Invariant varieties for polynomial dynamical systems

By Alice Medvedev and Thomas Scanlon

Abstract

We study algebraic dynamical systems (and, more generally, σ-varieties) $\Phi : \mathbb{A}_{\mathbb{C}}^n \to \mathbb{A}_{\mathbb{C}}^n$ given by coordinatewise univariate polynomials by refining an old theorem of Ritt on compositional identities amongst polynomials. More precisely, we find a nearly canonical way to write a polynomial as a composition of "clusters" from which one may easily read off possible compositional identities. Our main result is an explicit description of the (weakly) skew-invariant varieties, that is, for a fixed field automorphism $\sigma : \mathbb{C} \to \mathbb{C}$ those algebraic varieties $X \subseteq \mathbb{A}_{\mathbb{C}}^n$ for which $\Phi(X) \subseteq X^\sigma$. As a special case, we show that if $f(x) \in \mathbb{C}[x]$ is a polynomial of degree at least two that is not conjugate to a monomial, Chebyshev polynomial or a negative Chebyshev polynomial, and $X \subseteq \mathbb{A}_{\mathbb{C}}^2$ is an irreducible curve that is invariant under the action of $(x, y) \mapsto (f(x), f(y))$ and projects dominantly in both directions, then X must be the graph of a polynomial that commutes with f under composition. 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 disintegrated set defined by $\sigma(x) = f(x)$ for a polynomial f has Morley rank one and is usually strongly minimal, that model theoretic algebraic closure is a locally finite closure operator on the nonalgebraic points of this set unless the skew-conjugacy class of f is defined over a fixed field of a power of σ, and that nonorthogonality between two such sets is definable in families if the skew-conjugacy class of 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 $\Phi : \mathbb{A}_{\mathbb{C}}^n \to \mathbb{A}_{\mathbb{C}}^n$ 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 $X \subseteq \mathbb{A}^n_\mathbb{C}$ that are invariant under Φ. To do so, we solve a more general problem. We fix a field automorphism $\sigma : \mathbb{C} \to \mathbb{C}$, describe those algebraic varieties $X \subseteq \mathbb{A}^n_\mathbb{C}$ that are (weakly) 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. 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 [4] that we employ.

In [11] the first author refined the trichotomy theorems of [4], [6] for sets defined by formulas of the form $\sigma(x) = f(x)$, where f is a rational function, showing that they are *disintegrated*, or what is sometimes called *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 dominant rational function $\pi : G \to \mathbb{P}^1$ with $f \circ \pi = \pi^\sigma \circ \phi$. In this context, disintegratedness is a very strong property: it asserts that all algebraic relations amongst solutions to disintegrated equations are reducible to *binary* relations. This consequence and the fact that the dynamical systems arising from isogenies are well-understood 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 $C \subseteq \mathbb{A}^2_\mathbb{C}$ that are (f, g)-skew-invariant when f and g are disintegrated polynomials in the sense of the previous paragraph. It is not hard to reduce this problem to describing triples (h, π, ρ) of polynomials satisfying $f \circ \pi = \pi^\sigma \circ h$ and $g \circ \rho = \rho^\sigma \circ h$ (see Proposition 2.34). Possible compositional identities involving polynomials over $\mathbb{C}$ were explicitly classified by Ritt in [15], 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 quadruples of polynomials (a, b, c, d) in $\mathbb{C}[x]$ satisfy $a \circ b = c \circ d$ in [21].

Our combinatorial analysis of the ingredients of Ritt's theorem yields a refinement that is in some ways weaker and in other ways stronger than the ones in [21]. Applying our refinements of Ritt's theorem to the compositional equations involving f, g, h, π, and ρ, we explicitly describe all (f, g)-skew-invariant plane curves in terms of a decomposition of f as a compositional product.

We should say a few words as to what we mean by weaker and stronger. Ritt's theorem asserts that any one decomposition of a polynomial over $\mathbb{C}$ may be obtained from any other decomposition via a finite sequence of explicit identities, which we call *Ritt swaps*. From this theorem one might expect that it would be a routine matter to put a polynomial into a standard form as a composition of indecomposable polynomials. However, ambiguity as to the character of certain polynomials may be introduced through compositions with linear polynomials. A central part of our argument (as well as of [21]) consists of characterizing exactly how compositional identities involving the special polynomials appearing in Ritt's theorem and linear polynomials may hold. While the individual steps in these calculations are very easy, pinning down all of the possibilities requires an exhaustive analysis. From this point, our results on canonical forms diverge. While the formalism of [21] is well suited to studying decompositions of compositional powers, it is not well adapted to the problem of describing possible skew-invariant varieties. We discuss the comparison between our theorems on polynomial decompositions and those from [21] in detail in the body of the paper.

Our key technical innovation is the notion of a clustering of a decomposition whereby the various compositional factors are grouped, or "clustered," according to their combinatorial properties; for example, compatible Chebyshev polynomials are clustered together. Clusterings are not canonical, but using some invariants computed from clusterings we may read off properties of possible compositional identities. Specifically, with Theorem 4.37 we show that the number and types of the clusters (see Definition 4.2) appearing in a clustering of a decomposition, as well as the location of the "gates" (see Definition 4.15), are invariants of a polynomial, independent of a choice of decomposition.

In every reasonable sense, for almost every pair of polynomials (f, g), there are no (f, g)-skew-invariant curves other than products of the form $\{\xi\} \times \mathbb{A}^1$ or $\mathbb{A}^1 \times \{\zeta\}$ where $f(\xi) = \sigma(\xi)$ (respectively, $g(\zeta) = \sigma(\zeta)$). Indeed, even if $f = g = g^\sigma$, in most cases, the only additional (f, f)-invariant curves are graphs of iterates of f and their converse relations. For instance, it is easy to see that this holds for f indecomposable by using our reformulation of the existence of an (f, g)-skew-invariant curve in terms of compositional identities $f \circ \pi = \pi^\sigma \circ h$ and $g \circ \rho = \rho^\sigma \circ h$.

More generally, there are four basic sources for skew-invariant curves. Some come 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. If f is polynomial of degree at least two, then the set of linear polynomials L that skew commute with f in the sense that

$f \circ L = L^\sigma \circ f$ is generally finite but sometimes is nontrivial. If L skew-commutes with f, then the curve defined by $y = L(x)$ is necessarily (f, f)-skew-invariant. When f is expressible as a nontrivial compositional product, $f = a \circ b$, then considering what we call a *plain skew-twist* of f, $g := b^\sigma \circ a$, we see that the graph of b is (f, g)-skew-invariant. While all possible plain skew-twists can be easily read off from one expression of f as a composition of indecomposable polynomials, it takes more work to characterize the possible sequences of plain skew-twists that originate from f. 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 version of the assertion that these examples exhaust the possibilities for skew-invariant curves.

Our characterization of the skew-invariant varieties appears as a combination of Theorem 2.30 with Theorem 6.26. Given a regular map $\Phi : \mathbb{A}^N \to \mathbb{A}^N$ of the form $(x_1, \ldots, x_N) \mapsto (f_1(x_1), \ldots, f_N(x_N))$ where each f_i is a nonconstant polynomial, the coordinates may be partitioned according the trichotomy theorem for difference fields. That is, Φ may be realized as a Cartesian product of three maps of this form where in the first map each polynomial f_i is linear, for the second map each f_i is linearly conjugate to a power map, a Chebyshev polynomial or a negative Chebyshev polynomial, and for the third map each f_i is disintegrated. Then, the Φ-skew-invariant varieties are products of the skew-invariant varieties for each of these three components. It is a routine matter to classify the skew-invariant varieties for sequences of linear polynomials. It follows from the theory of one-based groups (or a straightforward degree computation) that the skew-invariant subvarieties for sequences of power maps and Chebyshev polynomials come from algebraic tori. Moreover, a skew-invariant variety for sequences of power maps and Chebyshev polynomials may be further decomposed into products of skew-invariant subvarieties for the subsequences consisting of power maps and Chebyshev polynomials of the same degree. We collect all of these observations in Theorem 2.30.

The most complicated class of skew-invariant varieties appear as skew-invariant subvarieties of $(\mathbb{A}^N, \Phi)$ where for some sequence of disintegrated polynomials $f_1, \ldots, f_N$ the map Φ takes the form

$$(x_1, \ldots, x_N) \mapsto (f_1(x_1), \ldots, f_N(x_N)).$$

Using disintegratedness, we see that any Φ-skew-invariant variety must be a component of the intersection of pullbacks of (f_i, f_j)-skew-invariant varieties ranging over all pairs (i, j) with $1 \leq i < j \leq n$. Thus, the classification of Φ-skew-invariant varieties reduces to the case that $N = 2$. With Theorem 6.22 we clarify the sense in which such skew-invariant curves must come from compositions of skew-twists, monomial equations, and graphs of twisted iterates.

In particular, we give very tight bounds on the degrees of the monomial equations that might appear. These bounds are in terms of some refined degrees of the indecomposable polynomials appearing in some complete decomposition of f_1 and are bounded by the degree of f_1 in the usual sense. The following is an especially notable special case.

THEOREM (see Theorem 6.24). *Let $f(x) \in \mathbb{C}[x]$ be a polynomial of degree at least two that is not conjugate to a monomial, a Chebyshev polynomial or a negative Chebyshev polynomial. Let $N \in \mathbb{Z}_+$ be a positive integer and $X \subseteq \mathbb{A}_\mathbb{C}^N$ be an irreducible subvariety of affine N-space over the complex numbers that is invariant under the coordinatewise action of f. Then X is defined by equations over the form $x_i = g(x_j)$ and $x_k = \xi$, where g is a polynomial that commutes with f and ξ is a fixed point of f. Moreover, g takes the form $L \circ h^{\circ m}$ for some $m \in \mathbb{N}$ where $h^{\circ \ell} = f$ for some ℓ and L is a linear polynomial that commutes with a compositional power of h.*

We apply our results on skew-invariant varieties to address problems of two different kinds. We prove variants of two conjectures of Zhang [20] on the arithmetic of dynamical systems. We also pin down definable structure on, and definable relations between, sets defined by $\sigma(x) = f(x)$ for some polynomial f in ACFA$_0$.

Zhang conjectured that if $\phi : X \to X$ is a polarizable dynamical system over some number field K, then there is a point $a \in X(K^{\mathrm{alg}})$ whose forward orbit under ϕ is Zariski dense (Conjecture 4.1.6 of [20]). We consider a situation inspired by Zhang's conjecture, but which is at one level more general in that we drop the polarizability hypothesis and strengthen the conclusion in that one need not pass to the algebraic closure to find the desired point with a Zariski dense orbit, but in another sense is more special in that the map ϕ is assumed to be given by a sequence of univariate polynomials. Let us note here a somewhat special case of our Theorem 7.16.

THEOREM. *If K is any field of characteristic zero and $\Phi : \mathbb{A}_K^n \to \mathbb{A}_K^n$ is given by a sequence of univariate polynomials each of degree at least two, then there is a point $a \in \mathbb{A}^n(K)$ with a Zariski dense Φ-forward orbit.*

In fact, we prove a somewhat stronger result in which some of the f_i's 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. Zhang conjectured that if $\phi : X \to X$ were a polarized dynamical system over $\mathbb{C}$ and $Y \subseteq X$ were an irreducible closed subvariety for which the ϕ-preperiodic points lying on Y were Zariski dense in Y, then Y would be a ϕ-preperiodic variety (Conjecture 1.2.1 of [20]). Counterexamples to this statement have been advanced, and the conjecture itself has been reformulated [7]. As with our theorem on the density of

rational orbits, our dynamical Manin-Mumford theorem is both more and less general than what is predicted by the corrected dynamical Manin-Mumford conjecture. We do not require polarizability, though we do consider only periodic points and dynamical systems arising from liftings of the Frobenius. Our precise statement is given as Theorem 7.33. Let us mention a special case here.

THEOREM. *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 $f(x) \equiv x^q \mod p\mathbb{Z}[x]$ but f is not conjugate to x^q itself, nor to the qth Chebyshev polynomial, nor to the negative Chebyshev polynomial of degree q. Then any irreducible variety $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ containing a Zariski dense set of n-tuples of f-periodic points is defined by equations of the following two forms. Some equations are of the form $x_i = \xi$ for some f-periodic point ξ. The rest are of the form $x_j = L \circ \alpha^{\circ m}(x_k)$ for some $m \in \mathbb{N}$ where $\alpha^{\circ N} = f$ for some N and L commutes with some compositional power of α.*

In the case of differential fields, Hrushovski and Itai showed that there are model complete theories of differential fields other than the theory of differentially closed fields [9]. It is still open whether or not there are model complete theories of difference fields other than ACFA. But if there were some formula $\theta(x)$ defining in a difference closed field a set of D-rank one having only finitely many algebraic realizations such that for every other formula $\eta(y, z)$, the set of parameters $\{b : \theta(x) \text{ is nonorthogonal to } \eta(y, b)\}$ were definable, then one could produce a new model complete difference field by omitting the nonalgebraic types in θ. Towards this goal, we prove Theorem 7.6.

THEOREM (Theorem 7.6). *For a nonconstant polynomial f, the set of polynomials g with $(\mathbb{A}^1, g) \not\perp (\mathbb{A}^1, f)$ is definable if and only if f is not skew-conjugate to f^{σ^n} for every positive integer $n \in \mathbb{Z}_+$.*

Two byproducts of this analysis are an explicit characterization of the algebraic closure operator on disintegrated sets defined by $\sigma(x) = f(x)$ and the computation of Morley rank and degree of these sets. As long as f is disintegrated, this set has Morley rank one, and if f is also not skew-conjugate to $x^k u(x)^n$ for some polynomial u and some $n > 1$, then the set is strongly minimal.

This paper is organized as follows. In Section 2 we lay out our notation and begin our analysis by reducing the problem to that of describing skew-invariant curves for pairs of disintegrated polynomials. We then convert this problem to one about compositional identities and lay out the problem in terms of certain monoid actions. The technical work on compositional identities is spread over the next four sections. In Section 3 we identify all of the possible ways in which one Ritt polynomial may be linearly related to another Ritt polynomial. In Section 4 we study clusterings of decompositions, ending the

section with a theorem on invariants of polynomials computed from these clusterings. In Section 5 we intensify the study of the monoid actions producing canonical forms. In Section 6, we finally obtain the characterization of the skew-invariant curves for actions of pairs of polynomials from the results on clusterings and monoid actions. In Section 7 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 [21] and for discussing issues around compositional identities of polynomials and rational functions. We thank the referee for subjecting our manuscript to a thorough review and suggesting numerous improvements.

2. **Coarse structure of skew-invariant varieties**

In this section we compare the formalism of algebraic dynamical systems and of σ-varieties to establish the relevance of the model theory of difference fields to the study of algebraic dynamics. We then interpret the fundamental trichotomy theorem for difference fields in terms of skew-invariant varieties. Using this interpretation, we reduce the analysis of skew-invariant varieties for maps given by coordinatewise actions of univariate polynomials on affine space to that of describing the skew-invariant curves in $\mathbb{A}^2$ for pairs of disintegrated polynomials. We close out this section by recalling Ritt's theorem on polynomial decompositions in detail and by formalizing that theorem in terms of actions of various monoids. In so doing, we convert the problem of describing invariant varieties into questions about canonical forms for these monoid actions.

2.1. *Algebraic dynamics and σ-varieties.*

Definition 2.1. A *difference field* (L, σ) is a field L given together with a distinguished field endomorphism $\sigma : L \to L$. The *fixed field* of a difference field (L, σ) is the subfield $\mathrm{Fix}(\sigma) := \{a \in L : \sigma(a) = a\}$.

If X is an algebraic variety over the difference field (L, σ), then the σ-transform of X, X^σ is the base change of X to L via σ. More concretely, if X is a closed subvariety of some affine space, then X^σ is defined by the same equations as X but with σ applied to the coefficients. At the level of rational points, $a \in X(K) \leftrightarrow \sigma(a) \in X^\sigma(K)$. The σ-transform gives a endofunctor of the category of algebraic varieties over L. That is, if $f : X \to Y$ is a morphism of varieties over L, then there is a uniquely defined morphism $f^\sigma : X^\sigma \to Y^\sigma$ of varieties over L where again, concretely, on affine charts the polynomials defining f^σ are the images under σ of the polynomials defining f.

Definition 2.2. Following Pink and Rössler [13], a σ-*variety* over the difference field (L, σ) is a pair (X, f) where X is an algebraic variety over L and $f : X \to X^\sigma$ is a dominant morphism from X to its σ-transform X^σ. A morphism of σ-varieties $\alpha : (X, f) \to (Y, g)$ is given by a morphism of varieties $\alpha : X \to Y$ for which $\alpha^\sigma \circ f = g \circ \alpha$:

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

We say that two σ-varieties (X, f) and (Y, g) are *skew-conjugate* if they are isomorphic as σ-varieties.

In particular, univariate polynomials f and g (which give σ-varieties on $\mathbb{A}^1$) are skew-conjugate if there is a linear polynomial α such that $g = \alpha^\sigma \circ f \circ \alpha^{-1}$.

