 0 & 0 \\ -y + C_{1,0}^{1,1} & x - C_{1,0}^{2,0} & x + C_{0,1}^{1,1} \\ -y - C_{1,0}^{1,1} & x - C_{1,0}^{2,0} & x - C_{0,1}^{1,1} \\ C_{1,0}^{0,2} & 0 & 0 \\ C_{1,0}^{0,2} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \\
\mathcal{D}_2 &= \det \begin{pmatrix} 1 & 0 & 0 \\ -y + C_{1,0}^{1,1} & x - C_{1,0}^{2,0} & x + C_{0,1}^{1,1} \\ -y - C_{1,0}^{1,1} & x - C_{1,0}^{2,0} & x - C_{0,1}^{1,1} \\ C_{1,0}^{0,2} & 0 & 0 \\ C_{1,0}^{0,2} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \\
\mathcal{D}_3 &= \det \begin{pmatrix} 1 & 0 & 0 \\ -y + C_{1,0}^{1,1} & x - C_{1,0}^{2,0} & x + C_{0,1}^{1,1} \\ -y - C_{1,0}^{1,1} & x - C_{1,0}^{2,0} & x - C_{0,1}^{1,1} \\ C_{1,0}^{0,2} & 0 & 0 \\ C_{1,0}^{0,2} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix};
\end{align*}
\]

whence,

\[
\begin{align*}
\mathcal{D}_1 &= x^2 - C_{1,0}^{2,0} x - C_{0,1}^{2,0} y - (C_{0,1}^{1,1})^2 + C_{0,1}^{0,2} C_{0,1}^{2,0} - C_{1,0}^{1,1} C_{0,1}^{2,0} C_{1,0}^{2,0}, \\
\mathcal{D}_2 &= x y - C_{1,0}^{1,1} x - C_{0,1}^{1,1} y + C_{0,1}^{1,1} C_{1,0}^{1,1} - C_{1,0}^{0,2} C_{0,1}^{2,0}, \\
\mathcal{D}_3 &= y^2 - C_{1,0}^{0,2} x - C_{0,1}^{0,2} y - C_{0,1}^{0,2} C_{1,0}^{1,1} + C_{0,1}^{0,2} C_{1,0}^{1,1} - (C_{1,0}^{1,1})^2 + C_{1,0}^{0,2} C_{1,0}^{2,0}. \\
\end{align*}
\]

Remark 6.2.1. Note that the summands \( N_j \) (48) of the \( D_j \) in (53) all vanish. This rather special phenomenon has some interesting consequences that, being tangential to our present purposes, will not be discussed in this paper. (For example, when the \( N_j \) all vanish, it can be shown that the pseudosyzygies \( H_k \) are invariant under the substitutions \( C_{h,k}^{r_i,s_i} \rightarrow D_{h,k}^{r_i,s_i} \).)
We will, however, “explain” why the phenomenon occurs in this and similar cases in the next section (Corollary 7.2.2).

Computing the functions \( \delta_{hk}^{\tau_j,s_j} \) (50), we find that

\[
\begin{align*}
\delta_{0,0}^{2,0} &= C_{0,0}^{2,0} - (C_{0,1}^{0,1})^2 + C_{0,1}^{0,2} C_{1,0}^{2,0} - C_{1,0}^{1,1} C_{0,1}^{2,0} + C_{0,1}^{1,1} C_{1,0}^{2,0}, \\
\delta_{0,0}^{1,0} &= C_{0,0}^{1,1} + C_{0,1}^{1,1} C_{1,0}^{1,0} - C_{1,0}^{0,2} C_{0,1}^{2,0}, \\
\delta_{0,0}^{0,2} &= C_{0,0}^{0,2} - C_{1,0}^{0,2} C_{0,1}^{1,1} + C_{0,1}^{0,2} C_{1,0}^{1,1} - (C_{1,0}^{1,1})^2 + C_{0,0}^{0,2} C_{1,0}^{2,0}, \\
\delta_{hk}^{\tau_j,s_j} &= 0, \text{ otherwise.}
\end{align*}
\]

(54)

We now express the \( \rho_{hk}^{d} \), listed in (43), in terms of the \( \delta_{hk}^{\tau_j,s_j} \), using the idea in the proof of Theorem 6.1.1. In light of Remark 6.2.1 and (54), we see that (52) yields

\[
(G_1 - D_1, G_2 - D_2, G_3 - D_3) = (-\delta_{0,0}^{2,0}, -\delta_{0,0}^{1,1}, -\delta_{0,0}^{0,2}).
\]

Computing the dot products of the \( \mu \)-pseudosyzygies (42) with the the preceding vector, as in (51), we obtain

\[
\begin{align*}
\rho_{0,0}^1 + \rho_{1,0}^1 x + \rho_{0,1}^1 y &= (-y + C_{1,0}^{1,1}, x - C_{1,0}^{2,0} + C_{0,1}^{1,1}, -C_{0,1}^{2,0}) \cdot \\
&\quad (-\delta_{0,0}^{2,0}, -\delta_{0,0}^{1,1}, -\delta_{0,0}^{0,2}) \\
&\quad = (-C_{1,0}^{1,1} \delta_{0,0}^{2,0} + (C_{1,0}^{2,0} - C_{0,1}^{1,1}) \delta_{0,0}^{1,1} + C_{0,0}^{2,0} \delta_{0,0}^{0,2}) - \\
&\quad \delta_{0,0}^{2,0} x + \delta_{0,0}^{1,1} y, \\
\rho_{0,0}^2 + \rho_{1,0}^2 x + \rho_{0,1}^2 y &= (C_{1,0}^{0,2}, -y + C_{1,0}^{1,1}, C_{0,1}^{0,2}, x - C_{0,1}^{1,1}) \cdot \\
&\quad (-\delta_{0,0}^{2,0}, -\delta_{0,0}^{1,1}, -\delta_{0,0}^{0,2}) \\
&\quad = (-C_{1,0}^{0,2} \delta_{0,0}^{2,0} + (C_{1,0}^{1,1} - C_{0,1}^{0,2}) \delta_{0,0}^{1,1} + C_{0,0}^{1,1} \delta_{0,0}^{0,2}) - \\
&\quad \delta_{0,0}^{2,0} x + \delta_{0,0}^{1,1} y,
\end{align*}
\]

from which we deduce

\[
\begin{align*}
\rho_{1,0}^1 &= -\delta_{0,0}^{1,1}, \\
\rho_{0,1}^1 &= \delta_{0,0}^{2,0}, \\
\rho_{1,0}^2 &= -\delta_{0,0}^{0,2}, \\
\rho_{0,0}^1 &= -C_{1,0}^{1,1} \delta_{0,0}^{2,0} + (C_{1,0}^{2,0} - C_{0,1}^{1,1}) \delta_{0,0}^{1,1} + C_{0,0}^{2,0} \delta_{0,0}^{0,2}, \\
\rho_{0,0}^2 &= -C_{1,0}^{0,2} \delta_{0,0}^{2,0} + (C_{1,0}^{1,1} - C_{0,1}^{0,2}) \delta_{0,0}^{1,1} + C_{0,0}^{1,1} \delta_{0,0}^{0,2}, \\
\rho_{0,1}^2 &= \delta_{0,0}^{1,1}.
\end{align*}
\]

Note that the \( \rho_{hk}^{d} \) in the first three of these relations are the elements of the generating set (44) of the ideal \( \mathfrak{R}_\mu \); one may check that these equations are consistent with the values given in (43) and (54). By replacing the \( \delta \)'s in the
last three relations with their $\rho$-equivalents from the first three relations, we
derive the relations (45), as we promised to do.

7. Smaller generating sets for $\mathcal{O}_{U_\mu}$ and affine cell criteria.

In this, the final section of the paper, we use Theorem 6.1.1 to identify a
subset of $c_\mu$ that generates $\mathcal{O}_{U_\mu}$ as a $k$-algebra, as promised at the end of
Section 2; we thereby obtain a smaller explicit presentation of $\mathcal{O}_{U_\mu}$ as a
quotient of a polynomial ring than those obtained in Sections 5 and 6. We
then derive sufficient conditions on $\mu$ for $U_\mu$ to be a $2n$-dimensional affine
cell in $H^n$; more precisely, for the map (15)
$$
\varepsilon_\mu : U_\mu \to \text{Spec}(k[p_\mu]) = A^2_k
$$
to be an isomorphism.

7.1. A “smaller” set of generators for $\mathcal{O}_{U_\mu}$. Recall that $p_\mu$ denotes the
$2n$-member subset (14) of $c_\mu$ that gives local parameters at the monomial
ideal $I_\mu \subseteq U_\mu$ (8), and is accordingly a $k$-algebraically independent set. We
define
$$
P_\mu = \{ C_{rj,sj}^{tj,sk} \in c_\mu \mid c_{tk}^{tj,sk} \in p_\mu \},
$$
and observe that the map
$$
u_\mu : k[p_\mu] \to \mathcal{O}_{U_\mu}, \quad C_{rj,sj}^{tj,sk} \mapsto C_{tj,sk}^{rj,sj},
$$
the restriction of the surjection (36), is injective.

Recall further that $M_\mu$ denotes the matrix whose rows are the $\mu$-pseudo-
syzygies $H_j$ of $(G_k)$ discussed in Section 5.1. We define
$$
\text{Ex}_\mu = \{ C_{rj,sj}^{tj,sk} \in c_\mu \mid C_{rj,sj}^{tj,sk} \text{ appears in } M_\mu \};
$$
$$
\text{ex}_\mu = \{ c_{rj,sj}^{tj,sk} \mid C_{rj,sj}^{tj,sk} \in \text{Ex}_\mu \} \subseteq c_\mu;
$$
the notation recalls that these are the sets of exposed coefficients in the
language of Section 4.1. We shall soon show that $\text{ex}_\mu$ generates $\mathcal{O}_{U_\mu}$ as a
$k$-algebra; we first pause to establish:

Lemma 7.1.1. $P_\mu \subseteq \text{Ex}_\mu$; whence, $p_\mu \subseteq \text{ex}_\mu$.

Proof. Let $C_{rj,sj}^{tj,sk} \in P_\mu$. Then, by definition of $p_\mu$ (see Section 2.4), one of
the following conditions holds:

- $x^{tj}y^{sk}$ is a top leading monomial and $x^hy^k$ is the right-
  most partition monomial in row $h$ (in the diagram of
  $\mu$), with $s_j \leq k$, or

- $x^{tj}y^{sk}$ is a side leading monomial and $x^hy^k$ is the top-
  most partition monomial in column $k$, with $s_j > k$.

In the first case, $C_{rj,sj}^{tj,sk}$ will appear in the $\mu$-pseudo-syzygy generated by
Algorithm 4.2.1 when $G_j$ is multiplied by $-y_i$; that is, $H_j$; in the second case,
$C_{hk}^{r_j,s_j}$ will appear in the $\mu$-pseudosyzygy generated when $G_j$ is multiplied by $x$; that is, $H_{j-1}$. In either case, $C_{hk}^{r_j,s_j}$ appears in $M_{\mu}$ (or is exposed), and is therefore a member of $\mathbb{E}_\mu$, as desired.

We write

$$\overline{\pi}_\mu^* : k[\mathbb{E}_\mu] \to \mathcal{O}_{U_\mu}, \quad C_{hk}^{r_j,s_j} \mapsto c_{hk}^{r_j,s_j}$$

for the restriction of the map (36) to $k[\mathbb{E}_\mu] \subseteq k[\mathfrak{c}_\mu]$.

**Proposition 7.1.2.** The map $\overline{\pi}_\mu^*$ is surjective, and its kernel is generated by the set $\{\delta_{hk}^{r_j,s_j} \mid C_{hk}^{r_j,s_j} \in \mathbb{E}_\mu\}$. In particular, the set $\mathfrak{e}_\mu$ generates $\mathcal{O}_{U_\mu}$ as a $k$-algebra.

**Proof.** Recall that $\delta_{hk}^{r_j,s_j} = C_{hk}^{r_j,s_j} - D_{hk}^{r_j,s_j}$ (50), where $D_{hk}^{r_j,s_j} \in k[\mathbb{E}_\mu]$ (since a coefficient of a subdeterminant of $M_{\mu}$ can only involve $C_{hk}^{r_j,s_j} \in \mathbb{E}_\mu$). We define a surjective map

$$\tau : k[\mathfrak{c}_\mu] \to k[\mathbb{E}_\mu], \quad C_{hk}^{r_j,s_j} \mapsto C_{hk}^{r_j,s_j} \quad \text{for} \quad C_{hk}^{r_j,s_j} \in \mathbb{E}_\mu,$$

$$\quad \quad C_{hk}^{r_j,s_j} \mapsto D_{hk}^{r_j,s_j} \quad \text{for} \quad C_{hk}^{r_j,s_j} \notin \mathbb{E}_\mu.$$ 

Since $C_{hk}^{r_j,s_j}$ and $D_{hk}^{r_j,s_j}$ have the same image under the surjection $u_\mu^*$ (49), we see that $u_\mu^*$ factors as

$$k[\mathfrak{c}_\mu] \xrightarrow{\tau} k[\mathbb{E}_\mu] \xrightarrow{\overline{\pi}_\mu^*} \mathcal{O}_{U_\mu};$$

it follows at once that $\overline{\pi}_\mu^*$ is a surjection.

Now let $\kappa$ be an element of $\text{ker}(\overline{\pi}_\mu^*)$, and let $\kappa' \in k[\mathfrak{c}_\mu]$ be any preimage of $\kappa$. Since $\kappa' \in \text{ker}(u_\mu^*)$, Theorem 6.1.1 allows us to write

$$\kappa' = \sum f_{hk}^{r_j,s_j} \cdot \delta_{hk}^{r_j,s_j},$$

where the coefficients $f_{hk}^{r_j,s_j} \in k[\mathfrak{c}_\mu]$. Since the $\delta_{hk}^{r_j,s_j}$ corresponding to $C_{hk}^{r_j,s_j} \notin \mathbb{E}_\mu$ map to 0 under $\tau$, we may apply $\tau$ to both sides of (57) to obtain a representation of $\kappa$ as a $k[\mathbb{E}_\mu]$-linear combination of the $\delta_{hk}^{r_j,s_j}$ corresponding to $C_{hk}^{r_j,s_j} \in \mathbb{E}_\mu$; this completes the proof of the proposition.

We indicated at the end of Section 2 that $\mathfrak{e}_\mu$ is typically a much smaller set of $k$-algebra generators of $\mathcal{O}_{U_\mu}$ than is $\mathfrak{c}_\mu$ (but not a minimal generating set, in general). To make this precise, we count the number of elements of $\mathfrak{e}_\mu$, using the notation (16): Each of the $p_1$ top monomials, when multiplied by $-y$, contributes $\ell$ exposed coefficients to $\mathfrak{e}_\mu$; likewise, each of the $\ell$ side monomials, when multiplied by $x$, contributes $p_1$ exposed coefficients. However, there are $d_\mu - 1$ monomials $x^r y^s$ that are both top and side monomials, and each of these exposes twice the $d_\mu$ “northeast corner” coefficients $C_{hk}^{r_j,s_j}$, where $x^h y^k$ is both the highest member of its column and the rightmost
member of its row (see Figure 4). Therefore, the cardinality of $\text{ex}_\mu$ is given by

\[(58) \quad |\text{ex}_\mu| = 2(p_1 \cdot \ell) - d_\mu(d_\mu - 1).\]

For the example $\mu = (5, 3, 2, 2)$ shown in Figure 4, we have that $|\epsilon_\mu| = \Lambda_\mu \cdot n = 7 \cdot 12 = 84$, by (20), whereas $|\text{ex}_\mu| = 2(5 \cdot 4) - 3 \cdot 2 = 34$.

7.2. Sufficient conditions for $U_\mu$ to be an affine cell. We seek conditions under which the map (15)

$$\varepsilon_\mu : U_\mu \rightarrow \text{Spec}(k[p_\mu]) = \mathbb{A}_k^{2n}$$

is an isomorphism, or, equivalently, conditions under which the inclusion $k[p_\mu] \subseteq \mathcal{O}_{U_\mu}$ is an equality. Since $\mathcal{O}_{U_\mu}$ is generated as a $k$-algebra by the set $\text{ex}_\mu$ (Proposition 7.1.2), we obtain the following:

**Corollary 7.2.1.** If every element of $\text{ex}_\mu$ lies in $k[p_\mu]$, then the map $\varepsilon_\mu$ is an isomorphism; in particular, $U_\mu$ is a $2n$-dimensional affine cell in $H^n$.

A special case of particular interest occurs when $P_\mu = \text{Ex}_\mu$ (which implies $p_\mu = \text{ex}_\mu$; recall from Lemma 7.1.1 that $\subseteq$ always holds):

**Corollary 7.2.2.** If $P_\mu = \text{Ex}_\mu$, then the map $\varepsilon_\mu$ is an isomorphism, and consequently $U_\mu$ is an affine cell. Furthermore, the summands $N_j$ (48) of the polynomials $D_j$ (47), $1 \leq j \leq \Lambda_\mu$, all vanish.

**Proof.** Only the last statement requires proof, to which end we write

$$N_j = \sum N_j^{p,q} x^p y^q, \quad N_j^{p,q} \in k[\text{ex}_\mu],$$

and recall that $\widetilde{u}_\mu^*(N_j) = 0$, where $\widetilde{u}_\mu^*$ is the map (37) given by applying $u_\mu^*$ to each coefficient $N_j^{p,q}$. We therefore have that $u_\mu^*(N_j^{p,q}) = 0$, and, by hypothesis, $N_j^{p,q} \in k[\text{ex}_\mu] = k[P_\mu]$. But we noted earlier that the map $u_\mu^*[k[P_\mu]]$ (55) is injective; whence, all the $N_j^{p,q}$ are 0, as desired. □

The case $\mu = (2,1)$, studied in Sections 5.2 and 6.2, provides an example of Corollary 7.2.2. One sees by inspection of (42) that $\text{Ex}_{(2,1)} = P_{(2,1)}$, and we noted both that $U_{2,1}$ is an affine cell (46) and that the polynomials $D_j$, $j = 1, 2, 3$, all have vanishing summands $N_j$ (Remark 6.2.1). We generalize this example in the following section.

7.3. Necessary and sufficient conditions on $\mu$ for $P_\mu = \text{Ex}_\mu$. Recall that we write a partition $\mu = (p_1, p_2, \ldots, p_\ell)$ of $n$ with parts in decreasing order, and that we use the notations (16); in particular, $\mu(i)$ denotes the number of occurrences of the integer $i$ in $\mu$. 

Proposition 7.3.1. Let \( \mu \) be a partition of \( n \). In order that \( P_\mu = \text{Ex}_\mu \), it is necessary and sufficient that

\[
\begin{align*}
P_\mu(p_1+1) &= p_1 - 1, \\
P_\mu(p_1+2) &= p_1 - 2, \\
\vdots \hspace{2cm} & \\
P_\ell &= p_\mu(p_1) + (\ell - \mu(p_1)) = p_1 - (\ell - \mu(p_1));
\end{align*}
\]

that is, the diagram of \( \mu \) has the “sawtooth” form shown in Figure 6, with every step width except possibly the topmost, and every step height except possibly the rightmost, being of size 1.

\[\begin{array}{cccc}
\vdots \\
x^h y^s & x^{h+1} y^{k-1} & x^{h+1} y^k \\
\end{array}\]

Figure 6. Diagram of \( \mu \) satisfying the hypotheses of Proposition 7.3.1.

Proof. We first prove the sufficiency. Suppose that \( \mu \) has the indicated form, and consider the top leading monomial \( m = x^{r_j} y^{s_j} \). The monomial \( m \) either lies above the topmost horizontal edge of the diagram or in one of the sawtooth “notches.” In the former case, each term \( C_{hk}^{r_j,s_j} x^h y^k \) that is exposed when the polynomial \( G_j \) is multiplied by \(-y\) is positioned at the rightmost end of row \( h \) of the diagram, with \( s_j \leq k \); therefore, by (56), \( C_{hk}^{r_j,s_j} \in P_\mu \). In the latter case, the exposed terms in \(-y \cdot G_j\) are either of the form just described or of the form \( C_{hk}^{r_j,s_j} x^h y^k \), with \( x^h y^k \) positioned at the top of column \( k \), and \( s_j > k \); since \( m \) is in this case both a top and a side monomial, we conclude that all the exposed \( C_{hk}^{r_j,s_j} \in P_\mu \). Similarly, one checks that if \( m \) is a side monomial, then all the exposed coefficients in \( x \cdot G_j \) lie in \( P_\mu \); whence, \( \text{Ex}_\mu \subseteq P_\mu \), which yields \( \text{Ex}_\mu = P_\mu \).

It remains to prove the necessity. Arguing by contradiction, we suppose that \( \text{Ex}_\mu = P_\mu \), but that the diagram of \( \mu \) does not have the “sawtooth” form. Suppose that the intermediate horizontal step at height \( h \) has width two or greater, that is, the diagram of \( \mu \) at height \( h \) is as in Figure 7.

\[\begin{array}{cccc}
\vdots \\
x^h y^s & x^{h+1} y^{k-1} & x^{h+1} y^k \\
\end{array}\]

Figure 7. A horizontal step of size \( > 1 \).
Then the coefficient $C_{h+1, s_{h,k-1}}$ is exposed (and therefore in $E_{\mu}$) when the $G$-polynomial with leading monomial $x^{h+1}y^s$ is multiplied by $x$, but according to (56), this coefficient is not in $P_{\mu}$, contradicting $E_{\mu} \subseteq P_{\mu}$. A similar contradiction is obtained if an intermediate vertical step has height two or greater.

Combining the proposition and Corollary 7.2.2, we obtain:

**Corollary 7.3.2.** If $\mu$ is a partition of $n$ satisfying the hypothesis of Proposition 7.3.1, then the map $\varepsilon_{\mu}$ (15) is an isomorphism; consequently, $U_{\mu}$ is an affine cell. Furthermore, the summands $N_j$ (48) of the polynomials $D_j$ (47), $1 \leq j \leq \Lambda_{\mu}$, all vanish.

**Remarks 7.3.3.**

1) Haiman discusses the case $\mu = (1, 1, \ldots, 1)$ in detail, from another direction, in [7, p. 214]; his discussion implies that $U_{\mu}$ is an affine cell, with $O_{U_{\mu}} = k[p_{\mu}]$. Since the diagram of $\mu$ consists of a single column of width 1, with no “sawteeth,” it satisfies the hypothesis of Proposition 7.3.1, and therefore Haiman’s results also follow from Corollary 7.3.2.

2) In case $\mu = (r, r-1, \ldots, 1)$, a partition of $n = r(r+1)/2$, it is clear that the hypothesis of Proposition 7.3.1 is satisfied; whence, $U_{\mu}$ is an affine cell (this is a direct generalization of the case $\mu = (2,1)$ considered earlier). The affine cellularity of $U_{\mu}$ can be obtained in another way by suitably modifying [8, Corollary 4.11 and Example 4.13]. (Similarly, [8, Example 4.12] concerns $\mu = (1, 1, \ldots, 1)$, and yields yet another proof of the affine cellularity in that case.) Note that the monomial ideal $I_{\mu} \subseteq U_{\mu}$ is the “fat point” ideal $(x, y)^r = (x^r, x^{r-1}y, \ldots, y^r) \subseteq k[x, y]$.

**7.4. A sufficient condition on $\mu$ for $E_{\mu} \subseteq k[p_{\mu}]$.** We now further generalize the criterion for affine cellularity proved in Section 7.3 by exploiting Corollary 7.2.1, the hypothesis of which is that $E_{\mu} \subseteq k[p_{\mu}]$. Since it is inconvenient for our current purpose to use the indexing of the leading monomials by $j$, $1 \leq j \leq \Lambda_{\mu}$, we will drop this index from the notations $x^{r_j}y^{s_j}$, $C_{r_k, s_k}^{r_j, s_j}$, etc., and we will denote the polynomials $G_j, g_j$ by $G_{(r,s)}, g_{(r,s)}$, respectively. Recall from (16) that we write $d_{\mu}$ for the number of distinct parts of the partition $\mu$, and $\mu(i)$ for the number of times the integer $i$ occurs in $\mu$.

**Theorem 7.4.1.** Let $\mu$ be a partition of $n$ with distinct parts $p_1 = s_1$, $s_2$, \ldots, $s_{d_{\mu}} = p_{d_{\mu}}$ such that

$$s_2 = s_1 - 1, \ s_3 = s_2 - 1, \ \ldots, \ s_{d_{\mu}} = s_{d_{\mu}-1} - 1, \ \text{and} \ \mu(s_{d_{\mu}}) \leq \mu(s_{d_{\mu}-1}) \leq \cdots \leq \mu(s_2);$$
that is, such that the diagram of \( \mu \) has the shape shown in Figure 8: Every step except possibly for the first is of width 1, and the step heights \( \mu(s_i) \) are nondecreasing as one moves to the right, except for the possibility that at the last stage \( \mu(s_2) > \mu(s_1) \). Then \( \mathsf{ex}_\mu \subseteq k[p_\mu] \).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{diagram.png}
\caption{Diagram of \( \mu \) satisfying the hypothesis of Theorem 7.4.1.}
\end{figure}

7.4.1. Start of the proof of Theorem 7.4.1. Recall that every leading monomial \( x^r y^s \) has either one or two opportunities to contribute coefficient functions to \( \mathsf{ex}_\mu \) as the syzygies of the \( g_{(r',s')} \) are formed: Top leading monomials contribute the exposed coefficients in \( -y \cdot g_{(r,s)} \), side leading monomials contribute the exposed coefficients in \( x \cdot g_{(r,s)} \), and some leading monomials contribute in both ways. We say that a coefficient function \( c^{r,s}_{h,k} \) is covered provided that it is a member of \( k[p_\mu] \). Accordingly, we say that a leading monomial is covered for \( y \) (resp. \( x \)) if its contributions to \( \mathsf{ex}_\mu \) via \( -y \cdot g_{(r,s)} \) (resp. \( x \cdot g_{(r,s)} \)) are all covered. To prove the theorem, we must prove that every top leading monomial is covered for \( y \), and every side leading monomial is covered for \( x \).

To do this, it is convenient to group the leading and partition monomials by columns; that is, by \( y \)-degree \( k \), \( 0 \leq k \leq p_1 = s_1 \). We use the following notation: In the column of \( y \)-degree \( k \), we write \( x^r y^k \) for the leading monomial of least \( x \)-degree in the column; if \( k < p_1 \), this is the unique top leading monomial in the column (see Figure 9). One checks easily that

\[ r_k = \mu(k+1) + r_{(k+1)} = \mu(k+1) + \mu(k+2) + \cdots + \mu(p_1), \]

and that the number \( L(k) \) of leading monomials in column \( k \) is given by

\[ L(k) = \begin{cases} 1, & \text{for } 0 \leq k \leq p_\ell - 1; \\ \mu(k) \geq 1, & \text{for } p_\ell \leq k \leq p_1. \end{cases} \]

To help keep track of the coefficient functions \( c^{r,s}_{h,k} \) that we have shown are covered, we introduce, for each pair of column indices \( (s,k) \) with \( 0 \leq s \leq p_1 \),...
Note that the exposed coefficients in all covered for \( x^s \leq C \) follows: Every entry in the first column of \( \mathbf{C} \). Expressed in terms of the matrices \( \mathbf{C} \), that is, the matrix with \( L(s) \) rows, arbitrarily many columns, and entries given by

\[
\mathbf{C}(s, k) = \begin{bmatrix}
    c_{r_k-1,1}^{r_s+L(s)-1,s} & c_{r_k-2,1}^{r_s+L(s)-1,s} & \cdots \\
    c_{r_k-1,2}^{r_s+L(s)-2,s} & c_{r_k-2,2}^{r_s+L(s)-2,s} & \cdots \\
    \vdots & \vdots & \ddots \\
    c_{r_k-1,s}^{r_s,s} & c_{r_k-2,s}^{r_s,s} & \cdots 
\end{bmatrix};
\]

that is, the matrix with \( L(s) \) rows, arbitrarily many columns, and entries given by

\[
\mathbf{C}(s, k)_{i,j} = c_{r_k-j,k}^{r_s+L(s)-i,s},
\]

where we define the entries to be 0 whenever \( r_k - j < 0 \) (in other words, the columns to the right of the \( r_k \)-th column are all 0).

We now catalogue the current state of our knowledge and summarize what remains to be proved in terms of the matrices \( \mathbf{C}(s, k) \) (initially, we know that the coefficient functions \( c_{r_k}^{r,s} \) are trivially covered). We first consider a side leading monomial \( x^r y^s \) (the hypothesis of Theorem 7.4.1 implies that \( x^r y^s \) is a side monomial \( \iff s \geq p_\ell \)). Note that the exposed coefficients in \( x \cdot g(r, s) \) are the \( c_{r_k-1,k}^{r,s} \), \( 0 \leq k < p_1 \). By (56), \( c_{r_k-1,k}^{r,s} \in \mathbb{P}_n \) provided that \( s > k \). Expressed in terms of the matrices \( \mathbf{C}(s, k) \), the last conclusion reads as follows: Every entry in the first column of \( \mathbf{C}(s, k) \) is covered provided that \( p_\ell \leq s \leq 0 \leq k < s \). To prove that every side leading monomial is covered for \( x \), it remains to prove that the entries in the first column of \( \mathbf{C}(s, k) \) are all covered for \( p_\ell \leq s \leq k \leq p_1 - 1 \).

We next consider a top leading monomial \( x^r y^s = x^r s y^s \), \( 0 \leq s \leq p_1 - 1 \). Note that the exposed coefficients in \( -y \cdot g(r, s) \) are the \( c_{h,k}^{r,s} \) with \( p_\ell - 1 \leq k \).
$\leq p_1 - 1$, $r_k - \mu(k+1) \leq h \leq r_k - 1$ (since these correspond to the partition monomials in column $k$ that are rightmost in their respective rows $h$, as illustrated in Figure 9). By (56), these coefficients lie in $p_\mu$ provided that $s \leq k$; in other words, we have that the first $\mu(k+1)$ entries in the bottom row of the matrix $C(s,k)$ are covered for $s, k$ in the given ranges and $s \leq k$. (Note that when $k = p_1 - 1$, every entry in the bottom row of $C(s,k)$ is covered, since those of index $j > \mu(k+1)$ ($= \mu(p_1) = r_{p_1-1} = r_k$) are equal to 0 by definition.) To prove that every top leading monomial is covered for $y$, it remains to prove that the first $\mu(k+1)$ entries in the bottom row of the matrix $C(s,k)$ are covered for $p_\ell - 1 \leq k < s \leq p_1 - 1$.

In summary, we know that:

- If $p_\ell \leq s \leq p_1$, $0 \leq k < s$, then the $c$’s in the first column of $C(s,k)$ are all covered, and
- if $0 \leq s \leq p_1 - 1$, $p_\ell - 1 \leq k \leq p_1 - 1$, and $s \leq k$, then the first $\mu(k+1)$ $c$’s on the bottom row of $C(s,k)$ are all covered, and all the entries on the bottom row of $C(s,k)$ are covered if $k = p_1 - 1$,

and it remains to prove that:

- If $p_\ell \leq s < p_1$, $s \leq k < p_1$, then the $c$’s in the first column of $C(s,k)$ are all covered, and
- if $p_\ell - 1 \leq k < s \leq p_1 - 1$, then the first $\mu(k+1)$ $c$’s on the bottom row of $C(s,k)$ are all covered.

We can dispose of one case right away: If the partition $\mu$ contains only one distinct part (i.e., $p_1 = p_\ell$), then the hypothesis of Theorem 7.4.1 is clearly satisfied and (60) is vacuously true, so the desired conclusion ($\text{ex}_\mu \subseteq k[p_\mu]$) follows. (Note that Proposition 7.3.1 yields this conclusion in the form $p_\mu = \text{ex}_\mu$.) Therefore, we may henceforth assume that $\mu$ has at least two distinct parts.

7.4.2. Reduction of the proof of Theorem 7.4.1 to a lemma. To prove (60), we will in fact prove more: In each case the goal is to establish that the first column (resp. initial segment of the bottom row) of a matrix $C(s,k)$ is covered. We will do this by showing that the entire lower-left triangular portion of the matrix having the first column (resp. initial segment of the bottom row) in question at its vertical (resp. horizontal) leg is covered; see Figure 10.

More precisely, Theorem 7.4.1 is an immediate consequence of the following:

**Lemma 7.4.2.** Let the partition $\mu$ satisfy the hypothesis of Theorem 7.4.1 and have at least two distinct parts. Then, for each $s$, $p_\ell \leq s \leq p_1 - 1$, we have that:
1) If \( s \leq k \leq p_1 - 1 \), the entries \( C(s, k)_{(i,j)} \) of the matrix \( C(s, k) \) lying in the lower-left triangular region defined by \( i - j \geq 0 \) (which includes the entire first column) are all covered.

2) If \( s - 1 \leq k < s' \leq p_1 - 1 \), the entries \( C(s', k)_{(i,j)} \) of the matrix \( C(s', k) \) lying in the lower-left triangular region defined by \( i - j \geq \mu(s') - \mu(k+1) \) (which includes the first \( \mu(k+1) \) entries on the bottom row) are all covered.

The proof of the lemma proceeds by descending induction on \( s \) in the nonempty range \( p_\ell \leq s \leq p_1 - 1 \). A key ingredient is the identity (11), which when recast in terms of the matrices \( C(s, k) \) reads as follows (for \( 2 \leq i \leq \mu(s) \)):

\[
C(s, k)_{(i-1,j)} = C(s, k)_{(i,j+1)} + \sum_{k'=0}^{p_1-1} C(s, k')_{(i,1)} \cdot C(k', k)_{(L(k'), j)}.
\]

### 7.4.3. Proof of Lemma 7.4.2.

Setting aside the base case for the moment, we make the inductive hypothesis that for some \( s, p_\ell \leq s \leq p_1 - 1 \), we have shown that the statements of the lemma hold for all integers in the interval \([s+1, p_1 - 1] \). We proceed to show that the statements of the lemma hold for \( s \).

We begin with the proof of the first statement: We must show that in each of the matrices \( C(s, k) \) with \( s \leq k \leq p_1 - 1 \), the lower-left triangular region having the entire first column as vertical leg is covered. At the outset, we know by (59) that an initial segment of the bottom row of each \( C(s, k) \) is covered; in particular, the element \( C(s, k)_{(L(s), 1)} \) in the lower-left corner is covered. If \( L(s) = \mu(s) = 1 \), we are done, so we assume that \( \mu(s) > 1 \) and that for each \( k' \) in the interval \([s, p_1 - 1] \), the lower-left triangular region of \( C(s, k') \) defined by \( i - j \geq \mu(s) - t > 0 \) is covered. Choose a particular \( k \) in \([s, p_1 - 1] \) and entry \((i, j)\) of \( C(s, k) \) such that \( i - j = \mu(s) - t \), and consider

---

**Figure 10.** Lower-left triangular portion of the matrix \( C(s, k) \) having the first column as its vertical leg.
the identity (61) specialized to this situation:

\[ \mathcal{C}(s, k)_{(i-1,j)} = \mathcal{C}(s, k)_{(i,j+1)} + \sum_{k^* = 0}^{s-1} \mathcal{C}(s, k^*)_{(i,1)} \cdot \mathcal{C}(k^*, k)_{(L(k^*), j)} \\
+ \sum_{k' = s}^{p_1-1} \mathcal{C}(s, k')_{(i,1)} \cdot \mathcal{C}(k', k)_{(L(k'), j)}. \]

Note that \( \mathcal{C}(s, k)_{(i-1,j)} \) lies immediately “northwest” of \( \mathcal{C}(s, k)_{(i,j+1)} \) along the diagonal lying immediately above the lower-left triangular region that we have inductively assumed is covered. We claim that the remaining matrix entries on the right-hand side of (62) are all covered:

- \( \mathcal{C}(s, k^*)_{(i,1)} \) is covered by (59), since it is in the first column and \( k^* < s \);
- \( \mathcal{C}(k^*, k)_{(L(k^*), j)} \) is covered by (59), since it lies on the bottom row, \( k^* < s \leq k \), and for \( k < p_1 - 1, j < \mu(s) \leq \mu(k+1) \) (the penultimate inequality holds since \( i \leq L(s) = \mu(s) \) and \( i - j = \mu(s) - t > 0 \), and the last reflects the nondecreasing stepsize hypothesis on the partition \( \mu \), which is here invoked for the first time in the proof);
- \( \mathcal{C}(s, k')_{(i,1)} \) is covered because it lies in the lower-left triangular region that we are inductively assuming is covered, since \( i - 1 \geq i - j = \mu(s) - t \);
- \( \mathcal{C}(k', k)_{(L(k'), j)} \) is covered for one of two reasons: If \( k' \leq k \), this matrix entry is covered by (59), since it lies on the bottom row and, for \( k < p_1 - 1, j < \mu(s) \leq \mu(k+1) \) as in the second bullet. If \( k' > k \), this matrix entry is covered because of the inductive hypothesis concerning \( s \): Such a \( k' \) is in the interval \([s+1, p_1-1]\) and \( k \geq s \), so the second statement of the lemma applies and tells us that the lower-left triangular region of \( \mathcal{C}(k', k) \) defined by \( i - j \geq \mu(k') - \mu(k+1) \) is covered. Our entry lies in this region since \( L(k') - j = \mu(k') - j \geq \mu(k') - \mu(s) \geq \mu(k') - \mu(k+1) \), where we have again used the inequalities \( j < \mu(s) \leq \mu(k+1) \) that follow as before.

It follows that whenever \( \mathcal{C}(s, k)_{(i,j+1)} \) is covered, so too is its northwest neighbor \( \mathcal{C}(s, k)_{(i-1,j)} \). We claim that the matrix entry lying farthest to the southeast (and therefore in the last row) on the diagonal containing the latter entries is covered: This entry has indices \( (\mu(s), (t + 1)) \), where \( i - j = \mu(s) - t > 0 \). By (59), the first \( \mu(k+1) \) entries in the last row of \( \mathcal{C}(s, k) \) are covered, and all entries in the last row are covered if \( k = p_1 - 1 \).

If \( k < p_1 - 1 \), we have \( t + 1 \leq \mu(s) \leq \mu(k+1) \) (as before); whence, the claim. Therefore, starting at the southeasternmost element and proceeding stepwise to the northwest along our diagonal, we obtain that the lower-left triangular region of \( \mathcal{C}(s, k) \) defined by \( i - j \geq \mu(s) - (t + 1) \) is covered. Since \( k \) was chosen arbitrarily in \([s, p_1 - 1]\), we may conclude by induction on \( t \).
that the lower-left triangular region defined by \( i-j \geq 0 \) is covered for each of the matrices \( C(s, k) \), as \( k \) ranges over the interval \([s, p_1 - 1]\). This completes the proof (under the inductive hypothesis on \( s \)) that the first statement of the lemma holds for \( s \).

We now prove that the second statement of the lemma holds for \( s \). In light of the induction hypothesis on \( s \), it remains to prove that for every \( s' \) in the interval \([s, p_1 - 1]\), the lower-left triangular region of the matrix \( C(s', s-1) \) defined by \( i-j \geq \mu(s')-\mu(s) \) is covered. By (59), the first column of each \( C(s', s-1) \) is covered; in particular, the lower left-hand element \( C(s', s-1)_{(L(s'),1)} \) comprising the lower-left triangular region defined by \( i-j \geq L(s')-1 = \mu(s')-1 \) is covered. If \( \mu(s) = 1 \), we are done, so suppose that \( \mu(s) > 1 \) and that for all \( s' \) in \([s, p_1 - 1]\), the lower-left triangular region of \( C(s', s-1) \) defined by \( i-j \geq \mu(s')-t > \mu(s')-\mu(s) \geq 0 \) is covered (the last inequality follows from the nondecreasing stepsize hypothesis on \( \mu \)). Choose a particular value \( \tilde{s} \) among the \( s' \) and an entry \((i, j)\) in \( C(\tilde{s}, s-1) \) such that \( i-j = \mu(\tilde{s})-t \), and consider the identity (61) specialized to this situation:

\[
C(\tilde{s}, s-1)_{(i-1,j)} = C(\tilde{s}, s-1)_{(i,j+1)} + \sum_{k^*=0}^{s-1} C(\tilde{s}, k^*)_{(i,1)} \cdot C(k^*, s-1)_{(L(k^*),j)} + \sum_{k'=s}^{p_1-1} C(\tilde{s}, k')_{(i,1)} \cdot C(k', s-1)_{(L(k'),j)}.
\]

We claim that the matrix entries on the right-hand side of (63), except possibly for the first, are covered:

- \( C(\tilde{s}, k^*)_{(i,1)} \) is covered by (59), since it lies in the first column and \( k^* < \tilde{s} \);
- \( C(k^*, s-1)_{(L(k^*),j)} \) is covered by (59), since it lies in the last row; \( k^* \leq s-1 \), and \( j \leq t < \mu(s) = \mu((s-1) + 1) \) (the double inequality holds since \( i-j = \mu(\tilde{s})-t > \mu(\tilde{s})-\mu(s) \), and \( \mu(s) \geq 1 \)), being a row index in \( C(\tilde{s}, s-1) \), is \( \leq L(s) = \mu(s) \);
- \( C(\tilde{s}, k')_{(i,1)} \) is covered for one of two reasons: If \( k' < \tilde{s} \), this entry is covered by (59). If \( \tilde{s} \leq k' \), this entry is covered by the first statement of Lemma 7.4.2, already verified by induction to hold for all \( s' \) in the interval \([s, p_1 - 1]\);
- \( C(k', s-1)_{(L(k'),j)} \) is covered by our induction hypothesis (namely, that the lower-left triangular region of \( C(s', s-1) \) defined by \( i-j \geq \mu(s')-t \) is covered for all \( s' \) in the interval \([s, p_1 - 1]\)), since the inequality \( j \leq t \) noted in the second bullet implies that \( L(k') - j = \mu(k') - j \geq \mu(k') - t \).
It follows that if $\mathcal{C}(\tilde{s}, s - 1)_{(i-1,j)}$ is covered, then so is $\mathcal{C}(\tilde{s}, s - 1)_{(i,j+1)}$; therefore, beginning with the extreme northwest member of the diagonal containing these entries (which, being in the first column, is covered, as noted earlier), and proceeding stepwise from northwest to southeast, we obtain that the entire diagonal is covered; whence, the larger lower-left triangular region defined by $i - j \geq \mu(\tilde{s}) - (t + 1) \geq \mu(s) - \mu(s)$ is covered. Since $\tilde{s}$ was arbitrarily chosen among the $s'$, the last conclusion applies to them all; whence, we may conclude by induction on $t$ that, for all $s'$ in the interval $[s, p_1 - 1]$, the lower-left triangular region of the matrix $\mathcal{C}(s', s - 1)$ defined by $i - j \geq \mu(s') - \mu(s)$ is covered. This completes the proof (under the inductive hypothesis on $s$) that the second statement of Lemma 7.4.2 holds for $s$.

To complete the proof of the lemma, it remains to prove the base case for the descending induction, that is, that the statements of the lemma hold for $s = p_1 - 1$. One sees easily that these statements reduce to the following:

1) The entries in the lower-left triangular region of $\mathcal{C}(p_1 - 1, p_1 - 1)$ defined by $i - j \geq 0$ are all covered.

2) The entries in the lower-left triangular region of $\mathcal{C}(p_1 - 1, p_1 - 2)$ defined by $i - j \geq 0$ are all covered.

We leave the proofs as exercises for the reader; simpler versions of the arguments used earlier suffice. This completes the proof of Lemma 7.4.2 and Theorem 7.4.1.

7.5. Further sufficient conditions for $U_\mu$ to be an affine cell. The following result is an immediate consequence of Theorem 7.4.1 and Corollary 7.2.1.

Corollary 7.5.1. If $\mu$ is a partition satisfying the hypothesis of Theorem 7.4.1, then the map $\varepsilon_\mu$ (15) is an isomorphism; consequently, $U_\mu$ is an affine cell.

A simple observation affords a slight generalization: The isomorphism $\tau: \mathbb{A}^2_\mathbb{K} \to \mathbb{A}^2_\mathbb{K}$, defined by

$$\tau^*: \mathbb{K}[x,y] \to \mathbb{K}[x,y], \quad x \mapsto y, \quad y \mapsto x,$$

induces (by pullback of subschemes) an isomorphism $\mathbf{H}^n \to \mathbf{H}^n$, under which $I \in \mathbf{H}^n$ maps to $\tau^*(I)$. If $I \in U_\mu$, then $\tau^*(I) \in U_{\mu'}$, where $\mu'$ is the conjugate partition to $\mu$, that is, the partition whose diagram is obtained from the diagram of $\mu$ by interchanging rows and columns, so that $(k, h) \in \mu$ if and only if $(h, k) \in \mu'$. It follows easily that $\tau$ induces an isomorphism

$$\tau_\mu: U_\mu \to U_{\mu'}$$

with comorphism $\tau_\mu^*: \mathcal{O}_{U_{\mu'}} \to \mathcal{O}_{U_\mu}$ defined by $c_{hk}^{pq} \mapsto c_{kh}^{qp}$
(we write \( \hat{c} \) for the \( c \)'s in \( O_{U_{\mu'}} \)) and inducing a bijection \( p_{\mu'} \rightarrow p_{\mu} \); whence, we obtain:

**Proposition 7.5.2.** \( U_{\mu} \) is an affine cell if and only if \( U_{\mu'} \) is an affine cell; moreover, the map \( \varepsilon_{\mu'} \) is an isomorphism if and only if \( \varepsilon_{\mu} \) is an isomorphism.

For example, the partition \( \mu = (4, 3, 3) \) satisfies the hypothesis of Theorem 7.4.1; therefore, both \( U_{\mu} \) and \( U_{\mu'} \) are affine cells, where the conjugate partition \( \mu' = (3, 3, 3, 1) \). Note that the latter partition does not satisfy the hypothesis of Theorem 7.4.1.

