POLYNOMIAL DYNAMICS

ALICE MEDVEDEV AND THOMAS SCANLON

Abstract. We study algebraic dynamical systems (and, more generally, σ-varieties) $\Phi : \mathbb{A}^n_{\mathbb{C}} \to \mathbb{A}^n_{\mathbb{C}}$ given by coordinatewise univariate polynomials, $\Phi(x_1, \ldots, x_n) = (f_1(x_1), \ldots, f_n(x_n))$ by refining an old theorem of Ritt on compositional identities amongst polynomials. Our main result is an explicit description of the skew-invariant varieties, that is, those algebraic varieties $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ for which there is a field automorphism $\sigma : \mathbb{C} \to \mathbb{C}$ with $\Phi(X) = X^\sigma$. As consequences, we deduce a variant of a conjecture of Zhang on the existence of rational points with Zariski dense forward orbits and a strong form of the dynamical Manin-Mumford conjecture for liftings of the Frobenius.

We also show that in models of ACFA_0, a trivial set defined by $\sigma(x) = f(x)$ for f a polynomial has Morley rank 1 and is usually strongly minimal, that the induced structure on this set is $\aleph_0$-categorical unless f is defined over a fixed field of a power of σ, and that nonorthogonality between two such sets is definable in families if f is defined over a fixed field of a power of σ.

1. Introduction

Let $f_1, \ldots, f_n \in \mathbb{C}[x]$ be a finite sequence of polynomials over the complex numbers and let $\Phi : \mathbb{A}^n_{\mathbb{C}} \to \mathbb{A}^n_{\mathbb{C}}$ be the map $(x_1, \ldots, x_n) \mapsto (f_1(x_1), \ldots, f_n(x_n))$ given by applying the polynomials coordinatewise. We aim to explicitly describe those algebraic varieties $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ which are invariant under Φ. To do so, we solve a more general problem. We fix a field automorphism $\sigma : \mathbb{C} \to \mathbb{C}$ and then describe those algebraic varieties $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ which are skew-invariant in the sense that $\Phi(X) \subseteq X^\sigma$ and recover the solution to the initial problem by taking σ to be the identity map.

We consider this more general problem of classifying the skew invariant varieties in order to import some techniques from the model theory of difference fields and because we are motivated by some fine structural problems in the model theory of difference fields. Recall that a difference field is a field K equipped with a distinguished endomorphism $\sigma : K \to K$. The theory of difference fields, expressed in the first-order language of rings expanded by a unary function symbol for the endomorphism, admits a model companion, ACFA, the models of which we call *difference closed*, and it is the rich structure theory of the definable sets in difference closed fields developed in [6] which we shall employ.

In [12] the first author refined the trichotomy theorems of [6, 8] to show that sets defined by formulas of the form $\sigma(x) = f(x)$ where f is a rational function are *trivial* unless f is covered by an isogeny of algebraic groups in the sense that there is a one-dimensional algebraic group G, an isogeny $\phi : G \to G^\sigma$, and a rational function $\pi : G \to \mathbb{P}^1$ with $f \circ \pi = \pi^\sigma \circ \phi$. In this context, triviality is a very strong property and, essentially, all algebraic relations amongst solutions to trivial

Scanlon is partially supported by NSF CAREER grant DMS-0450010 and a Templeton Infinity grant.

equations are reducible to *binary* relations. It is this consequence and the fact that the dynamical systems arising from isogenies are well-understood that we shall use to reduce the problem of describing general Φ-skew invariant varieties to that of describing skew invariant curves in the affine plane.

Thus, the bulk of the technical work in this paper concerns the problem of describing those affine plane curves $C \subseteq \mathbb{A}^2_{\mathbb{C}}$ which are (f,g)-skew invariant where f and g are trivial polynomials in the sense of the previous paragraph. Via some easy reductions one sees that this problem is really the same as that of describing those polynomials h for which there are polynomials π and ρ satisfying $f \circ \pi = \pi^\sigma \circ h$ and $g \circ \rho = \rho^\sigma \circ h$. Possible compositional identities involving polynomials over $\mathbb{C}$ were explicitly classified by Ritt in [17] and Ritt's work has been given a conceptually cleaner presentation and has been refined to give a very sharp answer to the question of which polynomials $a(x), b(x), c(x), d(x) \in \mathbb{C}[x]$ satisfy $a \circ b = c \circ d$ by Müller and Zieve in [13]. Here, we perform a combinatorial analysis of Ritt's theorem to explicitly describe the possible h, π, and ρ in terms of decompositions of f and g as compositional products.

We find that there are three basic sources for skew-invariant curves. The easiest to see are those coming from (skew) iteration. If f is any polynomial and $g = f^{\sigma^n}$, then the graph of $f^{\diamond n} := f^{\sigma^{n-1}} \circ f^{\sigma^{n-2}} \circ \cdots \circ f$ is (f,g)-skew-invariant. In particular, when $f = f^\sigma$ is fixed by σ, the graphs of iterates of f (and their converse relations) are (f,f)-invariant. When f is expressible as a nontrivial compositional product, $f = g \circ h$, then considering what we call a *plain skew-twist* of f, $\check{f} := h^\sigma \circ g$, we see that the graph of h is $(f, \check{f})$-skew invariant. In most cases, by computing one expression of f as a composition of indecomposable polynomials, one may explicitly and quickly describe all of the possible plain skew twists. Finally, it can happen that graphs of monomial identities (or their conjugates via some linear change of variables) may be (f,g)-invariant. For example, if $f(x) = x \cdot (1 + x^3)^2$ and $g(y) = y \cdot (1 + y^2)^3$, then the curve defined by $y^2 = x^3$ is (f,g)-invariant. Our primary task is to prove a precise a version of the assertion that these examples exhaust the possibilities for skew-invariant curves.

We apply our results on skew invariant varieties to address problems of two different characters. First, we use them to prove variants of two conjectures of Zhang [21] on the arithmetic of dynamical systems. In another direction, we use our results to resolve the question of definability of nonorthogonality between types containing a formula of the form $\sigma(x) = f(x)$ for some polynomial f in ACFA_0.

In the case of the diophantine questions, we show that if K is a finitely generated subfield of $\mathbb{C}$ and $\Phi : \mathbb{A}^n_K \to \mathbb{A}^n_K$ is given by a sequence of univariate polynomials each of degree at least two, then there are points $a \in \mathbb{A}^n(K)$ with a Zariski dense Φ-forward orbit. In fact, we prove a somewhat stronger result in which some of the f_is are allowed to be linear. In another direction we prove a refined version of Zhang's Manin-Mumford conjecture for dynamical systems lifting a Frobenius. Again, the precise statement will have to wait, but we can note a special case. Suppose that q is a power of a prime p and that $f(x) \in \mathbb{Z}[x]$ is a polynomial of degree q for which $f(x) \equiv x^q \pmod{p}$ but f is not linearly conjugate to either x^q itself or the q^{th} Chebyshev polynomial and f is not a compositional power, then any irreducible variety $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ containing a Zariski dense set of n-tuples of f-periodic points must be defined by finitely many equations of the form $x_i = \zeta$ for some f-periodic point ζ and $x_j = f^{\circ m}(x_k)$.

In the case of differential fields, Hrushovski and Itai showed that there are theories of model complete differential fields other than the theory of differentially closed fields [10]. It is still open whether or not there are model complete theories of difference fields other than ACFA, but if one could show that there were some formula $\theta(x)$ defining in a difference closed field a set of D-rank 1 having only finitely many algebraic realizations such that for for every other formula $\eta(y,z)$ the set of parameters $\{b :$ some nonalgebraic type $p(x)$ extending $\theta(x)$ is nonorthogonal to $\eta(y,b)\}$ is definable, then one could produce a new model complete difference field by omitting the nonalgebraic types in θ. We do not achieve this goal, but we do show that in characteristic zero for $\theta(x)$ given by $\sigma(x) = f(x)$ where f is a polynomial defined over the fixed field, for any $\eta(y,z)$ of the same form (*ie* $\eta(y,z)$ is defined by $\sigma(y) = g(y,z)$ where g is a polynomial), then the set of parameters b for which $\theta(x)$ and $\eta(x,b)$ are nonorthogonal is definable. In fact, we show that even in the cases where nonorthogonality is not definable, the only real obstruction comes from graphs of the distinguished automorphism. A byproduct of this analysis is an explicit characterization of the algebraic closure operator on trivial D-rank 1 sets defined by $\sigma(x) = f(x)$, and the observation that this set is strongly minimal unless $f(x)$ is skew-conjugate to $x^k \cdot u(x)^n$ for some polynomial u and some $n > 1$, and in any case has Morley rank 1 if u is non-constant.

This paper is organized as follows. In Section 2 we lay out our conventions and notation. We begin with our technical work on polynomial compositional identities in Section 3. In Section 4 we completely describe the possible skew-invariant varieties as a consequence of theorems on the model theory of difference fields and the results of Section 3. In Section 5 we conclude with three applications of our results to definability of orthogonality, Zhang's conjecture on the density of dynamical orbits, and a version of the dynamical Manin-Mumford conjecture for Frobenius lifts.

We thank M. Zieve for sharing a preliminary version of [13] and for discussing issues around compositional identities of polynomials and rational functions.

2. Notation and conventions

For the most part, our notation is standard.

Unless explicitly stated to the contrary, we work over an algebraically closed field L of characteristic zero which the reader may take to be $\mathbb{C}$ without loss of generality. Thus, for example, when we say "polynomial" without qualification we mean "polynomial with coefficients from L."

With the exception of the results in Section 5.3 for which the language of schemes is necessary, when we speak of algebraic varieties we really mean closed subvarieties of affine space. As such, the reader is welcome to read for "variety" the phrase "subset of some $\mathbb{A}^n(K) = K^n$ defined by the vanishing of finitely many polynomials." Note that for us a variety need not be irreducible. We write $\mathbb{G}_a$ for the additive group considered as an algebraic group and $\mathbb{G}_m$ for the multiplicative group considered as an algebraic group.

Recall that a difference field is a field K given together with a distinguished field endomorphism $\sigma : K \to K$. On occasion, we shall endow L with a distinguished endomorphism $\sigma : L \to L$ making L into a difference field. We shall explain some results from the model theory of difference fields in Section 4. By *the fixed field* we shall mean $\{x \in L : \sigma(x) = x\}$.

If X is an algebraic variety over L, then X^σ is the σ-transform of X. Formally, X^σ is the base change of X from $\mathrm{Spec}(L)$ to $\mathrm{Spec}(L)$ via σ^*. Less categorically, if X is given as a closed subvariety of some affine space defined by the vanishing of some polynomials, then X^σ is defined by the vanishing of the polynomials obtained by applying σ to the coefficients of the polynomials defining X. In terms of rational points, if $X \subseteq \mathbb{A}^n_L$, then $X^\sigma(L) = \{(\sigma(a_1), \ldots, \sigma(a_n)) : (a_1, \ldots, a_n) \in X(L)\}$. Likewise, if $f : X \to Y$ is a morphism of varieties, then one obtains a morphism $f^\sigma : X^\sigma \to Y^\sigma$.

Following Pink and Rößler [16], a *σ-variety* is a pair (X, f) where X is an algebraic variety and $f : X \to X^\sigma$ is a morphism from X to the σ-transform of X. Some authors require f to be dominant and in practice we are really only interested in this case. When $X^\sigma = X$ and $f^\sigma = f$, so that $f : X \to X$, we call this σ-variety an *algebraic dynamical system* or an *AD* for short. A morphism $\psi : (X, f) \to (Y, g)$ of σ-varieties is given by a morphism of varieties $\psi : X \to Y$ for which the following diagram commutes

$$\begin{array}{ccc} X & \xrightarrow{f} & X^\sigma \\ {\scriptstyle\psi}\downarrow & & \downarrow{\scriptstyle\psi^\sigma} \\ Y & \xrightarrow{g} & Y^\sigma \end{array}$$

A morphism $\psi : (X, f) \to (Y, g)$ of ADs is also given by a morphism of varieties $\psi : X \to Y$ for which a related diagram commutes

$$\begin{array}{ccc} X & \xrightarrow{f} & X \\ {\scriptstyle\psi}\downarrow & & \downarrow{\scriptstyle\psi} \\ Y & \xrightarrow{g} & Y \end{array}$$

When all of the objects in question are defined over the fixed field, a morphism of ADs is the same as a morphism of σ-varieties, but in general, these notions are different and it may happen that two ADs are isomorphic as σ-varieties but not as ADs.

Given an AD (X, f) we define $(X, f^{\circ n})$ by recursion with $f^{\circ 0} := \mathrm{id}_X$ and $f^{\circ(n+1)} := f \circ f^{\circ n}$. If (X, f) is a σ-variety, then we define $f^{\Diamond n} : X \to X^{\sigma^n}$ by recursion with $f^{\Diamond 0} := \mathrm{id}_X$ and $f^{\Diamond(n+1)} := f^{\sigma^n} \circ f^{\Diamond n}$. If we need to discuss Cartesian powers we might use the notation $f^{\times n}$. That is, if $f : X \to Y$, we might write $f^{\times n} : X^{\times n} \to Y^{\times n}$ for the map $(x_1, \ldots, x_n) \mapsto (f(x_1), \ldots, f(x_n))$.

If (X, Φ) is a σ-variety, then we say that a subvariety $Y \subseteq X$ is *skew-invariant*, or *Φ-skew-invariant* if we wish to emphasize Φ, if Φ maps Y dominantly to Y^σ. Likewise, if (X, Φ) is simply an AD, then a subvariety Y of X, also defined over the fixed field of σ, is *Φ-invariant* just in case $\Phi(Y) = Y$.

Note that it could happen that a variety is (skew-)invariant even though none of its irreducible components are. However, they will be (skew-)invariant for $\Phi^{\circ n}$ (respectively, $\Phi^{\Diamond n}$ and σ^n). In general, one might prefer to study those varieties for which $\Phi(Y) \subseteq Y^\sigma$ or for which $\Phi \restriction Y : Y \to Y^\sigma$ is dominant. For the questions we study here, these distinctions are immaterial.

3. Polynomial compositional identities

In this the longest section of the paper we recall Ritt's theorem on compositions of polynomials in detail, derive our theorems on the canonical sequences of Ritt swaps, and then use these results on canonical sequences of Ritt swaps to describe the possible (f, g)-skew invariant curves when f and g are polynomials of degree at least two.

We should say a something about attributions before launching into the technical details. Of course, the key result, Theorem 3.1 already mentioned, is due to Ritt [17]. Julia and Fatou are usually credited with having founded the study of identities amongst iterations of rational functions, but Ritt's contemporaneous work contains stronger results and his theorem on compositional identities amongst polynomials stands as one of the highest achievements in algebraic dynamics. We take Ritt's theorem as basic and then deduce via a mostly combinatorial analysis (mixing in some elementary properties of polynomials) some canonical forms for recompositions of polynomials. The end result of Section 3.5.2 is an easy consequence of the main theorem of Müller and Zieve in [13] and the proof there is much cleaner. The reader might well ask why we have bothered to retain this material. Our main goal is to describe the skew-invariant varieties for polynomial ADs. To do so we need to continue the combinatorial analysis to an investigation of skew-twists (see Section 3.6.1) and the internal structure of our combinatorial proof of the refinements of Ritt's theorem seems to be necessary. In private communication, Zieve suggested a method to deduce our final results from the main theorem of [13], but as these arguments were at least as long as those presented here, we chose to follow our original approach. It may be the case that Zieve's arguments can be simplified and also given a geometric presentation in which case they might adapt well to rational functions or positive characteristic.

3.1. **Ritt's theorem.** We begin by recalling the definitions needed for the statement of Ritt's theorem.

Definition 3.0.1. A polynomial L is *linear* if there are $A \neq 0$ and B, with $L(x) = Ax + B$. We say that L is a *scaling* if $B = 0$ and write $L = (\cdot A)$. We say that L is a *translation* if $A = 1$ and write $L = (+B)$. We write $(-C)$ for $(+(-C))$.

We use the word "linear" even though "affine" might be more appropriate. The linear polynomials correspond to the automorphisms of the affine line. In algebraic terms, the linear polynomials are the elements of the polynomial ring invertible under composition.

Definition 3.0.2. A polynomial f is *indecomposable* if $\deg(f) \geq 2$ and it cannot be written as a composition $f = g \circ h$ of two non-linear polynomials g and h.

We define decompositions to have indecomposable factors, since that is the only notion used in this paper. In a different context, one might wish to speak about decompositions some of whose factors are themselves decomposable, and then one might wish to call our decompositions *complete*.

Definition 3.0.3. For $f \in K[x]$ we say that the finite sequence $(f_k, \ldots, f_1)$ of polynomials f_i is a *decomposition of* f if $f = f_k \circ \cdots \circ f_1$ and each f_i is indecomposable. We often write $\vec{f}$ instead of $(f_k, \ldots, f_1)$.

As $\deg(g\circ h) = \deg(g)\cdot\deg(h) > \deg(g), \deg(h)$ for non-linear g and h, the degree of a polynomial acts as an analogue of a Euclidean norm. Thus, all nonlinear polynomials have decompositions. (Linear polynomials do not have decompositions.) Some polynomials admit many decompositions. For example, if $a, b \in \mathbb{Z}$ are distinct primes and $f(x) = x^{ab}$, then (x^a, x^b) and (x^b, x^a) are two different decompositions of f.

Ritt proved [17] that these different decompositions are in some sense unique up to permutations. We need a few more definitions before we can state his theorem. We will then prove a stronger version of the "up to permutations" part.

Definition 3.0.4. For $n \in \mathbb{Z}_+$, the n^{th} Chebyshev polynomial is the polynomial $C_n(x) \in \mathbb{Z}[x]$ defined by the relation

$$x^n + \frac{1}{x^n} = C_n(x + \frac{1}{x})$$

Other definitions of the Chebyshev polynomials appear in the literature. Note that with our definition, C_n is monic. We have included the case of $n = 1$ for the sake of uniformity, but when we speak of Chebyshev polynomials, we will usually mean C_n with $n \geq 2$. In fact, we will usually only consider C_p with p an odd prime.

Definition 3.0.5. The following polynomials, when they are indecomposable (and this is an extra condition only in the last case), are called *ritty*.

- $Q(x) := x^2$
- $P_p(x) := x^p$, p an odd prime
- $C_p(x)$, p an odd prime
- $S_{(k,\ell,n,u)} := x^k \cdot u(x^\ell)^n$ where $k \neq 0$, $\gcd(k, \ell) = 1$, $\gcd(k, n) = 1$, $u(0) \neq 0$, u is monic non-constant, and at least one of ℓ and n is greater than one.

Lemma 3.0.1. *Any ritty polynomial α of the last kind has a unique maximal k, ℓ, and n, and u is unique up to multiplication by nth roots of* 1.

Proof. Of course, the number k is the order of vanishing of α at 0. Let $f(x) := u(x^\ell)^n$ and let ℓ_2 be maximal such that $f(x) = g(x^{\ell_2})$ for some polynomial g. This is the maximal ℓ_2 such that the zeros of f are a union of multiplicative cosets of the group of ℓ_2th roots of 1. Let n_2 be maximal such that $f(x) = h(x)^{n_2}$ for some h. This n_2 is least common multiple of the multiplicities of roots of f. □

Definition 3.0.6. If f is a ritty polynomial of the last kind, we define the *in-degree of f* to be the maximal ℓ, and the *out-degree of f* to be the maximal n from the last lemma.

Remark 3.0.1. Occasionally, we write P_2 for Q, but as the properties of Q are distinct from those of the monomials of odd degree, we generally treat Q and P_p for p an odd prime separately. Occasionally, we write C_2 for the Chebyshev polynomial of degree 2, but note that it is not ritty. Though we give a separate name to the Chebyshev polynomials, it is clear from the definition that they are odd functions, so for each odd p there is some u such that $C_p(x) = x \cdot u(x^2)$; so they are in fact a special case of the last kind of ritty polynomials.

Remark 3.0.2. The condition that u be monic is not necessary and does not appear in the original statement of Ritt's theorem. We will show in Lemma 3.1.1 that including this hypothesis does not affect the truth of that result.

Definition 3.0.7. The following identities involving ritty polynomials are called the *basic Ritt identities.*

- $P_p \circ P_q = P_q \circ P_p$ for $p \neq q$
- $C_p \circ C_q = C_q \circ C_p$ for $p \neq q$
- $P_p \circ (S_{k,\ell p,n,u}) = S_{k,\ell,pn,u} \circ P_p$

We formally exclude the tautological identities $f \circ f = f \circ f$ from the first two kinds of basic Ritt identities because they are clearly useless, and because including them would make the combinatorial results and proofs even more cumbersome. In particular, Lemma 3.2.6 would not be true as stated.

Other compositional identities involving indecomposable polynomials can be obtained from these by carefully inserting linear factors: if $a \circ b = c \circ d$ and L, M, and N are linear, then

$$(L^{-1} \circ a \circ M) \circ (M^{-1} \circ b \circ N) = (L^{-1} \circ c) \circ (d \circ N)$$

Before giving a formal definition of this phenomenon, we note that it is even possible to obtain new compositional relations between *ritty* polynomials in this way.

Remark 3.0.3. From the fact that each C_p commutes with $C_2(x) = x^2 - 2$, it is easy to see that $\hat{C}_p := (+2) \circ C_p \circ (-2)$ is of the form $\hat{C}_p(x) = x \cdot u(x)^2$, so it is ritty; it is also easy to see that these commute with each other. Another example comes from using scalings L, M, and N on the basic Ritt identity $C_p \circ C_q = C_q \circ C_p$.

Definition 3.0.8. Suppose that $(f_k, \ldots, f_1)$ is a decomposition and that $1 \leq i < k$. If there are linear polynomials L, M, and N such that $(L^{-1} \circ f_{i+1} \circ M) \circ (M^{-1} \circ f_i \circ N) = R \circ S$ is a basic Ritt identity, then $(g_k, \ldots, g_1)$ given by $g_i := S \circ N^{-1}$, $g_{i+1} = L \circ R$, and $g_j := f_j$ for the other $j \leq k$ is another decomposition of the same polynomial and we say that this latter decomposition is *obtained from the former by a Ritt swap at i.*

In light of this and looking ahead to Ritt's theorem, we define

Definition 3.0.9. An indecomposable polynomial f is *swappable* if there are linear polynomials L and M such that $L \circ f \circ M$ is ritty. In general, we say that polynomials f and g are *linearly related* if there are linear L and M such that $L \circ f \circ M = g$.

Remark 3.0.4. It is clear from the definitions that if some decomposition may be obtained from $\vec{f}$ by a Ritt swap at i, then one of the following must happen:

- both f_i and f_{i+1} are linearly related to monomials
- both f_i and f_{i+1} are linearly related to odd-degree Chebyshev polynomials;
- f_i is linearly related to a monomial P_p and f_{i+1} is linearly related to a ritty polynomial whose out-degree is a multiple of p;
- f_{i+1} is linearly related to a monomial P_p and f_i is linearly related to a ritty polynomial whose in-degree is a multiple of p.

Remark 3.0.5. In the definition of one decomposition having been obtained from another via a Ritt swap at i we suppress the auxiliary choices of the linear polynomials L, M, and N. At the level of decompositions, there is a real ambiguity attributable to these choices. For example, $(x \cdot (x+1)^5, x^5)$ may be swapped to $(x^5, x \cdot (x^5+1))$ taking $L(x) = M(x) = N(x) = x$. It may also be swapped to $((32)^6 x^5, \frac{x}{2} \cdot \frac{x^5+1}{32})$ taking $L(x) = (32)^6 x$, $M(x) = 32x$, and $N(x) = 2x$. Although the factors in this decomposition are not monic, the basic Ritt identity used does

involve monic ritty polynomials.
However, the two resulting decompositions are *linearly equivalent*, defined four lines below. Our first fundamental result, Proposition 3.1.3, is that at most one decomposition can be obtained from a given one by a Ritt swap at a given i, up to linear equivalence.

Definition 3.0.10. The decompositions $(f_k, \dots, f_1)$ and $(g_k, \dots, g_1)$ are *linearly equivalent* if there are linear polynomials $L_{k-1}, \dots, L_1$ for which $g_k = f_k \circ L_{k-1}$, $g_i = L_i^{-1} \circ f_i \circ L_{i-1}$ for $k > i > 1$, and $g_1 = L_1^{-1} \circ f_1$.

If $(f_k, \dots, f_1)$ and $(g_k, \dots, g_1)$ are linearly equivalent, then they are decompositions of the same polynomial. Linear equivalence, as the name suggests, is an equivalence relation.

Definition 3.0.11. We often write $\vec{f}$ for a decomposition $(f_k, \dots, f_1)$; $[\vec{f}]$ is the linear-equivalence class of this decomposition. For a polynomial f, D_f is the set of linear-equivalence classes of decompositions of f.

Ritt's theorem says basically that Ritt swaps act transitively on D_f. Our first fundamental result is that this slogan can be formalized. Here is our statement of Ritt's theorem:

Theorem 3.1 (Ritt)**.** *Over $\mathbb{C}$, any two decompositions of the same polynomial have the same length. Indeed, if $\vec{f}$ and $\vec{g}$ are decompositions of the same polynomial, then $\vec{g}$ is linearly equivalent to a decomposition obtained from $\vec{f}$ by a finite sequence of Ritt swaps.*

A diligent reference hunter will note that our definition of ritty polynomials differs from Ritt's in that we require u to be monic in the last case; this does not make a difference:

Lemma 3.1.1. *The notion of "Ritt swap" is not changed if in the last case of the definition of ritty polynomial we drop the requirement that u be monic.*

Proof. It suffices to show that if p is a prime, k, n, and ℓ are integers, and U is a polynomial, and $x^k \cdot U(x^{p\ell})^n$ satisfies the requirements in the definition of ritty polynomial, then $(x^k \cdot U(x^\ell)^{pn}, P_p)$ is obtained by a Ritt swap from $(P_p, x^k \cdot U(x^{p\ell})^n)$. Let a be the leading coefficient of U. Set $\nu(x) := x$, $\mu(x) := a^n x$, and $\lambda(x) := a^{pn} x$. Then $\lambda^{-1} \circ P_p \circ \mu = P_p$ and $\mu^{-1} \circ (x^k \cdot U(x^{p\ell})^n) \circ \nu = (x^k \cdot u(x^{p\ell})^n)$ where $u(y) := \frac{U(y)}{a}$ is monic. The result is now clear. □

The next lemma shows that the relation "$\vec{g}$ is obtained from $\vec{f}$ by a Ritt swap at i" is invariant under linear equivalence.

Lemma 3.1.2. *If $\vec{f}$, $\vec{g}$, and $\vec{h}$ are decompositions of the same polynomial, $\vec{g}$ is obtained from $\vec{f}$ by a Ritt swap at i, and $\vec{h}$ is linearly equivalent to $\vec{f}$, then there is a decomposition obtained from $\vec{h}$ by a Ritt swap at i and linearly related to $\vec{g}$.*

Proof. Let $R_{k-1}, \dots, R_1$, L, M, and N be linear polynomials witnessing our hypotheses. That is, the Rs witness that $\vec{h}$ is linearly related to $\vec{f}$:
$h_k = f_k \circ R_{k-1}$, $h_j = R_j^{-1} \circ f_j \circ R_{j-1}$ for $1 < j < k$, $h_1 = R_1^{-1} \circ f_1$ and the other linear polynomials witness the Ritt swap: $(L^{-1} \circ f_{i+1} \circ M) \circ (M^{-1} \circ f_i \circ N) = T \circ S$ is a basic Ritt identity, $g_i := S \circ N^{-1}$, $g_{i+1} = L \circ T$, and $g_j := f_j$ for the other $j \leq k$. To simplify the notation, we define $R_k(x) = R_0(x) = x$.

Define $L' := R_{i+1}^{-1} \circ L$, $M' := R_i^{-1} \circ M$, and $N' := R_{i-1} \circ N$. It is now routine to check that this choice of L', M', and N' witnesses that $\vec{h}$ admits a Ritt swap at i and that the resulting decomposition is linearly equivalent to $\vec{g}$. □

To finish formalizing our slogan, that is defining an action of Ritt swaps on the linear-equivalence classes in D_f, we need "$\vec{g}$ is obtained from $\vec{f}$ by a Ritt swap at i" to be a function rather than a relation. Our next lemma has a surface appearance of yet another elementary result about compositions of polynomials, but its proof while locally elementary is surprisingly complicated. In fact, it is the crucial result required to show that sequences of Ritt swaps form a monoid that acts on linear-equivalence classes of decompositions. This is one the two crucial results towards canonical sequences of Ritt swaps, which give the near-action of the permutations group.

Lemma 3.1.3. *If two decompositions $\vec{h}$ and $\vec{g}$ are both obtained from $\vec{f}$ by a Ritt swap at i, then $\vec{h}$ is linearly equivalent to $\vec{g}$.*

Our proof of this crucial Lemma 3.1.3 occupies most of Section 3.2. Let us give a hint of the difficulties involved. Suppose there are two polynomials u and w, integers k, k' and p, and a field element A such that $f_1 := (x^k \cdot u(x^p)) \circ (+A) = x^{k'} \cdot w(x^p)$. (We will show that such polynomials do not actually exist.) Then the decomposition (P_p, f_1) admits two Ritt swaps: one with $L = M = N = \mathrm{id}$ produces the decomposition $(x^{k'} \cdot w(x)^p, P_p)$; the other, with $L = M = \mathrm{id}$ and $L = (-A)$, produces $(x^k \cdot u(x)^p, P_p \circ (+A))$. It is evident that these two are not linearly equivalent. As this example suggests, the bulk of our work will consist of classifying linearly related ritty polynomials.

We start the proof of the lemma here, indicating precisely what we need to resolve. The next section is devoted to resolving it.

Proof. Suppose that we can perform a Ritt swap at i in two different ways. That is, for $j = 1$ or 2 we can find linear polynomials L_j, M_j, and N_j and ritty polynomials G_j, H_j, $\widehat{G}_j$ and $\widehat{H}_j$ such that

- $G_j = L_j^{-1} \circ f_{i+1} \circ M_j$
- $H_j = M_j^{-1} \circ f_i \circ N_j$
- $G_j \circ H_j = \widehat{H}_j \circ \widehat{G}_j$ is a basic Ritt identity
- $g_{i+1} = L_1 \circ \widehat{H}_1$
- $g_i = \widehat{G}_1 \circ N_1^{-1}$
- $h_{i+1} = L_2 \circ \widehat{H}_2$, and
- $h_i = \widehat{G}_2 \circ N_2^{-1}$.

We are charged with finding a linear R for which $(L_1 \circ \widehat{H}_1) \circ R = (L_2 \circ \widehat{H}_2)$ and $R^{-1} \circ (\widehat{G}_1 \circ N_1^{-1}) = \widehat{G}_2 \circ N_2^{-1}$.

Now then, if we set $L := L_2^{-1} \circ L_1$, $M := M_2^{-1} \circ M_1$, and $N := N_2^{-1} \circ N_2$, then we have $L \circ G_1 \circ M^{-1} = G_2$ and $M \circ H_1 \circ N^{-1} = H_2$. We claim that it is enough to find a linear R for which $(L \circ \widehat{H}_1) \circ R = \widehat{H}_2$ and $R^{-1} \circ (\widehat{G}_1 \circ N^{-1}) = \widehat{G}_2$. Indeed, apply L_2 to the left of the first equation and N_2^{-1} to the right of the second.

It remains to analyze the possible ways to express ritty polynomials as linear polynomials composed with ritty polynomials. We carry out this analysis in Section 3.2 and complete this proof at the end of that section. Note that in light of Lemma 3.0.1, we may assume that at least one of the linear factors is non-trivial. □

3.2. **Linear relations amongst ritty polynomials.** This section is devoted to characterizing linearly related ritty polynomials. We first reduce the study of linearly related polynomials to the cases where the linear functions are both translations or both scalings.

First, some notation:

Definition 3.1.1. We say that the polynomials f and g are *scaling related* if they are linearly related witnessed by scalings. We say that f and g are *translation related* if they are linearly related witnessed by translations.

A special case of scaling related polynomials will be seen so often that it deserves its own notation.