Definition 2.3. An *algebraic dynamical system* over a field K is a pair (X, f) consisting of an algebraic variety X over K and a dominant regular map $f : X \to X$. A morphism $\alpha : (X, f) \to (Y, g)$ of algebraic dynamical systems is given by a regular map $\alpha : X \to Y$ for which $\alpha \circ f = g \circ \alpha$.

The algebraic dynamical system (X, f) over K may be regarded as a σ-variety over (K, id_K). An algebraic dynamical system (X, f) gives rise to a monoid action of $\mathbb{N}$ on X via iteration of f. We define $f^{\circ n}$ by recursion on n with $f^{\circ 0} := \mathrm{id}_X$ and $f^{\circ(n+1)} := f \circ f^{\circ n}$. For a rational point $a \in X(K)$, we define the forward orbit of a under f to be $\mathcal{O}_f(a) := \{f^{\circ n}(a) : n \in \mathbb{N}\}$. The point a is said to be *periodic* if $f^{\circ n}(a) = a$ for some $n \in \mathbb{Z}_+$ and to be *pre-periodic* if $\mathcal{O}_f(a)$ is finite (or, equivalently, if $f^{\circ n}(a)$ is periodic for some $n \in \mathbb{N}$).

If (X, f) is an algebraic dynamical system and $Y \subseteq X$ is a subvariety, then we say that Y is an f-*invariant variety* if $f(Y) = Y$. Equivalently, Y is f-invariant when $f(Y(K))$ is Zariski dense in Y when K is an algebraically closed field. We say that Y is *weakly f-invariant* if $f(Y) \subseteq Y$. If (X, f) is an algebraic dynamical system and $a \in X(K)$ is any point, then the Zariski closure of $\mathcal{O}_f(a)$ is a weakly f-*invariant subvariety* of X. Thus, $\mathcal{O}_f(a)$ is Zariski dense in X if and only if for no $n \in \mathbb{N}$ does $f^{\circ n}(a)$ lie on a (possibly reducible) proper f-invariant variety. In this way Zhang's conjecture on the existence of algebraic points with Zariski dense forward orbits (see Section 7.2) may be understood as an assertion that there are few f-invariant varieties.

In the more general context of a σ-variety, iteration need not give rise to maps from X back to itself, but it still makes sense. For a σ-variety (X, f) over (L, σ), we define the *skew-iteration* of (X, f) by recursion on n setting $f^{\lozenge 0} := \mathrm{id}_X$ and $f^{\lozenge(n+1)} := f^{\sigma^n} \circ f^{\lozenge n}$. Observe that $(X, f^{\lozenge n})$ is a σ^n-variety over (L, σ^n).

To distinguish Cartesian powers from $f^{\Diamond n}$ and $f^{\circ n}$, we sometimes write $f^{\times n}$ for the map $f^{\times n} : X^{\times n} \to (X^{\sigma})^{\times n}$ given by

$$(x_1, \dots, x_n) \mapsto (f(x_1), \dots, f(x_n)).$$

For (X, f) a σ-variety over the difference field (L, σ), a sub-variety $Y \subseteq X$ is *weakly f-skew-invariant* if $f(Y) \subseteq Y^{\sigma}$. The subvariety Y is *f-skew-invariant* if $f(Y) = Y^{\sigma}$, or, equivalently, if $(Y, f \restriction Y)$ is a σ-variety. A weakly skew-invariant variety need not be skew-invariant, but there is a naturally associated maximal f-skew-invariant subvariety.

Definition 2.4. Let (X, f) be a σ-variety and $Y \subseteq X$ a subvariety of X. The *f-skew-invariant part of* Y is the subvariety $Y_{\text{inv}} := \bigcap_{n=0}^{\infty} (f^{\Diamond n}(Y))^{\sigma^{-n}}$.

PROPOSITION 2.5. *If (X, f) is a σ-variety and $Y \subseteq X$ is a subvariety of X, then Y_{inv} is the maximal f-skew-invariant subvariety of Y. If Y is weakly f-skew-invariant, then $Y_{\text{inv}} = (f^{\Diamond n}(Y))^{\sigma^{-n}}$ for $n \gg 0$.*

Proof. From its definition, we have

$$Y_{\text{inv}}^{\sigma} = \Big(\bigcap_{n=0}^{\infty} (f^{\Diamond n}(Y))^{\sigma^{-n}} \Big)^{\sigma} \subseteq \Big(\bigcap_{n=1}^{\infty} (f^{\Diamond n}(Y))^{\sigma^{-n}} \Big)^{\sigma}$$

$$= \bigcap_{m=0}^{\infty} f^{\Diamond m+1}(Y)^{\sigma^{-m}} = f\Big(\bigcap_{m=0}^{\infty} f^{\Diamond m}(Y)^{\sigma^{-m}} \Big) = f(Y_{\text{inv}}).$$

As $\dim(Y_{\text{inv}}^{\sigma}) \geq \dim(f(Y_{\text{inv}}))$ and the number of components of Y_{inv}^{σ} is at least that of $f(Y_{\text{inv}})$, we conclude that $f(Y_{\text{inv}}) = Y_{\text{inv}}^{\sigma}$. On the other hand, if $Z \subseteq Y$ were f-skew-invariant, then for every n we would have $Z = f^{\Diamond n}(Z)^{\sigma^{-n}} \subseteq f^{\Diamond n}(Y)^{\sigma^{-n}}$. Hence, $Z \subseteq Y_{\text{inv}}$ so that Y_{inv} is the maximal f-skew-invariant subvariety of Y.

If Y were weakly f-skew-invariant, then the intersection defining Y_{inv} would be an intersection over a decreasing chain and, thus, equal to $f^{\Diamond n}(Y)^{\sigma^{-n}}$ for $n \gg 0$ by Noetherianity. □

PROPOSITION 2.6. *If $\pi : (X, f) \to (Y, g)$ is a map of σ-varieties, $Z \subseteq X$ is a subvariety of X and $W \subseteq Y$ is a subvariety of Y, then $\pi(Z_{\text{inv}}) = \pi(Z)_{\text{inv}}$ and $\pi^{-1}(W_{\text{inv}}) = \pi^{-1}(W)_{\text{inv}}$.*

Proof. From the fact that $\pi^{\sigma} \circ f = g \circ \pi$, we conclude that $\pi^{\sigma^n} \circ f^{\Diamond n} = g^{\Diamond n} \circ \pi$ for any $n \in \mathbb{N}$. Thus, from the definition of the skew-invariant part we have

$$\pi(Z)_{\text{inv}} = \bigcap_{n=0}^{\infty} (g^{\Diamond n}(\pi(Z)))^{\sigma^{-n}} = \bigcap_{n=0}^{\infty} (\pi^{\sigma^n} \circ f^{\Diamond n}(Z))^{\sigma^{-n}}$$

$$= \pi\Big(\bigcap_{n=0}^{\infty} (f^{\Diamond n}(Z))^{\sigma^{-n}} \Big) = \pi(Z_{\text{inv}}),$$

as claimed.

Likewise,

$$\pi^{-1}(W)_{\mathrm{inv}} = \bigcap_{n=0}^{\infty} (f^{\Diamond n} \pi^{-1} W)^{\sigma^{-n}} = \bigcap_{n=0}^{\infty} ((\pi^{\sigma^n})^{-1} g^{\Diamond n}(W))^{\sigma^{-n}}$$

$$= \pi^{-1} \bigcap_{n=0}^{\infty} (g^{\Diamond n}(W))^{\sigma^{-n}} = \pi^{-1}(W_{\mathrm{inv}}). \qquad \square$$

Let us note that if (X, f) is a σ-variety over (K, id_K), then (X, f) is simply an algebraic dynamical system over K, $f^{\Diamond n} = f^{\circ n}$ for each $n \in \mathbb{N}$, and a subvariety $Y \subseteq X$ defined over K is f-skew-invariant just in case it is f-invariant. However, if we start with an algebraic dynamical system (X, f) over some field K and then form the base change $(X, f)_L$ to some difference field (L, σ) where $\sigma \upharpoonright K = \mathrm{id}_K$, the notions of an f-invariant subvariety and of an f-skew-invariant variety need not coincide. Moreover, there are algebraic dynamical systems (X, f) and (Y, g) that are nonisomorphic as algebraic dynamical systems (and remain so after any field extension) but that become isomorphic as σ-varieties after an appropriate base extension. For example, $(\mathbb{A}^1, x \mapsto x + 1)$ and $(\mathbb{A}^1, \mathrm{id})$ are clearly not isomorphic as algebraic dynamical systems, but after base change to a difference field containing a solution to the difference equation $\sigma(b) = b + 1$, they become isomorphic as σ-varieties.

2.2. *Model theory of difference fields.* In this section we translate some of the fundamental theorems on the model theory of difference fields to a more geometric language. The reader can find a more thorough treatment of these connections in [5]. All of the theorems on the model theory of difference fields that we require can be found in [4].

A difference field (K, σ) is *difference closed* if it is existentially closed in the class of difference fields. That is, if a finite system of difference equations and inequations over K has a solution in some difference field extending (K, σ), then it already has a solution in (K, σ). By successively adjoining solutions to such systems of difference equations and inequations, one sees that every difference field embeds into a difference closed field. The class of difference closed fields is axiomatized by three schemata of axioms expressible in the language of difference fields, the language of rings augmented by a unary function symbol for the distinguished endomorphism. It is obvious that the first schema is given by a first-order sentence. A routine argument expresses the second as a countable list of sentences. However, the last schema requires absolute irreducibility of a variety to be a first-order property of the coefficients of the defining equations. This is attained by bounding the degrees for the ideal membership problem in polynomial rings.

FACT 2.7 (Theorem 1.1 of [4]). *A difference field (K, σ) is difference closed if and only if*

(1) σ is an automorphism of K;
(2) K is algebraically closed; and
(3) for any irreducible affine variety X defined over K and any irreducible subvariety $Y \subseteq X \times X^\sigma$ for which the two projections $Y \to X$ and $Y \to X^\sigma$ are dominant, there is a point $a \in X(K)$ with $(a, \sigma(a)) \in Y(K)$.

From the axioms for difference closed fields, we see that if (K, σ) is a difference closed field, X is an irreducible variety over K, $\Gamma \subseteq X \times X^\sigma$ is an irreducible subvariety of $X \times X^\sigma$ for which the two projections $\Gamma \to X$ and $\Gamma \to X^\sigma$ are dominant, and we define the (K, σ) points of (X, Γ) by

$$(X, \Gamma)^\sharp(K, \sigma) := \{a \in X(K) : (a, \sigma(a)) \in \Gamma(K)\},$$

then $(X, \Gamma)^\sharp(K, \sigma)$ is Zariski dense in X. In particular, if (X, f) is an irreducible σ-variety over a difference closed field (K, σ), then

$$\{a \in X(K) : f(a) = \sigma(a)\} = (X, \Gamma(f))^\sharp(K, \sigma)$$

is Zariski dense in X, where $\Gamma(f)$ is the graph of f. Moreover, an absolutely irreducible subvariety $Y \subseteq X$ is f-skew-invariant if and only if $Y(K) \cap (X, \Gamma(f))^\sharp(K, \sigma)$ is Zariski dense in Y. In this sense, there are enough rational points defined over a difference closed field to reflect the geometry of σ-varieties, or even of algebraic dynamical systems regarded as σ-varieties obtained by base change from the fixed field. We use this observation to translate results from the structure theory of definable sets in difference closed fields to the language of algebraic dynamical systems and σ-varieties.

Let us recall the notion of orthogonality, specialized to the case of σ-varieties. That σ-varieties of different characters (e.g., those coming from group actions versus those unrelated to groups) are orthogonal is the first step in the reduction of the study of skew-invariant varieties in general to the special case of skew-invariant curves in the plane.

Definition 2.8. Two absolutely irreducible σ-varieties (X, f) and (Y, g) over a difference field (K, σ) are *almost orthogonal*, written $(X, f) \perp_K^a (Y, g)$, if every (f, g)-skew-invariant subvariety of $X \times Y$ is a finite union of products of components of f-skew-invariant and g-skew-invariant varieties. If for every extension of difference fields $(L, \sigma) \supseteq (K, \sigma)$ we have $(X_L, f) \perp_L^a (Y_L, g)$, then (X, f) and (Y, g) are *orthogonal*, written $(X, f) \perp (Y, g)$,

Remark 2.9. What we are calling (almost) orthogonality is usually called *full (almost) quantifier-free orthogonality* in the model theory literature. The subtler notions of orthogonality for types, while present in the background, are not directly relevant to the problems we consider here. In fact, nonorthogonality of (X, f) and (Y, g) need not imply that some type in $(X, f)^\sharp$ is nonorthogonal to some type in $(Y, g)^\sharp$, as the (f, g)-skew-invariant subvariety witnessing

nonorthogonality in our sense may have no sharp points at all. Our notion does correspond to eventual nonorthogonality: if (X, f) and (Y, g) are nonorthogonal in our sense, then for some m, some type in $(X, f^{\Diamond m})^{\sharp}$, where $(X, f^{\Diamond m})$ is regarded as a σ^m-variety, is nonorthogonal to some type in $(Y, g^{\Diamond m})^{\sharp}$, even in the reduct to σ^m. We return to issues around quantifier elimination in Section 7.1.

Remark 2.10. The distinction between almost orthogonality and orthogonality is real. For example, if K is any field of characteristic zero, then the σ-varieties $(\mathbb{A}^1, \mathrm{id}_{\mathbb{A}^1})$ and $(\mathbb{A}^1, x \mapsto x + 1)$ are almost orthogonal over (K, id_K). However, after base extension to any difference field containing a solution a to the difference equation $\sigma(x) = x + 1$, these σ-varieties are isomorphic as σ-varieties via the map $x \mapsto x + a$. Then the graph of this isomorphism gives a skew-invariant variety not expressible as a product, witnessing the nonorthogonality of these two σ-varieties. As a general rule, such instances of almost orthogonality but nonorthogonality are mediated by the action of a definable group. Thus, for the σ-varieties of principal concern to us, the *disintegrated* σ-varieties (see Definition 2.20), at least when working over an algebraically closed base, there is no difference between almost orthogonality and orthogonality.

The nonorthogonality relation defines an equivalence relation on the set of σ-varieties whose underlying varieties are irreducible curves.

PROPOSITION 2.11. *Suppose that X, Y, and Z are absolutely irreducible curves and (X, f), (Y, g), and (Z, h) are σ-varieties over some difference field K. If $(X, f) \not\perp (Y, g)$ and $(Y, g) \not\perp (Z, h)$, then $(X, f) \not\perp (Z, h)$.*

Remark 2.12. From the model theoretic perspective, Proposition 2.11 is almost a special case of the fact that the nonorthogonality relation is an equivalence relation on minimal types. The provisos in Remark 2.9 explain the sense in which this remark is only approximately true.

When we view a subvariety $\Gamma \subseteq X \times Y$ as a many valued function from X to Y, we call it a *correspondence* from X to Y. Before proving Proposition 2.11, we recall what it means to compose correspondences and record some basic properties of this operation.

Definition 2.13. Let X, Y, and Z be three varieties over some field K, and let $\Gamma \subseteq X \times Y$ and $\Xi \subseteq Y \times Z$ be subvarieties of $X \times Y$ and $Y \times Z$, respectively. Let $\pi : X \times Y \times Z \to X \times Z$ be the projection map onto the first and third coordinates. We define $\Xi \circ \Gamma := \pi((\Gamma \times Z) \cap (X \times \Xi))$, the projection of the fibre product of Γ and Ξ over Y. If $W \subseteq X$ is any subvariety, then $\Gamma(W)$ is the projection to Y of $\Gamma \circ \Delta_W$, where $\Delta_W \subseteq W \times X$ is the graph of the embedding of W in X.

Remark 2.14. At the level of points, provided that $K = K^{\text{alg}}$, $\Xi \circ \Gamma$ is the Zariski closure of the set

$$\{(a, c) \in (X \times Z)(K) : (\exists b \in Y(K))\ (a, b) \in \Gamma(K)\ \&\ (b, c) \in \Xi(K)\}.$$

Remark 2.15. One treats a correspondence $\Gamma \subseteq X \times Y$ as a many valued function from X to Y. Provided that the projection map $\pi : \Gamma \to X$ is dominant, this "function" is defined almost everywhere. If $f : X \to Y$ is a rational function, then one regards f as a correspondence by identifying f with its graph $\Gamma(f)$. Switching the roles of input and output in $\Gamma(f)$ gives what we call *the converse relation to the graph of* f. If the projection map to the output coordinate is finite, then Γ may be regarded as a finite valued function. In the cases of interest to us, X and Y are irreducible curves and Γ is a curve, each of whose components projects dominantly to X and to Y. Here, $\Gamma(K)$ really is a finite-to-finite correspondence between $X(K)$ and $Y(K)$.