We end this paper with two brief remarks that complement the results obtained here:

**Remarks 7.5.3.**

1) By direct computation it can be shown that the map \( \varepsilon_{\mu} \) is an isomorphism (and therefore \( U_{\mu} \) is an affine cell) in case \( \mu = (3, 2, 1, 1) \); however, neither \( \mu \) nor its conjugate \( \mu' = (4, 2, 1) \) satisfy the hypothesis of Theorem 7.4.1.

2) Haiman observes that \( U_{\mu} \) need not be an affine cell for every \( \mu \) [7, footnote, p. 207].

**References**


Received September 6, 2000 and revised July 17, 2001. The author thanks the Department of Mathematics at Texas A&M University for its hospitality during the writing of the first version of this paper, and Skidmore College for sabbatical support. He also thanks the referee for many helpful suggestions.

**Department of Mathematics and Computer Science**

**Skidmore College**

**Saratoga Springs, New York 12866**

**E-mail address:** mhuibreg@skidmore.edu
Suppose \( X \) is a simply connected mod \( p \) \( H \)-space such that the mod \( p \) cohomology \( H^*(X; \mathbb{Z}/p) \) is finitely generated as an algebra. Our first result shows that if \( X \) is an \( A_n \)-space, then \( X \) is the total space of a principal \( A_n \)-fibration with base a finite \( A_n \)-space and fiber a finite product of \( \mathbb{C}P^\infty \)s. As an application of the first result, it is shown that if \( X \) is a quasi \( C_p \)-space, then \( X \) is homotopy equivalent to a finite product of \( \mathbb{C}P^\infty \)s.

1. Introduction.

The theory of \( H \)-spaces is a generalization of the homotopy theory of Lie groups, and it has been investigated as one of the most important objects of study in algebraic topology. It is useful to consider the \( H \)-spaces at a prime by using the completion of Bousfield-Kan \[2\]. Given a prime \( p \), an \( H \)-space which is completed at \( p \) is called a mod \( p \) \( H \)-space. In this paper, homotopy equivalence means mod \( p \) homotopy equivalence and cohomology is mod \( p \) cohomology unless otherwise specified.

In recent decades, many theorems have been proved about mod \( p \) finite \( H \)-spaces (cf. \[6\], \[12\] and \[17\]), which suggest that they are similar to Lie groups. In this paper, we study mod \( p \) \( H \)-spaces which need not be finite, but whose cohomology rings are finitely generated. It is known that the three-connected cover \( G(3) \) of a Lie group \( G \) is such an \( H \)-space. Another example is the infinite dimensional complex projective space \( \mathbb{C}P^\infty \). Recently, Broto and Crespo \[3\] and \[4\] have obtained remarkable results about mod \( p \) \( H \)-spaces with finitely generated cohomology. It follows from their results that such an \( H \)-space is the total space of a principal \( H \)-fibration with base a mod \( p \) finite \( H \)-space and fiber a product of \( \mathbb{C}P^\infty \)s. One of the purposes of this paper is to generalize their results to the case of higher homotopy associative mod \( p \) \( H \)-spaces with finitely generated cohomology.

Stasheff \[22\] introduced the notion of the higher homotopy associativity of \( H \)-spaces as a series of intermediate stages between \( H \)-spaces and loop spaces. An \( A_2 \)-space is an \( H \)-space with a multiplication \( M_2 : X \times X \to X \), and an \( A_3 \)-space is a homotopy associative \( H \)-space. Now we denote \( M_2(x, y) = xy \) for \( x, y \in X \). Let \( X \) be an \( A_3 \)-space, and let \( M_3 : I \times X^3 \to X \)
be a map satisfying that $M_3(0, x, y, z) = (xy)z$ and $M_3(1, x, y, z) = x(yz)$ for $x, y, z \in X$. By using the map $M_3$, we can define a map $\tilde{M}_4 : S^1 \times X^4 \to X$ such that $\tilde{M}_4(t, x, y, z, w)$ is the pentagon in Figure 1.

\begin{center}
\begin{tikzpicture}
    \node (a) at (0,0) {$(x(yz))w$};
    \node (b) at (2,1) {$(x(y)zw)$};
    \node (c) at (2,-1) {$(xy)(zw)$};
    \node (d) at (4,0) {$(xy)zw$};
    \node (e) at (6,0) {$(xy)z)w$};
    \node (f) at (4,-2) {$x((yz)w)$};
    \node (g) at (2,-2) {$x(yzw)$};
    \draw (a) -- (b);
    \draw (b) -- (c);
    \draw (c) -- (d);
    \draw (d) -- (e);
    \draw (e) -- (f);
    \draw (f) -- (g);
    \draw (g) -- (a);
\end{tikzpicture}
\end{center}

Figure 1. $A_4$-form on $X$.

$X$ is said to be an $A_4$-space if there exists a map $M_4 : D^2 \times X^4 \to X$ with $M_4|_{S^1 \times X^4} = \tilde{M}_4$. In general, an $H$-space $X$ is called an $A_n$-space if there exists an $A_n$-form $\{M_i : D^{i-2} \times X^i \to X\}_{2 \leq i \leq n}$ satisfying some conditions. Figure 2 denotes the $A_5$-form on $X$ (see §2). Furthermore, an $A_\infty$-space has the homotopy type of a loop space. Similarly, an $A_n$-map is defined as an $H$-map between $A_n$-spaces preserving the $A_n$-forms (see §2). An $H$-fibration consisting of $A_n$-spaces and $A_n$-maps is called an $A_n$-fibration.

Our first result is stated as follows:

**Theorem A.** If $X$ is a simply connected $A_n$-space such that the mod $p$ cohomology $H^*(X; \mathbb{Z}/p)$ is finitely generated as an algebra, then we have a simply connected finite $A_n$-space $Y$ and a principal $A_n$-fibration

$$F \longrightarrow X \longrightarrow Y,$$

where the fiber $F$ is the direct product of a finite number of $\mathbb{C}P^\infty$s.

The above theorem is regarded as a generalization of [3, Thm. 1.1] and [4, Thm. 1.1] since $H$-space is the same as $A_2$-space. From Theorem A, it is possible to reduce a problem about $A_n$-spaces with finitely generated cohomology to the case of finite $A_n$-spaces. In this paper, Theorem A is used to study the higher homotopy commutativity of $H$-spaces with finitely generated cohomology. In the case of $p = 2$, Slack has shown the following result:

**Theorem 1.1** ([21, Thm. 0.1]). If $X$ is a simply connected homotopy commutative mod 2 $H$-space such that the mod 2 cohomology $H^*(X; \mathbb{Z}/2)$ is finitely generated as an algebra, then $X$ is homotopy equivalent to a finite product of $\mathbb{C}P^\infty$s.

Here we note that Broto-Crespo [3, Cor. 1.5] gave another proof of Theorem 1.1. On the other hand, the odd prime version of Theorem 1.1 does
not hold. In fact, it was shown by Iriye-Kono [9, Thm. 1.3] that for an odd prime $p$, any connected mod $p$ $H$-space possesses a multiplication which is homotopy commutative. Furthermore, one may expect that a simply connected homotopy commutative mod $p$ loop space with finitely generated cohomology has the homotopy type of a product of $\mathbb{C}P^\infty$s. However, we see that by McGibbon [18, Thm. 2], $Sp(2)$ for $p = 3$ and $S^3$ for $p \geq 5$ are counterexamples.

To describe an odd prime version of Theorem 1.1, we need to generalize the homotopy commutativity of $H$-spaces to the higher ones. Such notions were first considered by Sugawara [24] and Williams [25] in the case of loop spaces. Later Hemmi [8] introduced the higher homotopy commutativity of $H$-spaces. Let $X$ be an $A_n$-space, and let $P_i(X)$ denote the $i$-th projective space of $X$ for $1 \leq i \leq n$. A quasi $C_n$-form on $X$ is defined by a system of maps $\{\lambda_i : (\Sigma X)^i \rightarrow P_i(X)\}_{1 \leq i \leq n}$ satisfying some conditions (see §3). An $A_n$-space which has a quasi $C_n$-form is called a quasi $C_n$-space. In [8, Thm. 1.1], Hemmi has shown that if $X$ is a simply connected finite quasi $C_p$-space, then $X$ is contractible (see Theorem 3.3). Now we generalize his result to the case of quasi $C_p$-spaces with finitely generated cohomology.

**Theorem B.** If $X$ is a simply connected quasi $C_p$-space such that the mod $p$ cohomology $H^*(X; \mathbb{Z}/p)$ is finitely generated as an algebra, then $X$ is homotopy equivalent to a finite product of $\mathbb{C}P^\infty$s.

Theorem B was first conjectured by Slack [21, pp. 4-5], and was suggested to the author by Lin. In the above theorem, it is impossible to relax the condition of quasi $C_p$-space to quasi $C_{p-1}$-space. In fact, by [8, Thm. 2.4], the odd dimensional sphere $S^{2n-1}$ is a quasi $C_{p-1}$-space for any $n \geq 1$. Now we note that Theorem B implies Theorem 1.1 since a homotopy commutative $H$-space is a quasi $C_2$-space by [8, Prop. 2.3] (see also [23, Thm. 13.6]). Furthermore, in the case that $X$ is a loop space, by [8, Thm. 2.2], quasi $C_n$-space is the same condition as $C_n$-space in the sense of Williams [25, Def. 5] (see also [20, Thm. 3.2]), and so Theorem B implies [14, Thm. C]. In particular, since the loop space of an $H$-space is a quasi $C_n$-space for all $n \geq 1$, we have the following result (see also [15] and [16]):

**Theorem 1.2** ([13, Thm. A]). Let $X$ be a simply connected mod $p$ $H$-space with finitely generated mod $p$ cohomology. If $X$ has the homotopy type of the loop space of an $H$-space, then $X$ is homotopy equivalent to a finite product of $\mathbb{C}P^\infty$s.

This paper is organized as follows: In §2, we recall the nullification functor and the colocalization functor introduced by Dror Farjoun [5]. It is shown that those homotopy functors preserve the higher homotopy associativity of $H$-spaces (see Theorem 2.1). §3 is devoted to the proofs of Theorem A and Theorem B. First we recall the results of Broto and Crespo [3] and
about H-spaces with finitely generated cohomology. By combining their results with Theorem 2.1 obtained in §2, we can prove Theorem A. To prove Theorem B, we show that the nullification functor $L_{BZ/p}$ with respect to $BZ/p$ preserves a quasi $C_n$-form (see Theorem 3.5). As a consequence of Theorem A and Theorem 3.5, we obtain the proof of Theorem B.

The content of the paper was first presented in JAMI conference on homotopy theory at Johns Hopkins University in March 2000. The author is grateful to the organizers for their kind invitation and hospitality. I would also like to thank Jim Stasheff for many helpful comments on the manuscript of the paper. Furthermore, I wish to express my appreciation to Yutaka Hemmi and Jim Lin for many discussions about the higher homotopy commutativity of H-spaces. In particular, Theorem B was suggested to me by Jim Lin. Finally, I am grateful to Takao Matumoto and Mitsunori Imaoka for their encouragements.

2. Nullification functor and colocalization functor.

Dror Farjoun [5] introduced the nullification functor and the colocalization functor with respect to spaces. Let $S_*$ denote the category of pointed spaces having the homotopy types of CW-complexes, and let $A \in S_*$. A space $X \in S_*$ is called $A$-null if the pointed mapping space $Map_*(A, X)$ is contractible. By Dror Farjoun [5, Thm. 1.A.3], we have the $A$-nullification functor $L_A : S_* \to S_*$. Given a space $X$, the $A$-nullification $L_A(X)$ is $A$-null, and we have the natural map $\phi_X : X \to L_A(X)$. By [5, Thm. 1.C.1], $\phi_X$ is homotopically universal, that is, for any $A$-null space $Z$ and a map $\zeta : X \to Z$, there exists a map $\tilde{\zeta} : L_A(X) \to Z$ unique up to homotopy such that $\tilde{\zeta}\phi_X \simeq \zeta$.

Furthermore, it is known by [5, p. 18] that the natural map $\phi_X$ induces a homotopy equivalence

$$\phi_X^* : Map_*(L_A(X), Z) \longrightarrow Map_*(X, Z)$$

for an $A$-null space $Z$.

Let $X, Y \in S_*$. A map $f : Y \to X$ is called an $A$-equivalence if the induced map

$$f_* : Map_*(A, Y) \longrightarrow Map_*(A, X)$$

is a homotopy equivalence. In [5, Prop. 2.B.1], Dror Farjoun constructed the $A$-colocalization functor $CW_A : S_* \to S_*$. Given a space $X$, we have the natural map $\psi_X : CW_A(X) \to X$ which is an $A$-equivalence. By [5, Thm. 2.B.3], the natural map $\psi_X$ is homotopically universal among all $A$-equivalences, that is, for any $A$-equivalence $\xi : Y \to X$, there exists a map $\tilde{\xi} : CW_A(X) \to Y$ unique up to homotopy such that $\tilde{\xi}\psi_X \simeq \psi_X$. If the natural $A$-equivalence $\psi_X : CW_A(X) \to X$ is a homotopy equivalence, then $X$ is called an $A$-cellular space.
To prove Theorem A, we show that those homotopy functors $L_A$ and $CW_A$ preserve the higher homotopy associativity of $H$-spaces (see Theorem 2.1). Now we recall the definitions of an $A_n$-space and an $A_n$-map introduced by Stasheff [22] and Iwase-Mimura [11], respectively.

Stasheff [22] introduced the notion of the higher homotopy associativity of $H$-spaces. To describe an $A_n$-form on an $H$-space, he defined a special complex $K_i$ which is homeomorphic to the $(i-2)$-dimensional disk for $i \geq 2$. Let $L_i = \partial K_i$ denote the boundary of the complex $K_i$. Then $L_i$ is the union of $(i(i-1)/2 - 1)$-faces $K_k(r,s)$ for $r,s \geq 2$, $1 \leq k \leq r$, $r+s = i+1$, and the face $K_k(r,s)$ is homeomorphic to $K_r \times K_s$ by the face operator $\partial_k(r,s) : K_r \times K_s \to K_k(r,s)$.

An $A_n$-form on $X$ consists of a system of maps $\{M_i : K_i \times X^i \to X\}_{2 \leq i \leq n}$ satisfying the following conditions:

\begin{equation}
M_2(x,*) = M_2(*,x) = x,
\end{equation}

that is, $M_2 : X^2 \to X$ is a multiplication with unit, where $K_2 \times X^2 = \{\ast\} \times X^2$ is identified with $X^2$.

\begin{equation}
M_i(\partial_k(r,s)(\rho,\sigma),x_1,\ldots,x_i) = M_r(\rho,x_1,\ldots,x_{k-1},M_s(\sigma,x_k,\ldots,x_{k+s-1}),x_{k+s},\ldots,x_i)
\end{equation}

for $(\rho,\sigma) \in K_r \times K_s$.

\begin{equation}
M_i(\tau,x_1,\ldots,x_{j-1},*,x_{j+1},\ldots,x_i) = M_{i-1}(s_j(\tau),x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_i),
\end{equation}

where $s_j : K_i \to K_{i-1}$ denotes the degeneracy map for $1 \leq j \leq i$ (see [22, I Prop. 3]). A space $X$ together with an $A_n$-form is called an $A_n$-space. If $X$ has a system of maps $\{M_i : K_i \times X^i \to X\}_{i \geq 2}$ such that $\{M_i\}_{2 \leq i \leq n}$ is an $A_n$-form on $X$ for any $n \geq 2$, then $X$ is called an $A_\infty$-space.

![Figure 2. $A_5$-form on $X$.](image-url)
It is natural to consider the notion of the higher homotopy associativity of maps between \(A_n\)-spaces. Such notions were first considered by Sugawara [24, §2] and Stasheff [23, Def. 11.9] under some restricted situations. The full generality was described by Iwase-Mimura [11]. To construct an \(A_n\)-form of a map, they defined a special complex \(\Gamma_i\) which is homeomorphic to the \((i - 1)\)-dimensional disk for \(i \geq 1\). Let \(\Lambda_\ell = \partial \Gamma_i\) denote the boundary of the complex \(\Gamma_i\). Then \(\Lambda_\ell\) is the union of the following \((i(i - 1)/2 + 2^{i-1} - 1)\)-faces:

\[
\begin{align*}
\Gamma_k(r, s) & \quad \text{for } 1 \leq k \leq r, 1 \leq r \leq i - 1, r + s = i + 1, \\
\Gamma(t; r_1, \ldots, t_l) & \quad \text{for } 2 \leq t \leq i, r_j \geq 1, r_1 + \cdots + r_t = i,
\end{align*}
\]

and the faces \(\Gamma_k(r, s)\) and \(\Gamma(t; r_1, \ldots, t_l)\) are homeomorphic to \(\Gamma_r \times K_s\) and \(K_t \times \Gamma_{r_1} \times \cdots \times \Gamma_{r_l}\), respectively. The homeomorphisms \(\delta_k(r, s) : \Gamma_r \times K_s \to \Gamma_k(r, s)\) and \(\delta(t; r_1, \ldots, r_l) : K_t \times \Gamma_{r_1} \times \cdots \times \Gamma_{r_l} \to \Gamma(t; r_1, \ldots, r_l)\) are called the face operators on \(\Gamma_i\). Furthermore, we have the degeneracy operations \(d_j : \Gamma_i \to \Gamma_{i-1}\) for \(1 \leq j \leq i\) satisfying certain conditions (see [11, (2-d)]).

From the construction of \(\Gamma_i\), there exists a homeomorphism \(\zeta_i : I \times K_i \to \Gamma_i\) such that \(\zeta_i(\{1\} \times K_i) = \Gamma_1(1, i), \zeta_i(\{0\} \times K_i) = \Gamma(0; i, \ldots, 1)\) and

\[
\zeta_i(I \times L_i) = \bigcup_{(k, r, s) \in \Phi_i} \Gamma_k(r, s) \cup \bigcup_{(t; r_1, \ldots, r_l) \in \Psi_i} \Gamma(t; r_1, \ldots, r_l),
\]

where \(\Phi_i = \{(k, r, s) \mid 1 \leq k \leq r, 2 \leq r \leq i - 1, r + s = i + 1\}\) and \(\Psi_i = \{(t; r_1, \ldots, r_l) \mid 2 \leq t \leq i - 1, r_j \geq 1, r_1 + \cdots + r_t = i\}\). By using the homeomorphism \(\zeta_i\), we identify the complex \(\Gamma_i\) with \(I \times K_i\).

Let \(X\) and \(Y\) be \(A_n\)-spaces, and let \(\phi : X \to Y\) be a map. Then we have the \(A_n\)-forms \(\{M_i : K_i \times X^i \to X\}_{2 \leq i \leq n}\) and \(\{N_i : K_i \times Y^i \to Y\}_{2 \leq i \leq n}\) on \(X\) and \(Y\), respectively. An \(A_n\)-form on the map \(\phi : X \to Y\) is a system of maps \(\{F_i : \Gamma_i \times X^i \to Y\}_{1 \leq i \leq n}\) satisfying the following conditions:

\[
F_i = \phi : X \to Y,
\]

where \(\Gamma_1 \times X = \{\ast\} \times X\) is identified with \(X\).

\[
F_i(\delta_k(r, s)(\rho, \sigma), x_1, \ldots, x_i) = F_i(\rho, x_1, \ldots, x_{k-1}, M_s(\sigma, x_k, \ldots, x_{k+s-1}), x_{k+s}, \ldots, x_i)
\]

for \((\rho, \sigma) \in \Gamma_r \times K_s\).

\[
F_i(\delta(t; r_1, \ldots, r_l)(\tau, \rho_1, \ldots, \rho_l), x_1, \ldots, x_i) = N_l(\tau, F_{r_1}(\rho_1, x_1, \ldots, x_{r_1}), \ldots, F_{r_l}(x_{r_1 + \cdots + r_{l-1} + 1}, \ldots, x_i))
\]

for \((\tau, \rho_1, \ldots, \rho_l) \in K_t \times \Gamma_{r_1} \times \cdots \times \Gamma_{r_l}\).

\[
F_i(\gamma, x_1, \ldots, x_{j-1}, \ast, x_{j+1}, \ldots, x_i) = F_{i-1}(d_j(\gamma), x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_i)
\]

for \(1 \leq j \leq i\).
Figure 3. $A_3$-form on $\phi$.

Figure 4. $A_4$-form on $\phi$.

Figure 3 and Figure 4 denote the $A_3$-form and the $A_4$-form on $\phi$, respectively. A map together with an $A_n$-form is called an $A_n$-map. From the definition, we see that an $A_2$-map and an $A_3$-map are an $H$-map and an $H$-map preserving the homotopy associativity of $H$-spaces, respectively. Furthermore, an $A_\infty$-map is homotopic to a loop map, and we have the induced map between classifying spaces (see [12, §6.4]).

In the proof of Theorem A, we need the following result:

**Theorem 2.1.** Let $A, X \in \mathcal{S}_*$. Then we have the following:

1. If $X$ is an $A_n$-space, then the $A$-nullification $L_A(X)$ is an $A_n$-space and the natural map $\phi_X : X \to L_A(X)$ is an $A_n$-map.
2. If $X$ is an $A_n$-space, then the $A$-colocalization $CW_A(X)$ is an $A_n$-space and the natural map $\psi_X : CW_A(X) \to X$ is an $A_n$-map.
A functor $F : \mathcal{S}_* \to \mathcal{S}_*$ is called continuous if the map
\[
\lambda_F : \text{Map}_*(X,Y) \to \text{Map}_*(F(X),F(Y))
\]
defined by $\lambda_F(g) = F(g)$ is continuous when the compact-open topology are assigned to those mapping spaces. Let $(\mathbb{Z}/p)_\infty$ denote the $p$-completion functor of Bousfield-Kan [2]. It is shown by Iwase [10, Thm. 3.7] that $(\mathbb{Z}/p)_\infty$ is a continuous functor. He used the result to show that $(\mathbb{Z}/p)_\infty$ strictly preserves the higher homotopy associativity of $H$-spaces (see [10, Cor. 3.10]). Since the continuity of $L_A$ and $CW_A$ are proved by Dror Farjoun [5, Thm. 1.3, Thm. 2.B.3], we also use a similar way to the proof of [10, Thm. 3.7, Cor. 3.10] to prove Theorem 2.1. But we need to argue more precisely than [10] since $L_A$ and $CW_A$ are in general homotopy functors. Theorem 2.1 shows that $L_A$ and $CW_A$ preserve the higher homotopy associativity of $H$-spaces up to homotopy.

Kawamoto [14, Thm. 2.14] has shown that the nullification functor $L_A$ preserves the higher homotopy commutativity in the sense of Williams [25] of loop spaces. Since the proof of Theorem 2.1(1) is similar to the one of [14, Thm. 2.14], we give an outline of the proof.

**Proof of Theorem 2.1(1).** Since $X$ is an $A_n$-space, there is a system of maps $\{M_i : K_i \times X^i \to X\}_{2 \leq i \leq n}$ satisfying Conditions (2.2)-(2.4). By using induction on $i$, we construct $A_n$-forms $\{N_i : K_i \times L_A(X)^i \to L_A(X)\}_{2 \leq i \leq n}$ and $\{F_i : \Gamma_i \times X^i \to L_A(X)\}_{1 \leq i \leq n}$ on $L_A(X)$ and $\phi_X$ satisfying Conditions (2.2)-(2.4) and (2.5)-(2.8), which implies the required conclusion.

First we put $F_1 = \phi_X : X \to L_A(X)$, where $\Gamma_1 \times X$ is identified with $X$. Next we identify $K_2 \times L_A(X)^2$ and $\Gamma_2 \times X^2$ with $L_A(X)^2$ and $I \times X^2$, respectively. By [5, 1.A.8 e.4], there is a homotopy equivalence $\gamma : L_A(X)^2 \to L_A(X)$ with $\gamma(\phi_X) \simeq \phi_X^2$. If we define a map $\tilde{N}_2 : L_A(X)^2 \to L_A(X)$ by the composite $\tilde{N}_2 = L_A(M_2)\gamma$, then $\tilde{N}_2(\phi_X)^2 \simeq \phi_X M_2$. Since $\tilde{N}_2 \phi_X \simeq \phi_X$, we have by (2.1) that $\tilde{N}_2 \phi_X \simeq 1_{L_A(X)}$ for $j = 1, 2$, where $\iota_j : L_A(X) \to L_A(X)^2$ denotes the inclusion on the $j$-th factor. From the homotopy extension property, there is a map $N_2 : L_A(X)^2 \to L_A(X)$ with $N_2 \iota_j = 1_{L_A(X)}$ for $j = 1, 2$ and $N_2(\phi_X) \simeq \phi_X M_2$. Here we can choose a homotopy $F_2 : I \times X^2 \to L_A(X)$ such that $F_2(t,x,) = F_2(t,*,x) = \phi_X(x)$, $F_2|[0] \times X = N_2(\phi_X)$ and $F_2|[1] \times X = \phi_X M_2$.

By the inductive hypothesis, we have systems of maps $\{N_j : K_j \times L_A(X)^j \to Y\}_{2 \leq j \leq i-1}$ and $\{F_j : \Gamma_j \times X^j \to L_A(X)\}_{1 \leq j \leq i-1}$ satisfying Conditions (2.2)-(2.4) and (2.5)-(2.8). Now we put that $S_i = I \times (L_i \times X^i \cup K_i \times X^i) \cup \{1\} \times K_i \times X^i$ and $T_i = (\zeta_i \times 1_{X^i})(S_i) \subset \Gamma_i \times X^i$, where $X^i$ denotes the $i$-fold fat wedge of $X$ given by

\[
X^i = \{(x_1, \ldots, x_i) \in X^i \mid x_j = * \text{ for some } 1 \leq j \leq i\}.
\]
Let $E_i : T_i \to L_A(X)$ be the map defined by

$$E_i(\delta_k(r, s)(\rho, \sigma), x_1, \ldots , x_i)$$

$$= F_r(\rho, x_1, \ldots , x_{k-1}, M_s(\sigma, x_k, \ldots , x_{k+s-1}), x_{k+s}, \ldots , x_i),$$

$$E_i(\delta(t; r_1, \ldots , r_t)(\tau, \rho_1, \ldots , \rho_t), x_1, \ldots , x_i)$$

$$= N_t(\tau, F_{r_1}(\rho_1, x_1, \ldots , x_{r_1}), \ldots , F_{r_t}(x_{r_1+\cdots+r_t-1+1}, \ldots , x_i))$$

for $(k, r, s) \in \Phi_i$ and $(t; r_1, \ldots , r_t) \in \Psi_i,$

$$E_i(\gamma, x_1, \ldots , x_{j-1}, *, x_{j+1}, \ldots , x_i)$$

$$= F_{i-1}(d_j(\gamma), x_1, \ldots , x_{j-1}, x_{j+1}, \ldots , x_i)$$

for $\gamma \in \Gamma_i$ and $1 \leq j \leq i,$ and

$$E_i(\zeta_i(1, \tau), x_1, \ldots , x_i) = \phi_X(M_i(\tau, x_1, \ldots , x_i))$$

for $\tau \in K_i.$ From the homotopy extension property, there exists a map $\bar{E}_i : \Gamma_i \times X^i \to L_A(X)$ with $\bar{E}_i|_{T_i} = E_i.$ Let $Q_i : K_i \times X^i \to L_A(X)$ be the map given by $Q_i(\tau, x_1, \ldots , x_i) = \bar{E}_i(0, \tau, x_1, \ldots , x_i).$ By using a similar argument to the proof of [14, Thm. 2.14], we can construct a map $N_i : K_i \times L_A(X)^i \to L_A(X)$ which satisfies (2.2)-(2.4) and $N_i(1_K \times (\phi_X)^i) \simeq Q_i$ rel $L_i \times X^i.$ By using the same argument as the proof of [14, Thm. 2.14] again, we have a map $\tilde{F}_i : \Gamma_i \times X^i \to L_A(X)$ satisfying Conditions (2.5)-(2.8), which implies the required conclusion. This completes the proof. \[ \square \]

In the case of the colocalization functor $CW_A,$ we can prove Theorem 2.1(2) by using a similar argument to the proof of Theorem 2.1(1), and so we omit the proof. In the proof, we need the following result instead of [14, Prop. 2.3]:

**Proposition 2.2.** Let $A, X \in S_\ast,$ and let $\psi_X : CW_A(X) \to X$ denote the natural map. Then we have the following homotopy equivalences:

$$\psi_X^* : \text{Map}_e(CW_A(X)^i, CW_A(X)) \to \text{Map}_e(CW_A(X)^i, X),$$

$$\psi_X^* : \text{Map}_e(CW_A(X)^{i(\ast)}, CW_A(X)) \to \text{Map}_e(CW_A(X)^{i(\ast)}, X)$$

for $i \geq 1,$ where $Z^{(\ast)}$ denotes the $i$-fold smash product of a space $Z.$

From the proof of [5, Thm. 2.1E.1], we have the following lemma:

**Lemma 2.3.** Let $A \in S_\ast.$ If $f : Y \to X$ is an $A$-equivalence and $W$ is an $A$-cellular space, then the induced map

$$f_* : \text{Map}_e(W, Y) \to \text{Map}_e(W, X)$$

is a homotopy equivalence.
Proof of Proposition 2.2. By Lemma 2.3, the induced map

\[(\psi_X)_* : \text{Map}_*(W, CW_A(X)) \to \text{Map}_*(W, X)\]

is a homotopy equivalence for an A-cellular space W. Then it is sufficient to show that CW_A(X)i and CW_A(X)(i) are A-cellular. If B and C are A-cellular, then by [5, Cor. 2.D.17], B × C is A-cellular. Since B ∨ C is represented as a homotopy colimit space, by [5, Def. 2.D.1], B ∨ C is A-cellular, and by using [5, Def. 2.D.1] again, so is B ∧ C. From these facts, we have the required conclusion. This completes the proof. □

Let Sm denote the m-dimensional sphere for m ≥ 1. From the definition, one can see that a space X is Sm-null if and only if \(\pi_i(X) = 0\) for i ≥ m, and a map \(f : X \to Y\) is Sm-equivalence if and only if the induced homomorphism \(f_* : \pi_i(X) \to \pi_i(Y)\) is an isomorphism for i ≥ m (see [5, 1.A.1.1, 2.D.2.6]). Furthermore, it is shown by Dror Farjoun [5, 1.E.1, 2.A.3.1] that the Sm-nullification \(L_{Sm}(X)\) is the \((m-1)\)-th stage \(X_{m-1}\) of the Postnikov system of X, and the Sm-colocalization \(CW_{Sm}(X)\) is the \((m-1)\)-connected cover \(X_{m} \langle m-1 \rangle\) of X. Then as direct consequences of Theorem 2.1, we have the following results:

**Corollary 2.4.** If X is an An-space, then the m-th stage \(X_m\) of the Postnikov system is an An-space and the natural projection \(p_m : X \to X_m\) is an An-map.

**Corollary 2.5.** If X is an An-space, then the m-connected cover \(X \langle m \rangle\) is an An-space and the natural inclusion \(i_m : X \langle m \rangle \to X\) is an An-map.

**Remark 2.6.** By using a result of Stasheff [22, II Cor. 10.6], we have a similar result to Corollary 2.4 (see [22, II Thm. 6.2]). His result implies that we can choose An-forms on X and \(X_m\), so that the projection \(p_m : X \to X_m\) is an An-homomorphism, where An-homomorphism is a map between An-spaces strictly preserving the An-forms (see [22, II Def. 4.1]). Corollary 2.4 has the advantage in that we need not change the given An-form on X.

### 3. Proofs of Theorem A and Theorem B.

In this section, we give the proofs of Theorem A and Theorem B. First we prove Theorem A by combining Theorem 2.1 with the results of Broto and Crespo about H-spaces with finitely generated cohomology. Next we show that the \(BZ/p\)-nullification functor \(L_{BZ/p}\) preserves a quasi Cm-form on an An-space (see Theorem 3.5). By using Theorem A and Theorem 3.5, we prove Theorem B.

Lin [17] posed a question whether a simply connected H-space whose mod p cohomology is finitely generated as an algebra has the same mod p cohomology as a product of \(\mathbb{C}P\infty\)s with three-connected covers of finite
$H$-spaces and finite $H$-spaces (see [17, p. 1105]). Broto and Crespo [3] and [4] answered positively the question as follows:

**Theorem 3.1 ([3, Thm. 1.1], [4, Thm. 1.1]).** Let $p$ be a prime. If $X$ is a simply connected mod $p$ $H$-space such that the mod $p$ cohomology $H^*(X; \mathbb{Z}/p)$ is finitely generated as an algebra, then we have a simply connected mod $p$ finite $H$-space $Y$ and a principal $H$-fibration

\[
F \xrightarrow{\alpha} X \xrightarrow{\beta} Y,
\]

where the fiber $F$ is the direct product of a finite number of $\mathbb{C}P^\infty$s.

In [3] and [4], they also remarked the next fact without the proof (see [3, p. 354]). Since it is an essential point in the proof of Theorem A, we will explain the proof in detail.

**Proposition 3.2.** Suppose that $X$ satisfies the same conditions as Theorem 3.1. Then we have the following homotopy commutative diagram of fibrations:

\[
\begin{array}{ccc}
F & \xrightarrow{\alpha} & X \\
\gamma \downarrow \cong & & \beta \\
\text{CW}_{BZ/p^\infty}(X) & \xrightarrow{\psi_X} & X \xrightarrow{\phi_X} L_{BZ/p}(X),
\end{array}
\]

where the top horizontal sequence is the principal $H$-fibration (3.1).

**Proof.** By a result of Dwyer-Wilkerson [7, Thm. 9.3], we have that $\text{Map}_*(BZ/p^\infty, Y)$ is contractible since $Y$ is a mod $p$ finite $H$-space. Then $\alpha : F \to X$ is a $BZ/p^\infty$-equivalence. By using the universality of the natural map $\psi_X : \text{CW}_{BZ/p^\infty}(X) \to X$, we have a map $\xi : \text{CW}_{BZ/p^\infty}(X) \to F$ with $\alpha \xi \simeq \psi_X$. Since $F$ is $BZ/p^\infty$-cellular, by Lemma 2.3, there exists a map $\gamma : F \to \text{CW}_{BZ/p^\infty}(X)$ with $\psi_X \gamma \simeq \alpha$. By using Lemma 2.3 again, we have a bijection

\[ (\psi_X)_* : [\text{CW}_{BZ/p^\infty}(X), \text{CW}_{BZ/p^\infty}(X)] \to [\text{CW}_{BZ/p^\infty}(X), X], \]

which implies that $\gamma \xi \simeq 1_{\text{CW}_{BZ/p^\infty}(X)}$. Similarly, we have $\xi \gamma \simeq 1_F$ by using a bijection $\alpha_* : [F, F] \to [F, X]$, and thus $\gamma : F \to \text{CW}_{BZ/p^\infty}(X)$ is a homotopy equivalence.

By Miller [19, Thm. A], $Y$ is $BZ/p$-null, and so the natural map $\phi_Y : Y \to L_{BZ/p}(Y)$ is a homotopy equivalence. By a result of Dror Farjoun [5, Cor. 3.D.3], the $BZ/p$-nullification functor $L_{BZ/p}$ preserves the fibration (3.1). Furthermore, by [1, Remark 9.5], $L_{BZ/p}(F)$ is contractible, which implies that $L_{BZ/p}(\beta) : L_{BZ/p}(X) \to L_{BZ/p}(Y)$ is a homotopy equivalence. Let $\delta = L_{BZ/p}(\beta)^{-1} \phi_Y : Y \to L_{BZ/p}(X)$, where $L_{BZ/p}(\beta)^{-1} : L_{BZ/p}(Y) \to L_{BZ/p}(X)$ denotes the homotopy inverse of $L_{BZ/p}(\beta)$. Then $\delta$ is a homotopy
equivalence with $\delta \phi_X \simeq \beta$, and we have the required conclusion. This completes the proof. \hfill \square

Now we prove Theorem A as follows:

**Proof of Theorem A.** Let $X$ be a simply connected $A_n$-space such that the mod $p$ cohomology $H^*(X; \mathbb{Z}/p)$ is finitely generated as an algebra. By Theorem 3.1 and Proposition 3.2, we have the following principal $H$-fibration:

$$CW_{B\mathbb{Z}/p}(X) \xrightarrow{\psi_X} X \xrightarrow{\phi_X} L_{B\mathbb{Z}/p}(X),$$

where $CW_{B\mathbb{Z}/p}(X)$ is homotopy equivalent to a finite product of $CP^\infty$s. From Theorem 2.1, the above fibration is a principal $A_n$-fibration, and we have the required conclusion. This completes the proof of Theorem A. \hfill \square

Next we proceed to the proof of Theorem B. Hemmi [8] introduced the concept of a quasi $C_n$-form on an $A_n$-space. Let $X$ be an $A_n$-space in the sense of Stasheff [22], and let $P_i(X)$ denote the $i$-th projective space of $X$ for $1 \leq i \leq n$. From the construction of $P_i(X)$, we have the following cofiber sequence:

$$P_{i-1}(X) \xrightarrow{\kappa_{i-1}} P_i(X) \xrightarrow{\rho_i} (\Sigma X)^{(i)}$$

for $1 \leq i \leq n$. To describe the definition of a quasi $C_n$-form, let $\kappa_i : (\Sigma X)^{(i-1)} \to (\Sigma X)^{(i)}$ and $\epsilon_i : \Sigma X \to (\Sigma X)^{(i)}$ be the inclusion maps given by

$$\kappa_i(x_1, \ldots, x_{i-1}) = (x_1, \ldots, x_{i-1}, *)$$

and

$$\epsilon_i(x) = (*, \ldots, *, x)$$

for $2 \leq i \leq n$. Let $\zeta_i : (\Sigma X)^{(i)} \to (\Sigma X)^{(i)}$ denote the natural projection for $1 \leq i \leq n$. A quasi $C_n$-form on $X$ is defined by a system of maps $\{\lambda_i : (\Sigma X)^{(i)} \to P_i(X)\}_{1 \leq i \leq n}$ satisfying the following conditions:

(3.2) $\lambda_1 = 1_{\Sigma X} : \Sigma X \to \Sigma X,$

(3.3) $\lambda_i \kappa_i = \iota_{i-1} \lambda_{i-1}$ for $2 \leq i \leq n,$

(3.4) $\lambda_i \epsilon_i = \iota_{i-1} \cdots \iota_1$ for $2 \leq i \leq n,$

(3.5) $\rho_i \lambda_i \simeq \left( \sum_{\sigma \in \Sigma_i} \sigma \right) \zeta_i$ for $1 \leq i \leq n,$

where the action of the symmetric group $\Sigma_i$ on $(\Sigma X)^{(i)}$ is given by the permutation of the coordinates, and the summation on the right hand side is defined by using the natural co-$H$-structure on $(\Sigma X)^{(i)}$. An $A_n$-space which has a quasi $C_n$-form is called a quasi $C_n$-space. Hemmi has shown the following result:
Theorem 3.3 ([8, Thm. 1.1]). If $X$ is a simply connected finite quasi $C_p$-space, then $X$ is contractible.

Remark 3.4. The above definition of a quasi $C_n$-form is slightly weaker than the original definition due to Hemmi. In fact, he defined a quasi $C_n$-form by a system of maps $\{\psi_i : J_i(\Sigma X) \to P_i(X)\}_{1 \leq i \leq n}$ satisfying some conditions (see [8, Def. 2.1]), where $J_i(\Sigma X)$ denotes the $i$-th James reduced product space of $\Sigma X$ for $1 \leq i \leq n$. Given a quasi $C_n$-form $\{\psi_i : J_i(\Sigma X) \to P_i(X)\}_{1 \leq i \leq n}$ in the sense of Hemmi, by composing $\psi_i$ with the natural projection $\pi_i : (\Sigma X)^i \to J_i(\Sigma X)$, we have a system $\{\lambda_i : (\Sigma X)^i \to P_i(X)\}_{1 \leq i \leq n}$ satisfying Conditions (3.2)-(3.5). Hemmi has shown Theorem 3.3 under the assumption that $X$ is a quasi $C_p$-space in his definition. However, one can see from his proof that our definition is also sufficient to prove Theorem 3.3.

Let $X$ and $Y$ be quasi $C_n$-spaces which have the quasi $C_n$-forms $\{\lambda_i^X : (\Sigma X)^i \to P_i(X)\}_{1 \leq i \leq n}$ and $\{\lambda_i^Y : (\Sigma Y)^i \to P_i(Y)\}_{1 \leq i \leq n}$, respectively. If $\phi : X \to Y$ is an $A_n$-map, then by Iwase-Mimura [11, Thm. 3.1], we have the induced map $P_i(\phi) : P_i(X) \to P_i(Y)$ for $1 \leq i \leq n$. Now $\phi : X \to Y$ is called a quasi $C_n$-map if the following diagram is homotopy commutative:

\[
\begin{array}{ccc}
(\Sigma X)^i & \xrightarrow{\lambda_i^X} & P_i(X) \\
(\Sigma \phi)^i \downarrow & & \downarrow P_i(\phi) \\
(\Sigma Y)^i & \xrightarrow{\lambda_i^Y} & P_i(Y)
\end{array}
\]

for $1 \leq i \leq n$.

In the proof of Theorem B, we need the following result:

Theorem 3.5. If $X$ is a simply connected quasi $C_n$-space such that the mod $p$ cohomology $H^*(X; \mathbb{Z}/p)$ is finitely generated as an algebra, then $L_{B\mathbb{Z}/p}(X)$ is a quasi $C_n$-space and $\phi_X : X \to L_{B\mathbb{Z}/p}(X)$ is a quasi $C_n$-map.

Lemma 3.6. Suppose $X$ satisfies the same conditions as Theorem 3.5. Then we have the following:

1. There exists a homotopy equivalence

   $\nu_i : (\Sigma L_{B\mathbb{Z}/p}(X))^i \to L_{B\mathbb{Z}/p}(\Sigma X)^i$

   with $\phi_X(\Sigma X)^i \simeq \nu_i(\Sigma \phi_X)^i$ for $i \geq 1$.

2. Given a $B\mathbb{Z}/p$-null space $Z$, the induced map

   \[
   ((\Sigma \phi_X)^i)^* : \text{Map}_*(((\Sigma L_{B\mathbb{Z}/p}(X))^i, Z) \to \text{Map}_*((\Sigma X)^i, Z)
   \]

   is a homotopy equivalence for $i \geq 1$.

Proof. First we show (1). By Dror Farjoun [5, 1.A.8 e.4], there exists a homotopy equivalence $\gamma_j : L_{B\mathbb{Z}/p}(\Sigma X)^j \times L_{B\mathbb{Z}/p}(\Sigma X) \to L_{B\mathbb{Z}/p}(\Sigma X)^{j+1}$ such that $\gamma_j(\phi_X(\Sigma X)^j \times \phi_X \Sigma X) \simeq \phi_X((\Sigma X)^{j+1})$ for $j \geq 1$. If we define a map
\[ \theta_i : L_{BZ/p}(\Sigma X)^i \to L_{BZ/p}(\Sigma X)^i \text{ by } \theta_i = \gamma_i - 1(\gamma_i - 2 \times 1_{L_{BZ/p}(\Sigma X)}) \cdots (\gamma_1 \times 1_{L_{BZ/p}(\Sigma X)^{i-2}}), \text{ then } \theta_i \text{ is a homotopy equivalence satisfying that } \theta_i(\phi_{\Sigma X})^i \simeq \phi_{\Sigma X}^i \text{ for } i \geq 1. \]

From [14, Lemma 2.6], we have a homotopy equivalence \( \omega : L_{BZ/p}(\Sigma X) \to L_{BZ/p}(L_{BZ/p}(S) \wedge L_{BZ/p}(X)) \) such that
\[
\omega \phi_{\Sigma X} \simeq \phi_{L_{BZ/p}(S) \wedge L_{BZ/p}(X)}(\phi S \wedge \phi X). 
\]
Let \( \omega^{-1} : L_{BZ/p}(L_{BZ/p}(S) \wedge L_{BZ/p}(X)) \to L_{BZ/p}(\Sigma X) \) denote the homotopy inverse of \( \omega \). By Theorem A, \( L_{BZ/p}(X) \simeq Y \) is finite, and so is \( L_{BZ/p}(S) \wedge L_{BZ/p}(X) \simeq \Sigma L_{BZ/p}(X) \). Then \( \phi_{L_{BZ/p}(S) \wedge L_{BZ/p}(X)} \) is a homotopy equivalence. Let \( \zeta : \Sigma L_{BZ/p}(X) \to L_{BZ/p}(\Sigma X) \) be the map defined by \( \zeta = \omega^{-1}\phi_{L_{BZ/p}(S) \wedge L_{BZ/p}(X)}(\phi S \wedge 1_{L_{BZ/p}(X)}) \). Then \( \zeta \) is a homotopy equivalence and \( \zeta(\Sigma \phi_X) \simeq \phi_{\Sigma X} \). If we put \( \nu_i = \theta_i \zeta^i : (L_{BZ/p}(X))^i \to L_{BZ/p}(\Sigma X)^i \), then \( \nu_i \) satisfies the required conditions.