Definition 3.1.2. Given any polynomial f and nonzero scalar c, we define

$$c * f := \frac{f(cx)}{c^{\deg(f)}}$$

Note that if f is monic, then $c * f$ is also monic. And now, the reduction:

Lemma 3.1.4. *If f and g are linearly related ritty polynomials, then there is another ritty polynomial h which is scaling related to f and translation related to g.*

Proof. Since the group of automorphisms of the affine line is the semidirect product of the group of translations by the group of scalings, if f and g are linearly related, then we can find a third polynomial h which is scaling related to f and translation related to g.

That is, if $L(x) = Ax + B$ and $M(x) = Cx + D$ are linear polynomials with $L \circ f \circ M = g$, then as $L = (+B) \circ (\cdot A)$ and $M = (\cdot C) \circ (+\frac{D}{C})$, we may take $h := (\cdot A) \circ f \circ (\cdot C)$ and then $g = (+B) \circ h \circ (+\frac{D}{C})$. We need to show that h is itself ritty.

Since h is translation-related to a monic polynomial g, h is itself monic. It is clear that if f and h are scaling related monic polynomials, then $h = \lambda * f$ for some λ.

It remains to show that a monic polynomial scaling related to a ritty polynomial is itself ritty:

- $\lambda * P_p = P_p$ for every p, including $p = 2$.
- As noted above, for p odd, C_p being an odd function, fits into the next case.
- Given a polynomial u and natural numbers k, ℓ, and n, if $w := (\lambda^\ell) * u$, then $\lambda * (x^k \cdot u(x^\ell)^n) = x^k \cdot w(x^\ell)^n$

□

Before the last part of the above proof is forgotten, we exploit it to show that carefully inserting scalings into a basic Ritt identity produces another basic Ritt identity. We will prove something of a converse to this lemma in Lemma 3.2.1.

Corollary 3.1.1. *If $b \circ a = d \circ c$ is a basic Ritt identity, and a and b are not both Chebyshev, then for any non-zero λ, μ, there are non-zero ν and η such that*

$$(\mu * b) \circ (\lambda * a) = (\eta * d) \circ (\nu * c)$$

is also a basic Ritt identity.

Proof. Since a and b are not both Chebyshev, one of them must be a monomial. If both are monomials, the result is immediate as $\lambda * P_p = P_p$ for any p and non-zero λ.
If only $a = P_p$ is a monomial, then b must be of the form $(x^k \cdot u(x^\ell)^{pn})$. Then looking at the proof of the lemma, $\mu * b = \mu * (x^k \cdot u(x^\ell)^{pn}) = x^k \cdot w(x^\ell)^{pn}$ for $w := (\lambda^\ell) * u$. Since $(\lambda * a) = (\lambda * P_p) = P_p$, we see that

$$(\mu * b) \circ (\lambda * a) = (x^k \cdot w(x^\ell)^{pn}) \circ P_p = P_p \circ (x^k \cdot w(x^{p\ell})^n)$$

Is a basic Ritt identity, and we may take $\eta = \mathrm{id}$, and $\nu^p = \lambda$.

If only $b = P_p$ is the monomial, then $(\mu * b) = b = c = (\nu * c) = P_p$ for any ν, so we might as well take $\nu = \mathrm{id}$. Now a must be of the form $(x^k \cdot u(x^{p\ell})^n)$, so

$$\lambda * a = \lambda * (x^k \cdot u(x^{p\ell})^n) = x^k \cdot w(x^{p\ell})^n$$

where $w = \lambda^{p\ell} * u$. It is easy to check as above that $\eta = \lambda^p$ works. □

Remark 3.1.1. In particular, if a and b are not both linearly related to Chebyshev polynomials, and (d, c) is obtained from (b, a) by a Ritt swap, then some $(\widetilde{d}, \widetilde{c})$ linearly equivalent to (d, c) can be obtained from (b, a) by a Ritt swap witnessed by translations.

We now return to characterizing linearly related ritty polynomials. With the reduction in Lemma 3.1.4, it is sufficient to separately characterize scaling-related ritty polynomials, and translation-related ritty polynomials. We have just done the first in the proof of the last Lemma 3.1.4, and we can immediately give the whole answer for monomials:

Proposition 3.1.1. *For all p including $p = 2$, P_p is not linearly related to any ritty polynomial other than itself, nor is it translation-related to itself. For every c and every p, $c * P_p = P_p$.*

Proof. Only the first of the three assertions is not immediate. It is true because P_p is the only ritty polynomial of degree p with a unique 0. □

As we have noted that Chebyshev polynomials are a special case of the last kind of ritty polynomials, we are left with the following problem.

Problem 3.1.1. *When are two ritty polynomials of the form $x^k \cdot u(x^\ell)^n$ translation related? That is, we need to solve*

$$(+B) \circ (x^k \cdot u(x^\ell)^n) \circ (+A) = (x^{k_2} \cdot u_2(x^{\ell_2})^{n_2})$$

The brunt of the work in this section is devoted to solving this. We make a few easy observations and summarize our results before diving into the necessary computations.

Remark 3.1.2. In light of Lemma 3.0.1, we are not interested in the case $A = B = 0$; evaluating both sides at 0 shows that the case $A = 0 \neq B$ is impossible; thus we assume that $A \neq 0$ and examine separately the cases where $B = 0$ and where $B \neq 0$.

The next two propositions are proved later in this section.

Proposition 3.1.2. *All solutions of Problem 3.1.1 with $B = 0$ are of the form*

$$(x^l \cdot (x - A)^m \cdot u(x)^n) \circ (+A) = (x + A)^l \cdot x^m \cdot u(x + A)^n$$

Proof. This follows imediately from Proposition 3.2.1 below and Lemma 3.1.5. □

Definition 3.1.3. A ritty polynomial of the form $f(x) = x^\ell \cdot (x - A)^m u(x)^n$ with both $\gcd(\ell, n) > 1$ and $\gcd(m, n) > 1$ is called a *type J ritty polynomial*. A polynomial linearly related to a type J ritty polynomial is called a *type J swappable polynomial*.

Remark 3.1.3. Since a ritty polynomial is, by definition, indecomposable, if $f(x) = x^\ell \cdot (x - A)^m \cdot u(x)^n$ is a type J ritty polynomial, then $\gcd(\ell, m, n) = 1$. Unless $\gcd(m, n) > 1$, f is not a ritty polynomial; the condition that $\gcd(\ell, n) > 1$ implies that $f \circ (+A)$ is also a ritty polynomial. Since indecomposability and swappability are invariant under linear relatedness, a type J swappable polynomial is necessarily swappable and indecomposable.

Remark 3.1.4. The representation of a type J ritty polynomial need not be unique: the polynomial

$$\prod_{i \leq n} (x - A_i)^{\prod_{j \neq i} p_j}$$

where p_j are distinct primes, is translation-related to n different type J ritty polynomials.

Remark 3.1.5. The existence of type J ritty polynomials rules out the possibility that in-degree and out-degree might be well-defined up to linear relatedness, and thus might be a property of a swappable indecomposable. However, we can at least say that the in-degree of a type J swappable indecomposable f is 1, in the sense that any ritty polynomial linearly related to f has in-degree 1. In light of Remark 3.0.4, we can see that if a factor f_i of a decomposition $\vec{f}$ is type J, then no decomposition may be obtained from $\vec{f}$ by a Ritt swap at i.

Proposition 3.1.3. *For each degree d and for each $A \neq 0$, there is a unique triple of polynomials U_A, V_A, W_A of degree d such that for some $B, C \neq 0$*

$$(+B) \circ (x \cdot U_A(x^2)) \circ (+A) = x \cdot V_A(x)^2$$
$$(+C) \circ (x \cdot V_A(x)^2) \circ (-2A) = x \cdot W_A(x)^2$$

Proof. This follows immediately from Propositions 3.2.4 and 3.2.3 and Lemma 3.1.5. □

Before explaining how these solutions come from Chebyshev polynomials, we make an easy observation

Lemma 3.1.5. *If f and g are ritty polynomials giving a solution to problem 3.1.1, that is if $(+B) \circ f \circ (+A) = g$, then $A * f$ is a solution of the same problem with* 1 *in place of A: $(+C) \circ (A * f) \circ (+1) = A * g$.*

So it is sufficient to characterize solutions for one particular A, as we do in the next proposition.

Proposition 3.1.4. *In Proposition 3.1.3, $x \cdot U_{-2}(x^2) = C_{2d+1}$ is the $(2d+1)$st Chebyshev polynomial.*
For any A, $U_{-A} = U_A$ and $W_A = V_{-A} = (-1) * V_A$.
Further, $(x \cdot V_{-2}(x)^2) \circ Q = Q \circ C_{2d+1}$.

Proof. For odd n, both $(+2) \circ C_n \circ (-2)$ and $(-2) \circ C_n \circ (+2)$ are of the form $x \cdot v(x)^2$. Both come from the fact that C_n commutes with $C_2(x) = x^2 - 2$. Here is the reason for the first:

$$C_n \circ C_2 = C_2 \circ C_n$$
$$C_n \circ (-2) \circ Q = (-2) \circ Q \circ C_n$$
$$(+2) \circ C_n \circ (-2) \circ Q = Q \circ C_n$$

For the second, first observe that

$$i * C_2 = \frac{1}{i^2}((ix)^2 - 2) = -(-x^2 - 2) = x^2 + 2$$

Now $C_n \circ C_2 = C_n \circ (\cdot - 1) \circ (\cdot - 1) \circ C_2 \circ (\cdot i) \circ (\cdot - i) =$

$$= (\cdot - 1) \circ C_n \circ (x^2 + 2) \circ (\cdot - i) = C_2 \circ C_n$$

So, bringing all outside linear factors to the right and introducing (-2) on the left,

$$(-2) \circ C_n \circ (+2) \circ Q = (-2) \circ (\cdot - 1) \circ C_2 \circ C_n \circ (\cdot i)$$

Now, $[(-2)(\cdot - 1) \circ C_2](x) = -(x^2 - 2) - 2 = -x^2 = [Q \circ (\cdot \pm i)](x)$, so

$$(-2)C_n \circ (+2) \circ Q = Q \circ (\cdot \pm i) \circ C_n \circ (\cdot i) = Q \circ (i * C_n)$$

□

We now name these polynomials, and the solutions for other different A which are scaling-related to these.

Definition 3.1.4. Continuing to use the notation from Proposition 3.1.3, $\widehat{C}_p(x) := x \cdot V_{-2}(x)^2$. For $\lambda \neq 0$,

$$C_{p,\lambda} := \lambda * C_p \text{ and } \widehat{C}_{p,\lambda} := \lambda * \widehat{C}_p$$

There ritty polynomials are called a *type C ritty polynomials*.

A (necessarily indecomposable and swappable) polynomial linearly related to a Chebyshev polynomial is called a *type C swappable polynomial*.

Remark 3.1.6. A Ritt swap involving two monomials or two Chebyshev polynomials really swaps two factors of a decomposition. While the non-monomial factors on the two sides of the last kind of a Ritt swap are different, it is useful to think of one of them *becoming* the other while simultaneously swapping places in the decomposition with the monomial.

To study indecomposables that might in this sense *become* type C or type J after a sequence of Ritt swaps, we introduce a formal notion and collect some taxonomy.

Definition 3.1.5. Given two indecomposable polynomials a and b, the relation "*a may become b after a sequence of Ritt swaps*" is the transitive closure of the relation "there exist indecomposable c and d such that $c \circ a = b \circ d$ or $a \circ c = d \circ b$.

Definition 3.1.6. (A list of different kinds of ritty polynomials)

- Monomials are self-explanatory.
- *Type C ritty polynomials* $C_{p,\lambda}$ and $\hat{C}_{p,\lambda}$ are defined above.
- *Type J ritty polynomials* are defined above.
- A ritty polynomial that is not type J but can become type J after a sequence of Ritt swaps is called a *type coJ ritty polynomial*.
- A ritty polynomial that is not a monomial, type C, J, or coJ is called a *type B ritty polynomial*.
- A (necessarily indecomposable and swappable) polynomial linearly related to a type X ritty polynomial is called a *type X swappable polynomial*, for X=C, J, coJ, B.

Remark 3.1.7. The symbols C, J and B are not entirely arbitrary, but we chose to spare the reader from our eccentric naming conventions: C for Chebyshevichi, J for Janus (Jani in the plural), and B for boring.

Definition 3.1.7.
- A class $\mathcal{C}$ of swappable polynomials is *closed under Ritt swaps* if whenever $f_i \in \mathcal{C}$ and $\vec{g}$ is obtained from $\vec{f}$ by a Ritt swaps at i (resp., $i-1$), then $g_{i+1} \in \mathcal{C}$ (resp., $g_{i-1} \in \mathcal{C}$). In other words, if a is in the class, and a may become b after a sequence of Ritt swaps in the sense of Remark 3.1.6, the b is also in the class.
- A class of ritty polynomials is *closed under Ritt swaps* if the same holds under the additional assumption that g_{i+1} (resp., g_{i-1}) is ritty.

Theorem 3.2. *The only ritty polynomials among the type C (respectively, J, coJ) swappable polynomials are the type C (respectively, type J, coJ) ritty polynomials. These classes of ritty polynomials - monomials, types C, J, coJ, and B - are disjoint and closed under linear relatedness, and cover all ritty polynomials.*
The class of type C swappable polynomials is closed under Ritt swaps; so is the class of type J and coJ swappable polynomials.
All translation relations among ritty polynomials are listed in propositions 3.1.2 and 3.1.3.

Proof. For the most part this theorem merely collects and translates into new notation the results of Propositions 3.1.3, 3.1.4 and 3.1.2 above. One bit worth explaining is that the classes of type C, type J, and type coJ swappables are disjoint. Type coJ ritty polynomials are not translation related to any other ritty polynomials, because their in-degree is too high for them to be type J, and their k is too high for them to be type C. In particular, type J or coJ ritty polynomials are not type C. The last part of the theorem is proved in Proposition 3.2.2 □

We now prove some useful consequences of this theorem, including Lemma 3.1.3. We begin with something of a converse to Corollary 3.1.1, as we use it in the proof of Lemma 3.1.3, which in turn is used to prove Lemma 3.2.2 which looks very similar to Lemma 3.2.1.

Lemma 3.2.1. *If a and b are ritty and not both type C, and L, M, and N are linear, and $(L \circ b \circ M^{-1}) \circ (M \circ a \circ N^{-1}) = \tilde{d} \circ \tilde{c}$ is a basic Ritt identity, then L, M, and N are scalings.*

Furthermore, there are ritty c and d such that $b \circ a = d \circ c$ is another basic Ritt identity, which is linearly equivalent to the first one, and in particular $(\tilde{d}, \tilde{c})$ is linearly equivalent to (d, c).

Proof. Since a and b are not both type C, one of them must be (linearly related to, and therefore equal to) a monomial.

If a is a monomial, then M and N must be scalings, since monomials are not translation-related to any ritty polynomial. Then, since both b and $L \circ b \circ M^{-1}$ are ritty and M is a scaling, L must also be a scaling because the equation in Problem 3.1.1 has no solutions with $B \neq 0 = A$.

If b is a monomial, then L and M must be scalings. Since both a and $(M \circ a \circ N^{-1})$ must be ritty, either N is a scaling or both a and $(M \circ a \circ N^{-1})$ must be type J. But $(L \circ b \circ M^{-1}, M \circ a \circ N^{-1})$ is swappable, contradicting remark 3.1.5.

"Furthermore" follows immediately from Corollary 3.1.1. □

Now we can finish the proof of Lemma 3.1.3. Afterwards, we must go back and prove the two propositions 3.1.2 and 3.1.3.

Proof. Let us recall the situation. We have ritty polynomials G_1, G_2, H_1, H_2, $\widehat{G}_1$, $\widehat{G}_2$, $\widehat{H}_1$, and $\widehat{H}_2$ and linear polynomials L,M, and N such that

$$L \circ G_1 \circ M^{-1} = G_2 \text{ and } M \circ H_1 \circ N^{-1} = H_2$$

and $G_j \circ H_j = \widehat{H}_j \circ \widehat{G}_j$ are basic Ritt identities for $j = 1$ and 2. We need to find a linear polynomial R such that

$$L \circ \widehat{H}_1 \circ R = \widehat{H}_2 \text{ and } R^{-1} \circ \widehat{G}_1 \circ N^{-1} = \widehat{G}_2$$

In fact, we show that $R = M^{-1}$ always works, and that it is always a scaling. We consider separately the three cases that none, one, or both of G_1 and H_1 are monomials. Since G_2 is linearly related to G_1, G_2 is a monomial if and only if G_1 is, and if both are monomials, then $G_1 = G_2$, and similarly for H_i.

(none) In this case, $\widehat{G}_i = G_i$ and $\widehat{H}_i = H_i$ are Chebyshev polynomials of odd degree, since commuting Chebyshevs are the only basic Ritt identity not involving any monomials. Then $R = M^{-1}$ works. (In fact, $L = M = N = (\cdot \pm 1)$ in this case, as Chebyshevs are not non-trivially linearly related to themselves except via $(-1) * C_p = C_p$.)

(two) In this case, $\widehat{G}_i = G_i$ and $\widehat{H}_i = H_i$ are monomials, since this is the only basic Ritt identity with two monomials on one side. Then $R = M^{-1}$ works. (In fact, L, M, and N are scalings in this case, as monomials are not non-trivially translation related to themselves.)

(one) This is done in Lemma 3.2.1, with $b := G_1$, $a := H_1$ with one less assumption. □

This proof does not use the hardest part of our analysis: it suffices to know that Problem 3.1.1 has no solutions with $A = 0 \neq B$, and to have a characterization of solutions with $B = 0 \neq A$, the type J ritty polynomials. Rest assured that we do need the full answer for the second crucial Lemma 3.2.6.

We end this section with a lemma closely resembling Lemma 3.2.1 because its proof is similar to the last two, though it will not be used until much later.

Lemma 3.2.2. *Suppose that a and b are ritty and neither is type C; and D, C, B and A are linear; and $\tilde{b} := D \circ b \circ C$ and $\tilde{a} := B \circ a \circ A$ are ritty, and $\tilde{b} \circ \tilde{a} = \tilde{d} \circ \tilde{c}$ is a basic Ritt identity.*

Then D, B, and A are scalings; and there are a translation T, a scaling C', and a ritty $b' := b \circ T$ such that $b \circ C = b' \circ C'$; and for some d and c, $b' \circ a = d \circ c$ is a basic Ritt identity. Unless b is type J, $T = \mathrm{id}$ and $b' = b$.

Proof. Since neither a nor b is type C, D and B must be scalings. If A is not a scaling, then a and $\tilde{a}$ must be type J, but this contradicts remark 3.1.5. So A is a scaling by λ and $\tilde{a} = \lambda * a$.

Write $C = T \circ C'$ for a scaling $C' =: (\cdot\mu)$ and a translation T. Then $b' := b \circ T$ is a monic polynomial scaling-related to the ritty $\tilde{b}$, so it is itself ritty. So if $T \neq \mathrm{id}$, then b is type J, and in any case $\tilde{b} = \mu * b'$.

So $(\mu * b') \circ (\lambda * a) = \tilde{d} \circ \tilde{c}$ is a basic Ritt identity. By Corollary 3.1.1, there are η and ν such that $b' \circ a = (\eta * \tilde{d}) \circ (\nu * \tilde{c})$ is a basic Ritt identity. □

3.2.1. *Zeros of $f(x) = x^k \cdot u(x^\ell)^n$ and its derivative.* We now turn to proving Propositions 3.1.2 and 3.1.3. It may be interesting to note that the only consequence of the indecomposability of f we use here is that $\gcd(k, \ell) = \gcd(k, n) = 1$, that is that no monomial is a compositional factor of f.

Some of our preliminary computations appear in [3].

In this subsection we fix a ritty polynomial $f(x)$ of the form $f(x) = x^k \cdot u(x^\ell)^n$ satisfying the requirements of Definition 3.0.5 and study its zeros as well as those of its derivative. We introduce several auxiliary polynomials and integer parameters and the association from f to these auxiliary objects is intended to be notationally uniform.

The following notation remains in force only throughout this Section 3.2.1.

Notation/Assumption 3.2.1. Given a nonzero polynomial $g(y)$ we define:

$$\widehat{g}(y) := \prod_{\{a : g(a) = 0\}} (y - a)$$
$$\widetilde{g}(y) := g(y)/\widehat{g}(y)$$
$$\bar{g}(y) := g'(y)/\widetilde{g}(y)$$

Let us return now to the case that $f(x) = x^k \cdot u(x^\ell)^n$ with the usual restrictions on the data defining f. As we will be studying the relations between f and other polynomials of a similar form, we fix $f_2(x) = x^{k_2} \cdot u_2(x^{\ell_2})^{n_2}$. For each of the terms defined for f, we have a corresponding term defined for f_2 noted with a subscript of "2."

Throughout this section we write $s := \deg(u)$ and $t := \deg(\widehat{u})$. We define two more associated polynomials

$$v(y) := u(y)^{n-1} \cdot \widetilde{u}(y)$$

$$w(y) := k\widehat{u}(y) + n\ell y \bar{u}(y)$$

Note that every zero of v is also a zero of u while w shares no zeros with u (and, hence, with v). Note also that $\deg(v) = (n-1)s + (s-t) = ns - t$ and $\deg(w) = t$.

A simple calculation shows that

$$f'(x) = x^{k-1} u(x^\ell)^{n-1} \widetilde{u}(x^\ell)[k\widehat{u}(x^\ell) + n\ell x^\ell \bar{u}(x^\ell)] = x^{k-1} \cdot v(x^\ell) \cdot w(x^\ell)$$

Thus, the zeros of f' may be described as follows.

(1) If $k > 1$, then $f'(0) = 0$ and $\mathrm{ord}_0(f') = k - 1$.

(2) Each ℓ^{th} root of a zero of v is a zero of both f and f'. Hence, counting multiplicity, there are $\ell(ns-t)$ zeros of $f'(x)$ arising as ℓ^{th} roots of zeros of v.
(3) There are ℓt zeros of f' arising as the ℓ^{th} roots of zeros of w. None of these are zeros of f.

From the above calculation, one sees that in some sense most of the ramification of f occurs above zero. (When we say that a zero a of f' *lies above* b, we simply mean that $f(a) = b$.)

Note that the zeros of f' of the last two kinds come in batches of size ℓ consisting of a (multiplicative) translate of the group of roots of unity over order dividing ℓ. Moreover, if a and b are zeros of the third kind and are in the same batch, then they lie above different points. Indeed, write $b = \zeta a \neq a$ for some ℓ^{th} root of unity ζ. We know that $w(a^\ell) = 0$ implies that $u(a^\ell) \neq 0$. As $\gcd(k, \ell) = 1$, we have $f(b) = (\zeta a)^k u((\zeta a)^\ell)^n = \zeta^k a^k u(a^\ell)^n \neq a^k u(a^\ell)^n = f(a)$.

Let us note some numerical consequences of these observations.

Observation 3.2.1. With the notation given above:

- # ramification points of f lying above zero $= (k-1) + \ell(ns-t)$
- # ramification points of f lying above other points $= \ell t$ and this set of ramification points is a union of t cosets of the group of ℓ^{th} roots of unity.

Let us return to the equation

$$(+B) \circ (x^k \cdot u(x^\ell)^n) \circ (+A) = x^{k_2} \cdot u_2(x^{\ell_2})^{n_2} \tag{1}$$

As the linear operations are invertible, we see that if $f_2(a) = f_2(b)$, then $f(a + A) = f(b + A)$.

Differentiating Equation 1, we have

$$f' \circ A = f_2'$$

Hence, for any point a, we have $\operatorname{ord}_a f_2' = \operatorname{ord}_{a+A} f$. That is, $(+A)$ translates the zeros of the derivative of f_2 onto the zeros of the derivative of f respecting multiplicities. If $B = 0$, then the ramification of f_2 above zero is matched precisely with the ramification of f above zero. If $B \neq 0$, then the ramification of f above one other point is matched with the ramification of f_2 above zero and *vice versa*. It is this consequence which makes these seemingly trivial observations useful.

Lemma 3.2.3. *If $\ell > 1$, then the sum of the roots of f is zero as is the sum of the roots of f'.*

Proof. f has a zero at 0 which contributes nothing to the sum while the rest of its zeros are the ℓ^{th} roots of the zeros of u. For each zero c of u pick one ℓ^{th} root b. Then the others have the form ζb for some ℓ^{th} root of unity. As the sum of the ℓ^{th} roots of unity is zero, the sum over all these roots is zero. From our expression for $f'(x)$, we see that it, too, may be expressed as a power of x times a function of x^ℓ. Hence, by the same reasoning the sum of its roots is zero. □

Corollary 3.2.1. *If $A \neq 0$ and $\ell > 1$, then the roots of $g(x) := (x^k \cdot u(x^\ell)^n) \circ (+A)$ do not sum to zero, and neither do the roots of $h'(x)$ where $h(x) = (+B) \circ (x^k \cdot u(x^\ell)^n) \circ (+A)$.*

Proof. Combining the above observations, we see that the sum of the roots of g is $A \cdot \deg(g)$ while the sum of the roots of $h'(x)$ is $A \cdot (\deg(h) - 1)$ (which is not zero as $\deg(h) \geq 3$). □

The following Corollary already appears in [3].

Corollary 3.2.2. *In the Equation 1 one of ℓ and ℓ_2 must be* 1.

Proof. Use the previous two results, remembering that we are only interested in the case $A \neq 0$. (Evaluating Equation 1 at 0 shows that $A = 0$ implies that $B = 0$.) □

Notation/Assumption 3.2.2. In light of Lemma 3.1.5 above, we may and do assume that $A = 1$ from now on.
In light of the last result, we may and do assume that $\ell = 1$ from now on, breaking the symmetry of the two sides of the equation.

In the next section, we address the possibility that $B = 0$; after that, we return to the other, significantly more difficult possibility.

3.2.2. *isolating type J.* In this section, we find all solutions to the equation

$$(x^k \cdot u(x)^n) \circ (+1) = (x^{k_2} \cdot u_2(x^{\ell_2})^{n_2})$$

that is, the special case of Problem 3.1.1 where $B = 0$ or, equivalently, where $u(1) = 0$. In light of Lemma 3.1.5, we have set $A = 1$ from Problem 3.1.1. We just proved that we may assume without loss of generality that $\ell = 1$, which is why it is missing.

Lemma 3.2.4. *Suppose that $f(x) := (x^k \cdot u(x)^n)$ and $f_2(x) := (x^{k_2} \cdot u_2(x^{\ell_2})^{n_2})$ and $f(x+1) = f_2(x)$. Then $\ell_2 = 1$.*

Proof. Since $f_2(x-1) = f(x)$, we see that $k = \operatorname{ord}_0(f) = \operatorname{ord}_{-1}(f_2)$. As $(-1)^{k_2} \neq 0$, it follows that -1 is a k-fold zero of $u_2(x^{\ell_2})^{n_2})$. Then any other ℓ_2nd root ζ of $(-1)^{\ell_2}$ is also a k-fold zero of $f_2(x)$. Thus unless $\ell_2 = 1$, some $(\zeta + 1) \neq 0$ is an exactly k-fold zero of f, and therefore of $u(x)^n$. Hence, n divides k. As $\gcd(n, k) = 1$, it follows that $n = 1$. But $\ell = n = 1$ so that f is not ritty. With this contradiction we see that $\ell_2 = 1$. □

Thus we only need to solve the equation $(x^k \cdot u(x)^n) \circ (+1) = (x^{k_2} \cdot u_2(x)^{n_2})$, with $u(1) = 0$.

Lemma 3.2.5. *If $u(1) = 0$ and the equation*

$$(x^k \cdot u(x)^n) \circ (+1) = (x^{k_2} \cdot u_2(x)^{n_2})$$

holds, then there are integers m and m_2 and a monic polynomial U such that $k = m_2 n_2$, $k_2 = mn$, $u(x) = (x-1)^m \cdot U(x)^{n_2}$, and $u_2(x) = (x+1)^{m_2} \cdot U(x+1)^n$.

Proof. We have $k_2 = \operatorname{ord}_0 f_2 = \operatorname{ord}_1 f$ implying that n divides k_2. Setting $m := k_2/n$ we see that $\operatorname{ord}_1 u = m$. Likewise, $k = \operatorname{ord}_0 f = \operatorname{ord}_{-1} f_2$ so that n_2 divides k. Write $m_2 := k/n_2$ and note that $\operatorname{ord}_{-1} u_2 = m_2$.

Write $u(x) = (x-1)^m P(x)$ and $u_2(x) = (x+1)^{m_2} P_2(x)$. Then, our equation is

$$(x+1)^{m_2 n_2} \cdot [x^m P(x+1)]^n = x^{mn} \cdot [(x+1)^{m_2} P_2(x)]^{n_2}$$

Cancelling $(x+1)^{m_2 n_2} \cdot x^{mn}$, we obtain

$$P(x+1)^n = P_2(x)^{n_2}$$

Recalling that $k = n_2 m_2$ and n are relatively prime, so that $\gcd(n, n_2) = 1$, it must be that P is an n_2^{th} power and P_2 an n^{th} power. Write $P(x) = U(x)^{n_2}$ and $P_2(x) = U_2(x)^n$. As $U(x+1)^{nn_2} = U_2(x)^{nn_2}$, if we take these polynomials to be monic, then we have $U(x+1) = U_2(x)$. □

Together, these lemmata prove the Proposition 3.1.2 above and vindicate Definition 3.1.3.

Proposition 3.2.1. *All solutions to the equation*

$$(+B) \circ (x^k \cdot u(x^\ell)^n) \circ (+1) = (x^{k_2} \cdot u_2(x^{\ell_2})^{n_2})$$

with $u(1) = 0$ *are of the form*

$$(x^{m_2 n_2} \cdot (x-1)^{mn} U(x)^{n_2 n}) \circ (+1) = x^{mn} \cdot (x+1)^{m_2 n_2} \cdot U(x+1)^{n_2 n}$$

for some monic polynomial U *and integers* m *and* m_2 *satisfying* $\gcd(mn, n_2) = \gcd(m_2 n_2, n) = 1$.

3.2.3. $B \neq 0$, *a first reduction.* We are left with the possibility that $u(1) \neq 0$ in the equation in Problem 3.1.1. Recall that we have reduced to the case that $A = 1$ (Lemma 3.1.5) and $\ell = 1$ (Corollary 3.2.2). Thus we need to solve the equation

$$(+B) \circ (x^k \cdot u(x)^n) \circ (+1) = (x^{k_2} \cdot u_2(x^{\ell_2})^{n_2})$$

when $B \neq 0$. This section is devoted to showing that $n = n_2 \ell_2 = 2$.

Then, the following two sections examine the two possibilities, $n_2 = 2 = 2\ell_2$ and $\ell_2 = 2 = 2n_2$.

Proposition 3.2.2. *If* $u(1) \neq 0$ *and the following equation holds*

$$(+B) \circ (x^k \cdot (u(x^\ell)^n)) \circ (+1) = (x^{k_2} \cdot u_2(x^{\ell_2})^{n_2})$$

then $k = k_2 = 1$, $s = t$, $s_2 = t_2$ *(that is, all roots of* u *and of* u_2 *are simple), and* $n = n_2 \ell_2 = 2$.

Proof. Note that $u(1) \neq 0$ if and only if $B \neq 0$.

Recall that for any ξ we have $\operatorname{ord}_\xi f_2' = \operatorname{ord}_{\xi+1} f'$ and that if $\xi \neq 0$ and $f_2'(\xi) = 0$, then for any ζ with $\zeta^{\ell_2} = 1$ we have $\operatorname{ord}_{\zeta\xi} f_2'$. Recall our observation that as long as $f_2(\xi) \neq 0$ and $\zeta \neq 1$ is an ℓ_2^{nd} root of unity as above, $f_2(\xi) \neq f_2(\zeta\xi)$. In particular, for each of the batches of ramification points of f_2 described in in Observation 3.2.1 there is at most one which lies above B. Translation by 1 takes the ramification of f_2 to ramification of f and it takes the points lying above B to point lying above 0. Thus,

- at most one point in each batch of the ramification of f_2 above a point other than zero goes to the ramification of f over 0 and
- the rest of the ramification points of f_2 must go to the ramification of f above other points.