Remark 2.16. Even if X, Y, Z, Γ, and Ξ are all irreducible, then $\Xi \circ \Gamma$ may be reducible. For example, if $X = Y = Z = \mathbb{A}^1$, $f : \mathbb{A}^1 \to \mathbb{A}^1$ is any polynomial of degree at least two, Γ is the graph of f, and Ξ is its converse relation, then $\Xi \circ \Gamma$ is defined by $f(x) = f(z)$, which always has the diagonal as one component and other components corresponding to the factors of the polynomial $\frac{f(x) - f(z)}{x - z}$.

LEMMA 2.17. *Let (X, f), (Y, g) and (Z, h) be σ-varieties over some difference field (K, σ). Suppose that $\Gamma \subseteq X \times Y$ is weakly (f, g)-skew-invariant and that $\Xi \subseteq Y \times Z$ is weakly (g, h)-skew-invariant. Then $\Xi \circ \Gamma$ is weakly (f, h)-skew-invariant and $(\Xi \circ \Gamma)_{\text{inv}} = (\Xi_{\text{inv}} \circ \Gamma_{\text{inv}})_{\text{inv}}$.*

Proof. Clearly, the intersection of two (weakly) (f, g, h)-skew-invariant varieties is (f, g, h)-skew-invariant so that $(\Gamma \times Z) \cap (X \times \Xi)$ is weakly (f, g, h)-skew-invariant and $(\Gamma_{\text{inv}} \times Z) \cap (X \times \Xi_{\text{inv}})$ is (f, g, h)-skew-invariant. Let $\pi : X \times Y \times Z \to X \times Z$ be the projection map. By Proposition 2.6, $(\Xi \circ \Gamma)_{\text{inv}} = (\pi((\Gamma \times Z) \cap (X \times \Xi)))_{\text{inv}} = (\pi((\Gamma_{\text{inv}} \times Z) \cap (X \times \Xi_{\text{inv}})))_{\text{inv}} = (\Xi_{\text{inv}} \circ \Gamma_{\text{inv}})_{\text{inv}}$. □

With our observations on compositions in place, we prove Proposition 2.11.

Proof. Taking a base change if need be, we find $\Gamma \subseteq X \times Y$ and $\Xi \subseteq Y \times Z$ that are (f, g)-skew-invariant (respectively, (g, h)-skew-invariant) curves witnessing $(X, f) \not\perp (Y, g)$ and $(Y, g) \not\perp (Z, h)$. By Lemma 2.17, $(\Xi \circ \Gamma)_{\text{inv}}$ is an (f, h)-skew-invariant subvariety of $X \times Z$. Since Ξ and Γ are curves for which the restriction of the various projection maps are all finite, $\Xi \circ \Gamma$ is a curve all of whose components project dominantly onto X and Z. Because the maps f, g, and h are finite, $\dim(\Xi \circ \Gamma)_{\text{inv}} = \dim(\Xi \circ \Gamma) = 1$. Hence, $(\Xi \circ \Gamma)_{\text{inv}}$ witnesses $(X, f) \not\perp (Z, h)$. □

We employ the theory of orthogonality to reduce the study of skew-invariant varieties to that of plane curves. To this end we use a simple, but powerful, observation that orthogonality of products of σ-varieties follows from pairwise orthogonality.

PROPOSITION 2.18. *Given a difference field (K, σ) and two sequences of σ-varieties $(X_1, f_1), \ldots, (X_n, f_n)$ and $(Y_1, g_1), \ldots, (Y_m, g_m)$ for which $(X_i, f_i) \perp (Y_j, g_j)$ for each $i \leq n$ and $j \leq m$, we have*

$$\prod_{i=1}^{n}(X_i, f_i) \perp \prod_{j=1}^{m}(Y_j, g_i).$$

Remark 2.19. In model theoretic stability theory, Proposition 2.18 is usually deduced as an immediate consequence of transitivity for the independence relation coming from nonforking.

Proof. Working by induction, one sees that it suffices to show that if (X, f), (Y, g), and (Z, h) are σ-varieties for which $(X, f) \perp (Y, g)$ and $(X, f) \perp (Z, h)$, then $(X, f) \perp (Y \times Z, (g, h))$. Now let (L, σ) be some difference field extension of (K, σ) and $U \subseteq (X \times (Y \times Z))_L$ an (f, g, h)-skew-invariant variety over L. For any difference field extension (M, σ) of (L, σ) and point $a \in (Z, h)^{\sharp}(M, \sigma)$, the fibre U_a of U is a (f, g)-skew-invariant subvariety of $(X \times Y)_M$. Since $(X, f) \perp (Y, g)$, we know that U_a is a finite union of varieties of the form $V(a) \times W(a)$, where $V(a) \subseteq X_M$ is f-skew-invariant and $W(a) \subseteq Y_M$ is g-skew-invariant. Since this is true for every point in $(Z, h)^{\sharp}$, it follows from compactness that there are finite sequences of locally closed sets $V_i \subseteq X \times Z$ and $W_i \subseteq Y \times Z$ (for $i \leq n$) so that for any $a \in (Z, h)^{\sharp}(M, \sigma)$, there is some $J \subseteq \{1, \ldots, n\}$ with

$$U_a = \bigcup_{i \in J}((V_i)_a \times (W_i)_a).$$

Taking the sequences to be minimal, we see that each V_i is a component of an (f, h)-skew-invariant subvariety of $X \times Z$, and and each W_i is a component of a (g, h)-skew-invariant subvariety of $Y \times Z$. Hence, by orthogonality, we may write each V_i as a finite union of products of f-skew-invariant varieties with h-skew-invariant varieties. Hence, U itself is a finite union of products of components of f-skew-invariant varieties with components of (g, h)-skew-invariant varieties. □

It is difficult to determine whether two given σ-varieties are orthogonal, though one expects that "most" pairs of σ-varieties are orthogonal. However, we exhibit a procedure to determine orthogonality in the special case of σ-varieties of the form $(\mathbb{A}^1, f)$. On the other hand, there are some easily verified sufficient conditions for orthogonality. For example, if $f : \mathbb{P}^1 \to \mathbb{P}^1$

and $g : \mathbb{P}^1 \to \mathbb{P}^1$ are two rational functions and $\deg(f) \neq \deg(g)$, then $(\mathbb{P}^1, f) \perp (\mathbb{P}^1, g)$. (This follows immediately from degree computations.) In a different direction, the dichotomy between σ-varieties coming from group actions and *disintegrated* σ-varieties gives a basic instance of orthogonality.

Two contradictory notions of triviality for σ-varieties appear in the literature. Sometimes (see, for example, [5]) one says that (X, f) is *trivial* if it is isomorphic (as a σ-variety) to a σ-variety of the form (Y, id_Y). On the other hand, sometimes (see, for instance, the preprint version of this very paper [12]) one says that (X, f) is trivial if every type in $(X, f)^\sharp$ is *trivial* in the sense of its forking geometry. Since this latter property also goes under the name of *disintegratedness*, we use this term. Just as orthogonality is usually defined using the theory of forking, so is disintegratedness, but we give a geometric definition for σ-varieties.

Definition 2.20. Let (X, f) be a σ-variety over the difference field (K, σ). We say that (X, f) is *disintegrated* if for each natural number $n \in \mathbb{N}$ and each algebraically closed difference field (L, σ) extending (K, σ), each component of an $f^{\times n}$-skew-invariant subvariety $Z \subseteq X_L^{\times n}$ is a component of the intersection $\bigcap_{1 \le i \le j \le n} \pi_{i,j}^{-1} \pi_{i,j}(Z)$, where $\pi_{i,j} : X^{\times n} \to X^{\times 2}$ is the projection map $(x_1, \ldots, x_n) \mapsto (x_i, x_j)$.

With the next proposition we note that for any product of disintegrated σ-varieties, the algebraic relations are essentially binary.

PROPOSITION 2.21. *If $(X_1, f_1), \ldots, (X_n, f_n)$ is a finite sequence of disintegrated σ-varieties over the difference field (K, σ) where each X_i is an absolutely irreducible curve, then for every difference field (L, σ) extending (K, σ), every component Z of a skew-invariant subvariety of $\prod_{i=1}^n (X_i, f_i)$ is a component of $\bigcap_{1 \le i \le j \le n} \pi_{i,j}^{-1} \pi_{i,j}(Z)$, where $\pi_{i,j} : \prod_{i=1}^n X_i \to X_i \times X_j$ is the projection map $(x_1, \ldots, x_n) \mapsto (x_i, x_j)$.*

Proof. By Proposition 2.11 we may partition the components of this product so that the factors are nonorthogonal within each block of the partition but are orthogonal between blocks. By Proposition 2.18 we may assume that for every pair, we have $(X_i, f_i) \not\perp (X_j, f_j)$. In particular, for each $i \le n$, there is an (f_i, f_1)-skew invariant curve $Y_i \subseteq X_i \times X_1$, none of whose components is a vertical or horizontal line. Let $Y := \prod Y_i$ be regarded as an $(f_1, \ldots, f_n; f_1^{\times n})$-skew-invariant subvariety of $\prod_{i=1}^n X_i \times X_1^{\times n}$. Let $\rho : Y \to \prod_{i=1}^n X_i$ be the restriction of the projection map onto the first n-coordinates and $\eta : Y \to X_1^{\times n}$ the projection onto the last n coordinates.

Suppose now that $W \subseteq \prod_{i=1}^n X_i$ is an $(f_1, \ldots, f_n)$-skew-invariant variety and that $Z \subseteq W$ is an irreducible component. By Lemma 2.17, $Y(W)$ is a weakly $f_1^{\times n}$-skew-invariant variety and $W = W_{\mathrm{inv}} = (Y^{-1}(Y(W)_{\mathrm{inv}}))_{\mathrm{inv}}$. Thus, there is a component $V \subseteq (Y(W))_{\mathrm{inv}}$ with $Z \subseteq Y^{-1}(V)$.

Since (X_1, f_1) is disintegrated, V is a component of $\bigcap_{1 \leq i \leq j \leq n} \pi_{i,j}^{-1} \pi_{i,j} V$. Since Y respects the product decomposition, it follows that $Y^{-1}(V)$ is contained in $\bigcap_{1 \leq i \leq j \leq n} \pi_{i,j}^{-1} \pi_{i,j} Y^{-1}(V)$. As Z is a component of $Y^{-1}(V)$, the result follows. $\qquad\square$

As we noted above, σ-varieties coming from algebraic groups are never disintegrated.

PROPOSITION 2.22. *Let (K, σ) be a difference field, G a connected positive dimensional algebraic group over K, $\phi : G \to G^\sigma$ a dominant map of algebraic groups, and $g \in G(K)$ a K-rational point. Let $\tau_g : G \to G$ be defined by $\tau_g(x) := gx$. Let $f : G \to G^\sigma$ be given by $f := \phi \circ \tau_g$. Then (G, f) is not disintegrated.*

Proof. Let (L, σ) be a difference field extending (K, σ) and containing a solution h to the difference equation $\sigma(h) = \phi(h) \cdot \phi(g)^{-1}$. One checks immediately that the subvariety Γ of $G_L^{\times 3}$ defined by the equation $z = x \cdot h \cdot y$ is an irreducible, proper closed $f^{\times 3}$-skew-invariant variety that projects onto $G^{\times 2}$ for each pair of coordinate projections, witnessing that (G, f) is not disintegrated. $\qquad\square$

More generally, quotients of such σ-varieties and σ-varieties coming from actions of algebraic groups are never disintegrated. In a precise sense, the main theorem of [4], [6] asserts that the presence of a group action is the only obstruction to disintegratedness. Specializing to the case of σ-varieties of the form $(\mathbb{A}^1, f)$ over a difference field of characteristic zero, the main theorem of the first author's doctoral dissertation [11] characterizes the nondisintegrated σ-varieties as exactly those coming from monomials and Chebyshev polynomials.

Definition 2.23. For each positive integer $n \in \mathbb{Z}_+$, we write $P_n(x) := x^n \in \mathbb{Z}[x]$ for the standard n^{th} power monomial. We define $C_n(x) \in \mathbb{Z}[x]$ to be the unique polynomial satisfying the functional equation $C_n \circ \pi = \pi \circ P_n$ where $\pi : \mathbb{G}_m \to \mathbb{A}^1$ is given by $x \mapsto x + \frac{1}{x}$. We call C_n the nth *Chebyshev polynomial*. By a *negative Chebyshev polynomial* we mean a polynomial of the form $-C_n$ for some $n \in \mathbb{Z}_+$. In practice, when we speak of a power function, Chebyshev polynomial, or negative Chebyshev polynomial we mean one of degree at least two.

Remark 2.24. What we call the nth Chebyshev polynomial is sometimes called the nth Dickson polynomial. Moreover, our normalization differs from that of the Chebyshev polynomials of the first kind, $T_n(x)$, defined by the relation $T_n(\cos(\theta)) = \cos(n\theta)$, in that $C_n(x) = 2T_n(\frac{1}{2}x)$.

In the following theorem and throughout this paper we abuse notation by saying that f is a disintegrated polynomial (respectively, rational function) when we mean that $(\mathbb{A}^1, f)$ (respectively, $(\mathbb{P}^1, f)$) is a disintegrated σ-variety.

FACT 2.25 (Theorem 10 of [11]). *Over a difference field of characteristic zero, a polynomial of degree greater than one is disintegrated unless it is (possibly after base change) skew-conjugate to a Chebyshev polynomial, negative Chebyshev polynomial, or a monomial.*

Using Proposition 2.18, the observation that polynomials of different degrees are orthogonal, and Fact 2.25 we see that for a σ-variety $(\mathbb{A}^n, \Phi)$ where $\Phi : \mathbb{A}^n \to \mathbb{A}^n$ takes the form $(x_1, \ldots, x_n) \mapsto (f_1(x_1), \ldots, f_n(x_n))$, we may partition the coordinates so that $(\mathbb{A}^n, \Phi)$ is a product of pairwise orthogonal σ-varieties, each of which has the form $(\mathbb{A}^m, \Psi)$ where $\Psi(x_1, \ldots, x_m) = (g_1(x_1), \ldots, g_n(x_n))$ with the g_i's univariate polynomials for which exactly one of the following occurs:

- each g_i is linear,
- there is some $N \in \mathbb{Z}_+$ so that each g_i is skew-conjugate to $\pm C_N$ or P_N, or
- $(\mathbb{A}^m, \Psi)$ is disintegrated and the polynomials g_i are pairwise nonorthogonal.

The skew-invariant varieties for σ-varieties of the first two kinds are very easy to describe.

After base change, a σ-variety of the form $(\mathbb{A}^1, g)$ with g linear is skew-conjugate to $(\mathbb{A}^1, \mathrm{id})$. Clearly, the skew-invariant subvarieties of $(\mathbb{A}^m, \mathrm{id})$ are precisely those varieties that are defined over the fixed field. Thus, if $\Phi : \mathbb{A}^n \to \mathbb{A}^n$ is any dominant affine map, then, after a base change required to find an isomorphism of σ-varieties $\alpha : (\mathbb{A}^n, \Phi) \to (\mathbb{A}^n, \mathrm{id})$, the Φ-skew-invariant varieties are precisely the varieties of the form $\alpha^{-1} Y$, where $Y \subseteq \mathbb{A}^n$ is a variety defined over the fixed field. Which of these descend to Φ-skew-invariant varieties defined over our base field can be an interesting question best addressed through the Picard-Vessiot theory for difference equations. We do not pursue the matter here other than to spell out what happens in the category of algebraic dynamics.

Definition 2.26. Let K be a field of characteristic zero. A sequence $f_1, \ldots, f_n \in K[x]$ of linear polynomials over K is *independent* if one of the following conditions holds: (1) the numbers $f_1'(0), \ldots, f_n'(0)$ are multiplicatively independent or (2) the multiplicative group generated by $f_1'(0), \ldots, f_n'(0)$ has rank $n - 1$ and $f_j(x) = x + b$ for some $j \leq n$ and some $b \neq 0$

PROPOSITION 2.27. *Let K be a field of characteristic zero, let $f_1, \ldots, f_n \in K[x]$ be an independent sequence of linear polynomials over K, and define $\Phi : \mathbb{A}_K^n \to \mathbb{A}_K^n$ by $\Phi(x_1, \ldots, x_n) := (f_1(x_1), \ldots, f_n(x_n))$. Then there is some $a = (a_1, \ldots, a_n) \in \mathbb{A}^n(K)$ for which $\mathcal{O}_\Phi(a)$ is Zariski dense.*

Proof. In general, if K is an algebraically closed field, G is an algebraic group over K, $\mu : G \times X \to X$ is a morphism of varieties giving an action of G on X, and $g \in G(K)$ is any K-rational point, then we have an algebraic

dynamical system $(X, \mu(g, \cdot))$ given by the action of g on X. Let $H \subseteq G$ be the Zariski closure of the group generated by g, which is itself an algebraic group. Note that a subvariety $Y \subseteq X$ is $\mu(g, \cdot)$-invariant just in case it is H-invariant. Thus, the $\mu(g, \cdot)$-invariant varieties correspond exactly to the H-orbits. Specializing to the case that G is the affine group acting on $\mathbb{A}^n$ and $\mu(g, \cdot)$ is given by a sequence of univariate linear polynomials, it is easy to see that we may make a change of variables so that each such component has the form $f_i(x) = \lambda_i \cdot x$ or $f_j(x) = x + 1$. For the remainder of this calculation, we shall assume that the polynomials do have this form. The Zariski closure H of the group generated by g is then isomorphic to either $\mathbb{G}_m^r$ or $\mathbb{G}_m^r \times \mathbb{G}_a$, where r is the rational rank of the multiplicative group generated by the scalars λ_i and there is a $\mathbb{G}_a$ factor just in case at least one of the f_j is $x + 1$. For any point $a \in \mathbb{G}_m^n(K)$, the stabilizer of a in H is trivial. Hence, as long as we arrange for $a_i \neq 0$ when $f_i(x) = \lambda_i x$, the dimension of the Zariski closure of the $\mu(g, \cdot)$-orbit of $a = (a_1, \ldots, a_n)$ is $\dim(H) = r$ or $r + 1$. $\square$

PROPOSITION 2.28. *Let K be a field of characteristic zero and let $f_1, \ldots, f_n$ $\in K[x]$ be a sequence of polynomials of degree at least two such that each f_i is a power function, a Chebyshev polynomial, or a negative Chebyshev polynomial. Let $\Phi : \mathbb{A}_K^n \to \mathbb{A}_K^n$ be defined by*

$$\Phi(x_1, \ldots, x_n) := (f_1(x_1), \ldots, f_n(x_n)).$$

Then there is a point $a \in \mathbb{A}^n(K)$ for which $\mathcal{O}_\Phi(a)$ is Zariski dense.

