The Journal of Geometric Analysis
Volume 13, Number 4, 2003

# A Sharp Bound for the Degree of Proper Monomial Mappings Between Balls

By *John P. D'Angelo, Šimon Kos, and Emily Riehl*

*ABSTRACT. The authors prove that a proper monomial holomorphic mapping from the two-ball to the N-ball has degree at most $2N - 3$, and that this result is sharp. The authors first show that certain group-invariant polynomials (related to Lucas polynomials) achieve the bound. To establish the bound the authors introduce a graph-theoretic approach that requires determining the number of sinks in a directed graph associated with the quotient polynomial. The proof also relies on a result of the first author that expresses all proper polynomial holomorphic mappings between balls in terms of tensor products.*

## 1. Introduction

The purpose of this article is to demonstrate a sharp relationship between the degree of a polynomial $p$ (satisfying a certain natural condition described below) and the number of distinct monomials in $p$. This result implies a special case of a conjecture concerning proper holomorphic mappings between balls in complex Euclidean spaces of different dimensions.

After stating the main result, we continue the introduction by describing the connection between this result and the more general situation.

**Theorem 1.1.** *Suppose that $p$ is a polynomial in two real variables $(x, y)$ such that*

1) *$p(x, y) = 1$ on the set $x + y = 1$, and*

2) *each coefficient of $p$ is nonnegative.*

*Let $N$ be the number of distinct monomials in $p$, and let $d$ be the degree of $p$. Then $d \leq 2N - 3$ and this result is sharp. (That is, for each $N \geq 2$, there is a polynomial satisfying 1) and 2) whose degree is $2N - 3$.*

**Corollary 1.2.** *Suppose that $f$ is a proper holomorphic monomial mapping from the unit ball in two complex dimensions to the unit ball in $N$ complex dimensions. Then the degree of $f$ does not exceed $2N - 3$, and this result is sharp.*

*Math Subject Classifications.* 32B99, 32H02, 11B309.
*Key Words and Phrases.* Proper holomorphic mappings, unit ball, Lucas polynomials.
*Acknowledgements and Notes.* The first author was supported by NSF grant DMS 0200551; the third author was supported by the same grant with a REU for the summer of 2002. The second author was supported by NSF grant DMR-94-24511.


ISSN 1050-6926

In Section 2 we give a (non-obvious) family of polynomials for which $d = 2N - 3$. These polynomials appear in [3]; they yield group-invariant proper mappings between balls. A one-variable version of these polynomials goes back to Lucas; see [8] for properties of Lucas polynomials in one variable. In Section 3 we use graph-theoretic techniques to show that $d \leq 2N - 3$. The methods of Section 3 apply more generally, but we do not discuss such generalizations in this article. Two comments justify the article in its current form where $n = 2$ and $f$ is a monomial. First, the graph-theoretic approach from Section 3 yields considerably sharper information than has yet been obtained by using CR vector fields or Chern–Moser invariants. Second, there is an abundance of monomial examples; for $N \geq 4$ there are already infinitely many inequivalent examples, and as $N$ increases to infinity the dimension of the parameter space of monomial examples is unbounded.

We now make the connection to proper holomorphic mappings between balls. Let $\mathbf{C}^n$ denote complex Euclidean space of dimension $n$, and let $||z||^2 = \sum |z_j|^2$ denote the Euclidean norm of $z$. The unit ball $B_n$ is defined to be the set of $z$ for which $||z|| < 1$.

Let $f : \mathbf{C}^n \to \mathbf{C}^N$ be a (holomorphic) polynomial mapping. Then $f$ is a proper mapping from $B_n$ to $B_N$ if and only if $||f(z)|| < 1$ for $||z|| < 1$ and

$$||f(z)||^2 = 1 \tag{1.1}$$

whenever $||z||^2 = 1$. It is natural to seek all such examples; more generally one wants to find all rational proper mappings between balls. In this article we restrict our consideration to polynomials.

Let now $f : B_n \to B_N$ be a polynomial mapping that satisfies (1.1) and whose component functions are monomials. We write, using multi-index notation,

$$f(z) = \left(\ldots, c_\alpha z^\alpha, \ldots\right) .$$

Then (1.1) simplifies to the condition

$$\sum_\alpha |c_\alpha|^2 |z|^{2\alpha} = 1 \tag{1.2}$$

on $||z||^2 = 1$. It is natural to write $x_j$ for $|z_j|^2$, and we therefore obtain a polynomial $F$ satisfying

$$F(x) = \sum_\alpha |c_\alpha|^2 x^\alpha = 1 \tag{1.3}$$

on the hyperplane $\sum x_j = 1$. Note that the coefficients of $F$ are nonnegative. Conversely, given a nonconstant polynomial $F$ in $n$ real variables, with nonnegative coefficients, and satisfying (1.3), there is a corresponding proper mapping $f$.

When $n = 2$, it is convenient to write $|z_1|^2 = x$ and $|z_2|^2 = y$. We obtain a polynomial $F$ in two variables; the coefficients of $F$ are nonnegative and $F$ satisfies

$$F(x, y) = \sum |c_{ab}|^2 x^a y^b = 1 \tag{1.4}$$

on $x + y = 1$.