Translating into inequalities involving the various degrees of f and f_2 using the calculations in Observation 3.2.1 and recalling that $\ell = 1$ we have the following

- $(k-1) + (ns - t) \leq t_2$
- $(k_2 - 1) + \ell_2(n_2 s_2 - t_2) + (\ell_2 - 1)t_2 \leq t$

Adding and collecting ts on the right, we have

$$(k_2 - 1) + (k-1) + \ell_2 n_2 s_2 + ns \leq 2t + 2t_2$$

Recall that $t \leq s$, $t_2 \leq s_2$, $n = n\ell \geq 2$, $\ell_2 n_2 \geq 2$, $k \geq 1$, and $k_2 \geq 1$. Thus, all of these inequalities must be equalities. □

Since the numbers n_2 and ℓ_2 are positive integers, the inequality $\ell_2 n_2 = 2$ implies $\ell_2 = 1$ and $n_2 = 2$ or $\ell_2 = 2$ and $n_2 = 1$.

3.2.4. *uniqueness of solutions with $B \neq 0$.* The next two propositions prove Proposition 3.1.3.

Proposition 3.2.3. *There is exactly one u of each degree satisfying*

$$(+B) \circ (x \cdot u(x)^2) \circ (+1) = (x \cdot u_2(x)^2)$$

Proof. Consider the equation

$$(+B) \circ (x \cdot u(x)^2) \circ (+1) = (x \cdot u_2(x)^2) \tag{2}$$

with $B \neq 0$ and u and u_2 having only simple roots. Since $B \neq 0$, it follows that $u(x+1)$ and $u_2(x)$ are coprime. Differentiating Equation 2 and using the notation from the beginning of this section we have

$$v(x+1)w(x+1) = v_2(x)w_2(x) \tag{3}$$

or

$$u(x+1) \cdot (u(x+1) + 2(x+1)u'(x+1)) = u_2(x) \cdot (u_2(x) + 2xu_2'(x)) \tag{4}$$

From the coprimality of $u(x+1)$ and $u_2(x)$ and noting the leading coefficients we obtain two equations.

$$(2s+1)u_2(x) = u(x+1) + 2(x+1)u'(x+1) \tag{5}$$

$$(2s+1)u(x+1) = u_2(x) + 2xu_2'(x) \tag{6}$$

Differentiating Equation 5 again we get

$$(2s+1)u_2'(x) = 3u'(x+1) + 2(x+1)u''(x+1) \tag{7}$$

Multiplying Equation 6 by $(2s+1)$, and then using Equations 5 and 7 to eliminate u_2 and u_2', we obtain

$$(2s+1)^2 u(x+1) = u(x+1) + 2(x+1)u'(x+1) + 2x(3u'(x+1) + 2(x+1)u''(x+1)) \tag{8}$$

Collecting terms, we see that $u(x+1)$ must satisfy the following ODE

$$(2s^2 + 2s)Y + (3 - 4z)Y' + 2(z - z^2)Y'' = 0 \tag{9}$$

A routine calculation shows that if $u(x+1)$ is a solution to Equation 9 and we define $u_2(x)$ via Equation 5 and set $B := -u(1)^2$, then these data satisfy Equation 2.

The linear differential operator $L = 2(z - z^2)\partial^2 + (3 - 4z)\partial + (2s^2 + 2s)$ defines a linear operator on the $s+1$-dimensional space of polynomials of degree s. With respect to the standard monomial basis of this space, the matrix $M = (M_{i,j})$ of L is upper triangular. On the main diagonal, we have $M_{i,i} = 2(1-i)i - 4i + (2s^2 + 2s) =$

$(2s^2+2s)-(2i^2+2i)$ and just above the diagonal we have $M_{i,i+1}=(i+1)(3+2i)$. In particular, $M_{s,s}=0$ so that $\operatorname{rank}(L)\leq s$ whilst the (s,s)-minor is invertible. Thus, the rank of L is s and the dimension of the space of solutions to Equation 9 is exactly one. As we require u to be monic, there is exactly one solution of degree s. □

Proposition 3.2.4. *For each positive integer s, there is a unique monic polynomial u of degree s and nonzero parameter B for which there is another monic polynomial u_2 satisfying*

$$(+B)\circ(x\cdot u(x)^2)\circ(+1)=(x\cdot u_2(x^2)) \tag{10}$$

Proof. As before, since $B\neq 0$, $u(x+1)$ and $u_2(x^2)$ are coprime. Differentiating, we obtain

$$u(x+1)\cdot(u(x+1)+2(x+1)u'(x+1))=u_2(x^2)+2x^2u_2'(x^2)=(u_2+2x\cdot u_2')\circ Q \tag{11}$$

The zeros of the righthand side of Equation 11 come in $\pm$-pairs. We claim that for each such pair one is a root of $u(x+1)$ and the other is a root of $(u(x+1)+2(x+1)u(x+1))$. Indeed, it cannot happen that $u(c+1)=0$ and $u(-c+1)=0$ for Equation 10 would yield $cu_2(c^2)=B=-cu_2((-c)^2)=-cu_2(c^2)$ contrary to the fact that $B\neq 0$. Thus, at most one of each pair of roots of the righthand side is also a root of $u(x+1)$. As the degree of the righthand side of Equation 11 is twice that of u, it follows that at least one root from each pair must be a root of $u(x+1)$. Matching leading coefficients, we conclude:

$$(-1)^s(2s+1)u(x+1)=(u(-x+1)+2(-x+1)u'(-x+1)) \tag{12}$$

Substituting $z:=-x+1$, we see that u satisfies the following difference-differential equation:

$$0=u(z)+2zu'(z)-(2s+1)(-1)^s u(2-z) \tag{13}$$

The difference-differential operator in Equation 13 is a linear operator on the space of degree s polynomials and it is given by an upper triangular matrix relative to the standard monomial basis. The entries along the main diagonal are

$$1+2i-(-1)^{i+s}(2s+1)$$

Hence, the rank of this operator is exactly s meaning that there is a unique monic solution. □

We are now done with linearly related ritty polynomials, and the proof of the crucial Lemma 3.1.3.

3.3. **The Ritt monoid.** With Lemma 3.1.3 in place we can describe the sequences of Ritt swaps as a monoid acting on the set of linear-equivalence classes of decompositions. Unfortunately, we will need to return to some messy computations to verify that this action works as required.

Definition 3.2.1. For a non-linear polynomial f we let D_f be the set of distinct decompositions of f up to linear equivalence. That is, $D_f := \{[f_k, \dots, f_1] : f = f_k \circ \cdots \circ f_1, \text{ each } f_i \text{ indecomposable } \}$ where $[f_k, \dots, f_1]$ is the equivalence class of the decomposition with respect to linear equivalence.

For the remainder of this section we will work with a fixed polynomial f admitting a decomposition of length k. By Ritt's theorem, all decompositions of f have length k.

Definition 3.2.2. Let M_k be the free monoid on the $(k-1)$ generators $t_1, \dots, t_{k-1}$.

Our work from the previous section allows us to define an action of M_k on D_f whereby t_i acts by converting a (linear equivalence class of a) decomposition into the (linear equivalence class of a) decomposition obtained by a Ritt swap at i, if possible. A word in the t_is (that is, an element of M_k) corresponds to a sequence of Ritt swaps. This action is very close to the action of the symmetric group S_k with t_i identified with the transposition $(i\ i+1)$.

Definition 3.2.3. We define an action of M_k on $D_f^* := D_f \cup \{\infty\}$ as follows. If some decomposition $\vec{h}$ may be obtained from $\vec{f}$ by a Ritt swap at i, we say that $t_i \star [\vec{f}] = [\vec{h}]$; if no decomposition may be obtained from $\vec{f}$ by a Ritt swap at i, we say that $t_i \star [\vec{f}] = \infty$; and $t_i \star \infty = \infty$ for all i.

For $w \in M_k$ and $[\overrightarrow{f}] \in D_f$ we say that $w \star [\overrightarrow{f}]$ is *defined* if $w \star [\vec{f}] \neq \infty$.

We often abuse notation writing $w \star \overrightarrow{f} = \overrightarrow{g}$ for $w \star [\overrightarrow{f}] = [\overrightarrow{g}]$.

Ritt's theorem may be restated as saying that this action is transitive on D_f.

Our goal now is to show that this action is more or less the same as an action of the permutation group S_k on a set with k elements. Of course, this cannot be literally true, but the result is close enough for our purposes.

Fact 3.2.1. *The following is a presentation of S_k, the symmetric group on k letters.*
Generators*:*

$$(i\ i+1) \text{ for } 1 \leq i < k$$

Relations*:*

$$(i\ i+1)^2 = \mathrm{id}$$
$$(i\ i+1)(j\ j+1) = (j\ j+1)(i\ i+1) \text{ for } j \neq i \pm 1$$
$$(i\ i+1)(i+1\ i+2)(i\ i+1) = (i+1\ i+2)(i\ i+1)(i+1\ i+2)$$

To make the connection more precise, we define

Definition 3.2.4. The *permutation represented by* a word $t_{a_r} \dots t_{a_2} t_{a_1}$ in M_k is $(a_r + 1\, a_r) \dots (a_2 + 1\, a_2)(a_1 + 1\, a_1)$.

Our fundamental result is that our action satisfies the relations satisfied by S_k:

Lemma 3.2.6. *For any $\overrightarrow{f} \in D_f$ and $i < k$*

- *If $t_i \star \overrightarrow{f}$ is defined, then $t_i^2 \star \overrightarrow{f} = \overrightarrow{f}$.*
- *For $j \neq i \pm 1$, $t_i t_j \star \overrightarrow{f} = t_j t_i \star \overrightarrow{f}$. In particular, one is defined if and only if the other is.*
- *$t_i t_{i+1} t_i \star \overrightarrow{f} = t_{i+1} t_i t_{i+1} \star \overrightarrow{f}$. In particular, one is defined if and only if the other is.*

The first two parts are immediate from the definition. The last part will be proved in the next section. It is clear that the qualifier 'if defined' is necessary in the first clause of the lemma. In most cases this rules out a true true action of the permutation group on D_f, i.e. the possibility that any two words in M_k representing the same permutation act the same way. However, we will show in Corollary 3.2.3 that if two words w and w' represent the same permutation and both $w \star \vec{f}$ and $w' \star \vec{f}$ are defined, then they are equal.

3.4. **Descalings, double-jumps, and the crucial combinatorial lemma.** The main aim of this section is

Proposition 3.2.5 (Crucial combinatorial lemma)**.** *For any decomposition $\vec{f}$ and for any i, $t_i t_{i+1} t_i \star \vec{f} = t_{i+1} t_i t_{i+1} \star \vec{f}$, i.e. that one is defined if and only if the other is, and when both are defined, the result is the same.*

The second part follows immediately form the results of Muller and Zieve [13], but we do not see how to obtain the first part without our intermediate results. We also use descalings and double-jumps, defined and described in this section, in section 3.6.1.

3.4.1. *Descalings.* What we do here is overkill for the Crucial Combinatorial Lemma at hand, but we will need it in the proof of Theorem 3.3, and it makes sense to develop the ideas here.

Definition 3.2.5. Given a decomposition $\vec{f}$, we deine a *descaling of $\vec{f}$* to be a k-tuple $(M_i, h_i, L_i)_{i \le k}$ of triples where h_i are ritty, and linear L_i and M_i are translations except possibly for M_k and L_1, such that $((M_k \circ h_k \circ L_k), \dots (M_1 \circ h_1 \circ L_1))$ is linearly equivalent to $\vec{f}$.

Lemma 3.2.7. *If all factors f_i in a decomposition $\vec{f}$ are swappable, then $\vec{f}$ admits a descaling (M_i, h_i, L_i), in which we may choose, for each $1 \le i < k$, to have $M_i = \mathrm{id}$ or $L_{i+1} = \mathrm{id}$, and also choose to have one of M_k and L_1 to be a translation.*

Proof. We sketch the proof for making L_1 a translation and $M_i = \mathrm{id}$ for all $i < k$. Other options are identical. First, find ritty g_i and linear B_i and A_i such that $f_i = B_i \circ g_i \circ A_i$. To make L_1 a translation, write $A_1 = (\cdot\lambda) \circ L_1$, let $h_i := \lambda * g_i$. Now $\vec{f}$ is linearly-equivalent to $(f_k, \dots f_2 \circ (\cdot\lambda^{\deg(f_1)}), h_1 \circ L_1)$. Inducting on k finishes the proof. □

Definition 3.2.6.

- A *left descaling* of a decomposition is a descaling (M_i, h_i, L_i) where $M_i = \mathrm{id}$ for all $i \ne k$ and L_1 is a translation.
- A left descaling (M_i, h_i, L_i) *has no loose translations at j* if neither h_j nor h_{j+1} is type C, and if h_{j+1} is type J, then $L_{j+1} = M_j = \mathrm{id}$.
- A left descaling (M_i, h_i, L_i) *has no essential translations at j* if neither h_j nor h_{j+1} is type C, and $L_{j+1} \circ M_j = \mathrm{id}$.
- A left descaling (M_i, h_i, L_i) *has no loose (resp. essential) translations* if it has no loose (resp. essential) translations at j for any $1 \le j < k$.
- If (M_i, h_i, L_i) is a descaling of $\vec{f}$, then a *compatible descaling of $t_j \star \vec{f}$* is (M_i, h'_i, L_i) where $h'_i = h_i$ for all $i \ne j, j+1$, and $L_{j+1} = M_j = \mathrm{id}$, and $h'_{j+1} \circ h'_j = h_{j+1} \circ h_j$.

With two compatible descalings, the linear factors witnessing the Ritt swap may be taken to be identity.

Lemma 3.2.8. *If a left descaling (M_i, h_i, L_i) of $\vec{f}$ has no loose translations at j and $t_j \star \vec{f}$ is defined, then (M_i, h_i, L_i) has no essential translations at j and $t_j \star \vec{f}$ admits a descaling compatible with this Ritt swap.*

Proof. Let us unwrap the definitions. Since $j < k$ and this is a left descaling, we are assuming that $M_j = \mathrm{id}$. We are assuming that neither h_{j+1} nor h_j is of type C, and that if h_{j+1} is type J, then $L_{j+1} = M_j = \mathrm{id}$. We need to show that $L_{j+1} = \mathrm{id}$, and that there are ritty g_{j+1} and g_j such that $g_{j+1} \circ g_j = h_{j+1} \circ h_j$ is a Ritt swap.

We first show that $L_{j+1} = \mathrm{id}$. We already know this when h_{j+1} is type J, so we assume it is not. So we have ritty h_{j+1} and h_j and a translation L_{j+1} such that $(h_{j+1} \circ L_{j+1}, h_j)$ is swappable, but h_{j+1} is neither type C not type J, and h_j is not type C. This is clearly impossible unless $L_{j+1} = \mathrm{id}$.

Now neither h_{j+1} nor h_j is type C, and (h_{j+1}, h_j) is swappable, and lemma 3.2.1 finds the desired ritty g_{j+1} and g_j for us. □

Proposition 3.2.6. *Suppose that $(M_i, h_i, L_i)_{i \leq k}$ is a left descaling of $\vec{f}$ which has no loose translations, let $v := (t_{k-1} t_{k-2} \dots t_1)$ and suppose that $v \star \vec{f}$ is defined. Then (M_i, h_i, L_i) has no essential translations, and $v \star \vec{f}$ admits a left descaling with no essential translations, compatible with the Ritt swaps in v.*

Proof. Since $t_1 \star \vec{f}$ is defined and (M_i, h_i, L_i) has no loose translations at 1, the last lemma says that it has no essential translations at 1 and $t_1 \star \vec{f}$ admits a compatible descaling (M_i, h_i, L_i). That is, $(M_i, h_i, L_i) = (M_k \circ h_k \circ L_k, h_{k-1} \circ L_{k-1}, \dots h_3 \circ L_3, h_2, h_1 \circ L_1)$ and there are ritty g_2, g_1 such that $M_k \circ h_k \circ L_k, h_{k-1} \circ L_{k-1}, \dots h_3 \circ L_3, g_2, g_1 \circ L_1)$ is a descaling of $t_1 \star \vec{f}$, and $h_2 \circ h_1 = g_2 \circ g_1$. Note that this new decomposition is still a left descaling. Since neither h_2 nor h_1 were type C, and the class of type C ritty polynomials is closed under Ritt swaps, the new ritty factors g_2 and g_1 are not type C. Also, this new descaling does not have any loose translations: at all $j \neq 1$, it inherits this property from the old descaling, and at 1 is has no translation as required.

Thus, the new decomposition satisfies the hypotheses of this proposition, so we can apply the lemma again to show that $L_3 = \mathrm{id}$ and obtain ritty $\widetilde{g}_3$ and g_2 such that $\widetilde{g}_3 \circ g_2 = h_3 \circ g_2$ and $(M_k \circ h_k \circ L_k, h_{k-1} \circ L_{k-1}, \dots h_4 \circ L_4, \widetilde{g}_3, g_2, g_1 \circ L_1)$ is a descaling of $t_2 t_1 \star \vec{f}$, again satisfying the hypotheses of the proposition.

Repeating this process $(k-1)$ times we end up with a descaling $(M_k \circ g_k, g_{k-1}, \dots g_2, g_1 \circ L_1)$ of $v \star \vec{f}$ that satisfies the conclusions of the Proposition. □

Lemma 3.2.9. *Again, let $v := (t_{k-1} t_{k-2} \dots t_1)$ and suppose that $v \star \vec{f}$ is defined. If no factor of $\vec{f}$ is type C, then $\vec{f}$ admits a left descaling with no loose translations.*

Proof. We showed before that $\vec{f}$ admits a left descaling $(M_k \circ h_k \circ L_k, h_{k-1} \circ L_{k-1}, \dots h_2 \circ L_2, h_1 \circ L_1)$. We would like to show that for each $j \geq 2$, if h_j is type J, then $L_j = \mathrm{id}$. Instead, we will show that if there is a left descaling of $\vec{f}$ where this first fails at j, then there is another left descaling of $\vec{f}$ where this first fails at $j+1$. Thus, we will show by induction that there is a left descaling where this never fails.

To start from the beginning, suppose this fails at 1, so that h_2 is type J but $L_2 \neq \mathrm{id}$. However, $t_1 \star \vec{f}$ is defined, so $(h_2 \circ L_2, h_1 \circ L_1)$ is swappable. Since h_2 is type J, h_1 must be a monomial, so $h_2' := h_2 \circ L_2$ must itself be ritty. So we have found

a new left descaling of $\vec{f}$, namely $(M_k \circ h_k \circ L_k, h_{k-1} \circ L_{k-1}, \dots h_3 \circ L_3, h_2', h_1 \circ L_1)$ with no loose translations at 1. It is clear that we can continue inductively. □

3.4.2. *Double-jumps and the proof of the crucial combinatorial lemma.* We first make a few observations about words of the form $t_i t_{i+1}$ or $t_{i+1} t_i$, called *double-jumps*. Most of the time, their action is undefined.

Definition 3.2.7. A *double-jump* a sequence of two Ritt swaps of the form $t_i t_{i+1}$ or $t_{i+1} t_i$.

Since only three compositional factors are involved in a double-jump, it is sufficient to characterize the case $i = 1$.

If $j_l \circ \hat{l} \circ r = l \circ j \circ r = l \circ \hat{r} \circ j_r$ witnesses the fact that $t_1 t_2 \star (j_l, \hat{l}, r) = (l, \hat{r}, j_r)$ then we say that j_l *double-jumps* $(\hat{l}, r)$ *from the left* and so $(\hat{l}, r)$ is *double-jumpable from the left*.

Reading the same equation right-to-left, $j_l \circ \hat{l} \circ r = l \circ j \circ r = l \circ \hat{r} \circ j_r$ witnesses the fact that $t_2 t_1 \star (l, \hat{r}, j_r) = (j_l, \hat{l}, r)$; we say that j_r *double-jumps* $(l, \hat{r})$ *from the right*, and so $(l, \hat{r})$ is *double-jumpable from the right*.

Remark 3.2.1. In order for a double-jump to be defined, all of the factors above must be swappable, i.e. linearly related to a ritty polynomial.

Lemma 3.2.10. *If in the middle of a double-jump j is an odd Chebyshev polynomial, then (l, j, r) is linearly equivalent to $(A \circ C_p, C_q, C_r \circ B)$ for some prime, not necessarily odd, p and r and some linear A and B. (We allow the possibility that $p = 2$ and/or $r = 2$.)*

Proof. We already know that l and r must be linearly related to (possibly degree 2) Chebyshev polynomials, so we may write $(l, j, r) = (A' \circ C_p \circ A'', C_q, B'' \circ C_r \circ B')$. Our purpose is to get rid of A'' and B''. We show how to get rid of A'' and assure you that the other half of the proof is exactly the same. Since (l, j) is swappable, there are linear L, M, and N such that $L^{-1} \circ A' \circ C_p \circ A'' \circ M$ and $M^{-1} \circ C_q \circ N$ are ritty and $[L^{-1} \circ A' \circ C_p \circ A'' \circ M] \circ [M^{-1} \circ C_q \circ N]$ is one side of a basic Ritt identity. If p is also odd, the basic Ritt identity must be $C_p \circ C_q = C_q \circ C_p$. Since odd Chebyshev polynomials are not non-trivially linearly related to themselves except for $(-1) * C_p = C_p$, we must have $L^{-1} \circ A' = A'' \circ M = (\cdot \pm 1)$ and $M^{-1} = N = (\cdot \pm 1)$, in particular making $A'' = (\cdot \pm 1)$, as wanted.

If $p = 2$, then the basic Ritt identity must be $Q \circ C_{q,\lambda} = \widehat{C}_{q,\lambda} \circ Q$ for some λ. Then $M^{-1} = (\cdot \mu)$ where $\mu = \frac{1}{\lambda^{2q+1}}$; and $L^{-1} \circ A' \circ C_2 \circ A'' \circ M = Q$.

Write $A'' = (+D) \circ (\cdot \nu)$ and apply $* \frac{1}{\mu\nu}$ to both sides to get $(\cdot \mu^2 \nu^2) \circ L^{-1} \circ A' \circ C_2 \circ (+D) = Q$. Therefore, $D = 2$, and

$$A' \circ C_2 \circ A'' = A' \circ C_2 \circ (+2) \circ (\cdot \mu) = A' \circ (+2\mu^2 - 2) \circ (\cdot \mu^2) \circ C_2$$

So we let $A = A' \circ (+2\mu^2 - 2) \circ (\cdot \mu^2)$ and obtain the desired conclusion. □

Lemma 3.2.11. *If j in a double-jump is not quadratic, then there are linear L and M and ritty $\widetilde{l}$, $\widetilde{j}$ and $\widetilde{r}$ such that $(L \circ \widetilde{l}, \widetilde{j}, \widetilde{r} \circ M)$ is linearly equivalent to (l, j, r).*

Proof. If $j = A \circ C_p \circ B$ is odd type C, we apply the previous lemma to $(l \circ A, C_p, B \circ r)$. Since $t_2 \star (l, j, r)$ is defined, j cannot be type J, as per remark 3.1.5. The hypothesis of the lemma explicitly rules out quadratic j. The only possibilities left are that j is type B, type coJ, or an odd monomial. In any case, since (l, j)

and (j, r) are both swappable, neither l nor r can be type C, so Proposition 3.2.6 applies and the rest is routine. □

We now give a proof of one of the directions of Proposition 3.2.5, and note that the same proof gives the other direction simply by reading equation 14 below right-to-left instead of left-to-right.

Proof. Going back to the original question and naming everything, suppose that $t_1t_2t_1 \star (a_0, b_0, c_0)$ is defined; we need to show that $t_2t_1t_2 \star (a_0, b_0, c_0)$ is also defined and equal. In

$$(a_0, b_0, c_0) \overset{t_1}{\mapsto} (a_0, c_1, b_1) \overset{t_2}{\mapsto} (c_2, a_2, b_1) \overset{t_1}{\mapsto} (c_2, b_3, a_3) \tag{14}$$

c_0 double-jumps (a_0, b_0) from the right, and a_0 double-jumps (c_1, b_1) from the left, with c_1 playing the role of j in the first case, and a_2 in the second case.

Since $a_0 \circ c_1 = c_2 \circ a_2$ is a Ritt swap, c_1 and a_2 cannot both be quadratic. Then, by Lemma 3.2.11 we get compatible descalings of (a_0, c_1, b_1) and (c_2, a_2, b_1), and the rest is routine. □

3.5. **canonical forms.**

3.5.1. *motivations and definitions.* We now work on finding canonical forms for words in M_k, such that every word has an *equivalent* (defined below) word of this canonical form. In particular, equivalent words represent the same permutation (see Definition 3.2.4). We obtain these canonical words $\hat{w}$ from the original word w by a sequence of syntactic operations on substrings of w. We just proved Lemma 3.2.6, which shows that we may

(1) replace t_it_i by the empty string;
(2) replace t_it_j by t_jt_i for non-consecutive i and j;
(3) replace $t_{i+1}t_it_{i+1}$ by $t_it_{i+1}t_i$, or vice versa.

These operations allow us to obtain equivalent words of two canonical forms, one roughly corresponding to an insert-sort, the other to a merge-sort, if one thinks of permuting factors as putting them in a particular order. Surely these combinatorial computations have been worked out by computer scientists, but we failed to find a reference in the literature. To the best of our understanding, our results on canonical forms do not follow easily from [13]. Although our main result of this section, Theorem 3.3, can be immediately deduced from their work, we need these canonical forms in the next section 3.6.1, whose results cannot be obtained easily from their work.

Remark 3.2.2. If a word v is obtained from a word w via finitely many applications of rules (1), (2) and (3) above, then they represent the same permutation, the length of v is less than or equal to the length of w, and for any decomposition $\vec{f}$, if $w \star \vec{f}$ is defined, the $v \star \vec{f} = w \star \vec{f}$. It may be that $v \star \vec{f}$ is defined while $w \star \vec{f}$ is not.

Definition 3.2.8. If two words v and w in M_k can be obtained from each other by operations (2) and (3) above, we write $w \approx v$.

Definition 3.2.9. A word w *is length-minimal* if no strictly shorter word v may be obtained from w via finitely many applications of rules (1), (2) and (3) above.

To shorten our proofs, we will assume that we start with a length-minimal word, and reach a contradiction every time we get a chance to cancel $t_i t_i$. Equivalently, we could induct on the length of the word and cite the inductive hypothesis every time we get a chance to cancel $t_i t_i$ obtaining a shorter equivalent word. We do not, because there are already far too many inductions in these proofs.

3.5.2. *The first canonical form:* A sequence of Ritt swaps in the first canonical form acts by moving several factors f_{b_i} in the decomposition some number of steps to the left; the first factor it moves, f_{b_1}, begins to the left of the next factor moved, f_{b_2}, which begins left of f_{b_3}, etc. In computer science, this is called an insert-sort: having arranged f_k through f_{i+1} in the right order, this sequence *inserts* f_i in the required place among f_k through f_{i+1}, and then proceeds to f_{i-1}, and so on, until all factors are ordered the right way.

Proposition 3.2.7. *For every $w \in M_k$ there exists a unique $\hat{w} \in M_k$ which represents the same permutation as w and has the form*

$$\hat{w} = (t_{a_k} t_{a_k - 1} \dots t_{b_k})(t_{a_{k-1}} t_{a_{k-1}-1} \dots t_{b_{k-1}}) \dots (t_{a_1} t_{a_1 - 1} \dots t_{b_1})$$

with $a_k \geq b_k$ for all k, and $b_k < b_{k-1} < \dots < b_1$.

This $\hat{w}$ is obtained from w by operations (1), (2), and (3) above, so for any decomposition $\vec{f}$ such that $w \star \vec{f}$ is defined, $\hat{w} \star \vec{f} = w \star \vec{f}$.

Proof. We begin by replacing w by some w' that has the shortest length among words that can be obtained from w by operations (1), (2), and (3) above. This means that as we continue to perform these operations on w', we should never be able to perform operation (1) as that would shorten the word.

Without the requirement that $b_k < b_{k-1} < \dots < b_1$, the proposition follows trivially, by cutting the word w' into maximal consecutive-decreasing-index substrings. So it is only when $b_{i+1} \geq b_i$ that we need to do anything. Note that any substring of w' is also length-minimal:

Lemma 3.2.12. *If $w' = tuv$ is length-minimal, then u is length-minimal.*

Note that if we can fix one pair of out-of-order b_i's, we can fix everything in finitely many steps; so it suffices to prove the proposition for

$$w' = (t_a t_{a-1} \dots t_b)(t_c t_{c-1} \dots t_d)$$

If $b < d$, then w' is already of the desired form. So assume $b \geq d$.

Now compare b and c:

- If $b > c + 1$, then $\hat{w} = (t_c t_{c-1} \dots t_d)(t_a t_{a-1} \dots t_b)$ works, because in this case each t_i in the first chunk of w' commutes with each t_j in the second chunk, and $b > c + 1 > d$.
- If $b = c + 1$, w' is already of the desired form, a single consecutive-descending-index string.
- If $b = c$, operation (1) shortens the word w' contradicting length-minimality.
- This leaves the case where $c > b \geq d$, which needs a lemma:

Lemma 3.2.13.
- *if $r+1 > r \geq s$, then $t_r(t_{r+1} t_r t_{r-1} \dots t_s) \approx (t_{r+1} t_r t_{r-1} \dots t_s) t_{r+1}$*
- *if $p > r \geq s$, then $t_r(t_p t_{p-1} \dots t_s) \approx (t_p t_{p-1} \dots t_s) t_{r+1}$*

Proof. For (1), $t_r t_{r+1} t_r \approx t_{r+1} t_r t_{r+1}$, and then t_{r+1} commutes with t_{r-1} through t_s. For (2), note that t_r commutes with t_p through t_{r+2} and then (1) applies. □

Returning to the special case of the proposition, we compare a and c:

- If $a < c$, then the lemma can be applied to each t_i for $a \geq i \geq b$ giving $w' \approx (t_c t_{c-1} \dots t_d) t_{a+1} t_a \dots t_{b+1} =: \hat{w}$, of the desired form because $d \leq b$ implies $d < b + 1$.
- If $a \geq c$, the lemma can still be applied to each t_i for $c - 1 \geq i \geq b$ giving

$$w' = (t_a t_{a-1} \dots t_b)(t_c t_{c-1} \dots t_d) =$$
$$= t_a \dots t_c (t_{c-1} \dots t_b t_c t_{c-1} \dots t_d) \approx$$
$$\approx t_a \dots \mathbf{t_c} (\mathbf{t_c} t_{c-1} \dots t_d t_c \dots t_{b+1})$$

  contradicting length-minimality.

This finishes the proof of the special case of the proposition, and the whole proposition follows. □

Corollary 3.2.3. *If two words w and w' represent the same permutation and both $w \star \vec{f}$ and $w' \star \vec{f}$ are defined, then $w \star \vec{f} = w' \star \vec{f}$.*

Proof. For every permutation there is a unique word in the first canonical form representing it. □

The first canonical for is used a lot, in particular to obtain words in the second canonical form in the next section 3.5.3. One immediate consequence is a bound on the length of words and the number of decompositions of a given polynomial.