Proof. In the case of the power functions of degree $N > 1$, a fairly routine argument with degrees shows that any irreducible skew-invariant subvariety of $(\mathbb{G}_m^g, x \mapsto x^N)$ is a translate of an algebraic group [8]. From the point of view of the model theory of difference fields, this result is a special case of the classification of definable groups [3]. Since the map $\pi : (\mathbb{G}_m, P_N) \to (\mathbb{A}^1, C_N)$ is a dominant map of σ-varieties, we see that any skew-invariant subvariety of $(\mathbb{A}^n, (f_1, \ldots, f_n))$, where each f_i is either P_N or C_N, pulls back to a weakly skew-invariant variety for $(\mathbb{A}^n, (P_N, \ldots, P_N))$ and thus comes from images of multiplicative translates of algebraic tori. In general, if each $f_i : \mathbb{A}^1 \to \mathbb{A}^1$ is merely (after base change) skew-conjugate to P_N or $\pm C_N$, then as with the linear polynomials, after base change, the skew-invariant varieties are precisely the images under the isomorphism with the standard polynomials of certain images of torsion translates of algebraic tori, but the question of which ones descend to skew-invariant varieties over our base field reduces to problems in difference Galois theory. Since it is easy to find points in $\mathbb{G}_m$ not contained in any proper algebraic subgroups, for example, take $a = (p_1, \ldots, p_n) \in \mathbb{G}_m^n(\mathbb{Q})$ where the p_i's are distinct primes, one sees that for dynamical systems given by sequences of power maps and Chebyshev polynomials, there are rational points with Zariski dense orbits. $\square$

Let us collect all of these observations into a single theorem in which we reduce the problem of describing skew-invariant subvarieties for σ-varieties on $\mathbb{A}^n$ given by sequences of univariate polynomials to the study of disintegrated polynomials.

Notation 2.29. If we are given a finite sequence of polynomials $f_1, \ldots, f_n$ and a subset $S \subseteq \{1, \ldots, n\}$, then we write $(\mathbb{A}^S, f_S)$ for the σ-variety $\prod_{i \in S}(\mathbb{A}^i, f_i)$.

THEOREM 2.30. *Suppose that (K, σ) is an algebraically closed difference field of characteristic zero and $f_1, \ldots, f_n$ is a sequence of nonconstant polynomials. Then there is a partition P of $\{1, \ldots, n\}$ so that for distinct S and T from P, $(\mathbb{A}^S, f_S) \perp (\mathbb{A}^T, f_T)$, implying that if $X \subseteq \mathbb{A}^n$ is a component of an $(f_1, \ldots, f_n)$-skew-invariant variety, then it is a product of components of f_S-skew-invariant varieties as S ranges through P and for each $S \in \mathsf{P}$ exactly one of the following is true:*

- *the polynomial f_i has degree one for each $i \in S$ and the f_S-skew-invariant varieties are obtained (after base change) by pullback from an isomorphism with $(\mathbb{A}^S, \mathrm{id})$ from the varieties defined over the fixed field;*
- *there is a number $N > 1$ so that each f_i is skew-conjugate to P_N or $\pm C_N$ for $i \in S$ and the f_S-skew-invariant varieties are obtained from algebraic tori; or*
- *all of the polynomials f_i are pairwise nonorthogonal and disintegrated for $i \in S$, implying that the irreducible f_S-skew-invariant varieties are components of intersections of pullbacks of (f_i, f_j)-skew-invariant curves in $\mathbb{A}^2$ and (i, j) ranges through S^2.*

Remark 2.31. A version of [Theorem 2.30](#) holds for rational functions in arbitrary characteristic. The first case must include purely inseparable maps and the second case must include Lattès maps and their additive analogues in positive characteristic (see [11]).

2.3. *From curves to polynomials.* We now convert the problem of describing (f, g)-skew-invariant curves to a question about polynomial compositional identities.

Notation 2.32. In what follows we work with an algebraically closed difference field (K, σ) of characteristic zero on which σ is an automorphism. When we speak of a polynomial f we mean a polynomial with coefficients from K. For the associated σ-variety, we may write $(\mathbb{A}^1, f)$ or in some cases $(\mathbb{P}^1, f)$. A rational function $g : \mathbb{P}^1 \to \mathbb{P}^1$ is a polynomial if ∞ is a totally ramified fixed point for g.

LEMMA 2.33. *Let f be a disintegrated polynomial. If C is a smooth, projective, irreducible curve, (C, h) is a σ-variety and $\gamma : (C, h) \to (\mathbb{P}^1, f)$ is a nonconstant morphism of σ-varieties, then $C = \mathbb{P}^1$ and h and γ are polynomials.*

Proof. The preimage $S := \gamma^{-1}(\{\infty\})$ of ∞ under γ is finite and totally invariant for h (that is, $h^{-1}(S) = S$). Thus, by an easy Riemann-Hurwitz argument (see Theorem 1.6 of [19]), we see that $C = \mathbb{P}^1$ and either h and γ are polynomials, or S has exactly two elements and $(\mathbb{P}^1, h)$ is isomorphic to $(\mathbb{P}^1, x \mapsto \frac{1}{x^{\deg(f)}})$. However, such a σ-variety cannot be disintegrated as the restriction of this map to $\mathbb{G}_m$ is an isogeny. $\qquad\square$

It follows from Lemma 2.33 that all (f, g)-invariant curves for f a disintegrated polynomial come from solutions to polynomial compositional identities.

PROPOSITION 2.34. *If f and g are disintegrated polynomials and $C \subseteq \mathbb{A}^2$ is an irreducible (f, g)-skew-invariant curve, then there are a polynomial h and polynomial morphisms of σ-varieties $\pi : (\mathbb{A}^1, h) \to (\mathbb{A}^1, f)$ and $\rho : (\mathbb{A}^1, h) \to (\mathbb{A}^1, g)$ for which C is parametrized by the map $t \mapsto (\pi(t), \rho(t))$. That is, there are polynomials h, ρ, and π satisfying the compositional equations $f \circ \pi = \pi^\sigma \circ h$ and $g \circ \rho = \rho^\sigma \circ h$:*

$$
\begin{array}{ccccc}
\mathbb{A}^1 & \xleftarrow{\ \pi\ } & \mathbb{A}^1 & \xrightarrow{\ \rho\ } & \mathbb{A}^1 \\
{\scriptstyle f}\downarrow & & {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle g} \\
\mathbb{A}^1 & \xleftarrow{\ \pi^\sigma\ } & \mathbb{A}^1 & \xrightarrow{\ \rho^\sigma\ } & \mathbb{A}^1.
\end{array}
$$

Proof. Passing to the closures in $\mathbb{P}^1 \times \mathbb{P}^1$, the projective curve $\overline{C}$ is (f, g)-skew-invariant. Let $\overline{h}$ be the restriction of (f, g) to $\overline{C}$. Let $\beta : C' \to \overline{C}$ be the normalization map. Since C' is a smooth curve and β is an isomorphism off a finite set, there is regular map $h : C' \to C'$ for which $\beta : (C', h) \to (\overline{C}, \overline{h})$ is a map of σ-varieties.

Let $\alpha_i : C \to \mathbb{A}^1$ be the projection map onto the ith coordinate for $i = 1$ or 2. Since the result is obvious if either projection map is constant, we shall assume that both α_1 and α_2 are nonconstant. By Lemma 2.33 applied to $\gamma = \alpha_1 \circ \beta$ (or $\gamma = \alpha_2 \circ \beta$), $C' = \mathbb{P}^1$ and h is a polynomial. Take $\pi := \alpha_1 \circ \beta$ and $\rho := \alpha_2 \circ \beta$. $\qquad\square$

Combining these observations we see that nonorthogonality between disintegrated polynomials is always witnessed by a solution to a system of polynomial compositional identities.

COROLLARY 2.35. *Given two disintegrated polynomials f and g, then $(\mathbb{A}^1, f) \not\perp (\mathbb{A}^1, g)$ if and only if there are a natural number M and nonconstant polynomials π, ρ and h for which $f^{\Diamond M} \circ \pi = \pi^{\sigma^M} \circ h$, and $g^{\Diamond M} \circ \rho = \rho^{\sigma^M} \circ h$:*

$$
\begin{array}{ccccc}
\mathbb{A}^1 & \xleftarrow{\ \pi\ } & \mathbb{A}^1 & \xrightarrow{\ \rho\ } & \mathbb{A}^1 \\
{\scriptstyle f^{\Diamond M}}\downarrow & & {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle g^{\Diamond m}} \\
\mathbb{A}^1 & \xleftarrow{\ \pi^{\sigma^M}\ } & \mathbb{A}^1 & \xrightarrow{\ \rho^{\sigma^M}\ } & \mathbb{A}^1.
\end{array}
$$

Proof. If $(\mathbb{A}^1, f) \not\perp (\mathbb{A}^1, g)$, then, possibly after base change, we find an (f, g)-skew-invariant curve $C \subseteq \mathbb{A}^2$ none of whose components is horizontal or vertical. Taking M sufficiently divisible, we find a component C' of C that is $(f^{\Diamond M}, g^{\Diamond M})$-skew-invariant (with respect to σ^m). The existence of π, ρ, and h now follows from Proposition 2.34.

In the other direction, the curve $C := (\pi, \rho)(\mathbb{A}^1) \subseteq \mathbb{A}^2$ witnesses that $(\mathbb{A}^1, f^{\Diamond M}) \not\perp (\mathbb{A}^1, g^{\Diamond M})$ as σ^M-varieties. The curve $C' := \bigcup_{j=0}^{M-1}(f^{\Diamond j}, g^{\Diamond j})(C)^{\sigma^{-j}}$ witnesses that $(\mathbb{A}^1, f) \not\perp (\mathbb{A}^1, g)$. □

2.4. *Decompositions and actions.*

In this section we analyze the identities of Corollary 2.35 through the combinatorics of decompositions of polynomials.

Definition 2.36. A polynomial f is *indecomposable* if $\deg(f) \geq 2$, and it cannot be written as a composition $f = g \circ h$ of two nonlinear polynomials g and h.

A finite sequence $\vec{f} := (f_k, \ldots, f_1)$ of polynomials f_i is a *decomposition* of a polynomial f if $f = f_k \circ \cdots \circ f_1$ and each f_i is indecomposable.

Remark 2.37. What we call decompositions are called "complete decompositions" in the literature (see, for example, [21]). Note how we index the factors of a decomposition: decreasing from left to right, so that the first factor to be applied to an input has index 1.

Remark 2.38. Induction on degree shows that every nonlinear polynomial has a decomposition. Linear polynomials are compositional units. As such, if L is a linear polynomial, then we write L^{-1} for its compositional inverse. More concretely, if $L(x) = ax + b$, then $L^{-1}(x) = \frac{1}{a}x - \frac{b}{a}$.

Definition 2.39. The decompositions $(f_k, \ldots, f_1)$ and $(g_k, \ldots, g_1)$ are *linearly equivalent* if there are linear polynomials $L_{k-1}, \ldots, L_1$ 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 $g_1 = L_1^{-1} \circ f_1$.

Polynomials a and b are *linearly related* if there are linear L and M such that $L \circ a \circ M = b$.

If $\vec{f}$ and $\vec{g}$ are linearly equivalent, then they are decompositions of the same polynomial. Linear equivalence, as the name suggests, is an equivalence relation. Corresponding factors of linearly equivalent decompositions are linearly related.

Definition 2.40. The linear-equivalence class of a decomposition $\vec{f}$ is denoted by $[\vec{f}]$. For a polynomial f, LE_f is the set of linear-equivalence classes of decompositions of f.

Not all decompositions of a polynomial are linearly equivalent; for example, (x^2, x^3+x) and (x^3+2x^2+x, x^2) are both decompositions of $(x \cdot (x^2+1))^2$.

Ritt's theorem [15] gives a precise sense in which all decompositions of a polynomial may be obtained from one given decomposition.

Definition 2.41. A *Ritt polynomial* is an indecomposable polynomial of one of the following kinds:

- Monomial: $P_p(x) := x^p$, p a prime;
- Chebyshev: $C_p(x)$, p an odd prime;
- $x^k \cdot u(x^\ell)^n$, where $k \neq 0$, $\gcd(k, \ell) = 1$, $\gcd(k, n) = 1$, $u(0) \neq 0$, u is monic nonconstant, and at least one of ℓ and n is greater than one.

The following identities involving Ritt polynomials are the *basic Ritt identities*:

- $P_p \circ P_q = P_q \circ P_p$ for prime $p \neq q$,
- $C_p \circ C_q = C_q \circ C_p$ for odd prime $p \neq q$,
- $P_p \circ (x^k \cdot u(x^{\ell p})^n) = (x^k \cdot u(x^\ell)^{pn}) \circ P_p$ for prime p.

Remark 2.42. These notions are closely related but not identical to "Ritt moves" and "Ritt neighbors" in [21].

Definition 2.43. If $\vec{g}$ and $\vec{f}$ are two decompositions of the same polynomial, we say that $\vec{g}$ is *obtained from $\vec{f}$ by a Ritt swap at i* if there are linear polynomials L, M, and N such that

$$g_i := S \circ N^{-1}, \quad g_{i+1} = L \circ R, \quad g_j := f_j \quad \text{for } j \neq i, i+1,$$

and $(L^{-1} \circ f_{i+1} \circ M) \circ (M^{-1} \circ f_i \circ N) = R \circ S$ is a basic Ritt identity.

An indecomposable polynomial f is *swappable* if it is linearly related to a Ritt polynomial.

Remark 2.44. The compositional identity $C_2 \circ C_p = C_p \circ C_2$ is not a basic Ritt identity, but (C_p, C_2) can be obtained by a Ritt swap at 1 from (C_2, C_p) as follows. As C_p is an odd polynomial, it is of the form $x \cdot u(x^2)$, and $C_2(x) = x^2 - 2 = L \circ P_2$ where $L(x) = x - 2$. Now taking $M = N = \text{id}$ makes $(L^{-1} \circ C_2 \circ M) \circ (M^{-1} \circ C_p \circ N)$ look like the left side of a basic Ritt identity. This is pursued in great detail in Section 3.

Remark 2.45. While it may be possible to obtain many different decompositions from the same $\vec{f}$ by a Ritt swap at the same i by choosing different linear witnesses L, M, and N, we show (see page 123) that all decompositions so obtained are linearly equivalent. This invariance result is also proved in [21] and is already implicit in Ritt's work.

Remark 2.46. The term "swap" should suggest that when a decomposition is obtained from another via a Ritt swap, then the factors involved swap places. However, a Ritt swap arising from a basic Ritt identity of the third kind is not really a swap in that one of the factors, linearly related to $x^k \cdot u(x^p)$, not only

switches places with the monomial but also "becomes" a different polynomial, linearly related to $x^k \cdot u(x)^p$.

Remark 2.47. We depart from [15] in requiring Ritt polynomials to be monic. An easy computation verifies that this has no effect on the meaning of "Ritt swap" and the truth of Ritt's Theorem below.

FACT 2.48 (Ritt, [15]). *Over* $\mathbb{C}$, *any two decompositions of the same polynomial have the same number of factors. 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.*

Ritt's Theorem may be stated loosely as "decompositions of polynomials are unique up to permutations," and indeed it is tempting to look for an action by the symmetric group, identifying the Ritt swap at i with the transposition $\tau_i := (i \ \ i+1) \in \mathrm{Sym}_k$ in the symmetric group on k elements.

Since often nothing can be obtained from $\vec{f}$ by a Ritt swap at i (for example, when one of the factors f_i and f_{i+1} is not swappable), at best this is a partial action. In light of Remark 2.45, Ritt swaps can only act on decompositions up to linear equivalence, that is on LE_f. The next two results show that this action is well defined when it is defined.