Thus Theorem 1.1 yields Corollary 1.2, and a proper monomial mapping from $B_2$ to $B_N$ has degree at most $2N - 3$. When $N = 2$ as well, we see that the degree is at most 1. This conclusion has been known since the 1970's. Pinchuk [9] proved that a proper holomorphic self-mapping of a strongly pseudoconvex domain is necessarily an automorphism. For the ball

this result was proved earlier by Alexander [1]; the automorphisms of the ball are linear fractional transformations, and hence are rational functions of degree 1. When $N = 3$, the result is also known; Faran [6] found all the rational proper mappings from $B_2$ to $B_3$, and all are of degree at most 3. The first author [3] listed the monomial mappings from $B_2$ to $B_4$, and found that the maximum degree was 5. (One monomial example was inadvertently omitted from that list, but it is of degree 3.) For $N = 5$ Wono [10] found a complicated ad hoc method to show that the degree of a proper monomial mapping is at most 7. There are heuristic reasons to believe that the monomial case is the most complicated, and hence one is tempted to conjecture that a rational proper mapping from $B_2$ to $B_N$ has degree at most $2N - 3$.

The polynomials giving the sharp bound are group-invariant, and were discovered by the first author in [3]. See also [4] and [5] for more information on group-invariant mappings. Perhaps one could somehow use the group-invariance to obtain a simpler proof, but we have been unable to do so. The coefficients arising in these polynomials have many interesting properties, and also arose in the Physics thesis (Section 2.9) of the second author [7].

The situation for $n \geq 3$ is different. The first author has conjectured in that case that the degree of a rational proper mapping between balls cannot exceed $\frac{N-1}{n-1}$. We do not consider $n \geq 3$ here.

In general, the unit sphere is a model for strongly pseudoconvex CR geometry; therefore sharp results for the sphere are likely to extend to the case of strongly pseudoconvex hypersurfaces. Hence the result in this article may indicate how to apply combinatorial considerations to CR geometry.

## 2. Polynomials for which $d = 2N - 3$

First we will exhibit a family of polynomials for which the inequality in Theorem 1.1 is sharp. We define these polynomials by a recurrence formula, although explicit formulas for them are available and appear in [2] and [4].

Put $g_0(x, y) = x$ and $g_1(x, y) = x^3 + 3xy$. Put

$$g_{n+2}(x, y) = \left(x^2 + 2y\right) g_{n+1}(x, y) - y^2 g_n(x, y) \,. \tag{2.1}$$

Now define $p_n(x, y)$ by

$$p_n(x, y) = g_n(x, y) + y^{2n+1} \,. \tag{2.2}$$

Observe that $p_0(x, y) = x + y$ and that $p_1(x, y) = x^3 + 3xy + y^3$. It is easy to check that $p_0$ and $p_1$ have the value one on the line given by $x + y = 1$. We prove by induction that the same holds for each $p_n$. Less obvious is that the coefficients of each $p_n$ are positive integers. We collect the properties of $p_n$ in the following result:

**Proposition 2.1.** *The polynomials $p_n(x, y)$ defined by (2.2) satisfy the following properties:*

1) *The degree $d$ of $p_n$ is $2n + 1$.*

2) *$p_n(x, y) = 1$ on $x + y = 1$.*

3) *Each non-zero coefficient of $p_n$ is a positive integer.*

4) *The number $N$ of nonzero monomials in $p_n$ is $n + 2$. Thus $d = 2N - 3$.*

***Proof.*** Most of this information is proved by induction on $n$, using as the basis step the explicit formulas for $p_0$ and $p_1$. Statement 1) follows easily by this approach, and is left to the reader.

We prove statement 2), using the special cases $p_0$ and $p_1$ as the basis step. Assume then that statement 2) holds for $n$ and $n+1$. We verify it for $n+2$:

$$p_{n+2}(x,y) = g_{n+2}(x,y) + y^{2n+5} = \left(x^2+2y\right) g_{n+1}(x,y) - y^2 g_n(x,y) + y^{2n+5}$$
$$= \left(x^2+2y\right)\left(q_{n+1}(x,y) - y^{2n+3}\right) - y^2\left(q_n(x,y) - y^{2n+1}\right) + y^{2n+5}\,. \quad (2.3)$$

Now set $x = 1-y$ and substitute in (2.3). Using the inductive hypothesis we obtain, after some simplification,

$$p_{n+2}(1-y,y) = \left((1-y)^2+2y\right)\left(1-y^{2n+3}\right) - y^2 + y^{2n+3} + y^{2n+5}$$
$$= \left(1+y^2\right) - y^{2n+3} - y^{2n+5} - y^2 + y^{2n+3} + y^{2n+5} = 1\,.$$