Next we show (2). By taking the mapping spaces, we have the following homotopy commutative diagram:
\[
\begin{array}{ccc}
\text{Map}_*(L_{BZ/p}((\Sigma X)^i), Z) & \xrightarrow{\nu_i^*} & \text{Map}_*((\Sigma L_{BZ/p}(X))^i, Z) \\
\Downarrow \text{ (\phi_{\Sigma X})^i}^* & \simeq & \Downarrow \text{ ((\Sigma \phi_X)^i)^*} \\
\text{Map}_*((\Sigma X)^i, Z) & = & \text{Map}_*((\Sigma X)^i, Z)
\end{array}
\]
for a \( BZ/p \)-null space \( Z \). Since \( \nu_i^* \) and \( (\phi_{\Sigma X})^i \) are homotopy equivalences by (1) and (2.1), so is \( ((\Sigma \phi_X)^i)^i \). This completes the proof. \( \square \)

**Proof of Theorem 3.5.** Let \( X \) be an \( A_n \)-space with finitely generated mod \( p \) cohomology. From Theorem 2.1, the \( BZ/p \)-nullification \( L_{BZ/p}(X) \) is an \( A_n \)-space and the natural map \( \phi_X : X \to L_{BZ/p}(X) \) is an \( A_n \)-map. Then by Iwase-Minura [11, Thm. 3.1], we have the induced map \( P_i(\phi_X) : P_i(X) \to P_i(L_{BZ/p}(X)) \) for \( 1 \leq i \leq n \) satisfying the following homotopy commutative diagram of cofiber sequences:
\[
\begin{array}{ccc}
P_{i-1}(X) & \xrightarrow{\iota_{i-1}} & P_i(X) \\
\downarrow P_i(\phi_X) & & \downarrow (\Sigma \phi_X)^i \\
P_{i-1}(L_{BZ/p}(X)) & \xrightarrow{\tilde{\iota}_{i-1}} & P_i(L_{BZ/p}(X)) \\
\end{array}
\]
(3.6)
where \( \tilde{\iota}_{i-1} : P_{i-1}(L_{BZ/p}(X)) \to P_i(L_{BZ/p}(X)) \) is the natural inclusion and \( \tilde{\rho}_i : P_i(L_{BZ/p}(X)) \to (\Sigma L_{BZ/p}(X))^{(i)} \) denotes the natural projection. Since \( L_{BZ/p}(X) \) is finite, so is \( (\Sigma L_{BZ/p}(X))^{(i)} \) for \( i \geq 1 \). By using induction on \( i \), the projective space \( P_i(L_{BZ/p}(X)) \) is finite, and so \( P_i(L_{BZ/p}(X)) \) is \( BZ/p \)-null. Then there exists a map \( \eta_i : L_{BZ/p}(P_i(X)) \to P_i(L_{BZ/p}(X)) \)
such that \( \eta_i \phi P_i(X) \simeq P_i(\phi_X) \). Now we define a map \( \hat{\lambda}_i : (\Sigma L_{BZ/p}(X))^i \to P_i(L_{BZ/p}(X)) \) by the composite \( \hat{\lambda}_i = \eta_i L_{BZ/p}(\lambda_i) \nu_i \). Then by Lemma 3.6, we see that \( \hat{\lambda}_i(\Sigma \phi_X)^i \simeq P_i(\phi_X)\lambda_i \) for \( 1 \leq i \leq n \). If we show that the system \( \{ \hat{\lambda}_i : (\Sigma L_{BZ/p}(X))^i \to P_i(L_{BZ/p}(X))\}_{1 \leq i \leq n} \) is a quasi \( C_n \)-form on \( L_{BZ/p}(X) \), then the result follows.

First we consider Conditions (3.3) and (3.4). Let \( \tilde{\kappa}_i : (\Sigma L_{BZ/p}(X))^{i-1} \to (\Sigma L_{BZ/p}(X))^i \) be the inclusion given by

\[
\tilde{\kappa}_i(y_1, \ldots, y_{i-1}) = (y_1, \ldots, y_{i-1}, *)
\]

for \( 2 \leq i \leq n \). From the definition of \( \hat{\lambda}_i \), and by using Diagram (3.6), we have that \( (\Sigma \phi_X)^i ([\hat{\lambda}_i \tilde{\kappa}_i]) = [P_i(\phi_X)\lambda_i \kappa_i] \) and \( ((\Sigma \phi_X)^{i-1}) ([\hat{\lambda}_i \tilde{\kappa}_i]) = [P_i(\phi_X)\tilde{\lambda}_{i-1} \lambda_{i-1}] \). Then \( ((\Sigma \phi_X)^i ([\hat{\lambda}_i \tilde{\kappa}_i]) = ((\Sigma \phi_X)^{i-1}) ([\hat{\lambda}_i \tilde{\kappa}_i]) \) since \( \lambda_i \kappa_i = \tilde{\lambda}_{i-1} \lambda_{i-1} \). By Lemma 3.6,

\[
((\Sigma \phi_X)^i) : ((\Sigma L_{BZ/p}(X))^{i-1}, P_{i-1}(L_{BZ/p}(X))) \to ((\Sigma L_{BZ/p}(X))^i, P_{i-1}(L_{BZ/p}(X)))
\]

is a bijection since \( P_{i-1}(L_{BZ/p}(X)) \) is \( BZ/p \)-null, which implies that \( \hat{\lambda}_i \tilde{\kappa}_i \simeq \tilde{\lambda}_{i-1} \hat{\lambda}_{i-1} \) for \( 2 \leq i \leq n \). By using the same argument as above, we have \( \hat{\lambda}_i \tilde{\kappa}_i \simeq \tilde{\lambda}_{i-1} \cdots \tilde{\lambda}_1 \) for \( 2 \leq i \leq n \), where \( \tilde{\epsilon}_i : (\Sigma L_{BZ/p}(X)) \to (\Sigma L_{BZ/p}(X))^i \) denotes the inclusion given by

\[
\tilde{\epsilon}_i(y) = (*, \ldots, *, y)
\]

for \( 2 \leq i \leq n \). From the homotopy extension property, we can choose a map \( \hat{\lambda}_i : (\Sigma L_{BZ/p}(X))^i \to P_i(L_{BZ/p}(X)) \) such that \( \hat{\lambda}_i \simeq \hat{\lambda}_{i-1} \cdots \hat{\lambda}_1 \hat{\lambda}_1 \) for \( 1 \leq i \leq n \) and the system \( \{ \hat{\lambda}_i \}_{1 \leq i \leq n} \) satisfies Conditions (3.3) and (3.4).

Next we show Condition (3.5). Let \( \tilde{\zeta}_i : (\Sigma L_{BZ/p}(X))^i \to (\Sigma L_{BZ/p}(X))^{(i)} \) denote the natural projection for \( 1 \leq i \leq n \). Since \( X \) is a quasi \( C_n \)-space, by using Condition (3.5) and the homotopy commutativity of Diagram (3.6), we have that

\[
((\Sigma \phi_X)^i) \cdot ([p_i \hat{\lambda}_i]) = (\Sigma \phi_X)^{(i)}(\sum_{\sigma \in \Sigma_i} \sigma) \tilde{\zeta}_i.
\]

On the other hand, we see that

\[
((\Sigma \phi_X)^i) \cdot \left( \left( \sum_{\sigma \in \Sigma_i} \sigma \right) \tilde{\zeta}_i \right) = (\Sigma \phi_X)^{(i)} \left( \sum_{\sigma \in \Sigma_i} \sigma \right) \tilde{\zeta}_i.
\]
since \((\Sigma \phi_X)^{(i)} : (\Sigma X)^{(i)} \to (\Sigma L_{BZ/p}(X))^{(i)}\) is a co-\(H\)-map. By applying Lemma 3.6 to a \(BZ/p\)-null space \((\Sigma L_{BZ/p}(X))^{(i)}\), we have a bijection

\[
((\Sigma \phi_X)^{(i)})^* : \left(\left(\Sigma L_{BZ/p}(X)\right)^{(i)}, (\Sigma L_{BZ/p}(X))^{(i)}\right) \to \left(\left(\Sigma X\right)^{(i)}, (\Sigma L_{BZ/p}(X))^{(i)}\right),
\]

which implies that

\[
\tilde{\rho}_i \tilde{\lambda}_i \simeq \left(\sum_{\sigma \in \Sigma_i} \sigma\right) \tilde{\zeta}_i
\]

for \(1 \leq i \leq n\). This completes the proof of Theorem 3.5. \(\square\)

Now we can prove Theorem B as follows:

**Proof of Theorem B.** Let \(X\) be a simply connected quasi \(C_p\)-space such that the mod \(p\) cohomology \(H^*(X; \mathbb{Z}/p)\) is finitely generated as an algebra. From Theorem A, we have a principal \(A_p\)-fibration

\[
F \longrightarrow X \overset{\phi_X}{\longrightarrow} L_{BZ/p}(X),
\]

where the fiber \(F\) is homotopy equivalent to the direct product of a finite number of \(CP^\infty\)s. By Theorem 3.5, the \(BZ/p\)-nullification \(L_{BZ/p}(X)\) is a simply connected finite quasi \(C_p\)-space, and by Theorem 3.3, \(L_{BZ/p}(X)\) is contractible. Hence \(X\) is homotopy equivalent to the fiber \(F\), and we have the required conclusion. This completes the proof of Theorem B. \(\square\)

**References**


Received July 14, 2000 and revised July 3, 2001.

DEPARTMENT OF MATHEMATICS
NATIONAL DEFENSE ACADEMY
1–10–20, HASHIRIMIZU
YOKOSUKA 239–8686, JAPAN
E-mail address: yusuke@cc.nda.ac.jp
DIMENSION GROUPS OF TOPOLOGICAL JOININGS AND NON-COALESCEENCE OF CANTOR MINIMAL SYSTEMS

HIROKI MATUI

When we have two extensions of a Cantor minimal system which are both one-to-one on at least one orbit, we can construct new Cantor minimal systems called topological joinings. We compute the dimension group of the joining in a special case. As an application, we show that a non-invertible endomorphism can induce the identity map on the dimension group of a Cantor minimal system.

1. Introduction.

By a topological dynamical system \((Y, \psi)\), we mean a compact Hausdorff space \(Y\) endowed with a homeomorphism \(\psi\). When \((Y_i, \psi_i), i = 0, 1\) are two topological dynamical systems, \(\psi_0 \times \psi_1\)-invariant closed subsets of \(Y_0 \times Y_1\) are called (topological) joinings, and when \((Y_0, \psi_0)\) equals \((Y_1, \psi_1)\), they are called self-joinings. In the measure-theoretical setting, the notion of self-joinings was introduced by D. Rudolph in [R], and it was proved that the minimal self-joining property implies coalescence and zero entropy. In this paper, we will compute the dimension group of joinings of Cantor minimal systems.

When \(Y\) is the Cantor set and a homeomorphism \(\psi\) on \(Y\) has no nontrivial invariant closed set, \((Y, \psi)\) is called a Cantor minimal system. We define the dimension group \(K^0(Y, \psi)\) of \((Y, \psi)\) as the quotient of \(C(Y, \mathbb{Z})\) by the coboundary subgroup

\[
B_\psi = \{f - f \circ \psi^{-1}; \ f \in C(Y, \mathbb{Z})\}.
\]

In [GPS], it was proved that the dimension group \(K^0(Y, \psi)\), as an ordered group with a distinguished order unit, characterizes the strong orbit equivalence class of \((Y, \psi)\).

We would like to consider the case that a joining \((Z, \tau)\) of Cantor minimal systems \((Y_0, \psi_0)\) and \((Y_1, \psi_1)\) is also a Cantor minimal system. (We must distinguish the property of minimal self-joinings and minimal systems in the joinings.) We don’t know a necessary and sufficient condition so that the joining is minimal. In a special case, however, we will prove that the joining becomes a Cantor minimal system, and compute its dimension group. Our
main result is Theorem 5, in which we will show that the dimension group of the joining is order isomorphic to the relative direct sum of the $K^0(Y_i, \psi_i)$'s.

In the last section, we will consider the non-coalescence of Cantor minimal systems. We denote by $C(\psi)$ the set of continuous maps on $Y$ which commute with $\psi$ and call it the centralizer of $(Y, \psi)$. If $C(\psi)$ consists of homeomorphisms, we say that the system $(Y, \psi)$ is topologically coalescent, and say that $(Y, \psi)$ is non-coalescent, if $C(\psi)$ contains a non-invertible endomorphism. For an element $\gamma \in C(\psi)$, we can define an order homomorphism $\text{mod}(\gamma)$ on $K^0(Y, \psi)$ by $\text{mod}(\gamma)([f]) = [f \circ \gamma]$ for $[f] \in K^0(Y, \psi)$. The map $\text{mod}$ gives a homomorphism from the automorphism group of $(Y, \psi)$ to the automorphism group of the dimension group. Finite subgroups of the kernel of the $\text{mod}$ map were studied by the author in [M2]. In the present paper, we will prove that there exist a lot of non-coalescent Cantor minimal systems whose non-invertible endomorphisms induce automorphisms on the dimension groups. Especially, we construct a non-coalescent Cantor minimal system which is strong orbit equivalent to an odometer system. For detailed information on odometer groups, we refer to [HR].

2. Dimension groups of joinings.

Let $\{(Y_i, \psi_i)\}_{i \in I}$ be a family of topological dynamical systems. For $(y_i)_{i \in I}$ in $Y^I$, the orbit closure of $(y_i)_i$ by the diagonal action is called the joining generated by $(y_i)_i$.

Lemma 1. Let $\{(Y_i, \psi_i)\}_{i \in I}$ and $(X, \phi)$ be topological dynamical systems and $\pi_i : (Y_i, \psi_i) \to (X, \phi)$ be factor maps. If $y = (y_i)_{i \in I}$ satisfies $\pi_i(y_i) = \pi_j(y_j)$ and $\pi_i^{-1}(\pi_i(y_i)) = \{y_i\}$ for all $i, j \in I$, and $(X, \phi)$ is minimal, then the joining generated by $(y_i)_{i \in I} \in Y^I$ is minimal.

Proof. It suffices to prove the case $I = \{0, 1\}$. Let $\psi$ be the diagonal action $\psi_0 \times \psi_1$. For every open neighborhood $O$ of $y = (y_0, y_1)$, there exists an open set $U$ of $X$ such that $V = \pi_0^{-1}(U) \times \pi_1^{-1}(U)$ is a neighborhood of $y$ contained in $O$. To prove the minimality, it is enough to show the almost periodicity of $y$ for $V$, that is,

$$\{k \in \mathbb{Z} : \psi^k(y) \in V\}$$

has a bounded gap in $\mathbb{Z}$. But, we have $\psi^k(y) \in V$ if and only if $\phi^k(\pi_0(y_0)) \in U$, and so the assertion comes from the minimality of $(X, \phi)$. \hfill \Box

When an unperforated ordered group $G$ satisfies the Riesz interpolation property, $G$ is called a dimension group in the abstract sense ([GPS, Section 1]). Let $\pi : H \to G$ be an order homomorphism between dimension groups. We say that $\pi$ is an order embedding, if $\pi$ is injective and $\pi(h) \geq 0$ if and only if $h \geq 0$. 
Definition 2. Let \( \pi_i : H \to G_i, \ i = 0, 1 \), be order embeddings between simple dimension groups, none of them equal to \( \mathbb{Z} \). Assuming \( G_i/\pi_i(H) \) is torsion-free for \( i = 0, 1 \), we call the quotient \( D \) of \( G_0 \oplus G_1 \) by
\[
\{(\pi_0(h), -\pi_1(h)); \ h \in H\}
\]
the relative direct sum of \( G_i \) with respect to \( \pi_i(H) \). Define the positive cone \( D^+ \) such as
\[
D^+ = \{[g_0, g_1] \in D; \ g_i \in G_i^+, \ i = 0, 1\},
\]
where \([\cdot, \cdot]\) means the quotient map. When \( \pi_0 \) and \( \pi_1 \) preserve the distinguished order units, the order unit of \( (D, D^+) \) can be defined in the obvious way.

We can define the relative direct sum of more than two dimension groups in a similar fashion.

Lemma 3. Let \( (D, D^+) \) be the relative direct sum of \( G_i \) with respect to \( \pi_i(H) \). Then, \( (D, D^+) \) is an unperforated ordered group, and for \( x \in D^+ \setminus \{0\} \) and \( y \in D^+ \), there exists a natural number \( n \in \mathbb{N} \) such that \( 0 \leq y \leq nx \).

Proof. It is easy to see that \( (D, D^+) \) is an ordered group. Suppose \( n[g_0, g_1] \geq 0 \) for \([g_0, g_1] \in D \) and \( n \in \mathbb{N} \). By the definition, there exists an element \( h \in H \) such that \( ng_0 + \pi_0(h) \geq 0 \) and \( ng_1 - \pi_1(h) \geq 0 \). If \( ng_0 + \pi_0(h) \) is equal to zero, \( h \) is divisible by \( n \) in \( H \), because \( G_0/\pi_0(H) \) is torsion-free. Hence, we get \([g_0, g_1] \geq 0 \) in \( D \). We have the same conclusion, if \( ng_1 - \pi_1(h) \) is zero. Therefore, we may assume the strict inequality. Since \( H \) is a simple dimension group \((\neq \mathbb{Z})\), we can find \( \epsilon_i \in H \) for \( i = 0, 1 \) such that \( 0 < \pi_0(\epsilon_0) < ng_0 + \pi_0(h) \) and \( 0 < \pi_1(\epsilon_1) < ng_1 - \pi_1(h) \), by Corollary 4.10 of [GH]. By using 4.10 of [GH] again, we get an element \( h' \) such that
\[
h - \epsilon_0 < nh' < h + \epsilon_1.
\]
Then, \( g_0 + \pi_0(h') \) and \( g_1 - \pi_1(h') \) are positive, which implies \([g_0, g_1]\) are positive in \( D \).

To prove the second statement, let \( x = [g_0, g_1] \) be in \( D^+ \setminus \{0\} \). We may assume \( g_0 > 0 \) and \( g_1 \geq 0 \). Because we can choose \( h \in H^+ \setminus \{0\} \) such that \( g_0 - \pi_0(h) > 0 \), the assertion has been proved. \( \square \)

The relative direct sum \( D \) may not satisfy the Riesz interpolation property. For example, let \( G \) be \( \mathbb{Q}^2 \) and define the positive cone by
\[
G^+ = \{(x, y); \ x > |y|\} \cup \{(0, 0)\}.
\]
If \( H = \{(x, 0)\} \cong \mathbb{Q} \) is a subgroup of \( G \), it can be seen that the relative direct sum of two copies of \( G \) with respect to \( H \) doesn’t satisfy the Riesz interpolation property.

We say \( \pi(H) \) is order dense in \( G \), if there is \( h \in H \) for every \( g < g' \) such that \( g < \pi(h) < g' \). By [GPS2, Proposition 1.1], \( \pi(H) \) is order dense in \( G \),
Lemma 4. The relative direct sum \((D, D^+)\) of \(G_i\) with respect to \(\pi_i(H)\) satisfies the Riesz interpolation property, if \(\pi_1(H)\) is order dense in \(G_1\).

Proof. We check the Riesz decomposition property. Assume the inequality
\[
0 \leq [a, a'] < [b, b'] + [c, c'] \quad \text{in } D \quad \text{for } a, b, c \in G_0^+ \quad \text{and } a', b', c' \in G_1^+.
\]
We must show that there exist \(x, y \in D^+\) so that \([a, a'] = x + y\), \(x \leq [b, b']\), \(y \leq [c, c']\). We may also assume \([b, b'] \neq 0\) and \([c, c'] \neq 0\). Because \([b + c, b' + c']\) is strictly larger than \([a, a']\), we can find \(v, \epsilon \in H\) such that
\[
0 \leq (b + c) - a + \pi_0(v), \quad 0 < 2\pi_1(\epsilon) < (b' + c') - a' - \pi_1(v).
\]
If \(\pi_1(H)\) is order dense in \(G_1\), we also get \(t, u \in H^+\) satisfying,
\[
0 \leq b + \pi_0(t), \quad 0 < b' - \pi_1(t) < \pi_1(\epsilon)
\]
and
\[
0 \leq c + \pi_0(u), \quad 0 < c' - \pi_1(u) < \pi_1(\epsilon).
\]
Take \(\epsilon' \in H\) such as \(0 < \pi_1(\epsilon') < b' - \pi_1(t)\) and \(0 < \pi_1(\epsilon') < c' - \pi_1(u)\). By using the order density again, we have \(s \in H^+\) satisfying
\[
0 \leq a + \pi_0(s), \quad 0 < a' - \pi_1(s) < \pi_1(\epsilon').
\]
Then, we obtain
\[
\pi_1(\epsilon') < (b' + c') - a' - \pi_1(t + u - s) < 2\pi_1(\epsilon),
\]
and so \(t + u - s\) is greater than \(v\). If we replace \([a, a'], [b, b']\) and \([c, c']\) with \([a + \pi_0(s), a' - \pi_1(s)], [b + \pi_0(t), b' - \pi_1(t)]\) and \([c + \pi_0(u), c' - \pi_1(u)]\), the assertion follows by the Riesz decomposition property of \(G_i\).

We would like to state the main theorem. Let \(\pi_i : (Y_i, \psi_i) \to (X, \phi)\) be factor maps between Cantor minimal systems for \(i = 0, 1\). Define
\[
E_i = \{x \in X; \#\pi_i^{-1}(x) \neq 1\}
\]
for each \(i\) and assume that \(y_{i, \text{max}} \in Y_i\) and \(x_{\text{max}} \in X\) satisfy
\[
\pi_0(y_{0, \text{max}}) = \pi_1(y_{1, \text{max}}) = x_{\text{max}} \in (E_0 \cup E_1)^c.
\]
By Lemma 1, the joining \((Z, \tau)\) generated by \((y_{0, \text{max}}, y_{1, \text{max}})\) is a Cantor minimal system. In general, a factor map \(\pi_i\) between Cantor minimal systems induces the order embedding \(\pi_i^* : K^0(X, \phi) \to K^0(Y_i, \psi_i)\). In Lemma 6, we will prove that the quotient of \(K^0(Y_i, \psi_i)\) by \(\pi_i^*(K^0(X, \phi))\) is torsion-free. We denote by \(p_i\) the canonical factor maps from \((Z, \tau)\) to \((Y_i, \psi_i)\) for \(i = 0, 1\). The direct sum of \(p_0^*\) and \(p_1^*\) induces the homomorphism from \(K^0(Y_0, \psi_0) \oplus K^0(Y_1, \psi_1)\) to \(K^0(Z, \tau)\).
Theorem 5. In the above setting, we further assume that $E_0$ and $E_1$ are disjoint. Then, the dimension group $K^0(Z, \tau)$ is isomorphic to the relative direct sum of $K^0(Y_i, \psi_i)$ with respect to $\pi^*_i(K^0(X, \phi))$ as a unital ordered group. Moreover, the state space of $K^0(Z, \tau)$ is equal to

$$\{ (\mu_0, \mu_1); \quad \mu_i \in S(K^0(Y_i, \psi_i)), \quad \mu_0 \circ \pi^*_0 = \mu_1 \circ \pi^*_1 \},$$

where $S(K^0(Y_i, \psi_i))$ is the state space of $K^0(Y_i, \psi_i)$.

If $E_0$ and $E_1$ are disjoint, the topological joining $Z$ coincides with the pull-back associated with $\pi$ where $\pi$ is the pull-back associated with $\pi_1$. Therefore, the theorem asserts that the dimension group of the pull-back system is isomorphic to the push-out of the dimension group.

For an ergodic measure $\mu$ on $(X, \phi)$, either of $\mu(E_0)$ and $\mu(E_1)$ must be zero when $E_0$ and $E_1$ are disjoint. If $\mu(E_0) = 0$ for every ergodic measure $\mu$, then $\pi^*_1(K^0(X, \phi))$ is order dense in $K^0(Y_0, \psi_0)$. But, we must remark that even if $E_0$ and $E_1$ are disjoint, $\pi^*_i(K^0(X, \phi))$ may not be order dense in $K^0(Y_i, \psi_i)$ for each $i = 0, 1$. See the example at the last of this section.

In order to prove the above theorem, at first, we need to fix Kakutani-Rohlin partitions for $(X, \phi)$ and $(Y_i, \psi_i)$.

Let $\{X_n\}_{n \in \mathbb{N}}$ be a decreasing sequence of clopen sets of $X$ such that the intersection of all $X_n$’s is $\{x_{\text{max}}\}$. By Theorem 4.2 of [HPS], for each $X_n$, we can find a finite set $V_n$, a map $h : V_n \to \mathbb{N}$ and a clopen partition

$$P_n = \{ X(n, v, k); \quad v \in V_n, \quad 1 \leq k \leq h(v) \}$$

of $X$, which satisfy the following properties:

- For every $v \in V_n$ and $k \neq h(v)$, we have $\phi(X(n, v, k)) = X(n, v, k + 1)$.
- The clopen set $X_n$ equals the disjoint union of $X(n, v, h(v))$ for all $v \in V_n$.
- The partition $P_{n+1}$ is finer than $P_n$ for every $n \in \mathbb{N}$, and the family of partitions $\{P_n\}_n$ generates the topology of $X$.

The clopen set $X_n$ is called the top set and the map $h$ is called the height function. In order to compute the dimension group $K^0(X, \phi)$, we must consider the free abelian group $\mathbb{Z}^{V_n}$. For convenience we denote the canonical basis of $\mathbb{Z}^{V_n}$ by the vertices $v_1, v_2, \cdots, v_j \in V_n$. When we define the rectangular matrix $A_n \in M_{V_n \times V_{n+1}}(\mathbb{Z})$ by

$$A_n(v, v') = \# \{ 1 \leq k \leq h(v'); \quad X(n + 1, v', k) \subset X(n, v, h(v)) \}$$

for every $v \in V_n$ and $v' \in V_{n+1}$, the dimension group $K^0(X, \phi)$ is computed by the inductive limit of $(\mathbb{Z}^{V_n})_n$ with the connecting maps $(A_n)_n$.

The partition $P_n$ induces the clopen partition of $Y_i$ for each $i = 0, 1$ and $n \in \mathbb{N}$. This family of partitions, however, does not generate the topology of $Y_i$. When we divide every clopen set $\pi_i^{-1}(X(n, v, k))$ so that the topology
is generated, we obtain a finite set $W_{i,n}$, a map $\rho_i : W_{i,n} \to V_n$ and a clopen partition

$$Q_{i,n} = \{ Y_i(n, w, k); w \in W_{i,n}, 1 \leq k \leq h(\rho_i(w)) \}$$

of $Y_i$ and the following holds for each $i = 0, 1$:

- For every $w \in W_{i,n}$ and $k \neq h(\rho_i(w))$, we have $\psi_i( Y_i(n, w, k)) = Y_i(n, w, k + 1)$.
- For every $v \in V_n$ and $1 \leq k \leq h(v)$, $\pi_i^{-1}(X(n, v, k))$ is the disjoint union of $Y_i(n, w, k)$ for all $w \in \rho_i^{-1}(v)$.
- As in the case of $(X, \phi)$, $Y_{i,n} = \pi_i^{-1}(X_n)$ is the disjoint union of $Y_i(n, w, h(\rho_i(w)))$ for all $w \in W_{i,n}$.
- The partition $Q_{i,n+1}$ is finer than $Q_{i,n}$ and the family of partitions $\{ Q_{i,n} \}$ generates the topology of $Y_i$.

Notice that the intersection of all $Y_{i,n}$'s equals $\{ y_{i,\text{max}} \}$ because $x_{\text{max}}$ is not contained in $E_i$. As above we denote the canonical basis of $Z^{W_{i,n}}$ by the vertices $w$ in $W_{i,n}$. We can compute the dimension group $K^0(Y_i, \psi_i)$ by the rectangular matrix $B_{i,n} \in M_{W_{i,n} \times W_{i,n+1}}(Z)$ determined in a similar fashion to the case of $A_n$. The homomorphisms sending $v \in Z^{W_{i,n}}$ to $\sum_{w \in \rho_i^{-1}(v)} w \in Z^{W_{i,n}}$ induce the order embedding $\pi^*: K^0(X, \phi) \to K^0(Y_i, \psi_i)$.

**Lemma 6.** Let $\pi : (Y, \psi) \to (X, \phi)$ be a factor map between Cantor minimal systems. If there is a point $y \in Y$ such that $\pi^{-1}(\pi(y)) = \{ y \}$, then the quotient $K^0(Y, \psi)/\pi^*(K^0(X, \phi))$ is torsion-free.

**Proof.** We will prove the lemma for $\pi_0 : (Y_0, \psi_0) \to (X, \phi)$ above. The embedding of $K^0(X, \phi)$ to $K^0(Y_0, \psi_0)$ is obtained by the natural injection from $Z^{W_{0,n}}$ to $Z^{W_{0,n}}$ induced by $\rho_0$. Since clearly $Z^{W_{0,n}}/Z^{W_{0,n}}$ is torsion-free, we get the conclusion. \(\square\)

We return to the setting of Theorem 5. We would like to describe Kakutani-Rohlin partitions of $(Z, \tau)$. When $w_0 \in W_{0,n}$ and $w_1 \in W_{1,n}$ are preimages of the same $v \in V_n$, we set

$$Z(n, w_0, w_1, k) = Z \cap (Y_0(n, w_0, k) \times Y_1(n, w_1, k))$$

for all $1 \leq k \leq h(v)$. Let $W_n$ be the set of pairs $(w_0, w_1)$ such that the clopen set $Z(n, w_0, w_1, h(v))$ is not empty. Then, we get the partition

$$\mathcal{R}_n = \{ Z(n, w_0, w_1, k); (w_0, w_1) \in W_n, 1 \leq k \leq h(\rho_i(w_i)) \}$$

of $Z$. We can check that the sequence of clopen partitions $\{ \mathcal{R}_n \}$ satisfies the similar properties as in the case of $\{ \mathcal{P}_n \}$ or $\{ \mathcal{Q}_{i,n} \}$.$\{ \mathcal{Q}_{i,n} \}$. Hence, we can compute the dimension group $K^0(Z, \tau)$ by the inductive limit of $(Z^{W_{i,n}})$. We denote the canonical basis of $Z^{W_{i,n}}$ by the vertices $(w_0, w_1)$ in $W_n$. The connecting map from $Z^{W_{i,n}}$ to $Z^{W_{i,n+1}}$ is given by a matrix $C_n$ and the entry
of \( C_n \) corresponding to \((w_0, w_1) \in W_n \) and \((w'_0, w'_1) \in W_{n+1}\) is equal to the cardinality of the set
\[
\{1 \leq k \leq h(\rho_i(w'_i)); \ Z(n, w'_0, w'_1, k) \subset Z(n, w_0, w_1, h(\rho_i(w_i)))\}.
\]
The order embedding \( p_0^* \) is induced from the homomorphism sending \( w_0 \in Z^{W_{0,n}} \) to \( \sum_{(w_0,w)\in W_n}(w_0,w) \in Z^{W_n} \). The order embedding \( p_1^* \) is described in the same way.

Define
\[
E_i(m, Q) = \{ P \in P_m; \ \pi_i^{-1}(P) \cap Q \text{ and } \pi^{-1}(P) \cap Q^c \text{ are both nonempty}\}
\]
for \( Q \in Q_{i,n}, n \leq m \) and \( i = 0, 1 \). If \( P' \in P_{m+1} \) is a subset of \( P \in P_m \) and \( P' \in E_i(m+1, Q) \), then we have \( P \in E_i(m, Q) \).

**Lemma 7.** For every \( Q_0 \in Q_{0,n} \) and \( Q_1 \in Q_{1,n} \), there exists a natural number \( m \) greater than \( n \) such that \( E_0(m, Q_0) \) and \( E_1(m, Q_1) \) are disjoint.

**Proof.** The proof is by contradiction. Let \( U_{i,m} \) be the disjoint union of \( P \in E_i(m, Q_i) \), for every \( m \geq n \) and \( i = 0, 1 \). The clopen set \( U_m = U_{0,m} \cap U_{1,m} \) is nonempty and forms a decreasing sequence. Therefore, there exists a point \( x \in X \) contained in all \( U_m \). Then, \( \pi^{-1}(x) \) contains at least two distinct points for each \( i = 0, 1 \), which contradicts the disjointness of \( E_0 \) and \( E_1 \). \( \square \)

We are now ready to prove the main theorem.

**Proof of Theorem 5.** By virtue of Lemma 7, we may assume that \( E_0(n+1, Q_0) \) and \( E_1(n+1, Q_1) \) are disjoint for all \( Q_i \in Q_{i,n} \) and \( n \in \mathbb{N} \). We denote the factor map \( \pi_0 \circ p_0 = \pi_1 \circ p_1 \) by \( \pi \).

At first, we prove the surjectivity of the map \( p_0^* \oplus p_1^* \) from \( K^0(Y_0, \psi_0) \oplus K^0(Y_1, \psi_1) \) to \( K^0(Z, \tau) \). Fix \( n \in \mathbb{N} \) and \((w_0, w_1) \in W_n \) arbitrarily. Let \( v = \rho_0(w_0) = \rho_1(w_1) \in V_n \). Set
\[
S = \{ R \in R_{n+1}; R \cap Z(n, w_0, w_1, h(v)) \text{ is not empty}\},
\]
\[
S_i = \{ R \in S; R \subset \pi^{-1}(P) \text{ for some } P \in E_i(n+1, Y_i(n, w_i, h(v)))\}
\]
for each \( i = 0, 1 \), and
\[
S_2 = \{ R \in S; R \cap \pi^{-1}(P) \text{ is empty for all } P \in E_i(n+1, Y_i(n, w_i, h(v)))\}, \ i = 0, 1\}
\]
Then, we have
\[
Z(n, w_0, w_1, h(v)) = \bigcup_{R \in S} R = \bigcup_{R \in S_0} R \cup \bigcup_{R \in S_1} R \cup \bigcup_{R \in S_2} R.
\]
The first summand is equal to
\[ \bigcup Z \cap (Y_0(n + 1, w'_0, k) \times Y_1), \]
where the union runs over \( w'_0 \) and \( k \) satisfying \( Z(n + 1, w'_0, w'_1, k) \in \mathcal{S}_0 \) for some \( w'_1 \in W_{1,n+1} \). The second summand can be written in the same way.

The last summand is equal to \( \pi^{-1}(U) \) for some clopen set \( U \) of \( X \). Hence, the surjectivity has been proved.

Let us consider the kernel of \( p_0^* \oplus p_1^* \). Suppose \( a_0 = \sum_{w_0 \in W_{0,n}} \lambda_{0,w_0} w_0 \) and \( a_1 = \sum_{w_1 \in W_{1,n}} \lambda_{1,w_1} w_1 \) satisfy
\[ p_0^*(a_0) = p_1^*(a_1) \]
in \( K^0(Z, \tau) \). We may assume that
\[ \sum_{(w_0,w_1) \in W_n} \lambda_{0,w_0}(w_0, w_1) = \sum_{(w_0,w_1) \in W_n} \lambda_{1,w_1}(w_0, w_1), \]
and so we have \( \lambda_{0,w_0} = \lambda_{1,w_1} \), if \( (w_0, w_1) \) exists in \( W_n \). Therefore, we may further assume that there exists a subset \( F_i \subset W_{i,n} \) for each \( i = 0, 1 \), such that:

- If \( w_0 \in F_0 \) and \( (w_0, w_1) \in W_n \), then \( w_1 \in F_1 \).
- If \( w_1 \in F_1 \) and \( (w_0, w_1) \in W_n \), then \( w_0 \in F_0 \).
- The element \( a_i \) is equal to \( \sum_{w_i \in F_i} (w_0, w_1) \) for each \( i = 0, 1 \).

Define the clopen set
\[ W = \bigcup_{w_0 \in F_0} Z(n, w_0, w_1, h(\rho_i(w_i))) = \bigcup_{w_1 \in F_1} Z(n, w_0, w_1, h(\rho_i(w_i))). \]
Let \( Z(n + 1, w'_0, w'_1, k) \in \mathcal{R}_{n+1} \) be a clopen set which is contained in \( W \). Then, there exist \( w_0 \in F_0 \) and \( w_1 \in F_1 \) such that \( Z(n + 1, w'_0, w'_1, k) \) is contained in \( Z(n, w_0, w_1, h(\rho_i(w_i))) \). Without loss of generality, we may assume that \( X(n + 1, \rho_0(w'_0), k) \) is not contained in \( E_0(n + 1, Y_0(n, w_0, h(\rho_0(w_0)))) \). Take a pair \( (w'_0, w'_1) \in W_{n+1} \) satisfying \( \rho_i(w'_j) = \rho_i(w'_j) \). There exists a pair \( (\tilde{w}_0, \tilde{w}_1) \in W_n \) satisfying \( Z(n + 1, w'_0, w'_1, k) \subset Z(n, \tilde{w}_0, \tilde{w}_1, h(\rho_i(\tilde{w}_i))). \)

By the assumption, we have \( \tilde{w}_0 = w_0 \), and so \( \tilde{w}_1 \) must be in \( F_1 \). Hence, \( \pi^{-1}(X(n + 1, \rho_0(w'_0), k)) \) is contained in \( W \). By repeating this argument, we see that \( p_0^*(a_0) = p_1^*(a_1) \) comes from \( K^0(X, \phi) \).

We have proved that \( p_0^* \oplus p_1^* \) induces an isomorphism from the relative direct sum to \( K^0(Z, \tau) \). It is obvious from the above argument that the isomorphism preserves the positive cones and the distinguished order units, and so the proof is completed.

**Corollary 8.** Let \( \{(Y_i, \psi_i)\}_{i \in I} \) be a family of at most countable Cantor minimal systems and \( \pi_i : (Y_i, \psi_i) \to (X, \phi) \) be factor maps to a Cantor minimal system \( (X, \phi) \). Define
\[ E_i = \{ x \in X; \#\pi_i^{-1}(x) \neq 1 \} \]
and assume \( y_i \in Y_i \) and \( x \in X \) satisfies \( \pi_i(y_i) = x \) and \( x \in E_i^c \) for all \( i \in I \). When we further assume that \( E_i \)'s are disjoint for all \( i \), the dimension group of the joining \((Z, \tau)\) generated by \((y_i)\), is isomorphic to the relative direct sum of \( K^0(Y_i, \psi_i)'s \) with respect to \( \pi_i^*(K^0(X, \phi)) \).

**Proof.** Use Theorem 5 repeatedly. \(\square\)

**Example.** Let \( \{q_n\}_{n \in \mathbb{N}} \) be a sequence of natural numbers satisfying

\[
\alpha = \lim_{n \to \infty} \prod_{k=1}^{n} \frac{q_k + 1}{q_k - 1} < \infty.
\]

For convenience, we assume that, for every \( p \in \mathbb{N} \), there exists \( n \in \mathbb{N} \) such that \( \prod_{k=1}^{n} (q_k + 1), \prod_{k=1}^{n} (q_k - 1) \) and \( \prod_{k=1}^{n} (q_k - 2) \) are divisible by \( p \).

We would like to define a simple ordered Bratteli diagram \((V, E, \leq)\) ([HPS]). Let the vertex set \( V_n \) be \( \{v_n, v'_n\} \) and define the edge set \( E_n \) so that the connecting matrix from \( \mathbb{Z}^V \) to \( \mathbb{Z}^{V_{n+1}} \) equals

\[
A_n = \begin{bmatrix} q_n & 1 \\ 1 & q_n \end{bmatrix}.
\]

Let the source vertex of the initial edge in \( r^{-1}(v_{n+1}) \) and \( r^{-1}(v'_{n+1}) \) be \( v_n \), and the source vertex of the final edge be \( v'_n \). Then, we obtain a simple ordered Bratteli diagram. Let \((X, \phi)\) be the Bratteli-Vershik system of \((V, E, \leq)\). The dimension group \( K^0(X, \phi) \) is isomorphic to \( \mathbb{Q}^2 \) and the positive cone is

\[
K^0(X, \phi)^+ = \{(x, y) ; \alpha x > |y|\} \cup \{(0, 0)\},
\]

and so there are two ergodic measures. The distinguished order unit \([1]\) is equal to \((2, 0)\).

Let us construct two extensions of \((X, \phi)\), namely \((Y_0, \psi_0)\) and \((Y_1, \psi_1)\). We would like to define two simple ordered Bratteli diagrams \((W_0, F_0, \leq)\) and \((W_1, F_1, \leq)\). Set the vertex set \( W_{0,n} \) be \( \{w_{0,n}, w'_{0,n}, w''_{0,n}\} \) and the vertex set \( W_{1,n} \) be \( \{w_{1,n}, w'_{1,n}, w''_{1,n}\} \). To describe the partial order on the edge set, for each vertex \( w \), we denote by \( \theta(w) \) the ordered list of the source vertices of edges in \( r^{-1}(w) \). Define the partial order on \( F_0 \) by

\[
\theta(w_{0,n+1}) = \left( w_{0,n}, w_{0,n}, \cdots, w_{0,n}, w'_{0,n}, w''_{0,n} \right)
\]

and

\[
\theta(w'_{0,n+1}) = \left( w_{0,n}, w'_{0,n}, \cdots, w'_{0,n}, w''_{0,n}, w'_{0,n} \right)
\]

and

\[
\theta(w''_{0,n+1}) = \left( w_{0,n}, w''_{0,n}, \cdots, w''_{0,n}, w''_{0,n}, w'_{0,n} \right)
\]
Similarly, define the partial order on $F_1$ by

$$\theta(w_{1,n+1}) = \left(w_{1,n}, w_{1,n}, \ldots, w_{1,n}, w'_{1,n}\right),$$

$$\theta(w'_{1,n+1}) = \left(w_{1,n}, w'_{1,n}, \ldots, w'_{1,n}, w''_{1,n}\right),$$

and

$$\theta(w''_{1,n+1}) = \left(w_{1,n}, w''_{1,n}, \ldots, w''_{1,n}, w'_{1,n}\right),$$

for each $n \in \mathbb{N}$. Then, we get two Cantor minimal systems $(Y_0, \psi_0)$ and $(Y_1, \psi_1)$ associated with simple ordered Bratteli diagrams $(W_0, F_0, \leq)$ and $(W_1, F_1, \leq)$. There exists a factor map $\pi_0 : (Y_0, \psi_0) \to (X, \phi)$ which sends the tower $w_{0,n}$ and $w''_{0,n}$ to $v_n$, and the tower $w'_{0,n}$ to $v'_n$. In the same way, there exists a factor map $\pi_1 : (Y_1, \psi_1) \to (X, \phi)$. For $i = 0, 1$, we set

$$E_i = \{x \in X; \#\pi_i^{-1}(x) \neq 1\}.$$

It is not hard to see that $E_0$ and $E_1$ are disjoint.

Let $\beta$ be the limit of $\prod_{k=1}^n (q_k - 2)/(q_k - 1)$. The dimension group $K^0(Y_i, \psi_i)$ is isomorphic to $\mathbb{Q}^3$ and the positive cone is

$$K^0(Y_i, \psi_i)^+ = \left\{(a, b, c); \begin{array}{l}
aa - b - \beta c > 0 \\
\aa + b + 2\beta c > 0 \\
\aa + b - \beta c > 0
\end{array}\right\} \cup \{(0, 0, 0)\},$$

for both $i = 0, 1$, hence, $(Y_i, \psi_i)$ has three ergodic measures. The order embeddings $\pi_0^*$ and $\pi_1^*$ send $(x, y) \in \mathbb{Q}^3$ to $(x, y, 0)$ and $(x, -y, 0)$, respectively.

Choose $x \in X \setminus (E_0 \cup E_1)$ and $y_i \in Y_i$ such that $\pi_i(y_i) = x$. The topological joining $(Z, \tau)$ generated by $(y_0, y_1)$ is a Cantor minimal system. The dimension group $K^0(Z, \tau)$ is isomorphic to the relative direct sum of $K^0(Y_0, \psi_0)$ and $K^0(Y_1, \psi_1)$ with respect to $K^0(X, \phi)$. Therefore, $K^0(Z, \tau)$ equals $\mathbb{Q}^4$ and the positive cone is determined by four linear inequalities. In this case, $K^0(Z, \tau)$ satisfies the Riesz interpolation property, although neither $\pi_0(K^0(X, \phi))$ nor $\pi_1(K^0(X, \phi))$ is order dense.

3. Non-coalescence and dimension groups.

Let $\alpha$ and $\beta$ be two irrational numbers which are linearly independent over $\mathbb{Q}$. By cutting the circle $S^1 \cong \mathbb{R}/\mathbb{Z}$ at the points $n\alpha - m\beta$ for $n \in \mathbb{Z}$ and $m \in \mathbb{N} \cup \{0\}$, a Cantor set $Y$ is obtained. Then, the $\alpha$-rotation on $Y$ gives a minimal homeomorphism $\psi$, and the system $(Y, \psi)$ is called the Denjoy system (PSS). It is easy to see that the $\beta$-rotation induces a non-invertible endomorphism $\gamma$ on $(Y, \psi)$. The dimension group $K^0(Y, \psi)$ is isomorphic to the countable direct sum of $\mathbb{Z}$, and $\text{mod}(\gamma)$ is an endomorphism on it.
Lemma 9. Let \((X, \phi)\) be a Cantor minimal system, \(\gamma \in C(\phi)\) be an automorphism which has essentially infinite order and \(x_{\text{max}} \in X\) be a point. Then, there exist a Cantor minimal system \((Y, \psi)\) and a factor map \(\pi : (Y, \psi) \to (X, \phi)\) such that the following properties are satisfied:

(i) \(\pi^*\) is an order isomorphism between \(K^0(X, \phi)\) and \(K^0(Y, \psi)\).

(ii) When we set \(E = \{ x \in X ; \# \pi^{-1}(x) \neq 1 \}, \gamma^i(x_{\text{max}})\) and \(\gamma^j(E)\) are all disjoint for \(i, j \in \mathbb{Z}\).

Proof. We may assume that \(X\) is the infinite path space of an ordered Bratteli diagram \((V, E, \leq)\), \(x_{\text{max}}\) is the unique maximal path and \(\phi\) is the Bratteli-Vershik map. Let \(x_{\text{min}}\) be the unique minimal path. We denote the source map by \(s : E_n \to V_n\) and the range map by \(r : E_n \to V_{n+1}\). The symbol \(E_{n,m}\) means the set of finite paths from \(V_n\) to \(V_m\) and \([e]\) for \(e \in E_{n,m}\) means the corresponding clopen set of \(X\). Let \(A_n\) be the incidence matrix from \(Z^{V_n}\) to \(Z^{V_{n+1}}\), and \(h(v)\) be the number of edges connecting \(v_0 \in V_0\) to \(v \in V_n\).