LEMMA 2.49. *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 $R_{k-1}, \ldots, R_1$, L, M, and N be linear polynomials witnessing our hypotheses. That is, the R's witness that $\vec{h}$ is linearly related to $\vec{f}$:

$$h_k = f_k \circ R_{k-1}, \quad h_j = R_j^{-1} \circ f_j \circ R_{j-1} \quad \text{for } 1 < j < k, h_1 = R_1^{-1} \circ f_1.$$

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 $j \le k$. To simplify the notation, we define $R_k(x) = R_0(x) = x$.

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

With the following theorem, whose proof is delayed to page 123, we show that the action of Ritt swaps on linear equivalence classes of decompositions is well defined. Stronger versions of this result are obtained in [21] and [15].

THEOREM 2.50. *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 $\vec{g}$.*

In the symmetric group, the adjacent transpositions $\tau_i := (i,\, i+1)$ have order 2 and satisfy the braid relations $\tau_i \tau_{i+1} \tau_i = \tau_{i+1} \tau_i \tau_{i+1}$ for all i and $\tau_i \tau_j = \tau_j \tau_i$ for $j \neq i \pm 1$. While Ritt swaps do satisfy the braid relations, they do not quite have order two. We formalize this symmetric group-like action via Ritt swaps as an action of a certain monoid.

Definition 2.51. Let RM_k be the free monoid on the $(k-1)$ generators $t_1, \ldots, t_{k-1}$. The *permutation represented by* a word $t_{a_r} \cdots t_{a_2} t_{a_1}$ in RM_k is the product $\tau_{a_r} \cdots \tau_{a_2} \tau_{a_1} \in \mathrm{Sym}_k$.

The action $\star$ of RM_k on $\mathrm{LE}_f^* := \mathrm{LE}_f \cup \{\infty\}$ is defined by

- $t_i \star [\vec{f}]$ is the linear equivalence class of a decomposition obtained from $\vec{f}$ by a Ritt swap at i, if one exists;
- otherwise, $t_i \star [\vec{f}] := \infty$;
- $t_i \star \infty = \infty$ for all i.

For $w \in \mathrm{RM}_k$ and $[\vec{f}] \in \mathrm{LE}_f$, we say that $w \star [\vec{f}]$ is *defined* if $w \star [\vec{f}] \neq \infty$. We often abuse notation writing $w \star \vec{f} = \vec{g}$ for $w \star [\vec{f}] = [\vec{g}]$.

With the following theorem, whose proof is completed on page 142, we show that Ritt swaps satisfy the braid relations and that t_i has order 2 except when $t_i \star w = \infty$. The first two parts of Theorem 2.52 are immediate given Theorem 2.50, but the last is not so easy.

THEOREM 2.52. *For any $[\vec{f}] \in \mathrm{LE}_f$ and $i < k$,*

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

With these identities, a purely combinatorial analysis yields (see Section 5) normal forms for words in the Ritt monoid, roughly corresponding to insert-sort and to merge-sort. That is, for each $w \in \mathrm{RM}_k$, we find another word w' of a special form, representing the same permutation and such that $w' \star \vec{f}$ is defined and equal to $w \star \vec{f}$ whenever $w \star \vec{f}$ is defined. For example, w' is the empty word when $w = t_i t_i$. This implies 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 $w \star [\vec{f}] = w' \star [\vec{f}]$ (see Corollary 5.12). This also provides an invaluable explicit computational tool for the rest of the paper.

Remark 2.53. A stronger version of Corollary 5.12, that the polynomial f and the sequence of degrees of the factors f_i already determine the linear

equivalence class of the decomposition $\vec{f}$, is shown in [21]. It seems that our stronger Theorem 2.52 is not a simple consequence of the work in [21]. The canonical forms in the present paper are substantially different from those in [21] and are better suited to our purposes.

2.5. *Skew-twists.* Recall that the purpose of studying decompositions is to characterize polynomial identities in Proposition 2.34:

$$
\text{(1)} \qquad
\begin{array}{ccc}
\mathbb{A}^1 & \xrightarrow{\ f\ } & \mathbb{A}^1 \\
{\scriptstyle \pi} \uparrow & & \uparrow {\scriptstyle \pi^\sigma} \\
\mathbb{A}^1 & \xrightarrow{\ g\ } & \mathbb{A}^1
\end{array}
$$

for disintegrated polynomials f and g. We eventually show that all such identities come from those where π is indecomposable or linear. For indecomposable π, these identities are either rare exceptions characterized in Proposition 5.23, or *single skew-twists* where π is an initial compositional factor of f, and also a terminal compositional factor of g twisted by σ. The rest of this section is devoted to the study of sequences of single skew-twists.

Definition 2.54. The decomposition $(f_1^\sigma, f_k, \ldots, f_2)$ is called *the single-skew-twist* of the decomposition $\vec{f} := (f_k, \ldots, f_2, f_1)$ and denoted $\phi \star \vec{f}$. (Here, 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.

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." That is, f is a skew-twist of g if there are $f = f_0, f_1, \ldots, f_n = g$ such that each f_i is a single-skew-twists of f_{i+1}, or vice versa.

To undo what ϕ does, we define $\beta \star \vec{f} := (f_{k-1}, \ldots, f_1, f_k^{(\sigma^{-1})})$. (Here, β stands for "back.")

When $n < k$ and $\vec{g} = \phi^n \star \vec{f}$, we call g a *plain skew-twist* of f.

Remark 2.55. A polynomial may have several single-skew-twists, coming from different decompositions. In composing a correspondence from f to g with one from g to h, both of which come from single skew-twists, the decompositions of g used to represent the skew-twists may differ. Thus, to describe correspondences obtained from sequences of skew-twists, we need to keep track of decompositions of intermediate polynomials.

Definition 2.56. For a given positive integer $k \in \mathbb{Z}_+$, the *skew-twist monoid*, ST_k, is the free monoid generated by the symbols ϕ, β, $t_1, \ldots, t_{k-1}$. If $\vec{f} = (f_k, \ldots, f_1)$ is a decomposition of a polynomial f and $w := w_n \cdots w_2 w_1 \in \mathrm{ST}_k$ where each w_i is a generator, then a sequence of decompositions $\vec{f} = \vec{f}^0, \vec{f}^1, \ldots, \vec{f}^n$ is *a witnessing sequence for* $w \star \vec{f}$ if for each j,

- if $w_j = t_i$, then $[\vec{f}^{j+1}] = w_j \star [\vec{f}^j]$; and
- if w_j is ϕ or β, then $\vec{f}^{j+1} = w_j \star \vec{f}^j$ in the sense of Definition 2.54.

The correspondence $\mathcal{A}$ encoded by this witnessing sequence is the composite of the curves $\mathcal{B}_n \circ \cdots \circ \mathcal{B}_1$ where

- if $w_j = t_i$ for some i, then $\mathcal{B}_j = \Delta_{\mathbb{A}^1}$ is the graph of the identity map on $\mathbb{A}^1$;
- if $w_j = \phi$, then $\mathcal{B}_j$ is the graph of f_1^j; and
- if $w_j = \beta$, then $\mathcal{B}_j$ is the converse relation of the graph of f_1^{j+1}.

We also say that $\mathcal{A}$ is *a correspondence encoded by $w \star \vec{f}$*.

While the witnessing sequence uniquely determines the correspondence, w and $\vec{f}$ do not uniquely determine the witnessing sequence because Ritt swaps are only defined up to linear equivalence, and even linearly equivalent decompositions may produce different single-skew-twists. We define skew-linear-equivalence and then formalize an action of the skew-twists monoid.

Remark 2.57. Suppose that $w = vu \in \mathrm{ST}_k$ and $w \star \vec{f} = \vec{h}$ is defined. Let $\{\vec{f}^j\}$ be a witnessing sequence for this, let $\mathcal{D}$ be the correspondence from f to h encoded thereby, and let $\vec{g}$ be the element of this sequence coming from $u \star \vec{f}$. Then $\mathcal{D} = \mathcal{B} \circ \mathcal{A}$, where $\mathcal{A}$ and $\mathcal{B}$ are the curves encoded by the two witnessing sequences $(\vec{f}, \ldots, \vec{g})$ and $(\vec{g}, \ldots, \vec{h})$, respectively.

Definition 2.58. 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 $(L^\sigma \circ f_k, f_{k-1}, \ldots, f_2, f_1 \circ L^{-1})$.

Remark 2.59. Skew-linear-equivalence is an equivalence relation. Skew-linearly-equivalent decompositions may be decompositions of different, but always skew-conjugate, polynomials. Indeed,

$$\vec{f} \mapsto \vec{g} := (L^\sigma \circ f_k, f_{k-1}, \ldots, f_2, f_1 \circ L^{-1})$$

is a bijection between decompositions of f and decompositions of $g := L^\sigma \circ f \circ L^{-1}$, and this bijection respects linear equivalence.

Definition 2.60. Let SE_f be the set of skew-linear-equivalence classes of decompositions of skew-twists of f. We write $[[\vec{f}]]$ for the skew-linear equivalence class of $\vec{f}$.

The action $\star$ of ST_k on $\mathrm{SE}_f^* := \mathrm{SE}_f \cup \{\infty\}$ is given by

- t_i still acts by the Ritt swap at i as in Definition 2.51;
- $\phi \star [[f_k, \ldots, f_1]] := [[f_1^\sigma, f_k, \ldots, f_2]]$ and $\phi \star \infty = \infty$;
- $\beta \star [[\vec{f}]] := [[f_{k-1}, \ldots, f_1, f_k^{(\sigma^{-1})}]]$ and $\beta \star \infty = \infty$.

For $w \in \mathrm{ST}_k$ and $[[\vec{f}]] \in \mathrm{SE}_f$, we say that $w \star [\vec{f}]$ is *defined* if $w \star [\vec{f}] \neq \infty$.

LEMMA 2.61.

(1) *Ritt swaps are well defined up to skew-linear-equivalence.*

(2) *Single skew-twists are well defined up to skew-linear-equivalence.*
(3) *Suppose that $w \in \mathrm{ST}_k$ and $\vec{f}$ is a decomposition of a polynomial f. Suppose that $\vec{g}$ and $\vec{h}$ are witnessing sequences for $w \star \vec{f}$ with corresponding encoded correspondences $\mathcal{A}$ and $\mathcal{B}$ between $(\mathbb{A}^1, f)$ and $(\mathbb{A}^1, g)$ (and $(\mathbb{A}^1, f)$ and $(\mathbb{A}^1, h)$, respectively). Then there is a linear L with $h = L^\sigma \circ g \circ L^{-1}$ and $\mathcal{B} = L \circ \mathcal{A}$. That is, h is skew-conjugate to g and the correspondence is off by the same linear factor.*

Proof. The proof of the first two parts serves as the induction (on the length of w) step for the proof of the last part. For the first part, note that if a decomposition $(h_k, h_{k-1}, \ldots, h_2, h_1)$ of f is obtained from $\vec{f}$ by a Ritt swap at i, then the decomposition $(L^\sigma \circ h_k, h_{k-1}, \ldots, h_2, h_1 \circ L^{-1})$ of $g := L^\sigma \circ f \circ L^{-1}$ is obtained from $\vec{g} := (L^\sigma \circ f_k, f_{k-1}, \ldots, f_2, f_1 \circ L^{-1})$ by a Ritt swap at i.

For the second part, take

$$\vec{g} := (L^\sigma \circ f_k \circ L_{k-1}^{-1}, L_{k-1} \circ f_{k-1} \circ L_{k-2}^{-1}, \ldots, L_2^{-1} \circ f_2 \circ L_1, L_1^{-1} \circ f_1 L^{-1})$$

skew-linearly equivalent to $\vec{f}$. The same linear factors, reindexed, witness that single skew-twists of $\vec{g}$ are skew-linearly equivalent to the corresponding single skew-twists of $\vec{f}$.

For the third part, let n be the length of w and take witnessing sequences $\vec{g}^j$ and $\vec{h}^j$ for $w \star \vec{f} = \vec{g}$ and for $w \star \vec{f} = \vec{h}$; so $\vec{g}^0 = \vec{f} = \vec{h}^0$ and $\vec{g}^n = \vec{g}$ and $\vec{h}^n = \vec{h}$. We induct on n, strengthening the induction hypothesis from $h = L^\sigma \circ g \circ L^{-1}$ to $\vec{h} = L^\sigma \circ \vec{g} \circ L^{-1}$.

Let $v := w_{n-1} \cdots w_1$. and let $\mathcal{A}_0$ and $\mathcal{B}_0$ be the curves encoded by $v \star \vec{f}$ via these witnessing sequences. By induction hypothesis and the first two parts, there is a linear factor L such that $\vec{h}^{n-1}$ is linearly equivalent to $L^\sigma \circ \vec{g} \circ L^{-1}$ and $\mathcal{B} = L \circ \mathcal{A}$. If w_n is a Ritt swap, the same L works: look at the proof of the first part of this lemma to prove the first part of the induction hypothesis, and note that the curve encoded is the same for v and w to prove the second part of the induction hypothesis. If w_n is a single skew-twist, then composing the graph of the first or last factor of $\vec{h}$ with $L \circ \mathcal{A}$ cancels L and introduces a new linear factor, one of the witnesses of the linear equivalence of $\vec{h}^{n-1}$ and $L^\sigma \circ \vec{g} \circ L^{-1}$. $\qquad\square$

Remark 2.62. The definition of witnessing sequences and encoded correspondences allows linear equivalence and skew-conjugacy in some cases but not in others. Because of this inconsistency, it is safest to artificially reintroduce the linear factor L at the end, as we do in Theorems 6.22, 6.24 and 6.26.

COROLLARY 2.63. *If two correspondences $\mathcal{A}$ and $\mathcal{B}$ between the polynomials f and g are both encoded by $w \in \mathrm{ST}_k$, then they are off by a skew-symmetry L of g, that is, $\mathcal{B} = L \circ \mathcal{A}$ and $L^\sigma \circ g \circ L^{-1} = g$.*

Our characterization of correspondences encoded by words in ST_k comes from the canonical form for such words, obtained in Proposition 2.70 and Lemmata 6.16 through 6.18. Here we state an imprecise nontechnical version as motivation.

Remark 2.64. This is a motivational imprecise nontechnical version of Proposition 2.70 and Lemmata 6.16 through 6.18.

Any word $w \in \mathrm{ST}_k$ such that $w \star \vec{f}$ is defined is *equivalent* to a word of the form $\phi^{Nk} w_0$ or $\beta^{Nk} w_0$ where the length of w_0 is bounded by a constant depending only on the degree of f.

Any (f, g)-skew-invariant curve coming from skew-twists is a composition of the graph of $f^{\Diamond N}$ for some $N \in \mathbb{N}$ with a correspondence both of whose degrees are bounded by $2 \cdot \deg f$; or a composition of a correspondence both of whose degrees are bounded by $2 \cdot \deg f$ with the converse relation to the graph of $g^{\Diamond N}$ for some $N \in \mathbb{N}$.

We define equivalence for words in the skew-twist monoid so as to make the second part of Remark 2.64 a consequence of the first. Thus, it must take into account the curves encoded by the words in the monoid, but it need not keep track of their strictly skew-pre-periodic components.

Definition 2.65. Fix $v, w \in \mathrm{ST}_k$ and a decomposition $\vec{f} = (f_k, \ldots, f_1)$. We say that v and w are *equivalent with respect to* $\vec{f}$ and write $v \approx_{\vec{f}} w$ if $v \star [[\vec{f}]] = w \star [[\vec{f}]]$ and there are witnessing sequences $(\vec{g}^j)$ and $(\vec{h}^j)$ for $v \star \vec{f}$ and $w \star \vec{f}$, respectively so that the final $\vec{g}^n$ and $\vec{h}^n$ are decompositions of the same polynomial g, and $(\mathcal{A}_v)_{\mathrm{inv}} = (\mathcal{A}_w)_{\mathrm{inv}}$ for the curves $\mathcal{A}_v$ and $\mathcal{A}_w$ encoded by v (respectively, w) via $(\vec{g}^j)$ (respectively, $(\vec{h}^j)$).

When $v \approx_{\vec{f}} w$ for all $\vec{f}$, we write $v \approx w$ and say that the two words are *equivalent*.

This notion is weaker than the purely syntactic one in Definition 5.3 of $v \simeq w$ for $v, w \in \mathrm{RM}_k$.

Lemma 2.66.

(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* ST_k, $w\phi^k \approx \phi^k w$ *and* $w\beta^k \approx \beta^k w$.
(4) $t_i\phi \approx \phi t_{i+1}$ *for* $i < k-1$ *while* $t_i\beta \approx \beta t_{i-1}$ *for* $i > 1$

Proof. (1) The (f, f)-skew-invariant correspondence $\mathcal{A}_{\beta\phi}$ encoded by $\beta\phi$ is defined by $f_1(x) = f_1(y)$. The diagonal is one of its irreducible components, is (f, f)-skew-invariant, and is equal to the image $(f, f)(\mathcal{A}_{\beta\phi})$ of the whole curve. Thus, $\beta\phi \approx_{\vec{f}} \mathrm{id}$ for any decomposition $\vec{f}$.