Statement 3) is difficult. First of all note, using induction and the recurrence relation, that the coefficients of $g_n$ (and hence of $p_n$) are integers. The hard part is to show that the nonzero coefficients are positive. It suffices to show this statement for $g_n$. We have the second order recurrence relation

$$g_{n+2} = \left(x^2+2y\right) g_{n+1} - y^2 g_n$$

and the initial conditions $g_0 = x$ and $g_1(x) = x^3 + 3xy$. Using the standard method for solving second order recurrences, we obtain the characteristic roots

$$\lambda = \frac{x^2+2y \pm x\sqrt{x^2+4y}}{2}\,.$$

Therefore we may write

$$g_n = c\left(\frac{x^2+2y+x\sqrt{x^2+4y}}{2}\right)^n + d\left(\frac{x^2+2y-x\sqrt{x^2+4y}}{2}\right)^n$$

for appropriate expressions $c$ and $d$, determined by the initial conditions. We omit the detailed computations and state the result:

$$g_n(x,y)$$
$$= \frac{x+\sqrt{x^2+4y}}{2}\left(\frac{x^2+2y+x\sqrt{x^2+4y}}{2}\right)^n + \frac{x-\sqrt{x^2+4y}}{2}\left(\frac{x^2+2y-x\sqrt{x^2+4y}}{2}\right)^n\,. \quad (2.4)$$

It is not obvious at first glance that (2.4) defines a polynomial; it must of course define a polynomial, because $g_0$ and $g_1$ are polynomials and the recurrence relation has polynomial coefficients. On the other hand, expanding (2.4) results in the cancellation of all terms involving odd powers of $\sqrt{x^2+4y}$ and all the other terms have positive coefficients. Therefore the coefficients of $g_n$ are nonnegative; we already know they are integers, so we have proved 3).

Finally we prove 4). The proof uses the group-invariance of the polynomials $p_n$. Let $\omega$ be a primitive $2n+1$-st root of unity. From (2.4) it is evident that polynomial $p_n$ has the following invariance property:

$$p_n\left(\omega x, \omega^2 y\right) = p_n(x,y)\,.$$

It follows easily from this that the only monomials arising in $p$ must also be invariant under this substitution. Since $p_n$ is of degree $2n+1$, the only possible monomials that can arise are $y^{2n+1}$ and $x^{2n+1-2b}y^b$ for $0 \leq b \leq n$. Assuming that 2) holds, it is proved in [3] and [2] that these $n+2$ monomials all have positive coefficients; therefore there are $n+2$ terms in $p_n$. See p. 175 of [2] for several formulas for these coefficients. □

***Example 2.2.*** For the reader's convenience we list the first few of these polynomials.

$$p_0(x, y) = x + y$$
$$p_1(x, y) = x^3 + 3xy + y^3$$
$$p_2(x, y) = x^5 + 5x^3y + 5xy^2 + y^5$$
$$p_3(x, y) = x^7 + 7x^5y + 14x^3y^2 + 7xy^3 + y^7$$
$$p_4(x, y) = x^9 + 9x^7y + 27x^5y^2 + 30x^3y^3 + 9xy^4 + y^9 .$$

***Remark 2.3.*** The proof used the invariance of $p_n$ under the substitution $(x, y) \to (\omega x, \omega^2 y)$ where $\omega^{2n+1} = 1$. In fact these polynomials were discovered by seeking proper polynomial mappings invariant under various representations of cyclic groups. See [4] and [5].

## 3. Proof of Theorem 1.1

The proof here combines an idea about labeled Newton diagrams with two facts, proved in much more generality in [2]. These facts are special cases of general statements about proper polynomial mappings between balls. Here we state and prove them only in the simple case needed in this article.

**Notation.** We let $\mathcal{P}$ denote the collection of polynomials in two variables satisfying 1) and 2) of Theorem 1.1. We also sometimes write $s$ for $x + y$.

**Lemma 3.1.** *If $p \in \mathcal{P}$ is homogeneous of degree $d$, then $p(x, y) = (x + y)^d = s^d$.*

***Proof.*** Since $p \in \mathcal{P}$, we have $p(x, y) = 1$ on the line $x + y = 1$. Therefore $p(x, y) = 1 = (x + y)^d$ on this line. By homogeneity, the equality $p(x, y) = (x + y)^d$ holds everywhere. □

Let $p$ be a polynomial. We can write $p = \sum p_j$, where $p_j$ denotes the sum of the monomials of degree $j$ in $p$. Thus $p_j$ is homogeneous of degree $j$; the resulting formula is called the expansion of $p$ into homogeneous parts.

**Lemma 3.2.** *Suppose that $p \in \mathcal{P}$ has degree $d$. Let $p = \sum p_j$ denote the expansion of $p$ into homogeneous parts. Then*

$$\sum p_j s^{d-j} = \sum_{j=0}^{d} p_j(x, y)(x + y)^{d-j} = (x + y)^d = s^d . \tag{3.1}$$

***Proof.*** The polynomial on the far left-hand side in (3.1) has the value 1 on the line $s = 1$, because it agrees with $p$ there. It is also homogeneous; by Lemma 3.1 it must be $s^d$. □

For later purposes we provide an intuitive explanation of Lemma 3.2. Starting with $h = s^d$ one constructs $p$ by a finite number of operations of the following form: replace the expression

$c_{ab}(x^{a+1}y^b + x^a y^{b+1})$ by $c_{ab}x^a y^b$, and keep the other terms the same. This operation is a special case of the *undoing* of a tensor product operation discussed in [2]. The notion of undoing determines a partial ordering on $\mathcal{P}$; we say that $p$ is an *ancestor* of $p'$, or that $p'$ is a *descendant* of $p$, if we can reach $p'$ from $p$ by finitely many such operations. For a fixed degree $d$, the mapping $h = s^d$ is maximal, and by Lemma 3.4 is the initial ancestor of each element of $\mathcal{P}$ of degree $d$.