Corollary 3.2.4. *For any given polynomial P and decomposition $(f_k, \dots f_1)$ of P, there are at most $k!$ other decompositions $\vec{g}$ of P (up to linear equivalence, of course), and any one of them can be reached by a sequence of at most $\frac{k(k-1)}{2}$ Ritt swaps.*

3.5.3. *second canonical form.* The second canonical form is for refactoring decompositions of polynomials that come pre-factored into chunks: suppose $P = P_t \circ \dots \circ P_1$ and we have a decomposition $(f_{i,r_i}, \dots, f_{i,1})$ for each P_i. Then we want to first do as much shuffling as possible within the decompositions of P_i, and only then move factors between them. This corresponds to a merge-sort, where first the factors in each chunk are put into the order in which they will appear in the final decomposition, and then the chunks are *merged*. More precisely,

Proposition 3.2.8. *Given a polynomial $P = F_t \circ \dots \circ F_1$, decompositions $(f_{i,r_i}, \dots, f_{i,1})$ for each F_i, let $\vec{f} = (f_{t,r_t}, \dots f_{t,1}, f_{t-1,r_{t-1}}, \dots f_{1,1})$ be the decomposition of P obtained by concatenating the decompositions of the F_i.*

For every word w such that $w \star \vec{f}$ is defined, there is another word $\hat{w} = v w_1 w_2 \dots w_t$ such that:

- *$\hat{w} \star \vec{f} = w \star \vec{f}$*
- *each w_i only permutes factors $f_{i,j}$, so $w_1 w_2 \dots w_t \star \vec{f}$ is still a concatenation of decompositions of P_i.*
- *each of w_i and v is in the first canonical form.*
- *v never switches factors $f_{i,j}$ and $f_{i,j'}$ originating in the decomposition of the same F_i.*

The last item in the conclusion of the proposition is a bit vague, to be made more precise in Lemma 3.2.15 below. The rest of this section constitutes the proof of this proposition.

Remark 3.2.3. Although the statement of the proposition seems to be about a decomposition, this is in fact a purely combinatorial result that only depends on the numbers r_i and not on the particular polynomials. The new word will be obtained from the old word by operations (1), (2), and (3) defined at the beginning of this section, so the same new word will work for any decomposition. Given a tuple of positive integers r_i and a permutation of $k = \sum_i r_i$ elements, there is a unique word in M_k representing that permutation and satisfying the last three requirements of the proposition, so $\hat{w}$ is unique and does not depend on $\vec{f}$.

Lemma 3.2.14. *It is enough to prove the Proposition 3.2.8 for $t = 2$.*

Proof. We induct on t with $t = 2$ as base case. To prove the case $t = s + 1$ from $t = s$, first write

$$P = G_2 \circ G_1 \text{ where } G_2 = F_{s+1} \circ F_s \circ \ldots \circ F_2 \text{ and } G_1 = F_1$$

We first apply the case $t = 2$ to w and $P = G_2 \circ G_1$ getting $w \approx v_Q \circ u_1 \circ u_2$. We then apply the case $t = s$ to u_2 and $G_2 = F_{s+1} \circ F_s \circ \ldots \circ F_2$, getting $u_2 \approx v' w_2' w_3' \ldots w_{s+1}'$. So $w \approx v_Q u_1 v' w_2' w_3' \ldots w_{s+1}' \approx v_Q v' u_1 w_2' w_3' \ldots w_{s+1}'$, the second equivalence because u_1 and v' act on disjoint sets of factors. Finally, we notice that letting $v = v_Q v'$, $w_1 = u_1$, and $w_i = w_i'$ for $i \geq 2$ is of the desired form. □

Let us restate Proposition 3.2.8 for $t = 2$ with fewer subscripts and more precision:

Lemma 3.2.15. *(Proposition 3.2.8 for $t = 2$) Suppose that we have a polynomial $F = H \circ G$, decompositions $(g_k, \ldots g_1)$ of G and $(h_r, \ldots h_1)$ of H, and a word $w \in M_{k+r}$. Let $\vec{f} := (h_r, \ldots h_1, g_k, \ldots g_1)$, so $\vec{f}$ is a decomposition of F.*
Then there is another word $\hat{w} = v w_G w_H$ such that:

- $\hat{w} \star \vec{f} = w \star \vec{f}$
- *w_G it a word in t_1 through t_{k-1}, so it only permutes factors of G;*
- *w_H it a word in t_{k+1} through t_{r+k-1}, so it only permutes factors of H;*
- *all three pieces v, w_G, and w_H are in the first canonical form, so in particular*

  $$v = (t_{a_m} t_{a_m - 1} \ldots t_{b_m})(t_{a_{m-1}} t_{a_{m-1}-1} \ldots t_{b_{m-1}}) \ldots (t_{a_1} t_{a_1 - 1} \ldots t_{b_1})$$

- *v never switches factors that both originate in H or both originate in G.*

$$b_1 = k, b_2 = k - 1, \ldots b_m = k - m + 1 \text{ and } a_1 > a_2 > \ldots a_m$$

Proof. We may assume without loss of generality that w is already in the first canonical form:

$$w = (t_{c_n} t_{c_n - 1} \ldots t_{d_n})(t_{c_{n-1}} t_{c_{n-1}-1} \ldots t_{d_{n-1}}) \ldots (t_{c_1} t_{c_1 - 1} \ldots t_{d_1})$$

with $c_k \geq d_k$ for all k, and $d_n < d_{n-1} < \ldots < d_1$.

Now w_H is going to be the largest right substring of w that doesn't touch the factors of G. More precisely, Let j be the greatest index for which $d_j > k$, and let

$$w_H := (t_{c_j} t_{c_j - 1} \ldots t_{d_j}) \ldots (t_{c_1} t_{c_1 - 1} \ldots t_{d_1})$$

Then

$$w = (t_{c_n} t_{c_n - 1} \ldots t_{d_n})(t_{c_{n-1}} t_{c_{n-1}-1} \ldots t_{d_{n-1}}) \ldots (t_{c_{j+1}} t_{c_{j+1}-1} \ldots t_{d_{j+1}}) w_H := w' w_H$$

Rewriting w' (what's left of w) as $v w_G$ requires actual reordering for two reasons corresponding to the two new requirements: it is possible that some d is too small:

$d_{j+1} < k$ or $d_{i+1} < d_i - 1$ for some i; or that $c_{i+1} \geq c_i$. As we did in the proof of the first canonical form, we start unwrapping w' from the right, maintaining the following inductive hypotheses:

- $w' \approx v_{bad} v_{good} u_G$,
- u_G only permutes the factors of G,
- $v_{good} = (t_{a_l} t_{a_l - 1} \dots t_{k+1-l}) \dots (t_{a_2} t_{a_2 - 1} \dots t_{k-1})(t_{a_1} t_{a_1 - 1} \dots t_k)$ satisfies the requirements for v, i.e. has $a_1 > a_2 > \dots a_l$,
- v_{bad} is in the first canonical form in chunks ($t_{c_i} \dots t_{d_i}$ with all $d_i < k+1-l$ (l is from the previous item on this list).

We initiate the induction by collecting as much as possible in u_G, i.e. setting u_G to be the maximal right substring of w' which only uses t_1 through t_{k-1}. Let w'' be what's left, i.e. such that $w' = w'' u_G$. Since the indices in the ith chunk $(t_{c_i} t_{c_i - 1} \dots t_{d_i})$ of w' begin with $d_i \leq k$ and increase rightwards, the first from the right index $\geq k$ is k. So the rightmost chunk of w'' indeed has k as its lowest index, so we can set v_{good} to be that chunk. We set v_{bad} to be what's left, still in the first canonical form, with every chunk beginning with $d_i < k$.

The induction step will shorten v_{bad} by one chunk. So we need to find a word equivalent to

$$q := (t_c \dots t_d)(t_{a_l} t_{a_l - 1} \dots t_{k+1-l}) \dots (t_{a_2} t_{a_2 - 1} \dots t_{k-1})(t_{a_1} t_{a_1 - 1} \dots t_k)$$

that looks like a $\widetilde{v_{good}} \widetilde{u_G}$

We first deal with the possibility that d is too low, namely that $d < k - l$. If $c < k - l$ also, then the whole left chunk $(t_c \dots t_d)$ commutes with v_{good}, so if we let $\widetilde{v_{good}} := v_{good}$ and $\widetilde{u_G} := (t_c \dots t_d)$, we're done. Otherwise, $(t_c \dots t_d) = (t_c \dots t_{k-l})(t_{k-l-1} \dots t_d)$, the right half of which commutes with v_{good}, so now, with $\widetilde{u_G} := (t_{k-l-1} \dots t_d)$,

$$q \approx (t_c \dots t_{k-l})(t_{a_l} t_{a_l - 1} \dots t_{k+1-l}) \dots (t_{a_2} t_{a_2 - 1} \dots t_{k-1})(t_{a_1} t_{a_1 - 1} \dots t_k) \widetilde{u_G}$$

Now the worry is that $c \geq a_l$. Since $a_l \geq k + 1 - l > k - l$,

$$t_c \dots t_{k-l} = (t_c \dots t_{a_l + 1})(t_{a_l} \dots t_{k-l})$$

Computations just like in the proof of the first canonical form give

$$(t_{a_l} \dots t_{k-l})(t_{a_l} t_{a_l - 1} \dots t_{k+1-l}) \approx (t_{a_l - 1} \dots t_{k-l})(t_{a_l} \dots t_{k+1-l} t_{k-l})$$

Now the rightmost t_{k-l} commutes with the rest of v_{good}, and $(t_c \dots t_{a_l + 1})$ commutes with $(t_{a_l - 1} \dots t_{k-l})$, so we get that

$$q \approx (t_{a_l - 1} \dots t_{k-l})(t_c \dots t_{a_l + 1})(t_{a_l} \dots t_{k+1-l})(t_{a_{l-1}} t_{a_{l-1} - 1} \dots t_{k+1-(l-1)}) \dots$$

$$\dots (t_{a_2} t_{a_2 - 1} \dots t_{k-1})(t_{a_1} t_{a_1 - 1} \dots t_k) t_{k-l}$$

The rightmost t_{k-l} can be added into u_G. As long as $c \geq a_i$, we repeat this procedure, moving $(t_c \dots t_{a_i + 1})$ past the ith chunk of v_{good}, until he's finally in his rightful place.

We are now done with the induction step, and hence with the proof.

□

Combining the two lemmas proves Proposition 3.2.8.

3.6. **Using second canonical form to show that almost everything comes form skew-twists.** Finally, we put our technical tools to use. Though you may have forgotten over the past twenty pages (assuming we told you in the first place), our quest is to characterize (f, g)-skew-invariant curves for trivial polynomials f and g, i.e. non-linear f and g that are not skew-conjugate to monomials or Chebyshevs. Remember (if we ever told you), for linear L, two polynomials f and $g := L^\sigma \circ f \circ L^{-1}$ are called *skew-conjugate*. Most of these skew-invariant curves come from skew-twists, described in the next section. Here, we mop up the few curves that do not. Actually, we will say nothing about invariant curves until much later; for now, we characterize certain commutative diagrams of polynomials with coefficients in a difference field. More precisely, we characterize triples of polynomials (f, g, π) satisfying $\pi^\sigma \circ f = g \circ \pi$, where f and π share no initial compositional factors, and π^σ and g share no terminal compositional factors.

The main result of this section, Theorem 3.3 follows immediately from Lemma 2.8 of [13]. However, the new tool we develop to prove this theorem, chebyclumps, will be used extensively in the next section, and is not present in [13].

Theorem 3.3. *If two trivial polynomials f and g satisfy $g \circ \pi = \pi^\sigma \circ f$, and f and π share no initial compositional factors, and π^σ and g share no terminal compositional factors, then there are linear L and M such that $M \circ \pi \circ L$ is a monomial and both $L^\sigma \circ f \circ L^{-1}$ and $(M^\sigma)^{-1} \circ g \circ M$ have decompositions where all factors are ritty and the degree of π divides the out-degree of all factors of some decomposition of f and the in-degree of all factors of some decomposition of g.*

One can remove much of our terminology from this theorem, in particular removing all references to decompositions. Then the theorem would read
"If two polynomials f and g satisfy $g \circ \pi = \pi^\sigma \circ f$, and f and π share no initial compositional factors, and π^σ and g share no terminal compositional factors, then there are linear L and M such that either $L^\sigma \circ f \circ L^{-1}$ and $(M^\sigma)^{-1} \circ g \circ M$ are both monomials or Chebyshev polynomials (and then we say nothing about π); or $M \circ \pi \circ L(x) = x^n$ is a monomial, $L^\sigma \circ f \circ L^{-1}(x) = x^k \cdot u(x^n)$, and $(M^\sigma)^{-1} \circ g \circ M(x) = x^k \cdot u(x)^n$ for some polynomial u."

We break down the proof into three propositions: a translation into the language of decompositions and canonical forms, a proof of the theorem in case none of these polynomials have any type C factors, and a proof of the theorem in case one of the polynomials does have a type C factor.

Proposition 3.3.1. *(Translating the theorem)*
Suppose that polynomials f, g, and π satisfy $g \circ \pi = \pi^\sigma \circ f$, and that f and π share no initial compositional factors, and π^σ and g share no terminal compositional factors. Let m be the number of factors in (any) decomposition of π, and let l be the number of factors in (any) decomposition of f (or g). Then there are decompositions $\vec{\pi}$ of π, $\vec{f}$ is f, $\vec{g}$ of g, and $\vec{\rho}$ of π^σ (which $\vec{\rho}$ need not be $(\vec{\pi})^\sigma$) such that

$$(t_l \dots t_1) \dots (t_{l+m-2} \dots t_{m-1})(t_{l+m-1} \dots t_m) \star \vec{g}\vec{\pi} = \vec{\rho}\vec{f}$$

Proof. Let $(\pi_m, \dots, \pi_1)$ be a decomposition of π, and $(g_l, \dots, g_1)$ be a decomposition of g. Let $w = vw_1w_2$ be the word in the second canonical form that yields a decomposition of f followed by a decomposition of π^σ. Since we were free to choose the decompositions of π and g, we may assume, losing this freedom, that w_i are

empty. So we get decompositions as above and

$$v = (t_{a_k} t_{a_k-1} \dots t_{b_k})(t_{a_{k-1}} t_{a_{k-1}-1} \dots t_{b_{k-1}}) \dots (t_{a_1} t_{a_1-1} \dots t_{b_1})$$

with $a_i \geq b_i - 1$ for all i ($a_i = b_i - 1$ means that the word $(t_{a_i}, \dots t_{b_i})$ is empty); $b_i = \text{length}(\vec{\pi}) + 1 - i$, and $a_k < \dots < a_2 < a_1$; and

$$v \star \vec{g}\vec{\pi} = \vec{\rho}\vec{f}$$

Now it follows immediately that $k = \text{length}(\vec{\pi})$, for otherwise t_1 does not occur in v, so the rightmost factor π_1 in $\vec{g}\vec{\pi}$ is untouched by the action of v, so it is a shared initial factor of π and f, contradicting a hypothesis of the proposition.

For exactly the same reasons, unless $a_i = \text{length}(\vec{g}) + \text{length}(\vec{\pi}) - i$ for all i, ρ and g will share a terminal factor, which is also not supposed to happen.

So $v = (t_l \dots t_1) \dots (t_{l+m-2} \dots t_{m-1})(t_{l+m-1} \dots t_m)$ as wanted. □

We have $v := v_1 v_2 \dots v_m$ where $v_i := t_{l+1-i} t_{l+2-i} \dots t_i$, and we have $v \star \vec{g}\vec{\pi}$ defined. In the next proposition where none of the g_i or π_i are type C, we will make repeated use of Proposition 3.2.6. We will then introduce chebyclumps and make a series of observations about them in order to handle the other case.

Proposition 3.3.2. *(the Descaling case)*
Theorem 3.3 holds if g and π have no type C factors.

Proof. We may without loss of generality replace $\vec{g}$ and $\vec{\pi}$ by linearly equivalent decompositions that admit descalings with only g_l and π_1 non-monic, so that the concatenation of the two descalings is a descaling of $\vec{g}\vec{\pi}$. Since none of the g_i or π_i are type C, Lemma 3.2.9 and Proposition 3.2.6 apply to the action of each v_i, and an easy induction on i shows that there are linear L and M and ritty h_i, o_i such that $g_l = M \circ h_l$, $g_i = h_i$ for all other i, $\pi_1 = o_1 \circ L$, and $\pi_i = o_i$ for all other i.

So

$$g \circ \pi = M \circ h_l \circ h_{l-1} \circ \dots \circ h_1 \circ o_k \dots \circ o_2 \circ o_1 \circ L$$

and

$$\pi^\sigma \circ f = M \circ \tilde{o}_k \dots \circ \tilde{o}_2 \circ \tilde{o}_1 \circ \tilde{h}_l \circ \tilde{h}_{l-1} \circ \dots \circ \tilde{h}_1 \circ L$$

where the ritty polynomials with tildes are the result of the Ritt swaps where the linear factors witnessing the Ritt swaps are all identity.

By assumption, none of the factors are type C. None of the o are type J, since those cannot swap in the necessary direction. So all non-monomial o_i have well-defined in- and out-degrees; for each i, the in-degree of $\tilde{o}_i$ is higher than the in-degree of the corresponding o_i (by a factor of $\deg g$), contradicting the fact that $\tilde{o}_i$ are the factors of π^σ and o_i are the factors of π. Therefore, $\tilde{o}_i = o_i$ are monomials for all i. For some linear N, $N \circ o_k \dots \circ o_2 \circ o_1 \circ L$ is a decomposition of π, so $L_1 := L^{-1}$ and $M_1 := N^{-1}$ are the unique translations such that $M_1 \circ \pi \circ L_1$ is a monomial. They witness the conclusion of the theorem. □

To treat the last possibility, that at least one factor of g or π is type C, we must turn aside and contemplate how type C factors play with each other. These observations will be used again heavily in the next section.

3.6.1. *Chebyclumps.* We show that clumps of compatibly-scaled type C factors persist (up to invading quadratic factors) under Ritt swaps, and that Ritt swaps involving two odd Chebyshevs can only occur within these clumps.

Definition 3.3.1. Let $\vec{f}$ be a decomposition of a polynomial f.

If $L \circ f_j \circ \ldots \circ f_i \circ M = C_n$ for some integer n that is not a power of 2, and some linear L and M, we call $(f_j, \ldots f_i)$ a *chebyclump* of the decomposition.

If in addition neither $(f_{j+1}, f_j \ldots f_i)$ nor $(f_j, \ldots f_i, f_{i-1})$ is a chebyclump, we call $(f_j, \ldots f_i)$ a *maximal chebyclump* of the decomposition.

A chebyclump is called *odd* if n is odd.

Note that this definition is invariant under linear equivalence. Let us see how this notion interacts with Ritt swaps.

Observation 3.3.1.

(1) If one of the f_1 and f_2 is type C, then (f_2, f_1) is swappable if and only if the pair is a chebyclump.
(2) Every decomposition of C_n is linearly equivalent to a decomposition all of whose factors are C_p for prime p, including $p = 2$.
(the property "all factors are C_p" is invariant under Ritt swaps.)
(3) If $(f_j, \ldots, f_i)$ is a chebyclump so that $L \circ f_j \circ \ldots \circ f_i \circ M = C_n$, let $g_i := C_{\deg(f_i)}$; then $(L \circ f_j, f_{j-1}, \ldots, f_{i+1}, f_i \circ M)$ is linearly equivalent to $\vec{g}$.
(immediate from previous)
(4) The converse to the previous item is obvious.
(5) For every type C indecomposable c, there are unique up to ± 1 linear T and S such that $T \circ c \circ S = C_p$. For C_2, there is at most one such S given such a T, and vice versa: $(+2A^2 - 2) \circ (\cdot A^2) \circ C_2 \circ (\cdot A) = C_2$ for each non-zero A.
(6) L and M in the definition of chebyclump are unique up to multiplication by ± 1.
(Combine (2) and (4) and the fact that at least one factor in a chebyclump has odd degree.)

Lemma 3.3.1. *If $(f_j, \ldots f_{i+1})$ and $(f_i, \ldots, f_k)$ are both chebyclumps, witnessed by L, M for the first and by $\hat{L}, \hat{M}$ for the second, then the concatenation $(f_j, \ldots f_k)$ is a chebyclump if and only if $\hat{L} \circ M = (\cdot \pm 1)$.*

Proof. The back directions is immediate; we prove the forward direction. In other words, we have polynomials f and g and linear A, B, C, D, L, and M, such that $B^{-1} \circ f \circ A^{-1} = C_m$ and $D^{-1} \circ g \circ C^{-1} = C_n$ and $L \circ g \circ f \circ M = C_{mn}$.

So $f = B \circ C_m \circ A$ and $g = D \circ C_n \circ C$.

Since chebyclumps must have at least one factor of odd degree, and factors can move freely within the chebyclump, we choose decompositions $(B \circ C_p, f_{k-1}, \ldots f_1)$ of f and $(g_l, \ldots, g_2, C_q \circ C)$ of g.

Then $(g_l, \ldots, g_2, C_q \circ C, B \circ C_p, f_{k-1}, \ldots f_1)$ is a decomposition of $g \circ f = L^{-1} \circ C_{mn} \circ M^{-1}$. According to observation 3 above, it must be linearly equivalent to the decomposition $(L^{-1} \circ C_{\deg(g_l)}, \ldots, C_{\deg(g_2)}, C_q, C_p, C_{\deg(f_{k-1})} \ldots C_{\deg(f_1)} \circ M^{-1})$. But type C ritty polynomials are not non-trivially linearly related to themselves except for $(-1) * C_p = C_p$, so $C \circ B = (\cdot \pm 1)$. □

Corollary 3.3.1. *The unique L and M are invariant under Ritt swaps within the chebyclump.*

Proof. Without loss of generality, all of $L \circ f_j, f_{j-1}, \ldots, f_{i+1}, f_i \circ M$ are actually Chebyshevs, in which case the assertion is verified immediately by writing out the Ritt swap. □

Lemma 3.3.2. *If f_b has odd degree, then $(f_c, \ldots, f_b, \ldots, f_a)$ is a chebyclump if and only if both $(f_c, \ldots, f_b)$ and $(f_b, \ldots, f_a)$ are chebyclumps.*

Proof. The forward direction is obvious, so we assume that both $(f_c, \ldots, f_b)$ and $(f_b, \ldots, f_a)$ are chebyclumps and prove that $(f_c, \ldots, f_b, \ldots, f_a)$ is one.

Without loss of generality, we may assume that, for some Chebyshev polynomials g_c and g_b and linear A and B, $f_c = A \circ g_c$, $f_b = g_b \circ B$, and that f_i are themselves Chebyshev for $c > i > b$.

On the other hand, $(f_b, \ldots f_a)$ is linearly equivalent to $(C \circ g_b, g_{b-1}, \ldots g_{a+1}, g_a \circ D$ for linear C and D and Chebyshev g_i. The occurrence of g_b here and in the previous paragraph is not an accident: it is indeed the same Chebyshev of the same degree as f_b.

Let $L_b, \ldots L_{a+1}$ witness this:

$$f_b \circ L_b = C \circ g_b,\ L_b^{-1} \circ f_{b-1} \circ L_{b-1} = g_{b-1},\ \ldots\ , L_{a+2} \circ f_{a+1} \circ L_{a+1} = g_{a+1},\ L_{a+1}^{-1} \circ f_a = g_a \circ D$$

Since $f_b = g_b \circ B$, the first of the above gives $g_b \circ B \circ L_b = f_b \circ L_b = C \circ g_b$. But g_b is a Chebyshev polynomial of odd degree, only linearly related to himself by $(-1) * g_b = g_b$, so $B \circ L_b = C = (\cdot \pm 1)$.

Now the same linear factors L_i inserted in the same places witness that $(f_c, \ldots, f_b, \ldots, f_a)$ is linearly equivalent to $(f_c, \ldots, f_{b+1}, f_b \circ L_b, L_b^{-1} \circ f_{b-1} \circ L_{b-1}, \ldots, L_{a+1}^{-1} \circ f_a) = (A \circ g_c, g_{c-1}, \ldots, g_{b+1}, g_b \circ (\cdot \pm 1), g_{b-1}, \ldots g_{a+1}, g_a \circ D$

The $(\cdot \pm 1)$ linear factor can be pulled left out of the middle of a chebyclump: $C_p \circ (\cdot - 1) = (\cdot - 1) \circ C_p$ for odd p, and $C_2 \circ (\cdot - 1) = C_2$, so this is indeed a chebyclump as wanted. □

The purpose of all those technical bits is:

Lemma 3.3.3. *Suppose $(f_j, \ldots f_i)$ is a maximal chebyclump of the decomposition $\vec{f}$ of f. Let $\vec{g}$ be another decomposition of f obtained from $\vec{f}$ by a single Ritt swap. Then $(g_r, \ldots g_s)$ is a maximal chebyclump of $\vec{g}$ for some $r \in \{j-1, j, j+1\}$ and some $s \in \{i-1, i, i+1\}$. Further, the odd parts of the degrees of the two chebyclumps are the same.*

Proof. First note that Ritt swaps within the chebyclump, and Ritt swaps among factor neither in nor adjacent to the chebyclump have no effect. Thus we have nothing to prove except for the Ritt swaps at $(j+1)$, j, $(i-1)$ and $(i-2)$. Note that in any case $(g_{j-1}, \ldots g_{i+1})$ is still a chebyclump. Thus a maximal chebyclump cannot lose more than one factor though a single Ritt swap; since Ritt swaps are invertible, this means that a maximal chebyclump also cannot gain more than one factor. It remains to show that the factor gained (or, symmetrically, lost) cannot be odd. Suppose towards contradiction that a Ritt swap at $(j+1)$ adds an odd factor to the chebyclump; that is, $t_{j+1} \star (b, a, f_j, \ldots f_i) = (d, c, f_j, \ldots f_i)$, where c is odd type C and $(c, f_j, \ldots f_i)$ is a chebyclump. Since (d, c) is swappable, by observation 1 above, (d, c) is a chebyclump. Now by the last lemma the whole $(d, c, f_j, \ldots f_i)$ is a chebyclump, giving the desired contradiction. The proof for the Ritt swap at $(i-2)$ is identical. □

The next (hard) technical lemma describes the very rare situations when two quadratic factors can get all the way across a decomposition.

Lemma 3.3.4. *Let $\vec{f}$ be a decomposition of length k of f with $\deg(f_2) = \deg(f_1) = 2$; let $v_1 := t_{k-1}t_{k-2}\dots t_2$ and $v_2 := t_{k-2}t_{k-3}\dots t_1$ and $v = v_2v_1$; suppose that $v \star \vec{f}$ is defined. Then either $\vec{f}$ is a chebyclump, or $\vec{f}$ is linearly equivalent to $(A_k \circ \gamma_k, \gamma_{k-1} \dots, \gamma_3, Q, Q \circ L)$ for some linear L and A_k, where each γ_i is a monomial or a ritty polynomial whose out-degree a multiple of 4.*

Proof. Up to linear equivalence, we may take $\vec{f} := (f_k, \dots f_3, Q, M \circ Q \circ L)$ for some linear L and translation M.

Since $t_2 \star \vec{f}$ is defined, there is a linear A_3, an integer r_3, and a monic polynomial u_3 such that either $A_3^{-1} \circ f_3 = x^{r_3} \cdot u_3(x)^2$ or $A_3^{-1} \circ f_3$ is a monomial. Let $\alpha_3 := x^{r_3} \cdot u_3(x^2)$ in the first case, and the same monomial in the second case. Then $t_2 \star \vec{f} = (f_k, \dots f_4, A_3 \circ Q, \alpha_3, M \circ Q \circ L)$.

Since $t_3t_2 \star \vec{f}$ is defined, there is a linear A_4, an integer r_4, and a monic polynomial u_4 such that either $A_4^{-1} \circ f_4 \circ A_3 = x^{r_4} \cdot u_4(x)^2$ or $A_4^{-1} \circ f_3$ is a monomial. Let $\alpha_4 := x^{r_4} \cdot u_4(x^2)$ in the first case, and the same monomial in the second case. Then $t_3t_2 \star \vec{f} = (f_k, \dots f_5, A_4 \circ Q, \alpha_4, \alpha_3, M \circ Q \circ L)$.

Inducting, since $t_i \dots t_3t_2 \star \vec{f}$ is defined, there is a linear A_{i+1}, an integer r_{i+1}, and a monic polynomial u_{i+1} such that either $A_{i+1}^{-1} \circ f_{i+1} \circ A_i = x^{r_{i+1}} \cdot u_{i+1}(x)^2$ or it is a monomial. Let $\alpha_{i+1} := x^{r_{i+1}} \cdot u_{i+1}(x^2)$ in the first case, and the same monomial in the second case. Then $t_i \dots t_3t_2 \star \vec{f} = (f_k, \dots f_{i+2}, A_{i+1} \circ Q, \alpha_i \dots \alpha_3, M \circ Q \circ L)$.

And at the end we get

$$v_1 \star \vec{f} = (A_k \circ Q, \alpha_k, \dots \alpha_3, M \circ Q \circ L)$$

In the rest of the proof we will name many linear factors witnessing Ritt swaps between *ritty* indecomposables. In each of these Ritt swaps one of the indecomposables is Q, so none of these Ritt swaps will involve commuting odd Chebyshevs, so we may and will always choose these witnesses to be translations (see Remark 3.1.1).

Now since $t_1v_1 \star \vec{f}$ is defined, $(\alpha_3, M \circ Q \circ L)$ is swappable. Thus, there must be some translation N, integer s, and polynomial v such that $N \circ \alpha_3 \circ M = x^s \cdot v(x)^2$ or $N \circ \alpha_3 \circ M$ is a monomial. There are two distinct possibilities: either $M = N = \mathrm{id}$, or α_3 and $N \circ \alpha_3 \circ M$ are two non-trivially translation-related ritty polynomials. The rest of the proof goes differntly in the two cases, yeilding the two different conclusions of the lemma.

Case 1: Suppose that $M = N = \mathrm{id}$. Then either α_3 is a monomial, or $\alpha_3 = x^{s_3} \cdot v_3(x)^2$; let $\beta_3 := \alpha_3$ if it is a monomial, and $\beta_3 := x^{s_3} \cdot v_3(x^2)$ otherwise. Then $t_1v_1 \star \vec{f}$ is linearly equivalent to $(A_k \circ Q, \alpha_k, \dots \alpha_4, Q, \beta_3, L)$.

Now since $t_2t_1v_1 \star \vec{f}$ is defined, (α_4, Q) is swappable. Since there are no solutions to translation $\circ$ ritty = ritty, this means that either α_4 is a monomial, or $\alpha_4 = x^{s_4} \cdot v_4(x)^2$; let $\beta_4 := \alpha_4$ if it is a monomial, and $\beta_4 := x^{s_4} \cdot v_4(x^2)$ otherwise. Then $t_2t_1v_1 \star \vec{f}$ is linearly equivalent to $(A_k \circ Q, \alpha_k, \dots \alpha_5, Q, \beta_4, \beta_3, L)$.

Inducting, for each $4 \le i \le k$, since $t_{i-1} \dots t_2t_1v_1 \star \vec{f}$ is defined, (α_i, Q) is swappable. Since there are no solutions to translation $\circ$ ritty = ritty, this means that either α_i is a monomial, or $\alpha_i = x^{s_i} \cdot v_i(x)^2$; let $\beta_i := \alpha_i$ if it is a monomial, and $\beta_i := x^{s_i} \cdot v_i(x^2)$ otherwise. Then for $i < k$ we get $t_{i-1} \dots t_2t_1v_1 \star \vec{f}$ is linearly

equivalent to $(A_k \circ Q, \alpha_k, \dots \alpha_{i+1}, Q, \beta_i, \dots \beta_3, L)$, and finally

$$v_2 v_1 \star \vec{f} = (A_k \circ Q, Q, \beta_k, \dots \beta_3, L)$$

Note that in this case, for each i we have that either $\beta_i = \alpha_i$ is a monomial, or $\alpha_i = x^{s_i} \cdot v_i(x)^2$ and $\beta_i = x^{s_i} \cdot v_i(x^2)$. Remember from the first half of the proof that $\alpha_i = A_i^{-1} \circ f_i \circ A_{i-1}$ if it is a monomial, and that otherwise $\alpha_i = x^{r_i} \cdot u_i(x^2)$ and $A_i^{-1} \circ f_i \circ A_{i-1} = x^{r_i} \cdot u_i(x)^2$. So either $\beta_i = \alpha_i = A_i^{-1} \circ f_i \circ A_{i-1}$ is a monomial; or $\alpha_i = x^{s_i} \cdot v_i(x)^2 = x^{r_i} \cdot u_i(x^2)$, i.e. $r_i = s_i$ and there is some polynomial w_i such that $\alpha_i = x^{r_i} \cdot w_i(x^2)^2$ and $A_i^{-1} \circ f_i \circ A_{i-1} = x^{r_i} \cdot w_i(x)^4$. Let $\gamma_i = \alpha_i$ if it is a monomial, and otherwise let $\gamma_i := x^{r_i} \cdot w_i(x)^4 = A_i^{-1} \circ f_i \circ A_{i-1}$. Now A_{k-1} through A_3 witness the conclusion of the lemma.