(2) This is an immediate consequence of Lemma 2.17, Lemma 2.61 and Corollary 2.63.

(3) Since $\phi^k \star \vec{f} = \vec{f}^\sigma$ and $\beta^k \star \vec{f} = \vec{f}^{(\sigma^{-1})}$, it is clear that ϕ^k and β^k commute with Ritt swaps. Part (1) ensures that they commute with ϕ and β.

(4) After a shift, the same two factors participate in the Ritt swap on the two sides of each equation. $\qquad\square$

LEMMA 2.67. *For all $w \in \mathrm{ST}_k$, there is some $u \in \mathrm{ST}_k$ 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 on the left. $\qquad\square$

Remark 2.68. Here is the geometry behind this bit of combinatorics. When $w \star \vec{f}$ is defined, the correspondence $\mathcal{A}$ encoded is (an irreducible component of) the fiber product of a diagram

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

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

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

or

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

where we know one arrow, $g^{\Diamond N}$, exactly, and the other arrow is a sequence of plain skew-twists.

In most cases, it is also possible to bring together all the ϕ's in u in Lemma 2.67, and then F must be skew-compositional power of g composed with (a not necessarily indecomposable) factor of g. However, it is not always possible to do this. Consider the following example:

$$\phi t_1 \phi \star (x \cdot (x^5+1), x^5) = \phi t_1 \star (x^5, x \cdot (x^5+1)) = \phi \star (x \cdot (x+1)^5, x^5) = (x^5, x \cdot (x+1)^5).$$

The encoded correspondence, defined by $y = x^{25}$, is not a compositional power of $x^5 \cdot (x^5)^4$ in any sense. The trouble is that Lemma 2.66 does not give a way to simplify $t_{k-1}\phi$ and βt_{k-1}. We deal with this issue by introducing what we call the *border guard monoid* whose action on decompositions leaves the leftmost factor fixed, though possibly altering it via Ritt swaps in the sense of Remark 2.46.

Definition 2.69. For a fixed positive integer $k \geq 2$, BG_k is the free monoid on the symbols ψ, γ, and $t_1, \ldots t_{k-2}$. (There are no generators of the form t_i if $k = 2$.) Regard BG_k as a submonoid of ST_k by mapping t_i to t_i, ψ to $(t_{k-1}\phi)$, and γ to (βt_{k-1}).

The action of BG_k on SE_f is the restriction of the action of ST_k. More concretely,

$$\psi \star (f_k, \ldots, f_1) = t_{k-1}\phi \star (f_k, \ldots, f_1)$$
$$= t_{k-1} \star (f_1^{\sigma}, f_k, \ldots, f_2) = (\widehat{f_k}, \widehat{f_1^{\sigma}}, \ldots, f_2),$$
$$\gamma \star (f_k, f_{k-1}, \ldots, f_1) = \beta t_{k-1} \star (f_k, \ldots, f_1)$$
$$= \beta \star (\widehat{f_{k-1}}, \widehat{f_k}, \ldots, f_1) = (\widehat{f_k}, \ldots, f_1, \widehat{f_{k-1}}^{\sigma^{-1}}).$$

We use BG_k to establish the bounds in Remark 2.64. Indeed, finding the word w' of the following proposition goes a long way towards producing the short word w_0 of Remark 2.64.

PROPOSITION 2.70. *Any word w in ST_k is equivalent to $\phi^N w'$ or to $\beta^N w'$ for some $N \in \mathbb{N}$ and some word $w' \in \mathrm{BG}_k$.*

Proof. We take $w \in \mathrm{ST}_k$, start from the right, and move to the left. At every step, we have a word $w_{\mathrm{bad}}\beta^a\phi^b w_{\mathrm{good}}$ with $w_{\mathrm{bad}} \in \mathrm{ST}_k$ and $w_{\mathrm{good}} \in \mathrm{BG}_k$. Working by induction on the length of w_{bad}, it is clearly sufficient to prove that if s is a generator of ST_k, then there are natural numbers a' and b' and some $u \in \mathrm{BG}_k$ with $s\beta^a\phi^b \approx \beta^{a'}\phi^{b'}u$. If $s = \beta$, then we may take $a' := a + 1$, $b' = b$, and u the empty word. If $s = \phi$ and $a = 0$, then we take $a' = 0$, $b' = b+1$, and u the empty word. If $s = \phi$ and $a > 0$, then then we take $a' = a - 1$, $b' = b$, and u the empty word.

We work by induction on $(a + b)$ for the case that $s = t_i$ for some $i < k$. In the base case of $a = b = 0$, if $i < k - 1$, then we may take $a' = b' = 0$ and $u = t_i$. For $i = k - 1$, we note $t_{k-1} \approx \phi\beta t_{k-1} = \phi\gamma$, so that we may take $a' = 0$, $b' = 1$, and $u = \gamma$. 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 $t_{k-1}\beta^{a-2}\phi^b$. $\square$

Remark 2.71. It is sometimes helpful to think of SE_f as a bunch of indecomposable factors arranged in a circle, rather than a line, with f_k standing

next to f_1. In that spirit, both γ and ψ act by a Ritt swap between these two. To be more precise, for any decomposition $\vec{f}$, the following three are equivalent:

- $\gamma \star \vec{f}$ is defined,
- $\psi \star \vec{f}$ is defined,
- $t_1 \star (f_1^\sigma, f_k)$ is defined.

COROLLARY 2.72. *If $w \in \mathrm{ST}_k$ and $w \star \vec{f}$ is defined and f_k is not swappable, then the correspondence encoded by $w \star \vec{f}$ is already encoded by $\phi^N \star \vec{g}$ or $\beta^N \star \vec{g}$ for some $N \in \mathbb{N}$ and some decomposition $\vec{g}$ of f.*

Proof. Get the $\phi^N w' \approx w$ or $\beta^N w' \approx w$ from Proposition 2.70, with $w' \in \mathrm{BG}_k$. Because f_k is not swappable, $w' \star \vec{f}$ is only defined if $w' \in \mathrm{RM}_k$. Let $\vec{g} := w' \star \vec{f}$. Then the correspondence encoded by $w \star \vec{f}$ is the same as the one encoded by $\phi^N \star \vec{g}$, or $\beta^N \star \vec{g}$, as the case may be.   □

Of course, the hypothesis that f_k is the special unswappable factor is purely artificial.

COROLLARY 2.73. *If $w \in \mathrm{ST}_k$ and $w \star \vec{f}$ is defined and f_i is not swappable for some i, then the correspondence encoded by $w \star \vec{f}$ is already encoded by $\phi^N u \phi^i \star \vec{f}$ or $\beta^N u \phi^i \star \vec{f}$ for some $N \in \mathbb{N}$ and some $u \in \mathrm{RM}_k$.*

Proof. Recall that $w \beta^i \phi^i \approx w$. Since the kth factor f_i^σ of $\vec{h} := \phi^i \star \vec{f}$ is unswappable, Corollary 2.72 applies to $(w \beta^i) \star \vec{h}$, with the sequence u of Ritt swaps giving the potentially necessary new decomposition $\vec{g}$ in the statement of that corollary.   □

Remark 2.74. The hypothesis that f_k is not swappable is unnecessarily strong. Requiring merely that $t_1 \star (g_1^\sigma, g_k)$ is not defined where $\vec{g} := u \star \vec{f}$ for some $u \in \mathrm{RM}$ would suffice. Many explicit examples satisfying this requirement appear in a previous draft of this paper [12] related to the concept of a "crack."

Outline of the technical Sections 3–6

The next four sections constitute technical proofs of the results described in Section 2. Three of the four sections are devoted to refinements of Ritt's Fact 2.48, and the last one uses these refinements to obtain the desired characterization of skew-invariant curves.

The characterization of linear relatedness between Ritt polynomials in our Section 3 is also carried out in [21] and is implicit in [15]. We include our analysis because we use many of the intermediate results in the two following Sections 4 and 5.

In Section 4, we describe a nearly unique way to write a polynomial as a composition of *clusters*. One of our two kinds of clusters is the same as one of the two kinds of blocks in [21], but our C-free clusters are nothing like their

monomial blocks. Again, similar technical issues come up for us and for them, such as the fact that no more than one quadratic factor may cross a boundary between clusters in the same direction. Our first use of these clusters is to prove our fundamental Theorem 2.52 for the Ritt monoid action, that $t_i t_{i+1} t_i \star \vec{f}$ is defined if and only of $t_{i+1} t_i t_{i+1} \star \vec{f}$ is defined. While it follows immediately from [21] that the two are equal when defined, it is not clear to us whether our stronger result follows from their work.

In Section 5, we use the fact that the action of the Ritt monoid on linear-equivalence classes of decompositions factors through the "braid monoid" to find canonical forms for sequences of Ritt swaps, roughly corresponding to insert-sort and to merge-sort. To the best of our understanding, our results on canonical forms do not follow easily from [21], where different canonical forms are used to obtain tighter bounds on the number of Ritt swaps necessary to obtain one decomposition from another. Applying the second canonical form to a clustering produces particularly strong results. We end that section with a characterization (see Proposition 5.23) of those rare polynomial identities $\pi^\sigma \circ f = g \circ \pi$ that have nothing to do with skew-twists. A slight weakening of it follows immediately from [21], and the full version can be deduced with a little more work.

Section 6 combines all of our technical tools and finally characterizes skew-invariant curves. In Section 6.1, we introduce more generators into our monoids in order to encode correspondences coming from Proposition 5.23 rather than from skew-twists. Within this formalism, we describe precisely how the correspondences arising from Proposition 5.23 interact (commute) with those arising from skew-twists. In Section 6.2, we then combine our work on clusterings with our understanding of skew-twists in order to obtain a characterization of correspondences encoded by $w \star \vec{f}$ for $w \in \mathrm{ST}_k$ for those rare $\vec{f}$ that are not subject to Corollary 2.73. In Section 6.3, Theorem 6.22 is a complete, precise, and technical characterization of (f, g)-invariant curves for disintegrated polynomials f and g. The technical conclusion of Theorem 6.22 becomes much more readable in the special case of (h, h)-invariant curves. It is stated in Theorem 6.24 and then used to obtain a more readable but less tight characterization for the general case in Theorem 6.26.

Notation 2.75. Throughout the next four technical sections we work over a fixed difference-closed field of characteristic zero with automorphism σ. We reserve the symbol "x" for the variable in the polynomial ring. When we speak of a polynomial, linear polynomial, scalar, etc., we mean a polynomial over this field, linear polynomial over this field, element of this field, etc. Occasionally, and especially towards the end, we explicitly note how our results specialize to the category of algebraic dynamical systems defined over the fixed field of σ.

3. Linear relations between Ritt polynomials

In this section we identify the possible linear relations between Ritt polynomials and identify certain classes of Ritt polynomials admitting extra linear relations. Using these results on linear relations we complete the proof of Theorem 2.50 showing that the action of a Ritt swap at i is well-defined on the linear equivalence classes of decompositions of a polynomial. Much of the basic work on linear relations appears also in [21] (see Lemmas 3.20–3.22) and is implicit in [15].

3.1. Definitions and examples.

Definition 3.1. A *scaling* is a linear polynomial of the form $(\cdot \lambda) := \lambda x$ for some nonzero scalar λ. A *translation* is a linear polynomial of the form $(+A) := x + A$ for some scalar A. Two linearly related polynomials f and g are *translation related* (respectively, *scaling related*) if $g = L \circ f \circ M$ for some translations (respectively, scalings) L and M.

Remark 3.2. The group of automorphisms of $\mathbb{A}^1_K$ may be identified with the semidirect product of the group of translations by the group of scalings.

Definition 3.3. Given a polynomial f and a nonzero scalar λ, we define $\lambda * f := (\cdot \lambda^{-\deg(f)}) \circ f \circ (\cdot \lambda)$.

Remark 3.4. If f is monic, then so is $\lambda * f$. On the other hand, if f and g are monic polynomials and $(\cdot \mu) \circ f \circ (\cdot \lambda) = g$, then $\mu = \lambda^{-\deg(f)}$. That is, $g = \lambda * f$.

Remark 3.5. For any $n \in \mathbb{N}$ and scalar λ, we have $\lambda * P_n = P_n$. More generally, if $f = x^k \cdot U(x^\ell)$ for some polynomial U and λ a scalar, then $\lambda * f = x^k \cdot (\lambda^\ell * U)(x^\ell)$. Thus if f is a Ritt polynomial, then so is $\lambda * f$ for any nonzero λ. In particular, if ℓ is maximal for which f takes this form, then $\lambda * f = f$ if and only if λ is an ℓth root of unity.

The above observations imply that to describe all instances of linear relatedness between Ritt polynomials, it suffices to separately describe those witnessed by translations and those witnessed by scalings.

LEMMA 3.6. *If f and g are linearly related Ritt polynomials, then there is a third Ritt polynomial h which is translation related to f and scaling related to g.*

Proof. Let L and M be linear polynomials with $L \circ f \circ M = g$. Write $L = (\cdot \lambda) \circ (+B)$ and $M = (+A) \circ (\cdot \mu)$ for appropriate scalars A, B, λ, and μ. Set $h := (+B) \circ f \circ (+A)$. Since translations preserve the highest degree term, h is still monic and translation related to f. As $h = (\cdot \lambda^{-1}) \circ g(\cdot \mu^{-1})$ and both g and h are monic, we conclude by Remark 3.4 that $h = \mu^{-1} * g$. From Remark 3.5 we see that h is a Ritt polynomial. $\square$

By similar reasoning, the class of basic Ritt identities other than $C_p \circ C_q = C_q \circ C_p$ is closed under scalings.

PROPOSITION 3.7. *If $b \circ a = d \circ c$ is a basic Ritt identity, at least one of a or b is not a Chebyshev polynomial, and λ and μ are nonzero scalars, then there are scalars η and ν for which $(\mu * b) \circ (\lambda * a) = (\eta * d) \circ (\nu * c)$ is a basic Ritt identity.*

Proof. At least one of a or b must be a monomial P_p for some prime p. If they are both monomials, then the result is immediate as $\lambda * P_p = P_p$. Suppose now that $a = P_p$ and b takes the form $x^k \cdot u(x^\ell)^n$ for some monic u with nonzero constant term. Then $d = P_p$ and $c = x^k \cdot u(x^{p\ell})^{\frac{n}{p}}$. We saw in Remark 3.5 that $\mu * b = x^k \cdot (\mu^\ell * u)(x^\ell)^n$ and $\lambda * a = a$. Thus, $(\mu * b) \circ (\lambda * a) = (1 * d) \circ (\sqrt[p]{\mu} * c)$. Likewise, if $b = P_p$, we may take $\eta = \lambda^p$ and $\nu = 1$. $\square$

Since Chebyshev polynomials of odd degree are odd functions, every Ritt polynomial is of the form $x^k \cdot u(x^\ell)^n$ with $k\ell n > 1$ and therefore is involved in a nontrivial scaling relation to a Ritt polynomial. We focus on translation relations amongst Ritt polynomials; it turns out that these appear in only two special classes, what we call types A (for "adaptable") and C (for "Chebyshev-like").

Definition 3.8. A *type A Ritt polynomial* is a Ritt polynomial of the form $f(x) = x^\ell \cdot (x-A)^m u(x)^n$, where u is a monic polynomial with nonzero constant term, A is some nonzero scalar, and both $\gcd(\ell, n) > 1$ and $\gcd(m, n) > 1$. A *type A swappable polynomial* is a polynomial that is linearly related to a type A Ritt polynomial.

Remark 3.9. Since a Ritt polynomial must be indecomposable, in Definition 3.8 we must have $\gcd(\ell, m, n) = 1$.

Remark 3.10. Lemma 3.6, the observation that for any λ and f, either both f and $\lambda * f$ are type A Ritt polynomials, or neither one is, and Theorem 3.15 together imply that a Ritt polynomial that happens to be a type A swappable polynomial is, in fact, a type A Ritt polynomial.

Definition 3.11. A *type C swappable polynomial* is a polynomial of odd prime degree that is linearly related to a Chebyshev polynomial.

Definition 3.12. For a natural number n and scalar λ, we define $C_{n,\lambda} := \lambda * C_n$ and $\widehat{C}_{n,\lambda} := \lambda * ((+2) \circ C_n \circ (-2))$. For odd prime n, these are the *type C Ritt polynomials*.

It follows from Remark 3.5 that $C_{n,\lambda}$ is a Ritt polynomial for odd prime n and nonzero λ. For odd n, we show (Proposition 3.13) that both $\widehat{C}_{n,1} = (+2) \circ C_n \circ (-2)$ and $\widehat{C}_{n,-1} = (-2) \circ C_n \circ (+2)$ are of the form $x \cdot u(x)^2$ as a consequence of the fact that C_n commutes with $C_2(x) = x^2 - 2$. It then

follows from Remark 3.5 that $\widehat{C}_{n,\lambda}$ are Ritt polynomials for all odd prime n and nonzero λ. It follows from Theorem 3.15 that these are the only Ritt polynomials amongst type C swappable polynomials.