The polynomial $p$ given by $x^3 + 3xy + y^3$ is obtained by undoing $(x + y)^3$ once. It is easy to prove that $p$ cannot be obtained by starting with 1 and replacing factors of 1 with $x + y$. Thus $p$ is not an ancestor of 1.

***Definition 3.3.*** For $p \in \mathcal{P}$, we define its *quotient* $q_p$ by

$$q_p(x, y)(x + y - 1) = p(x, y) - 1 . \tag{3.2}$$

It is elementary to see that $q_p$ is a polynomial. When $h(x, y) = (x + y)^d$, one sees easily that

$$q_h(x, y) = 1 + (x + y) + \ldots + (x + y)^{d-1} .$$

The main idea in this proof is to compare $q_p$ with $q_h$. There is an algebraic part and a graph-theoretic part of this relationship. Recall that $p_j$ denotes the sum of the monomials of degree $j$ in $p$. For each $(a, b)$ we let $Q_{ab}(p)$ denote the coefficient of $x^a y^b$ in $q_p$. First we note the following elementary result.

**Lemma 3.4.** *Suppose $p \in \mathcal{P}$ is of degree $d$, and $h(x, y) = (x + y)^d$. Then*

$$q_h = q_p + \sum_{j=0}^{d-1} p_j \left( \sum_{m=0}^{d-j-1} s^m \right) . \tag{3.3}$$

*It follows, for each $(a, b)$, that*

$$Q_{ab}(p) \leq Q_{ab}(h) . \tag{3.4}$$

***Proof.*** From (3.1) we obtain

$$h(x, y) - 1 = (x + y)^d - 1 = p(x, y) - 1 - \sum_{j=0}^{d} p_j(x, y) \left(1 - (x + y)^{d-j}\right) . \tag{3.5}$$

Dividing both sides of (3.5) by $x + y - 1$ yields the formula

$$\begin{aligned} q_h(x, y) &= q_p(x, y) - \frac{\sum_{j=0}^{d-1} p_j(x, y) \left(1 - (x + y)^{d-j}\right)}{(x + y - 1)} \\ &= q_p(x, y) + \frac{\sum_{j=0}^{d-1} p_j(x, y) \left(1 - (x + y)^{d-j}\right)}{(1 - (x + y))} . \end{aligned} \tag{3.6}$$

Formula (3.6) and the formula for the finite sum of a geometric series yield (3.3). Since each term in the sum in (3.3) has a positive coefficient, inequality (3.4) follows. □

We next turn to the graph-theoretic aspect, which enables us to see geometrically how $q_p$ determines $p$. Given $p \in \mathcal{P}$, we determine from $q_p$ a labeled Newton diagram, written $G(p)$, in

the following fashion. Let $(a, b)$ be a lattice point with $a$ and $b$ nonnegative. We assign $P$ to it when $Q_{ab}(p) > 0$; we assign $N$ to it when $Q_{ab}(p) < 0$, and we assign 0 to it when $Q_{ab}(p) = 0$.

We call $(a', b')$ an immediate predecessor of $(a, b)$ if $a'$ and $b'$ are nonnegative and either $(a, b) = (a' + 1, b')$ or $(a, b) = (a', b' + 1)$.

***Definition 3.5.*** A lattice point $(a, b)$ in $G(p)$ is called a sink if either S1) or S2) holds:

S1) $Q_{ab}(p) < 0$ but $Q_{(a-1)b}(p) \geq 0$ and $Q_{a(b-1)}(p) \geq 0$.

S2) $Q_{ab}(p) = 0$ but $Q_{(a-1)b}(p) \geq 0$ and $Q_{a(b-1)}(p) \geq 0$ and at least one of these is strictly positive.

In other words, a lattice point $m = (a, b)$ is a sink if either S1') or S2') holds:

S1') $m$ has the label $N$, and its immediate predecessors have the labels $P$ or 0.

S2') $m$ has the label 0, its immediate predecessors have the labels $P$ or 0, and at least one of these is a $P$.

We sketch all possible pictures of sinks, using the $P$, $N$, 0 notation. Each picture shows only that part of the Newton diagram near the sink, which is located at the top right. The sinks in (3.7) satisfy S1), while those in (3.8) satisfy S2).

$$\begin{pmatrix} P & N \\ & P \end{pmatrix} \quad \begin{pmatrix} P & N \\ & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & N \\ & P \end{pmatrix} \quad \begin{pmatrix} 0 & N \\ & 0 \end{pmatrix} \tag{3.7}$$

$$\begin{pmatrix} P & 0 \\ & P \end{pmatrix} \quad \begin{pmatrix} 0 & 0 \\ & P \end{pmatrix} \quad \begin{pmatrix} P & 0 \\ & 0 \end{pmatrix}. \tag{3.8}$$

There is a similar definition of source, where all the signs in the definition of sink are reversed. We picture one source that is relevant in this problem.

$$\begin{pmatrix} 0 & P \\ & 0 \end{pmatrix}. \tag{3.9}$$

We show how these pictures work for two elements of $\mathcal{P}$.