Case 2: The other possibility is that $N \circ \alpha_3 \circ M = x^s \cdot v(x)^2$ with at least one of N and M not identity. Then α_3 is not a monomial, so we see form the first part of the proof that its in-degree is at least 2. Thus, α_3 is not type J, so it must be type C, and in fact $\lambda_3 * C_{p_3}$ for some λ_3. We now wish to insert scalings into the decomposition so as to maintain the monicity of all α_i and turn α_3 into C_{p_3}. For $i \geq 3$, let $\lambda_{i+1} = \lambda_i^{\deg(\alpha_i)}$ so that $(\cdot \frac{1}{\lambda_{i+1}}) \circ \alpha_i \circ (\cdot \lambda_i) = \lambda_i * \alpha_i$ is monic. Let $\tilde{u}_i := (\lambda_i^2) * u_i$ so that $\lambda_i * \alpha_i = x^{r_i} \cdot \tilde{u}_i(x^2)$. Putting it all together, let

$$\vec{g} := (A_k \circ (\cdot \lambda_{k+1}^2) \circ Q, \lambda_k * \alpha_k, \dots, \lambda_4 * \alpha_4, C_{p_3}, M' \circ Q \circ L')$$

where M' is a translation such that $(\cdot \frac{1}{\lambda_3}) \circ M = M' \circ (\cdot \frac{1}{\lambda_3})$ and $L' := (\cdot \frac{1}{\sqrt{\lambda_3}}) \circ L$. Since $\vec{g}$ is linearly equivalent to our previous decomposition of $v_1 \star \vec{f}$, we may say that

$$v_1 \star \vec{f} = \vec{g}$$

And now we do the whole thing all over again, applying v_2 to $\vec{g}$ one swap at a time. But now we are applying these swaps to g_i, about whom we know a whole lot, rather than to f_i, about whom we knew nothing. This is why it takes two quadratic factors to get the conclusion of the lemma.

So, back to $t_1 \star \vec{g}$ being defined: for some N, s, v, $N^{-1} \circ C_{p_3} \circ M' = x^s \cdot v(x)^2$. But then we must have $N = M' = (-2)$ or $N = M' = (+2)$. In the first case, $t_i \star \vec{g} = (\dots, \lambda_4 \circ \alpha_4, (-2) \circ Q, C_{p_3} \circ L')$. In the second case, $t_i \star \vec{g} = (\dots, \lambda_4 * \alpha_4, (+2) \circ Q, i * C_{p_3} \circ L')$.

Now $t_2 t_1 \star \vec{g}$ is defined, so for some N, s, v, $N^{-1} \circ \lambda_4 * \alpha_4 \circ M' = x^s \cdot v(x)^2$. Whether $M' = (+2)$ or $M' = (-2)$, this still forces $\lambda_4 * \alpha_4 = C_{p_4}$ and $N = M'$. So the two options for $t_2 t_1 \star \vec{g}$ are $(\dots, \lambda_5 * \alpha_5, (+2) \circ Q, i * C_{p_4}, i * C_{p_3} \circ L')$ and $(\dots, \lambda_5 \circ \alpha_5, (-2) \circ Q, C_{p_4}, C_{p_3} \circ L')$.

Induct as before to obtain $v_2 * \vec{g} =$

$$(A \circ (\cdot \lambda_{k+1}^2) \circ Q, (-2) \circ Q, C_{p_k}, \dots C_{p_4}, C_{p_3} \circ L')$$

or

$$(A \circ (\cdot \lambda_{k+1}^2) \circ Q, (+2) \circ Q, i * C_{p_k}, \dots i * C_{p_4}, i * C_{p_3} \circ L')$$

In either case the whole decomposition is a chebyclump. □

Let us now return to the question at hand, namely to decompositions $\vec{g}$, $\vec{\pi}$ at least one of which has at least one type C factor, and to the word $v := (t_l \dots t_1) \dots (t_{l+m-2} \dots t_{m-1})(t_{l+m-1} \dots t_m)$ such that $v \star \vec{g}\vec{\pi}$ is defined.

Proposition 3.3.3. *(type C lemma)*
Theorem 3.3 holds when at least one of g and π has a type C factor.

Proof. Note that if one of g and π has a type C factor, all factors of the other one must be either type C or degree 2. Since each factor of g must at some point swap with each factor of π, they cannot both have quadratic factors. This leaves the following options:

(1) Both g and π have a type C factor, and one of g and π is linearly related to an odd-degree (decomposable) Chebyshev polynomial.
(2) g has a type C factor and no quadratic factors, while all factors of π are quadratic.
(3) π has a type C factor and no quadratic factors, while all factors of g are quadratic.

In the first case, the whole $\vec{g}\vec{\pi}$ must be one giant chebyclump, which forces the unique linear L and M such that $M \circ \pi \circ L = C_n$ for $n = \deg(\pi)$ to also make $L^\sigma \circ f \circ L^{-1}$ a Chebyshev polynomial, contradicting its triviality.

In the second case, if π is itself quadratic, then $M \circ \pi \circ L = Q$ for some linear M and L and the theorem is clearly true. Otherwise, the previous lemma applies to $\vec{g}\vec{\pi}$. If π is linearly related to x^4, the conclusion of the theorem follows. If π is not linearly related to P_4, then chasing the linear factors in the lemma above, one sees that g must be skew-conjugate to a Chebyshev polynomial.

For the third case, we only skecth the proof. First, we break π into maximal chebyclumps: $\pi := M^{-1}C_{n_a} \circ T_a \circ C_{n_{a-1}} \circ \ldots \circ C_{n_2} \circ T_2 \circ C_{n_1} \circ L^{-1}$ for linear M, L, and non-trivial linear T_i. If $\deg g > 2$, the lemma above applied to $\vec{\pi}^\sigma \vec{f}$ implies that π must be a single chebyclump, and chasing the translations in the proof of the lemma shows that g is skew-conjugate to $C_{\deg(g)}$. If g is quadratic, and m is the number of indecomposable factors of π, then $t_1 t_2 \ldots t_m \star \vec{g}\vec{\pi} = \vec{\rho}\vec{f}$ and $\vec{\rho}$ will break into maximal chebyclumps the same way that $\vec{\pi}$ did, but where T_a had to be scalings in order for $t_1 t_2 \ldots t_m \star \vec{g}\vec{\pi}$ to be defined, the corresponding linear factors in ρ will not be scalings, contradicting the fact that ρ is a decomposition of π^σ. □

3.7. **skew-twist monoid.** In trying to make this section more readable, we have tried to motives the rather technical results as we go along. Unfortunately, the motivation often involves concepts not defined until Section 4. The worst offenders are *trivial polynomials*, which are simply polynomials that are not skew-conjugate to any monomial or Chebyshev polynomial; and *correspondences between σ-varieties $(\mathbb{A}^1, f)$ and $(\mathbb{A}^1, g)$* which are simply (f, g)-skew-invariant curves.

3.7.1. *skew-linear equivalence.*

Definition 3.3.2. Remember, for linear L, two polynomials f and $g := L^\sigma \circ f \circ L^{-1}$ are called *skew-conjugate*.

In this case, $L : (\mathbb{A}^1, f) \to (\mathbb{A}^1, g)$ is an isomorphism of σ-varieties and, when f, g, adn L are defined over the fixed field of σ, of the corresponding dynamical systems. For fixed f and g, L also gives rise to a bijection between the set of decompositions of f and the set of decompositions of g: if $(f_k, f_{k-1}, \ldots, f_2, f_1)$ is a decomposition of f, then $(L^\sigma \circ f_k, f_{k-1}, \ldots, f_2, f_1 \circ L^{-1})$ is a decomposition of g.

Lemma 3.3.5. *Given a linear polynomial L and a decomposition $\vec{f}$ of a polynomial f, let $\vec{g} := (L^\sigma \circ f_k, f_{k-1}, \ldots, f_2, f_1 \circ L^{-1})$, a decomposition of $g := L^\sigma \circ f \circ L^{-1}$. Thus for fixed f and g, L gives rise to a bijection between decompositions of f and decompositions of g, which respects linear equivalence.*

Definition 3.3.3. Two decompositions $\vec{f}$ and $\vec{h}$ are *skew-linearly-equivalent* if there is a linear L such that $\vec{h}$ is linearly equivalent to $(L^\sigma \circ f_k, f_{k-1}, \dots, f_2, f_1 \circ L^{-1})$.

Skew-linearly-equivalent decompositions may be decompositions of different, but always skew-conjugate, polynomials. It is immediate that

Lemma 3.3.6. *Skew-linear-equivalnce is an equivalence relation.*

3.7.2. *skew-twists.*

Definition 3.3.4. The decomposition $(f_1^\sigma, f_k, \dots, f_2)$ is called *the single-skew-twist* of the decomposition $\vec{f} := (f_k, \dots, f_2, f_1)$ and denoted $\phi \star \vec{f}$. (ϕ stands for "forward".)
If $\vec{f}$ is a decomposition of a polynomial f, then $\phi \star \vec{f}$ is a decomposition of a (probably different) polynomial h; we call h *a single-skew-twist* of f.
To undo what ϕ does, we define $\beta \star \vec{f} := (f_{k-1}, \dots, f_1, f_k^{(\sigma^{-1})})$. ($\beta$ for "back".)

Note that while a decomposition has a unique single-skew-twist, a polynomial may have several single-skew-twists, coming from different decompositions. In particular, single-skew-twists of linearly-equivalent decompositions will be discussed shortly.

If h is a single skew-twist of f, then $h \circ f_1 = f_1^\sigma \circ f$, so the graph of f_1 is an (f, h)-skew-invariant subvariety of $\mathbb{A}^2$; and f_1 is a morphism of σ-varieties from the one defined by f to the one defined by h. We have shown in Theorem 3.3 that under most circumstances, all skew-invariant curves come from *composing* many such morphisms, possibly in *different directions*. This suggests the following definition:

Definition 3.3.5. For polynomials f and g, the relation "f is a *skew-twist* of g" is the symmetric-transitive closure of the relation "f is a single-skew-twist of g".

3.7.3. *monoid.* Similar to the monoid M_k of Ritt swaps acting on linear-equivalence classes of decompositions, we now define a monoid of B_k of Ritt swaps and single skew-twists, acting on skew-linear-equivalence classes of decompositions. While the action of M_k always produced a new decomposition of the same polynomial, the action of B_k will produce decompositions of skew-twists of the original polynomial.

We start with an analog of the first crucial Lemma 3.1.3:

Lemma 3.3.7. *Ritt swaps are well-defined up to skew-linear-equivalence.*
Single skew-twists are well-defined up to skew-linear-equivalence.

Proof. In Lemma 3.1.3, we proved that Ritt swaps are well-defined up to linear equivalence; so for the first statement, we only need to prove that if a decomposition $(h_k, h_{k-1}, \dots, h_2, h_1)$ of f is obtained from $\vec{f}$ by a Ritt swap at i, then the decomposition $(L^\sigma \circ h_k, h_{k-1}, \dots, h_2, h_1 \circ L^{-1})$ of $g := L^\sigma \circ f \circ L^{-1}$ is obtained from $\vec{g} := (L^\sigma \circ f_k, f_{k-1}, \dots, f_2, f_1 \circ L^{-1})$ by a Ritt swap at i. This is immediate from the definition of Ritt swap, with the same linear factor witnesses.

For the second part of the Lemma, we need two things (g and $\vec{g}$ stay the same)

- The decomposition $(f_1^\sigma, f_k, \dots f_2)$ obtained from $\vec{f}$ by a plain skew-twist is linearly equivalent to the decomposition $((f_1 \circ L^{-1})^\sigma, L^\sigma \circ f_k, f_{k-1}, \dots f_2)$ obtained from $\vec{g}$ by a plain skew twist.
- If $\vec{a}$ is obtained from $\vec{c}$ by a plain skew-twist,
  and $\vec{b}$ is obtained from $\vec{d}$ by a plain skew-twist,

and $\vec{d}$ is linearly equivalent to $\vec{c}$,
then $\vec{a}$ is skew-linearly equivalent to $\vec{b}$.

The first is immediate, and so is the second once the assumtions are written out:

$$\vec{a} = (c_1^\sigma, c_k, \dots, c_2)$$
$$\vec{d} = (c_k L_k, L_k^{-1} c_{k-1} L_{k-1}, \dots L_3^{-1} c_2 L_2, L_2^{-1} c_1)$$
$$\vec{b} = ((L_2^{-1} c_1)^\sigma, c_k L_k, L_k^{-1} c_{k-1} L_{k-1}, \dots L_3^{-1} c_2 L_2)$$

Now let $L := L_2^{-1}$ and note that $\vec{b}$ is linearly equivalent to $(L^\sigma \circ a_k, a_{k-1}, \dots a_2, a_1 \circ L)$. □

Definition 3.3.6. Let B_k be the free monoid generated by $\{t_i : 1 \le i \le k-1\}$ and ϕ and β.
Let S_f be the set of skew-linear-equivalence classes of decompositions of skew-twists of f.

Lemma 3.3.8. *If $\vec{f}$ is a decomposition of f, then any decomposition of any skew-twist of f may be obtained by a finite sequence of the following operations:*

- *t_i: Ritt swap at i defined long ago;*
- *ϕ defined by $\phi \star (f_k, \dots, f_1) = (f_1^\sigma, f_k, \dots, f_2)$;*
- *β defined by $\beta \star \vec{f} := (f_{k-1}, \dots, f_1, f_k^{(\sigma^{-1})})$.*

This defines a action of B_k on $S_f \cup \{\infty\}$, transitive on S_f.

3.7.4. *correspondences encoded.* In the case of the monoid of Ritt swaps acting on the decompositions of a single polynomial, the only question was which decompositions could be obtained; how they could be obtained only mattered as a tool. Now things are different: a word w in B_k applied to a decomposition $\vec{f}$ not only gives another decomposition $\vec{g} := w \star \vec{f}$, but also encodes a correspondence between σ-varieties $(\mathbb{A}^1, f)$ and $(\mathbb{A}^1, g)$. We care which correspondences. Intuitively, this correspondence comes from a long commutative diagram (here, superscripts are merely names and not any kind of power)

$$\begin{array}{ccc}
\mathbb{P}^1 & \xrightarrow{f=f^0} & \mathbb{P}^1 \\
\downarrow & & \downarrow \\
\mathbb{P}^1 & \xrightarrow{f^1} & \mathbb{P}^1 \\
\downarrow & & \downarrow \\
\mathbb{P}^1 & \xrightarrow{f^2} & \mathbb{P}^1 \\
\downarrow & & \downarrow \\
\vdots & & \vdots \\
\downarrow & & \downarrow \\
\mathbb{P}^1 & \xrightarrow{g=f^n} & \mathbb{P}^1
\end{array} \tag{15}$$

where the horizontal arrows are the polynomials whose decompositions are obtained from $\vec{f}$ by longer and longer subwords of w, and the vertical arrows are either

identity, if the corresponding symbol in w is one of the t_i, or a morphism down for ϕ and up for β. More precisely,

Definition 3.3.7. Given a word $w := w_n \dots w_2 w_1 \in B_k$ and a decomposition $\vec{f}$ of a polynomial f, let $\vec{g} := w \star \vec{f}$ be a decomposition of a polynomial g, and for each $j \leq n$ let $\vec{f^j} := (w_j \dots w_1) \star \vec{f}$ be a decomposition of a polynomial f^j; so $f^0 = f$ and $f^n = g$.

Then *the correspondence encoded by w and $\vec{f}$* is a correpondence $\mathcal{A}_w$ between σ-varieties $(\mathbb{A}^1, f)$ and $(\mathbb{A}^1, g)$ defined by
$(a, b) \in \mathcal{A}_w$ if and only if there are $a_0 = a, a_1, a_2, \dots a_n = b$ such that for each j

- if $w_j = t_i$ for some i, then $a_{j+1} = a_j$;
- if $w_j = \phi$, then $a_{j+1} = f_1^j(a_j)$;
- if $w_j = \beta$, then $a_j = f^{j+1}(a_{j+1})$.

Remark 3.3.1. A careful reader will not permit us speak of *the* (f, g)-invariant correspondence encoded by $w \star \vec{f}$, when even $\vec{g} := w \star \vec{f}$ is only defined up to skew-linear equivalence, and so g is only defined up to skew-conjugacy! She would worry that these linear factors entering at every step of the inductive definition might make an intractable mess. These concerns are addressed by Lemma 3.3.7, which show that g and C_w are well-defined up to a *single* linear factor.

It is clear that $\mathcal{A}_w$ is an (f, g)-skew-invariant set: one simply pushes the witnesses forward by $a_j \mapsto f^j(a_j)$ and notes that every box in the diagram above commutes. On the other hand, $\mathcal{A}_w$ will usually be reducible, and its irreducible components may be skew-invariant, skew-periodic, or strictly skew-pre-periodic. We only care about the skew-invariant components, which suggests the following definition

Definition 3.3.8. Two words w and w' in B_k are *equivalent with respect to $\vec{f}$*, written $w \approx_{\vec{f}} w'$, if $w \star \vec{f} = w' \star \vec{f} =: \vec{g}$ (in particular, defined), and the two correpondences $\mathcal{A}_w$ and $\mathcal{A}_{w'}$ have the same skew-invariant irreducible components (i.e. for any invariant irreducible $\mathcal{D}$, we have $\mathcal{D} \subset \mathcal{A}_w$ if and only if $\mathcal{D} \subset \mathcal{A}_{w'}$), and for any irreducible component $\mathcal{E}$ of one but not the other, $(f \times g)(\mathcal{E})$ is skew-invariant.

The last bit of the definition is needed for $\approx$ to be preserved under concatenation. It is harmless, as the only source of equivalent words whose correspondences are not identical is the following remark.

Remark 3.3.2. In particular, the (f, f)-skew-invariant corrsepondence $C_{\beta\phi}$ encoded by $\beta\phi$ is defined by $f_1(a) = f_1(b)$. One of its irreducible components, the diagonal $a = b$, is clearly (f, f)-(skew)-periodic. However, the image of the whole $C_{\beta\phi}$ under $(f \times f)$ is just the diagonal, so all other irreducible components are strictly *pre*-skew-periodic.

We now show that ϕ and β commute up to $\approx$; that ϕ^k and β^k commute with everything up to $\approx$; and that $\approx$ is preserved under concatenation.

Lemma 3.3.9. (1) $\phi\beta \approx \mathrm{id} \approx \beta\phi$

(2) *Suppose $u_1 \approx_{\vec{f}} v_1$, and so let $\vec{g} := u_1 \star \vec{f} = v_1 \star \vec{f}$, and suppose $u_2 \approx_{\vec{g}} v_2$; then $u_2 u_1 \approx_{\vec{f}} v_2 v_1$.*

(3) *For any word w in M_k, $w\phi^k \approx \phi^k w$ and $w\beta^k \approx \beta^k w$.*

Later, we will also want:

- *For $i < k-1$, $t_i\phi \approx \phi t_{i+1}$*
- *For $i > 1$, $t_i\beta \approx \beta t_{i-1}$*

Proof. We indicate the ideas of the proof:

(1) One part is explained in the remark above, the other comes from the fact that "$\exists x\, f(x) = a$ and $f(x) = b$" is equivalent to $a = b$.
(2) Push the witnesses in the definition of $\mathcal{A}_w$ forward.
(3) With part (1), we only need to show that ϕ^k and β^k commute with Ritt swaps. This is so because $\phi^k \star \vec{f} = \vec{f}^{\sigma}$ and $\beta^k \star \vec{f} = \vec{f}^{(\sigma^{-1})}$.

For the last two parts, note that the same two factors Ritt-swap on the two sides of each equation. □

Lemma 3.3.10. *For all $w \in B_k$, there is some $u \in B_k$ that does not contain β nor ϕ^k as a substring, and such that $w \approx \phi^{mk}u$ or $w \approx \beta^{nk}u$.*

Proof. We may introduce extra $\beta^i\phi^i$ pairs into the word w. We introduce enough of them to obtain $w' \approx w$ so that β only occurs in multiples of k in w'. Then we pull all β^k to the left, and obtain $\beta^{Nk}w'' \approx w'$ where w'' contains no instances of β. Then we can also pull all ϕ^k to the left and obtain $\beta^{Nk}\phi^{Mk}u \approx \beta^{Nk}w''$ where u contains no instances of β, and no instances of ϕ^r for $r \geq k$. Then we cancel $\beta\phi$ pairs in the beginning. □

What geometry is hiding behind this bit of combinatorics? Naturally, $\mathcal{A}_w$ comes with a diagram

$$(\mathbb{A}^1, f) \leftrightarrow (\mathbb{A}^1, ?) \leftrightarrow \ldots \leftrightarrow (\mathbb{A}^1, g)$$

each arrow corrseponding to an occurrence of ϕ or β in w. What we just proved is that, for correspondences coming from skew-twists, we can instead look at irreducible components of the fiber produt of the diagram

$$(\mathbb{A}^1, f) \xrightarrow{F} (\mathbb{A}^1, g^{\sigma^N}) \xleftarrow{g^{\diamond N}} (\mathbb{A}^1, g)$$

where we know exactly what one of the morphisms is. In the next two sections we describe the usual situation where u in Lemma 3.3.10 can be taken to be of the form $\phi^i v$ where v only contains Ritt swaps, so once appropriate decompositions are chosen, we know both arrows in the above diagram. Afterwards, we will need more combinatorics to mop up vicious special cases.

3.8. **cracked.**

3.8.1. *definitions.*

Definition 3.3.9. Suppose that $h = a \circ b$ for non-linear a and b; let $g := b^{\sigma} \circ a$; we then call the triple (h, g, b) a *plain-skew-twist*.

Translating the definition of plain-skew-twists into the language of our monoid action, we have the following lemma.

Lemma 3.3.11. *The triple (h, g, b) a* plain-skew-twist *if and only if there are decompositions $\vec{h}$ of h and $\vec{g}$ of g and $i \leq k$ such that $\vec{g} = \phi^i \star \vec{h}$ and the graph of $b = f_i \circ \ldots \circ f_1$ is the correspondence from h to g encoded by ϕ^i.*

Definition 3.3.10. A pair (b, a) of non-linear polynomials is called a *crack* of f if $f = b \circ a$ and for any decomposition $\vec{f}$ of f, there is an integer m and a linear polynomial L such that

$$b = f_k \circ \dots f_{m+1} \circ L^{-1} \text{ and } a = L \circ f_m \circ \dots \circ f_1$$

We say that a polynomial h is *cracked at the edge* if for any decomposition $\vec{h}$ of h and for any i

$$((h_i^\sigma \circ \dots \circ h_1^\sigma), (h_k \circ h_{i+1}))$$

is a crack. Beware the universal quantifier on decompositions!

If h is cracked at the edge and g is a plain skew-twist of h, we say that g is *cracked*.

The purpose of this notion is a transitivity for plain skew-twists.

Proposition 3.3.4. *If h is cracked at the edge, and both (h, g, b) and (g, f, d) are plain skew-twists, then either $(h, f, d \circ b)$ is a plain skew-twist, or there is some π for which (h^σ, f, π) is a plain skew-twist, and $\pi \circ h = d \circ b$.*

Proof. Here's the diagram witnessing the assumptions of the proposition:

$$\begin{array}{ccc} \mathbb{P}^1 & \xrightarrow{h=a\circ b} & \mathbb{P}^1 \\ {\scriptstyle b}\downarrow & & {\scriptstyle b^\sigma}\downarrow \\ \mathbb{P}^1 & \xrightarrow{g=b^\sigma\circ a=c\circ d} & \mathbb{P}^1 \\ {\scriptstyle d}\downarrow & & {\scriptstyle d^\sigma}\downarrow \\ \mathbb{P}^1 & \xrightarrow{f=d^\sigma\circ c} & \mathbb{P}^1 \end{array} \tag{16}$$

Let $(h_n, \dots, h_1)$ be a decomposition of h obtained by concatenating decompositions of a and b, so that $b = b' \circ h_1$ and $a = h_n \circ a'$; then $g = b'^\sigma \circ h_1^\sigma \circ h_n \circ a' = c \circ d$. Decompose c and d in any which way: $d = d_l \circ \dots \circ d_1$ and $c = c_m \circ \dots \circ c_1$; then $g = c_m \circ \dots \circ c_1 \circ d_l \circ \dots \circ d_1$ is a decomposition of g, so a right part of it must be a: this is where we use the assumption that h is cracked at the edge to obtain that for some linear L

- either $a = L \circ d_r \circ \dots \circ d_1$ for some $r \leq l$, and then $b^\sigma = c \circ d_l \circ \dots d_{r+1} \circ L^{-1} := c \circ d'$, and $d = d' \circ a$;
- or $a = L \circ c_r \circ \dots \circ c_1 \circ d_l \circ \dots \circ d_1 \circ L^{-1} := c' \circ d$ for some $r \leq m$, and then $b^\sigma = c_m \circ \dots \circ c_{r+1}$ and $c = b^\sigma \circ c'$.

In the second case, $h = a \circ b = c' \circ d \circ b$ and $f = d^\sigma \circ c = d^\sigma \circ b^\sigma \circ c'$, so $(h, f, d \circ b)$ is a plain skew-twist.

In the first case, let's insert the information that $d = d' \circ a$ into the diagram:

(17)
$$\begin{array}{ccc}
\mathbb{P}^1 & \xrightarrow{h=a\circ b} & \mathbb{P}^1 \\
{\scriptstyle b}\downarrow & & {\scriptstyle b^\sigma}\downarrow \\
\mathbb{P}^1 & \xrightarrow{g=b^\sigma\circ a=c\circ d'\circ a} & \mathbb{P}^1 \\
{\scriptstyle a}\downarrow & & {\scriptstyle a^\sigma}\downarrow \\
{\scriptstyle d'}\downarrow & & {\scriptstyle d'^\sigma}\downarrow \\
\mathbb{P}^1 & \xrightarrow{f=(d'\circ a)^\sigma\circ c} & \mathbb{P}^1
\end{array}$$

Now forget all about g, but insert a horizontal arrow one level below:

(18)
$$\begin{array}{ccc}
\mathbb{P}^1 & \xrightarrow{h=a\circ b} & \mathbb{P}^1 \\
{\scriptstyle b}\downarrow & & {\scriptstyle b^\sigma}\downarrow \\
{\scriptstyle a}\downarrow & & {\scriptstyle a^\sigma}\downarrow \\
\mathbb{P}^1 & \xrightarrow{h^\sigma=a^\sigma\circ b^\sigma} & \mathbb{P}^1 \\
{\scriptstyle d'}\downarrow & & {\scriptstyle d'^\sigma}\downarrow \\
\mathbb{P}^1 & \xrightarrow{f=(d'\circ a)^\sigma\circ c} & \mathbb{P}^1
\end{array}$$

Now in the bottom square of this diagram (h^σ, f, d') is a plain skew-twist, and $d \circ b = d' \circ (a \circ b) = d' \circ h$ as wanted. □

3.8.2. *translating back to combinatorics.*

Corollary 3.3.2. *Suppose that $\vec{h}$ is a decomposition of h, h is cracked at the edge, $u \in M_k$ is a sequence of Ritt swaps, and $0 \leq n, m < k$; then there is another sequence of Ritt swaps $v \in M_k$ such that $\phi^n u \phi^m \star \vec{h} = \phi^{m+n} v \star \vec{h}$.*

Proof. To relate this to the proposition above, let $\vec{g} := \phi^m \star \vec{h}$ and $\vec{g'} := u \star \vec{g}$ be two decompositions of the same polynomial g, and let $\vec{f} := \phi^n u \phi^m \star \vec{h}$ be a decomposition of f. Further, let $b := h_m \circ \dots h_1$ and let $d := g'_n \circ \dots g'_1$. Now h, g, f, b, and d satisfy the hypotheses of the proposition. If $(h, f, d \circ b)$ is a plain skew-twist, take v to be the sequence or Ritt swaps that turns $\vec{h}$ into the decomposition of h that has $d \circ b$ as an initial segment.
Otherwise, take v to be the sequence of Ritt swaps that turns $(\vec{h})^\sigma$ into the decomposition that has d' as an initial segment, and note that $\phi^i v \phi^k \approx \phi^i \phi^k v$. □

Corollary 3.3.3. *Suppose that $\vec{h}$ is a decomposition of h, h is cracked at the edge, and $w \in B_k$ does not contain β. Then there is some $v \in M_k$ such that $\phi^M v \approx_{\vec{h}} w$.*

Proof. Use the previous corrolary repeatedly and carefully. Write
$w := \phi^{a_m} u_m \phi^{a_{m-1}} u_{m-1} \dots u_1 \phi^{a_0}$. Use the previous corollary to find v_1 such that $\phi^{a_1+a_0} v_1 \approx_{\vec{h}} \phi^{a_1} u_1 \phi^{a_0}$. Let $\vec{g} := v_1 \star \vec{h}$, still cracked at the edge because of the universal quantifier on decompositions in the definition of "cracked at the edge". Now we're looking at $\phi^{a_m} u_m \phi^{a_{m-1}} u_{m-1} \dots u_2 \phi^{a_1+a_0} \star \vec{g}$, essentially decreasing m by 1. □

Corollary 3.3.4. *Suppose that $\vec{h}$ is a decomposition of h, h is cracked at the edge, and $w \in B_k$; then there is some $n \in \mathbb{N}$ and some $v \in M_k$ such that $w \approx_{\vec{h}} \beta^n v$ or $w \approx_{\vec{h}} \phi^n v$.*

Proof. First use Lemma 3.3.10 to pull all the instances of β out left, then use the previous corrolary, and then cancel any $\beta\phi$ pairs. □

Having inducted, let us forget combinatorics and restate the above result:

Corollary 3.3.5. *If h is cracked at the edge, and an irreducible correspondence $\mathcal{A}$ between h and g comes from skew-twists, then there are decompositions $\vec{h}$ of h and $\vec{g}$ of g and an integer n such that $\mathcal{A}$ is the correspondence encoded by $\beta^n \star \vec{h} = \vec{g}$ or by $\phi^n \star \vec{h} = \vec{g}$.*

Proof. Where did the v from the above Corrolary go? Into the freedom to choose a decomposition of g. □

We rewrite the above without naming decompositions.

Corollary 3.3.6. *If g is cracked, and an irreducible correspondence $\mathcal{A}$ between g and f is encoded by a sequence of skew-twists, then $\mathcal{A}$ is either a graph of $a^{\sigma^{n+1}} \circ f^{\Diamond n}$ for some n and for some initial compositional component a of f; or the corresponding thing in the other direction.*

3.8.3. *who is cracked?* Now that we have a theorem about cracked polynomials, we should prove that some polynomials are indeed cracked. First, some reminders and one new definition:

Definition 3.3.11.
- A pair of indecomposables (a, b) is called *swappable* if there is a Ritt swap $a \circ b = c \circ d$.
- An indecomposable is called *swappable* if it is linearly related to a ritty polynomial.
- A pair of indecomposables (a, b) is called *double-jumpable from the right* if there is some indecomposable c such that $t_2 t_1 \star (a, b, c)$ is defined. Similarly for *double-jumpable from the left.*
- A pair of indecomposables (a, b) is called *a mixing bowl* if there are indecomposables c and d such that $t_2 t_3 t_1 \star (c, a, b, d)$ is defined.

Note that in each of the above cases, something is not a crack: (a, b) for swappable, $(a, b \circ c)$ for double-jumpable from the right, $(c \circ a, b \circ d)$ for a mixing bowl.

The first result will cover the bulk of polynomials and is very easy to prove.

Proposition 3.3.5. *Suppose $\vec{g}$ is a decomposition of a polynomial g and one of the g_i is not swappable; then g is cracked.*

Proof. Let $\vec{h}$ be the plain skew-twist of $\vec{g}$ such that h_1 is unswappable. Then for any decomposition $\vec{f}$ of h, f_1 is (the same up to linear-relatedness) unswappable. It is easy to see that whenever a is an unswappable indecomposable, $(b \circ a, c)$ is a crack for any b and c. So h is cracked at the edge, and g is cracked. □

Now a not-so-hard sufficient condition.