PROPOSITION 3.13. *For every odd prime p and scalar λ, the polynomial $\widehat{C}_{p,\lambda}$ is a Ritt polynomials of the form $x \cdot u(x)^2$. Moreover, for any number n, we have $\widehat{C}_{n,-1} = (-4) \circ \widehat{C}_{n,1} \circ (+4)$.*

Proof. For odd n, we show that both $\widehat{C}_{n,1} = (+2) \circ C_n \circ (-2)$ and $\widehat{C}_{n,-1} = (-2) \circ C_n \circ (+2)$ are of the form $x \cdot u(x)^2$ as a consequence of the fact C_n commutes with $C_2(x) = x^2 - 2$. For the first observation, we compute

$$C_n \circ C_2 = C_2 \circ C_n,$$
$$C_n \circ (-2) \circ P_2 = (-2) \circ P_2 \circ C_n,$$
$$(+2) \circ C_n \circ (-2) \circ P_2 = P_2 \circ C_n.$$

Thus, since $\widehat{C}_n = (+2) \circ C_n \circ (-2)$ appears in a basic Ritt identity with P_2, it must be of the form $x \cdot u(x)^2$ for some polynomial u. By Remark 3.5, it follows that the same holds of $\widehat{C}_{n,\lambda}$ for all nonzero λ.

For the second, first observe that

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

Now

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

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

$$(-2) \circ C_n \circ (+2) \circ P_2 = (-2) \circ (\cdot - 1) \circ C_2 \circ C_n \circ (\cdot i).$$

Now, $[(-2) \circ (\cdot - 1) \circ C_2](x) = -(x^2 - 2) - 2 = -x^2 = [P_2 \circ (\cdot \pm i)](x)$, so

$$(-2) \circ C_n \circ (+2) \circ P_2 = P_2 \circ (\cdot \pm i) \circ C_n \circ (\cdot i) = P_2 \circ (i * C_n). \qquad \square$$

Although C_2 is not a Ritt polynomial, how it might be linearly related to itself or to the monomial P_2 is important in Section 4 and is summarized with the following remark.

Remark 3.14. Since the only way P_2 is linearly related to itself is by scalings $\lambda * P_2 = P_2$, the only way $C_2(x) = x^2 - 2$ is linearly related to itself is by $A_\lambda \circ C_2 \circ (\cdot \lambda) = C_2$ for $A_\lambda(x) := \frac{1}{\lambda^2}x + \frac{2}{\lambda^2} - 2$. Note the immediate consequence that if $L \circ P_2 \circ M = C_2$, then $M = \cdot \lambda$ is a scaling, and $L = B_\lambda$ where $B_\lambda(x) := \frac{1}{\lambda^2}x - 2 = (-2) \circ (\cdot \frac{1}{\lambda^2})(x)$. Note that $A_\lambda(x)$ is never a scaling unless $\lambda = \pm 1$ and $A_\lambda = \mathrm{id}$, and B_λ is never a scaling.

3.2. *Characterization of translation related Ritt polynomials.* In the next theorem, whose proof occupies the rest of this subsection, we collect all instances of linear relatedness amongst Ritt polynomials via translations. Using Lemma 3.6, a general description follows.

THEOREM 3.15. *If f and g are Ritt polynomials and A and B are scalars, not both zero, for which $(+B) \circ f \circ (+A) = g$, then either*

- $B = 0$, f *and g are type* A *Ritt polynomials, or*
- $B \neq 0$, f *and g are type* C *Ritt polynomials.*

In fact, if $B \neq 0$, then either $f = C_{p,\lambda}$ and $g = \widehat{C}_{p,\lambda}$, where $\lambda = \frac{-2}{A} = \sqrt[p]{\frac{2}{B}}$ and p is an odd prime, or $f = \widehat{C}_{p,\mu}$ and $g = \widehat{C}_{p,-\mu}$, where $\mu = \frac{4}{A} = \sqrt[p]{\frac{-4}{B}}$ and p is an odd prime.

We turn to the task of proving Theorem 3.15, reformulating its statement as the solution of the following problem.

PROBLEM 3.16. *For which Ritt polynomials f and g and scalars A and B can we have*

$$(+B) \circ f \circ (+A) = g?$$

In the solution of Problem 3.16 and in the course of the analysis of the monoid actions introduced in Section 2.4, we make use of some refined degrees of Ritt polynomials.

Definition 3.17. If f is any polynomial that is not a monomial, then f may be expressed as $x^k \cdot u(x^\ell)^n$, where u is a polynomial with a nonzero constant term and n and ℓ are maximal. The number k is the order of vanishing of f at 0. The number n, which we call the *out-degree* of f, is the greatest common divisor of the orders of vanishing of f at points other than 0. The number ℓ, which we call the *in-degree* of f, is the size of the multiplicative stabilizer of the set of roots of f.

Remark 3.18. Of course, it is true that a monomial may be expressed in the above form, taking $u = 1$, but then no maximal n nor ℓ would exist. If f is a nonmonomial Ritt polynomial, then either its in-degree or its out-degree must be at least two.

Remark 3.19. By considering type A Ritt polynomials, one sees that even for Ritt polynomials, the out-degree and in-degree are *not* invariants of the linear relatedness class of a polynomial. However, two scaling related Ritt polynomials f and $\lambda * f$ clearly have the same in-degrees and out-degrees.

It is natural to state the next lemma here, even though its proof requires some later results. Rest assured that the lemma is not used until after those later results are proved.

LEMMA 3.20. *All $C_{p,\lambda}$ have in-degree 2 and out-degree 1. All $\widehat{C}_{p,\lambda}$ have in-degree 1 and out-degree 2.*

Proof. Since C_p is an odd function, its in-degree is divisible by 2. From the computations in the proof of Proposition 3.13, it follows that the out-degree of $\widehat{C}_p$ is divisible by 2. The rest of the result for C_p and $\widehat{C}_p$ follows by Proposition 3.24 and Lemma 3.29, and Remark 3.19 finishes the proof.   □

Returning to Problem 3.16 we observe that $A = B = 0$ and $f = g$ always gives a trivial solution. On the other hand, evaluating both sides at 0 we see that there are no solutions with $A = 0 \neq B$. Thus, we may and do assume that $A \neq 0$ and examine the cases where $B = 0$ and where $B \neq 0$ separately. We have already found some solutions of these problems: type A Ritt polynomials for the case when $B = 0$ and type C Ritt polynomials for the case $B \neq 0$. Our task is to prove that there are no others. Using an appropriate scaling, we reduce to the case that $A = 1$.

LEMMA 3.21. *If A, B, f, and g give a solution to Problem 3.16, then 1, $\frac{B}{A^{\deg(f)}}$, $A * f$, $A * g$ is also a solution to Problem 3.16.*

Proof. $(+\frac{B}{A^{\deg(f)}}) \circ (A * f) \circ (+1) = A^{-\deg(f)} f(A(x + 1)) + A^{-\deg(f)} B = A^{-\deg(f)}(f(Ax + A) + B) = A * ((+B) \circ f \circ (+A)) = A * g.$ □

REDUCTION 3.22. *For the remainder of this section, we assume that $A = 1$. Thus, we seek solutions to*

$$(+B) \circ f \circ (+1) = g,$$

where f and g are Ritt polynomials. By way of notation, we write $f = f_1 = x^{k_1} u_1(x^{\ell_1})^{n_1}$ and $g = f_2 = x^{k_2} u_2(x^{\ell_2})^{n_2}$, where ℓ_i is the in-degree of f_i and n_i is the out-degree of f_i. We write $s_i := \deg(u_i)$ and t_i for the number of zeros of u_i, not counted with multiplicity.

Let us record a simple ramification calculation.

LEMMA 3.23. *Let k, ℓ, n be natural numbers with $\gcd(k, \ell) = \gcd(k, n) = 1$ and u a polynomial with $u(0) \neq 0$. Set $f := x^k \cdot u(x^\ell)^n$. Let t be the number of zeros of u not counted with multiplicity, and let $s := \deg(u)$. Then the following hold:*

- *The number of points (counted with multiplicity) at which both f and f' vanish, that is, the number of ramification points above zero, is $(k - 1) + \ell(ns - t)$.*
- *The number of points at which f' vanishes but f does not, that is, the number of ramification points lying above points other than zero, is ℓt. Moreover, this set of points is closed under multiplication by the group of ℓth roots of unity.*

Proof. This is a straightforward computation, which we include for completeness:

$$f'(x) = kx^{k-1}u(x^\ell)^n + x^k nu(x^\ell)^{n-1}u'(x^\ell)\ell x^{\ell-1}$$
$$= x^{k-1}u(x^\ell)^{n-1}(ku(x^\ell) + \ell nu'(x^\ell)x^\ell).$$

Since $u(0) \neq 0$, we see that $\text{ord}_0 f' = k - 1$. For any other root a of f, we must have $u(a^\ell) = 0$. If $u'(a^\ell) = 0$, then $\text{ord}_a f' = (n-1)\text{ord}_{a^\ell} u + \text{ord}_{a^\ell} u' = n\,\text{ord}_{a^\ell}(u) - 1$. Otherwise, $\text{ord}_{a^\ell}(u) = 1$ and $\text{ord}_{a^\ell}(f) = n - 1$. Summing over the distinct roots of u, we finish the calculation of the total ramification over zero. If we let $\tilde{u} := \gcd(u, u')$, by which we mean the monic polynomial that generates the ideal generated by u and u', then the other zeros of f' come from the zeros of $k\frac{u}{\tilde{u}}(x^\ell) + \ell n\frac{u'}{\tilde{u}}(x^\ell)x^{\ell-1}$, which has degree exactly ℓt.   □

Differentiating the equation $(+B) \circ f_1 \circ (+1) = f_2$, we see that $f_1' \circ (+1) = f_2'$. Hence, for any point a, we have $\text{ord}_a f_2' = \text{ord}_{a+1} f_1'$. That is, $(+1)$ translates the zeros of f_2' to the zeros of f_1' respecting multiplicities. If $B = 0$, then the ramification above zero is matched. If $B \neq 0$, then there is one nonzero point for which the ramification of f_2 above zero is matched with the ramification of f_1 above that point and vice versa. It is this consequence that makes these seemingly trivial observations useful.

PROPOSITION 3.24. *In the notation from* Reduction 3.22, *either* $\ell_1 = 1$ *or* $\ell_2 = 1$.

Proof. If $\ell_1 > 1$, then the sum of the roots of f_1 is zero as is the sum of the roots of f_1'. Indeed, zero contributes nothing to the sum. The other roots both of f_1 and of f_1' are partitioned into cosets of the ℓ_1^{th} roots of unity over which the sum is zero. Because $f_2' = f_1' \circ (+1)$, we see that the sum of the roots of f_2' is $(1 - \deg(f_1)) \neq 0$ (as $\deg(f_1) \geq 3$).   □

REDUCTION 3.25. *For the remainder of this section, we take* $\ell_1 = 1$.

LEMMA 3.26. *If* $B = 0$, *then* $\ell_2 = 1$.

Proof. As $f_2(x + 1) = f_1(x)$, we see that $k_1 = \text{ord}_0 f_1 = \text{ord}_{-1} f_2$. That is, -1 is a k_1-fold zero of $u_2(x^{\ell_2})^{n_2}$. We thus have $\text{ord}_{-\zeta} f_2 = k_1$ for any other ℓ_2^{th} root of unity ζ. Unless $\ell_2 = 1$, we can choose ζ so that $-\zeta + 1 \neq 0$, but then $k_1 = \text{ord}_{-\zeta+1} f_1 = \text{ord}_{-\zeta+1} u_1(x)^{n_1}$, so n_1 divides k_1. If $n_1 > 1$, this contradicts the indecomposability of f_1. Otherwise, $n_1 = \ell_1 = 1$, so f_1 is not a Ritt polynomials, again a contradiction.   □

We first complete the solution for the case where $B = 0$.

PROPOSITION 3.27. *In* Problem 3.16, *if* $B = 0$ *and* $\ell_1 = \ell_2 = 1$, *then there are positive integers* m_1, m_2 *and a monic polynomial* U *for which* $u_1(x) = (x-1)^{m_1}U(x)^{n_2}$ *and* $u_2(x) = (x+1)^{m_2}U(x+1)^{n_1}$. *In particular,* f_1 *and* f_2 *are type* A *Ritt polynomials.*

Proof. As $k_2 = \mathrm{ord}_0 f_2 = \mathrm{ord}_1 f_1$, we see that $k_2 \mid n_1$. Set $m_1 := \frac{k_2}{n_1}$. Observe that $\mathrm{ord}_1 u_1 = m_1$. Likewise, since $k_1 = \mathrm{ord}_0 f_1 = \mathrm{ord}_{-1} f_2$, n_2 divides k_1. Write $m_2 := \frac{k_1}{n_2}$, and observe that $\mathrm{ord}_{-1} u_2 = m_2$. Express $u_1(x) = (x-1)^{m_1} V_1(x)$ and $u_2(x) = (x+1)^{m_2} V_2(x)$. Specializing Problem 3.16, we have the following equation:

$$(x+1)^{m_2 n_2} \cdot [x^{m_1} V_1(x+1)]^{n_1} = x^{m_1 n_1} \cdot [(x+1)^{m_2} V_2(x)]^{n_2}.$$

Canceling $(x+1)^{m_2 n_2} x^{m_1 n_1}$ we obtain $V_1(x+1)^{n_1} = V_2(x)^{n_2}$. Recalling that $n_2 m_2 = k_1$ and n_1 are relatively prime, so that $\gcd(n_1, n_2) = 1$, it must be that V_1 is an n_2^{th} power and V_2 an n_1^{th} power. Write $V_1 = U_1^{n_2}$ and $V_2 = U_2^{n_1}$. As f_1 is monic, we may take each of U_1 and U_2 to be monic. As $U_1(x+1)^{n_1 n_2} = U_2(x)^{n_1 n_2}$, we have $U_1(x+1) = U_2(x)$, as required. $\qquad\square$

REDUCTION 3.28. *In what follows, we assume that $B \neq 0$.*

LEMMA 3.29. *Given our reductions, $k_1 = k_2 = 1$, all roots of u_1 and u_2 are simple, and $n_1 = n_1 \ell_1 = n_2 \ell_2 = 2$.*

Proof. Concretely, we are considering the equation $f_1(x+1) + B = f_2(x)$. Since $\gcd(k_2, \ell_2) = 1$, if $f_2(a) \neq 0$ and $\zeta \neq 1$ is an ℓ_2^{th} root of unity, then $f_2(\zeta a) \neq f_2(a)$. Thus, in each of the cosets of the ℓ_2^{th} roots of unity contained in the critical points of f_2 there can be at most one point that maps to B under f_2. As translation by 1 takes the critical points of f_2 over B to the critical points of f_1 over 0, we conclude from Lemma 3.23 (taking into account that $\ell_1 = 1$) that

$$(k_1 - 1) + (n_1 s_1 - t_1) \leq t_2.$$

On the other hand, since translation by 1 induces a (multiplicity preserving) bijection between the critical points of f_2 with those of f_1, we see that the other critical points of f_2 must be mapped to critical points of f_1 not above 0. From Lemma 3.23 again we see that

$$(k_2 - 1) + \ell_2(n_2 s_2 - t_2) + (\ell_2 - 1)t_2 \leq t_1.$$

Combining these two inequalities we obtain

$$(k_1 - 1) + (k_2 - 1) + n_1 s_1 + \ell_2 n_2 s_2 \leq 2t_1 + 2t_2.$$

Bearing in mind that $t_i \leq s_i$, $2 \leq n_1 \ell_1 = n_1$, $2 \leq n_2 \ell_2$, and $1 \leq k_i$, we see that all of these inequalities must be equalities. $\qquad\square$

Thus, we are left with describing those solutions where $\ell_2 = 2$ and $n_2 = 1$ and where $\ell_2 = 1$ and $n_2 = 2$. We already have examples of these in Definition 3.12 and Proposition 3.13; the next two propositions say that there are no others.

PROPOSITION 3.30. *For each positive integer s, there is a unique monic polynomial u of degree s for which there is some nonzero scalar B and polynomial v satisfying*

$$(2) \qquad (+B) \circ (x \cdot u(x)^2) \circ (+1) = (x \cdot v(x)^2).$$

Proof. The polynomials u and v have only simple roots by Lemma 3.29. Since u is monic, we may assume that v is monic as well.