***Example 3.6.*** Let $p(x, y) = x + xy + y^2$. Then $q_p(x, y) = 1 + y$. Therefore $G(p)$ has strictly positive coefficients at $(0, 0)$ and at $(0, 1)$, and vanishing coefficients at $(1, 0)$, $(2, 0)$, $(1, 1)$, and $(0, 2)$. There is one source at $(0, 0)$, and there are three sinks, at $(0, 2)$, $(1, 1)$, and $(1, 0)$. Notice that the point $(2, 0)$ satisfies neither S1) nor S2).

***Example 3.7.*** Let $p(x, y) = x^3 + 3xy + y^3$. Thus $p$ is the simplest case from Section 2 where $d = 2N - 3$. Here $q_p(x, y) = 1 + x + y + x^2 - xy + y^2$. We assign $N$ to $(1, 1)$, and $P$ to the other lattice points $(a, b)$ with $a + b \leq 2$. We obtain a sink satisfying S1) at $(1, 1)$ and sinks satisfying S2) at $(3, 0)$ and $(0, 3)$. We illustrate the diagram by the following picture, where we have put the sinks in bold:

$$\begin{pmatrix} \mathbf{0} & & & \\ P & 0 & & \\ P & \mathbf{N} & 0 & \\ P & P & P & \mathbf{0} \end{pmatrix}. \tag{3.10}$$

**Proposition 3.8.** *Let $p \in \mathcal{P}$, and let $G(p)$ denote its labeled diagram.*

1) *If $G(p)$ has a sink at $(a, b)$, then the coefficient of $x^a y^b$ in $p$ must be positive.*

2) *The number of terms in $p$ is at least as large as the number of sinks.*

3) *Then $G(p)$ has a unique source, which is located at $(0, 0)$.*

***Proof.*** 1). Suppose S1) holds. The negative coefficient at $(a, b)$ get multiplied by $-1$ and hence contributes positively in determining the coefficient $\lambda$ of $x^a y^b$ in $p$. The nonnegative coefficients of the immediate predecessors contribute nonnegatively to $\lambda$. Thus $\lambda > 0$.

Suppose S2) holds. The zero coefficient at $(a, b)$ does not contribute to the coefficient $\lambda$ of $x^a y^b$ in $p$; if either immediate predecessor is positive, and both are nonnegative, then $\lambda > 0$.

In either case a sink at $(a, b)$ forces a positive coefficient of $x^a y^b$ in $p$.

2) follows from 1).

3) The source at $(0, 0)$ is necessary to account for the $-1$ in $p - 1$. Any other source would force $p$ to have a negative coefficient, contradicting the definition of $\mathcal{P}$. This statement follows from the same argument used to prove 1), with all signs switched. □

Let $h(x + y) = (x + y)^d$. The diagram $G(h)$ is easy to describe. All lattice points $(a, b)$ with $a + b < d$ have positive coefficients. There is one source, at the origin. There are $d + 1$ sinks, at the lattice points $(a, b)$ with $a + b = d$. Each of these sinks is of type S2). The main idea in this part of the article is to relate $G(p)$ to $G(h)$.

Since $q_p$ is of degree $d - 1$, all points $(a, b)$ with $a + b = d$ are assigned 0. In many examples $G(p)$ has many additional zeroes. We define the *boundary* of $G(p)$ to be the collection of $d + 1$ points $(a, b)$ such that

B1) $(a', b')$ is assigned 0 whenever $a' \geq a$ and $b' \geq b$.

B2) In each row and column we choose the minimal $(a, b)$ satisfying B1).

We give an example, where we have written the boundary in bold.

$$\begin{pmatrix} \mathbf{0} & & & & \\ P & \mathbf{0} & & & \\ P & \mathbf{0} & 0 & & \\ P & P & \mathbf{0} & 0 & \\ P & P & \mathbf{0} & 0 & 0 \end{pmatrix}. \tag{3.11}$$

For later convenience we consider those lattice points with negative coefficients in $G(p)$. Recall that the degree of $p$ is $d$. We say that a lattice point with nonpositive coefficient is *connected to the boundary* if there is a collection of lattice points $L_j$ for $j = 0, 1, \ldots, k$ such that

1) $L_0 = (a, b)$, and if $L_k = (a', b')$, then $a' + b' = d - 1$.

2) $Q_p(L_j) \leq 0$ for $0 \leq j < k - 1$.

3) $L_{j+1} = L_j + (1, 0)$ or $L_{j+1} = L_j + (0, 1)$ for each $j$.

We say that a lattice point with nonpositive coefficient is *separated from the boundary* if it is not connected to the boundary. This concept is not strictly needed for the proof; it is useful for the following reason: although the procedure of Lemma 3.2 can create sinks, it cannot create sinks separated from the boundary. The simplest example of a sink separated from the boundary is a

lattice point $(a, b)$ with negative coefficient, but where the coefficients of $(a+1, b)$ and $(a, b+1)$ are both positive. The relevant part of the picture here is:

$$\begin{pmatrix} P & \\ N & P \end{pmatrix}. \tag{3.12}$$

The diagrams $G(h)$ and $G(p)$ are related via a propagation of sinks. The next observation follows immediately from Lemma 3.4.