By telescoping the diagram, we would like to choose \(v_n \in V_n\) and \(f_n, f'_n \in E_n\) for each \(n \in \mathbb{N}\) which satisfy the following:

- The source and range vertex of \(f_n\) and \(f'_n\) are \(v_n\) and \(v_{n+1}\), respectively.
- The clopen sets \(\gamma^j([f_n])\) and \(\gamma^k([f'_n])\) for \(|j|, |k| \leq n\) are all disjoint.
- The clopen set \([f_n] \cup [f'_n]\) doesn’t contain \(\gamma^k(x_{\text{max}})\) and \(\gamma^k(x_{\text{min}})\) for \(|k| \leq n\).

The construction of \(v_n, f_n\) and \(f'_n\) is by induction. At first, we choose \(v_1 \in V_1\) arbitrarily. Let us assume that \(v_n, f_{n-1}\) and \(f'_{n-1}\) have already been chosen. Since \(\gamma\) has essentially infinite order, we can find \(m > n\) and \(v' \in V_m\) satisfying

\[
[e] \cap \gamma^k[e] \text{ is empty for all } e \in E_{v_n,v'}, \ 0 < |k| \leq n
\]

and

\[
\#_{E_{v_n,v'}} > 2(2n + 1) + h(v_n)(4n + 1) + 1,
\]

where \(E_{v_n,v'}\) means the set of edges from \(v_n\) to \(v'\). Choose \(f \in E_{v_n,v'}\) so that \([f]\) does not contain \(\gamma^k(x_{\text{max}})\) and \(\gamma^k(x_{\text{min}})\) for \(|k| \leq n\). For a large number \(l > m\), we may assume that

\[
\forall e \in E_{0,l}, \ \forall |k| \leq 2n, \ \exists e' \in E_{n,m}, \ \text{s.t.} \ \gamma^k([e]) \subset [e'].
\]

Then, let \(f_n \in E_{n,l}\) be an arbitrary extension of \(f\) and choose \(f'_n \in E\) as an extension of an edge of \(E_{v_n,v'}\) so that the above property is satisfied. Hence, the induction is completed.
Let $F$ be the set of infinite paths consisting of $f_n$’s and $f'_n$’s and $E$ be the union of all $\phi^k(F)$ for $k \in \mathbb{Z}$. We construct a new ordered Bratteli diagram $(W, E', \leq)$. Define the vertex set $W_n$ as the disjoint union of $V_n$ and $V'_n$, and let the map $\rho : W_n \to V_n$ be $\rho(v_n) = v_n$ and $\rho(v) = v$ for $v \in V_n$. Define the edge set $E'_n$ so that the incidence matrix $B_n$ satisfies

$$B_n(w, w') = \begin{cases} 0 & w = v'_n, w' \neq v_{n+1}, v'_{n+1} \\ 1 & w = v'_n, w' = v_{n+1}, v'_{n+1} \\ A_n(v_n, v_{n+1}) - 1 & w = v_n, w' = v_{n+1}, v'_{n+1} \\ A_n(\rho(w), \rho(w')) & \text{otherwise} \end{cases}$$

for $w \in W_n$ and $w' \in W_{n+1}$. It is not hard to check that the inductive limit of $(Z^{W_n})_n$ with the connecting maps $(B_n)_n$ is order isomorphic to $K^0(X, \phi)$. Let us define the linear order on the set $r^{-1}(w')$ for each $w' \in W_{n+1}$. When $w'$ is not equal to $v_{n+1}$ nor $v'_{n+1}$, we define the linear order by exactly the same way as in $r^{-1}(\rho(w'))$. If $w'$ is $v_{n+1}$ or $v'_{n+1}$, we change the source vertex of the edge $f_n$ or $f'_n$, respectively, to $v_n$. Then, we get a well-defined ordered Bratteli diagram.

Let $(Y, \psi)$ be the Cantor minimal system determined by $(W, E', \leq)$. Obviously, there is a factor map $\pi : (Y, \psi) \to (X, \phi)$, and $\pi$ is not one-to-one exactly on the subset $E$. By the construction of $E$, we see that $E$ satisfies the condition (ii) above. □

For an automorphism $\gamma$ which induces the identity on the dimension group, we introduced a new invariant $\eta(\gamma)$ in the Ext group ([M]). When $\gamma$ is a non-invertible endomorphism, we can define $\eta(\gamma)$ in the same way, by replacing $\gamma^{-1}$ with $\gamma$ in [M].

**Theorem 10.** Let $(X, \phi)$ be a Cantor minimal system, $\gamma \in C(\phi)$ be an automorphism which has essentially infinite order. Then, there exists a Cantor minimal system $(Z, \tau)$, a factor map $\Phi : (Z, \tau) \to (X, \phi)$ and a non-invertible endomorphism $\widetilde{\gamma} \in C(\tau)$ such that $\gamma \circ \Phi = \Phi \circ \widetilde{\gamma}$ and $\Phi$ induces the isomorphism between the dimension groups $K^0(X, \phi)$ and $K^0(Z, \tau)$. Moreover, when mod($\gamma$) is the identity, $\eta(\widetilde{\gamma})$ equals $\eta(\gamma)$ in the Ext group.

**Proof.** Choose $x_{\max} \in X$ arbitrarily and let $(Y, \psi)$ and $\pi$ as in Lemma 9. For each $i \in \mathbb{N} \cup \{0\}$, let $y_i \in Y$ be the unique point such that $\pi(y_i) = \gamma^i(x_{\max})$ and $(Z, \tau)$ be the topological joining generated by $(y_i)_i$. By applying Lemma 1 to $\gamma^{-1} \circ \pi$, we see that $(Z, \tau)$ is a Cantor minimal system. Let $\Phi$ be the composition of $\pi$ and the projection from $Z$ to the first summand. By Corollary 8, the factor map $\Phi$ induces an isomorphism between $K^0(X, \phi)$ and $K^0(Z, \tau)$. It is easily seen that the one-sided subshift on $Y^{\mathbb{N}\cup\{0\}}$ gives the well-defined centralizer $\widetilde{\gamma} \in C(\tau)$ and $\widetilde{\gamma}$ is a non-invertible endomorphism satisfying $\gamma \circ \Phi = \Phi \circ \widetilde{\gamma}$. □
By applying the above theorem to an odometer system \((X, \phi)\), we get a Cantor minimal system \((Z, \tau)\) which is strong orbit equivalent to \((X, \phi)\) and has a non-invertible endomorphism \(\gamma\). Of course, \(\gamma\) induces the identity map on the dimension group \(K^0(Z, \tau)\). Actually, in the proof of Lemma 9, we can find the Cantor minimal system \((Y, \psi)\) in the class of Toeplitz subshifts, if \((X, \phi)\) is an odometer system. Therefore, the Cantor minimal system \((Z, \tau)\) is the projective limit of Toeplitz minimal subshifts. But, the system \((Z, \tau)\) itself is not expansive. In [D], T. Downarowicz constructed an example of a Toeplitz minimal subshift which admits a non-invertible endomorphism. We can check that the endomorphism in his example does not induce the identity on the dimension group. The author doesn’t know non-coalescent minimal subshifts except for Downarowicz’s example.

Acknowledgments. The author is grateful to Professor Izumi for his constant encouragement.

References


Received June 10, 2000 and revised June 26, 2001.

**Department of Mathematics and Informations**  
**Chiba University**  
**Yayoityo 1-33, Inageku**  
**Chiba 263-8522**  
**JAPAN**  
**E-mail address:** matui@math.s.chiba-u.ac.jp
RATIONAL POLYNOMIALS OF SIMPLE TYPE

WALTER D. NEUMANN AND PAUL NORBURY

We classify two-variable polynomials which are rational of simple type. These are precisely the two-variable polynomials with trivial homological monodromy.

1. Introduction.

A polynomial map \( f : \mathbb{C}^2 \to \mathbb{C} \) is rational if its generic fibre, and hence every fibre, is of genus zero. It is of simple type if, when extended to a morphism \( \tilde{f} : X \to \mathbb{P}^1 \) of a compactification \( X \) of \( \mathbb{C}^2 \), the restriction of \( \tilde{f} \) to each curve \( C \) of the compactification divisor \( D = X - \mathbb{C}^2 \) is either degree 0 or 1. The curves \( C \) on which \( \tilde{f} \) is non-constant are called horizontal curves, so one says briefly “each horizontal curve is degree 1”.

The classification of rational polynomials of simple type gained some new interest through the result of Cassou-Nogues, Artal-Bartolo, and Dimca [4] that they are precisely the polynomials whose homological monodromy is trivial (it suffices that the homological monodromy at infinity be trivial by an observation of Dimca).

A classification appeared in [12], but it is incomplete. It implicitly assumes trivial geometric monodromy (on page 346, lines 10-11). Trivial geometric monodromy implies isotriviality (generic fibres pairwise isomorphic) and turns out to be equivalent to it for rational polynomials of simple type. The classification in the non-isotrivial case was announced in the final section of [17]. The main purpose of this paper is to prove it. But we recently discovered that there are also isotrivial rational polynomials that are not in [12], so we have added a classification for the isotrivial case using our methods. This case can also be derived from Kaliman’s classification [9] of all isotrivial polynomials. The fact that his list includes rational polynomials of simple type that are not in [12] appears not to have been noticed before (it also includes rational polynomials not of simple type).

In general, the classification of polynomial maps \( f : \mathbb{C}^2 \to \mathbb{C} \) is an open problem with extremely rich structure. One notable result is the theorem of Abhyankar-Moh and Suzuki [1, 23] which classifies all polynomials with one fibre isomorphic to \( \mathbb{C} \). The analogous result for the next simplest case, where one fibre is isomorphic to \( \mathbb{C}^* \), is open except in special cases where the
genus of the generic fibre of the polynomial is given. Kaliman [10] classifies all rational polynomials with one fibre isomorphic to \( \mathbb{C}^* \).

The basic tool we use in our study of rational polynomials is to associate to any rational polynomial \( f : \mathbb{C}^2 \to \mathbb{C} \) a compactification \( X \) of \( \mathbb{C}^2 \) on which \( f \) extends to a well-defined map \( \tilde{f} : X \to \mathbb{P}^1 \) together with a map \( X \to \mathbb{P}^1 \times \mathbb{P}^1 \). The map to \( \mathbb{P}^1 \times \mathbb{P}^1 \) is not in general canonical. We will exploit the fact that for a particular class of rational polynomials, there is an almost canonical choice.

Although we give explicit polynomials, the classification is initially presented in terms of the splice diagram for the link at infinity of a generic fibre of the polynomial (Theorem 4.1). This is called the regular splice diagram for the polynomial (since generic fibres are also called “regular”). See [15] for a description of the link at infinity and its splice diagram. The regular splice diagram determines the embedded topology of a generic fibre and the degree of each horizontal curve. Hence we can speak of a “rational splice diagram of simple type”.

The first author has asked if the moduli space of polynomials with given regular splice diagram is connected. For a rational splice diagram of simple type we find the answer is “yes”. We describe the moduli space for our polynomials in Theorem 4.2 and use it to help give explicit normal forms for the polynomials. We also describe how the topology of the irregular fibres varies over the moduli space.

The more general problem of classifying all rational polynomials, which would cover much of the work mentioned above, is still an open and interesting problem. It is closely related to the problem of classifying birational morphisms of the complex plane since a polynomial is rational if and only if it is one coordinate of a birational map of the complex plane. Russell [20] calls this a “field generator” and defines a good field generator to be a rational polynomial that is one coordinate of a birational morphism of the complex plane. A rational polynomial is good precisely when its resolution has at least one degree one horizontal curve, [20]. Daigle [5] studies birational morphisms \( \mathbb{C}^2 \to \mathbb{C}^2 \) by associating to a compactification \( X \) of the domain plane a canonical map \( X \to \mathbb{P}^2 \). A birational morphism is then given by a set of curves and points in \( \mathbb{P}^2 \) indicating where the map is not one-to-one. The approach we use in this paper is similar.

The full list of rational polynomials \( f : \mathbb{C}^2 \to \mathbb{C} \) of simple type is as follows. We list them up to polynomial automorphisms of domain \( \mathbb{C}^2 \) and range \( \mathbb{C} \) (so-called “right-left equivalence”).
Theorem 1.1. Up to right-left equivalence a rational polynomial $f(x,y)$ of simple type has one of the following forms $f_i(x,y)$, $i = 1, 2, or 3$.

\[ f_1(x,y) = x^{q_1}s^q + x^{p_1}s^p \prod_{i=1}^{r-1} (\beta_i - x^{q_1}s^q)^{a_i} \quad (r \geq 2) \]

\[ f_2(x,y) = x^{p_1}s^p \prod_{i=1}^{r-1} (\beta_i - x^{q_1}s^q)^{a_i} \quad (r \geq 1) \]

\[ f_3(x,y) = y \prod_{i=1}^{r-1} (\beta_i - x)^{a_i} + h(x) \quad (r \geq 1). \]

Here:

$0 \leq q_1 < q$, $0 \leq p_1 < p$, \[ \frac{p}{q} \begin{vmatrix} p_1 & q_1 \end{vmatrix} = \pm 1; \]

$s = yx^k + P(x)$, with $k \geq 1$ and $P(x)$ a polynomial of degree $< k$;

$a_1, \ldots, a_{r-1}$ are positive integers;

$\beta_1, \ldots, \beta_{r-1}$ are distinct elements of $\mathbb{C}^*$;

$h(x)$ is a polynomial of degree $< \sum_{i=1}^{r-1} a_i$.

Moreover, if $g_1(x,y) = g_2(x,y) = x^{q_1}s^q$ and $g_3(x,y) = x$ then $(f_i, g_i) : \mathbb{C}^2 \to \mathbb{C}^2$ is a birational morphism for $i = 1, 2, 3$. In fact, $g_i$ maps a generic fibre $f_i^{-1}(t)$ biholomorphically to $\mathbb{C} - \{0, t, \beta_1, \ldots, \beta_{r-1}\}$, $\mathbb{C} - \{0, \beta_1, \ldots, \beta_{r-1}\}$, or $\mathbb{C} - \{\beta_1, \ldots, \beta_{r-1}\}$, according as $i = 1, 2, 3$. Thus $f_1$ is not isotrivial and $f_2$ and $f_3$ are.

In [12] the isotrivial case is subdivided into seven subcases, but these do not include any $f_2(x,y)$ with $p, q, p_1, q_1$ all $> 1$.

2. Resolution.

Given a polynomial $f : \mathbb{C}^2 \to \mathbb{C}$, extend it to a map $\tilde{f} : \mathbb{P}^2 \to \mathbb{P}^1$ and resolve the points of indeterminacy to get a regular map $\tilde{f} : X \to \mathbb{P}^1$ that coincides with $f$ on $\mathbb{C}^2 \subset X$. We call $D = X - \mathbb{C}^2$ the divisor at infinity. The divisor $D$ consists of a connected union of rational curves. An irreducible component $E$ of $D$ is horizontal if the restriction of $\tilde{f}$ to $E$ is not a constant mapping. The degree of a horizontal curve $E$ is the degree of the restriction $\tilde{f}|E$. Although the compactification defined above is not unique, the horizontal curves are essentially independent of choice.

Note that a generic fibre $F_c := f^{-1}(c)$ is a punctured Riemann surface with punctures precisely where $F_c$ meets a horizontal curve. Thus $f$ has simple type if and only if $\overline{F_c}$ meets each horizontal curve exactly once, so the number of punctures equals the number of horizontal curves. For non-simple type the number of punctures will exceed the number of horizontal curves.
We say that a rational polynomial is \textit{ample} if it has at least three degree one horizontal curves. Those polynomials with no degree one horizontal curves, or bad field generators \cite{20}, are examples of polynomials that are not ample. The classification of Kaliman \cite{10} mentioned in the introduction gives examples of polynomials with exactly one degree one horizontal curve so they are also not ample. Nevertheless, ample rational polynomials will be the focus of our study in this paper. We will classify all ample rational polynomials that are also of simple type.

3. Curves in $\mathbb{P}^1 \times \mathbb{P}^1$.

If $\tilde{f} : X \rightarrow \mathbb{P}^1$ is a regular map with rational fibres then $X$ can be blown down to a Hirzebruch surface, $S$, so that $\tilde{f}$ is given by the composition of the sequence of blow-downs $X \rightarrow S$ with the natural map $S \rightarrow \mathbb{P}^1$; see \cite{2} for details. Moreover, by first replacing $X$ by a blown-up version of $X$ if necessary, we may assume that $S = \mathbb{P}^1 \times \mathbb{P}^1$ and the natural map to $\mathbb{P}^1$ is projection onto the first factor.

A rational polynomial $f : \mathbb{C}^2 \rightarrow \mathbb{C}$, once compactified to $\tilde{f} : X = \mathbb{C}^2 \cup D \rightarrow \mathbb{P}^1$, may thus be given by $\mathbb{P}^1 \times \mathbb{P}^1$ together with instructions how to blow up $\mathbb{P}^1 \times \mathbb{P}^1$ to get $X$ and how to determine $D$ in $X$. For this we give the following data:

- A collection $\mathcal{C}$ of irreducible rational curves in $\mathbb{P}^1 \times \mathbb{P}^1$ including $L_\infty := \infty \times \mathbb{P}^1$;
- a set of instructions on how to blow up $\mathbb{P}^1 \times \mathbb{P}^1$ to obtain $X$;
- a sub-collection $\mathcal{E}$ of the curves of the exceptional divisor of $X \rightarrow \mathbb{P}^1 \times \mathbb{P}^1$;

satisfying the condition:

- If $D$ is the union of the curves of $\mathcal{E}$ and the proper transforms of the curves of $\mathcal{C}$ then $X - D \cong \mathbb{C}^2$.

If $C \subset \mathbb{P}^1 \times \mathbb{P}^1$ is an irreducible algebraic curve we associate to it the pair of integers $(m,n)$ given by degrees of the two projections of $C$ to the factors of $\mathbb{P}^1 \times \mathbb{P}^1$. Equivalently, $(m,n)$ is the homology class of $C$ in terms of $H_2(\mathbb{P}^1 \times \mathbb{P}^1) = \mathbb{Z} \oplus \mathbb{Z}$. We call $C$ an $(m,n)$ curve. The intersection number of an $(m,n)$ curve $C$ and an $(m',n')$ curve $C'$ is $C \cdot C' = mn' + nm'$.

The above collection $\mathcal{C}$ of curves in $\mathbb{P}^1 \times \mathbb{P}^1$ will consist of some \textit{vertical curves} (that is, $(0,1)$ curves; one of these is $L_\infty$) and some other curves. These non-vertical curves give the horizontal curves for $f$, so they all have $m = 1$ if $f$ is of simple type. Note that a $(1,n)$ curve is necessarily smooth and rational (since it is the graph of a morphism $\mathbb{P}^1 \rightarrow \mathbb{P}^1$).

The image in $\mathbb{P}^1 \times \mathbb{P}^1$ of the fibre over infinity is the $(0,1)$ curve $L_\infty$ and the image of a degree $m$ horizontal curve is an $(m,n)$ curve. This view allows one to see as follows a geometric proof of the result of Russell \cite{20}
that a rational polynomial \( f \) is good precisely when its resolution has at least one degree one horizontal curve. A degree one horizontal curve for \( f \) has image in \( \mathbb{P}^1 \times \mathbb{P}^1 \) given by a \((1,n)\) curve. Call this image \( C \) and let \( P \) be its intersection with \( L_\infty \). The \((1,n)\) curves that do not intersect \( C - P \) form a \( \mathbb{C} \)-family that sweeps out \( \mathbb{P}^1 \times \mathbb{P}^1 - (L_\infty \cup C) \) so they lead to a map \( X \to \mathbb{P}^1 \) which takes values in \( C \) at points that do not lie over \( L_\infty \cup C \).

Restricting to \( C^2 = X - D \) we obtain a meromorphic function \( g_1 \) that has poles only at points that belong to exceptional curves that were blown up on \( C \) (and do not belong to \( E \)). However the polynomial \( f \) is constant on each such curve, so if \( c_1, \ldots, c_k \) are the values that \( f \) takes on these curves, then \( g := g_1(f-c_1)^{a_1} \cdots (f-c_k)^{a_k} \) will have no poles, and hence be polynomial, for \( a_1, \ldots, a_k \) sufficiently large. Then \((f,g)\) is the desired birational morphism \( C^2 \to C^2 \). For the converse, given a birational morphism \((f,g): C^2 \to C^2\), we compactify it to a morphism \((\tilde{f}, \tilde{g}): X \to \mathbb{P}^1 \times \mathbb{P}^1 \). Then the proper transform of \( \mathbb{P}^1 \times \infty \) is the desired degree one horizontal curve for \( f \).

We shall use the usual encoding of the topology of \( D \) by the dual graph, which has a vertex for each component of \( D \), an edge when two components intersect, and vertex weights given by self-intersection numbers of the components of \( D \). We will sometimes speak of the valency of a component \( C \) of \( D \) to mean the valency of the corresponding vertex of the dual graph, that is, the number of other components that \( C \) meets.

The approach we will take to get rational polynomials will be to start with any collection \( \mathcal{C} \) of \( k \) curves in \( \mathbb{P}^1 \times \mathbb{P}^1 \) and see if we can produce a divisor at infinity \( D \) for a map from \( C^2 \) to \( \mathbb{C}^2 \). In order to get a divisor at infinity we must blow up \( \mathbb{P}^1 \times \mathbb{P}^1 \), say \( m \) times, and include some of the resulting exceptional curves in the collection so that this new collection gives a divisor \( D \) whose complement is \( C^2 \). The exceptional curves that we “leave behind” (i.e., do not include in \( D \)) will be called cutting divisors.

**Lemma 3.1.**

(i) \( D \) must have \( m + 2 \) irreducible components, so we must include \( m - k + 2 \) of the exceptional divisors in the collection leaving \( k - 2 \) behind as cutting divisors;

(ii) \( D \) must be connected and have no cycles;

(iii) \( D \) must reduce to one of the “Morrow configurations” by a sequence of blow-downs. The Morrow configurations are the configurations of rational curves with dual graphs of one of the following three types, in which, in the last case, after replacing the central \((n,0,-n-1)\) by a single \((-1)\) vertex the result should blow down to a single \((+1)\) vertex by a sequence of blow-downs:

```
1
\( \circ \)

0 \( \ldots \) \( \circ \)
0 \( \circ \)

l_m \( \cdots \) \( l_1 \) \( n \) \( 0 \) \( -n-1 \) \( t_1 \) \( \cdots \) \( t_k \)
```

---

RATIONAL POLYNOMIALS OF SIMPLE TYPE 181
These conditions are also sufficient that $X - D \cong \mathbb{C}^2$.

Proof. The first property follows from the fact that each blow-up increases the rank of second homology by 1. Thus $H_2(X)$ has rank $m + 2$, so $D$ must have $m + 2$ irreducible components. Notice that this implies easily the well-known result [11, 12, 23] that

$$\delta - 1 = \sum_{a \in \mathcal{C}} (r_a - 1),$$

where $\delta$ is the number of horizontal curves of $f$ and $r_a$ is the number of irreducible components of $f^{-1}(a)$. (Both sides are equal to $k - 1 - \{\text{number of finite curves at infinity}\}$.)

The second property follows from the third property. For the third property and sufficiency see [13, 19]. \[\square\]

Now assume that $\tilde{f}$ has at least three degree one horizontal curves. Take these three horizontal curves and use them to map $X$ to $\mathbb{P}^1 \times \mathbb{P}^1$ as follows. The three horizontal curves define three points in a generic fibre of $\tilde{f}$. We can map this generic fibre to $\mathbb{P}^1$ by mapping these three points to $0, 1, \infty \in \mathbb{P}^1$. This defines a map from a Zariski open set of $X$ to $\mathbb{P}^1$ which then extends to a map $\pi$ from $X$ to $\mathbb{P}^1$. If $\pi$ is not a morphism then we blow up $X$ to get a morphism. Rather than introducing further notation for this blow-up we will assume we began with this blow-up and call it $X$. Together with the map $\tilde{f}$ this gives us the desired morphism

$$X \xrightarrow{(f, \pi)} \mathbb{P}^1 \times \mathbb{P}^1$$

with the property that the three horizontal curves map to $(1, 0)$ curves.

If all horizontal curves for $f$ are of type $(1, 0)$ then the generic fibres form an isotrivial family (briefly “$f$ is isotrivial”). Thus if $f$ is of simple type but not isotrivial, there must be a horizontal curve of type $(1, n)$ in $\mathcal{C}$ with $n > 0$. From now on, therefore, we assume that there are at least three $(1, 0)$ curves and at least one $(1, n)$ curve in $\mathcal{C}$ with $n > 0$.

Lemma 3.2. Any curve of $D$ that is beyond a horizontal curve from the point of view of $\tilde{L}_\infty$ has self-intersection $\leq -2$.

Proof. If the curve is an exceptional curve then it has self-intersection $\leq -1$. If $-1$, then the curve must have valency at least three (since any $-1$ exceptional curve that could be blown down is a cutting divisor). Any three adjacent curves must include two horizontal curves, which contradicts the fact that the dual graph of $D$ has no cycles. If the curve is not exceptional then it is the proper transform of a vertical curve. But we must have blown up at least three times on the vertical curve to get rid of cycles in the dual graph of $D$ so in this case the self-intersection is $\leq -3$. \[\square\]
3.1. **Horizontal curves.** The next few lemmas will be devoted to finding restrictions on the horizontal curves in the configuration $C \subset \mathbb{P}^1 \times \mathbb{P}^1$, culminating in Proposition 3.9.

**Lemma 3.3.** A horizontal curve of type $(1, n)$ in $C$ must be of type $(1, 1)$.

*Proof.* Assume we have a horizontal curve $C \in C$ of type $(1, n)$ with $n > 1$. It intersects each of the three $(1, 0)$ curves $n$ times (counting with multiplicity) so in order to break cycles—Lemma 3.1 (ii)—we have to blow up at least $n$ times on each $(1, 0)$ horizontal curve, so the proper transforms of the three $(1, 0)$ curves have self-intersection at most $-n$ and the proper transform of the $(1, n)$ curve has self-intersection at most $2n - 3n = -n$.

By Lemma 3.1 (iii), $D$ must reduce to a Morrow configuration by a sequence of blow-downs. Thus $D$ must contain a $-1$ curve $E$ that blows down. By Lemma 3.2, the curve $E$ must be a proper transform of a horizontal curve. The proper transform of each $(1, 0)$ curve has self-intersection at most $-n < -1$. Thus $E$ must come from one of the $(1, *)$ horizontal curves. As mentioned above, the proper transform of a $(1, k)$ curve has self-intersection $\leq -k$ so $E$ must be the proper transform of a $(1, 1)$ curve, $E_0$. But $E_0$ would intersect $C$, the $(1, n)$ curve, $2n$ times and hence $E.E \leq 2 - 2n < -1$ since $n > 1$. This is a contradiction so any horizontal curve of type $(1, n)$ must be a $(1, 1)$ curve. \hfill \Box

Hence, the horizontal curves consist of a collection of $(1, 0)$ curves and $(1, 1)$ curves. Figure 1 shows an example of a possible configuration of horizontal curves in $\mathbb{P}^1 \times \mathbb{P}^1$.

![Figure 1. Configuration of horizontal curves.](image)

**Lemma 3.4.** $\tilde{L}_\infty \cdot \tilde{L}_\infty = -1$.

*Proof.* We blow up at a point on $L_\infty$ precisely when at least two horizontal curves meet in a common point there. In general, if a horizontal curve meets $L_\infty$ with a high degree of tangency then we blow up repeatedly there. But, since all horizontal curves are $(1, 0)$ and $(1, 1)$ curves, they meet $L_\infty$ transversally, so a point on $L_\infty$ will be blown up at most once.
If there are two such points to be blown up, then after blowing up there will be (in the dual graph) two non-neighbouring $-1$ curves with valency $> 2$. The complement of such a configuration cannot be $\mathbb{C}^2$. This is proven by Kaliman [11] as Corollary 3. Actually the result is stated for two $-1$ curves of valency 3 but it applies to valency $\geq 3$.

Thus, at most one point on $L_\infty$ is blown up and $\tilde{L}_\infty \cdot \tilde{L}_\infty = 0$ or $-1$. We must show $0$ cannot occur.

Since there are at least four horizontal curves, if $\tilde{L}_\infty \cdot \tilde{L}_\infty = 0$, then $\tilde{L}_\infty$ has valency at least 4 and every other curve has negative self-intersection. Furthermore, the only possible $-1$ curves must be horizontal curves, and these intersect $\tilde{L}_\infty$ in $D$. As we attempt to blow down $D$ to get to a Morrow configuration, the only curves that can be blown down will always be adjacent to $\tilde{L}_\infty$. Thus the intersection number of $\tilde{L}_\infty$ will become positive and all other intersection numbers remain negative, so a Morrow configuration cannot be reached. Hence, $\tilde{L}_\infty \cdot \tilde{L}_\infty = -1$. □

Lemma 3.5. A configuration of curves that contains two branches consisting of curves of self-intersection $< -1$ that meet at a valency $> 2$ curve of self-intersection greater than or equal to $-1$ as in Figure 2 (where the meeting curve is drawn with valency 3 for convenience) cannot be blown down to a Morrow configuration.

Proof. Since the two branches consist of curves of self-intersection $< -1$, they cannot be reduced before the other branches are reduced. If the rest of the configuration of curves is blown down first then the valency $> 2$ curve becomes a valency 2 curve with non-negative self-intersection and no more blow-downs can be done. Since there is no 0 curve, we have not reached a Morrow configuration. □

![Figure 2](image-url)

Figure 2. The branches $B_1$ and $B_2$ consist of curves of self-intersection $< -1$ and $e \geq -1$.

Lemma 3.6. The intersection of any two $(1, 1)$ curves in $C$ consists of two distinct points contained in the union of the $(1, 0)$ curves in $C$. 
Proof. We will assume otherwise and reduce to the situation of Lemma 3.5 to give a contradiction. Thus, assume that two \((1, 1)\) curves do not intersect in two points contained in the union of the \((1, 0)\) curves. Then in order to break cycles these curves must be blown up at least four times—once each for at least three of the \((1, 0)\) curves and at least another time for the intersection of the two \((1, 1)\) curves. Thus they have self-intersection \(< -1\).

Case 1: Suppose two \((1, 1)\) curves meet on \(L_\infty\). Then after blowing up (twice if the \((1, 1)\) curves meet at a tangent), the exceptional curves are retained and the final exceptional curve has self-intersection \(-1\), valency 3 and two branches, which we will call \(B_1\) and \(B_2\), consisting of the proper transforms of the two \((1, 1)\) curves and any other curves beyond these proper transforms all of which have self-intersection \(< -1\). Thus we are in the situation of Lemma 3.5 and we get a contradiction.

Case 2: Suppose two \((1, 1)\) curves meet \(L_\infty\) at distinct points. Then at least one of the \((1, 1)\) curves, \(D\), must meet \(L_\infty\) at a point away from the \((1, 0)\) curves by Lemma 3.4. Also one of the \((1, 0)\) curves, \(H\), must meet \(L_\infty\) away from the \((1, 1)\) curves and contain at least two points where it intersects the \((1, 1)\) curves and thus have self-intersection \(< -1\) after blowing up to break cycles. We are once more at the situation of Lemma 3.5 where the valency \(> 2\) curve is \(\tilde{L}_\infty\) which has self-intersection \(-1\) by Lemma 3.4, and the branches \(B_1\) and \(B_2\) are the proper transform of \(D\) and any curves beyond it, respectively the proper transform of \(H\) and any curves beyond it. Thus we have a contradiction.

Notice that both cases apply to two \((1, 1)\) curves that may intersect at a tangent point, and shows that this situation is impossible. □

Lemma 3.7. If there is more than one \((1, 1)\) curve in \(C\) then there are exactly three \((1, 0)\) horizontal curves in \(C\).

Proof. Assume that there are more than three \((1, 0)\) horizontal curves in \(C\) and at least two \((1, 1)\) curves, say \(C_1\) and \(C_2\).

Case 1: \(C_1\) and \(C_2\) meet on \(\tilde{L}_\infty\). Then they meet each of at least two \((1, 0)\) curves in distinct points, so after blowing up to destroy cycles, these \((1, 0)\) curves have self-intersection \(\leq -2\) and Lemma 3.5 applies.

Case 2: \(C_1\) and \(C_2\) meet \(\tilde{L}_\infty\) at distinct points. Then one of them, say \(C_1\), meets \(\tilde{L}_\infty\) at a point not on a \((1, 0)\) curve by Lemma 3.4. At least one \((1, 0)\) curve \(C_3\) meets \(C_1\) and \(C_2\) in distinct points. After breaking cycles, \(C_1\) and \(C_3\) have self-intersections \(\leq -2\) so Lemma 3.5 applies again. □

Lemma 3.8. A family of \((1, 1)\) horizontal curves in \(C\) must pass through a common pair of points.
Proof. The statement is trivial for one \((1,1)\) horizontal curve so assume there are at least two \((1,1)\) horizontal curves in \(\mathcal{C}\). By the previous lemma, there are exactly three \((1,0)\) horizontal curves.

If there are exactly two \((1,1)\) horizontal curves in \(\mathcal{C}\) then the lemma is clear since the curves cannot be tangent by Lemma 3.6.

When there are more than two \((1,1)\) curves in \(\mathcal{C}\), apply Lemma 3.6 to two of them. If another \((1,1)\) horizontal curve in \(\mathcal{C}\) does not intersect these two \((1,1)\) curves at their common two points of intersection then, by Lemma 3.6, it must meet both these \((1,1)\) curves at the third \((1,0)\) horizontal curve of \(\mathcal{C}\). So the first two \((1,1)\) curves would meet there, which is a contradiction. □

Proposition 3.9. Any configuration of horizontal curves in \(\mathcal{C}\) is equivalent to one of the form in Figure 1.

Proof. By assumption and Lemma 3.3 there are at least three \((1,0)\) horizontal curves and some \((1,1)\) horizontal curves in \(\mathcal{C}\). If there is exactly one \((1,1)\) horizontal curve then the proposition is clear. If there is more than one \((1,1)\) horizontal curve, then by Lemmas 3.7 and 3.8 there are precisely three \((1,0)\) horizontal curves and two of the \((1,0)\) horizontal curves contain the common intersection of the \((1,1)\) curves. Each \((1,1)\) curve also contains a distinguished point where the curve meets the third \((1,0)\) horizontal curve.

A Cremona transformation can bring such a configuration to that in Figure 1 by blowing up at the two points of intersection of the \((1,1)\) curves and blowing down the two vertical lines containing the two points. This sends two of the \((1,0)\) horizontal curves and each \((1,1)\) curve to \((1,0)\) horizontal curves and one of the \((1,0)\) curves to a \((1,1)\) curve that intersects each of the other horizontal curves exactly once. Note that since we blow up \(\mathbb{P}^1 \times \mathbb{P}^1\) to get the polynomial map, two configurations of curves \(\mathcal{C}, \mathcal{C}'\) in \(\mathbb{P}^1 \times \mathbb{P}^1\) related by a Cremona transformation give rise to the same polynomial, so we are done. □

3.2. The configuration \(\mathcal{C}\). The image \(\mathcal{C}\) of \(D \subset X \to \mathbb{P}^1 \times \mathbb{P}^1\) will consist of the configuration of horizontal curves in Figure 1 plus some \((0,1)\) vertical curves. The next two lemmas show that in fact the only \((0,1)\) vertical curve we need to include in \(\mathcal{C}\) is \(L_\infty\) and furthermore that \(\mathcal{C}\) can be given by Figure 4.

Lemma 3.10. The configuration \(\mathcal{C}\) appears in Figure 3 or Figure 4.

Proof. Let \(r+2\) denote the number of horizontal curves and \(k+1\) denote the number of \((0,1)\) vertical curves in \(\mathcal{C}\). Thus \(\mathcal{C}\) consists of \(k+r+3\) irreducible components and by Lemma 3.1 \((i)\), when blowing up to get \(D\) from \(\mathcal{C}\) we must leave \(k+r+1\) exceptional curves behind as cutting divisors.

By Lemma 3.1 \((ii)\) we must break all cycles. The minimum number of cutting divisors needed to do this is \(kr+k+r-2\min\{k,r\}\). This is because each of the \(k\) \((0,1)\) vertical curves different from \(L_\infty\) must be separated from
all but one of the $r+1$ $(1,0)$ horizontal curves, so we need $kr$ cutting divisors. Also, the $(1,1)$ horizontal curve meets each of the $r+1$ $(1,0)$ horizontal curves and each of the $k$ $(0,1)$ vertical curves once, so that requires $k+r$ cutting divisors (by Lemma 3.4 the $(1,1)$ curve must meet $L_{\infty}$ at a triple point with a $(1,0)$ horizontal curve, so this intersection does not produce a cycle to be broken). We would thus require $kr+k+r$ cutting divisors except that the $(1,1)$ curve may pass through intersections of the $(1,0)$ horizontal curves and the $(0,1)$ vertical curves, so some of the cutting divisors may coincide. The most such intersections possible is $\min\{k,r\}$ and we have then over-counted required cutting divisors by $2\min\{k,r\}$. Hence we get at least $kr + k + r - 2\min\{k,r\}$ cutting divisors.

Since the number $k + r + 1$ of cutting divisors is at least $kr + k + r - 2\min\{k,r\}$, we have $k + r + 1 \geq kr + k + r - 2\min\{k,r\}$, so

$$1 \geq k(r - 2) \quad \text{and} \quad 1 \geq (k - 2)r, \quad k \geq 0, \ r \geq 2.$$

(1)

The solutions of (1) are $(k,r) = \{(0,r), (1,2), (1,3), (2,2)\}$.

Recall by Lemma 3.4 that the $(1,1)$ curve must meet $L_{\infty}$ at a triple point with a $(1,0)$ horizontal curve. Furthermore, by keeping track of when either inequality in (1) is an equality, or one away from an equality, we can see that the $(1,1)$ curve must meet any other $(0,1)$ vertical curves at a triple point with a $(1,0)$ horizontal curve. Thus, the only possible configurations for $C$ are given in Figures 3 and 4.

---

**Figure 3.** Configuration $C$.

**Figure 4.** Configuration $C$ with $r+2$ horizontal curves.
In the following lemmas we will exclude the configurations in Figure 3. Label the triple points in the first two configurations of Figure 3 by \( P_\infty \in L_\infty \) and \( P_1 \), and in the third configuration by \( P_\infty, P_1, P_2 \). Also, label the exceptional divisor obtained by blowing up the triple point \( P_i \) by \( E_i \) and its proper transform by \( \tilde{E}_i \).

**Lemma 3.11.** If \( E_i \) is a cutting divisor then the \((0, 1)\) vertical curve containing \( P_i \) can be removed from \( C \) by a birational transformation.

**Proof.** In each of the configurations of Figure 3 we can perform a Cremona transformation by blowing up \( P_\infty \) and \( P_i \) for \( i = 1 \) or 2 and then blowing down \( \tilde{L}_\infty \) and the proper transform of the \((0, 1)\) vertical curve that contains \( P_i \). The exceptional divisors \( E \) and \( E_i \) become \((0, 1)\) curves and the \((0, 1)\) vertical curve that contains \( P_i \) becomes an exceptional divisor in a new configuration \( \tilde{C} \). When \( E_i \) is a cutting divisor this operation essentially removes a \((0, 1)\) vertical curve from \( C \). \( \square \)

**Lemma 3.12.** In a configuration from Figure 3 with \((k, r) \in \{(1, 3), (2, 2)\}\) at least one of the exceptional divisors \( E_1 \) or \( E_2 \) is a cutting divisor.

**Proof.** Suppose otherwise, that \( E_1 \) is not a cutting divisor and for \((k, r) = (2, 2)\) nor is \( E_2 \) a cutting divisor. The exceptional curves \( E_i \) introduce an extra intersection and hence an extra cutting divisor is required. There is one such extra intersection in the configuration with \((k, r) = (1, 3)\) and two such extra intersections in the configuration with \((k, r) = (2, 2)\). As mentioned in the proof of Lemma 3.10 the solution \((k, r) = (1, 3)\) gives equality in (1) and so it cannot sustain an extra cutting divisor. Similarly the solution \((k, r) = (2, 2)\) is 1 away from equality in (1) and so it cannot sustain two extra cutting divisors. Hence we get a contradiction and the lemma is proven. \( \square \)

By the previous two lemmas we can simplify any configuration from Figure 3 to lie in Figure 4 or to be the first configuration from Figure 3 (the one with \((k, r) = (1, 2)\)) with the requirement that \( E_1 \) is not a cutting divisor. It is this last case that we will now exclude.

The next three lemmas suppose that we have the first configuration from Figure 3 and that \( E_1 \) is not a cutting divisor. We will denote the four horizontal curves by \( H_i, i = 1, \ldots, 4 \), and their proper transforms by \( \tilde{H}_i \) where \( H_4 \) is the \((1, 1)\) curve, \( H_1 \) contains \( P_1 \) and \( H_3 \) contains \( P_\infty \). Also denote the \((1, 0)\) vertical curve that contains \( P_1 \) by \( L_1 \) and its proper transform by \( \tilde{L}_1 \).

**Lemma 3.13.** At least one of \( \tilde{H}_1 \) and \( \tilde{H}_2 \) and at least one of \( \tilde{H}_3 \) and \( \tilde{H}_4 \) has self-intersection \(-1\).

**Proof.** The proper transform of each horizontal curve has self-intersection less than or equal to \(-1\) and all curves in \( D \) beyond horizontal curves have
self-intersection strictly less than $-1$. If the two horizontal curves that meet $\bar{L}_\infty$, $\bar{H}_1$ and $\bar{H}_2$, have self-intersection strictly less than $-1$, then since all curves beyond the two horizontal curves also have self-intersection strictly less than $-1$, and since $\bar{L}_\infty$ has self-intersection $-1$ and valence 3 this gives a contradiction by Lemma 3.5. The same argument applies to $\bar{H}_3$ and $\bar{H}_4$ together with $E$.

\textbf{Lemma 3.14.} $\bar{H}_4 \cdot \bar{H}_4 = -1$ if and only if $\bar{H}_2 \cdot \bar{H}_2 = -1$.

\textit{Proof.} Since $L_1$ must be separated from at least one of $H_2$ and $H_3$ then at most one of $\bar{H}_2 \cdot \bar{H}_2 = -1$ and $\bar{H}_3 \cdot \bar{H}_3 = -1$ can be true. Similarly $E_1$ must be separated from at least one of $H_1$ and $H_4$ so at most one of $\bar{H}_1 \cdot \bar{H}_1 = -1$ and $\bar{H}_4 \cdot \bar{H}_4 = -1$ can be true. By Lemma 3.13, if $\bar{H}_2 \cdot \bar{H}_2 \neq -1$ then $\bar{H}_1 \cdot \bar{H}_1 = -1$ so $\bar{H}_1 \cdot \bar{H}_4 \neq -1$. Similarly, $\bar{H}_1 \cdot \bar{H}_1 \neq -1$ implies that $\bar{H}_2 \cdot \bar{H}_2 = -1$ and $\bar{H}_4 \cdot \bar{H}_4 = -1$. \hfill $\Box$

\textbf{Lemma 3.15.} The configuration from Figure 3 with $(k, r) = (1, 2)$ together with the requirement that $E_1$ is not a cutting divisor cannot occur.

\textit{Proof.} Suppose otherwise. Assume that $\bar{H}_1 \cdot \bar{H}_1 = -1$ and $\bar{H}_3 \cdot \bar{H}_3 = -1$. If this is not the case, then by Lemmas 3.13 and 3.14 we may assume that $\bar{H}_4 \cdot \bar{H}_4 = -1$ and $\bar{H}_2 \cdot \bar{H}_2 = -1$ and argue similarly. The curves beyond $\bar{H}_1$ have self-intersection strictly less than $-1$. The curve immediately adjacent and beyond $\bar{H}_1$ is $E_1$ and this has self-intersection strictly less than $-2$. This is because we must blow up between $E_1$ and $H_4$ to separate cycles, and also between $E_1$ and $\bar{L}_1$ to break cycles and to maintain $\bar{H}_1 \cdot \bar{H}_1 = -1$ and $\bar{H}_3 \cdot \bar{H}_3 = -1$. Thus if we blow down $\bar{H}_1$ the remaining branch beyond $\bar{L}_\infty$ consists of curves with self-intersection strictly less than $-1$. Also $\bar{H}_2$ has self-intersection strictly less than $-1$ since we have to blow up the intersection between $H_2$ and $H_4$ and the intersection between $H_2$ and $L_1$ in order to break cycles and maintain $\bar{H}_3 \cdot \bar{H}_3 = -1$. After blowing down $\bar{H}_1$, $\bar{L}_\infty$ has self-intersection 0 and valency 3 with two branches consisting of curves of self-intersection strictly less than $-1$. Thus we can use Lemma 3.5 to get a contradiction. \hfill $\Box$


The configuration in Figure 4 is the starting point for any non-isotrivial rational polynomial of simple type. Notice that we can fill one puncture in each fibre of any such map to get an isotrivial family of curves and the puncture varies linearly with $c \in \mathbb{C}$. Notice also that there is an irregular fibre for each of the $r$ intersection points of the $(1, 1)$ curve with $(1, 0)$ horizontal curves away from $L_\infty$. In fact there is at most one more irregular fibre which can only occur in rather special cases, as we discuss in Subsection 4.1.
From now on the configuration $C$ is given by Figure 4 with $r + 2$ horizontal curves. Beginning with $C$ we will list all of the rational polynomials of simple type generated from this configuration. We shall give the splice diagrams for these polynomials first. Although we compute the polynomials later, geometric information of interest is often more easily extracted from the splice diagram or from our construction of the polynomials than from an actual polynomial.

The splice diagram encodes the topology of the polynomial. It represents the link at infinity of the generic fibre, or it can be thought of as an efficient plumbing graph for the divisor at infinity, $D \subset X$. It encodes an entire parametrised family of polynomials with the same topology of their regular fibres. See [7, 15, 16] for more details. Within this family, polynomials can still differ in the topology of their irregular fibres. Our methods also give all information about the irregular fibres, as we describe in Subsection 4.1.