Lemma 3.3.12. *Suppose that for every decomposition $\vec{h}$ of a given polynomial h the pair (h_1^σ, h_k) is not swappable, nor double-jumpable* by indecomposables appearing in some decomposition of *h, nor a mixing bowl* witnessed by indecomposables appearing in some decomposition of *h. Then h is cracked at the edge.*

Proof. We need to show that $((h_i^\sigma \circ \dots \circ h_1^\sigma), (h_k \circ h_{i+1}))$ is a crack. Any other decomposition of the same polynomial can be obtained from this one by a sequence of Ritt swaps in the second canonical form with respect to the three pieces $\vec{a} := (h_i^\sigma, \dots h_2^\sigma)$, (h_1^σ, h_k), and $\vec{b} := h_{k-1}, \dots h_{i+1})$. Swaps within $\vec{a}$ and $\vec{b}$ are irrelevant. No swaps are possible within (h_1^σ, h_k) since it is not swappable. Since (h_1^σ, h_k) is not double-jumpable from the left (by a_1), h_k cannot move left at all. So any decomposition of $h = a \circ h_1^\sigma \circ h_k \circ b$ can be obtained from $(a_r, \dots a_j, \widetilde{h_1^\sigma}, \widetilde{a_{j-1}}, \dots \widetilde{a_1}, h_k, b_s, \dots b_1)$ for some j by moving some of the b_i left as per second canonical form. In particular, $(a_r, \dots a_j, \widetilde{h_i^\sigma}, \widetilde{a_{j-1}}, \dots \widetilde{a_1})$ will be a left segment of the resulting decomposition unless b_s can double-jump $\widetilde{a_1} \circ h_k$, or $h_1^\sigma \circ h_k$ if $j = 1$, from the right. But b_s cannot double-jump $h_1^\sigma \circ h_k$ since it is not double-jumpable, and b_s cannot double-jump $\widetilde{a_1} \circ h_k$ from the right, since that would make $h_1^\sigma \circ h_k$ a mixing bowl. □

3.8.4. *essential translations.* All of these, except when all factors are monomials, are covered by the next borderguard section, though the result there is weaker. Here, we build on our previous analysis of descalings in section 3.4.1. In this section, we describe how to verify the hypotheses of Lemma 3.3.12 for decompositions without type C factors. Since we have already dealt with polynomials that have an unswappable factor, and we are not ready to deal with type C factors, we make the following assumption:

Notation/Assumption 3.3.1. Throughout this Section 3.8.4, all decompositions have no type C factors and no unswappable factors.

Definition 3.3.12. We say that a left descaling (M_i, h_i, L_i) of a decomposition is *clean* if M_k is a scaling and $L_i = \mathrm{id}$ whenever $h_i \circ L_i$ is already ritty.

The point of the definition is that whenever $L_i \neq \mathrm{id}$, the descaling has an essential translation at $(i-1) \mod (k)$; with skew-twists, it makes sense to speak of an essential translation at k. The following generalizes the definition of essential translation in section 3.4.1, in particular defining the notion of an essential translation at k.

Definition 3.3.13. We say that $\vec{h}$ *has an essential translation* at i if $L_{i+1 \mod k} \neq \mathrm{id}$ in some clean left descaling of $\vec{h}$.

In this case we also say that any decomposition $\vec{f}$ which is skew-linearly equivalent to $\vec{h}$ has an essentail translation before i, and even that the polynomial $f := f_k \circ \dots f_1$ has an essentail translation before i.

We now prove that this is a good definition: that every decomposition admits a clean left descaling up to skew-conjugacy; that this definition does not depend on the choice of a clean left descaling, nor on the choice of a decomposition of a given polynomial up to skew-conjugacy.

Lemma 3.3.13. *For every decomposition $\vec{f}$, some decomposition skew-linearly equivalent to $\vec{f}$ admits a clean left descaling.*

Proof. Lemma 3.2.7 proves that any decomposition $\vec{f}$ admits a left descaling. To obtain $L_i = \mathrm{id}$ whenever $h_i \circ L_i$ is already ritty, replace (M_i, h_i, L_i) by $(M_i, h_i \circ L_i, \mathrm{id})$ whenever possible. For the last bit, write $M_k = T \circ S$ for a translation T and scaling S and look at $T^{-1} \circ f \circ T^{(\sigma^{-1})}$. □

Any two clean left descalings of the same decomposition are almost equal:

Lemma 3.3.14. *If (M_i, h_i, L_i) and (B_i, g_i, A_i) are clean left descalings of the same decomposition; then $M_k = B_k$, and $h_i \circ L_i = g_i \circ A_i$ for all i. Further, $h_i = g_i$ and $L_i = A_i$ unless g_i (and therefore h_i) is type J and $A_i \neq \mathrm{id}$.*

Proof. First, observe that both M_k and B_k are the scaling by the leading coefficient of the polynomial decomposed, so they are equal. Now choose linear T_i for $1 \leq k-1$ witnessing linear equivalence: $T_1 \circ h_1 \circ L_1 = g_1 \circ A_1$, and $T_i \circ h_i \circ L_i \circ T_{i-1}^{-1} = g_i \circ A_i$ for $1 < i < k$, and $M_k \circ h_k \circ L_k \circ T_{k-1}^{-1} = B_k \circ g_k \circ A_k$.

Starting with T_1 and inducting up, it is clear that all T_i must be translations. Since $T_1 \circ h_1 \circ L_1 \circ A_1^1 = g_1$ and h_1 is not type C, $T_1 = \mathrm{id}$; inducting, we see that all $T_i = \mathrm{id}$. This gives $h_i \circ L_i = g_i \circ A_i$ for all i. Unless g_i is type J, $h_i = g_i \circ A_i \circ L_i^{-1}$ forces $A_i \circ L_i^{-1} = \mathrm{id}$. If g_i is type J and $A_i = \mathrm{id}$, then $h_i \circ L_i = g_i$ is ritty, so L_i must also be identity. □

Corollary 3.3.7.

- *One clean left descaling of $\vec{f}$ has an essential translation at i if and only if every clean left descaling of $\vec{f}$ has an essential translation at i.*
- *If $\vec{g}$ is skew-linearly equivalent to $\vec{f}$ and both admit clean left descalings, then clean left descalings of one have an essential translation at i if and only if clean left descalings of the other do.*

Proof. The first assertion follows immediately form the previous lemma. The second one is true because linear functions witnessing skew-linear equivalence must be scalings, as can be seen easily from the proof of the previous lemma. □

Lemma 3.3.15.

(1) *If $\vec{f}$ has an essential translation at i, then (f_{i+1}, f_i) is not swappable.*

(2) *If $\vec{f}$ has an essential translation at i, then any decomposition $t_j \star \vec{f}$ obtained from it by a Ritt swap at j also has an essential translation at i.*

Proof. (1) Replacing $\vec{f}$ by a skew-linearly equivalent decomposition if necessary, let (M_i, h_i, L_i) be a clean left descaling of $\vec{f}$. Then (f_{i+1}, f_i) is swappable if and only if $(h_{i+1} \circ L_{i+1}, h_i)$ is swappable. That requires the existence of linear L, M, and N such that $(L \circ h_{i+1} \circ L_{i+1} \circ M^{-1})$ and $(M \circ h_i \circ N$ are both ritty. Since h_i and h_{i+1} are not type C, $L = M = \mathrm{id}$, and then $h_{i+1} \circ L_{i+1}$ is ritty, contradicting the assumption that $\vec{f}$ has an essential translation ay i.

(2) We just proved that $j \neq i$. If $j \neq i+1, i, i-1$, the relevant factors are unchanged and the conclusion is immediate.

If $j = i+1$, let translations L, M, N witness the Ritt swap: $b := L \circ h_{i+2} \circ L_{i+2} \circ M^{-1}$ and $a := M \circ h_{i+1} \circ L_{i+1} \circ N^{-1}$ are ritty, and $d \circ c = b \circ a$ is a basic Ritt identity. Since no factors are type C, $L = M = \mathrm{id}$. Since $a = h_{i+1} \circ L_{i+1} \circ N^{-1}$ is ritty but $\vec{f}$ has an essential translation at i, $N \neq \mathrm{id}$. Now $t_{i+1} \star \vec{f} = (\ldots d, c \circ N, h_i \circ L_i \ldots)$ has an essential translation at i unless $c \circ N$ is ritty. Since $N \neq \mathrm{id}$, if $c \circ N$ is ritty , then c must be type J, which is impossible according to remark 3.1.5.

If $j = i-1$, let translations L, M, N witness the Ritt swap: $b := L \circ h_i \circ L_i \circ M^{-1}$ and $a := M \circ h_{i-1} \circ L_{i-1} \circ N^{-1}$ are ritty, and $d \circ c = b \circ a$ is a basic Ritt identity. Since no factors are type C, $L = M = \mathrm{id}$. Now $t_{i-1} \star \vec{f} = (\ldots h_{i+1} \circ L_{i+1}, d, c \circ N, \ldots)$ still has an essential translation at i.

□

That was the induction step for the following Lemma.

Lemma 3.3.16. *If f has an essential translation at k, then any clean left descaling of any decomposition of any polynomial skew-conjugate to f has an essential translation at i.*

Proof. The statement of this lemma is just like the definition except for a universal quantifier instead of the existential. We need to show that If *some* clean left descaling of *some* decomposition $\vec{g}$ of *some* polynomial g skew-conjugate to f has an essential translation at i, then *any* clean left descaling of *any* decomposition $\vec{h}$ of *any* polynomial h skew-conjugate to f has an essential translation at i. Now $\vec{h}$ is skew-linearly equivalent to $w \star \vec{g}$ for some sequence of Ritt swaps $w \in M_k$, so we induct on the length of w, using the last lemma as the induction step, and Lemma 3.3.7 for the base case where w is empty. □

Lemma 3.3.17. *If h has an essential translation before 1, then h is cracked at the edge.*

Proof. The previous lemma shows that in any clean left descaling of any decomposition $\vec{h}$ of any polynomial skew-conjugate to h, $L_1 \neq \mathrm{id}$. We will use Lemma 3.3.12 to obtain the conclusion of this lemma. In Lemma 3.3.15 above, we showed that (h_1^σ, h_k) is not swappable; it is also neither double-jumpable nor a mixing bowl, because in either of those, the first one or two Ritt swaps would preserve the essential translation (again, according to Lemma 3.3.15), and the last Ritt swap would have to be across an essential translation, which is impossible (again, according to Lemma 3.3.15). □

3.9. **border-guards.** Our analysis so far leaves two cases unexamined: the case when all factors of f are swappable, and one of them is type C; and the case when all factors of f are swappable, none are type C, and there are no essential translations. Both can be attacked with some further combinatorial machinery that we now develop.

3.9.1. *definition.* We define a submonoid G_k of B_k; all words in G_k will leave f_k in its place, though possibly altering it via Ritt swaps in the sense of Remark 3.1.6:

Definition 3.3.14. Let G_k be the free monoid generated by t_1 through t_{k-2} and ψ and γ.
Embed G_k in B_k by mapping t_i to t_i and ψ to $(t_{k-1}\phi)$ and γ to (βt_{k-1}).

We define the action of G_k on S_f by identifying G_k with its image in B_k and denote the action by the same sumbol $\star$. G_k gives just enough wiggle room to perform all the Ritt swaps:

Proposition 3.3.6. *Any word w in B_k is equivalent to $\phi^N w'$ or to $\beta^N w'$ for some integer N and some word w' in G_k.*

Proof. We take $w \in B_k$, start from the right, and move to the left. At every step, we have a word $w_{bad}\beta^a\phi^b w_{good}$ with $w_{bad} \in B_k$ and $w_{good} \in G_k$. We begin with $w_{bad} = w$ and induct on its length; we begin with $a = b = 0$ and w_{good} empty. At the induction step, we must make w_{bad} shorter. Thus, it is sufficient to prove

Claim 1: If $w_{good} \in G_k$, $a, b \in \mathbb{N}$, and s is a generator of B_k, then there are $a', b' \in \mathbb{N}$ and $w'_{good} \in G_k$ such that $s\beta^a\phi^b w_{good} \approx \beta^{a'}\phi^{b'} w'_{good}$.

If $s = \beta$, let $a' := a + 1$, $b' := b$, and $w'_{good} := w_{good}$.
If $s = \phi$ and $a > 0$, let $a' := a - 1$, $b' := b$, and $w'_{good} := w_{good}$.
If $s = \phi$ and $a = 0$, let $a' := a$, $b' := b + 1$, and $w'_{good} := w_{good}$.
If $s = t_i$ for some i, we need

Claim 2: For any t_i a generator of M_k and any $a, b \in \mathbb{N}$, there exists u a generator of G_k and $a', b' \in \mathbb{N}$ such that $t_i\beta^a\phi^b \approx \beta^{a'}\phi^{b'}u$.

We prove Claim 2 by induction on $(a + b)$.

Base cases:
If $a = b = 0$ and $i \neq k - 1$, we let $a' = a$, $b' = b$, and $u = t_i$.
If $a = b = 0$ and $i = k - 1$, note that $t_{k-1} \approx \phi\beta t_{k-1} = \phi\gamma$, so let $a' = 0$, $b' = 1$, and $u = \gamma$.

Inductive cases:
If $a = 0$ and $i \neq k - 1$, then $t_i\phi \approx \phi t_{i+1}$ and we can apply the inductive hypothesis to $t_{i+1}\phi^{b-1}$.
If $a = 0$ and $i = k - 1$, then $b \neq 0$. If $b = 1$, then we are looking at $(t_{k-1}\phi)$, so we let $a' = b' = 0$ and $u = \psi$. If $b \geq 2$, note that $t_{k-1}\phi^2 \approx \phi^2 t_1$ so we can apply the inductive hypothesis to $t_1\phi^{b-2}$.
If $a \neq 0$ and $i \neq 1$, then $t_i\beta \approx \beta t_{i-1}$ and we can apply the inductive hypothesis to $t_{i-1}\beta^{a-1}\phi^b$.
If $a \neq 0$ and $i = 1$, note that $t_1\beta \approx \beta^2 t_{k-1}\phi$. If $a = 1$, then we get $t_1\beta\phi^b \approx \beta^2 t_{k-1}\phi^{b+1}$ and we can apply the second inductive step to $t_{k-1}\phi^{b+1}$. If $a \geq 2$, we get $t_1\beta^a\phi^b \approx \beta^2 t_{k-1}\phi\beta^{a-1}\phi^b \approx \beta^2 t_{k-1}\beta^{a-2}\phi^b$, and we can apply the inductive hypothesis to $t_{k-1}\beta^{a-2}\phi^b$.

Now we have proved Claim 2, which is sufficient to prove Claim 1, which is suficient to prove the proposition. □

We get a weaker analog of Lemma 3.3.10:

Lemma 3.3.18. *For any word $w \in G_k$, there are $w_i \in G_k$ such that $w \approx w_2 w_1$ and γ does not appear in w_1 and ψ does not appear in w_2.*

Proof. Given $w \in G_k$, we find an equivalent word w' that has no substrings of the form $\psi u \gamma$ for some $u \in M_{k-1}$. Clearly, w' is the desired word. To construct w', we prove a

Claim: for any $u \in M_{k-1}$ there is a word $v' \in M_{k-1}$ such that $\psi u\gamma \approx v'$ or $\psi u\gamma \approx \gamma t_{k-2}\psi v'$.

Then replacing a substring $\psi u\gamma$ by one of these does not increase the number of instances of ψ and γ in a word, and straightens out one ψ, γ pair in the wrong order. Thus, after finitely many such operations we obtain the desired w'.

Proof of Claim: Without loss of generality, we may assume that $u \in M_{k-1}$ is in reverse first canonical form, i.e. either $u = v$ or $u = t_1 v$ where t_1 does not appear in $v \in M_{k-1}$. Then $\psi u\gamma \approx v'$ in the first case, and $\psi u\gamma \approx \gamma t_{k-2}\psi v'$ in the second, for some $v' \in M_{k-1}$. □

The purpose of the previous lemma is that with a sufficiently vicious f_k guarding the border, the number of instances of ψ in w_1 and γ in w_2 is severely limited.

Lemma 3.3.19. *If f_k is neither a monomial nor type C, then the correspondences encoded by the words w_i in Lemma 3.3.18 are (up to skew-conjugacy) graphs of P_{n_i}, where n_i are bounded by the product of the in- and out-degrees of f_k.*

Proof. The careful reader will complain that type J indecomposables do not have a well-defined out-degree; but we are only after a *bound*, and for a given type J or coJ indecomposable f_k, there certainly is a bound on the out-degrees it can pretend to have.

We sketch the proof for $w_1 \in G_k$ in which γ does not appear; the proof for w_2 in which ψ does not appear is analogous. We write $w_1 = v_n\psi v_{n-1}\psi \dots v_1\psi v_0$ for some sequences of Ritt swaps $v_i \in M_{k-1}$. We induct on n. Without loss of generality we may replace $\vec{f}$ by a decomposition where $f_k =: a$ is itself ritty and not a monomial.

Since $t_k v_0 \star \vec{f}$ is defined, the second leftmost factor of $v_0 \star \vec{f}$ is linearly equivalent to a monomial: for some linear L and M, and for some prime (not necessarily odd) p, we have $v_0 \star \vec{f} = (a, L \circ P_p \circ M, \dots)$ and $t_k v_0 \star \vec{f} = (P_p, b \circ M, \dots)$, where $(a \circ L) \circ P_p = P_p \circ b$ is a basic Ritt identity, so b cannot be type J because of remark 3.1.5. Then $\psi v_0 \star \vec{f} = (b \circ M, \dots, P_p) =: \vec{g}$, and the correspondence encoded by this equation is the graph of P_p.

Note that $\vec{g}$ satisfies the hypotheses of the lemma. The in- and out-degrees of its leftmost factor b are well-defined up to linear relatedness. Even if a was type J, the out-degree of $(a \circ L)$ is well-defined (not up to linear relatedness). Now the out-degree of b is $\frac{1}{p}$ times the out-degree of $(a \circ L)$. We now apply the inductive hypothesis to $v_n\psi \dots \psi v_1 \star \vec{g}$. □

As before, for the purpose of irreducible skew-invariant curves, we can cancel $\beta\phi$, and therefore $\gamma\psi$. Since monomials commute, if n_1 and n_2 in the lemma have a common factor, we may bring those two together, and then cancel them. So

Corollary 3.3.8. *If f_k is neither a monomial nor type C, and $w \in G_k$, then the only skew-invariant irreducible component of the correspondence encoded by w is defined by $x^{n_1} = y^{n_2}$ for relatively prime n_1 and n_2 bounded by the product of the in- and out-degrees of f_k.*

Together with the last item of Lemma 3.3.18, this characterizes correspondences coming from skew-twists for polynomials that have at least one factor that is neither a monomial nor type C.

Corollary 3.3.9. *For any decomposition $\vec{f}$ with at least one factor that is neither type C nor a monomial, there exists an integer N_f such that for any $w \in B_k$ with $w \star \vec{f}$ defined, there are $w_i \in G_k$ and integers $a < k$ and N such that $w \approx_{\vec{f}} \phi^N w_2 w_1 \phi^a$ or $w \approx_{\vec{f}} \beta^N w_2 w_1 \phi^a$; w_1 contains no occurrences of γ and at most N_f occurrences of ψ; w_2 contains no occurrences of ψ and at most N_f occurrences of γ; and the correspondences encoded by w_i are graphs of monomials of degree at most N_f.*

The N_f in the corollary is the product of in- and out-degrees of the factor that is neither type C nor a monomial, and the maximum possible such if the factor is type J so that his out-degree is not well-defined.

This leaves two questions. What about polynomials each of whose factors is type C or linearly related to a monomial? How do correpondences coming from Theorem 3.3 interact with those coming from skew-twists?

3.10. **Maximal Odd chebyclumps.** Now we deal with the possibility that all factors of f are monomials or type C; and at least one of the factors is type C. We make heavy use of chebyclumps introduced for the proof of Theorem 3.3.

There, we proved that in any decomposition, type C factors occur in clumps which are well-defined up to linear equivalence, and which persist (up to invading quadratic factors) under Ritt swaps. We also proved that Ritt swaps involving two odd Chebyshevs can only occur within these clumps; it is clear that the only other factor that can Ritt swap all the way through, or even any distance into, an odd chebyclump is Q. This makes odd chebyclumps effective borderguards. We begin by extending the technical results to the new context of skew-twists, showing that chebyclumps are invariant under skew-linear equivalence, and that they persist under single skew-twists. First, we must adjust the notion of maximality.

Definition 3.3.15. A maximal chebyclump $(g_k, \dots, g_j)$ of a decomposition $(g_k, \dots, g_1)$, with $j > 1$, is called *skew-maximal* if $(g_1^\sigma, g_k, \dots, g_j)$ is not a chebyclump.

Lemma 3.3.20. *Suppose that a decomposition $\vec{f}$ contains a maximal chebyclump. Then there is a plain skew twist $\vec{g} := \phi^i \star \vec{f}$ such that $(g_k, \dots, g_j)$ is a maximal chebyclump for some j.*

Proof. Suppose that $(f_b, \dots f_a)$ is a maximal chebyclump of $\vec{f}$; then $(g_k, \dots g_{k-b+a})$ is a maximal chebyclump of $\vec{g} := \phi^{k-b} \star \vec{f}$. □

Now we begin to look at correspondences encoded by the words w_i in Lemma 3.3.18 acting on a decomposition that has a maximal chebyclump on the left.

Lemma 3.3.21. *Suppose that $\vec{f}$ is a decomposition of a polynomial f of degree $o' \cdot 2^t$ for some odd o'; suppose that f is not skew-conjugate to a Chebyshev polynomial; suppose that $(f_k, \dots f_a)$ is a maximal chebyclump of degree $o \cdot 2^r$; and suppose that the degree of f_k is odd. Let $w_1 \in G_k$ not contain any instances of γ, and suppose that $w_1 \star \vec{f}$ is defined. Then the correspondence $\mathcal{A}_{w_1}$ is the graph of a chebyshev polynomial C_N where N divides $o \cdot 2^{t+1}$.*

Proof. Write $w_1 = v_n \psi \dots \psi v_0$, and replace $\vec{f}$ by a skew-conjugate decomposition such that $f_i = C_{p_i}$ for each $a \le i \le k$, where p_i may be 2.

We induct on the odd part of the degree of the chebyclump. In the decomposition $\vec{g} := v_0 \star \vec{f}$, $(g_k, \dots g_b)$ is still a chebyclump for some $b \le k-1$, with the same odd part of the degree as the chebyclump $(f_k, \dots f_a)$ in $\vec{f}$. We may again assume that $g_i = C_{q_i}$ for each $b \le i \le k$. Then $\psi v_0 \star \vec{f} = \psi \star \vec{g} = \phi t_{k-1} \star \vec{g} = \phi \star (C_{q_{k-1}}, C_{q_k}, \dots) = (C_{q_k}, \dots, C_{q_{k-1}}) =: \vec{h}$ The correspondence encoded is the graph of $C_{q_{k-1}}$. If q_{k-1} is odd, then the degree of the maximal chebyclump $(h_k, \dots, h_i)$ is $1/q_{k-1}$ times the degree of the chebyclump $(f_k, \dots f_a)$ in $\vec{f}$; in particular, the odd part of the degree of the new chebyclump is less than the odd part of the degree of the old chebyclump, so we have completed the induction step.

Note that, since f is not skew-conjugate to a Chebyshev polynomial, the whole $\vec{h}$ is not a chebyclump, so the odd $h_1 = C_{q_{k-1}}$ cannot rejoin the chebyclump $(h_k, \dots, h_i)$ via Ritt swaps.

We now tell a story in order to avoid gruesome notation. What is different when $q_{k-1} = 2$? What prevents us from inducting on the whole degree of the chebyclump is the fact that it is not invariant under Ritt swaps, and may indeed grow from a decomposition $\vec{h}$ to the decomposition $v_i \star \vec{h}$ as new quadratic factors join the

chebyclump. This can happen in two distinct ways: either some quadratic f_i for $i < a$ from the original decomposition finds his way into the chebyclump (which is accounted for by 2^t), or in some $\psi v_i \dots \psi v_0 \star \vec{f}$ the rightmost factor is a new quadratic who has just been pushed across the border by ψ, and v_{i+1} brings this quadratic all the way to the chebyclump on the left of the decomposition. This may happen once, but if it happens twice, Lemma 3.3.4, a little bit of care, and a whole lot of notation forces the whole $\vec{f}$ to be a chebyclump, and f to be skew-conjugate to a chebyshev polynomial. □

A similar proof yeilds the corresponding statement for $w_2 \in B_k$ in which ψ doesn't appear, and together with Lemma 3.3.18 they give

Proposition 3.3.7. *Given a word $w \in B_k$ and a decomposition $\vec{f}$ with at least one type C factor, let o be the degree of the largest maximal odd chebyclump in $\vec{f}$, and let t be maximal such that 2^t divides the degree of f. Suppose that $w \star \vec{f}$ is defined. Then there there are words w_i in G_k and integers $a < k$ and N, such that*

$$w \approx_{\vec{f}} \phi^N w_2 w_1 \phi^a \text{ or } w \approx_{\vec{f}} \beta^N w_2 w_1 \phi^a$$

and

- *w_1 contains to instances of γ and the correpondence from f to g encoded by $w_1 \star \vec{f} =: \vec{g}$ is the graph of C_A, where A divides $o2^{t+1}$; and*
- *w_2 contains no instances of ϕ and the correpondence from h to $\tilde{g}$ encoded by $w_2 \star \vec{\tilde{g}} =: \vec{h}$ is the graph of C_B, where B divides $o2^{t+1}$.*

3.11. **skew twists summary.** We have now described all correspondences that arise from skew-twists. We have three possible conclusions, the strongest for cracked polynomials in Corollary 3.3.6, and two weaker ones in Corollary 3.3.9 and in Proposition 3.3.7.

Corollary 3.3.6 applies to polynomials with unswappable factors and to polynomials with an essential translation (and, therefore, no type C factors).

Proposition 3.3.7 applies to trivial polynomials with at least one type C factor.

Corollary 3.3.9 applies to polynomials with at least one factor that is neither type C nor a monomial; it is useful if some factor is type C, or if there are no essential translations.

This exhausts trivial polynomials, as a polynomial all of whose factors are linearly related to monomials either has an essential translation, or is skew-conjugate to a monomial and therefore is not trivial.

Actually, the vast majority of decompositions are cracked, yielding the much stronger conclusion of Corollary 3.3.6, but we do not wish to bore the reader with more computations and a long list of exceptions.

3.12. **How skew-twists interact with correspondences from Theorem 3.3.** Correspondences from Theorem 3.3 are composed of pieces in each of which π is a monomial P_p for some prime p. When P_p is also a compositional component of f, these turn out to be skew-twists. Otherwise, they commute with skew-twists.

Proposition 3.3.8. *Suppose that $f_i = P_p$ in a decomposition $\vec{f}$ of f, and $\pi = P_p$ gives a morphism from f to some g. Then this morphism is actually a skew twist, i.e. there is a decomposition $\vec{h}$ of f such that $h_1 = P_p$.*

Proof. Since the foreign P_p from π cannot Ritt swap with the native P_p from f, it is the native P_p from f that must end up all the way at the beginning after the sequence of skew-twists that turns $(P_p, f_k, \dots f_1)$ into $(g_k, \dots, g_1, P_p)$. Putting that sequence into first canonical form shows that the native P_p can end up all the way at the beginning before the foreign P_p does anything, which is precisely what we want to show. □

So this correspondence is a single skew-twist for polynomials, and it is a sequence of Ritt swaps followed by a single skew-twist followed by a sequence of Ritt swaps for the decompositions. These have already been completely characterized in various cases, summarized in the next section. This leaves the case when $\pi = P_p$ is not a compositional factor of f.

Proposition 3.3.9. *Suppose $\mathcal{A}$ is a correspondence from g to h coming from skew-twists, and that π gives a morphism from f to g as in Theorem 3.3. Then there is some d which admits a correspondence $\mathcal{D}$ from f to d coming from skew twists, and such that d, π, and h are as in Theorem 3.3. Indeed, $\mathcal{D}$ is encoded by the same word $w \in B_k$ as $\mathcal{A}$.*
And conversely.

Proof. By induction, it is sufficient to show this for single Ritt swaps and single skew twists. Both are completely obvious: Ritt swaps with P_p only change the degrees of p in the in- and out-degrees of the factors of f, which does not affect the ability of other factors P_q of f to swap, because that ability depends on factors of q in the in- and out-degrees. □

This proposition shows that correspondences coming from skew-twists commute with those coming from Theorem 3.3.

Remark 3.3.3. Correspondences from Theorem 3.3 are always defined by $x^n = y^m$ for some integers m and n. If m and n are not relatively prime, this correspondence is reducible, as it is the correspondence $x^p = y^p$ composed on some other stuff. Its irreducible components are given by $x = \zeta y$ for pth roots of unity ζ. In the context of dynamics, when we assume that the algebraic closure of the prime field sits inside the fixed field of σ, only the diagonal among these is invariant, and the others are periodic, in contrast with correspondences of this form that came from skew twists, where the other components were strictly pre-periodic.

This is the last bit we needed for a complete description of correspondences between σ-varieties given by polynomials.

3.13. **answers.** We prove in the next section that all skew-invariant curves come from skew-twists and from Theorem 3.3, and thus are listed in this theorem.

Theorem 3.4. *Given two trivial polynomials f and h, any (f, h)-skew-invariant curve coming from skew twists and Theorem 3.3 is of the form $\mathcal{A}_3 \circ \mathcal{A}_2 \circ \mathcal{A}_1$ where*

- *$\mathcal{A}_3$ is the graph of an initial compositional factor a of f: it is an $(f, \tilde{f})$-skew-invariant curve, where $f = b \circ a$ and $\tilde{f} = b \circ a^\sigma$.*
- *$\mathcal{A}_2$ is a $(\tilde{f}, g)$-invariant curve, which is*
  - *the diagonal if f has an unswappable factor, or if it is linearly related to a monomial.*

  - *defined by $x^M = y^N$ if some indecomposable compositional factor of f is not linearly related to any monomial or Chebyshev polynomial. In this case, M and N are relatively prime and bounded by the degree of that factor.*
  - *defined by $C_M(x) = C_N(y)$ if some indecomposable compositional factor of f is linearly related to some Chebyshev polynomial. In this case, M and N are relatively prime and each divides $2 \deg f$.*
- *$\mathcal{A}_1$ is the graph of $c^{(\sigma^{n+1})} \circ g^{\Diamond n}$ for some initial compositional factor c of g, where $g = d \circ c$ and $h = d^{(\sigma^n)} \circ c^{(\sigma^{n+1})}$.*

Proof. With Proposition 3.3.9, it is sufficient to insert one correspondence coming from Theorem 3.3 in a place of our choice in the word in B_k in Corollary 3.3.6, Corollary 3.3.9, or Proposition 3.3.7. Correspondences from Theorem 3.3 do not appear for polynomials with an unswappable factor, nor for trivial polynomials linearly related to a monomial. In both other cases we choose to insert it between w_2 and w_1. To see that this works in Proposition 3.3.7, note that since there is an type C factor, the correspondence coming from Theorem 3.3 is the graph of x^2, which is not itself a skew-twist only if f has no quadratic factors. □

In fact, $\mathcal{A}_2$ is the diagonal for the vast majority of polynomials, but the precise characterization of exceptions is tedious. For example, $\mathcal{A}_2$ is the diagonal if f has an essential translation, or if some decomposition of f has two factors f_i and f_j with $\deg_{in}(f_i) \cdot \deg_{out}(f_i)$ and $\deg_{in}(f_j) \cdot \deg_{out}(f_j)$ are relatively prime.