**Observation.** For $p \in \mathcal{P}$ the diagram $G(p)$ is obtained from $G(h)$ by changing some of the $P$'s to $N$'s or $O$'s.

We remarked above that the procedure of Lemma 3.2 will not create sinks separated from the boundary; using this remark one could express the proof of Proposition 3.11 below slightly differently. This remark bears also on the following warning.

**Warning.** By changing some $P$'s in $G(h)$ to $N$'s for example, we generally end up with (the diagram of) a polynomial with some negative coefficients. The observation asserts only that, for $p \in \mathcal{P}$, we obtain $G(p)$ in this way.

**Lemma 3.9.** *Suppose that $p \in \mathcal{P}$ and $p$ has degree $d$. Then there are $A, B$ with $0 \leq A, B \leq d-1$ such that $(a, 0)$ is assigned $P$ for $0 \leq a \leq A$ and assigned 0 for $A < a \leq d$, and $(0, b)$ is assigned $P$ for $0 \leq b \leq B$ and assigned 0 for $B < a \leq d$. In particular $G(p)$ has two sinks on the axes.*

***Proof.*** Observe first that $(0, 0)$ is assigned $P$ because $Q_{00}(p) = 1$. Next we claim that no $(a, 0)$ can be assigned $N$. If this were true, then the point $(a+1, 0)$ with maximal such $a$ would be a source, contradicting Proposition 3.8.

Next we claim that, if $(a, 0)$ is assigned 0, then the point $(a', 0)$ must also be assigned 0 for each $a'$ with $a' > a$. Again, if this were false, then we would have a source at the point where the first $P$ after a 0 was assigned. Thus there is an $A$ such that $Q_{a0}(p) > 0$ for $0 \leq a \leq A$, but $Q_{a0}(p) = 0$ for $a > A$.

By symmetry the analogous statements hold for points $(0, b)$. The points $(A+1, 0)$ and $(0, B+1)$ are sinks in $G(p)$. □

**Lemma 3.10.** *Suppose that $p \in \mathcal{P}$ and $p$ has degree $d$. Then $G(p)$ has at least two sinks on the diagonal $a + b = d$.*

***Proof.*** Since $q_p$ has degree $d-1$, and the coefficients of $p$ are positive, there is at least one point $(a, b)$ such that $a + b = d - 1$ and $Q_{ab} > 0$. The points $(a+1, b)$ and $(a, b+1)$ are then necessarily sinks in $G(p)$. □

**Proposition 3.11.** *For $p \in \mathcal{P}$, the labeled Newton diagram $G(p)$ has at least $2 + \lceil \frac{d-1}{2} \rceil$ sinks.*

***Proof.*** First we sketch the idea of the proof. The diagram $G(h)$ has $d+1$ sinks. We obtain $G(p)$ from $G(h)$ by changing finitely many coefficients from $P$ to $N$ or to 0. Those operations from Lemma 3.4 that change a coefficient in $q$ from positive to positive do not change the diagram. As we flip coefficients, the sinks originally in $G(h)$ propagate. There are three possibilities.

1) A sink at $(a, b)$ in $G(h)$ moves eventually to a sink $(e, f)$ in $G(p)$ with $e \leq a$ and $f \leq b$.

2) A pair of sinks in $G(h)$ coalesces somewhere into one or zero sinks.

3) A sink in $G(h)$ eventually disappears.

The idea is to show that, when 2) or 3) occurs, away from the boundary, in passing from $p$ to $p^*$, then there is an ancestor $p'$ of $p$ such that $G(p')$ has the same number of sinks as $G(p^*)$. We will therefore be able to ignore these situations in seeking the minimal number of sinks. Hence the minimal number of sinks will arise when maximal coalescence takes place on the diagonal $a + b = d - 1$. We will see that two sinks must remain for points $(a, b)$ with $a + b = d$. When $d$ is odd the other $d - 1$ boundary sinks coalesce in pairs; the minimum number of sinks will be $2 + \frac{d-1}{2}$. When $d$ is even three sinks remain and the other $d - 2$ sinks coalesce in pairs. These statements combine to yield the desired statement.

We now fill in the details. First we study the coefficients along the axes. By Lemma 3.9 we obtain precisely two sinks on the axes. These have the pictures

$$\begin{pmatrix} P & P & \mathbf{0} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{0} \\ P \\ P \end{pmatrix}.$$

On the diagonal where $a + b = d$ we assign 0 to each point. We consider the diagonal where $a + b = d - 1$. Consider $Q_{ab}(p)$ for $a + b = d - 1$ and $a, b > 0$. If $Q_{ab}(p) < 0$, then $Q_{(a-1)(b+1)}(p)$ and $Q_{(a+1)(b-1)}(p)$ are positive. This conclusion follows by contradiction; if such a coefficient were not positive, then a source would exist at $(a, b + 1)$ or $(a + 1, b)$.