The configuration $C$ has $r + 3$ irreducible components so when we blow up to get $D$ by Lemma 3.1 (i) we will leave $r + 1$ exceptional curves behind as cutting divisors. By Lemma 3.1 (ii) we must break the $r$ cycles in $C$ with multiple blow-ups at the points of intersection leaving $r$ exceptional curves behind as cutting divisors. We blow up multiple times between the $r$th $(1, 0)$ horizontal curve and the $(1, 1)$ horizontal curve in order to break a cycle. Thus, we require those blow-ups to satisfy the condition that the exceptional curve will break the cycle if removed. Equivalently, each new blow-up takes place at the intersection of the most recent exceptional curve with an adjacent curve. We call such a multiple blow-up a **separating blow-up sequence**.

We have one extra cutting divisor. This will arise as the last exceptional curve blown up in a sequence of blow-ups that does not break a cycle. We will call this sequence of blow-ups a **non-separating blow-up sequence**. A priori, this non-separating blow-up sequence could be a sequence as in Figure 5, where the final $-1$ curve is the cutting divisor. However, we shall see that the extra nodes this introduces in the dual graph prohibit $D$ from blowing down to a Morrow configuration, so the sequence is simply a string of $-2$ exceptional curves followed by $-1$ exceptional curve that is the cutting divisor.

![Figure 5](image-url)
RATIONAL POLYNOMIALS OF SIMPLE TYPE 191

This arises from blowing up a point on a curve in the blow-up of \( C \) that does not lie on an intersection of irreducible components.

Let us begin by just performing the separating blow-up sequences at the points of intersection of \( C \) and leaving the non-separating blow-up sequence until later. This gives the dual graph in Figure 6 with the proper transforms of the \( r + 1 \) \((1,0)\) horizontal curves and the \((1,1)\) horizontal curve indicated along with \( L_\infty \) and the exceptional curve \( E \) arising from the blow-up of the triple point in \( C \). There are \( r \) branches heading out from the proper transform of \((1,1)\) consisting of curves of self-intersection less than \(-1\) and beyond each of the proper transforms of the \( r \) \((1,0)\) horizontal curves the curves have self-intersection less than \(-1\).

![Figure 6. Dual graph of \( C \) blown up at points of intersection.](image)

The self-intersection of each of \((\tilde{1},0)_0\), \( E \) and \( \tilde{L}_\infty \) is \(-1\). The self-intersections of \((\tilde{1},1)\) and \((\tilde{1},0)_i\), \( i = 1, \ldots, r \) are negative and depend on how we blow up at each point of intersection.

**Lemma 4.1.** There is at most one branch in \( D \) beyond \((\tilde{1},1)\), and \( r - 1 \) of the horizontal curves \((\tilde{1},0)_i\) (those with index \( i = 1, \ldots, r - 1 \) say) have self-intersection \(-1\) and only \(-2\) curves beyond.

**Proof.** Since the self-intersection of each of the curves beyond \((\tilde{1},1)\) is less than \(-1\) each branch beyond \((\tilde{1},1)\) cannot be blown down before \((\tilde{1},1)\). Thus, there are at most two branches.

Furthermore, since the self-intersection of each of the curves beyond \((\tilde{1},0)_i\), \( i = 1, \ldots, r \) is less than \(-1\), the branch beyond \((\tilde{1},0)_i\) can be blown down before \((\tilde{1},0)_i\) only if \((\tilde{1},0)_i\) has self-intersection \(-1\) and each curve beyond has self-intersection \(-2\). Thus, at most two branches beyond \((\tilde{1},0)_i\), \( i = 1, \ldots, r \) do not consist of a \(-1\) curve with a string of \(-2\) curves beyond. If there are two such branches then the blow-ups that create them create corresponding branches beyond \((\tilde{1},1)\) (or possibly just decrease the intersection number at \((\tilde{1},1)\)). These two branches cannot be fully blown down until everything else connecting to the \( \tilde{L}_\infty \) vertex are blown down, but the vertex \((\tilde{1},1)\) and any branches beyond it cannot blow down first. Thus \( D \) cannot blow down
to a Morrow configuration. Thus there is at most one such branch, proving the Lemma.

Figure 7 gives the dual graph of the partially blown up $C$ where the label of each curve is now its self-intersection number. The branch beyond $(\tilde{1},0)_i$ consists of a string of $a_i - 1$ $-2$ curves and $A = \sum_{i=1}^{r-1} a_i$. We have thus far only blown up once between the $r$th $(1,0)$ horizontal curve and the $(1,1)$ horizontal curve, indicating the exceptional divisor by $\otimes$. We may blow up many more times—perform a separating blow-up sequence—leaving behind the final exceptional curve as cutting divisor to get a branch beyond $(\tilde{1},1)$ and a branch beyond $(1,0)_r$. In addition, we still have to perform the non-separating blow-up sequence at some point on the divisor.

![Image of a dual graph of partially blown-up configuration of curves.]

**Figure 7.** Dual graph of partially blown-up configuration of curves.

**Lemma 4.2.** The non-separating blow-up sequence occurs beyond either $(\tilde{1},1)$, $(1,0)_r$, or $(1,0)_0$ and in the latter case $(1,1) \cdot (1,1) = -1$.

**Proof.** If the non-separating blow-up sequence occurs on the branch beyond $(\tilde{1},0)_i$, $i = 1, \ldots, r-1$ then that branch cannot be blown down. By the proof of Lemma 4.1, in order to obtain a linear graph we must blow down $r - 1$ of the branches beyond $(\tilde{1},0)_i$, $i = 1, \ldots, r$. Thus, if the non-separating blow-up sequence does occur beyond $(\tilde{1},0)_i$ for some $i \leq r - 1$, then the $(1,0)_r$ branch blows down, so we simply swap the labels $i$ and $r$.

The non-separating blow-up sequence cannot occur on $E$ or $\tilde{L}_\infty$ because the resulting cutting divisor would not be sent to a finite value.

If the non-separating blow-up sequence occurs on the branch beyond $(1,0)_0$ then we must be able to blow down the branch beyond $(1,1)$, hence the branch must consist of $(1,1)$ with self-intersection $-1$.

**Lemma 4.3.** We may assume the non-separating blow-up sequence does not occur beyond $(1,0)_0$.
Proof. By Lemma 4.2 if the non-separating blow-up sequence occurs beyond 
\((\tilde{1},0)_0\) then \((\tilde{1},1) \cdot (\tilde{1},1) = -1\). In particular, \(1 = A = \sum_1^{r-1} a_i\). Thus, 
\(r = 2, a_1 = 1\). With only four horizontal curves, we can perform a Cremona 
transformation to make 
\((\tilde{1},0)_0\) the \((1,1)\) curve and hence we are in the first 
case of Lemma 4.2. □

Lemma 4.4. The non-separating blow-up sequence occurs on either of the 
last curves beyond \((1,1)\) or \((1,0)_r\) and is a string of \(-2\) curves followed by 
the \(-1\) curve that is a cutting divisor.

Proof. Arguing as previously, if the non-separating blow-up sequence occurs 
anywhere else, or if it is more complicated, then it introduces a new branch 
preventing the divisor \(D\) from blowing down to a linear graph. □

We now know that our divisor \(D\) results from Figure 7 by doing a sepa-
rating blow-up sequence between the \((1,1)\) curve and the \(r\)-th \((1,0)\) curve, 
leaving behind the final \(-1\) exceptional curve as a cutting divisor and then 
performing a non-separating blow-up sequence on a curve adjacent to this 
cutting divisor to produce second cutting divisor.

A priori, it is not clear that this procedure always gives rise to a divisor 
\(D \subset X\) where \(X\) is a blow-up of \(\mathbb{P}^2\) and \(D\) is the pre-image of the line at 
infinity. The classification will be complete once we show it does.

Lemma 4.5. The above procedure always gives rise to a configuration that 
blows down to a Morrow configuration (see Lemma 3.1) and hence deter-
mines a rational polynomial of simple type.

Proof. The calculation involves the relation between plumbing graphs and 
splice diagrams described in [7] or [16], with which we assume familiarity. 
In particular, we use the continued fractions of weighted graphs described 
in [7]. If one has a chain of vertices with weights \(-c_0, -c_1, \ldots, -c_t\), its 
continued fraction based at the first vertex is defined to be

\[
\frac{c_0}{1 - \frac{1}{c_1 - \frac{1}{c_2 - \ddots - \frac{1}{c_t}}}}.
\]

The dual graph for the curve configuration of Lemma 4.5 has chains start-
ing at the vertex \((\tilde{1},1)\) and \((\tilde{1},0)_r\). We claim these chains have continued 
fractions evaluating to \(A - 1 + \frac{P}{Q}\) and \(\frac{2}{P}\) respectively, where \(P, Q, p, q\) are arbi-
trary positive integers with \(Pq - pqQ = 1\). We describe the main ingredients 
of this calculation but leave the details to the reader.

An easy induction shows that the initial separating blow-up sequence 
leads to chains at \((\tilde{1},1)\) and \((1,0)_r\) with continued fractions \(A - 1 + \frac{p}{m}\) and
with positive coprime $n$ and $m$. The non-separating blow-up sequence then changes the fraction $\frac{m}{n}$ or $\frac{n}{m}$ to $\frac{m'}{n'}$ with $\frac{m'}{n'} \neq \frac{m}{n}$ or $\frac{n}{m}$ that it operates on as follows. If the non-separating blow-up sequence consists of $k$ blow-ups at the end of the left chain then $\frac{m}{n}$ is replaced by $\frac{m'}{n'}$ with $Nm - nM = 1$ and $k \leq \frac{M}{M} < \frac{N}{N} \leq k + 1).$ If the non-separating blow-up sequence is on the right then $\frac{n}{m}$ is similarly changed instead.

Renaming, we can describe this in terms of our chosen names $p, q, P, Q$ as follows. We either have $P > p$ or $q > Q$. If $P > p$ the initial separating blow-up sequence leads to chains with continued fractions $A - 1 + \frac{p}{q}$ and $\frac{q}{p}$ and the non-separating blow-up sequence then consists of a sequence of $k := \left\lfloor \frac{Q}{P} \right\rfloor$ blowups extending the left chain (and changing its continued fraction to $A - 1 + \frac{P}{Q}$). If $q > Q$ the continued fractions are $A - 1 + \frac{P}{Q}$ and $\frac{Q}{P}$ after the separating blowup and the non-separating blow-up consists of $k := \left\lfloor \frac{p}{Q} \right\rfloor$ blow-ups extending the right chain (and changing its continued fraction to $\frac{q}{p}$).

To prove the Lemma we must show that the dual graph of our curve configuration blows down to a Morrow configuration. We can blow down the chains starting at $(\tilde{1}, 0), i = 0, \ldots, r - 1$, to get a chain. To check that this chain is a Morrow configuration we must compute its determinant, which we can do with continued fractions as in [7]. We first replace the two end chains by vertices with the rational weights $-A + 1 - \frac{P}{Q}$ and $-\frac{q}{p}$ determined by their continued fractions to get a chain of four vertices with weights

$$-A + 1 - \frac{P}{Q}, \ 0, \ -1 + A, \ -\frac{q}{p}.$$ Then, computing the continued fraction for this chain based at its right vertex gives $\frac{q}{p} - \frac{Q}{P} = \frac{Pq - pQ}{Pp} = \frac{1}{Pp}$, showing that the determinant is $-1$ as desired, and completing the proof.

**Theorem 4.1.** Given positive integers $P, Q, p, q$ with $Pq - pQ = 1$ and positive integers $a_1, \ldots, a_{r-1}$, the splice diagram of our rational polynomial $f$ of simple type with non-isotrivial fibres is given in Figure 8 with

$$A = a_1 + \cdots + a_{r-1},$$
$$B = AQ + P - Q,$$
$$C = Aq + p - q,$$
$$b_i = qQa_i + 1 \quad \text{for each } i.$$ The degree of $f$ is: $\deg(f) = A(Q + q) + P + p.$

(In [17] an “additional” case was given, which is, however, of the above type with $P = Q = p = 1, q = a_1 = 2.$)
Proof. For the following computations we continue to assume the reader is familiar with the relationship between resolution graphs and splice diagrams described in [16]. The arrows signify places at infinity of the generic fibre, one on each horizontal curve. The fact that $(\tilde{1},0)_r$ is next to $\tilde{L}_\infty$ in the dual graph says that the edge determinant of the intervening edge is $1$. This corresponds to the fact that $Pq - pQ = 1$, which we already know. Similarly, $(\tilde{1},0)_i$ is next to $\tilde{L}_\infty$ for $i = 1, \ldots, r - 1$ so the weight $b_i$ is determined by the edge determinant condition $b_i = qQa_i + 1$. The "total linking number" at the vertex corresponding to each horizontal curve (before blowing down $(\tilde{1},0)_0$) is zero (terminology of [16]); this reflects the fact that the link component corresponding to the horizontal curve has zero linking number with the entire link at infinity, since at almost all points on a horizontal curve, the polynomial has no pole. The weight $C$ is determined by the zero total linking number of $(\tilde{1},1)$, giving $C = Aq + p - q$. For any $i$ the fact that vertex $(\tilde{1},0)_i$ has zero total linking gives $B = AQ + P - Q$. □

It is worth summarising some consequences of our construction that will be useful later.

Lemma 4.6. The number of blow-ups in the final non-separating blow-up sequence is $k := \max(\lfloor \frac{Q}{q} \rfloor, \lfloor \frac{P}{p} \rfloor)$ and these blow-ups occurred at the $(Q,-q)$ branch or the $(p,-P)$ branch of the above splice diagram according as the first or second entry of this max is the larger. Moreover, the non-separating blow-ups occurred on the corresponding horizontal curve if and only if $q = 1$ resp. $P = 1$.

Proof. The first part was part of the proof of Lemma 4.5. For the second part, note that if $q = 1$ then certainly $q > Q$ must fail, so $P > p$ and the non-separating blow-ups were on the left. The continued fraction on the left was $A - 1 + \frac{p}{q} = A - 1 + p$ which is integral, showing that the left chain consisted only of the exceptional curve before the non-separating blow-up.
Conversely, if the non-separating blow-ups were adjacent to that exceptional curve then the left chain was a single vertex, hence had integral continued fraction, so \( q = 1 \). The argument for \( P = 1 \) is the same. \( \square \)

**Theorem 4.2.** The moduli space of polynomials \( f: \mathbb{C}^2 \to \mathbb{C} \) with the above regular splice diagram, modulo left-right equivalence (that is, the action of polynomial automorphisms of both domain \( \mathbb{C}^2 \) and range \( \mathbb{C} \)), has dimension \( r+k-2 \) with \( k \) determined in the previous Lemma. In fact it is a \( \mathbb{C}^k \)-fibration over the \((r-2)\)-dimensional configuration space of \( r-1 \) distinct points in \( \mathbb{C}^* \) labelled \( a_1, \ldots, a_{r-1} \), modulo permutations that preserve the labelling and transformations of the form \( z \mapsto az \).

**Proof.** The splice diagram prescribes the number of horizontal curves and the separating blow-up sequences at each point of intersection. The only freedom is in the placement of the horizontal curves in \( \mathbb{P}^1 \times \mathbb{P}^1 \), and in the choice of points, on prescribed curves, on which to perform the string of blow-ups we call the non-separating blow-up sequence. The \((1,1)\) horizontal curve is *a priori* the graph of a linear map \( y = ax + b \) but can be positioned as the graph of \( y = x \) by by an automorphisms of the image \( \mathbb{C} \).

The point in the configuration space of the Theorem determines the placement of the horizontal curves \((1,0)_1, \ldots, (1,0)_r \) (after putting the \((1,0)_0 \) curve at \( \mathbb{P}^1 \times \{\infty\} \) and the \((1,0)_r \) curve at \( \mathbb{P}^1 \times \{0\} \)). The fibre \( \mathbb{C}^k \) determines the sequence of points for the non-separating blow-up sequence.

This proves the Theorem, except that we need to be careful, since some diagrams occur in the form of Theorem 4.1 in two different ways, which might seem to lead to disconnected moduli space. But the only cases that appear twice have four horizontal curves and the configurations \( \mathcal{C} \) are related by Cremona transformations. \( \square \)

This completes the classification of non-isotrivial rational polynomials of simple type.

### 4.1. The irregular fibres.

We can read off the topology of the irregular fibres of the polynomial \( f \) of Theorem 4.1 from our construction, since any such fibre is the proper transform of a vertical \((0,1)\) curve together with any exceptional curves left behind as cutting divisors when blowing up on this vertical curve.

We shall use the notation \( \mathbb{C}(r) \) to mean \( \mathbb{C} \) with \( r \) punctures (so \( \mathbb{C}^* = \mathbb{C}(1) \)), and for the purpose of this subsection we used \( C \cup C' \) to mean disjoint union of curves \( C \) and \( C' \), and \( C + C' \) to mean union with a single normal crossing. The generic fibre of \( f \) is \( \mathbb{C}(r+1) \).

The irregular fibres of \( f \) arise through the breaking of cycles between the \((1,1)\) curve and the \((1,0)_i \) curve for \( i = 1, \ldots, r \), so there are \( r \) of them. The non-separating blow-up also contributes, but it usually contributes to the \( r \)-th irregular fibre. However, if \( P = 1 \) or \( q = 1 \) then the non-separating
blow-up occurs on a horizontal curve and can thus have any \( f \)-value, so it generically leads to an additional \((r + 1)\)-st irregular fibre.

The irregular fibres are all reduced except for the \( r \)-th irregular fibre, which is always non-reduced unless one of \( P, Q, p, q \) is 1.

We first assume \( q \neq 1 \) and \( P \neq 1 \), so there are exactly \( r \) irregular fibres. Then for each \( i = 1, \ldots, r - 1 \) the \( i \)-th irregular fibre is \( \mathbb{C}(r - 1) + \mathbb{C}^* \) if \( a_i = 1 \) and \( \mathbb{C}(r) \cup \mathbb{C}^* \) if \( a_i > 1 \). The \( r \)-th irregular fibre is \( \mathbb{C}(r) \cup \mathbb{C}^* \cup \mathbb{C} \) generically.

As mentioned above, this fibre is reduced if and only if \( Q = 1 \) or \( p = 1 \).

There is a single parameter value in the \( \mathbb{C}^k \) factor of the parameter space of Theorem 4.2 for which the \( r \)-th irregular fibre has different topology, namely \( \mathbb{C}(r) \cup (\mathbb{C} + \mathbb{C}) \). In this case it is non-reduced even if \( Q = 1 \) or \( p = 1 \).

4.2. Monodromy. We can also read off the monodromy for our polynomial \( f \). Consider a generic vertical \((0,1)\) curve \( C \) in our construction. Removing its intersections with the horizontal curves gives a regular fibre \( F \) of \( f \). Since we have positioned the horizontal curve \((1,0) \) at \( \infty \) we think of \( F \) as an \((r + 1)\)-punctured \( \mathbb{C} \). We call the intersection of the \((1,1)\) horizontal curve with \( C \) the 0-th puncture of \( F \) and for \( i = 1, \ldots, r \) we call the intersection of the \((1,0) \) curve with \( C \) the \( i \)-th puncture of \( F \).

If the \((r+1)\)-st irregular fibre exists the local monodromy around it is trivial. For \( i = 1, \ldots, r \) the monodromy around the \( i \)-th irregular fibre rotates the 0-th puncture of the regular fibre \( \mathbb{C}(r+1) \) around the \( i \)-th puncture. In terms of the braid group on the \( r+1 \) punctures, with standard generators \( \sigma_i \) exchanging the \((i-1)\)-st and \( i \)-th puncture for \( i = 1, \ldots, r \), the local monodromies are \( h_1 = \sigma_1^2, h_2 = \sigma_1 \sigma_2 \sigma_1^{-1}, \ldots, h_r = \sigma_1 \sigma_2 \cdots \sigma_r^{-2} \sigma_{r-1} \cdots \sigma_1^{-1} \).

The monodromy \( h_\infty = h_r \cdots h_1 \) at infinity is \( \sigma_1 \sigma_2 \cdots \sigma_r \sigma_r \cdots \sigma_1 \). It is not hard to verify that \( h_1, \ldots, h_r \) freely generate a free subgroup of the braid group.

5. Explicit polynomials.

The splice diagram gives sufficient information (Newton polygon, topological properties, etc.) that one can easily find the polynomial without significant
computation by making an educated guess and then confirming that the guess is correct. The answer is as follows:

Case 1. $k \leq \frac{p}{P} < k + 1$. (Then $\frac{p}{P} < \frac{Q}{q} \leq k + 1$.)

Let $s_1 = \alpha_0 + \alpha_1 x + \cdots + \alpha_{k-1} x^{k-1} + x^k$. Let $\beta_1, \ldots, \beta_{r-1}$ be distinct complex numbers in $\mathbb{C}^*$. 

$$f(x,y) = x^{q - Qk} s_1^Q + x^{p - Pk} s_1^P \prod_{i=1}^{r-1} (\beta_i - x^{q - Qk} s_1^Q)^{a_i}.$$ 

Case 2. $k \leq \frac{Q}{q} < k + 1$. (Then $\frac{Q}{q} < \frac{P}{p} \leq k + 1$.)

Let $s_2 = \alpha_0 + \alpha_1 y + \cdots + \alpha_{k-1} y^{k-1} + xy^k$. Let $\beta_1, \ldots, \beta_{r-1}$ be distinct complex numbers in $\mathbb{C}^*$. 

$$f(x,y) = y^{Q - qk} s_2^Q + y^{p - pk} s_2^P \prod_{i=1}^{r-1} (\beta_i - y^{Q - qk} s_2^Q)^{a_i}.$$ 

One can compute the splice diagram and see it is correct. One can verify that the generic fibres are rational by the explicit isomorphism:

$$f^{-1}(t) \to \mathbb{C} - \{0, \beta_1, \ldots, \beta_{r-1}, t\} \quad \begin{cases} (x,y) \mapsto x^{q - Qk} s_1^Q & \text{(Case 1)}, \\ (x,y) \mapsto y^{Q - qk} s_2^Q & \text{(Case 2)}, \end{cases}$$

for generic $t$. The irregular values of $t$ are $0, \beta_1, \ldots, \beta_{r-1}$ if $P \neq 1$ and $q \neq 1$. If $P = 1$ then $t = \alpha_0 \prod \beta_i$ is the additional irregular value that our earlier discussion predicts, and if $q = 1$ then $t = \alpha_0$ is the additional irregular value.

The space of parameters $(\alpha_0, \ldots, \alpha_{k-1}, \beta_1, \ldots, \beta_{r-1})$ maps to the moduli space we computed earlier with fibre of dimension 1. Indeed, with $B,C$ as in Theorem 4.1, the polynomial

$$f_\lambda(x,y) = \lambda^{-1} f(\lambda^B x, \lambda^{-C} y)$$

has the same form with the parameters $\beta_j$ replaced by $\lambda^{-1} \beta_j$ and $\alpha_j$ replaced by $\lambda^{iB + A^{-1}} \alpha_j$.

To put the above polynomials in the form of $f_1(x,y)$ of Theorem 1.1, in Case 1 we rename the exponents $q - Qk$ to $q_1$, $p - Pk$ to $p_1$, $Q$ to $q$, $P$ to $p$.

In Case 2 we rename $Q - qk$ to $q_1$, $P - pk$ to $p_1$, and then exchange $x$ and $y$.

6. The isotrivial case.

After the first version of this paper was completed we realised that the classification in [12] for the isotrivial case has omissions. In this section we therefore sketch the corrected classification using the techniques of this paper. The discussion of the parameter spaces and the irregular fibres for the resulting polynomials is similar to the non-isotrivial case, so we leave it to
the reader. One can give an alternative proof using Kaliman’s classification [9] of all isotrivial polynomials.

We will restrict ourselves to the case of ample rational polynomials, i.e., those with at least three \((1,0)\) horizontal curves. The case of one \((1,0)\) horizontal curve always gives a polynomial equivalent to a coordinate by the Abhyankar-Moh-Suzuki theorem [1, 23]. The case of two \((1,0)\) horizontal curves is dealt with from a splice diagram perspective in [15] and earlier by analytic methods in [21]. The result is included in our summary Theorem 1.1.

As before, compactify \(\mathbb{C}^2\) to \(X\) and construct a map \(X \to \mathbb{P}^1 \times \mathbb{P}^1\). The map is essentially canonical (up to an automorphism of one factor.) The image of the divisor at infinity \(D \subset X\) in \(\mathbb{P}^1 \times \mathbb{P}^1\) is given by a collection of \((1,0)\) curves since we used three of the horizontal curves to get a map to \(\mathbb{P}^1 \times \mathbb{P}^1\) and in order that the fibres give an isotrivial family, any other horizontal curves must also be \((1,0)\) curves.

When there are at least three \((1,0)\) horizontal curves, by the following lemma the original configuration of curves in \(\mathbb{P}^1 \times \mathbb{P}^1\) breaks into the two cases of no vertical curves or one vertical curve.

**Lemma 6.1.** An ample rational polynomial with isotrivial fibres has at most one vertical curve over a finite value.

**Proof.** We can argue as in the previous section. The curve over infinity, \(L_{\infty}\) is not blown up since there are no triple points. If there is more than one vertical curve over a finite value then there are precisely three \((1,0)\) horizontal curves since otherwise there would be at least two \((1,0)\) horizontal curves that would be blown up at least twice and since all curves beyond these horizontal curves (exceptional curves or vertical curves) have self-intersection \(< -1\) we would get two branches \(B_1\) and \(B_2\) made up of the proper transforms of these two \((1,0)\) horizontal curves and all curves beyond these, meeting at a valency \(> 2\) curve, \(L_{\infty}\), with self-intersection 0. This is the impossible situation of Lemma 3.5.

There can be at most two vertical curves since if there are \(l\) vertical curves we need to break \(2l\) cycles but since there are precisely three \((1,0)\) horizontal curves, we begin with \(l + 4\) curves so we can break at most \(l + 2\) cycles by Lemma 3.1 (i). Therefore \(2l \leq l + 2\) so \(l \leq 2\).

The lemma follows when we get rid of the case of two vertical curves and three \((1,0)\) horizontal curves. The few cases are easily dismissed by hand.

So the beginning configuration is given by Figure 9 or Figure 10. We analyse these below as Case 1 and Case 2.

**Case 1.** Denote by \(r\) the number of horizontal curves. In Figure 9 we must leave behind \(r - 1\) curves as cutting divisors. To do so we do a non-separating
blow-up sequence on each of \( r - 1 \) horizontal curves (anything else leads to a configuration of curves whose intersection matrix has determinant 0, and which can therefore not blow down to a Morrow configuration). Thus, on the \( i \)-th horizontal curve we blow up \( a_i \) times and then leave behind the final exceptional divisor, giving a string of \(-2\) curves of length \( a_i - 1 \).

The resulting splice diagram is as in Figure 11.

Figure 9. Configuration of horizontal curves with \( L_\infty \).

Figure 10. Configuration of horizontal curves with \( L_\infty \) and a vertical curve over a finite value.

Figure 11. Splice diagram for Case 1 of isotrivial fibres.
This splice diagram has been analyzed in [16], where it is shown that its general polynomial is

\[ f(x, y) = y \prod_{i=1}^{r-1} (x - \beta_i)^{a_i} + h(x), \]

where \( h(x) \) is a polynomial of degree \(< \sum_{i=1}^{r-1} a_i.\)

This case covers the following cases from [12]: Case 1 of Theorem 3.3., Theorem 3.7, Case I of Theorem 3.10.

**Case 2.** Denote by \( r + 1 \) the number of horizontal curves. In Figure 10 we must do separating blow-up sequences at \( r \) intersection points and then do an additional non-separating blow-up sequence. As in Section 4, one finds that each of \( r - 1 \) of the separating blow-up sequences creates a string of \(-2\) curves attached to the corresponding horizontal curve, while the last one can be arbitrary, as described in the proof of Lemma 4.5. In Figure 12 we show the situation after doing the first \( r - 1 \) separating blow-up sequences and doing the first step of the \( r \)-th one.

![Figure 12. Dual graph of partially blown-up configuration of curves for Fig. 10.](image)

Moreover, the non-separating blow-up sequence then occurs adjacent to the exceptional curve left behind in the final separating blow-up sequence. The analysis is almost identical to the proof of Lemma 4.5, with the resulting strings now having continued fractions \( A + \frac{P}{Q} \) and \( \frac{Q}{P} \) respectively, with notation as in that proof.

The resulting splice diagram is as in Figure 13, with notation exactly as in Theorem 4.1. The polynomial in this case is exactly as in Section 5 except that the first term \( x^{q-Qk} s_1^Q \) respectively \( y^{Q-qk} s_2^Q \) is omitted. Namely, let \( \beta_1, \ldots, \beta_{r-1} \) be distinct complex numbers in \( \mathbb{C}^* \) and let \( k \) be as in Theorem 4.2.
If \( k \leq \frac{P}{Q} < k+1 \) (so \( \frac{P}{Q} < \frac{Q}{p} \leq k+1 \)), let \( s_1 = \alpha_0 + \alpha_1 x + \cdots + \alpha_{k-1} x^{k-1} + x^k y \). Then
\[
f(x, y) = x^{p-Pk} y^{1} \prod_{i=1}^{r-1} (\beta_i - x^{q-Qk} s_1^Q)^{a_i}.
\]

If \( k \leq \frac{Q}{p} < k+1 \) (so \( \frac{Q}{p} < \frac{P}{Q} \leq k+1 \)), let \( s_2 = \alpha_0 + \alpha_1 y + \cdots + \alpha_{k-1} y^{k-1} + xy^k \). Then
\[
f(x, y) = y^{p-Pk} s_2^p \prod_{i=1}^{r-1} (\beta_i - y^{q-Qk} s_2^Q)^{a_i}.
\]

This case covers the following cases from [12]: Cases 2, 3, 4 of Theorem 3.3 and Case II of Theorem 3.10. However, [12] only has examples in which one of \( P, Q, p, q \) is equal to 1.

Note that the isotrivial splice diagrams of Case 1 and Case 2 can be considered to belong to one family: putting \( (P, Q) = (1, 0) \) in Figure 13 gives Figure 11. Nevertheless, the two cases have rather different geometric properties.

7. General rational polynomials.

In this section we will give a result for ample rational polynomials that are not necessarily of simple type.

**Proposition 7.1.** An ample rational polynomial contains a \((1, 0)\) horizontal curve whose proper transform has self-intersection \(-1\) and meets \( \bar{L}_\infty \).

*Proof.* By the classification of ample rational polynomials of simple type, the proposition is true in this case. So, we may assume that there is a horizontal curve of type \((m, n)\) for \( m > 1 \).

Suppose there is no \((1, 0)\) horizontal curve with the property of the proposition. Then by the proof of Lemma 3.4 there are at least two \((1, 0)\) horizontal curves whose proper transforms have self-intersection \(-1\) and meet.
any curves beyond these horizontal curves have self-intersection \(< -1\).

A horizontal curve of type \((m, n)\) must meet \(L_\infty\) at exactly one point, and hence with a tangency of order \(m\) or at a singularity of the curve. This is because if a horizontal curve were to meet \(L_\infty\) twice then we would not be able to break cycles since when we blow up next to \(L_\infty\), those exceptional curves are sent to infinity under the polynomial and hence must be retained in the configuration of curves. Thus we must blow up there to get a configuration of curves with normal intersections. The final exceptional curve in such a sequence of blow-ups will have self-intersection \(-1\) and valency \(> 2\).

If we can blow down the configuration of curves then eventually at least one curve adjacent to the \(-1\) curve is blown down and hence the \(-1\) curve ends up with non-negative self-intersection. But the final configuration is not a linear graph since the proper transforms of the two \((1, 0)\) horizontal curves and any curves beyond give two branches. Thus the final configuration is not a Morrow configuration which contradicts Lemma 3.1. □

The following result is a generalisation of Lemma 3.3.

**Corollary 7.1.** For any ample rational polynomial, a smooth horizontal curve of type \((m, n)\) with \(m > 0\) must be of type \((m, 1)\).

**Proof.** The statement is true for \(m = 1\) by Lemma 3.3 so will assume \(m > 1\). A curve of type \((m, n)\) will intersect the \((1, 0)\) horizontal curves \(m\) times, with multiplicity, unless possibly if the \((m, n)\) curve is singular at these points of intersection. The latter possibility is ruled out by the assumption of the corollary. Hence the \((1, 0)\) horizontal curves will be blown up at least \(m\) times and their proper transforms will have self-intersection \(< -m\). This contradicts the previous proposition so the result follows. □

When the rational polynomial is not ample, Russell has an example of a horizontal curve of type \((3, 2)\). See the examples in the next section. Note that smoothness of the horizontal curve is necessary in the corollary (at the points of intersection with the \((1, 0)\) horizontal curves) since we can always have two horizontal curves of types \((l, 1)\) and \((m, 1)\) and together they can be considered as a singular horizontal curve of type \((l + m, 2)\).

### 7.1. Adding horizontal curves.

Consider the following construction on \(\mathbb{C}^2\). Blow up repeatedly starting at a point on the \(y\)-axis so that the resulting exceptional curves form a chain from the \(y\)-axis to the last exceptional curve blown up. If we now remove the \(y\)-axis and all but the last exceptional curve from the blown-up \(\mathbb{C}^2\) we get a new \(\mathbb{C}^2\) that we call \(\mathbb{C}^2_1\). Any polynomial \(f: \mathbb{C}^2 \to \mathbb{C}\) induces a polynomial \(f_1: \mathbb{C}^2_1 \to \mathbb{C}\). Suppose the \(y\)-axis intersects generic fibres of \(f\) in \(d\) points. Then the generic fibres of \(f_1\) are simply generic fibres of \(f\) with \(d\) extra punctures. In fact, this construction simply
adds an extra degree \( d \) horizontal curve, namely the \( y \)-axis becomes a degree \( d \) horizontal curve for \( f_1 \).

From the point of view of the polynomials, what we have done is replaced \( f(x,y) \) by

\[
f_1(x,y) = f(x,s), \quad s = a_0 + a_1x + \cdots + a_{k-1}x^{k-1} + x^ky,
\]

that is, we have composed \( f \) with the birational morphism \((x, y) \mapsto (x, s)\) of \( \mathbb{C}^2 \).

Since one can compose \( f \) first with a polynomial automorphism to raise its degree, one can easily add horizontal curves of arbitrarily high degree by this construction. This makes clear that any classification of non-simple-type polynomials must take account of this sort of operation, including composition with more complicated birational morphisms.

Although this is a complication, it can also simplify some issues. Here is a simple illustrative example. We start with the simplest rational polynomial \( g(x,y) = x \), apply a polynomial automorphism to get \( f(x,y) = x + y^2 \) and then apply the above birational morphism to get \( f_1(x,y) = x + (a_0 + a_1x + \cdots + a_{k-1}x^{k-1} + x^ky)^2 \) with one degree one horizontal and one degree two horizontal. It is not hard to check (e.g., by listing possible splice diagrams) that this gives, up to equivalence, the only non-simple-type polynomials with generic fibre \( \mathbb{C} - \{0, 1\} \), so with the classification of simple type polynomials, we get:

**Proposition 7.2.** A polynomial with general fibre \( \mathbb{C} - \{0, 1\} \) is left-right equivalent to one of the form \( f_2(x,y) \) or \( f_3(x,y) \) of Theorem 1.1 with \( r = 2 \) or \( r = 3 \) respectively, or to \( f(x,y) = x + (a_0 + a_1x + \cdots + a_{k-1}x^{k-1} + x^ky)^2 \).

This proposition also follows from Kaliman’s classification [9] of isotrivial polynomials.

**8. Examples.**

It is worth including some interesting known examples of rational polynomials from the perspective used in this paper. These examples are neither of simple type nor ample.

Russell [20] (correctly presented in [3]) constructed an example of a rational polynomial with no degree one horizontal curves. This is an example of a bad field generator—a polynomial that is one coordinate in a birational transformation but not in a birational morphism. It is given by beginning with three curves in \( \mathbb{P}^1 \times \mathbb{P}^1 \) as in Figure 14. The \( (2, 1) \) curve and the \( (3, 2) \) curve intersect at an order three tangency and at the same point the \( (3, 2) \) intersects itself at a tangency. They are the two horizontal curves of the polynomial. The vertical curve is \( L_\infty \).

The actual polynomial in this case is, with \( s = xy + 1 \),

\[
f(x, y) = (y^2s^4 + g(s + xy)s + 1)(ys^5 + 2xy^2s^2 + x)
\]
Kaliman [10] classified all rational polynomials with one fibre isomorphic to $\mathbb{C}^*$. Figure 15 gives three curves in $\mathbb{P}^1 \times \mathbb{P}^1$, the two horizontal curves and $L_\infty$. The $(m, 1)$ curve has the property that when it is mapped downwards onto a $(1, 0)$ curve, there are only two points of ramification, both with maximal ramification of $m$, at $L_\infty$ and at the irregular fibre isomorphic to $\mathbb{C}^*$. Kaliman’s entire classification begins with this configuration of curves. The only points that can be blown up are those that are infinitely near to the point of intersection of the two horizontal curves (besides the unnecessary blowing up where the $(m, 1)$ curve meets $L_\infty$) and one exceptional curve is left behind as a component of the reducible fibre.
Figure 15. Classification of rational polynomials with a $\mathbb{C}^*$ fibre.

References


Received September 18, 2000 and revised February 21, 2001. This research was supported by the Australian Research Council.

DEPARTMENT OF MATHEMATICS
BARNARD COLLEGE, COLUMBIA UNIVERSITY
NEW YORK, NY 10027
E-mail address: neumann@math.columbia.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MELBOURNE
AUSTRALIA 3010
E-mail address: norbs@ms.unimelb.edu.au
DILATATION OF MAPS BETWEEN SPHERES

PENG CHIAKUEI AND TANG ZIZHOU

Dedicated to Professor Wu W.T. on his 80th birthday

For a smooth map between spheres, we are concerned with the relation between its homotopy class (topological complexity) and its dilatation (geometrical complexity). This paper (1) generalizes the results of Olivier and Roitberg on the dilatation of Hopf fibrations and the elements of the stable homotopy groups of spheres. (2) Disproves two conjectures of Olivier and Roitberg by showing that \( \delta(2, 4) < 3 \) and \( \delta(3, 4) = 2 \).

1. Introduction.

Let \( f : S^m \rightarrow S^n \) be a \( C^1 \) map of the standard unit spheres and let \( df \) be the differential of \( f \). Following Olivier [Ol], we define \( \delta(f) \), the dilatation (or “stretching”) of \( f \), by the formula

\[
\delta(f) = \sup \{ |df(V)| \mid |V| = 1 \}
\]

where \( V \) ranges over all unit tangent vectors of \( S^m \), and \(|\cdot|\) is the Euclidean length. Moreover, for a homotopy class \( \alpha \in \pi_m(S^n) \), we define

\[
\delta(\alpha) = \inf \{ \delta(f) \mid [f] = \alpha \},
\]

the infimum taken over all differentiable representatives \( f \) of \( \alpha \).

As stated in [Gr1], we ask how to estimate a measure of the topological complexity of a map \( f : S^m \rightarrow S^n \) by its geometry. It is natural to measure geometrical complexity of \( f \) by its dilatation. The topological complexity of \( f \) may be measured by its homotopy class or the Brouwer degree (when the degree makes sense). We should point out in fact that our present paper arose out of continuous attempt to represent homotopy class of spheres by harmonic maps or algebraic maps (cf. [PT2], [PT3] and [PT4]).

Olivier remarked in [Ol, p. 387] that the Hopf fibrations \( S^{15} \rightarrow S^8, S^7 \rightarrow S^4, S^3 \rightarrow S^2 \) have dilatation exactly 2. Furthermore, Roitberg investigated the elements of the stable homotopy groups of spheres \( \pi_k^s = \lim_{n \rightarrow \infty} \pi_{n+k}(S^n) \), and was able to completely analyze the situation for those elements in \( \pi_k^s \) which lie in the image of the stable \( J \)-homomorphism \( \pi_k(O) \rightarrow \pi_k^s \). The dilatation of any such nonzero element was shown to be always 2 [Ro, p. 202].
To generalize these results, we try to get information on polynomial maps between spheres. Using Bernstein-Szego Theorem, we can establish in Section 2 a succinct but nice estimate which implies immediately the statements mentioned above.

**Theorem 1.1.** Let $f : S^m \to S^n$ be a polynomial map of algebraic degree $k$. Then $\delta(f) \leq k$.

Several interesting consequences of this theorem will be given in Section 2.

In Section 3, we make a deep study of problem of obtaining upper bounds for the dilatation invariants of maps $S^2 \to S^2$, $S^3 \to S^3$ and $S^3 \to S^2$, respectively. In [Ol, p. 389], Olivier used the following definition ($n \geq 0, k \geq 0$):

$$\delta(n, k) = \inf \{ \delta(f) | f : S^n \to S^n \text{ differentiable, } \deg(f) = k \}$$

where $\deg(f)$ denotes the Brouwer degree of the map $f$, which determines completely the homotopy class $[f] \in \pi_n S^n = \mathbb{Z}$. It is worth summarizing the following fundamental characterizations.

(i) $\delta(1, k) = k$ and $\delta(n, k) \leq k$ [Ol];

(ii) $\delta(n, k) \geq \sqrt{n}k$ [Ol] and [Gr2];

(iii) $\delta(n, 2) = 2$ [Ol] and [He];

(iv) $\delta(n, k) \geq 2$ for $|k| \geq 2$ [Ol] and [Gr1].

It was conjectured in [Ol] that $\delta(2, k) = k$ for $k \geq 0$. Unfortunately, this conjecture will be shown to be false for $k = 4$ by the following inequality.

**Proposition 1.2.** $\delta(2, 4) \leq 2\sqrt{2}$.

It would be interesting to know the exact value of $\delta(2, 4)$.

In order to estimate the dilatation of elements in $\pi_3 S^3$, Olivier [Ol, p. 389] constructed explicitly a map $f : S^3 \to S^3$ and stated that this map has dilatation exactly 2. However, we will show that his map has dilatation more than $\sqrt{\frac{24}{3}}$. In fact, we can prove that any differentiable join: $S^3 = S^1 \ast S^1 \to S^3$ has dilatation more than 2. On the other hand, we will construct a family of differentiable maps $S^3 \to S^3$ of the Brouwer degree 4 whose dilatation approaches 2. These will imply the following:

**Proposition 1.3.** $\delta(3, 4) = 2$.

It was conjectured in [Ro, p. 202] that $\delta(\sigma \circ \tau) = \delta(\sigma) \cdot \delta(\tau) = 2\delta(\tau)$, at least if $\sigma$ is one of the Hopf classes. The final section will provide a counterexample.

**Theorem 1.4.** Let $\tau \in \pi_3 S^3$ be the homotopy class of the Brouwer degree 2, and $\sigma \in \pi_3 S^2$ the Hopf class, then $\delta(\sigma \circ \tau) < \sqrt{15} < 4 = 2\delta(\tau)$.
2. Polynomial maps.

Recall that a map \( f: S^m \to S^n \) is said to be a polynomial map if it is the restriction to \( S^m \) of a polynomial map from \( \mathbb{R}^{m+1} \) to \( \mathbb{R}^{n+1} \). If, in addition, the polynomials are all homogeneous of degree \( k \), then \( f \) is called a \( k \)-form (see, for example, [Wo]). In order to prove Theorem 1.1, we need to use the Bernstein-Szego Inequality which bounds the derivative of a trigonometric polynomial in terms of its supremum norm.

Bernstein-Szego Inequality ([BE, p. 232]). Let

\[
    f(\theta) = a_0 + \sum_{\lambda=1}^{k} (a_\lambda \cos \lambda \theta + b_\lambda \sin \lambda \theta), \quad \theta \in [0, 2\pi].
\]

Then

\[
    |f'(\theta)| \leq k \cdot \max_{\theta \in [0, 2\pi]} |f(\theta)|.
\]

We are now in a position to prove Theorem 1.1.

**Proof of Theorem 1.1.** Suppose that \( f: S^m \to S^n \) is a polynomial map of degree \( k \). Given any point \( x \) in \( S^m \) and a unit tangent vector \( V \) to \( x \), denote by \( S^1 \) the great circle spanned by \( x \) and \( V \). By a suitable orthogonal transformation of \( \mathbb{R}^{m+1} \), we may assume that \( x = (1, 0, \ldots, 0) \), \( V = (0, 1, \ldots, 0) \) and

\[
    S^1 = \{(\cos \theta, \sin \theta, 0, \ldots, 0) \mid 0 \leq \theta \leq 2\pi\}.
\]

Observe that any orthogonal transformation preserves the degree of the polynomials.

Furthermore, by a suitable orthogonal transformation of \( \mathbb{R}^{n+1} \), we may assume

\[
    f|_{S^1} (\cos \theta, \sin \theta, 0, \ldots, 0) = (f_0, f_1, \ldots, f_n)
\]

with derivatives \( f'_0 = a, f'_1 = f'_2 = \cdots = f'_n = 0 \), at \( \theta = 0 \). Thus the function \( f_0 = f_0(\theta) \) is represented by trigonometric polynomial of degree at most \( k \),

\[
    f_0(\theta) = a_0 + a_1 \cos \theta + b_1 \sin \theta + \cdots + a_k \cos k\theta + b_k \sin k\theta (a_i,b_j \in \mathbb{R}).
\]

Since \( f_0^2 + \cdots + f_n^2 = 1 \), \( |f_0(\theta)| \leq 1 \), by the Bernstein-Szego Inequality, we finally obtain the vital inequality

\[
    |a| = |f'_0| \leq k \cdot \max |f_0| \leq k,
\]

and thus \( |df_x(V)| = |a| \leq k \). \( \square \)

Modelling on the proof of Theorem 1 in [La], we may obtain:

**Lemma 2.1.** Let \( f : S^m \to S^n \) be a 2-form (quadratic map). Assume moreover that \( f \) is onto. Then \( \delta(f) \geq 2 \).
Proof. Suppose $\delta(f) < 2$. Take any point $x$ in $S^m$ so that $f(-x) = f(x)$. Then each meridian in $S^m$ from $x$ to $(-x)$ is mapped into a loop passing through $f(x)$ of length less than $2\pi$. Hence $-f(x)$ does not lie in the image of $f$, and therefore $f$ is not onto. □

Combining the preceding lemma with Theorem 1.1 we obtain at once:

**Corollary 2.2.** Let $f : S^m \to S^n$ be a 2-form. If $f$ is onto, then $\delta(f) = 2$.

**Remark 2.3.** In [Ro], Roitberg proved his Theorem 2 by computing the dilatation of map $\overline{f} : S^{n+k} \to S^m$, defined by

$$\overline{f}(a, b) = (2g(a) \cdot b, |b|^2 - |a|^2)$$

for $a = (a_1, \ldots, a_{k+1}) \in R^{k+1}, b = (b_1, \ldots, b_n) \in R^n$, where $g(S^k) \subset O(n)$. Note that this map $\overline{f}$ is a quadratic map, the assertion $\delta(\overline{f}) = 2$ follows immediately from Corollary 2.2.