4. Skew invariant varieties

In this section we complete our classification of the skew-invariant varieties for σ-varieties of the form $\Phi : \mathbb{A}^n \to \mathbb{A}^n$ where Φ is given by a sequence of univariate polynomials. The results of Section 3 can be used directly to describe the skew-invariant plane curves. To describe the skew-invariant varieties in higher dimensions we use the model theory of difference fields to reduce to the case of plane curves. After recalling two important ideas from model theory, triviality and orthogonality, we show how the problem of classifying skew-invariant varieties reduces to the cases of linear dynamics, dynamics defined by monomials, and the cases considered in Section 3. We then dispose of the first two cases and conclude by combining these results.

4.1. **Difference algebraic geometry.** Recall from Section 2 that a difference field is a field K given together with a distinguished endomorphism $\sigma : K \to K$. The first order theory of difference fields admits a model companion, ACFA, axiomatized by saying that K is *difference closed* in the sense that it is algebraically closed, σ is an automorphism, and for every irreducible algebraic variety X over K and irreducible subvariety $\Gamma \subseteq X \times X^\sigma$ which projects dominantly in both directions the set of points $(X, \Gamma)^\sharp(K, \sigma) := \{a \in X(K) : (a, \sigma(a)) \in \Gamma(K)\}$ is Zariski dense in X. To say that ACFA is a model companion of the theory of difference fields includes the assertion that every difference field extends to a difference closed field. Note that if X is an irreducible variety over a difference closed field (K, σ) and $f : X \to X^\sigma$ is a dominant map making (X, f) into a σ-variety, then the axioms for ACFA include the assertion that $\{a \in X(K) : \sigma(a) = f(a)\}$ is Zariski dense in X. These facts, that every difference field extends to a difference closed field and that the solutions to $\sigma(x) = f(x)$ become Zariski dense, are what allow us to deduce strong structure

theorems for algebraic dynamical systems and σ-varieties from the corresponding theorems about definable sets in difference closed fields proven by Chatzidakis and Hrushovski in [6].

Before we discuss orthogonality and triviality, let us introduce one new notion. If (X, f) is a σ-variety over some difference field (K, σ), then a *σ-subvariety* is a subvariety $Y \subseteq X$ for which $f \restriction Y : Y \to Y^\sigma$. This is an important notion, and one which is directly implicated by the problem of the classification of the skew-invariant varieties, but we need to consider a slightly stronger condition. A subvariety $Y \subseteq X$ is *difference subvariety* if in some difference closed field (L, σ) extending (K, σ) the set of points $\{a \in Y(L) : f(a) = \sigma(a)\}$ is Zariski dense in Y. Provided that K is algebraically closed, one may replace the phrase "in some difference closed field" by "in every difference closed field."

Two notions from geometric stability theory, a part of modern model theory, are important for us. Fortunately, their usually complicated definitions based on the theory of forking can be replaced by simpler statements in our special setting.

Definition 4.0.1. Let (K, σ) be an algebraically closed difference field and (X, f) and (Y, g) two irreducible σ-varieties over (K, σ) with f and g dominant. We say that (X, f) and (Y, g) are *almost orthogonal over K*, written $(X, f) \perp^a_K (Y, g)$, if every difference subvariey of $(X \times Y, (f, g))$ is a product of a difference subvariety of (X, f) with a difference subvariety of (Y, g). We say that (X, f) and (Y, g) are *orthogonal* if for every difference field extension (L, τ) of (K, σ) one has $(X_L, f) \perp^a_L (Y_L, g)$ where we have written X_L and Y_L for the base changes of these varieties to L.

Remark 4.1. Orthogonality is usually defined at the generic level. What we call (almost) orthogonality is usually called *full* (almost) orthogonality. As we are concentrating on the structure of difference varieties rather than on difference fields, which would be better encoded by generic behavior, we shall take full orthogonality as primitive.

Remark 4.2. The distinction between almost orthogonality and orthogonality is real, but in the cases that concern us, ADs defined by univariate polynomials of degree at least two over fields of characteristic zero, the phenomenon does not appear. However, it is relevant for linear dynamics: for instance, to obtain an isomorphism between $(\mathbb{A}^1, \mathrm{id})$ and $(\mathbb{A}^1, x \mapsto x+1)$ one needs parameters beyond the fixed field of σ. We will see shortly that some linear maps that are not isomorphic over the fixed field are almost-orthogonal over the fixed field, while others are not.

An important result for us is that orthogonality passes from pairs to products, though this result is not true for almost orthogonality.

Fact 4.2.1. *If (K, σ) is a difference field, $(X_1, f_1), \ldots (X_r, f_r)$ and $(Y_1, g_1), \ldots, (Y_s, g_s)$ are σ-varieties over (K, σ) for which $(X_i, f_i) \perp (Y_j, g_j)$ for all i and j, then $(\prod X_i, (f_1, \ldots, f_r)) \perp (\prod Y_j, (g_1, \ldots, g_s))$.*

Two contradictory notions of triviality for σ-varieties appear in the literature. Sometimes (see for instance [16, 15]), one says that a σ-variety closely related to one of the form (X, id_X) is trivial; this is *not* the notion we mean. Our triviality comes from the model-theoretic notion of forking triviality, first isolated in the context of stable theories (see [2]) and then successfully used in the context of difference fields (see [6] and [8] for the development of theory of forking in difference fields). The

following is really a theorem, but it suffices as a definition for the purposes of this paper:

Definition 4.2.1. Let (K, σ) be an algebraically closed difference field and (X, f) a σ-variety over (K, σ). We say that (X, f) is *trivial* if for every $n \in \mathbb{Z}_+$, every irreducible difference subvariety $Y \subseteq (X^{\times n}, f^{\times n})$ is a component of the intersection $\bigcap_{1 \leq i \leq j \leq n} \pi_{i,j}^{-1} \pi_{i,j} Z$ where $\pi_{i,j} : X^{\times n} \to X^{\times 2}$, given by $(x_1, \ldots, x_n) \mapsto (x_i, x_j)$, is the projection onto the i^{th} and j^{th} coordinates.

The main theorem of the first author's doctoral thesis [12] gives an explicit characterization of the rational functions f for which $(\mathbb{P}^1, f)$ is trivial. We abuse notation by saying "trivial polynomial f" to mean "polynomial f such that $(\mathbb{P}^1, f)$ is trivial." Similarly, we write $f \perp g$ to mean $(\mathbb{A}^1, f) \perp (\mathbb{A}^1, g)$. We state the (simpler) special case of that theorem, where f is a polynomial and the characteristic of the field is zero, which is relevant to this paper.

Theorem 4.3. *Suppose that f is a non-constant, non-linear polynomial over a difference field of characteristic zero, and suppose that f is not skew-conjugate to a monomial or a Chebyshev polynomial. Then the σ-variety $(\mathbb{A}^1, f)$ is trivial.*

The hypotheses of this theorem are as weak as possible: the graph of multiplication witnesses that monomials are not trivial. Each Chebyshev polynomial admits a 2-to-1 cover by the corresponding monomial, inheriting its rich structure. Triviality is invariant under isomorphisms, so in particular under skew-conjugation.

It is fairly easy to see that, when specialized to the case of X being a curve, triviality is equivalent to the nonexistence of families of difference subvarieties of X^2 other than horizontal and vertical lines (see Chapter 2 of [14] for details). From the technical results of this paper, one immediately concludes that polynomials satisfying the hypotheses of the theorem do not admit families of difference subvarieties of X^2 other than horizontal and vertical lines, in effect reproving Theorem 4.3.

Combining the Zilber trichotomy for minimal types in ACFA proved in [6] with Theorem 4.3 above and some easy observations, we obtain:

Proposition 4.3.1. *Linear polynomials are non-orthogonal to each other and orthogonal to all other polynomials. For polynomial f, $P_n \not\perp f$ if and only if f is skew-conjugate to P_n or C_n.*

From these observations, we conclude that the difference varieties for coordinatewise polynomial actions may be decomposed into pieces corresponding to each of the three classes from the Zilber trichotomy.

Proposition 4.3.2. *Suppose that polynomials the polynomials Φ_i are linear for $1 \leq i \leq a$, skew-conjugate to monomials and Chebyshevs of degree ≥ 2 for $a+1 \leq i \leq b$, and none of those for $b < i \leq n$ and that $\Phi : \mathbb{A}^n \to \mathbb{A}^n$ is given by $(x_1, \ldots, x_n) \mapsto (\Phi_1(x_i), \ldots, \Phi_n(x_n))$. Then any irreducible difference subvariety of $(\mathbb{A}^n, \Phi)$ is of the form $A \cap B \cap C$, where $A = A_0 \times \mathbb{A}^{(n-a)}$, $B = \mathbb{A}^a \times B_0 \times \mathbb{A}^{(n-b)}$, $C = \mathbb{A}^{a+b} \times C_0$, and each of A, B and C is a difference subvariety. Further, B_0 similarly breaks into pieces according to the degrees of the Φ_i.*

The possible B_0 were already classified via the study of one-based groups in difference closed fields (see [5, 7]). Indeed, non-linear monomials and Chebyshevs define nontrivial, modular difference varieties, which were the key tools in Hrushovski's

proof of the Manin-Mumford conjecture [11]. We pay little attention to them in this paper as their difference subvarieties, such as B_0 above, are well understood.

If $\phi : \mathbb{G}_m^g \to \mathbb{G}_m^g$ is given by $(x_1, \ldots, x_g) \mapsto (x_1^{M_1}, \ldots, x_g^{M_g})$ for integers $M_i \geq 2$, then every difference subvariety of $(\mathbb{G}_m^g, \phi)$ is a finite union of translates of algebraic subgroups of $\mathbb{G}_m^g$.

To deal with Chebyshevs, pull back the coordinates on which they act by $\pi(x) = x + \frac{1}{x}$ to obtain a sub-σ-variety of a σ-variety defined by monomials. It bears noting that not every connected algebraic subgroup of $\mathbb{G}_m^g$ is a difference subvariety of $(\mathbb{G}_m^g, \phi)$.

We have already classified the possible C_0 without having said so explicitly. Indeed, as each $f_i^\sharp$ is a trivial, C_0 must be a component of $\bigcap_{a+b<i\leq j\leq n} \pi_{i,j}^{-1}\pi_{i,j}(C_0)$. Thus, it suffices to describe the invariant curves for $(f_i, f_j) : \mathbb{A}^2 \to \mathbb{A}^2$.

An easy computation of ramification indices yields the following proposition.

Proposition 4.3.3. *If f and g are trivial polynomials and an irreducible curve $C \subset \mathbb{A}^2$ is a sub-σ-variety of $(\mathbb{A}^2, f \times g)$, then there are polynomials π, ρ, and h such that $(a, b) \in C$ if and only if there is some c with $\pi(c) = a$ and $\rho(c) = b$ and the following commutes:*

$$\begin{array}{ccc} \mathbb{A}^1 & \xrightarrow{f} & \mathbb{A}^1 \\ {\scriptstyle\pi}\uparrow & & \uparrow{\scriptstyle\pi^\sigma} \\ \mathbb{A}^1 & \xrightarrow{h} & \mathbb{A}^1 \\ {\scriptstyle\rho}\downarrow & & \downarrow{\scriptstyle\rho^\sigma} \\ \mathbb{A}^1 & \xrightarrow{g} & \mathbb{A}^1 \end{array} \tag{19}$$

Proof. Let α_i be the projection of C onto i^{th} coordinate and let $\tilde{h} : C \to C^\sigma$ be the restriction of $f \times g$ to C.

We need to show that there is a birational isomorphism $\beta : C \to \mathbb{A}^1$ such that all of $\alpha_i \circ \beta^{-1}$ and $\beta^\sigma \tilde{h} \circ \beta^{-1}$ are polynomials. First, we normalize the curve C, then we note that $\deg \tilde{h} = \deg f > 1$, so C must have genus 0 or 1.

Let $\{a_1, \ldots, a_m\} := \alpha_1^{-1}(\infty)$, and let n be the degree of f.

Note that $\tilde{h}$ must be a bijection from $\{a_1, ..., a_m\}$ to $\{a_1^\sigma, ..., a_m^\sigma\}$, since $\tilde{h}$ must take the α_1-fiber above ∞ to the α_1^σ-fiber above $f(\infty) = \infty$, and cannot take any other points into that fiber since f does not take any other points to ∞. Let τ be the permutation of $\{1, \ldots, m\}$ such that $g(a_i) = a_{\tau(i)}^\sigma$.

Let us compare the two ways to compute the ramification index of the diagonal of the diagram at a_i:

$$e_{\alpha_1}(a_i) \cdot e_f(\alpha_1(a_i)) = e_{\tilde{h}}(a_i) \cdot e_{\alpha_1^\sigma}(\tilde{h}(a_i))$$

Since $\alpha_1(a_i) = \infty$ and $e_f(\infty) = n$; $\tilde{h}(a_i) = a_{\tau(i)}^\sigma$ and $e_{\alpha_1^\sigma}(a_{\tau(i)}^\sigma) = e_{\alpha_1}(a_{\tau(i)})$; and $e_{\tilde{h}}(\text{any point}) \leq \deg(\tilde{h}) = n$, the equation becomes

$$e_{\alpha_1}(a_i) \cdot n = e_{\tilde{h}}(a_i) \cdot e_{\alpha_1}(a_{\tau(i)}) \leq n \cdot e_{\alpha_1}(a_{\tau(i)})$$

So for all i, $e_{\alpha_1}(a_i) \leq e_{\alpha_1}(a_{\tau(i)})$, with equality iff $e_{\tilde{h}}(a_i) = n$.

But $\sum_i e_{\alpha_1}(a_i) = \sum_i e_{\alpha_1}(a_{\tau(i)})$ since τ is a permutation.

Therefore, for all i, $e_{\alpha_1}(a_i) = e_{\alpha_1}(a_{\tau(i)})$ and $e_{\tilde{h}}(a_i) = n > 1$.

Notice that, in particular, $\tilde{h}$ ramifies: C admits a ramified, separable, non-constant morphism $\tilde{h}$ of degree grater than 1 to a curve C^σ of the same genus, so the genus of C must be 0.

Since $\tilde{h}$ totally ramifies at all a_i, there are at most two such points. If there are two points, they are either fixed or switched by $\tilde{h}$, in either case contradicting the triviality of f and g as in the first case $\tilde{h}$ is conjugate to x^n and in the second to a the rational funtion x^{-n}. So the unique point P in $\alpha_1^{-1}(\infty)$ is the unique point where $\tilde{h}$ is totally ramified, which by the same argument must also be the unique point in $\alpha_2^{-1}(\infty)$. Any $\beta : C \to \mathbb{A}$ such that $\beta(P) = \infty$ works.

□

Finally, we need to say something about A_0. For a coordinatewise linear $\Phi : \mathbb{A}^k \to \mathbb{A}^k$, the difference subvarieties of $(\mathbb{A}^k, \Phi)$ are very easy to describe: over a difference closed field the σ-variety $(\mathbb{A}^k, \Phi)$ is isomorphic to the σ-variety $(\mathbb{A}^k, \mathrm{id})$, whose difference subvarieties are exactly the subvarieties defined over the fixed field of σ. To exhibit the isomorphism, it suffices to find one solution of the equation $\sigma(x) = \Phi_i(x)$ for each i where we write $\Phi(x_1, \ldots, x_k) = (\Phi_1(x_1), \ldots, \Phi_k(x_k))$. The question is less easy for ADs as we work over the fixed field of σ, so these parameters are not available to us. This section clarifies this situation.

Every linear polynomial is a (possibly trivial) scaling or a translation by 1, up to conjugation by linear polynomials. Therefore, up to isomorphism of dynamical systems, a coordinate-wise linear polynomial action acts on each coordinate by either scaling or by adding 1. The dynamical system on $\mathbb{A}^2$ given by $\Phi(x, y) = (x + 1, y + 1)$ is isomorphic to the one given by $G(z, w) = (z, w + 1)$ via the isomorphism $(x, y) \mapsto (x - y, y)$, so we may also assume without loss of generality that the dynamical system acts by translation on at most one coordinate. We will reduce the more difficult case, where Φ indeed acts on one of the coordinates by translation, to the easier case where Φ acts only by scalings.

That is, we may reduce to the study of algebraic dynamical systems on $\mathbb{A}^r$ or $\mathbb{A}^{r+1}$ of the form $\Phi : \mathbb{A}^{r+1} \to \mathbb{A}^{r+1}$ given by $(x_1, \ldots, x_r, y) \mapsto (\lambda_1 x_1, \ldots, \lambda_r x_r, y + 1)$ where each λ_i is a nonzero scalar or $\Phi : \mathbb{A}^r \to \mathbb{A}^r$ given by $(x_1, \ldots, x_r) \mapsto (\lambda_1 x_r, \ldots, \lambda_r x_r)$.

It is clear that the coordinate hyperplanes defined by $x_i = 0$ are invariant and the the restriction of Φ to any such has the same form, but with one less scaling term. Hence, to analyze the invariant varieties, it suffices to consider Φ on $\mathbb{G}_m^r \times \mathbb{G}_a$. Let $H \leq \mathbb{G}_m^r$ be the smallest algebraic group containing $(\lambda_1, \ldots, \lambda_r)$. Then every Φ-orbit must be contained in a coset of $H \times \mathbb{G}_a$. If M is the index of H^0, the connected component of H, in H, then we see that every orbit of $\Phi^{\circ M}$ is contained in a coset of $H^0 \times \mathbb{G}_a$. As the Φ-invariant varieties are also $\Phi^{\circ M}$-invariant, it suffices to classify the latter. As H^0 is a connected algebraic torus, possibly after base change, it is isomorphic to $\mathbb{G}_m^t$ for some $t \leq r$ and as the action of Φ is semisimple, relative to some isomorphism with $\mathbb{G}_m^t \times \mathbb{G}_a$, the action of Φ on $H^0 \times \mathbb{G}_a$ takes the same form as that of Φ. Hence, we may reduce to the case that $\lambda_1, \ldots, \lambda_r$ are multiplicatively independent.

With the next lemma we use the Skolem-Chabauty [20, 4] method to deduce that there cannot be any interesting algebraic relations on a Φ orbit. In the following proof we make use of the p-adic exponential function and the consequence of the Weierstraß Division Theorem that a convergent p-adic power series in one variable

which vanishes on infinitely many p-adic integers must be identically zero (see chapter 6 of [9] for details).

Definition 4.3.1. If (X, F) is a dynamical system and $P \in X(K)$, then $\mathcal{O}_F(P) \subset X(K)$ is the (forward) orbit of P under F.

Lemma 4.3.1. *Let K be an algebraically closed field of characteristic zero and $\lambda_1, \dots, \lambda_r \in K^\times$ a sequence of multiplicatively independent elements of K. Let $\Phi : \mathbb{G}_m^r \times \mathbb{G}_a \to \mathbb{G}_m^r \times \mathbb{G}_a$ be given by $(x_1, \dots, x_r, y) \mapsto (\lambda_1 x_1, \dots, \lambda_r x_r, y+1)$. Let $P = (a_1, \dots, a_r, b) \in (\mathbb{G}_m^r \times \mathbb{G}_a)(K)$. Then $\mathcal{O}_\Phi(P)$ is Zariski dense in $\mathbb{G}_m^r \times \mathbb{G}_a$.*

Proof. Let us start with a few reductions.

First, the lemma is obvious when $r = 0$. So, we may assume $r > 0$.

Secondly, If $\Psi(x_1, \dots, x_r, y) := (a_1 x_1, \dots, a_r x_r, y + b)$, then Ψ is an automorphism of $\mathbb{G}_m^r \times \mathbb{G}_a$ (as an algebraic variety), takes $\mathcal{O}_\Phi((1, \dots, 1, 0))$ to $\mathcal{O}_\Phi(P)$ and preserves the product structure. Hence, we may, and do assume that $P = (1, \dots, 1, 0)$. Thus, $\mathcal{O}_\Phi(P) = \{(\lambda_1^n, \dots, \lambda_r^n, n) : n \in \mathbb{N}\}$.

Finally, if the lemma fails, then we may assume that there is an irreducible hypersurface Y for which $Y(K) \cap \mathcal{O}_\Phi(P)$ is Zariski dense in Y. Indeed, if $Z \subset \mathbb{G}_m^r \times \mathbb{G}_a$ is an irreducible subvariety which contains a Zariski dense set of points from $\mathcal{O}_\Phi(P)$, then the projection of Z to $\mathbb{G}_m^r$ must be either a point or all of $\mathbb{G}_m^r$ by the Mordell-Lang theorem for the multiplicative group. As $\mathcal{O}_\Phi(P)$ is not contained in $F \times \mathbb{G}_a(K)$ for any finite set F, we see that some component of $\overline{\mathcal{O}_\Phi(P)}$ must be a hypersurface.

So, we can find an irreducible polynomial $G(x_1, \dots, x_r, y) \in K[x_1, \dots, x_n, y]$ defining a hypersurface Y containing a Zariski dense set of points from $\mathcal{O}_\Phi(P)$.

As all of these data are defined over a finitely generated extension of $\mathbb{Q}$, by choosing an appropriate rational prime p we may assume that $G \in \mathbb{Z}_p[x_1, \dots, x_r, y]$ and that each λ_i is a nonzero p-adic number.

Let us write $\lambda_i = \omega_i \exp_p(\mu_i)$ where ω_i is a $(p-1)^{\text{th}}$ root of unity and $\exp_p$ is p-adic exponential function. The function $z \mapsto \lambda_i^z$ is p-adic analytic on each coset of $N + p\mathbb{Z}_p$ of $\mathbb{Z}_p$ (where $N \in \mathbb{Z}$) and is given by the formula $z \mapsto \omega^N \exp(\mu_i(z - N))$. Hence, the function $g : \mathbb{Z}_p \to \mathbb{Z}_p$ given by $g(z) := G(\lambda_1^z, \dots, \lambda_r^z, z)$ is itself analytic on each coset of $p\mathbb{Z}_p$.

From our hypotheses, g vanishes on infinitely many natural numbers, and thus, because it is piecewise analytic, on some coset of $p\mathbb{Z}_p$. Thus, on some coset of $p\mathbb{Z}_p$ the series expansion for g is identically zero.

Let $H(x_1, \dots, x_n, y) := \frac{\partial G}{\partial y} + \sum_{i=1}^r \mu_i X_i \frac{\partial G}{\partial x_i}$.

Differentiating, we see that $g'(z) = H(\lambda_1^z, \dots, \lambda_r^z, z)$. That is, H also vanishes on an infinite subset of $\mathcal{O}_\Phi(P)$ and is therefore an element of the ideal generated by G. So, there is a number α with $H = \alpha G$.

Let us write G in multi-index notation as $G(X, y) = \sum_{I \in \mathbb{N}^r, j \in \mathbb{N}} G_{I,j} x^I y^j$. Then we compute that $H(X, y) = \sum_{I \in \mathbb{N}^r, j \in \mathbb{N}} ((\sum_{i=1}^r \mu_i I_i) G_{I,j} + (j+1) G_{I,j+1}) x^I y^j$.

As Y is not horizontal, there must be some nonzero multi-index I for which there is some j with $G_{I,j} \neq 0$. Choosing j maximal with this property we have $(\sum_{i=1}^r \mu_i I_i) G_{I,j} = \alpha G_{I,j}$. If $\alpha = 0$, then we obtain a nontrivial linear dependence amongst the μ_i's contrary to the multiplicative independence of the λ_i's. In any case, as Y is irreducible, there must be some other multi-index K (possibly an n-tuple of zeros) for which there is some ℓ with $G_{K,\ell} \neq 0$. Again choosing ℓ maximal

with this property we deduce the $\sum_{i=1}^{r} \mu_i K_i = \alpha$. Hence, $\sum_{i=1}^{r} \mu_i (I_i - K_i) = 0$ is a nontrivial linear dependence. With this contradiction we conclude the proof. □

Our calculations in the case of linear dynamics yield the following proposition.

Proposition 4.3.4. *Let $\Phi : \mathbb{A}^n \to \mathbb{A}^n$ be given by $(x_1, \ldots, x_n) \mapsto (\Phi_1(x_1), \ldots, \Phi_n(x_n))$ where each Φ_i is linear. Let r be the dimension of the Zariski closure of the subgroup of GL_n generated by $(\Phi_1, \ldots, \Phi_n)$. Then there is an isomorphism of ADs $\Theta : (\mathbb{A}^n, \Phi) \to (\mathbb{A}^r, \Psi) \times (\mathbb{A}^{n-r}, \mathrm{id})$ and the irreducible Φ-invariant varieties are exactly those of the form $\Theta^{-1}(V \times Y)$ where $V \subseteq \mathbb{A}^r$ is an intersection of coordinate hyperplanes and Y is any irreducible subvariety of $\mathbb{A}^{n-r}$.*

Let us put all of these observations together into a refinement of Proposition 4.3.2.

Theorem 4.4. *Suppose that polynomials Φ_i are linear for $1 \leq i \leq a$, skew-conjugate to monomials and Chebyshevs of degree ≥ 2 for $a+1 \leq i \leq b$, and none of those for $b < i \leq n$ and that $\Phi : \mathbb{A}^n \to \mathbb{A}^n$ is given by $(x_1, \ldots, x_n) \mapsto (\Phi_1(x_i), \ldots, \Phi_n(x_n))$. Then any irreducible difference subvariety of $(\mathbb{A}^n, \Phi)$ is of the form $A \cap B \cap C$, where $A = A_0 \times \mathbb{A}^{(n-a)}$, $B = \mathbb{A}^a \times B_0 \times \mathbb{A}^{(n-b)}$, and each of A, B and C is a difference subvariety. Moreover:*

- *A_0 is described by Proposition 4.3.4.*
- *B_0 is a quotient by a finite group action of a translate of an algebraic torus.*
- *C_0 is a component of $\bigcap_{a+b<i<j\leq n} \pi_{i,j}^{-1} \pi_{i,j}(C_0)$ and $\pi_{i,j}(C_0)$ is a point, a line, or a curve described by Theorem 3.4.*

5. Applications

In this section we use the results from Section 4 to answer some open questions about the model theory of difference fields and the arithmetic of algebraic dynamical systems.

5.1. **Trivial minimal sets in ACFA.**

This section is intended mainly for logicians.

In this section we work in a sufficiently saturated model (L, σ) of ACFA_0. For a polynomial f we write $f^\sharp$ for $\{a \in L : \sigma(a) = f(a)\}$. For a 1-type p over some small substructure of L we write $p \in f^\sharp$ to mean that the formula $\sigma(x) = f(x)$ is an element of the type p, or, equivalently, that for any realization $a \models p$ one has $a \in f^\sharp$. Whenever we say that some property P of a polynomial is *definable* we mean that for each natural number d, the set $\{(a_0, \ldots, a_d) : \sum_{i=0}^{d} a_i x^i$ has property $\mathsf{P}\}$ is a definable set.

Theorem 4.4 allows us to describe the structure of trivial minimal sets of the form $(\mathbb{A}^1, f)^\sharp$ for a trivial polynomial f. In particular, we try to answer two questions definably:

Problem 5.0.1.
(1) *Given f, for what g are there non-orthogonal types $p \in f^\sharp$ and $q \in g^\sharp$?*
(2) *Given $p \in f^\sharp$ and $q \in g^\sharp$, are they non-orthogonal?*

Another theorem from the first author's thesis translates these questions into the language of this paper:

Theorem 5.1. *If $p(x) \in f^\sharp$ is non-orthogonal to $q(y) \in g^\sharp$, then there are polynomials π, ρ, and h such that*

$$\pi : h^\sharp \to f^\sharp \text{ and } \rho : h^\sharp \to g^\sharp$$

and the formula $\theta(x,y) := (\exists z \pi(z) = x \wedge \rho(z) = y)$ witnesses nonorthogonality.

Conversely, it is easy to see that given such h, π, and ρ, any $p \in f^\sharp$ and $q \in g^\sharp$ are non-orthogonal **as long as**

$$(\exists z \pi(z) = x \wedge \sigma(z) = h(z)) \in p \text{ and}$$

$$(\exists z \rho(z) = y \wedge \sigma(z) = h(z)) \in q$$

Otherwise, there might or might not be some other formula witnessing non-orthogonality between p and q.

Thus, the questions we need to answer are

Problem 5.1.1. (1) *Given f, for what g are there h, π, ρ such that $\pi : h^\sharp \to f^\sharp$ and $\rho : h^\sharp \to g^\sharp$?*
(2) *Given f and g, what are the possible h, π, ρ such that $\pi : h^\sharp \to f^\sharp$ and $\rho : h^\sharp \to g^\sharp$?*

As was mentioned above, morphisms of the form $f^{\Diamond n} : (\mathbb{A}, f) \to (\mathbb{A}, f^{(\sigma^n)})$ prevent the definability of answers, to the first question if f is not over the fixed field of any power of σ, and to the second question if f is over some power of σ. However, it follows from Theorem 4.4 that these are the only obstructions. Furthermore, this is an obstruction to question 5.0.1.2 if p is also defined over some fixed field. In particular, if f is not over any fixed field, there are only finitely many definable finite-to-finite correspondences from f to itself, so the model-theoretic algebraic closure on $f^\sharp$ is finite.

Lemma 5.1.1. *The following properties of two trivial polynomials f and g of the same degree are first-order definable in ACFA:*

$$\exists a, b \, f = b \circ a \text{ and } g = b \circ a^\sigma$$

$$\exists h \, \exists N, M \leq 2 \cdot \deg f \, (P_M \circ h = f \circ P_M \text{ and } P_N \circ h = g \circ P_N) \text{ or}$$
$$(C_M \circ h = f \circ C_M \text{ and } C_N \circ h = g \circ C_N)$$

If f is defined over the fixed field of σ^m for some m, $\exists n \, g = f^{(\sigma^n)}$ is also definable.

Lemma 5.1.2. *For fixed trivial polynomials f and g of the same degree, the following properties of a pair of points $A \in f^\sharp$ and $B \in g^\sharp$ are definable:*

$$\exists a, b \, f = b \circ a \text{ and } g = b \circ a^\sigma \text{ and } B = a(A)$$

$$\exists h \, \exists N, M \leq 2 \cdot \deg f (P_M \circ h = f \circ P_M \text{ and } P_N \circ h = g \circ P_N \text{ and } A^N = B^M) \text{ or}$$
$$(C_M \circ h = f \circ C_M \text{ and } C_N \circ h = g \circ C_N \text{ and } C_N(A) = C_N(B))$$

If f is not defined over the fixed field of σ^m for any m, then there is at most one n such that $g = f^{\Diamond n}$, so

$$\exists n \, g = f^{(\sigma^n)} \text{ and } B = f^{\Diamond n}(A)$$

is also definable.

Note that for the last items in both lemmata, the requirement on f is not only sufficient but also necessary: otherwise, there is a countable family of almost-disjoint difference subvarieties of $(\mathbb{A}^2, f \times g)^\sharp$, so definability would produce an infinite definable family of such, contradicting triviality. Combining these lemmata with Theorem 3.4 gives:

Proposition 5.1.1. *For a fixed trivial f, the property "$g \not\perp f$" is definable if and only if f is defined over the fixed field of σ^m for some m.*

For fixed trivial f and g with $f \not\perp g$ the property "$A \in f^\sharp$ and $B \in g^\sharp$ and $B \in \mathrm{acl}(A)$" is definable if and only if f is not defined over the fixed field of σ^m for any m.

Here $\mathrm{acl}(A)$ is the model-theoretic algebraic closure of $A \cup k$ where k is a small model of ACFA over which everything is defined.

A special case $f = g$ of the second part of the proposition gives

Corollary 5.1.1. *For a generic $A \in f^\sharp$, $\mathrm{acl}(A) \cap f^\sharp$ is finite if and only if f is not defined over the fixed field of σ^m for any m.*

In another direction, using Proposition 1.1 in [7] we show that these trivial minimal sets are in fact strongly minimal:

Proposition 5.1.2. `For` *f* `a nonlinear polynomial t`*he minimal set $f^\sharp$ is strongly minimal unless $(L^\sigma \circ f \circ L^{-1})(x) = x^k \cdot u(x)^p$ for some prime p, some integer k, and some linear polynomial L, and $\sigma(\zeta) = \zeta^k$ for some primitive pth root of unity ζ. In that case, it is a finite union of strongly minimal sets.*

Proof. With an easy computation of ramification indices one deduces from Proposition 1.1 in [7] that any infinite-coinfinite subset S of $f^\sharp$ is defined by "$\exists B \in g^\sharp\, \pi(B) = A$" for some polynomials π and g such that $\pi^\sigma \circ g = f \circ \pi$. Our computations show that it suffices to show that this is impossible in the case that π is indecomposable, and that $\pi^\sigma \circ g = f \circ \pi$ either is a single skew-twist, or comes from Theorem 3.3.