We noted the situation on the axes. It follows that if $(a, b)$ is a sink of type S1) in $G(p)$, then $a > 0$ and $b > 0$. Thus sinks cannot propagate to the edges of the diagram unless the coefficient there is zero. This situation occurs in Example 3.6.

We first consider coalescense. When $a + b = d - 1$ such coalescense takes place; we may reverse the signs at alternate $(a, b)$ beginning at $(d - 2, 1)$ and ending no later than $(1, d - 2)$. This procedure removes $\frac{d-1}{2}$ sinks when $d$ is odd, and removes $\frac{d-2}{2}$ sinks when $d$ is even. In either case we use the ceiling function to express the number of remaining sinks as

$$2 + \left\lceil \frac{d-1}{2} \right\rceil. \tag{3.13}$$

We claim that (3.13) gives a lower bound for the number of sinks in $G(p)$. The idea is that sinks can disappear or coalesce in the interior only when they have been unnecessarily created in an earlier step of the propagation; hence the minimum number possible is given by (3.13).

We next consider several cases of coalescence. The others are handled in the same manner. First we suppose $p \in \mathcal{P}$ and $G(p)$ has sinks of type S1) at $(a + 1, b)$ and $(a, b + 1)$. Suppose additionally that these sinks coalesce into one sink at $(a, b)$ with $a + b < d - 1$ when we form $p'$ by flipping the sign of the coefficient at $(a, b)$. We consider the behavior of $G(p)$ near $(a, b)$ under this assumption. The coefficients $Q_{(a+1)(b+1)}$, $Q_{a(b+1)}$, and $Q_{(a+1)b}$ must all be negative. The coefficients $Q_{(a-1)(b+1)}$, $Q_{(a-1)b}$, $Q_{ab}$, $Q_{a(b-1)}$, and $Q_{(a+1)(b-1)}$ are necessarily nonnegative. We provide a picture for this situation, where $(a, b)$ is at the center of the picture. There are similar pictures with the same $N$'s and where some of the $P$'s are 0's. There are also similar

pictures in case one or both of the sinks is of type S2). The proofs are essentially the same.

$$\begin{pmatrix} P & \mathbf{N} & N \\ P & P & \mathbf{N} \\ & P & P \end{pmatrix}. \tag{3.14}$$

We obtain an ancestor $p^*$ of $p$ by flipping the coefficients at both $(a+1, b)$ and $(a, b+1)$. Then $(a+1, b+1)$ must be a sink in $G(p^*)$. Here is a picture of this situation, where $x$ denotes an unknown coefficient:

$$\begin{pmatrix} & x & & \\ P & P & N & \\ P & P & P & x \\ & P & P & \end{pmatrix}. \tag{3.15}$$

If there is not a sink at $(a+1, b)$ nor one at $(a, b+1)$, then $G(p^*)$ and $G(p')$ have the same number of sinks. In this case an ancestor of $p$ has the same number of sinks in its diagram as $p'$ does, so the coalescence is irrelevant.

On the other hand, suppose flipping the coefficients at both $(a+1, b)$ and $(a, b+1)$ back to $P$ results in a sink at either $(a+2, b)$ or $(a, b+2)$, that is, at least one $x$ is an $N$ or 0, and its unspecified initial predecessor is nonnegative. If both are $N$ or 0, then flip one of them back to $P$. Otherwise one is an $N$ or a 0. In either case we have an ancestor with consecutive sinks on the same diagonal. Thus we have the same situation where coalescense occurs one diagonal further to the northeast. We may apply the same reasoning to this ancestor until we reach the diagonal given by $a + b = d - 1$. We conclude that interior coalescence of sinks occurs only when some ancestor had already the same number of sinks as there were after the coalescence.

We handle a second type of coalescence. Suppose that we have a picture (along the axis) such as

$$\begin{pmatrix} N & 0 & \mathbf{0} & 0 & \\ P & P & P & \mathbf{0} & 0 \end{pmatrix},$$

where the sinks are in bold. Replacing the $P$ farthest to the east by a 0 results in

$$\begin{pmatrix} N & 0 & 0 & 0 & \\ P & P & \mathbf{0} & 0 & 0 \end{pmatrix},$$

and the two sinks have coalesced into one. As in the above proof concerning interior coalescense, consider the ancestor

$$\begin{pmatrix} N & 0 & P & \mathbf{0} & \\ P & P & P & P & \mathbf{0} \end{pmatrix}.$$

We have the same situation where coalescence can occur, but one diagonal to the northeast. Hence by iteration we can make sure that coalescence of this type occurs at the boundary.

Next we handle the possible coalescense of two sinks into none, and the closely related disappearance of a sink. Suppose that a sink of type S1) at $(a, b)$ disappears somewhere in the interior when we flip the coefficient at $(a, b-1)$ or $(a-1, b)$ from $P$ to $N$. Then necessarily the coefficient at $(a, b-2)$ or $(a-2, b)$ must be negative. We give a sample illustration, where $x$ denotes an unknown coefficient. (The other cases are similar).

$$\begin{pmatrix} P & \mathbf{N} & x \\ P & P & x \\ P & \mathbf{N} & P \\ & P & \end{pmatrix}. \tag{3.16}$$

In this situation flipping the $P$ in the second row from the top creates

$$\begin{pmatrix} P & N & x \\ P & N & x \\ P & \mathbf{N} & P \\ & P & \end{pmatrix} \tag{3.17}$$

and the top sink has disappeared. In case each $x$ had been an $N$, there would have been another sink at the far right of the second row from the top of (3.16), and passing to (3.17) would amount to the coalescense of two sinks into zero sinks.