Theorem 1.1 provides upper bounds for the dilatation of polynomial maps between spheres. It is natural to ask the question: Given $m, n, k \in \mathbb{Z}^+$, when does there exist a polynomial map $f : S^m \to S^n$ of degree $k$, with $\delta(f) = k$? If $m > n$, according to [Wo], it is in general not easy to solve this question. However, if $m = n$, we can answer this question affirmatively with the following:

**Theorem 2.4.** Let $f_k : S^n \to S^n(k > 0)$ be defined by

$$(\cos \theta, \sin \theta \cdot x) \mapsto (\cos k\theta, \sin k\theta \cdot x),$$

where $0 \leq \theta \leq \pi$ and $x \in S^{n-1} \subset R^n$. Then $f_k$ can be characterized as follows:

1. $f_k$ is actually a $k$-form;
2. $\delta(f_k) = k$;
3. the Brouwer degree of $f_k$ is given by

$$\deg f_k = \begin{cases} k & \text{if } n \text{ is odd}, \\ 1 & \text{if } n \text{ is even and } k \text{ is odd}, \\ 0 & \text{otherwise}. \end{cases}$$
Proof. (1) It suffices to observe that
\[
\cos k\theta + i \sin k\theta = (\cos \theta + i \sin \theta)^k
\]
\[
= \sum_j \binom{k}{j} \cos^{k-j} \theta \cdot \sin^j \theta \cdot i^j
\]
\[
= \sum_p (-1)^p \binom{k}{2p} \cos^{k-2p} \theta \cdot |\sin \theta \cdot x|^{2p}
\]
\[
+ \sum_p (-1)^{p_0} \binom{k}{2p+1} \cos^{k-2p-1} \theta \cdot |\sin \theta \cdot x|^{2p} \cdot \sin \theta.
\]

(2) Since \( f_k \) is a \( k \)-form, it follows from Theorem 1.1 that \( \delta(f_k) \leq k \). It remains only to find a suitable point \( x \) in \( S^n \) and a suitable unit tangent vector \( V \) to \( x \) satisfying \( |d(f_k)_x(V)| = k \). The work is not difficult and left to the reader.

(3) The conclusion can be proved in several ways. We will make use of moving frames to calculate the Brouwer degree of \( f_k \). Let \( S^{n-1} = \{ x \in \mathbb{R}^n \mid \langle x, x \rangle = 1 \} \). We choose a local field of orthonormal frame \( e_1, e_2, \ldots, e_{n-1} \) in \( TS^{n-1} \). Define
\[
\omega_i = \langle dx, e_i \rangle, i = 1, 2, \ldots, n-1.
\]
Then the volume element of \( S^{n-1} \) can be written as
\[
dS^{n-1} = \omega_1 \wedge \cdots \wedge \omega_{n-1}.
\]
Now write \( S^n \) as
\[
S^n = \{ y \in \mathbb{R}^{n+1} \mid y = (\cos \theta, \sin \theta \cdot x), 0 \leq \theta \leq \pi, x \in S^{n-1} \}.
\]
Then
\[
dy = (d \cos \theta, d(\sin \theta \cdot x))
\]
\[
= (-\sin \theta, \cos \theta \cdot x)d\theta + (0, \sin \theta dx)
\]
\[
= (-\sin \theta, \cos \theta \cdot x)d\theta + \sum_{i=1}^{n-1} (0, e_i) \sin \theta \cdot \omega_i.
\]
For convenience we write
\[
e_0 = (-\sin \theta, \cos \theta \cdot x), e_i = (0, e_i), i = 1, 2, \ldots, n-1.
\]
It is clear to see that \( e_0, e_1, \ldots, e_{n-1} \) provides a local orthonormal frame field of \( TS^n \). In addition, define
\[
\phi_j = \langle dy, e_j \rangle, j = 0, 1, 2, \ldots, n-1.
\]
We are led to
\[
\phi_0 = d\theta, \phi_j = \sin \theta \cdot \omega_i (i = 1, 2, \ldots, n-1).
\]
Further, we find
\[ d\tilde{y} = df_k(y) \]
\[ = d(\cos k\theta, \sin k\theta \cdot x) \]
\[ = (-\sin k\theta, \cos k\theta)kd\theta + (0, e_i) \sin k\theta \cdot \omega_i. \]

It follows from the definition \( d\tilde{y} = \sum \psi_j \epsilon_j \) that
\[ \psi_0 = kd\theta, \psi_i = \sin k\theta \cdot \omega_i \quad (i = 1, 2, \ldots, n-1). \]

By the definition of the Brouwer degree, we conclude that
\[ \deg f_k \cdot V(S^n) = \int_{S^n} f_k^* d\tilde{V} \]
\[ = \int_{S^n} \psi_0 \wedge \cdots \wedge \psi_{n-1} \]
\[ = \int_{S^n} k \sin^{n-1} k\theta d\theta \wedge \omega_1 \wedge \cdots \wedge \omega_{n-1}, \]
and hence
\[ \deg f_k = \frac{k \int_0^\pi \sin^{n-1} k\theta d\theta}{\int_0^\pi \sin^{n-1} \theta d\theta}, \]
from which the required result follows immediately. \( \square \)

We turn to investigate the exact value of the dilatation of gradient maps of isoparametric polynomials. Let \( f \) be a homogeneous polynomial of degree \( g \) on the Euclidean space \( \mathbb{R}^{n+2} \). Recall that \( f \) is called an isoparametric polynomial if it satisfies the following Cartan-Münzner’s differential equations:
\[ |\nabla f|^2 = g^2 |x|^{2g-2}, \]
\[ \Delta f = \frac{1}{2} g^2 (m_2 - m_1) |x|^{g-2}, \]
where \( \nabla f \) and \( \Delta f \) denote the gradient and the Laplacian of \( f \) respectively, and \( m_1, m_2 \) two (possible equal) natural numbers (see, for example, [PT1], [CR]). E. Cartan has solved completely the classification problem for isoparametric polynomials for \( g = 1, 2 \) or 3. By using cohomological arguments, Münzner obtained the splendid result that the number \( g \) of distinct principal curvatures of level sets can be only 1, 2, 3, 4 or 6.

Note that the gradient map \( \Phi \), defined by \( \Phi = \frac{1}{k} \nabla f \), is a map from \( \mathbb{R}^{n+2} \) to \( \mathbb{R}^{n+2} \). Moreover, every component of \( \Phi \) is a homogeneous polynomial of degree \( g - 1 \). Since \( f \) satisfies the Cartan-Münzner’s differential equations, the restriction of \( \Phi \) on the unit sphere produces a \((g - 1)\)-form from \( S^{n+1} \) to \( S^{n+1} \). In [PT1], the Brouwer degree of \( \Phi \) was calculated. An analysis of
the gradient map of isoparametric polynomial, given in [PT1], will imply the following:

**Proposition 2.5.** Let $\Phi = \frac{1}{k} \nabla f$ and $f$ be an isoparametric polynomial of degree $g$. Then $\delta(\Phi) = g - 1$.

**Proof.** The conclusion follows obviously from (1.8) and (1.18) in [PT1]. $\square$

To conclude this section we wish to pose a natural question: Given $n,k \in \mathbb{Z}^+$, how many equivalence classes (under orthogonal transformations) of polynomial maps $f: S^n \to S^n$ of degree $k$ with $\delta(f) = k$ are there?

It is well-known that classifying isoparametric polynomials amounts classifying their gradient maps. Let $f_1$ and $f_2$ be the isoparametric polynomials of degree 4 with multiplicities $(m_1, m_2) = (1, 3)$ and $(2, 2)$, respectively. Obviously $f_1$ is not equivalent to $f_2$, hence the gradient maps $\frac{1}{4} \nabla f_1 : S^9 \to S^9$ and $\frac{1}{4} \nabla f_2 : S^9 \to S^9$ are not equivalent, although they have equal dilation 3.

For spheres of lower dimensions, we are able to show the following uniqueness result:

**Proposition 2.6.** Let $f: S^2 \to S^2$ be a 3-form. If $\delta(f) = 3$, then by suitable orthogonal transformations, $f$ coincides with the “standard” $f_3$ which is given in Theorem 2.4.

The proof is entirely elementary, but full of precise analysis. We wish to write down the detailed proof in a future paper.

### 3. Differentiable maps.

This section is devoted to the proofs of Propositions 1.2 and 1.3 and Theorem 1.4. We begin by stating the following approximation result.

**Lemma 3.1.** Suppose $a < c < b$, and $0 \leq k_1 < k_2$. Let $f$ be a piecewise linear function defined by

$$f(x) = \begin{cases} k_1 x + d, & \text{if } x \in (a, c); \\ k_2 (x - c) + (k_1 c + d), & \text{if } x \in [c, b). \end{cases}$$

Then for any $\epsilon > 0$ there exist $\delta > 0$ and a differentiable function $\tilde{f} \in C^1(a, b)$ such that:

1. $\tilde{f}(x) = f(x)$ for $x \in (a, c - \delta] \cup [c + \delta, b)$;
2. $|\tilde{f}(x) - f(x)| < \epsilon$ and $k_1 \leq \tilde{f}'(x) \leq k_2$ for $x \in (c - \delta, c + \delta)$.

**Proof.** Obvious. $\square$

We note that this lemma can be generalized to the general case where $f$ is a piecewise linear function of multiple pieces.
Proof of Proposition 1.2. To prove $\delta(2, 4) \leq 2\sqrt{2}$, it suffices to construct a $C^1$-map $f : S^2 \to S^2$ of the Brouwer degree 4, with $\delta(f) \leq 2\sqrt{2} + \epsilon$ for any given small $\epsilon > 0$. The proof is divided into three steps.

(i) Write the two dimensional sphere $S^2$ as

$$S^2 = \{(\cos \theta, \sin \theta \cdot e^{i\phi}) \mid 0 \leq \theta \leq \pi, 0 \leq \phi \leq 2\pi\}.$$ 

$S^2$ is then divided into two parts

$$S_1 = \left\{ 0 \leq \theta \leq \frac{\pi}{2} \right\} \text{ and } S_2 = \left\{ \frac{\pi}{2} \leq \theta \leq \pi \right\}$$

whose intersection is precisely the equator.

In this way we define $F_1 : S_1 \to S^2$ by

$$F_1(\cos \theta, \sin \theta \cdot e^{i\phi}) = (\cos 2\theta, \sin 2\theta \cdot e^{i2\phi}),$$

and $F_2 : S_2 \to S^2$ by the composition: $F_2 = F_1 \circ j$, where $j : S^2 \to S^2$ is defined by sending $(x_1, x_2, x_3) \in S^2 \subset \mathbb{R}^3$ to $(-x_1, x_2, -x_3)$. We finally get a map $F : S^2 \to S^2$ simply defined by $F |_{S_1} = F_1$ and $F |_{S_2} = F_2$. Since the diffeomorphism $j$ is orientation preserving, $F$ has the Brouwer degree 4. Intuitively, the map $F$ wraps the source $S^2$ around the target $S^2$ four times. We remark that $F$ is not differentiable.

(ii) Define $F_1 : S_1 \to S^2$ by

$$\tilde{F}_1(\cos \theta, \sin \theta \cdot e^{i\phi}) = (\cos \alpha(\theta), \sin \alpha(\theta) \cdot e^{i2\phi}),$$

where $\alpha : [0, \frac{\pi}{2}] \to [0, \pi]$ is a function given by

$$\alpha(\theta) = \begin{cases} 
0, & \theta \in [0, \delta\pi]; \\
\frac{5}{16\delta} \cdot (x - \delta\pi), & \theta \in [\delta\pi, \frac{\pi}{4}]; \\
\frac{11}{16\delta} \cdot (x - \frac{\pi}{2} + \delta\pi) + \pi, & \theta \in [\frac{\pi}{4}, \frac{\pi}{2} - \delta\pi]; \\
\pi, & \theta \in [\frac{\pi}{2} - \delta\pi, \frac{\pi}{2}].
\end{cases}$$

Here $\delta > 0$ is sufficiently small. In fact, $\alpha$ is a piecewise linear function joining the points $(0, 0), (\delta\pi, 0), (\frac{\pi}{4}, \frac{5\pi}{16}), (\frac{\pi}{2} - \delta\pi, \pi), (\frac{\pi}{2}, \pi)$ together.

Again, define $\tilde{F}_2 : S_2 \to S^2$ by $\tilde{F}_2 = \tilde{F}_1 \circ j$. Furthermore, we have $\tilde{F} : S^2 \to S^2$ by simply defining

$$\tilde{F} |_{S_1} = \tilde{F}_1 \quad \text{and} \quad \tilde{F} |_{S_2} = \tilde{F}_2.$$ 

It is easily seen that $\tilde{F}$ is homotopic to $F$ whose Brouwer degree is equal to 4. Hence $\deg \tilde{F} = 4$. On the other hand, it is evident to verify

$$\sup_{0 \leq \theta \leq \frac{\pi}{2}, \theta \neq 0, \delta\pi, \frac{\pi}{2} - \delta\pi, \frac{\pi}{2}} \left\{ 2 \left| \frac{\sin \alpha(\theta)}{\sin \theta} \right|, |\alpha'(\theta)| \right\} \leq 2\sqrt{2}$$

which implies that the dilatation of $\tilde{F}$ on the domain of smooth points is not bigger than $2\sqrt{2}$. It should be remarked that the particular $\alpha$ is not the
optimal choice. In fact, one can improve the estimate $2\sqrt{2}$ by altering $\alpha$. However, we don’t know the minimum.

(iii) It remains to uniformly approximate $\widetilde{F}$ by a $C^1$ map $f$ such that $\delta(f) \leq 2\sqrt{2} + \epsilon$. Of course, approximating does not change the homotopy class, and then preserves the Brouwer degree. The desired map is guaranteed by applying a variation of Lemma 3.1.

The proof of Proposition 1.2 is now complete. \hfill \Box

Remark 3.2. By making use of the sphere-packing method of Gromov in [Gr2], one may get that $\delta(2,4) \leq \frac{\pi}{\arcsin \sqrt{\frac{2}{3}}} \approx 3.289$. More precisely, inside the unit sphere $S^2$, there exists an inscribed regular tetrahedron with four vertices: $(0,0,1), \left( \sqrt{\frac{2}{3}}, 0, -\frac{1}{3} \right), \left( -\sqrt{\frac{2}{3}}, \sqrt{\frac{2}{3}}, -\frac{1}{3} \right)$ and $\left( -\sqrt{\frac{2}{3}}, -\sqrt{\frac{2}{3}}, -\frac{1}{3} \right)$. Wrapping the four geodesic spheres around the target $S^2$ by geodesic stretching respectively (see [Gr2] for details), we obtain a map $f : S^2 \to S^2$ such that $\deg f = 4$ and $\delta(f) = \frac{\pi}{\rho}$. By contrast with this conclusion, Hefter [He] stated that $\delta(2,4) < 3$ without detailed proof. On the other side, since $S^2$ is simply connected, and thus any map $f$ of degree different from $\pm 1$ has at least one critical point, it follows from an integration inequality [Gr2, p. 14] that there is no $C^1$ map $f : S^2 \to S^2$ satisfying both $\deg f = 4$ and $\delta(f) = 2$.

To estimate the value of $\delta(3,4)$, we first point out a mistake in [Ol, p. 389] which claimed that $\delta(F) = 2$, where $F$ is given as follows:

\begin{proposition}
As in [Ol], define $F : S^3 \to S^3$ by
\[
F(Z,W) = \left( \frac{Z^2}{\sqrt{|Z|^4 + |W|^4}}, \frac{W^2}{\sqrt{|Z|^4 + |W|^4}} \right)
\]
for $(Z,W) \in S^3 \subset C \times C$. Then $\delta(F) \geq \sqrt{\frac{24}{5}} > 2$.
\end{proposition}

\begin{proof}
Take $x = \left( \frac{\sqrt{3}}{2}, 0, \frac{1}{2}, 0 \right)$ in $S^3$ and $V = (0, 1, 0, 0)$ a tangent vector to $x$. It is not difficult to verify
\[
|d(F)_x(V)|^2 = \frac{24}{5}
\]
as required.

This result is not special. In fact, we can prove that $\delta(f_\alpha) > 2$ for general differentiable join $f_\alpha$. More precisely, we state:
Proposition 3.4. Define \( f_\alpha : S^3 \to S^3 \) by
\[
f_\alpha(\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) = (\cos \alpha(\theta) \cdot e^{i2\phi}, \sin \alpha(\theta) \cdot e^{i2\psi})
\]
for \((\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) \in S^3 \subset C \times C\), where \( \alpha = \alpha(\theta) : [0, \frac{\pi}{2}] \to [0, \frac{\pi}{2}] \) is a function so that \( f_\alpha \) is differentiable. Then \( \deg f_\alpha = 4 \) and \( \delta(f_\alpha) \geq 2 \).

Proof. As mentioned earlier, \( f_\alpha \) is homotopic to \( f : S^3 \to S^3 \) defined by
\[
f(\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) = (\cos \theta \cdot e^{i2\phi}, \sin \theta \cdot e^{i2\psi})
\]
whose Brouwer degree is easily seen to equal to 4. Therefore we obtain \( \deg f_\alpha = 4 \). Suppose now that \( \delta(f_\alpha) \leq 2 \). It follows that \( |df_\alpha(\frac{\partial}{\partial \phi})| \leq 2 |\frac{\partial}{\partial \phi}| \). Since \( \frac{\partial}{\partial \phi} = (\cos \theta \cdot i e^{i\phi}, 0) \) and \( df_\alpha(\frac{\partial}{\partial \phi}) = (2 \cos \alpha(\theta) \cdot i e^{i2\phi}, 0) \), we have therefore that \( |\cos \alpha(\theta)| \leq |\cos \theta| \) for \( \theta \in [0, \frac{\pi}{2}] \), equivalently \( \alpha(\theta) \equiv \theta \). Hence \( f_\alpha \) coincides \( f \), which is not differentiable. This contradicts the assumption. \( \square \)

Remark 3.5. The function \( \alpha(\theta) \) in [Ol, p. 389] is given by
\[
\alpha(\theta) = \arccos \frac{\cos^2 \theta}{\sqrt{\cos^4 \theta + \sin^4 \theta}}.
\]

It seems that one does not know if there is a \( C^1 \) map \( f : S^3 \to S^3 \) of the Brouwer degree 4 and \( \delta(f) = 2 \). However, we have the following:

Lemma 3.6. For any small \( \epsilon > 0 \), there exists a \( C^1 \) map \( f_\alpha : S^3 \to S^3 \) such that \( \deg f_\alpha = 4 \) and \( \delta(f_\alpha) \leq 2 + \epsilon \).

Proof. Define \( f_\alpha : S^3 \to S^3 \) by
\[
f_\alpha(\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) = (\cos \alpha(\theta) \cdot e^{i2\phi}, \sin \alpha(\theta) \cdot e^{i2\psi}),
\]
where \( \alpha = \alpha(\theta) : [0, \frac{\pi}{2}] \to [0, \frac{\pi}{2}] \) satisfies \( \alpha(0) = 0 \) and \( \alpha(\frac{\pi}{2}) = \frac{\pi}{2} \). As mentioned before, the Brouwer degree of \( f_\alpha \) is exactly 4. Assume that \( f_\alpha \) is differentiable. Since \( \frac{\partial}{\partial \phi} = (\cos \theta \cdot i e^{i\phi}, 0), \frac{\partial}{\partial \psi} = (0, \sin \theta \cdot i e^{i\psi}), \frac{\partial}{\partial \theta} = (-\sin \theta \cdot e^{i\phi}, \cos \theta \cdot e^{i\phi}) \), and then
\[
\delta(f_\alpha) = \sup_{0 \leq \theta \leq \frac{\pi}{2}} \left\{ 2 \left| \frac{\cos \alpha(\theta)}{\cos \theta} \right|, 2 \left| \frac{\sin \alpha(\theta)}{\sin \theta} \right|, |\alpha'(\theta)| \right\}.
\]

The inequality \( \delta(f_\alpha) \leq 2 + \epsilon \) is therefore equivalent to the following inequalities:

\((*1)\) \quad \left| \frac{\cos \alpha(\theta)}{\cos \theta} \right| \leq 1 + \frac{\epsilon}{2};

\((*2)\) \quad \left| \frac{\sin \alpha(\theta)}{\sin \theta} \right| \leq 1 + \frac{\epsilon}{2};

\((*3)\) \quad |\alpha'(\theta)| \leq 2 + \epsilon.
Finally, the existence of \( \lim_{\theta \to 0} \frac{\sin \alpha(\theta)}{\sin^2(\theta)} \) and \( \lim_{\theta \to \frac{\pi}{2}} \frac{\cos \alpha(\theta)}{\cos^2(\theta)} \) guarantee that \( f \) is a \( C^1 \) map. We may construct such a satisfactory function \( \alpha = \alpha(\theta) \) in two steps.

(1) For odd \( n \), let \( \beta = \beta(t) = \frac{1}{2n} t^n (1-t^{2n}), t \in [-1,1] \). Then, the following properties are easily verified for \( t \in [-1,1] \):

- (1a) \( \beta(\pm 1) = \beta(0) = 0 \);
- (1b) \( \beta'(\pm 1) = -1, |\beta'(t)| \leq 1 \), and \( \beta'(t) = -1 \) if and only if \( t = \pm 1 \);
- (1c) \( |\beta(t)| \leq \frac{1}{2n} \);
- (1d) \( -1 \leq t + \beta(t) \leq 1 \).

(2) Furthermore, let \( \alpha = \alpha(\theta) = \theta + \frac{n}{4} \beta(\frac{4}{n} \theta - 1) \) for \( \theta \in [0, \frac{\pi}{2}] \). Then it follows from (1) that

- (2a) \( \alpha(0) = 0, \alpha(\frac{n}{4}) = \frac{n}{4} \), \( \alpha(\frac{2n}{4}) = \frac{2n}{4} \);
- (2b) \( \alpha'(0) = 0, \alpha'(\frac{n}{4}) = 0 \), and \( \alpha'(\theta) \leq 2 \);
- (2c) \( \frac{\pi}{4} \beta'(\frac{4}{n} \theta - 1) \leq \frac{\pi}{8n} \);
- (2d) \( 0 \leq \alpha(\theta) \leq \theta \) if \( \theta \in [0, \frac{n}{4}] \), and \( \theta \leq \alpha(\theta) \leq \frac{2n}{4} \) if \( \theta \in [\frac{n}{4}, \frac{2n}{4}] \).

Summarizing the above arguments, we have finally

\[
\begin{align*}
|\sin \alpha(\theta)| - \sin \theta \leq 1 + \sin \frac{\pi}{8n}; \\
|\cos \alpha(\theta)| - \cos \theta \leq 1 + \sin \frac{\pi}{8n}; \\
\left| \frac{\alpha'(\theta)}{\sin \theta} \right| \leq 2; \\
\lim_{\theta \to 0} \frac{\sin \alpha(\theta)}{\sin^2 \theta} \quad \text{and} \quad \lim_{\theta \to \frac{\pi}{2}} \frac{\cos \alpha(\theta)}{\cos^2 \theta} \quad \text{exist.}
\end{align*}
\]

These imply that \((*) \), \((*) \) and \((*) \) hold, if \( n \) is sufficiently large. We get the desired function \( \alpha = \alpha(\theta) \), and hence the proof of the lemma.

Proposition 1.3 is an immediate consequence.

We should mention that Gromov in [Gr2] gave a map \( g : S^3 \to S^3 \) by
\[
g(\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) = (\cos \theta \cdot e^{i2\phi}, \sin \theta \cdot e^{i2\psi})
\]
whose dilatation is equal to 2 and the Brouwer degree is just 4. However, this map \( g \) is only continuous, not differentiable. Note that for continuous map, the dilatation is defined by Lipschitz constant instead of \( \sup \{|df(V)||V| = 1\} \) (for more details, see [He]). Indeed, Lemma 3.6 provides explicitly a smooth approximation to the map \( g \) of Gromov.

The last part of this section will be devoted to:

Proof of Theorem 1.4. Define \( f : S^3 \to S^3 \) by
\[
f(\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) = (\cos \theta \cdot e^{i2\phi}, \sin \theta \cdot e^{i2\psi})
\]
and denote by \( \tau \) the homotopy class of \( f \). It is easy to see that \( \deg f = 2 \), and it follows from [Ol, p. 389] that \( \delta(\tau) = 2 \).

Next, define \( \pi_\alpha : S^3 \to S^2 \) by

\[
\pi_\alpha(\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) = (\cos \alpha(\theta), \sin \alpha(\theta) \cdot e^{i(\phi+\psi)})
\]

where \( \alpha : [0, \frac{\pi}{2}] \to [0, \pi] \) satifies \( \alpha(0) = 0, \alpha(\frac{\pi}{2}) = \pi \). Note that if \( \alpha(\theta) \equiv 2\theta \), then \( f_\alpha \) is precisely the Hopf fibration \( \pi : S^3 \to S^2 \). Moreover, any such map \( \pi_\alpha \) (called \( \alpha \)-Hopf construction) is homotopic to the Hopf fibration (see, for example, [PT2] and [Ba]). It follows from Theorem 1 in [La, p. 433] that \( \delta(\sigma) = 2 \) for \( \sigma = [\pi_\alpha] = [\pi] \in \pi_3 S^2 \).

Now observe that the composition \( \pi_\alpha \circ f : S^3 \to S^2 \) sends \( (\cos \theta \cdot e^{i\phi}, \sin \theta \cdot e^{i\psi}) \) in \( S^3 \) to \( (\cos \alpha(\theta), \sin \alpha(\theta) \cdot e^{i(2\phi+\psi)}) \) in \( S^2 \). Using a similar argument as before, we get

\[
\delta(\pi_\alpha \circ f \mid_{0 < \theta < \frac{\pi}{2}}) = \sup_{0 < \theta < \frac{\pi}{2}} \left\{ \left| a_1'(\alpha'(\theta)) \right|^2 + \left( a_2 - \frac{2\sin \alpha(\theta)}{\cos \theta} \right)^2 + \frac{\sin \alpha(\theta)}{\sin \theta} \right\} \left| a_1^2 + a_2^2 + a_3^2 = 1 \right|.
\]

We are then left to construct a suitable function \( \alpha = \alpha(\theta) \) such that the composition \( \pi \circ f \) is differentiable and the inequality \( \delta(\pi \circ f) < 3 \) holds.

Let \( \delta > 0 \) be sufficiently small, and define \( \alpha = \alpha(\theta) \) by

\[
\alpha(\theta) = \begin{cases} 
0, & \theta \in [0, \delta \pi]; \\
(\theta - \delta \pi) \cdot \frac{3-16\delta}{1-4\delta}, & \theta \in [\delta \pi, \frac{\pi}{4}]; \\
(\theta - \frac{\pi}{2} + \delta \pi) \cdot \frac{1+16\delta}{1-4\delta} + \pi, & \theta \in \left[\frac{\pi}{4}, \frac{\pi}{2} - \delta \pi\right]; \\
\pi, & \theta \in \left[\frac{\pi}{2} - \delta \pi, \frac{\pi}{2}\right].
\end{cases}
\]

In other words, \( \alpha \) is a piecewise linear function joining the points \((0, 0), (\delta \pi, 0), (\frac{\pi}{2}, \frac{\pi}{2} - 4\delta \pi), (\frac{\pi}{2} - \delta \pi, \pi), (\frac{\pi}{2}, \pi)\) together. It is evident to see that the function \( \alpha \) has slopes \( 0, \frac{3-16\delta}{1-4\delta} (\equiv 3), \frac{1+16\delta}{1-4\delta} (\equiv \frac{3}{2}) \) and \( 0 \) on the four intervals, respectively. It is straightforward to verify that

\[
\sin^2 \alpha(\theta) \leq \frac{3}{2} \sin 2\theta;
\]

\[
\sup_{\theta \in (0, \frac{\pi}{2}), \theta \neq \delta \pi, \frac{\pi}{2} - \delta \pi} \left\{ \left| \alpha'(\theta) \right| \left| \frac{\sin \alpha(\theta)}{\cos \theta} \right|, \left| \frac{\sin \alpha(\theta)}{\sin \theta} \right| \right\} < 3.
\]

Hence \( \delta(\pi_\alpha \circ f \mid_{0 < \theta < \pi/2}) \leq \sqrt{15} \). It remains to make a smooth approximation to the map \( \pi_\alpha \circ f \). Applying a variation of Lemma 3.1 will produce the desired map, which lies in the homotopy class \( \sigma \circ \tau \in \pi_3 S^2 \) with dilatation less than \( 3 \). The proof is now complete.

To conclude this section we wish to point out that Theorem 1.4 still holds for \( \sigma \in \pi_7 S^4 \) (or \( \pi_{15} S^8 \)) the Hopf class and \( \tau \in \pi_7 S^7 \) (or \( \pi_{15} S^{15} \)) of the Brouwer degree 2. The proof is entirely analogous to that for \( \pi_3 S^2 \).
Acknowledgements. The authors are very grateful to Professor S.S. Chern and Professor M. Gromov for their encouragements, to Ms. Guan for careful typing, and to referee for nice comments. The second named author thanks Professor Yu J.T. for inviting to visit Hong Kong University. The research was supported by The Hong Kong Qiu-Shi Foundation, The National Natural Science Foundation of China and Academia Sinica, as well as the Education Foundation of Tsinghua University.

References


Received August 11, 2000 and revised February 18, 2001.
E-mail address: pengck@sun.ihep.ac.cn

Department of Mathematical Sciences
Tsinghua University
Beijing 100084, China
E-mail address: zztang@mx.cei.gov.cn
A BIRATIONAL INVARIANT FOR ALGEBRAIC GROUP ACTIONS

Z. REICHSTEIN AND B. YOSSIN

We construct a birational invariant for certain algebraic group actions. We use this invariant to classify linear representations of finite abelian groups up to birational equivalence, thus answering, in a special case, a question of E.B. Vinberg and giving a family of counterexamples to a related conjecture of P.I. Katsylo. We also give a new proof of a theorem of M. Lorenz on birational equivalence of quantum tori (in a slightly expanded form) by applying our invariant in the setting of PGL_n-varieties.

1. Introduction.

Let $G$ be an algebraic group and let $X$ be a smooth projective $G$-variety (i.e., an algebraic variety with a $G$-action) defined over an algebraically closed base field of characteristic zero. It is shown in [RY1] that for each finite abelian subgroup $H$, the presence of an $H$-fixed points is a birational invariant of $X$ as a $G$-variety. In other words, if $X$ and $Y$ are birationally isomorphic smooth projective $G$-varieties and $H$ is a finite abelian subgroup of $G$ then $X^H \neq \emptyset$ iff $Y^H \neq \emptyset$. (Here as usual, $X^H$ denotes the subvariety of $H$-fixed points of $X$.) Note that only nontoral finite abelian subgroups $H$ are of interest in this setting; if $H$ lies in a torus $T$ of $G$ then $X^T$ (and thus $X^H$) can never be empty by the Borel Fixed Point theorem. In [RY1], [RY2] and [RY3] we used $H$-fixed points for nontoral finite abelian subgroup $H$ of $G$ to study the geometry of $G$-varieties (their essential dimensions, splitting degrees, etc.) and properties of related algebraic objects (field extensions, division algebras, octonion algebras, etc.).

In this paper we will associate (under additional assumptions on $X$, $G$ and $H$) a more subtle invariant $i(X, x, H)$ to a point $x \in X^H$; the precise definition is given in Section 4. Our main result about $i(X, x, H)$ is stated below.

Recall that the rank of a finite abelian group $H$ is the minimal number of generators of $H$ (see Section 2) and that a $G$-variety $X$ is called generically free if $\text{Stab}(x) = \{1\}$ for $x$ in general position in $X$ (see Section 3).

**Theorem 1.1.** Let $G$ be an algebraic group of dimension $d$, $H$ be a finite abelian subgroup of $G$ of rank $r$, and $X, Y$ be birationally isomorphic smooth
projective irreducible generically free $G$-varieties of dimension $d+r$. Assume that $\text{Stab}(x)$ is finite for every $x \in X^H$ and $\text{Stab}(y)$ is finite for every $y \in Y^H$. Then for every $x \in X^H$ there exists a $y \in Y^H$ such that $i(Y, y, H) = i(X, x, H)$.

Informally speaking, the presence of $H$-fixed points $x$ with a prescribed value $i(X, x, H)$ (on a suitable model) is a birational invariant of $X$ as a $G$-variety. Our proof of Theorem 1.1, presented in Section 6, relies on canonical resolution of singularities. Note that Theorem 1.1 remains valid even if $X$ and $Y$ are not assumed to be irreducible; see Remark 6.4.

We give two applications of Theorem 1.1.

A birational classification of linear representations. For our first application recall that by the no-name lemma any two generically free linear representations of a given algebraic group $G$ are stably isomorphic as $G$-varieties; see, e.g., [P, 1.5.3]. Thus it is natural to try to classify such representations up to birational isomorphism. This problem was proposed by E. B. Vinberg [PV2, pp. 494-496]; see also [P, 1.5.1]. P. I. Katsylo has subsequently stated the following conjecture:

**Conjecture 1.2 ([K]; see also [P, 1.5.10]).** Let $V$ and $W$ be generically free linear representations of an algebraic group $G$. If $\dim(V) = \dim(W)$ then $V$ and $W$ are birationally isomorphic as $G$-varieties.

In this paper we will establish the following birational classification of faithful linear representations of a diagonalizable group. Recall that every diagonalizable group $G$ can be uniquely written in the form

$$G = \mathbb{G}_m(n_1) \times \cdots \times \mathbb{G}_m(n_r)$$

such that

$$\mathbb{G}_m(n_1) \subset \cdots \subset \mathbb{G}_m(n_r)$$

and each $n_i = 0$ or $\geq 2$,

where $\mathbb{G}_m = k^* = \mathbb{G}_m(0)$ denotes the 1-dimensional torus, $\mathbb{G}_m(n) \simeq \mathbb{Z}/n\mathbb{Z}$ is the $n$-torsion subgroup of $\mathbb{G}_m$, and $\mathbb{G}_m(a) \subset \mathbb{G}_m(b)$ iff either $a$ divides $b$ or $a = b = 0$; see, e.g., [Bo, Proposition III.8.7] and [DF, Theorem 5.2.3]. Recall also that a representation of a diagonalizable group is faithful iff it is generically free ([PV1, Proposition 7.2]) and that any such representation has dimension $\geq r$ ([Bo, Proposition III.8.2(d)]).

**Theorem 1.3.** Let $G$ be a diagonalizable group, as in (1.1).

(a) If $d \geq r + 1$ then any two faithful $d$-dimensional linear representations of $H$ are birationally equivalent.

(b) If $n_1 = 0$ or $2$ then any two faithful $r$-dimensional linear representations of $H$ are birationally equivalent.

(c) If $n_1 \geq 3$ then $H$ has exactly $\phi(n_1)/2$ birational equivalence classes of faithful $r$-dimensional representations. Here $\phi$ denotes the Euler $\phi$-function.
In particular, Conjecture 1.2 fails for $G$ if and only if $n_1 = 5$ or $\geq 7$.

The birational equivalence classes of faithful linear representations of $G$ are explicitly described in Theorem 7.1. Later in Section 7 we will show that Conjecture 1.2 also fails for some nonabelian finite groups $G$. On the other hand, we remark that P. I. Katsylo [K] proved Conjecture 1.2 for $G = SL_2$, $G = PGL_2$ and $G = S_n$ ($n \leq 4$), and that many interesting cases remain open, including $G = S_n$ ($n \geq 5$) and $G$ = arbitrary connected semisimple group.

Birational equivalence of quantum tori. Our second application is based on the fact that birational isomorphism classes of generically free $PGL_n$-varieties are in natural correspondence with central simple algebras of degree $n$; see, e.g., [Se1, X.5] or [RY2, Section 3]. Thus Theorem 1.1 (with $G = PGL_n$) will sometimes allow us to prove that certain division algebras are not isomorphic over $k$.

Let $\omega_1, \ldots, \omega_r$ be roots of unity and let $R(\omega_1, \ldots, \omega_r)$ be the associative $k$-algebra $k(x_1^\pm 1, \ldots, x_r^\pm 1)$, where the variables $x_1, \ldots, x_r$ are subject to relations $x_{2i-1}x_{2i} = \omega_i x_{2i}x_{2i-1}$ for $i = 1, \ldots, r$ and $x_ax_b = x_bx_a$ for all other pairs $x_a, x_b$. Denote the algebra of quotients of $R(\omega_1, \ldots, \omega_r)$ by $Q(\omega_1, \ldots, \omega_r)$. Note that $Q(\omega_1, \ldots, \omega_r)$ is obtained from $R(\omega_1, \ldots, \omega_r)$ by adjoining the inverses of all central elements and that $Q(\omega_1, \ldots, \omega_r)$ is a finite-dimensional division algebra (in fact, it is a tensor product of symbol algebras).

M. Lorenz [Lo1, Proposition 1.3] showed that $Q(\omega)$ and $Q(\omega')$ are isomorphic as $k$-algebras if and only if $\omega' = \omega^{\pm 1}$. In Section 8 we will give a geometric proof of the following variant of this result.

**Theorem 1.4.** Suppose $\omega_i$ is a primitive $n_i$th root of unity, $n_i$ divides $n_{i+1}$ for $i = 1, \ldots, r - 1$, $n_1 \geq 2$, and $(m_i,n_i) = 1$. Then $Q(\omega_1, \ldots, \omega_r)$ and $Q(\omega_1^{m_1}, \ldots, \omega_r^{m_r})$ are isomorphic as $k$-algebras if and only if $m_1 \cdots m_r \equiv \pm 1$ (mod $n_1$).

The centers of the algebras $Q(\omega_1, \ldots, \omega_r)$ and $Q(\omega_1^{m_1}, \ldots, \omega_r^{m_r})$ can be naturally identified with the field $K = k(x_1^{n_1}, \ldots, x_r^{2n_r})$. Note that these algebras may be $k$-isomorphic but not $K$-isomorphic, i.e., not Brauer equivalent. More precisely, $Q(\omega_1, \ldots, \omega_r)$ and $Q(\omega_1^{m_1}, \ldots, \omega_r^{m_r})$ are Brauer equivalent iff $m_i \equiv 1$ (mod $n_i$) for every $i = 1, \ldots, r$.

It is natural to think of $R(\omega_1, \ldots, \omega_r)$ and $Q(\omega_1, \ldots, \omega_r)$ as, respectively, the “coordinate ring” and the “function field” of a quantum torus. Using this terminology, one may view Theorem 1.4 as a result about birational isomorphism classes of quantum tori.

Finally we remark that M. Lorenz has communicated to us a proof of Theorem 1.4 based on the techniques of [Lo1] and [Lo2]. His argument works in arbitrary characteristic.
Acknowledgements. We are grateful to P.I. Katsylo, M. Lorenz and V.L. Popov for helpful communications related to the subject matter of this paper.


2. Linear algebra in abelian groups.

Recall that any finitely generated abelian group \((A, +)\) can be written in the form
\[
A \simeq \mathbb{Z}/n_1\mathbb{Z} \times \cdots \times \mathbb{Z}/n_r\mathbb{Z},
\]
where each \(n_i = 0\) or \(\geq 2\) (2.1) and \(n_{i+1} \in n_i\mathbb{Z}\) for every \(i = 1, \ldots, r\);
see, e.g., [DF, Theorem 5.2.3]. Here \(r\) and \(n_1, \ldots, n_r\) are uniquely determined by the isomorphism type of \(A\). We shall refer to the integer \(r\) as the rank of \(A\); equivalently, the rank of \(A\) equals the minimal possible number of generators of \(A\).

Recall that if \(B\) is an abelian group then the dual group \(B^*\) is defined as \(\text{Hom}(B, \mathbb{Q}/\mathbb{Z})\); we will often identify \(\mathbb{Q}/\mathbb{Z}\) with the multiplicative group of roots of unity in \(k\). The finitely generated group \(A\) of (2.1) and the diagonalizable group \(G\) of (1.1) are dual to each other. The rank of a diagonalizable group \(G\) is defined as the rank of the finitely generated group \(G^*\) (in particular, the group \(G\) of (1.1) has rank \(r\)). Note that this is consistent with the way we defined rank for a finitely generated group: Indeed, if \(A\) is both diagonalizable and finitely generated, i.e., is finite abelian, then \(A\) and \(A^*\) are isomorphic, so that their ranks coincide.

Skew-symmetric powers. We will write \(\bigwedge^d(A)\) for the \(d\)-th skew-symmetric power of \(A\), viewed as a \(\mathbb{Z}\)-module.

The proof of the following lemma is elementary; we leave it as an exercise for the reader.

Lemma 2.1. Let \(A\) be a finitely generated abelian group as in (2.1). Then

(a) \(\bigwedge^1(A) \simeq \mathbb{Z}/n_1\mathbb{Z}\).

(b) \(\bigwedge^d(A) = (0)\), if \(d \geq r + 1\).

Definition 2.2. Let \(A\) be a finite abelian group. Let \(\omega: A \times A \rightarrow \mathbb{Q}/\mathbb{Z}\) be a \(\mathbb{Z}\)-bilinear form. As usual, we shall say that

(a) \(\omega\) is alternating if \(\omega(a, a) = 0\) for every \(a \in A\),

(b) \(\omega\) is nondegenerate if \(\omega(a, \cdot)\) is not identically zero for any \(a \in A \setminus \{0\}\),

(c) \(\omega\) is symplectic if it is both alternating and nondegenerate.

Lemma 2.3. Let \(A\) be a finite abelian group of rank \(2r\), \(\omega\) be a symplectic form of \(A\), and \(\psi\) be an \(\omega\)-preserving automorphism \(A \rightarrow A\). Then

(a) \(\bigwedge^{2r} \psi\) is the trivial automorphism of \(\bigwedge^{2r}(A)\) and
(b) \( \wedge^{2r} \psi^* \) is the trivial automorphism of \( \Lambda^{2r}(A^*) \).

**Proof.** (a) It is well-known that \( A \) can be written in the form \( A = A_0 \oplus A_0^* \) such that

\[
\omega((a, a^*), (b, b^*)) = a^*(b) - b^*(a)
\]

for any \( a, b \in A_0 \) and \( a^*, b^* \in A_0^* \); see, e.g., [TA, Theorem 4.1]. Write \( A_0 \) as \( \mathbb{Z}/n_1 \mathbb{Z} \times \cdots \times \mathbb{Z}/n_r \mathbb{Z} \), where \( n_i \) divides \( n_{i+1} \) for \( i = 1, \ldots, r - 1 \) and \( n_1 \geq 2 \).

Let \( e_i \in A_0 \) be a generator of the of the \( i \)th factor, and let \( f_i \in A_0^* \) be given by \( f_i : A_0 \to \mathbb{Z}/n_i \mathbb{Z} \to \mathbb{Q}/\mathbb{Z} \), where the first map the projection to the \( i \)th factor, and the second map takes \( e_i \) to \( 1/n_i \). Then every \( a \in A \) can be written in the form \( a = \sum_{i=1}^n (\alpha_{2i-1} e_i + \alpha_{2i} f_i) \), where \( \alpha_{2i-1}, \alpha_{2i} \in \mathbb{Z} \), and (2.2) translates into

\[
\omega \left[ \sum_{i=1}^n (\alpha_{2i-1} e_i + \alpha_{2i} f_i), \sum_{i=1}^n (\beta_{2i-1} e_i + \beta_{2i} f_i) \right] = \frac{1}{n_i} (\alpha_{2i} \beta_{2i-1} - \alpha_{2i-1} \beta_{2i}).
\]

Suppose

\[
\begin{align*}
\psi(e_1) &= c_{11} e_1 + c_{12} f_1 + \cdots + c_{1,2r-1} e_r + c_{1,2r-1} f_r, \\
\psi(f_1) &= c_{21} e_1 + c_{22} f_1 + \cdots + c_{2,2r-1} e_r + c_{2,2r-1} f_r, \\
\vdots \\
\psi(e_r) &= c_{2r-1,1} e_1 + c_{2r-1,2} f_1 + \cdots + c_{2r-1,2r-1} e_r + c_{2r-1,2r-1} f_r, \\
\psi(f_r) &= c_{2r,1} e_1 + c_{2r,2} f_1 + \cdots + c_{2r,2r-1} e_r + c_{2r,2r} f_r,
\end{align*}
\]

where \( C = (c_{ij})_{i,j=1,\ldots,2r} \in M_n(\mathbb{Z}) \). Since \( \lambda = e_1 \wedge f_1 \wedge \cdots \wedge e_r \wedge f_r \) generates \( \Lambda^{2r}(A^*) \simeq \mathbb{Z}/n_1 \mathbb{Z} \) and \( \wedge^{2r}(\psi)(\lambda) = \det(C) \lambda \), it is sufficient to show that \( \det(C) = 1 \mod n_1 \).