In the first case, we show that $\pi : g^\sharp \to f^\sharp$ is onto , so $S = f^\sharp$ is not coinfinite. Indeed, in this case $f = \rho \circ \pi^\sigma$ and $g = \rho \circ \pi$ for some ρ. Given $A \in f^\sharp$, let $B := \sigma^{-1}(\rho(A))$. Note that ρ is a morphism of σ-varieties from $f^\sharp$ to $(g^\sigma)^\sharp$ because $g^\sigma = \rho^\sigma \circ \pi^\sigma$. Then $\rho(A) \in (g^\sigma)^\sharp$, so $B \in g^\sharp$. On the other hand, $\pi(B) = \pi(\sigma^{-1}(\rho(A))) = \sigma^{-1}(\pi(\rho(A))) = \sigma^{-1}(f(A)) = A$.

In order for the second case to be relevant, we must have $\tilde{f}(x) := (L^\sigma \circ f \circ L^{-1})(x) = x^k \cdot u(x)^p$ for some prime p, some integer k, and some linear polynomial L. To lighten notation, we work with $\tilde{f}$ instead of f. Then $\pi(x) = x^p$ and $g(x) = x^k \cdot u(x^p)$. If $A \in f^\sharp$, then σ must take points in the fiber of π above A to the points in the fiber of $\pi^\sigma = \pi$ above $\tilde{f}(A)$. Fix some $B \in \pi^{-1}(A)$ and a primitive pth root of unity ζ, and note that $g(B)$ is in the fiber of $\pi^\sigma = \pi$ above $\tilde{f}(A)$. So $\sigma(B) = \eta \cdot g(B)$ for some pth root of unity η.

What happens with other points in the π-fiber above A? Note that $\pi^{-1}(A) = \{\zeta^i B\}_{i \le k}$ for some primitive pth root of unity ζ, and that $g(\zeta^i B) = (\zeta^i B)^k \cdot u((\zeta^i B)^p) = \zeta^{ik} g(B)$ while $\sigma(\zeta^i \cdot B) = \sigma(\zeta^i) \cdot \sigma(B)$, so

$$\sigma(\zeta^i \cdot B) = \frac{(\sigma(\zeta))^i \eta}{\zeta^{ik}} g(\zeta^i \cdot B)$$

If $\frac{\sigma(\zeta)}{\zeta^k}$ is a primite pth root of unity, we can find i such that $\frac{(\sigma(\zeta))^i\eta}{\zeta^{ik}} = 1$ and then $\zeta^i \cdot B \in \pi^{-1}(A) \cap g^\sharp$ shows that $\pi : g^\sharp \to \tilde{f}^\sharp$ is onto.

Otherwise, the $\frac{\sigma(\zeta)}{\zeta^k} = 1$, and all points in $\pi^{-1}(A)$ belong to $(\eta \cdot g)^\sharp$, so the formulae "$\exists B \in (\eta \cdot g)^\sharp$ such that $B^p = A$" define p disjoint infinite coinfinite subsets of $f^\sharp$, one for each pth root of unity η.

However, this can only happen finitely many times, as each time the out-degree of g is strictly less than the out-degree of f. □

5.2. **Density of dynamical orbits.** In this section we apply Theorem 4.4 to deduce a version of a conjecture of Zhang on the density of dynamical orbits.

Let us recall Zhang's conjecture.

Conjecture 5.2 (Conjecture 4.1.6 of [21])**.** *Let K be a number field and $f : X \to X$ a polarizable dynamical system over K. Then there is point $a \in X(K^{alg})$ algebraic over K whose forward orbit $\mathcal{O}_f(a) := \{f^{\circ n}(a) : n \in \mathbb{Z}_+\}$ is Zariski dense in X.*

The dynamical systems we have been considering, namely, $(\mathbb{A}^n, \Phi)$ given by coordinatewise univariate polynomials as above, do not fit Conjecture 5.2 as stated for a couple of reasons. First, as $\mathbb{A}^n$ is affine, no dynamical system on $\mathbb{A}^n$ can be polarized. More seriously, even if we pass to a projective closure, the hypothesis of polarizability forces all of the polynomials involved to have the same degree. We shall prove that there are dense orbits without these restrictions.

In light of our results and a geometric version of Conjecture 5.2 due to Amerik and Campana [1], we propose a more general conjecture on the density of dynamical orbits.

Conjecture 5.3. *Let K be an algebraically closed field of characteristic zero, X an irreducible algebraic variety over K, and $\Phi : X \to X$ a* rational *self-map. We suppose that there does not exist a positive dimensional algebraic variety Y and dominant rational map $g : X \to Y$ for which $g \circ \Phi = g$ generically. Then there is some point $a \in X(K)$ with a Zariski dense forward orbit.*

We shall prove the instance of Conjecture 5.3 in which X is affine space and Φ is given by a sequence of univariate polynomials.

Theorem 5.4. *Let K be a field of characteristic zero, $f_1, \dots, f_n \in K[x]$ non-constant polynomials over K in one variable. Suppose that the linear polynomials amongst the f_i's are independent in the sense that if f_i are linear for $i \in I \subseteq \{1, \dots, n\}$ then the Zariski closure of the subgroup of $\mathrm{GL}_{|I|}$ generated by $(f_i)_{i\in I}$ has dimension $|I|$. Let $\Phi : \mathbb{A}^n_K \to \mathbb{A}^n_K$ be given by $(x_1, \dots, x_n) \mapsto (f_1(x_1), \dots, f_n(x_n))$. Then there is a point $a \in \mathbb{A}^n(K)$ for which $\mathcal{O}_\Phi(a)$ is Zariski dense.*

Remark 5.5. As one sees from the proof, in some sense almost every point in $\mathbb{A}^n(K)$ has a Zariski dense orbit. We do not pursue the issue of giving a quantitative treatment of this observation.

Remark 5.6. As the reader will see, the notion of *independence* is exactly what is required for Theorem 5.4 to holds for a sequence of linear polynomials. We do not pretend that the inclusion of linear polynomials in this statement is deep, but we have included them as there is little extra work involved in doing so and they round out the statement.

Remark 5.7. Theorem 5.4 may be read as saying that there are points $a \in \mathbb{A}^n(K)$ having the property that for no positive integer N is $\Phi^{\circ N}(a)$ contained in any proper difference subvariety of $(\mathbb{A}^n, \Phi)$ when K is treated as a difference field with $\sigma = \mathrm{id}_K$. In fact, we will prove Theorem 5.4 by explicitly describing the irreducible difference subvarieties of $(\mathbb{A}^n, \Phi^{\circ M})$ for all $M \in \mathbb{Z}_+$ and then observing that there are points in $\mathbb{A}^n(K)$ whose forward orbits miss all such difference subvarieties.

We prove Theorem 5.4 as a consequence of a number of simple lemmata.

Lemma 5.7.1. *Let $f : X \to X$ be an algebraic dynamical system over some field K with X being irreducible. A point $a \in X(K)$ has a Zariski dense forward orbit if and only if there is no natural number m and proper f-invariant subvariety (not necessarily irreducible) of X containing $f^{\circ m}(a)$.*

Proof. For any point $a \in X(K)$, as $f(\mathcal{O}_f(a)) = \mathcal{O}_f(f(a)) \subseteq \mathcal{O}_f(a)$, for $m \gg 0$ the variety $\overline{\mathcal{O}_f(f^{\circ m}(a))}$ is an f-invariant subvariety of X. Hence, if $\mathcal{O}_f(a)$ is not Zariski dense in X, then $\overline{\mathcal{O}_f(f^{\circ m}(a))}$ is a proper f-invariant subvariety of X. Conversely, if $f^{\circ m}(a) \in Y \subsetneq X$ and Y is f-invariant, then $\mathcal{O}_f(a) \subseteq Y(K) \cup \{f^{\circ i}(a) : 0 \leq i \leq m\}$ so that $\overline{\mathcal{O}_f(a)} \subseteq Y \cup \{f^{\circ i}(a) : 0 \leq i \leq m\} \subsetneq X$. □

Lemma 5.7.2. *If $f : X \to X$ is an algebraic dynamical system over some field K, X is irreducible, and $a \in X(K)$ has a Zariski dense forward orbit, then for any $m \in \mathbb{Z}_+$, $X = \overline{\mathcal{O}_{f^{\circ m}}(a)}$*

Proof. For $i = 0, \ldots, m-1$, let $Z_i := \overline{\mathcal{O}_{f^{\circ m}}(f^{\circ i}(a))}$. Then as $\mathcal{O}_f(a) = \bigcup_{i=0}^{m-1} \mathcal{O}_{f^{\circ m}}(f^{\circ i}(a))$, we have $X = \bigcup_{i=0}^{m-1} Z_i$. Hence, $X = Z_i$ for some i. As X has a dense f-orbit, the map $f : X \to X$ is necessarily dominant (otherwise, $\overline{\mathcal{O}_f(a)} \subseteq \{a\} \cup \overline{f(X)} \subsetneq X$). As f maps Z_j to $Z_{j+1 \pmod m}$, we must have $X = Z_j$ for all j. In particular, $X = Z_0 = \overline{\mathcal{O}_{f^{\circ m}}(a)}$. □

Lemma 5.7.3. *Suppose that $f : X \to X$ and $g : Y \to Y$ are algebraic dynamical systems over the field K, $(X, f) \perp (Y, g)$, and that there are rational points $a \in X(K)$ and $b \in Y(K)$ with $\overline{\mathcal{O}_f(a)} = X$ and $\overline{\mathcal{O}_g(b)} = Y$. Then $\overline{\mathcal{O}_{(f,g)}(a, b)} = X \times Y$.*

Proof. Let $Z := \overline{\mathcal{O}_{(f,g)}(a, b)}$ be the Zariski closure of the forward (f, g)-orbit of (a, b). As $(f, g)(\mathcal{O}_{(f,g)}(a, b)) \subseteq \mathcal{O}_{(f,g)}(a, b)$, the variety Z is (f, g)-invariant. As $(X, f) \perp (Y, g)$, Z must be a finite union of varieties of the form $A \times B$ where $A \subseteq X$ is $f^{\circ m}$-invariant for some m and $B \subseteq Y$ is $g^{\circ \ell}$-invariant for some ℓ. Taking a common multiple, we may assume that all such A and B are invariant for the same iterate m of f or g, respectively. Let $A \times B$ be a component containing (a, b). By Lemma 5.7.2, $X = \overline{\mathcal{O}_{f^m}(a)} \subseteq A \subseteq X$ and $Y = \overline{\mathcal{O}_{g^m}(b)} \subseteq B \subseteq Y$. Hence, $X \times Y = \overline{\mathcal{O}_{(f,g)}(a, b)}$. □

The next few lemmata (Lemma 5.7.4, Lemma 5.7.5, and Lemma 5.8.1) are all well-known, but we include them for completeness.

Lemma 5.7.4. *If $f_1, \ldots, f_n \in K[x]$ are independent linear polynomials over a field K of characteristic zero, then there is a point $a \in \mathbb{A}^n(K)$ for which $\mathcal{O}_{(f_1,\ldots,f_n)}(a)$ is Zariski dense in $\mathbb{A}^n$.*

Proof. After a making a linear a transformation, we may assume $(f_1(x), \ldots, f_n(x)) = (\lambda_1 x_1, \ldots, \lambda_n x_n)$ or $(f_1(x), \ldots, f_n(x)) = (\lambda_1 x_1, \ldots, \lambda_{n-1} x_{n-1}, x_n + 1)$ where the

λ_i's are multiplicatively independent. By Lemma 4.3.1, relative to this presentation, any $a \in \mathbb{G}_m^n(K)$ will do. □

Lemma 5.7.5. *If $N_1, \dots, N_n \in \mathbb{Z}_+$ are positive integers each greater than one and K is a field of characteristic zero, then here is a point $a \in (K^\times)^n = \mathbb{G}_m^n(K)$ with a dense f-orbit where $f : \mathbb{G}_m^n \to \mathbb{G}_m^n$ is given by $(x_1, \dots, x_n) \mapsto (x_1^{N_1}, \dots, x_n^{N_n})$.*

Proof. As we noted above each f-invariant variety is a union of translates of algebraic subgroups of $\mathbb{G}_m^n$ by torsion points. Thus, by Lemma 5.7.1, we need only find a point $a \in \mathbb{G}_m^n(K)$ which does not belong to any translate of an algebraic subgroup by a torsion point. Simply take $a = (a_1, \dots, a_n)$ so that the multiplicative group generated by $a_1, \dots, a_n$ has rank n. For example, let these points be n distinct rational primes. The unique factorization theorem for $\mathbb{Z}$ says that this choice works. □

Remark 5.8. In Lemma 5.7.5, it is not true that every translate by a torsion point of an algebraic torus is invariant.

Lemma 5.8.1. *If $N_1, \dots, N_n \in \mathbb{Z}_+$ are positive integers each greater than one, $t \leq n$, and K is a field of characteristic zero, then there is a point $b \in \mathbb{A}^n(K)$ with a dense $g := (C_{N_1}, \dots, C_{N_t}, P_{N_{t+1}}, P_{N_n})$-orbit.*

Proof. Let $a \in \mathbb{G}_m^n(K)$ be a point with a dense $(P_{N_1}, \dots, P_{N_n})$-orbit given by Lemma 5.7.5. The map $h : (\mathbb{G}_m^n, (P_{N_1}, \dots, P_{N_n})) \to (\mathbb{A}^n, (C_{N_1}, \dots, C_{N_t}, P_{N_{t+1}}, \dots, P_{N_n}))$ given by $(x_1, \dots, x_n) \mapsto (x_1 + \frac{1}{x_1}, \dots, x_t + \frac{1}{x_t}, x_{t+1}, \dots, x_n)$ is a dominant map of dynamical systems. Hence, if we set $b := h(a)$, we have $\overline{\mathcal{O}_g(b)} = \mathbb{A}^n$. □

Lemma 5.8.2. *Let K be a field of characteristic zero and f and g two polynomials over K of degree at least 2 neither of which is linearly conjugate to a monomial or a Chebyshev polynomial. Suppose that $R \subseteq K$ is a subring of K over which some decompositions*

$f = f_k \circ \cdots \circ f_1$ and $g = g_r \circ \cdots \circ g_1$ are defined and over which each of the leading coefficients of the polynomials in the decompositions is a unit. Then if $C \subseteq \mathbb{A}_K^2$ is an $(f^{\circ m}, g^{\circ m})$-invariant curve for some $m \in \mathbb{Z}_+$ and $(a, b) \in C(K)$ with $a \in R$, then b is integral over R.

Proof. This follows immediately from Theorem 3.4. The curve C is a component of a composite of correspondences coming encoded by the β and ϕ operators for the decomposition of f, hence as graphs of the given components of f and their converse relations, algebraic tori, and then skew twists of the decomposition of $g^{\circ m}$, each of which is given by a polynomial over R. Following a through these correspondences we see that in each step either we apply a polynomial defined over R (and thus maintain integrality over R) or we extract a root to an equation of the form $h(x) = c$ where c is integral over R and h is a polynomial over R with a unit as leading coefficient. □

Lemma 5.8.3. *Let K be a field of characteristic zero and $f_1, \dots, f_n \in K[x]$ a sequence of nonconstant polynomials over K. We assume that each f_i has degree at least two and is not linearly conjugate to a Chebyshev polynomial or to a monomial. Then there is a rational point $a = (a_1, \dots, a_n) \in \mathbb{A}^n(K)$ with a dense $(f_1, \dots, f_n)$-orbit.*

Proof. Let $R \subseteq K$ be some finitely generated subring over which complete decompositions of each f_i are defined and the leading coefficient of each indecomposable factor is a unit. We argue by induction on i that we can find some finitely generated ring B containing R and contained in K for which there is a point $(a_1, \dots, a_i) \in \mathbb{A}^i(B)$ with $\mathcal{O}_{(f_1,\dots,f_i)}(a)$ Zariski dense in $\mathbb{A}^i$. In the case of $i = 1$, the result follows by height considerations (for example, by embedding $R \subseteq \mathbb{C}$ if we take $a \in R$ with $|a| \gg 0$, then $\lim_{m\to\infty} f_1^{\circ}(a) = \infty$ so that, in particular, a is not preperiodic).

In the inductive case, we have $(a_1, \dots, a_i) \in \mathbb{A}^i(B)$ with a Zariski dense $(f_1, \dots, f_i)$-orbit. Let $a_{i+1} \in K$ be any element of K which is not integral over B. Then for every m, $f^{\circ m}(a_{n+1})$ is also non-integral so by Lemma 5.8.2 $(f^{\circ m}(a_j), f^{\circ m}(a_{i+1}))$ does not belong to any $(f_j^{\circ m}, f_{i+1}^{\circ m})$-invariant curve. By triviality, it follows that $(f_1^{\circ m}(a), \dots, f_{i+1}^{\circ m}(a))$ does not belong to any $(f_1^{\circ m}, \dots, f_{i+1}^{\circ m})$-invariant variety. □

Combining these lemmata and with Theorem 4.4, we conclude that Theorem 5.4 is true.

5.3. **Difference equations for Frobenius lifts.** In this section we observe that for dynamical systems lifting the Frobenius, one can capture the periodic points with a difference equation. Consequently, our results on the structure of difference varieties imply strong restrictions on the algebraic relations among the periodic points of such dynamical systems.

In what follows, K is a field with a valuation v, ring of integers $R := \{x \in K : v(x) \geq 0\}$, maximal ideal $\mathfrak{m} := \{x \in R : v(x) > 0\}$, and residue field $k := R/\mathfrak{m}$ of characteristic $p > 0$. We assume that $\sigma : K \to K$ is an automorphism lifting the p-power Frobenius in the sense that $v(\sigma(x)) = v(x)$ for all $x \in K$ and $\sigma(x) \equiv x^p \mod \mathfrak{m}$ for $x \in R$. We assume moreover that K is maximally complete and algebraically closed. The results we prove about periodic points descend from K to subfields, so the reader may comfortably drop these last two hypotheses, but some of our intermediate results require at least completeness. Ultimately, we shall assume that K has characteristic zero, but for now, this is not necessary.

If X is a scheme over R, then we write X_0 for the base change of X to k and X_η for the base change of X to K. We write $\pi : X(R) \to X_0(k)$ for the natural reduction map.

With Theorem 5.9 we show that difference equations given by liftings of the Frobenius give dynamical Teichmüller maps. Towards the end of this section we specialize to the case of dynamical systems given by sequences of univariate polynomials and thereby deduce form our earlier work that algebraic relations amongst periodic points of such systems are highly restricted.

Theorem 5.9. *Let X be a separated scheme of finite type over R. We assume that X is smooth over R. Suppose that $\Gamma \subseteq X \times X^\sigma$ is a closed subscheme of $X \times X^\sigma$ for which the projection $\Gamma \to X$ is étale. Suppose moreover that $q = p^n$ is a power of p and Γ lifts the Frobenius in the sense that some component of the special fibre Γ_0 is the graph of the geometric q-power Frobenius morphism $F : X_0 \to X_0^{(q)}$. Then the reduction map $\pi : X(R) \to X_0(k)$ restricts to a bijection between $(X, \Gamma)^\sharp(R, \sigma^n)$ and $X_0(k)$.*

Proof. To ease notation let us write $\rho := \sigma^n$.

Let us first show that $\pi : (X, \Gamma)^\sharp(R, \rho) \to X_0(k)$ is surjective. Let $a \in X_0(k)$ be any k-rational point on X_0. Pick any point $\widetilde{a} \in X(R)$ with $\pi(\widetilde{a}) = a$. From the

hypothesis that X is smooth over R, we may fix an étale covering $f : U \to \mathbb{A}^m_R$ where $\widetilde{a} \in U(R)$, $U \subseteq X$ is an affine open subset and $f(a) = \mathbf{0}$. Note that $f^\sigma : U^\sigma \to \mathbb{A}^m_R$ gives analytic coordinates on X^σ near $\sigma(\widetilde{a})$.

As $\Gamma \to X$ is étale, the set $(f \times f^\sigma)(\Gamma(R) \cap \pi^{-1}\{a\} \times (\pi^\sigma)^{-1}\{F(a)\})$ is the graph of an analytic function $g : \mathfrak{m}^m \to \mathfrak{m}^m$ where $g(x_1, \dots, x_m) = (x_1^q, \dots, x_m^q) \bmod \mathfrak{m} \cdot R[[x_1, \dots, x_m]]$. That we can find a solution to $g(\boldsymbol{x}) = \sigma(\boldsymbol{x})$ follows from Newton's method (see [19] in this context).

That is, if for some $\gamma > 0$ we have a solution to $g(x) \equiv \sigma(x) \mod I_\gamma$ where $I_\gamma := \{x \in R : v(x) \geq \gamma\}$, we can find some x' with $x \equiv x' \mod I_\gamma$ but $g(x) \equiv \sigma(x) \bmod I_{\gamma^+} := \{x \in R : v(x) > \gamma\}$ and then taking limits we find a true solution with in the given neighborhood. In our case, we already know that $g(\mathbf{0}) = \mathbf{0} \mod \mathfrak{m} = I_{0^+}$. Given an approximate solution x, suppose that $g(x) \equiv \sigma(x) \mod I_\gamma$ with $\gamma > 0$. Let $\epsilon \in R$ with $v(\epsilon) = \gamma$. We seek to find $x' = x + c\epsilon$ with $c = (c_1, \dots, c_m)$ and $v(c_i) \geq 0$ for each i. We have $g(x + c\epsilon) = g(x) + \sum_{i=1}^m \frac{\partial g}{\partial X_i}(x)c\epsilon + \epsilon^2 * \equiv g(x) \mod I_{\gamma^+}$ as $\frac{\partial g}{\partial X_i}(X) \equiv qX_i^q \mod \mathfrak{m}R[[X_1, \dots, X_m]]$. On the other hand, $\sigma(x + c\epsilon) = \sigma(x) + \sigma(c)\sigma(\epsilon) \equiv \sigma(x) + (c_1^q, \dots, c_m^q)\sigma(\epsilon) \mod I_{\gamma^+}$. Subtracting, we need only solve $\sigma(\epsilon)(c_1^q, \dots, c_m^q) \equiv g(x) - \sigma(x) \mod I_{\gamma^+}$. By hypothesis, each component of $g(x) - \sigma(x)$ has valuation at least $\gamma = v(\sigma(\epsilon))$. As k is perfect, we may solve these equations.

These calculations demonstrate that the restriction of π to $(X, \Gamma)^\sharp(R, \rho)$ is injective as well since the solution $c = (c_1, \dots, c_n)$ is uniquely determined modulo $\mathfrak{m}$. Since we know the residue of the solution, this shows that the reduction map is injective. □

Corollary 5.9.1. *With X and Γ as in Theorem 5.9, for any natural number N one has $(X, \Gamma)^\sharp(R, \rho) = (X, \Gamma^{\diamond N})^\sharp(R, \rho^N)$.*

Proof. A composite of étale extensions is étale. Hence, the hypothesis of Theorem 5.9 apply to X, $\Gamma^{\diamond N}$, and mN. So, $\pi : (X, \Gamma^{\diamond N})^\sharp(R, \rho^N) \to X_0(k)$ is also a bijection. As $(X, \Gamma)^\sharp(R, \rho) \subseteq (X, \Gamma^{\diamond N})^\sharp(R, \rho^N)$, these sets must be equal. □

Specializing Γ somewhat, we may use Theorem 5.9 to find a difference equation for periodic points.

Theorem 5.10. *Let X be a separated scheme of finite type over R, smooth over R and $f : X \to X$ a morphism lifting the $q = p^n$-power Frobenius. Let $\rho := \sigma^n$. We assume that $f = f^\rho$ and $X = X^\rho$. Then every f-periodic R-rational point belongs to $(X, f)^\sharp(R, \rho)$.*

Proof. Let $b \in X(R)$ be an f-periodic point of order M. There are only finitely many solutions to $f^{\circ M}(x) = x$ (as, for instance, this is true on the special fibre). Hence, $\rho^N(b) = b$ for some $N > 0$. Thus, b satisfies $\rho^{MN}(x) = f^{\circ MN}(x)$. That is, $b \in (X, f^{\circ MN})^\sharp(R, \rho^{MN})$ which is $(X, f)^\sharp(R, \rho)$ by Corollary 5.9.1. □

Remark 5.11. Theorem 5.10 holds for f analytic. This observation yields interesting information in the case that X is a moduli space of abelian varieties, $\Gamma \subseteq X \times X$ is a p-power Hecke correspondence, and $f : X \to X$ (or, really, f is defined on some dense open subset) is a branch of Γ lifting the Frobenius. In this case, the difference equation captures the canonical lifts. (See [18] for more details.)

Remark 5.12. If in Theorem 5.10 we assume that $k = \mathbb{F}_p^{\mathrm{alg}}$, then as every point in $X(k)$ is f-periodic, every point in $(X, f)^\sharp(R, \rho)$ is f-periodic.

Remark 5.13. This method of obtaining interesting difference equations for periodic points by lifting equations on the Frobenius has been used in the study of Manin-Mumford questions [11, 16]. When more structure (for instance, a group) is available, then more complicated equations beyond simply $f(x) = \sigma(x)$ may be used to give deeper information. We expect that these equations in the more general dynamical context will be useful, but we have not pursued this issue.

Let us conclude by specializing to the case of sequences of univariate polynomials.

Theorem 5.14. *Let $q = p^\ell$ be a power of p. We suppose that K has characteristic zero. Let $f_1, \ldots, f_n \in R[x]$ be polynomials with $f_i(x) \equiv x^p \mod \mathfrak{m}R[x]$ for each $i \leq n$. We suppose that for some $m > 0$ each $f_i = f_i^{\sigma^m}$ for each i. If $X \subseteq \mathbb{A}^n_K$ is an irreducible subvariety containing a Zariski dense set of points of the form $(\zeta_1, \ldots, \zeta_n)$ where $\zeta_i \in R$ is f_i-periodic, then X is a difference subvariety of $(\mathbb{A}^n, (f_1^{\diamond m}, \ldots, f_n^{\diamond m}))$ and has the shape described in Section 5.2. Moreover, if $\deg(f_i) = q$ for each i, then we may replace the hypothesis "$\zeta_i \in R$" by $\zeta_i \in K$."*

Proof. By Theorem 5.10, the $(f_1, \ldots, f_n)$-periodic points in $\mathbb{A}^n(R)$ are all contained in $(\mathbb{A}^n, (f_1^{\diamond m}, \ldots, f_n^{\diamond m}))^\sharp(R, \sigma^{\ell m})$. Hence, if X contains a Zariski dense set of periodic points from $\mathbb{A}^n(R)$, then $X \cap (\mathbb{A}^n, (f_1^{\diamond m}, \ldots, f_n^{\diamond m}))^\sharp(R, \sigma^{\ell m})$ is Zariski dense in X implying that X is a difference subvariety of $(\mathbb{A}^n, (f_1^{\diamond m}, \ldots, f_n^{\diamond m}))$. The description of X now follows from our description of such difference varieties.

For the "moreover" clause observe that if $\deg(f_i) = q$, then every f_i-periodic point is integral over R, and, hence, actually an element of R as R is integrally closed in K. □

Remark 5.15. Further specializing Theorem 5.14 one obtains statements about algebraic relations amongst the periodic points of polynomial without reference to valuations. For example, let q be a power of a prime number p. Suppose that $f(x) = x^q + pg(x)$ where $g(x) \in \mathbb{Z}[x]$ and $\deg(g) \leq q$. Suppose moreover that f is not linearly conjugate to a monomial or a Chebyshev polynomial and that f is not a compositional power. Then every irreducible variety $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ which contains a Zariski dense set of n-tuples of f-periodic points is defined by a sequence of equations of the form $f(x_i) = x_j$ or $f(x_\ell) = a$ for a some fixed f-periodic point.

The remark is true because all compositional factors of f have degree a power of p. In particular, no two have relatively prime degrees, so no Ritt swaps are possible amongst them.

References

[1] Ekaterina Amerik and Frédéric Campana. Fibrations méromorphes sur certaines variétés à fibré canonique trivial. *Pure Appl. Math. Q.*, 4(2, part 1):509–545, 2008.

[2] John T. Baldwin and Leo Harrington. Trivial pursuit: remarks on the main gap. *Ann. Pure Appl. Logic*, 34(3):209–230, 1987. Stability in model theory (Trento, 1984).

[3] A. F. Beardon. Symmetries of Julia sets. *Bull. London Math. Soc.*, 22(6):576–582, 1990.

[4] Claude Chabauty. Sur les points rationnels des courbes algébriques de genre supérieur à l'unité. *C. R. Acad. Sci. Paris*, 212:882–885, 1941.

[5] Zoé Chatzidakis. Groups definable in ACFA. In *Algebraic model theory (Toronto, ON, 1996)*, volume 496 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 25–52. Kluwer Acad. Publ., Dordrecht, 1997.

[6] Zoé Chatzidakis and Ehud Hrushovski. Model theory of difference fields. *Trans. Amer. Math. Soc.*, 351(8):2997–3071, 1999.

[7] Zoé Chatzidakis and Ehud Hrushovski. Difference fields and descent in algebraic dynamics. I. *J. Inst. Math. Jussieu*, 7(4):653–686, 2008.
[8] Zoé Chatzidakis, Ehud Hrushovski, and Ya'acov Peterzil. Model theory of difference fields. II. Periodic ideals and the trichotomy in all characteristics. *Proc. London Math. Soc. (3)*, 85(2):257–311, 2002.
[9] Fernando Q. Gouvêa. *p-Adic Numbers*. Universitext. Springer-Verlag, Berlin, 1993. An introduction.
[10] E. Hrushovski and M. Itai. On model complete differential fields. *Trans. Amer. Math. Soc.*, 355(11):4267–4296 (electronic), 2003.
[11] Ehud Hrushovski. The Manin-Mumford conjecture and the model theory of difference fields. *Ann. Pure Appl. Logic*, 112(1):43–115, 2001.
[12] A. Medvedev. *Minimal sets in ACFA*. PhD thesis, UC Berkeley, 2007.
[13] Peter Müller and Michael Zieve. On Ritt's polynomial decomposition theorems. `arXiv:0807.3578v1`, 38 pages, 2008.
[14] Anand Pillay. *Geometric Stability Theory*, volume 32 of *Oxford Logic Guides*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.
[15] Anand Pillay and Martin Ziegler. Jet spaces of varieties over differential and difference fields. *Selecta Math. (N.S.)*, 9(4):579–599, 2003.
[16] R. Pink and D. Roessler. On ψ-invariant subvarieties of semiabelian varieties and the Manin-Mumford conjecture. *J. Algebraic Geom.*, 13(4):771–798, 2004.
[17] J. F. Ritt. Prime and composite polynomials. *Trans. Amer. Math. Soc.*, 23(1):51–66, 1922.
[18] T. Scanlon. Local André-Oort conjecture for the universal abelian variety. *Invent. Math.*, 163(1):191–211, 2006.
[19] Thomas Scanlon. Analytic difference rings. In *International Congress of Mathematicians. Vol. II*, pages 71–92. Eur. Math. Soc., Zürich, 2006.
[20] Th. Skolem. Einige Sätze über π-adische Potenzreihen mit Anwendung auf gewisse exponentielle Gleichungen. *Math. Ann.*, 111(1):399–424, 1935.
[21] S. W. Zhang. Distributions in algebraic dynamics, a tribute to professor S.S. Chern. volume 10 of *Survey in Differential Geometry*, pages 381–430. International Press, 2006.

Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 322 Science and Engineering Offices (M/C 249), 851 S. Morgan Street, Chicago, IL 60607-7045

E-mail address: `alice@math.uic.edu`

University of California, Berkeley, Department of Mathematics, Evans Hall, Berkeley, CA 94720-3840

E-mail address: `scanlon@math.berkeley.edu`