Notice in all cases however that the bottom sink in (3.16) is separated from the boundary; therefore flipping the $N$ to $P$ there creates an ancestor $p'$ whose diagram has one fewer sink. In case both $x$'s are $N$, and there is another sink, the rest of the picture falls back into the case of coalescence of two sinks into one sink we treated above. In case this other sink is not there, the sink moved southward, but the number of sinks is unchanged.

In all cases there is an ancestor of $p$ with the same number of sinks as we would have after the double coalescense or disappearance.

We conclude that the number of sinks in $G(p)$ cannot be smaller than the number of sinks obtained from $G(h)$ by coalescing sinks only along the diagonal given by $a + b = d - 2$. We have noted earlier that we cannot remove the two sinks along the axes. We must also have two sinks on the diagonal $a + b = d$. We minimize in the case when the sinks on the axis are also on the diagonal. After this we minimize the number of sinks by coalescing the other $d - 1$ pairs when $d$ is odd, or $d - 2$ pairs when $d$ is even. This shows that the number of sinks in $G(p)$ is at least the number in (3.12). □

Propositions 3.8 and 3.11 yield Theorem 1.1. Suppose $p$ has $N$ terms and is of degree $d$. Then $N$ is at least the number of sinks in $G(p)$, so $N \geq 2 + \frac{d-1}{2}$, or $2N - 3 \geq d$.

The reader will find it instructive to study $G(p)$ for $p$ as in the next example. This polynomial has the maximum possible degree, namely seven, given that it has five terms, yet it is not one of the examples from Section 2. It can be shown that there are infinitely many odd positive integers $2N - 3$ for which there are distinct polynomials in $\mathcal{P}$ with $N$ terms and of degree $2N - 3$.

***Example 3.12.*** Let $p(x, y) = x^7 + y^7 + \frac{7}{2}(x^5y + xy^5 + xy)$. Then $G(p)$ has sinks at $(7, 0)$, $(5, 1)$, $(1, 1)$, $(1, 5)$, and $(0, 7)$. The diagram can be found from the function $q_p$, which satisfies the following formula:

$$\begin{aligned} q_p(x, y) = 1 &+ x + y + x^2 - \frac{3}{2}xy + y^2 + x^3 - \frac{1}{2}x^2y - \frac{1}{2}xy^2 + y^3 \\ &+ x^4 + \frac{1}{2}x^3y - x^2y^2 + \frac{1}{2}xy^3 + y^4 \\ &+ x^5 + \frac{3}{2}x^4y - \frac{1}{2}x^3y^2 - \frac{1}{2}x^2y^3 + \frac{3}{2}xy^4 + y^5 \\ &+ x^6 - x^5y + x^4y^2 - x^3y^3 + x^2y^4 - xy^5 + y^6 . \end{aligned} \tag{3.18}$$

## References

[1] Alexander, H. Proper holomorphic mappings in $\mathbf{C}^n$, *Indiana Univ. Math. J.*, **26**, 137–146, (1977).

[2] D'Angelo, J.P. *Several Complex Variables and the Geometry of Real Hypersurfaces*, CRC Press, Boca Raton, FL, (1993).

[3] D'Angelo, J.P. Proper polynomial mappings between balls, *Duke Math. J.*, **57**, 211–219, (1988).

[4] D'Angelo, J.P. Invariant holomorphic mappings, *J. Geom. Anal.*, **6**(2), 163–179, (1996).

[5] D'Angelo, J.P. and Lichtblau, D. Spherical space forms, CR mappings, and proper maps between balls, *J. Geom. Anal.*, **2**(5), 391–415, (1992).

[6] Faran, J.J. Maps from the two-ball to the three-ball, *Inv. Math.*, **68**, 441–475, (1982).

[7] Kos, Šimon, Two applications of the quasiclassical method to superfluids, Thesis, Dept. of Physics, University of Illinois, Urbana, (2001).

[8] Koshy, T. *Fibonacci and Lucas Numbers,* John Wiley & Sons, NY, (2001).

[9] Pinchuk, S. On the holomorphic continuation of holomorphic mappings, *Math. USSR-Sb.*, **27**. 375–392, (1975).

[10] Setya-Budhi, W. Proper holomorphic mappings in several complex variables, (thesis), University of Illinois, (1993).

---

Received September 6, 2002
Revision received June 3, 2003

Department of Mathematics, University of Illinois, 1409 W. Green St. Urbana IL 61801
e-mail: jpda@math.uiuc.edu

Center for Nonlinear Studies, Los Alamos National Laboratory, MS B258 Los Alamos, NM 87545
e-mail: s-kos@lanl.gov

Department of Mathematics, University of Illinois, 1409 W. Green St. Urbana IL 61801
e-mail: eriehl@fas.harvard.edu

Communicated by Steven Krantz