The condition that \( \psi \) preserves \( \omega \) translates into \( C J C^t = J \mod 1 \), where \( C^t \) is the transpose of \( C \) and

\[
J = \begin{pmatrix}
0 & 1/n_1 & \cdots & 0 & 0 \\
-1/n_1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 1/n_r \\
0 & 0 & \cdots & -1/n_r & 0
\end{pmatrix}.
\]

In other words,

\[
C J C^t = J + N,
\]

where \( N \) is a skew-symmetric integral matrix. We shall deduce the desired equality, \( \det(C) = 1 \mod n_1 \), by computing the Pfaffian on both sides.
of (2.3). On the one hand

$$\text{Pf}(CJC^d) = \det(C) \text{Pf}(J) = (-1)^r \det(C) \frac{1}{n_1 \ldots n_r};$$

see, e.g., [Lang, XIV, 10, Theorem 7]. On the other hand, suppose \( X = (x_{ij}) \), where \( x_{ji} = -x_{ij} \) for \( 1 \leq i, j \leq 2r \). Then \( \text{Pf}(X) \in \mathbb{Z}[x_{ij} \mid 1 \leq i < j \leq 2r] \) has degree 1 in every \( x_{ij} \), where \( i < j \). (Indeed, \( \det(X) \) has degree 2 in every \( x_{ij} \), and \( \text{Pf}(X)^2 = \det(X) \).) Consequently,

$$\text{Pf}(J + N) = \text{Pf}(J) + \frac{z}{n_2 \ldots n_r} = (-1)^r \frac{1}{n_1 \ldots n_r} + \frac{z}{n_2 \ldots n_r},$$

where \( z \) is an integer. (Here we used the fact that \( n_i \) divides \( n_{i+1} \) for every \( i = 1, \ldots, r - 1 \).) Putting (2.3), (2.4) and (2.5) together, we see that

\[ \det(C) = 1 + (-1)^r n_1 z, \quad \text{i.e., } \det(C) = 1 \pmod{n_1}, \]

as claimed.

(b) Let \( i : A \to A^* \) be the isomorphism \( a \mapsto i_a \), where \( i_a(b) = \omega(a, b) \).

It is easy to see that the automorphism \( \psi^* : A^* \to A^* \) preserves the symplectic form \( \omega^* \) on \( A^* \) given by \( \omega^*(a^*, b^*) = \omega(i^{-1}a^*, i^{-1}b^*) \).

Applying part (a) to \( \psi^* \), we obtain the desired result. \( \square \)

**Elementary operations.** Let \( A \) be an abelian group. We will say that two \( d \)-tuples \((a_1, \ldots, a_d)\) and \((b_1, \ldots, b_d) \in A^d\) are related by an **elementary operation** if \((b_1, \ldots, b_d) = (a_1, \ldots, a_{i-1}, a_i + ma_j, a_{i+1}, \ldots, a_d)\) for some \( i \neq j \) and \( m \in \mathbb{Z} \).

**Lemma 2.4.** Let \( a_1, \ldots, a_d \) and \( b_1, \ldots, b_d \) be two sets of generators for an abelian group \( A \). Then \( a = (a_1, \ldots, a_d) \) and \( (b_1, \ldots, b_d) \) are related by a sequence of elementary operations if and only if \( a_1 \wedge \cdots \wedge a_d = b_1 \wedge \cdots \wedge b_d \) in \( \Lambda^d(A) \).

**Proof.** It is clear that if \((a_1, \ldots, a_d)\) and \((b_1, \ldots, b_d)\) are related by a sequence of elementary operations then \( a_1 \wedge \cdots \wedge a_d = b_1 \wedge \cdots \wedge b_d \). We will prove the converse by induction on \( d \).

If \( d = 1 \) there is nothing to prove, since \( \Lambda^1(A) = A \). For the induction step, assume \( d \geq 2 \) and \( A \simeq (\mathbb{Z}/n_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/n_r\mathbb{Z}) \) as in (2.1). Here \( r = \text{rank}(A) \leq d \), since we are assuming \( A \) is generated by \( d \) elements.

A \( d \)-tuple of generators \((a_1, \ldots, a_d)\) of \( A \) can now be represented by a \( d \times r \)-matrix \( a = (a_{ij}) \) whose \( i \)th row is \( a_i \). Elements of the \( j \)th column of this matrix lie in \( \mathbb{Z}/n_i\mathbb{Z} \). Elementary operation on such matrices amount to adding an integer multiple of the \( j \)th row to the \( i \)th row or some \( i \neq j \). Elementary operations allow us to perform the Euclidean algorithm in the last column of \((a_{ij})\). Since \( a_{1r}, \ldots, a_{dr} \) generate \( \mathbb{Z}/n_r\mathbb{Z} \), after a sequence
of elementary operations, we may assume that

\[
(a_{ij}) = \begin{pmatrix}
0 \\
X \\
\vdots \\
0 \\
\ast & \ast & \ast & 1
\end{pmatrix},
\]

where \(X\) is a \((d-1) \times (r-1)\)-matrix. Since the rows of \((a_{ij})\) generate \(A\), the rows of \(X\) generate \(A = (Z/n_1Z) \times \cdots \times (Z/n_rZ)\). Thus after performing additional elementary operations, we may assume

\[
(a_{ij}) = \begin{pmatrix}
0 \\
X \\
\vdots \\
0 \\
0 & 0 & \ast & 1
\end{pmatrix},
\]

and similarly \((b_{ij}) = \begin{pmatrix}
0 \\
Y \\
\vdots \\
0 \\
0 & 0 & \ast & 1
\end{pmatrix}\).

In other words, we may assume \((a_1, \ldots, a_{d-1})\) and \((b_1, \ldots, b_{d-1})\) are \((d-1)\)-tuples of generators in \(A\) and \(a_d = b_d\) is the generator \(1 \in Z/n_rZ\).

We claim that

\[
a_1 \wedge \cdots \wedge a_{d-1} = b_1 \wedge \cdots \wedge b_{d-1} \text{ in } \bigwedge^{(d-1)}(A).
\]

Indeed, if \(d \geq r + 1\), this is obvious, since \(\bigwedge^{d-1}(A) = (0)\); see Lemma 2.1(b). If \(d = r\) then the isomorphism \(\bigwedge^d(A) \simeq Z/n_1Z\) identifies \(a_1 \wedge \cdots \wedge a_d\) with \(\det(a_{ij}) \pmod{n_1}\), and the isomorphism \(\bigwedge^{d-1}(A) \simeq Z/n_1Z\) identifies \(a_1 \wedge \cdots \wedge a_{d-1}\) with \(\det(X) \pmod{n_1}\). Since \(a_1 \wedge \cdots \wedge a_d = b_1 \wedge \cdots \wedge b_d\), we know that \(\det(a_{ij}) = \det(b_{ij}) \pmod{n_1}\); hence, \(\det(X) = \det(Y) \pmod{n_1}\), and (2.6) follows.

Now by the induction assumption \((a_1, \ldots, a_{d-1})\) and \((b_1, \ldots, b_{d-1})\) are related by a sequence of elementary operations. Since \(a_d = b_d\), so are \((a_1, \ldots, a_d)\) and \((b_1, \ldots, b_d)\), as desired.

\[\square\]

**Corollary 2.5.** Let \((a_1, \ldots, a_d)\) and \((b_1, \ldots, b_d)\) be two sets of generators for an abelian group \(A\). Then the following conditions are equivalent.

(a) There exists a matrix \(N = (n_{ij}) \in \text{GL}_d(Z)\) such that \(b_i = n_{i1}a_1 + \cdots + n_{id}a_d\) for \(i = 1, \ldots, d\).

(b) \(a_1 \wedge \cdots \wedge a_d = \pm b_1 \wedge \cdots \wedge b_d \) in \(\bigwedge^d(A)\).

**Proof.** The implication (a) \(\Rightarrow\) (b) is obvious. To prove the converse, note that we may assume without loss of generality that \(a_1 \wedge \cdots \wedge a_d = b_1 \wedge \cdots \wedge b_d\); indeed, if \(a_1 \wedge \cdots \wedge a_d = -b_1 \wedge \cdots \wedge b_d\) then we can simply replace \((a_1, a_2, \ldots, a_d)\) by \((-a_1, a_2, \ldots, a_d)\). Now Lemma 2.4 says, \((b_1, \ldots, b_d)\) is obtained from \((a_1, \ldots, a_d)\) by a sequence of elementary operations. Each elementary operation relates the two sets of generators as in (a), with \(N = \text{elementary matrix} \in \text{SL}(Z)\). Multiplying these matrices we obtain the desired conclusion. \[\square\]
3. $H$-slices.

In this section we establish several simple properties of slices on a $G$-variety $X$. Note that we do not assume that $X$ is affine or that $G$ is reductive. (Under these assumptions, one can prove quite a bit more than we do here; see [Lu] or [PV1, Section 6].)

**Definition 3.1.** Let $G$ be an algebraic group and $X$ be a $G$-variety. We will call a locally closed subvariety $V$ of $X$ a slice at $x \in V$ if $X$ and $V$ are smooth at $x$ and $T_x(X) = T_x(Gx) \oplus T_x(V)$. (Here $T_x(X)$ denotes the tangent space to $X$ at $x$.) If, moreover, $V$ is invariant under the action of a subgroup $H$ of Stab($x$), we will refer to $V$ as an $H$-slice.

**Remark 3.2.** Note that since $V$ is smooth at $x$, we may replace $V$ by its (unique) irreducible component passing through $x$ and thus assume that $V$ is irreducible.

**Example 3.3.** Let $G$ be an algebraic group, $H$ be an algebraic subgroup of $G$ and $V$ be an $H$-variety. Recall that the homogeneous fiber product $X = G \ast_H V$ is defined as the geometric quotient $X = (G \times V)/H$, where $H$ acts on $G \times V$ by $h(g,v) = (gh^{-1}, hv)$. (This geometric quotient exists under mild assumptions on $V$; in particular, it exists whenever $V$ is quasi-projective; see [PV1, Theorem 4.19].) Note that $X = G \ast_H V$ is naturally a $G$-variety, where $G$ acts by left multiplication on the first factor; the details of this construction can be found in [PV1, Section 4.8].

The points of $X$ are in 1-1 correspondence with $H$-orbits in $G \times V$; we shall denote the point $x \in X$ corresponding to the $H$-orbit of $(g,v)$ in $G \times V$ by $x = [g,v]$. Let $\bar{V}$ be the image of the $H$-equivariant map $V \to X$ given by $v \mapsto [1,v]$. With these notations, $\bar{V}$ is an $H$-slice for $X$ at $x = [1,v]$ for every smooth point $v$ of $V$; see [PV1, Proposition 4.22].

**Lemma 3.4.** Let $G$ be an algebraic group, $X$ be an irreducible $G$-variety, and $V$ be a slice at $x \in X$. Then $GV$ is dense in $X$.

**Proof.** Consider the map $\phi: G \times V \to X$, given by $\phi(g,v) = gv$. The differential $d\phi(1,x): T_1(G) \times T_x(V) \to T_x(X)$ is onto, since its image contains both $T_x(Gx)$ and $T_x(V)$. Consequently, $d\phi$ is onto at a general point of $G \times V$. Thus $\phi$ is dominant, and the lemma follows.

**Lemma 3.5.** Let $G$ be an algebraic group, $H$ be a subgroup, $X$ be a $G$-variety, and $x$ is a smooth $H$-fixed point of $X$. If $H$ is reductive then $X$ has an $H$-slice at $x$.

**Proof.** Let $\mathcal{M}_{x,X}$ be the maximal ideal of the local ring of $X$ at $x$. Consider the natural $H$-equivariant linear maps $\mathcal{M}_{x,X} \to T_x(X)^* \to T_x(Gx)^*$. Since $H$ is reductive these maps have $H$-equivariant splittings as maps of $k$-vector spaces. Thus we can choose a local coordinate system $u_1, \ldots, u_n$
in \( \mathcal{M}_{x,X} \) such that \( \text{Span}_k(u_1, \ldots, u_d) \) is an \( H \)-invariant \( k \)-vector subspace of \( \mathcal{M}_{x,X} \) and \( u_1, \ldots, u_d \) (restricted to \( Gx \)) form a local coordinate system in \( \mathcal{O}_{x,Gx} \). (Here \( n = \dim(X) \) and \( d = \dim(Gx) \).)

Note that \( u_1, \ldots, u_n \) are regular functions in some \( H \)-invariant open neighborhood of \( x \). In this neighborhood a slice \( V \) with desired properties is given by \( u_1 = \cdots = u_d = 0 \). \( \Box \)

Recall that a \( G \)-variety \( X \) is called generically free if \( \text{Stab}(x) = \{1\} \) for \( x \) in general position in \( X \).

**Proposition 3.6.** Let \( G \) be an algebraic group, \( H \) be a reductive subgroup, \( X \) be a generically free \( G \)-variety and \( x \) be a smooth \( H \)-fixed point of \( X \). Then \( H \) acts faithfully on \( T_x(X)/T_x(Gx) \).

**Proof.** Let \( X_0 \) be the unique component of \( X \) passing through \( x \) and \( G_0 \) be the subgroup of all elements of \( G \) that preserve \( X_0 \). Note that \( G_0 \) has finite index in \( G \) and \( H \subset G_0 \). After replacing \( X \) by \( X_0 \) and \( G \) by \( G_0 \), we may assume \( X \) is irreducible.

We now argue by contradiction. Assume the kernel \( K \) of the \( H \)-action on \( T_x(X)/T_x(Gx) \) is nontrivial. Since \( K \) is a normal subgroup of \( H \) and \( H \) is reductive, \( K \) is not unipotent. Hence, we can find a nonidentity element \( g \in K \) of finite order.

By Lemma 3.5 \( X \) has an \( H \)-slice \( V \) at \( x \). Since \( T_x(V) \simeq T_x(X)/T_x(Gx) \) as \( H \)-representations, \( g \) acts trivially on \( T_x(V) \). This implies that \( g \) acts trivially on \( V \); see, e.g., [RY1, Lemma 4.2]. On the other hand, by Lemma 3.4 \( GV \) is dense in \( X \); consequently, for every \( x \in X \) in general position \( \text{Stab}(x) \) contains a conjugate of \( g \). Thus the \( G \)-action on \( X \) is not generically free, contradicting our assumption. \( \Box \)

### 4. Definition and first properties of \( i(X,x,H) \).

Throughout this section we shall make the following assumptions:

- \( G \) algebraic group
- \( H \) finite abelian subgroup of \( G \) of rank \( r \)
- \( X \) \( G \)-variety of dimension \( \dim(G) + r \)
- \( x \) smooth \( H \)-fixed point of \( X \) whose stabilizer is finite.

**Definition 4.1.** The \( H \)-representation on \( T_x(X)/T_x(Gx) \) decomposes as a direct sum of \( r \) characters \( \chi_1, \ldots, \chi_r \in H^* \). We define

\[
i(X,x,H) = \chi_1 \wedge \cdots \wedge \chi_r \in \bigwedge^r(H^*).
\]

Since the collection of characters \( \chi_1, \ldots, \chi_r \) is well-defined, up to reordering, \( i(X,x,H) \) is well-defined in \( \bigwedge^r(H^*) \), up to multiplication by \( -1 \). Thus, properly speaking, \( i(X,x,H) \) should be viewed as an element of the factor set \( \bigwedge^r(H^*)/\sim \), where \( w_1 \sim w_2 \) iff \( w_1 = \pm w_2 \). By abuse of notation we will
sometimes write $i(X, x, H) \in \bigwedge^r (H^*)$; in such cases it should be understood that $i(X, x, H)$ is only defined up to sign.

**Remark 4.2.** It is clear from the definition that if $V$ is an $H$-slice for $X$ at $x$ then $i(X, x, H) = i(V, x, H)$. In particular, in the setting of Example 3.3, if $V$ is an $r$-dimensional $H$-variety, $X = G \ast_H V$, and $v$ is a smooth $H$-fixed point of $V$ then $i(X, [1, v], H) = i([1, v], H) = i(V, v, H)$.

**Remark 4.3.** Let $g$ be an element of the normalizer $N_G(H)$ and let $\phi_g$ be the automorphism of $H$ sending $h$ to $ghg^{-1}$. Then it is easy to see that $i(X, gx, H) = (\bigwedge^r \phi_g^*) \left(i(X, x, H)\right)$, where $\bigwedge^r \phi_g^*$ is the automorphism of $\bigwedge^r (H^*)$ induced by $\phi_g$.

**Example 4.4.** Let $G = H$ be a finite abelian group of rank $r$, $\chi_1, \ldots, \chi_r$ be a generating set for $H^*$, and $V = \bigwedge^r$ be a faithful $r$-dimensional linear representation of $H$, given by

$$h : (v_1, \ldots, v_r) \rightarrow (\chi_1(h)v_1, \ldots, \chi_r(h)v_r).$$

Then the origin $0_V$ is the only $H$-fixed point of $V$, and Definition 4.1 immediately implies $i(V, 0_V, H) = \chi_1 \wedge \cdots \wedge \chi_r$. The extended $H$-action on $V = \mathbb{P}^r$, given by

$$h(v_0 : v_1 : \cdots : v_d) = (v_0 : \chi_1(h)v_1 : \cdots : \chi_d(h)v_d),$$

has exactly $r + 1$ fixed points:

$$x_0 = (1 : 0 : \cdots : 0), \ldots, x_r = (0 : \cdots : 0 : 1).$$

Note that $x_0 = 0_V$. We claim that, up to sign,

$$i(V, x_j, H) = i(V, x_0, H) = \chi_1 \wedge \cdots \wedge \chi_r$$

for $j = 1, \ldots, r$. To prove this claim, say for $j = 1$, note that $v_0/v_1, v_2/v_1, \ldots, v_r/v_1$ form an affine coordinate system on $V$ near $x_1$. The $H$-action is diagonal in these coordinates, and the representation of $H$ on $T_{x_1}(V)$ is the direct sum of the characters $\chi_1^{-1}, \chi_2\chi_1^{-1}, \ldots, \chi_r\chi_1^{-1}$. Consequently, $i(V, x_1, H) = \pm \chi_1^{-1} \wedge \cdots \wedge \chi_r\chi_1^{-1} = \pm \chi_1 \wedge \cdots \wedge \chi_r$, as claimed.

**Lemma 4.5.** Suppose $X$ is a generically free $G$-variety. Then $i(X, x, H)$ generates $\bigwedge^r (H^*)$ as a group.

Note that the statement of the lemma makes sense, even though $i(X, x, H)$ is only defined up to sign: If $a$ generates $\bigwedge^r (H^*)$ then so does $-a$.

**Proof.** By Proposition 3.6 $H$ acts faithfully on $T_x(X)/T_x(Gx)$. Hence, the characters $\chi_1, \ldots, \chi_r$ introduced in Definition 4.1 generate $H^*$ as an abelian group, and the lemma follows. \qed
5. \( i(X,x,H) \) and birational morphisms.

The purpose of this section is to prove the following:

**Theorem 5.1.** Let \( G \) be an algebraic group of dimension \( d \), \( H \) be a finite abelian subgroup of \( G \) of rank \( r \), \( f: X \rightarrow Y \) be birational morphism of irreducible generically free \( G \)-varieties of dimension \( d + r \), \( x \) is a smooth \( H \)-fixed point of \( X \), \( y = f(x) \) is a smooth point of \( Y \), and \( \text{Stab}(y) \) is finite. Then \( i(X,x,H) = i(Y,y,H) \).

**Case I:** \( G = H \). As a first step towards proving Theorem 5.1, we will consider the case where \( G = H \) is a finite abelian group. In this case Theorem 5.1 can be restated as follows.

**Proposition 5.2.** Let \( H \) be a finite abelian group, and \( f: X \rightarrow Y \) be a birational morphism of irreducible generically free \( H \)-varieties of dimension \( r = \text{rank}(H) \). Assume that \( x \) is a smooth \( H \)-fixed point of \( X \) and \( y = f(x) \) is a smooth point of \( Y \). Then \( i(X,x,H) = i(Y,y,H) \).

Before proceeding with the proof of Proposition 5.2, we introduce some background material on the power series ring \( k[[u_1, \ldots, u_r]] \).

Given \( w \in k[[u_1, \ldots, u_r]] \) we shall denote by \( \text{lm}(w) \) the lowest degree monomial in \( u_1, \ldots, u_r \) which enters into \( w(u_1, \ldots, u_r) \) with a nonzero coefficient. Here “lowest degree” refers to a fixed lexicographic monomial order \( > \) given by, say, \( u_1 > \cdots > u_r \).

Suppose \( v_1, \ldots, v_m \) lie in the maximal ideal of \( k[[u_1, \ldots, u_r]] \), i.e., \( \text{lm}(v_i) > 1 \) for any \( i = 1, \ldots, m \). Then we can substitute \( v_1, \ldots, v_m \) into any power series \( p \in k[[z_1, \ldots, z_m]] \); in other words, \( p(v_1, \ldots, v_m) \) is a well-defined element of \( k[[u_1, \ldots, u_r]] \). If \( p = Z \) is a monomial in \( k[[z_1, \ldots, z_m]] \) then clearly

\[
\text{lm}(Z(v_1, \ldots, v_m)) = Z(\text{lm}(v_1), \ldots, \text{lm}(v_m)).
\]

We shall write \( \langle u_1, \ldots, u_r \rangle \) for the group of all Laurent monomials in \( u_1, \ldots, u_r \) (here we allow negative exponents).

**Lemma 5.3.** Suppose \( v_1, \ldots, v_m \in k[[u_1, \ldots, u_r]] \). If \( \text{lm}(v_1), \ldots, \text{lm}(v_m) \) generate a rank \( m \) subgroup \( \Lambda \) in \( \langle u_1, \ldots, u_r \rangle \simeq \mathbb{Z}^r \) then \( \text{lm}(p(v_1, \ldots, v_m)) \in \Lambda \) for any \( p \in k[[z_1, \ldots, z_m]] \).

Note that the conditions of the lemma imply \( m \leq r \); only the case \( m = r \) will be used in the subsequent application.

**Proof.** Suppose \( p(z_1, \ldots, z_m) = \sum c_z Z \), where \( Z \) ranges over all monomials in \( z_1, \ldots, z_m \) with nonnegative exponents and each \( c_z \in K \). By our assumption \( \text{lm}(v_1), \ldots, \text{lm}(v_m) \) are (multiplicatively) linearly independent, i.e.,

\[
Z(\text{lm}(v_1), \ldots, \text{lm}(v_m)) \neq Z'(\text{lm}(v_1), \ldots, \text{lm}(v_m))
\]
for any two distinct monomials \( Z \) and \( Z' \). Suppose \( Z_0(\text{lm}(v_1), \ldots, \text{lm}(v_m)) \) is the lexicographically smallest monomial (in \( u_1, \ldots, u_m \)) of the form \( Z(\text{lm}(v_1), \ldots, \text{lm}(v_m)) \), with \( c_Z \neq 0 \). Then (5.1) tells us that

\[
\text{lm}(Z(v_1, \ldots, v_m)) > \text{lm}(Z_0(v_1, \ldots, v_m)) = Z_0(\text{lm}(v_1), \ldots, \text{lm}(v_m))
\]

for any \( Z \neq Z_0 \) with \( c_Z \neq 0 \). Thus

\[
\text{lm}(p(v_1, \ldots, v_m)) = \text{lm}(Z_0(v_1, \ldots, v_m)) = Z_0(\text{lm}(v_1), \ldots, \text{lm}(v_m)) \in \Lambda,
\]
as claimed. \( \square \)

**Proof of Proposition 5.2.** Diagonalizing the action of \( H \) on the cotangent space \( T^*_x(X) \), we obtain a basis \( \bar{u}_1, \ldots, \bar{u}_r \in T^*_x(X) \) such that \( h\bar{u}_i = \chi_i(h)\bar{u}_i \)

for every \( h \in H \); here \( \chi_1, \ldots, \chi_r \in H^* \). Since the \( k \)-linear map \( \mathcal{M}_{x,X} \rightarrow \mathcal{M}_{x,X}/\mathcal{M}_{x,X}^2 = T^*_x(X) \) has an \( H \)-invariant \( k \)-linear splitting, we can find a local system of parameters \( u_1, \ldots, u_r \in \mathcal{M}_{x,X} \) such that

\[
(5.2) \quad hu_i = \chi_i(h)u_i
\]

for every \( h \in H \) and \( i = 1, \ldots, r \). Similarly, we can find a local coordinate system \( v_1, \ldots, v_r \in \mathcal{M}_{y,Y} \) for \( Y \) at \( y \) and \( \eta_1, \ldots, \eta_r \in H^* \) such that

\[
(5.3) \quad hv_i = \eta_i(h)v_i
\]

for every \( h \in H \) and \( i = 1, \ldots, r \). Clearly \( i(X, x, H) = \chi_1 \wedge \cdots \wedge \chi_r \) and \( i(Y, y, H) = \eta_1 \wedge \cdots \wedge \eta_r \).

We shall identify the elements \( v_1, \ldots, v_r \) with their images in \( O_{x,X} \) under \( f^*: O_{y,Y} \hookrightarrow O_{x,X} \). The \( H \)-action on \( O_{x,X} \) naturally extends to \( k[[u_1, \ldots, u_r]] \); in view of (5.2) the leading term map \( w \mapsto \text{lm}(w) \) is \( H \)-equivariant. Suppose \( \text{lm}(v_i) = u_1^{a_{i1}} \cdots u_r^{a_{ir}} \) for some nonnegative integers \( a_{ij} \). Then (5.2) and (5.3) imply \( \eta_i = \prod_j \chi_j^{a_{ij}} \); and thus, up to sign,

\[
i(Y, y, H) = \eta_1 \wedge \cdots \wedge \eta_r = \det(a_{ij})\chi_1 \wedge \cdots \wedge \chi_r = \pm \det(a_{ij})i(X, x, H).
\]

There are two conclusions we can draw from this formula. First of all, by Lemma 4.5 we know that both \( i(X, x, H) \) and \( i(Y, y, H) \) generate \( \Lambda^\ast(H^*) \); thus \( \det(a_{ij}) \neq 0 \). Secondly, in order to prove the proposition, it is sufficient to show that

\[
(5.4) \quad \det(a_{ij}) = \pm 1 \quad \text{in} \ \mathbb{Z}.
\]

We now proceed with the proof of (5.4). Let \( \langle u_1, \ldots, u_r \rangle \) be the free abelian multiplicative group generated by \( u_1, \ldots, u_r \). Since \( \det(a_{ij}) \neq 0 \), the leading monomials \( \text{lm}(v_1), \ldots, \text{lm}(v_r) \) generate a (free abelian) subgroup \( \Lambda \) of rank \( r \) in \( \langle u_1, \ldots, u_r \rangle \cong \mathbb{Z}^r \); in other words, \( \Lambda \) has finite index in \( \langle u_1, \ldots, u_r \rangle \). On the other hand, (5.4) holds if and only if \( \Lambda = \langle u_1, \ldots, u_r \rangle \). It is therefore sufficient to prove that \( u_i \in \Lambda \) for every \( i = 1, \ldots, r \).

Since \( O_{x,X} \) and \( O_{y,Y} \) have the same field of fractions, each \( u_i \) can be written as \( p/q \), where \( p, q \in O_{y,Y} - \{0\} \). Represent \( p \) and \( q \) by power series
in $v_1,\ldots,v_r$. By Lemma 5.3 $\text{lm}(p(v_1,\ldots,v_r))$ and $\text{lm}(q(v_1,\ldots,v_r))$ lie in $\Lambda$; thus taking the leading monomials on both sides of the equation

$$q(v_1,\ldots,v_r)u_i = p(v_1,\ldots,v_r),$$

we conclude that $u_i \in \Lambda$, as desired. \hfill \Box

**Case II: $G$ - arbitrary.** We are now ready to finish the proof of Theorem 5.1. The idea is to replace $X$ and $Y$ by suitable $H$-slices, then appeal to Proposition 5.2.

Diagonalizing the $H$-action on $T^*_y(Gy)$, we obtain a basis $\overline{v}_1,\ldots,\overline{v}_d$ such that $h\overline{v}_i = \alpha_i(h)\overline{v}_i$ for some characters $\alpha_1,\ldots,\alpha_d$ of $H$. Since the natural $H$-equivariant $k$-vector space maps $M_{y,Y} \rightarrow T^*_y(Y) \rightarrow T^*_y(Gy)$ have $H$-equivariant splittings, we can lift $\overline{v}_1,\ldots,\overline{v}_d$ to $v_1,\ldots,v_d \in M_{y,Y}$ such that such that $hv_i = \alpha_i(h)v_i$ for each $h \in H$. In other words, $v_1,\ldots,v_d$ form a local coordinate system for $Gy$ at $y$.

Since both $\text{Stab}(x)$ and $\text{Stab}(y)$ are finite, $df_x : T_x(Gx) \rightarrow T_y(Gy)$ is an isomorphism. Thus $f^*(v_1),\ldots,f^*(v_r)$ form a local coordinate system for $Gx$ at $x$. Define $W \subset Y$ as the irreducible component of $\{v_1 = \cdots = v_d = 0\}$ passing through $y$ and $V \subset X$ as the irreducible component of $\{f^*(v_1) = \cdots = f^*(v_d) = 0\}$ passing through $x$. Then $W$ is an $H$-slice for $Y$ at $y$ and $V$ is an $H$-slice for $X$ at $x$.

Clearly $f(V) \subset W$, i.e., $f|_V : V \rightarrow W$ is a well-defined morphism. We claim that $f|_V$ is, in fact, a birational morphism. Theorem 5.1 follows from this claim because

$$i(X,x,H) \overset{\text{Remark 4.2}}{=} i(V,x,H) \overset{\text{Proposition 5.2}}{=} i(W,y,H) \overset{\text{Remark 4.2}}{=} i(Y,y,H).$$

To show that $f|_V$ is dominant, assume, to the contrary, that $\dim(f(V)) < r$. Since $X$ is irreducible, $GV$ is dense in $X$ by Lemma 3.4 and thus $\dim(Y) = \dim(f(X)) = \dim(f(GV)) = \dim(Gf(V)) \leq d + \dim(f(V)) < d + r$, a contradiction.

It remains to show that $f|_V$ is generically 1-1 on closed points. Since $f$ is a birational morphism (i.e., has degree 1), there exists a dense $G$-invariant open subset $Y_0$ of $Y$ such that for every $y_0 \in Y_0$, $f^{-1}(y_0)$ is a single point in $X$. Since $GW$ is dense in $Y$ (by Lemma 3.4), $Y_0 \cap W$ is a dense open subset of $W$. Thus a general point of $W$ has exactly one preimage in $X$. On the other hand, a general point of $W$ has at least $\deg(f|_V) \geq 1$ preimages in $X$. This shows that $\deg(f|_V) = 1$, i.e., $f|_V$ is birational, as claimed. \hfill \Box

**6. Proof of Theorem 1.1.**

In this section we deduce Theorem 1.1 from Theorem 5.1. Our proof relies on canonical resolution of singularities. (We remark that canonical resolution of singularities is not used elsewhere in this paper.)

We begin with two simple preliminary results.
Lemma 6.1. Let $H$ be an algebraic group, $f: Z \to X$ be a birational morphism of complete irreducible $H$-varieties and $X_1$ be an irreducible $H$-invariant codimension 1 subvariety of $X$. If $X_1$ passes through a normal point of $X$ then there exists an $H$-invariant irreducible codimension 1 subvariety $Z_1 \subset Z$ such that $f_{|Z_1}: Z_1 \to X_1$ is a birational morphism.

Proof. Since $X_1$ contains a normal point of $X$, the rational map $f^{-1}: X \dasharrow Z$ is defined at a general point of $X_1$. Now $Z_1 = \text{the closure of } f^{-1}(X_1)$ in $Z$, has the desired properties. □

Proposition 6.2. Let $H$ be a diagonalizable group, $\alpha: Z \to X$ be a birational morphism of complete irreducible $H$-varieties and $x$ be a smooth $H$-fixed point of $X$. Then there exists an $H$-fixed point $z \in Z$ such that $\alpha(z) = x$.

This result can be established by the argument used in the proof of [RY1, Proposition A.4], due to J. Kollár and E. Szabó. (In fact, if $H$ is a $p$-group then Proposition 6.2 follows from [RY1, Proposition A.4].) We give a simple self-contained proof below.

Proof. We argue by induction on $\dim(X)$. The base case, $\dim(X) = 0$, is obvious. For the induction step, assume $\dim(X) \geq 1$. We claim that there exists codimension 1 $H$-invariant irreducible subvariety $X_1$ such that $x$ is a smooth point of $X_1$. Arguing as we did at the beginning of the proof of Proposition 5.2, we see that there exists a nonzero element $u \in M_{x,X}$ such that for every $h \in H$, $hu = \alpha(h)u$, where $\alpha$ is a character of $H$. Then the (locally closed) subvariety $\{u = 0\}$ of $X$ is $H$-invariant and smooth at $x$. Hence, its unique irreducible component passing through $x$ is also $H$-invariant, and we can define $X_1$ as the closure of this irreducible component. This proves the claim.

Now by Lemma 6.1 there exists a codimension 1 irreducible $H$-invariant subvariety $Z_1$ of $Z$ such that $\alpha_{|Z_1}: Z_1 \to X_1$ is a birational morphism of $H$-varieties. Applying the induction assumption to this morphism, we construct $z \in Z_1 \subset Z$ with desired properties. □

We are now ready to complete the proof of Theorem 1.1. The idea is to construct a complete smooth model $Z$ of $X$ (or $Y$) that dominates them both, i.e., fits into a diagram

$$
\begin{array}{ccc}
Z & \alpha & \beta \\
X & & Y
\end{array}
$$

where $\alpha$ and $\beta$ are birational morphisms of $G$-varieties. If we find such a $Z$, Theorem 1.1 will easily follow. Indeed, by Proposition 6.2 there exists an
$H$-fixed point $z \in Z$ such that $\alpha(z) = x$. Setting $y = \beta(z)$ and applying Theorem 5.1 to $\alpha$ and $\beta$, we conclude that $i(X, x, H) = i(Z, z, H) = i(Y, y, H)$, as desired.

It remains to construct $Z$. Let $W \subset X \times Y$ be the closure of the graph of a birational isomorphism $f$ between $X$ and $Y$. Then $W$ is a complete $G$-variety that dominates both $X$ and $Y$. In other words, $W$ satisfies all of our requirements for $Z$, with one exception: It may not be smooth. Let

\begin{equation}
\pi: Z = W_n \xrightarrow{\pi_n} \cdots \xrightarrow{\pi_1} W_0 = W,
\end{equation}

be the canonical resolution of singularities of $W$, as in [V, Theorem 7.6.1] or [BM, Theorem 1.6]. Here $Z$ is smooth, and each $\pi_i$ is a blowup with a smooth center; since these centers are canonically chosen, they are $G$-invariant. Thus the $G$ action can be lifted to $Z$ so that $\pi$ is a birational morphism of complete $G$-varieties. The smooth complete $G$-variety $Z$ constructed in this way has the desired properties. □

**Remark 6.3.** An alternative construction of $Z$ is given by the equivariant version of Hironaka’s theorem on elimination of points of indeterminacy (proved in [RY4]), which asserts that for every rational map $f: X \to Y$ of $G$-varieties there exists a sequence of blowups

$$
\pi: Z = X_n \xrightarrow{\pi_n} \cdots \xrightarrow{\pi_1} X_0
$$

with smooth $G$-equivariant centers such that $f \pi: Z \to Y$ is regular. The advantage of this approach is that it only uses Proposition 6.2 in the case where $\alpha$ is a single blowup with a smooth $G$-equivariant center (in which case the proof is immediate; see, e.g., [RY1, Lemma 5.1]). On the other hand, since the theorem on equivariant elimination of points of indeterminacy is itself deduced from canonical resolution of singularities in [RY4], we opted for a direct proof here.

**Remark 6.4.** A rational map $f: X \to Y$ (respectively, a morphism $f: X \to Y$) of reducible varieties is called a birational isomorphism (respectively, a birational morphism) if $X$ and $Y$ have irreducible decompositions $X_1 \cup \cdots \cup X_n$ and $Y_1 \cup \cdots \cup Y_n$ such that $f$ restricts to a birational isomorphism $X_i \to Y_i$ (respectively, a birational morphism) for each $i$. With this definition, the irreducibility assumption in Proposition 6.2 can be removed. Indeed, we can reduce to the case where $X$ and $Z$ are irreducible by replacing them with suitable irreducible components.

The irreducibility assumption in Theorem 1.1 (respectively, Theorem 5.1 and Proposition 5.2) can also be removed, if we assume $\dim_x(X) = d + r$ (respectively, $\dim_x(X) = d + r$ and $\dim_y(X) = r$). Indeed, if $X_1$ is the (necessarily unique) irreducible component of $X$ containing $x$ and $G_1 = \{ g \in G \mid g(X_1) = X_1 \}$, then $H \subset G_1$, $[G : G_1] < \infty$, and $i(X, x, H) = i(X_1, x, H)$, so that in each case we may replace $X, Y$, and $G$ by $X_1, Y_1$ and $G_1$. 
Remark 6.5. One may ask if the condition that \( \text{Stab}(x) \) is finite for every \( x \in X^H \) (and similarly for \( Y \)) of Theorem 1.1 is ever satisfied. Indeed, if \( H \) is contained in a torus \( T \) of \( G \) then the answer is “no”, since \( \text{Stab}(x) \) is infinite for every \( x \in X^T \subset X^H \), and \( X^T \neq \emptyset \) by the Borel Fixed Point Theorem. On the other hand, if the centralizer \( C_G(H) \) is finite, then we claim that every generically free \( G \)-variety \( X \) has a birational model satisfying this condition. Indeed, by [RY1, Theorem 1.1] \( X \) has a model with the property that the stabilizer of every point is of the form \( U \triangleleft D \), where \( D \) is diagonalizable and \( U \) is unipotent. Assume \( x \in X^H \). By the Levi decomposition theorem, we may choose \( D \) so that \( D \) contains \( H \). Now [RY1, Lemma 7.3] tells us that \( U = \{1\} \). Thus \( \text{Stab}(x) = D \subset C_G(H) \), which is finite, as claimed.

Examples of pairs \( H \subset G \), where \( G \) is a semisimple algebraic group and \( H \) is an abelian subgroups of \( G \) whose centralizer is finite can be found, e.g., in [Gr], [Se2] or [RY1, Section 8].


It is not difficult to see that Conjecture 1.2 fails if \( G \) is a finite cyclic group of order \( n = 5 \) or \( \geq 7 \). Indeed, let \( V_\omega \) be the 1-dimensional representation of \( G \) such that \( \sigma \) acts on \( V_\omega \) by the character \( x \mapsto \omega x \), where \( \omega \) is a primitive \( n \)th root of unity. Since any birational automorphism of \( \mathbb{A}^1 \) lifts to a regular automorphism of \( \mathbb{P}^1 \), it is easy to see that \( V_\omega \) is birationally isomorphic to \( V_{\omega'} \) iff \( \omega' = \omega \) or \( \omega' = \omega^{-1} \). (The two \( G \)-fixed points in \( \mathbb{P}^1 \) are preserved in the former case and interchanged in the latter.) If \( n = 5 \) or \( \geq 7 \), we can find two primitive \( n \)th roots of unity \( \omega \) and \( \omega' \) such that \( \omega' \neq \omega \pm 1 \), so that \( V_\omega \) and \( V_{\omega'} \) are not birationally isomorphic. (P.I. Katsylo has informed us that this observation was independently made by E.A. Tevelev.)

In this section we will classify faithful linear representations of diagonalizable group \( G \) up to birational equivalence and show that Conjecture 1.2 fails for a number of groups, both abelian and nonabelian. These results have the same general flavor as the observation in the previous paragraph but the arguments are more complicated due to the fact that we will be working with higher-dimensional varieties, rather than curves.

Representations of diagonalizable groups. Recall that every linear representation \( V \) of \( G \) decomposes as a sum of 1-dimensional character spaces ([Bo, Proposition III.8.2(d)]); if the associated characters of \( G \) are \( \chi_1, \ldots, \chi_d \), we shall write \( V = \chi_1 \oplus \cdots \oplus \chi_d \).

Theorem 7.1. Let \( G \) be a diagonalizable group of rank \( r \) and let \( V = \chi_1 \oplus \cdots \oplus \chi_d \) and \( W = \eta_1 \oplus \cdots \oplus \eta_d \) be faithful \( d \)-dimensional linear representations of \( G \). (In particular, \( d \geq r \).) Then \( V \) and \( W \) are birationally isomorphic as \( G \)-varieties if and only if \( \chi_1 \wedge \ldots \wedge \chi_d = \pm \eta_1 \wedge \cdots \wedge \eta_d \) in \( \bigwedge^d(G^*) \).
Proof. Since $G$ acts faithfully on $V$ and $W$, we have

\begin{equation}
\langle \chi_1, \ldots, \chi_d \rangle = \langle \eta_1, \ldots, \eta_d \rangle = G^*.
\end{equation}

Assume $\chi_1 \wedge \cdots \wedge \chi_d = \pm \eta_1 \wedge \cdots \wedge \eta_d$. Then by Corollary 2.5 there exists an $N = (n_{ij}) \in \text{GL}(Z)$ such that $\eta_i = \chi_1^{n_{i1}} \cdots \chi_d^{n_{id}}$. The desired birational isomorphism $V \longrightarrow W$ is now be explicitly given, in the natural coordinates on $V$ and $W$, by $(x_1, \ldots, x_d) \longrightarrow (y_1, \ldots, y_d)$, where $y_i = x_1^{n_{i1}} \cdots x_d^{n_{id}}$.

Conversely, suppose

\begin{equation}
\chi_1 \wedge \cdots \wedge \chi_d \neq \pm \eta_1 \wedge \cdots \wedge \eta_d \text{ in } \wedge^d(G^*).
\end{equation}

We want to prove that $V$ and $W$ are not birationally isomorphic as $G$-varieties. Note that (7.2) is impossible if $d \geq r + 1$, since in this case $\wedge^d(G^*) = (0)$; see Lemma 2.1(b). Thus we will assume from now on that $d = r = \text{rank}(G)$. We will consider three cases.

**Case 1.** $G = (G_m)^r$ is a torus. In this case $\chi_1 \wedge \cdots \wedge \chi_r$ and $\eta_1 \wedge \cdots \wedge \eta_r$ are both generators of $\wedge^r(G^*) = Z$, so that (7.2) is impossible.

**Case 2.** $G$ is a finite abelian group. The $G$-action on $V = A^r$ (respectively $W = A^r$) naturally extends to the projective space $\mathbb{V} = \mathbb{P}^r$ (respectively, $\mathbb{W} = \mathbb{P}^r$). Example 4.4 shows that for every $G$-fixed point $x \in \mathbb{V}$, $i(\mathbb{V}, x, G) = \pm \chi_1 \wedge \cdots \wedge \chi_r$ and for every $G$-fixed point $y \in \mathbb{W}$, $i(\mathbb{W}, y, G) = \pm \eta_1 \wedge \cdots \wedge \eta_r$. Thus in view of (7.2), Theorem 1.1 says that $\mathbb{V}$ and $\mathbb{W}$ (and, hence, $V$ and $W$) are not birationally isomorphic as $G$-varieties.

**Case 3.** $G$ is a diagonalizable group but not a torus. Write $G = \mathbb{G}_m(n_1) \times \cdots \times \mathbb{G}_m(n_r)$, as in (2.1). Since $G$ is not a torus, $n_1 \geq 2$. Let $H = \mathbb{G}_m(n_1)^r = (\mathbb{Z}/n_1\mathbb{Z})^r$ be the $n_1$-torsion subgroup of $G$. It is sufficient to show that $V$ and $W$ are not birationally isomorphic as $H$-varieties; then they certainly cannot be birationally isomorphic as $G$-varieties. By Case 2, it is enough to show that

\begin{equation}
\chi_1' \wedge \cdots \wedge \chi_r' \neq \pm \eta_1' \wedge \cdots \wedge \eta_r' \text{ in } \wedge^r(H^*)
\end{equation}

where $\chi_i'$ and $\eta_j'$ are the characters of $H$ obtained by restricting $\chi_i$ and $\eta_j$ from $G$ to $H$. Note that the inclusion $\phi: H \hookrightarrow G$ induces a surjection $\phi^*: G^* \longrightarrow H^*$ of the dual group, which, in turn, induced a map of cyclic groups $\wedge^r(\phi^*): \wedge^r(G^*) \longrightarrow \wedge^r(H^*)$. Elementary group theory tells us that $G^* = (\mathbb{Z}/n_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/n_r\mathbb{Z})$, $H^* = (\mathbb{Z}/n_1\mathbb{Z})^r$, $\phi^*: G^* \longrightarrow H^*$ is (componentwise) reduction modulo $n_1$, and $\wedge^r(\phi^*)$ is the identity map $\wedge^r(G^*) = \mathbb{Z}/n_1\mathbb{Z} \longrightarrow \mathbb{Z}/n_1\mathbb{Z} = \wedge^r(H^*)$. Applying $\wedge^r(\phi^*)$ to both sides of (7.2), we obtain (7.3), as desired. \hfill \Box

**Proof of Theorem 1.3.** Let $G$ be a diagonalizable group $G$ of rank $r$ of the form (1.1), and let $V = \chi_1 \oplus \cdots \oplus \chi_d$ and $W = \eta_1 \oplus \cdots \ominus \eta_d$ be faithful $d$-dimensional linear representations of $G$. 


(a) If \( d \geq r + 1 \) then \( \bigwedge^r(G^*) = (0) \), so that \( \chi_1 \wedge \ldots \wedge \chi_d = 0 = \eta_1 \wedge \ldots \wedge \eta_r \). Thus \( V \) and \( W \) are birationally isomorphic as \( G \)-varieties by Theorem 7.1.

From now on we will assume \( d = r \). Note that in this case both \( \chi_1 \wedge \ldots \wedge \chi_r \) and \( \eta_1 \wedge \ldots \wedge \eta_r \) are generators of \( \bigwedge^r(G^*) = \mathbb{Z}/n_1\mathbb{Z} \).

(b) Suppose \( n_1 = 2 \). Since \( \bigwedge^r(G^*) = \mathbb{Z}/2\mathbb{Z} \) has only one generator, \( \chi_1 \wedge \ldots \wedge \chi_r = \eta_1 \wedge \ldots \wedge \eta_r \). Thus \( V \) and \( W \) are birationally isomorphic as \( G \)-varieties by Theorem 7.1.

Now assume \( n_1 = 0 \). Then \( \bigwedge^r(G^*) = \mathbb{Z} \). The only generators of this group are \( \pm 1 \); thus \( \chi_1 \wedge \ldots \wedge \chi_r = \pm \eta_1 \wedge \ldots \wedge \eta_r \), and, once again, Theorem 7.1 tells us that \( V \) and \( W \) are birationally isomorphic.

(c) Suppose \( n_1 \geq 3 \). By Theorem 7.1, birational isomorphism classes of \( r \)-dimensional linear representations of \( H \) are in 1-1 correspondence with the generators of \( \bigwedge^r(H^*) \simeq \mathbb{Z}/n_1\mathbb{Z} \) (as an additive group), modulo multiplication by \( -1 \). Since \( n_r \geq 3 \), \( a \neq -a \) for any generator \( a \) of \( \mathbb{Z}/n_1\mathbb{Z} \). Thus in this case the number of isomorphism classes of faithful \( r \)-dimensional \( H \)-representations is \( \phi(n_r)/2 \), as claimed. \( \square \)

**Further counterexamples to Conjecture 1.2.** Theorem 1.3 shows that Conjecture 1.2 fails for many diagonalizable groups. We will now see that this conjecture also fails for some nonabelian finite groups.

**Proposition 7.2.** Let \( n \) and \( r \) be positive integers, \( P \) be a subgroup of \( S_r \), and \( G = (\mathbb{Z}/n\mathbb{Z})^r \ast P \), where \( P \) acts on \( (\mathbb{Z}/n\mathbb{Z})^r \) by permuting the factors. Assume there exists an \( m \in \mathbb{Z} \) such that \( (m, n) = 1 \) and \( m^r \not\equiv \pm 1 \mod n \). Then there exist two birationally inequivalent \( r \)-dimensional representations of \( G \). In particular, Conjecture 1.2 fails for this group.

We remark that an integer \( m \) satisfying the requirements of Proposition 7.2 always exists if the exponent of \( U_n \) does not divide \( 2r \); here \( U_n \) is the (multiplicative) group of units in \( \mathbb{Z}/n\mathbb{Z} \). In particular, \( m \) exists if there is a prime power \( p^e \) such that \( p^e \mid n \) but \( \phi(p^e) = (p - 1)p^{e-1} \nmid 2r \).

**Proof.** Let \( \omega \) be a primitive \( n \)-th root of unity. We define the \( r \)-dimensional representations \( V \) and \( W \) of \( G \) as follows:

\[
((a_1, \ldots, a_r), \sigma) : (v_1, \ldots, v_r) \mapsto (\omega^{a_1}v_{\sigma^{-1}(1)}, \ldots, \omega^{a_r}v_{\sigma^{-1}(r)})
\]

and

\[
((a_1, \ldots, a_r), \sigma) : (w_1, \ldots, w_r) \mapsto (\omega^{ma_1}w_{\sigma^{-1}(1)}, \ldots, \omega^{ma_r}w_{\sigma^{-1}(r)}).
\]

Here \( a_1, \ldots, a_n \in \mathbb{Z}/n\mathbb{Z} \), \( \sigma \in P \subset S_r \), \( (v_1, \ldots, v_r) \in V \) and \( (w_1, \ldots, w_r) \in W \). It is easy to see that \( V \) and \( W \) are, indeed, well-defined faithful \( r \)-dimensional representations of \( G \).

To prove the proposition it is sufficient to show that \( V \) and \( W \) are not birationally isomorphic as \( (\mathbb{Z}/n\mathbb{Z})^r \)-varieties. Let \( \chi_i \) be the character of
are at least $\phi$. In this section we will use the invariant $i_{X,x,H}$ and let $V = \chi_1 \oplus \cdots \oplus \chi_r$ and $W = \chi_1^m \oplus \cdots \oplus \chi_r^m$. By our assumption

$$x_1^m \wedge \cdots \wedge x_r^m = m^r x_1 \wedge \cdots \wedge x_r \neq \pm x_1 \wedge \cdots \wedge x_r$$

in $\Lambda^r(\mathbb{Z}/n\mathbb{Z}) \cong \mathbb{Z}/n\mathbb{Z}$. Thus Theorem 7.1 tells us that $V$ and $W$ are not isomorphic as $(\mathbb{Z}/n\mathbb{Z})^r$-varieties (and hence, as $G$-varieties). \hfill \Box

**Remark 7.3.** The same argument proves the following stronger result. Let $n_1, n_2, \ldots, n_s, r_1, \ldots, r_s$ be positive integers such that $n_i$ divides $n_{i+1}$ for $i = 2, \ldots, r$. Let $P_i$ be a subgroup of the symmetric group $S_{n_i}$ and let $G_i = (\mathbb{Z}/n_i\mathbb{Z})^{r_i} \rtimes P_i$. Assume there exist integers $m_1, \ldots, m_s$ such that $(m_i, n_i) = 1$ and $m_1^{r_1} \cdots m_s^{r_s} \equiv \pm 1 \pmod{n_1}$. Then $G = (G_m)^a \times G_1 \times \cdots \times G_s$ has two birationally inequivalent $(a + r_1 + \cdots + r_s)$-dimensional representations. In particular, Conjecture 1.2 fails for any $G$ of this form.

**Remark 7.4.** The proof of Proposition 7.2 shows that $G = (\mathbb{Z}/n\mathbb{Z})^r \rtimes P$ has at least $|\pm U^r_n|/2$ birational isomorphism classes of $r$-dimensional representations. Here, as before, $U^r_n$ denotes the multiplicative group of units in the ring $\mathbb{Z}/n\mathbb{Z}$, and $\pm U^r_n$ denotes the subset of $U_n$ consisting of elements of the form $\pm m^r$, as $m$ ranges over $U_n$.

A similar estimate can be given for the number of birational isomorphism classes of $(a + r_1 + \cdots + r_s)$-dimensional representations of the group $G$ in Remark 7.3. In particular, if $n_1 = \cdots = n_s$ and $(r_1, \ldots, r_s) = 1$ then there are at least $\phi(n_1)/2$ such classes.

### 8. Birational equivalence of quantum tori.

In this section we will use the invariant $i(X, x, H)$ to classify PGL$_n$-varieties (and consequently central simple algebras) of a certain form. In particular, we will prove Theorem 1.4.

**Abelian subgroups of PGL$_n$.** Let $A$ be a finite abelian group of order $n$ and let $V = k[A]$ be the group ring of $A$. For $a \in A$ and $\chi \in A^*$ define $P_a, D_\chi \in \text{GL}(V)$ by $P_a(b) = ab$ and $D_\chi(b) = \chi(b)b$ for every $b \in A$. It is easy to see that $D_\chi P_a = \chi(a) P_a D_\chi$. Thus if $p_a$ and $d_\chi$ denote the elements of PGL($V$) represented, respectively, by $P_a$ and $D_\chi \in \text{GL}(V)$, then

$$\phi: A \times A^* \hookrightarrow \text{PGL}(V) = \text{PGL}_n$$

$$\phi(a, \chi) \mapsto p_a d_\chi$$

defines an embedding of $A \times A^*$ in PGL$_n$.

Let $H$ be an abelian subgroup of PGL$_n$. Then $H$ is naturally equipped with an alternating bilinear form $\omega_H: H \times H \to \mu_n$ (cf. Definition 2.2(a)). Here $\mu_n$ is the group of $n$th roots of unity in $k$, identified with the center of SL$_n$, and $\omega_H(a, b) = ABA^{-1}B^{-1}$, where $A$ and $B \in \text{SL}_n$ represent $a$ and $b$, respectively.
Lemma 8.1. Let $A$ be a finite abelian group of rank $r$, $H = \phi(A \times A^*) = \{ p_a d_\chi \mid a \in A, \chi \in A^* \}$ be the subgroup of $\text{PGL}_n$ defined above. Then

(a) the elements of $P_a D_\chi$ span the matrix algebra $M_n = M_n(k)$ as a $k$-vector space; here as $a$ ranges over $A$ and $\chi$ ranges over $A^*$, and
(b) the alternating bilinear form $\omega_H$ is symplectic (i.e., nondegenerate).

Let $g$ be an element of the normalizer $N_{\text{PGL}_n}(H)$, and $\psi_g : H \rightarrow H$ be conjugation by $g$. Then

(c) $\psi_g$ preserves $\omega_H$, and
(d) $\psi_g$ induces the identity map $\wedge^{2r}(H^*) \rightarrow \wedge^{2r}(H^*)$.

Proof. (a) See [RY3, Lemma 3.2].
(b) See [RY2, Lemma 7.8].
(c) Choose $a$ and $b \in H \subset \text{PGL}_n$ and lift them to $A$ and $B \in \text{SL}_n$. Since $ABA^{-1}B^{-1}$ is a central element of $\text{SL}_n$, we have

$$
\omega_H(\psi_g(a), \psi_g(b)) = \omega_H(gag^{-1}, gbg^{-1})
= (gAg^{-1})(gBg^{-1})(gA^{-1}g^{-1})(gB^{-1}g^{-1})
= g(ABA^{-1}B^{-1})g^{-1} = ABA^{-1}B^{-1} = \omega_H(a, b),
$$
as claimed.
(d) Follows from (b), (c) and Lemma 2.3(b). \hfill \Box

PGL$_n$-varieties.

Proposition 8.2. Let $A$ be a finite abelian group of order $n$ and rank $r$ and let $H = \phi(A \times A^*)$ be the subgroup of $\text{PGL}_n$ defined above. Suppose $V = \chi_1 \oplus \cdots \oplus \chi_{2r}$ and $W = \eta_1 \oplus \cdots \oplus \eta_{2r}$ are faithful representations of $H$. Then the following are equivalent:

(a) $\chi_1 \wedge \cdots \wedge \chi_{2r} = \pm \eta_1 \wedge \cdots \wedge \eta_{2r}$ in $\wedge^{2r}(H^*)$,
(b) $V$ and $W$ are birationally isomorphic as $H$-varieties,
(c) $X = \text{PGL}_n *_H V$ and $Y = \text{PGL}_n *_H W$ are birationally isomorphic as $\text{PGL}_n$-varieties.

Here $\text{PGL}_n *_H V$ and $\text{PGL}_n *_H W$ are homogeneous fiber products; see Example 3.3.

Proof. (a) and (b) are equivalent by Theorem 7.1. The implication (b) $\implies$ (c) is obvious.

Thus we only need to show (c) $\implies$ (a). The idea of the proof is to appeal to Theorem 1.1. We begin by observing that $X$ and $Y$ naturally embed as dense open subsets in projective varieties $\overline{X} = (\mathbb{P}(M_n) \times \overline{V})/H$ and $\overline{Y} = (\mathbb{P}(M_n) \times \overline{W})/H$ respectively. Here $\overline{V} = \mathbb{P}^{2r}$ is the projective completion of $V = \mathbb{A}^{2r}$; $\text{PGL}_n$ acts on $\mathbb{P}(M_n) \times \overline{V}$ by left multiplication on
the first factor; this action commutes with the \( H \)-action on \( \mathbb{P}(M_n) \times V \) given by \( h: (x, y) \mapsto (xh^{-1}, hy) \) and thus descends to the geometric quotient \( \overline{X} = (\mathbb{P}(M_n) \times V)/H \). We shall denote the point \( x \in \overline{X} \) corresponding to the orbit of \((g, v) \in \mathbb{P}(M_n) \times V \) by \([g, v] \). The \( H \)-variety \( \overline{W} \) and the \( \text{PGL}_n \)-variety \( \overline{Y} \) are defined in a similar manner.

Our goal is to show that

(i) every \( H \)-fixed points of \( \overline{X} \) is of the form \( x = [g, v] \), where \( g \in N_{\text{PGL}_n}(H) \) and \( v \) is an \( H \)-fixed point of \( V \), and for any such point \( x \),

(ii) \( \text{Stab}(x) = H \) and

(iii) up to sign, \( i(X, x, H) = \chi_1 \wedge \cdots \wedge \chi_{2r} \).

These assertions, in combination with Theorem 1.1, will prove that if \( X \) and \( Y \) (and hence, \( \overline{X} \) and \( \overline{Y} \)) are birationally isomorphic then \( \chi_1 \wedge \cdots \wedge \chi_{2r} = \pm \eta_1 \wedge \cdots \wedge \eta_{2r} \), i.e., (c) \( \implies \) (a).

To prove (i), assume \( x = [g, v] \) is an \( H \)-fixed point of \( \overline{X} \) for some \( g \in \mathbb{P}(M_n) \) and \( v \in \overline{V} \). This means that for every \( h \in H \) there exists an \( h' \in H \) such that \((hg, v) = (gh', (h')^{-1}v) \) in \( \mathbb{P}(M_n) \times \overline{V} \). Equivalently, \( hg = gh' \) and \((h')^{-1}v = v \).

Consider the vector space \( k^n \) of \((n \times 1)\)-row vectors. The multiplication by \( g \) on the right yields a linear map \( k^n \to k^n \); let \( \text{RKer}(g) \) be the kernel of this map. Note that since \( g \in \mathbb{P}(M_n) \), this linear map is only defined up to a nonzero constant multiple but \( \text{RKer}(g) \) is well-defined.

The equality \( hg = gh' \) implies that \( \text{RKer}(g) \) is an \( H \)-invariant subspace of \( k^n \) with respect to the right action of \( H \); again, as \( H \subset \text{PGL}_n \), the right multiplication by an element \( h \in H \) is a linear map \( k^n \to k^n \) defined up to a nonzero constant multiple, but the notion of \( H \)-invariance of a linear subspace of \( k^n \) is well-defined.

Now recall that by Lemma 8.1(a) the \( n^2 \) elements of the form \( P_a D_\chi \) which represent the elements of \( H \subset \text{PGL}(V) = \text{PGL}_n \) in \( \text{GL}(V) = \text{GL}_n \) span \( M_n \) as a \( k \)-vector space. Thus the only \( H \)-invariant subspaces of \( k^n \) are the ones that are invariant under all of \( M_n \), namely \( k^n \) and \( 0 \). If \( \text{RKer}(g) = k^n \) then \( g \) is the zero matrix, which is impossible since \( g \in \mathbb{P}(M_n) \). Thus we conclude that \( \text{RKer}(g) = (0) \). This means that \( g \) is nonsingular, i.e., \( g \in \text{PGL}_n \). Now we can rewrite \( hg = gh' \) as \( g^{-1}hg = h' \in H \); this shows that \( g \in N_{\text{PGL}_n}(H) \). Moreover, as \( h \) ranges over \( H \), \( h' = g^{-1}hg \) also ranges over all of \( H \). Thus the equality \( (h')^{-1}v = v \) implies that \( v \) is an \( H \)-fixed point of \( \overline{V} \). This proves (i).

From now on let \( x = [g, v] \) be an \( H \)-fixed point of \( \overline{X} \), where \( g \in N_{\text{PGL}_n}(H) \) and \( v \) is an \( H \)-fixed point of \( \overline{V} \).

To prove (ii), assume \( g' \in \text{Stab}(x) \), i.e., \( g'[g, v] = [g, v] \). Then \( g'g = gh' \) for some \( h' \in H' \). Since \( g \in N_{\text{PGL}_n}(H) \), we conclude that \( g' = ghg^{-1} \in H \), as desired.
To prove (iii), first note that \( i(\overline{X}, [g, v], H) = \wedge^2 \psi_g^* \left( i(\overline{X}, [1, v], H) \right) \), where \( \psi_g : H \to H \) is conjugation by \( g \in N_{\text{PGL}_n}(H) \) and \( \wedge^2 \psi_g^* \) is the automorphism of \( \wedge^2 (H^*) \) induced by \( \psi_g \); see Remark 4.3. By Lemma 8.1(d) \( \wedge^2 \psi_g^* \) is the identity automorphism. Thus \( i(\overline{X}, [g, v], H) = i(\overline{X}, [1, v], H) \).

On the other hand, by Remark 4.2 \( i(\overline{X}, [1, v], H) = i(\overline{V}, [v], H) \). Finally, recall that for any \( v \in \overline{V} \), \( i(\overline{V}, [v], H) = i(V, [0v], H) = \chi_1 \wedge \cdots \wedge \chi_{2r} \); see Example 4.4. In summary,

\[
i(\overline{X}, [g, v], H) = i(\overline{X}, [1, v], H) = i(\overline{V}, [v], H) = \chi_1 \wedge \cdots \wedge \chi_{2r},
\]

as claimed.

\[\square\]

**Remark 8.3.** The exceptional group \( E_8 \) has a nontoral subgroup \( H \) isomorphic to \( (\mathbb{Z}/5\mathbb{Z})^3 \); see [Gr, Lemma 10.3]. Modifying the proof of Proposition 8.2, we can show the following:

Let \( V = \chi_1 \oplus \chi_2 \oplus \chi_3 \) and \( W = \eta_1 \oplus \eta_2 \oplus \eta_3 \) be faithful 3-dimensional representations of \( H \), where \( \chi_i \) and \( \eta_j \) are characters of \( H \). Then the following are equivalent:

- (a) \( \chi_1 \wedge \chi_2 \wedge \chi_3 = \pm \eta_1 \wedge \eta_2 \wedge \eta_3 \) in \( \wedge^3 (H^*) \cong \mathbb{Z}/5\mathbb{Z} \),
- (b) \( V \) and \( W \) are birationally isomorphic as \( H \)-varieties, and
- (c) \( E_8 *_H V \) and \( E_8 *_H W \) are birationally isomorphic as \( E_8 \)-varieties.

In particular, there are exactly two birational isomorphism classes of \( E_8 \)-varieties of the form \( E_8 *_H V \), where \( V \) is a faithful 3-dimensional representation of \( H \): One corresponds to \( \pm 1 \), and the other to \( \pm 2 \) in \( \mathbb{Z}/5\mathbb{Z} \cong \wedge^3 (H^*) \).

**Remark 8.4.** Note that the \( \text{PGL}_n \)-varieties \( X = \text{PGL}_n *_H V \) and \( Y = \text{PGL}_n *_H W \) of Proposition 8.2, are stably isomorphic (as \( \text{PGL}_n \)-varieties). In fact, \( X \times A^1 \cong Y \times A^1 \) because \( X \times A^1 = \text{PGL}_n *_H (V \times A^1), Y \times A^1 = \text{PGL}_n *_H (W \times A^1) \), and \( V \times A^1 \cong W \times A^1 \) as \( H \)-varieties by Theorem 7.1. For the same reason \( X \times A^1 \) and \( Y \times A^1 \) are isomorphic \( E_8 \)-varieties, if \( X \) and \( Y \) are as in Remark 8.3.

**Proof of Theorem 1.4.** Recall that birational isomorphism classes of generically free irreducible \( \text{PGL}_n \)-varieties \( X \) with \( k(X)^{\text{PGL}_n} = K \) are in 1-1 correspondence with central simple algebras of degree \( n \) over \( K \); see e.g., [Se1, X.5] or [RY2, Section 3].

In particular, by [RY3, Lemma 4.2], the algebra \( Q(\omega_1, \ldots, \omega_r) \) of Theorem 1.4 corresponds to the variety \( X = \text{PGL}_n *_H V \), where \( V \) is a faithful 2r-dimensional representations of \( H = A \times A^* \) constructed as follows.

Choose a set of generators \( a_1, \ldots, a_r \) for \( A = \mathbb{Z}/n_1\mathbb{Z} \times \cdots \times \mathbb{Z}/n_r\mathbb{Z} \) and a “dual” set of generators \( \chi_1, \ldots, \chi_r \) for \( A^* \) so that

\[
\chi_i(a_j) = \begin{cases} 
1 & \text{if } i \neq j \\
\omega_i & \text{if } i = j.
\end{cases}
\]
Now note that each $a \in A$ defines a character of $H = A \times A^*$ by $(b, \eta) \mapsto \eta(a)$. Similarly, each $\chi \in A^*$ gives rise to a character $H = A \times A^* \rightarrow k^*$ via $(b, \eta) \mapsto \chi(b)$; we shall denote these characters by $c(a)$ and $c(\chi)$ respectively. In these notations,

$$V = c(a_1) \oplus \cdots \oplus c(a_r) \oplus c(\chi_1)^{-1} \oplus \cdots \oplus c(\chi_r)^{-1},$$

see [RY3, Proof of Lemma 4.2].

Similarly the $\text{PGL}_n$-variety associated to $Q(\omega_1^m, \ldots, \omega_r^m)$ is $Y = \text{PGL}_n \ast_H W$, where $W = c(a'_1) \oplus \cdots \oplus c(a'_r) \oplus c(\chi'_1)^{-1} \oplus \cdots \oplus c(\chi'_r)^{-1}$. Here $a'_1, \ldots, a'_r$ are generators of $A$ and $\chi'_1, \ldots, \chi'_r$ are generators of $A^*$ such that

$$\chi'_i(a'_j) = \begin{cases} 1 & \text{if } i \neq j \\ \omega_i^m & \text{if } i = j. \end{cases}$$

A natural choice for $a'_i$ and $\chi'_i$ is $a'_i = a_i$ and $\chi'_i = \chi_i^m$, so that

$$W = c(a_1) \oplus \cdots \oplus c(a_r) \oplus c(\chi_1)^{-m_1} \oplus \cdots \oplus c(\chi_r)^{-m_r}.$$

As we mentioned above, $Q(\omega_1, \ldots, \omega_r)$ and $Q(\omega_1^m, \ldots, \omega_r^m)$ are isomorphic as $k$-algebras iff their associated $\text{PGL}_n$-varieties, $X = \text{PGL}_n \ast_H V$ and $Y = \text{PGL}_n \ast_H W$, are birationally isomorphic. By Proposition 8.2 $X$ and $Y$ are birationally isomorphic iff

$$c(a_1) \wedge \cdots \wedge c(a_r) \wedge c(\chi_1)^{-1} \wedge \cdots \wedge c(\chi_r)^{-1}$$

$$= \pm c(a_1) \wedge \cdots \wedge c(a_r) \wedge c(\chi_1)^{-m_1} \wedge \cdots \wedge c(\chi_r)^{-m_r} \text{ in } \bigwedge^{2r}(H^*) \cong \mathbb{Z}/n_r\mathbb{Z}.$$  

The last condition is equivalent to $m_1 \cdots m_r = \pm 1 \pmod{n_1}$.

\begin{thebibliography}{9}


\end{thebibliography}
246 Z. REICHSTEIN AND B. YOUSSIN


Received July 13, 2000 and revised February 16, 2001. Z. Reichstein was partially supported by NSF grant DMS-9801675

DEPARTMENT OF MATHEMATICS
OREGON STATE UNIVERSITY
CORVALLIS, OR 97331

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, B.C., CANADA V6T 1Z2
E-mail address: reichst@math.ubc.ca

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
UNIVERSITY OF THE NEGEV
BE’ER SHEVA’, ISRAEL

Hashofar 26/3
MA’ALE ADUMIM, ISRAEL
E-mail address: youssin@math.huji.ac.il
A LIOUVILLE TYPE THEOREM FOR SEMILINEAR ELLIPTIC SYSTEMS

Tomomitsu Teramoto and Hiroyuki Usami

Dedicated to Professor Masayuki Itô on his 60th birthday

This paper treats the second order semilinear elliptic systems of the form

\[ \begin{align*}
\Delta u &= p(x)v^\alpha, \\
\Delta v &= q(x)u^\beta,
\end{align*} \]

where \( N \geq 3 \), \( \alpha \), \( \beta > 0 \) are constants satisfying \( \alpha \beta > 1 \), and \( p, q \in C(\mathbb{R}^N; (0, \infty)) \). We obtain a Liouville type theorem for non-negative entire solutions of this system.

1. Introduction and statement of the result.

In this paper we consider second order semilinear elliptic systems of the form

\[ \begin{align*}
\Delta u &= p(x)v^\alpha, \\
\Delta v &= q(x)u^\beta, \\
x &\in \mathbb{R}^N,
\end{align*} \]

where \( N \geq 3 \), \( \alpha > 0 \) and \( \beta > 0 \) are constants satisfying \( \alpha \beta > 1 \), and \( p, q \in C(\mathbb{R}^N; (0, \infty)) \). An entire solution of system (1) is defined to be a function \((u, v) \in C^2(\mathbb{R}^N) \times C^2(\mathbb{R}^N)\) which satisfies (1) at every point in \( \mathbb{R}^N \).

In the previous paper [4] one of the authors has proved the following:

**Theorem 0.**

(i) Let \( \alpha > 1 \) and \( \beta > 1 \). Suppose that

\[ \liminf_{|x| \to \infty} |x|^\lambda p(x) > 0 \quad \text{and} \quad \liminf_{|x| \to \infty} |x|^\mu q(x) > 0 \]

hold for some constants \( \lambda \) and \( \mu \) satisfying

\[ \begin{align*}
\lambda &\leq \alpha(2 - \mu) + 2 \quad \text{or} \\
\mu &\leq \beta(2 - \lambda) + 2.
\end{align*} \]

Then (1) does not possess any positive entire solutions.

(ii) Let \( \alpha \beta > 1 \), and \( p \) and \( q \) have radial symmetry. Suppose moreover that

\[ \limsup_{|x| \to \infty} |x|^\lambda p(x) < \infty \quad \text{and} \quad \limsup_{|x| \to \infty} |x|^\mu q(x) < \infty \]
hold for some constants \( \lambda \) and \( \mu \) satisfying
\[
\begin{align*}
\lambda &> \alpha(2 - \mu) + 2 \\
\mu &> \beta(2 - \lambda) + 2.
\end{align*}
\]

Then (1) has infinitely many positive radial entire solutions.

From the above statement a natural question comes to the authors: When \( \alpha \beta > 1 \) (and one of \( \alpha \) and \( \beta \) is less than 1), and (2) and (3) hold for some constants \( \lambda \) and \( \mu \), does not (1) possess a positive entire solution? — We will answer this problem partially here. That is, when \( \alpha \beta > 1 \), we can obtain a Liouville type theorem for nonnegative entire solutions of system (1) to the effect that (1) cannot possess nonnegative entire solutions \((u, v)\) except for the trivial one \((u, v) \equiv (0, 0)\) if it satisfies a kind of growth condition at \( \infty \).

Our result is as follows:

**Theorem 1.** (i) Let \( \alpha \beta > 1 \), \( 0 < \alpha < 1 \), and (2) hold for some constants \( \lambda \) and \( \mu \) satisfying
\[
\lambda \leq \alpha(2 - \mu) + 2.
\]
If \((u, v)\) is a nonnegative entire solution of (1) satisfying
\[
u(x) = O(\exp |x|^\rho) \quad \text{as} \quad |x| \to \infty \quad \text{for some} \quad \rho > 0,
\]
then \((u, v) \equiv (0, 0)\).

(ii) Let \( \alpha \beta > 1 \), \( 0 < \beta < 1 \), and (2) hold for some constants \( \lambda \) and \( \mu \) satisfying
\[
\mu \leq \beta(2 - \lambda) + 2.
\]
If \((u, v)\) is a nonnegative entire solution of (1) satisfying
\[
u(x) = O(\exp |x|^\rho) \quad \text{as} \quad |x| \to \infty \quad \text{for some} \quad \rho > 0,
\]
then \((u, v) \equiv (0, 0)\).

Since the positive radial entire solutions \((u, v)\) constructed in [4] under the assumption of (ii) of Theorem 0 have the asymptotic growth
\[
u, \quad v = O(|x|^k) \quad \text{as} \quad |x| \to \infty \quad \text{for some} \quad k > 0,
\]
the assumption of Theorem 1 is best possible in some sense.

2. Preliminary lemmas.

Let \( w \) be a continuous function in \( \mathbb{R}^N \). We denote by \( \overline{w}(r) \), \( r \geq 0 \), the average of \( w(x) \) over the sphere \( |x| = r \), that is,
\[
\overline{w}(r) = \frac{1}{\omega_N r^{N-1}} \int_{|x|=r} w(x) dS,
\]
where \( \omega_N \) denotes the surface area of the unit sphere in \( \mathbb{R}^N \).

The next lemma is needed in proving Theorem 1:
**Lemma 2.** Let $\beta > 1$, $(u, v)$ be a nonnegative entire solution of \((1)\), and $b \in (0, 1)$ a constant. Then its spherical mean $(\bar{u}, \bar{v})$ satisfies the ordinary differential inequalities
\[
\begin{align*}
\pi'(r) &\geq \tilde{C} r p_\ast(r) \pi(br) \alpha, \quad r > 0, \quad \pi'(0) = 0, \\
(r^{N-1} \overline{\pi}'(r))' &\geq r^{N-1} \hat{q}(r) \overline{\pi}(r) \beta, \quad r > 0, \quad \overline{\pi}'(0) = 0,
\end{align*}
\]
where $\tilde{C} = \tilde{C}(N, \alpha, b) > 0$ is a constant and
\[
p_\ast(r) = \min_{|x| \leq r} p(x), \quad r \geq 0,
\]
and
\[
\hat{q}(r) = \left( \frac{1}{\omega_N r^{N-1}} \int_{|x| = r} \frac{dS}{q(x)^\beta} \right)^{-\beta/\beta'}, \quad r \geq 0
\]
with $1/\beta + 1/\beta' = 1$.

To prove Lemma 2, we prepare the following lemma; see [1, p. 244] or [3, p. 225].

**Lemma 3.** Let $D$ be a domain in $\mathbb{R}^N$. Suppose that $\sigma > 0$ is a constant and $x_0 \in D$ and $r > 0$ satisfy $B_{2r}(x_0) \equiv \{ x : |x - x_0| \leq 2r \} \subset D$. Then, we can find a constant $C = C(N, \sigma) > 0$ satisfying
\[
\left( \max_{B_{2r}(x_0)} u \right)^\sigma \leq \frac{C}{r^N} \int_{B_{2r}(x_0)} u^\sigma \, dx
\]
for any function $u \in C^2(D)$ satisfying $u \geq 0$ and $\Delta u \geq 0$ in $D$.

**Proof of Lemma 2.** Let $(u, v)$ be a nonnegative entire solution of \((1)\). Since $\beta > 1$, one can prove \((6)\) easily by the same computation as was used in [2, p. 508]. We will prove the validity of \((5)\). By taking the mean value of the first equation of \((1)\), we have
\[
\begin{align*}
(r^{1-N} (r^{N-1} \overline{\pi}'(r)))' &= \frac{1}{\omega_N r^{N-1}} \int_{|x| = r} p(x)v(x)^\alpha \, dS, \quad r \geq 0.
\end{align*}
\]
Since an integration of \((6)\) shows that $\overline{\pi}$ is nondecreasing on $[0, \infty)$, we may assume $b > 1/2$ in \((5)\). Put $b = 1 - a$, $a \in (0, 1/2)$. Integrating \((7)\) over $[0, r]$, we have
\[
\begin{align*}
\overline{\pi}'(r) &= \frac{1}{\omega_N r^{N-1}} \int_{|x| \leq r} p(x)v(x)^\alpha \, dx \\
&\geq \frac{p_\ast(r)}{\omega_N r^{N-1}} \int_{|x| \leq r} v(x)^\alpha \, dx.
\end{align*}
\]
Let $r > 0$ be fixed. We take $y \in \mathbb{R}^N$ such that
\[
v(y) = \max_{|x| = (1-a)r} v(x) \left( = \max_{|x| \leq (1-a)r} v(x) \right),
\]

and take \( z \in \mathbb{R}^N \) such that \( z = My, \ 0 < M < 1 \) and \( |y - z| = ar \). Then we can see that
\[
\int_{|x|\leq r} v^{\alpha} \, dx \geq \int_{|x-z|\leq 2ar} v^{\alpha} \, dx.
\]
Using Lemma 3, we obtain
\[
\int_{|x-z|\leq 2ar} v^{\alpha} \, dx \geq C_0 r^N \left( \max_{|x-z|\leq ar} v(x) \right)^{\alpha}
= C_0 r^N [v(y)]^{\alpha}
= C_0 r^N \left( \max_{|x|\leq (1-a)r} v(x) \right)^{\alpha}
\geq C_0 r^N \tau((1-a)r)^{\alpha},
\]
where \( C_0 = C_0(N, \alpha, a) > 0 \) is a constant. From this estimate and (8) we obtain (5). This completes the proof.

3. Proof of Theorem 1.

This section is entirely devoted to proving our Theorem 1. Assume that (2) hold. Then there exist positive constants \( K_1, K_2 \) and \( r_0 \) such that
\[
(9) \quad p(x) \geq \frac{K_1}{|x|^\lambda}, \quad q(x) \geq \frac{K_2}{|x|^{\mu}} \quad \text{for} \ |x| \geq r_0.
\]

Proof of Theorem 1. We prove only the statement (i); the proof of (ii) is similar. It suffices to treat the case that \( \lambda = \alpha(2-\mu)+2 \). The proof is done by contradiction. Suppose to the contrary that (1) has a nonnegative nontrivial entire solution \((u,v)\) satisfying (4). Then, by Lemma 2, its spherical mean \((\overline{u}, \overline{v})\) satisfies (5) and (6).

Let \( m > 1 \) be a number satisfying
\[
(10) \quad 1 < m < (1+\delta)^{1/\rho}, \quad \delta = \frac{\alpha\beta - 1}{\beta + 2},
\]
where \( \rho \) is the number appearing in (4). We choose the constant \( b \) in (5) such that \( 1/m < b < 1 \). The proof is decomposed into three steps.

Step 1. We show that
\[
(11) \quad \lim_{r \to \infty} \overline{u}(r) = \infty.
\]

Integrating (5) and (6) on \([0, r]\), we see that \( \overline{u} \) and \( \overline{v} \) are nondecreasing functions on \([0, \infty)\), and
\[
(12) \quad \overline{u}(r) \geq \overline{u}(0) + C \int_0^r sp(s) \overline{v}(bs)^{\alpha} \, ds, \quad r \geq 0,
\]
and
\begin{align}
(13) \quad \bar{v}(r) & \geq \bar{v}(0) + \frac{1}{N-2} \int_0^r s \bar{q}(s) \left[ 1 - \left( \frac{s}{r} \right)^{N-2} \right] \bar{u}(s)^\beta ds, \quad r \geq 0, \\
\end{align}
respectively. For some point \( x^* \in \mathbb{R}^N \) we have \( u(x^*) > 0 \) or \( v(x^*) > 0 \); that is \( \bar{u}(r^*) > 0 \) or \( \bar{v}(r^*) > 0 \), \( r^* = \| x^* \| \). Therefore we see from (12) and (13) that \( \bar{u}(r) > 0, \bar{v}(r) > 0 \) for \( r > r^* \). We may assume that \( r_0 > r^* \). From (9) and the monotonicity of \( \bar{u} \), we have
\[ \bar{u}(r) > 0, \quad \bar{v}(r) > 0 \quad \text{for} \quad r > r^*. \]
Accordingly we observe that, if \( \lambda \leq 2 \), then
\[ \bar{u}(r) \geq K_1 \bar{u}(r_0)^\alpha \int_{r_0/b}^r s^{1-\lambda} ds, \quad r \geq r_0/b. \]
for some constant \( K_1 > 0 \) and \( r \geq r_1 > 2r_0/b \). Therefore (11) holds if \( \lambda \leq 2 \).

It remains to consider the case of \( \lambda > 2 \). Since in this case \( \mu < 2 \), from (9) and (13), we have
\begin{align}
(14) \quad \bar{v}(r) & \geq \bar{v}(0) + \frac{1}{N-2} \int_{r_0}^r s \bar{q}(s) \left[ 1 - \left( \frac{s}{r} \right)^{N-2} \right] \bar{u}(s)^\beta ds \\
& \geq \bar{v}(0) + \frac{K_2 \bar{u}(r_0)^\beta}{N-2} \int_{r_0}^r s^{1-\mu} \left[ 1 - \left( \frac{s}{r} \right)^{N-2} \right] ds \\
& \geq \bar{v}(0) + \frac{K_2 \bar{u}(r_0)^\beta}{N-2} \left[ 1 - \left( \frac{1}{2} \right)^{N-2} \right] \int_{r_0}^{r/2} s^{1-\mu} ds \\
& \geq C_1 r^{2-\mu}, \quad r \geq r_2 > 2r_0
\end{align}
for some constant \( C_1 > 0 \). Let \( r \geq r_3 > r_2/b \). From (14) and (12), we have
\[ \bar{u}(r) \geq \bar{u}(0) + \bar{C} \int_0^r s \bar{q}(s) \bar{v}(bs)^\alpha ds \]
\[ \geq \bar{u}(0) + \bar{C} K_1 C_1^\alpha b^{\alpha(2-\mu)} \int_{r_3}^r s^{1-\lambda+\alpha(2-\mu)} ds \]
\[ = \bar{u}(0) + \bar{C} K_1 C_1^\alpha b^{\alpha(2-\mu)} \int_{r_3}^r s^{-1} ds \]
\[ \geq C_2 \log r, \quad r \geq r_4 > 2r_3 \]
for some constant \( C_2 > 0 \). Thus we obtain (11).

\textbf{Step 2.} We show that
\[ \bar{u}(mr) \geq M \bar{u}(r)^{\beta+1} \quad \text{near} \quad + \infty \]
for some constant \( M > 0 \), where \( m \) is the number appearing in (10).
Let us fix \( R > r_5 > \max\{r_1, r_4\} \) arbitrarily for a moment. Integrating (5) and (6) over \([R,r]\), we have

\[
\bar{u}(r) \geq \bar{u}(R) + \tilde{C} \int_R^r sp_*(s)\bar{v}(bs)^\alpha ds, \quad r \geq R,
\]

and

\[
\bar{v}(r) \geq \bar{v}(R) + \frac{1}{N-2} \int_R^r s \left[ 1 - \left( \frac{s}{r} \right)^{N-2} \right] \hat{q}(s)\bar{u}(s)^\beta ds, \quad r \geq R,
\]

respectively. Using (9) and the inequality

\[
s \left[ 1 - \left( \frac{s}{r} \right)^{N-2} \right] \geq \frac{N-2}{mN-2}(r - s), \quad R \leq s \leq r \leq mR,
\]

we have

\[
(16) \quad \bar{u}(r) \geq C_3 R^{1-\lambda} \int_R^r \bar{v}(bs)^\alpha ds, \quad R \leq r \leq mR,
\]

and

\[
(17) \quad \bar{v}(r) \geq C_4 R^{-\mu} \int_R^r (r - s)\bar{u}(s)^\beta ds, \quad R \leq r \leq mR,
\]

where \( C_3 \) and \( C_4 \) are positive constants independent of \( r \) and \( R \). Now let us define the functions \( f(r; R) \) and \( g(r; R) \) for \( R \leq r \leq mR \), by the right hand sides of (16) and (17), respectively. Then \( f \) and \( g \) satisfy \( f(R; R) = g(R; R) = 0 \). We denote simply \( f(r; R) = f(r) \) and \( g(r; R) = g(r) \), when there is no ambiguity. We then have

\[
(18) \quad f'(r) = C_3 R^{1-\lambda} \bar{v}(br)^\alpha \geq C_3 R^{1-\lambda} g(br)^\alpha, \quad b^{-1} R \leq r \leq mR;
\]

\[
g'(r) = C_4 R^{-\mu} \int_R^r \bar{u}(s)^\beta ds \geq 0, \quad R \leq r \leq mR; \quad g'(R) = 0,
\]

and

\[
(19) \quad g''(r) = C_4 R^{-\mu} \bar{u}(r)^\beta \geq C_4 R^{-\mu} f(r)^\beta, \quad R \leq r \leq mR.
\]
Multiplying (18) by \( g'(r) \geq 0 \) and integrating by parts the resulting inequality on \([b^{-1}R, r]\), we have

\[
f(r)g'(r) - f(b^{-1}R)g'(b^{-1}R) - \int_{b^{-1}R}^{r} f(s)g''(s)ds \\
\geq C_3 R^{1-\lambda} \int_{b^{-1}R}^{r} g(bs)^\alpha g'(s)ds \\
\geq C_3 R^{1-\lambda} \int_{b^{-1}R}^{r} g(bs)^\alpha g'(bs)ds \\
= \frac{C_3}{b(\alpha + 1)} R^{1-\lambda} g(br)^{\alpha + 1}, \quad b^{-1}R \leq r \leq mR.
\]

Hence

\[
f(r)g'(r)^\beta \geq C_5 R^{(1-\lambda)\beta} g(br)^{(\alpha + 1)\beta}, \quad b^{-1}R \leq r \leq mR,
\]

where \( C_5 = \left( \frac{C_3}{b(\alpha + 1)} \right)^\beta \). From now on, we use \( C \) to denote various positive constants independent of \( r \) and \( R \), as we will have no confusion. Combining this inequality with (19), we obtain

\[
g''(r)g'(r)^\beta \geq CR^{(1-\lambda)\beta - \mu} g(br)^{(\alpha + 1)\beta}, \quad b^{-1}R \leq r \leq mR.
\]

Multiplying this inequality by \( g'(r) \geq 0 \) and integrating over \([b^{-1}R, r]\), we have

\[
g'(r) \geq CR^{-(1-\lambda)(\beta + 1) - \mu} g(br)^{\delta + 1}, \quad b^{-1}R \leq r \leq mR.
\]

Let \( \varepsilon > 0 \) be sufficiently small fixed number. Integrating this relation over \([1 + \varepsilon R, mR]\), we see that

\[
(20) \quad g(mR; R) \geq CR^{(2-\lambda)\beta + 2 - \mu} g((1 + \varepsilon)R; R)^{\delta + 1}.
\]

On the other hand, from the definition of \( g \) and the monotonicity of \( \bar{u} \), we have

\[
(21) \quad g(mR; R) = \frac{C_4}{R^\mu} \int_{R}^{mR} (mR - s) \bar{u}(s)^\beta ds \\
\leq \frac{C_4}{R^\mu} \bar{u}(mR)^\beta \int_{R}^{mR} (mR - s)ds \\
= \frac{C_4(m - 1)^2}{2} R^{2-\mu} \bar{u}(mR)^\beta, \quad R > r_5,
\]
and

\begin{align}
g((1 + \varepsilon)R; R) &= \frac{C_4}{R^\mu} \int_R^{(1+\varepsilon)R} ((1 + \varepsilon)R - s) \pi(s)^\beta ds \\
&\geq \frac{C_4 \pi(R)^\beta}{R^\mu} \int_R^{(1+\varepsilon)R} ((1 + \varepsilon)R - s) ds \\
&= \frac{C_4 \varepsilon^2}{2} R^{2-\mu} \pi(R)^\beta, \quad R > r_5.
\end{align}

From (20), (21) and (22) we observe that

\begin{align}
\pi(mR) &\geq CR \frac{(2-\mu)\alpha + 2 - \lambda}{\beta + 2} \pi(R)^{\delta + 1} \\
&= C\pi(R)^{\delta + 1}, \quad R > r_5.
\end{align}

This implies that (15) holds.

**Step 3.** This is the final step. Let \( \tilde{r} \) be so large that

\begin{align}
M^{1/\delta} \pi(\tilde{r}) \geq e,
\end{align}

and

\begin{align}
\pi(m \tilde{r}) \geq M \pi(\tilde{r})^{1+\delta}, \quad r \geq \tilde{r}
\end{align}

hold, where \( M > 0 \) is the constant appearing in (15). This choice of \( \tilde{r} \) is possible by Steps 1 and 2. For \( l \in \mathbb{N} \) we obtain from (24)

\begin{align}
\pi(m^l \tilde{r}) &\geq M \pi(m^{l-1} \tilde{r})^{1+\delta} \\
&\geq M^{1+(1+\delta)} \pi(m^{l-2} \tilde{r})^{(1+\delta)^2} \\
&\geq \ldots \ldots \\
&\geq M^{1+(1+\delta)+\ldots+(1+\delta)^{l-1}} \pi(\tilde{r})^{(1+\delta)^l} \\
&= M^{-1/\delta} \left[ M^{1/\delta} \pi(\tilde{r}) \right]^{(1+\delta)^l}.
\end{align}

Thus (23) yields

\begin{align}
\pi(m^l \tilde{r}) &\geq M^{-1/\delta} \exp\{(1 + \delta)^l\}.
\end{align}

Let \( r \geq m \tilde{r} \). Then we can find a unique positive integer \( l = l(r) \) satisfying

\( m^l \tilde{r} \leq r < m^{l+1} \tilde{r} \), that is,

\( l > \frac{\log r - \log \tilde{r}}{\log m} - 1. \)
It follows therefore from (25) that
\[
\begin{align*}
\overline{u}(r) &\geq \overline{u}(m^{1/\delta}r) \geq M^{-1/\delta} \exp \left\{ (1 + \delta) \right\} \\
&\geq M^{-1/\delta} \exp \left\{ (1 + \delta) \frac{\log r}{\log m} - 1 \cdot (1 + \delta) \frac{\log r}{\log m} \right\} \\
&= M^{-1/\delta} \exp \left\{ (1 + \delta) \frac{\log r}{\log m} - 1 \cdot \frac{\log(1+\delta)}{\log m} \right\}.
\end{align*}
\]

On the other hand, because \( u(x) = O(\exp \|x\|^{\rho} \) as \( |x| \to \infty \), we obviously have
\[
\overline{u}(r) = O(\exp r^{\rho}) \quad \text{as} \quad r \to \infty.
\]
Since \( \log(1+\delta)/\log m > \rho \) from our choice of \( m \), (26) gives a contradiction. Therefore \( u \equiv v \equiv 0 \) in \( \mathbb{R}^N \). The proof is finished.

References


Received June 5, 2000.

Department of Mathematics  
Faculty of Science  
Hiroshima University  
Higashi-Hiroshima 739-8526  
JAPAN  
E-mail address: teramoto@math.sci.hiroshima-u.ac.jp

Department of Mathematics  
Faculty of Integrated Arts and Sciences  
Hiroshima University  
Higashi-Hiroshima 739-8521  
JAPAN  
E-mail address: usami@mis.hiroshima-u.ac.jp
Guidelines for Authors

Authors may submit manuscripts at pjm.math.berkeley.edu/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use \LaTeX, but papers in other varieties of \TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as \LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of 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.