Differentiating, we obtain

$$(3) \qquad u(x+1)(u(x+1) + 2(x+1)u'(x+1)) = v(x)(v(x) + 2xv'(x)).$$

Since $B \neq 0$, it follows that $u(x+1)$ and $v(x)$ are coprime. Hence, $u(x+1) + 2(x+1)u'(x+1)$ is a scalar multiple of $v(x)$ and $v(x) + 2xv'(x)$ is a scalar multiple of $u(x+1)$. Taking into account the leading coefficients, we deduce the following equations:

$$(4) \qquad (2s+1)v(x) = u(x+1) + 2(x+1)u'(x+1),$$

$$(5) \qquad (2s+1)u(x+1) = v(x) + 2xv'(x).$$

Differentiating equation (4), we obtain

$$(6) \qquad (2s+1)v'(x) = 3u'(x+1) + 2(x+1)u''(x+1).$$

Multiplying equation (5) by $(2s+1)$, and then using equations (4) and (6) to eliminate v and v', we obtain

(7)
$$(2s+1)^2 u(x+1) = u(x+1) + 2(x+1)u'(x+1) + 2x(3u'(x+1) + 2(x+1)u''(x+1)).$$

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

$$(8) \qquad (2s^2 + 2s)Y + (-4x - 1)Y' + 2(-x - x^2)Y'' = 0.$$

A routine calculation shows that if $u(x+1)$ is a solution to equation (8) and we define $v(x)$ via equation (4) and set $B := -u(1)^2$, then these data satisfy equation (2).

The linear differential operator $L = 2(-x-x^2)\frac{d^2}{dx^2} + (-4x-1)\frac{d}{dx} + (2s^2+2s)$ 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 $M_{j,j} = 2(1-j)j - 4j + (2s^2 + 2s) = (2s^2 + 2s) - (2j^2 + 2j)$, and just above the diagonal, we have $M_{j,j+1} = -(j+1)(3+2j)$. In particular, $M_{s,s} = 0$ so that $\text{rank}(L) \leq s$ while the (s,s)-minor is invertible. Thus, the rank of L is s and the dimension of the space of solutions to equation (8) is exactly one. As we require u to be monic, there is exactly one solution of degree s. $\qquad \square$

PROPOSITION 3.31. *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 v satisfying*

$$(9) \qquad (+B) \circ (x \cdot u(x)^2) \circ (+1) = (x \cdot v(x^2)).$$

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

$$(10) \quad u(x+1) \cdot (u(x+1) + 2(x+1)u'(x+1)) = v(x^2) + 2x^2 v'(x^2) = (v + 2x \cdot v') \circ P_2.$$

The zeros of the right-hand side of equation (10) 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 (9) would yield $cv(c^2) = B = -cv((-c)^2) = -cv(c^2)$, contrary to the fact that $B \neq 0$. Thus, at most one of each pair of roots of the right-hand side is also a root of $u(x+1)$. As the degree of the right-hand side of equation (10) 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:

$$(11) \qquad (-1)^s(2s+1)u(x+1) = (u(-x+1) + 2(-x+1)u'(-x+1)).$$

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

$$(12) \qquad 0 = u(z) + 2zu'(z) - (2s+1)(-1)^s u(2-z).$$

The difference-differential operator in equation (12) 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

$$1 + 2j - (-1)^{j+s}(2s+1).$$

Hence, the rank of this operator is exactly s implying that there is a unique monic solution. $\qquad \square$

This concludes the proof of Theorem 3.6. $\qquad \square$

3.3. *Proof of Theorem 2.50 and related results.* We collect some observations about Ritt swaps towards and around the proof of Theorem 2.50.

Remark 3.32. 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 Ritt polynomial whose out-degree is a multiple of p, or

- f_{i+1} is linearly related to a monomial P_p and f_i is linearly related to a Ritt polynomial whose in-degree is a multiple of p.

Remark 3.33. The in-degree of a type A swappable f is 1, in the sense that any Ritt polynomial linearly related to f has in-degree 1. Remark 3.32 then implies that if f_i is type A, then no decomposition may be obtained from $\vec{f}$ by a Ritt swap at i.

We now prove some useful consequences of Theorem 3.15, including Theorem 2.50. We begin with a few slightly more comprehensive results about Chebyshev polynomials.

COROLLARY 3.34. *If L and M are linear, $p \geq 3$ is prime, and $L \circ C_p \circ M = C_p$, then both M and L are $(\cdot \pm 1)$.*

Proof. As in the proof of Lemma 3.6, there are scalars A, B, λ, and μ such that $L = (\cdot \lambda) \circ (+B)$ and $M = (+A) \circ (\cdot \mu)$. Let $h := (+B) \circ C_p \circ (+A) = (\cdot \frac{1}{\lambda}) \circ C_p \circ (\cdot \frac{1}{\mu})$. By the first equality, h is monic, so $h = \frac{1}{\mu} * C_p = C_{p, \frac{1}{\mu}}$ is a Ritt polynomial. Since C_p has in-degree at least 2, so does h. By Theorem 3.15, h cannot be nontrivially translation related to another Ritt polynomial C_p with in-degree 2, so $A = B = 0$ and $h = C_p$. Since all complex roots of C_p are real, $C_p \neq C_{p, \frac{1}{\mu}}$ unless $\mu = \pm 1$.   □

LEMMA 3.35. *For any Ritt swap involving a type C swappable, the underlying basic Ritt identity is either of the form $C_p \circ C_q = C_q \circ C_p$ for odd prime p and q or of the form $P_2 \circ C_p = \widehat{C_p} \circ P_2$ for some odd prime p. In particular, if a type C swappable f_i "becomes" g_j through Ritt swaps, in the sense of Remark 2.46, then g_j is also a type C swappable.*

Proof. By Theorem 3.15, C_p (for odd prime p) is not linearly related to any Ritt polynomials except $C_{p,\lambda}$ and $\widehat{C}_{p,\lambda}$. By Lemma 3.20, all these have in- and out-degrees 1 and 2, so they can only participate in basic Ritt identities of the third kind with the quadratic P_2. It is easy to obtain the identity $P_2 \circ C_p = \widehat{C_p} \circ P_2$ from $C_2 \circ C_p = C_p \circ C_2$ and the definition of $\widehat{C_p}$. (See the proof of Proposition 3.13.)   □

Next, we generalize Corollary 3.34 to decomposable Chebyshev polynomials.

LEMMA 3.36. *If A and B are linear, $n \neq 2$, and $B \circ C_n \circ A = C_n$, then each of A and B are scalings by ± 1.*

Proof. Let $p_k, \ldots, p_1$ be the prime factors of n, with repetitions, with $p_k, \ldots, p_m$ equal to 2 and the rest odd. Now $(B \circ C_{p_k}, C_{p_{k-1}}, \ldots, C_{p_2}, C_{p_1} \circ A)$ must be linearly equivalent to $(C_{p_k}, C_{p_{k-1}}, \ldots, C_{p_2}, C_{p_1})$. Let $L_{k-1}, \ldots, L_1$ witness this. Induct right-to-left.

If any p_i are odd, then p_1 is odd, so in $L_1^{-1} \circ C_{p_1} \circ A = C_{p_1}$ we must have $L_1 = A = (\cdot(\pm 1))$ by Corollary 3.34. Then at each step, $L_i^{-1} \circ C_p \circ (\cdot(\pm 1))$ forces $L_i = (\cdot(\pm 1))$ (even for $p = 2$), and finally at the last step, $B = (\cdot(\pm 1))$.

If all $p_i = 2$, then $k \geq 2$. From $L_1^{-1} \circ C_2 \circ A = C_2$ we get (using Remark 3.14) that $A = (\cdot \lambda)$ is a scaling and $L_1^{-1}(x) := \frac{1}{\lambda^2} x + \frac{2}{\lambda^2} - 2$. From the next step (since $k \geq 2$, there is a next step), we see that L_1 must also be a scaling, so $\lambda = \pm 1$ and A is as desired, and $L_1 = \mathrm{id}$. Now inducting, at each step $L_i^{-1} \circ C_2 = C_2$ makes all $L_i = \mathrm{id}$, and at the last step, $B = \mathrm{id}$. □

The next lemma is something of a converse to Proposition 3.7.

LEMMA 3.37. *If a and b are Ritt polynomial and not both type C; 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 Ritt polynomial 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 other Ritt polynomial. Since both b and $L \circ b \circ M^{-1}$ are Ritt polynomial and M is a scaling, L must also be a scaling because the equation in Problem 3.16 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 Ritt polynomial, either N is a scaling or both a and $(M \circ a \circ N^{-1})$ must be type A. However $(L \circ b \circ M^{-1}, M \circ a \circ N^{-1})$ is swappable, contradicting Remark 3.33.

The "furthermore" clause follows immediately from Proposition 3.7. □

We complete the proof of Theorem 2.50: 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 $\vec{g}$.

Proof. (This is the proof of Theorem 2.50.) Let us collect and name the witnesses for the two Ritt swaps at i. That is, for $j = 1$ or 2, we have linear polynomials L_j, M_j, and N_j and Ritt polynomial 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 seek a linear R for which

$$(L_1 \circ \widehat{H_1}) \circ R = (L_2 \circ \widehat{H_2}) \text{ and } R^{-1} \circ (\widehat{G_1} \circ N_1^{-1}) = \widehat{G_2} \circ N_2^{-1}.$$

Let

$$L := L_2^{-1} \circ L_1, \quad M := M_2^{-1} \circ M_1, \quad \text{and} \quad N := N_1^{-1} \circ N_1.$$

Then

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

Applying L_2 to the left of the first equation below and N_2^{-1} to the right of the second one shows that it is sufficient to find R such that

$$(L \circ \widehat{H_1}) \circ R = \widehat{H_2} \text{ and } R^{-1} \circ (\widehat{G_1} \circ N^{-1}) = \widehat{G_2}.$$

Recall that $G_j \circ H_j = \widehat{H_j} \circ \widehat{G_j}$ are basic Ritt identities, so the above equations are linear relations between Ritt polynomial polynomials.

We claim that $R = M^{-1}$ always works and 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, $\widehat{G_1} = G_1 = G_2 = \widehat{G_2}$ and $\widehat{H_1} = H_1 = H_2 = \widehat{H_2}$ are Chebyshev polynomials of odd degree, since commuting Chebyshevs are the only basic Ritt identity not involving any monomials. Then by Corollary 3.34, L, M, and N are scalings by ± 1, and $R := M^{-1}$ works.

(one) This is done in Lemma 3.37, with $b := G_1$, $a := H_1$, with one less assumption.

(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 nontrivially translation related to themselves.) $\square$

We end this section with a lemma closely resembling Lemma 3.37. This time we demand that both, rather than at least one, of the Ritt polynomials not be type C, but we give more freedom to the linear factors.

LEMMA 3.38. *Suppose that a and b are Ritt polynomials and neither is type C; L, M, R, and S are linear; and $\tilde{b} := S \circ b \circ R$ and $\tilde{a} := M \circ a \circ L$ are Ritt polynomials; and $\tilde{b} \circ \tilde{a} = \tilde{d} \circ \tilde{c}$ is a basic Ritt identity. Then L, M, and S are scalings; and there are scalars A and λ, and a Ritt polynomial $\hat{b} := b \circ (+A)$ such that $b \circ R = \lambda * \hat{b}$; and for some d and c, $\hat{b} \circ a = d \circ c$ is a basic Ritt identity. Unless b is type A, $A = 0$ and $\hat{b} = b$.*

Proof. Since neither a nor b is type C, M and S must be scalings. If L is not a scaling, then a and $\tilde{a}$ must be type A, but this contradicts Remark 3.33. Thus, L is a scaling, say by λ, and $\tilde{a} = \lambda * a$.

Write $C = (+A) \circ (\cdot \mu)$ for scalars A and μ. Then $\hat{b} := b \circ (+A)$ is a monic polynomial scaling related to the Ritt polynomial $\tilde{b}$, so it is itself Ritt polynomial. Thus if $A \neq 0$, then b is type A, and in any case $\tilde{b} = \mu * \hat{b}$.

Thus $(\mu * \hat{b}) \circ (\lambda * a) = \tilde{d} \circ \tilde{c}$ is a basic Ritt identity. By [Proposition 3.7](), there are η and ν such that $\hat{b} \circ a = (\eta * \tilde{d}) \circ (\nu * \tilde{c})$ is a basic Ritt identity. □

4. **Clusters**

In this section, we describe a natural and nearly canonical way to break a decomposition of a disintegrated polynomial into *clusters* in a way that controls the linear factors floating amongst Ritt polynomials and makes it easy to see what other decompositions can be obtained via sequences of Ritt swaps.

One of our two kinds of clusters is the same as one of the two kinds of blocks in [21], but our C-free clusters are nothing like their monomial blocks.

Our first use of these clusters is to prove the fundamental [Theorem 2.52]() for the Ritt monoid action, that $t_i t_{i+1} t_i \star \vec{f}$ is defined if and only if $t_{i+1} t_i t_{i+1} \star \vec{f}$ is defined. While it follows immediately from [21] that the two are equal when defined, it is not clear to us whether our stronger result follows from their work.

We end up showing that any polynomial all of whose indecomposable factors are swappable can be written almost uniquely (not up to permutations!) as a composition of lower-degree polynomials (compositions of "clusters") each of which is linearly related to a (possibly decomposable) Chebyshev polynomial, or to a composition of several Ritt polynomials, none of them type C.

We do not work out the straightforward generalization that includes decompositions with unswappable factors in this analysis by allowing a third kind of cluster, a single unswappable indecomposable: our case-outs are unwieldy enough as it is. [Corollary 2.73]() suffices to characterize (f, g)-skew-invariant curves when one of f or g polynomials has at least one unswappable factor.

The polynomial then admits a decomposition where linear factors are collected outside these clusters, Ritt swaps can only take place inside these clusters, and can always be witnessed (with respect to this decomposition) by identity linear factors — almost. Factors of degree two are the only source of ambiguity in choosing these clusters, and one of them is always involved in any Ritt swap between two clusters.

Further refining our analysis of the linear factors between clusters, we show that, when the clusters are maximal enough, only one quadratic may cross the boundary between two clusters, and then only in one direction.

Throughout this section we constantly use the results of the [Section 3]() without explicit reference.

Definition 4.1. Given a sequence $\vec{f} := (f_k, \ldots, f_1)$ of polynomials and integers $k \geq b \geq a \geq 0$, we use the following notation:
$$\vec{f}_{[b,a)} := (f_b, f_{b-1}, \ldots, f_{a+1}), \quad \vec{f}^{\circ}_{[b,a)} := f_b \circ f_{b-1} \circ \cdots \circ f_{a+1}, \quad \vec{f}^{\circ} := f_k \circ f_{k-1} \circ \cdots \circ f_1.$$

Definition 4.2. Let $\vec{f}$ be a decomposition of a polynomial f.

- If $B \circ \vec{f}^{\circ}_{[b,a)} \circ A = C_n$ is a Chebyshev polynomial for some integer n that is not a power of 2, and some linear A and B, then $\vec{f}_{[b,a)}$ is a C *cluster*.
- If $\vec{f}_{[b,a)}$ is linearly equivalent to $(B \circ h_b, h_{b-1}, \ldots, h_{a+2}, h_{a+1} \circ A)$ for some linear A and B and some Ritt polynomials h_i none of which are type C, then $\vec{f}_{[b,a)}$ is an C-*free cluster*.

In either case, $\vec{f}_{[b,a)}$ is a *cluster* of $\vec{f}$.

A *preclustering* of a decomposition $\vec{f}$ is a sequence $k = a_r > a_{r-1} > \cdots > a_1 > a_0 = 0$ such that $\vec{f}_{[a_j,a_{j-1})}$ is a cluster for each j. We say that i is *a cluster boundary* of $\vec{a}$ if $i = a_j$ for some j.

Definition 4.3. If $k = a_r > a_{r-1} > \cdots > a_1 > a_0 = 0$ is a preclustering of a decomposition $\vec{f}$, the data $(h_k, \ldots, h_1; L_k, L_{k-1}, \ldots, L_0)$ is a *cleanup* of this preclustering if

(1) $(L_k \circ h_k \circ L_{k-1}, h_{k-1} \circ L_{k-2}, \ldots, h_1 \circ L_0)$ is linearly equivalent to $\vec{f}$;
(2) all L_i are linear, and $L_i = \mathrm{id}$ except when i is a cluster boundary of $\vec{a}$;
(3) inside C clusters (that is, whenever $\vec{f}_{[a_j,a_{j-1})}$ is a C cluster and $a_j \geq i > a_{j-1}$) $h_i = C_{\deg(f_i)}$ are Chebyshev polynomials;
(4) inside C-free clusters (that is, whenever $\vec{f}_{[a_j,a_{j-1})}$ is a C-free cluster and $a_j \geq i > a_{j-1}$) h_i are Ritt polynomials;
(5) the linear factor L_{a_j} to the right of any C-free cluster $\vec{f}_{[a_{j+1},a_j)}$ is a translation, and if $h_{a_j+1} \circ L_{a_j}$ is a Ritt polynomial, then $L_{a_j} = \mathrm{id}$.

Remark 4.4. Swappable factors of a decomposition $\vec{f}$ are linearly related to Ritt polynomials, and the linear factors witnessing this can be gathered outside clusters in the following somewhat canonical fashion. Applying the definition of "cluster" to all clusters of a preclustering puts a linear factor on each side of each cluster. Composing pairs of linear factors that sit between clusters, we may assume that only the leftmost cluster has a linear factor on the left of it. To obtain a cleanup, push all scalings through C-free clusters as far left as possible (Lemma 4.6). Generalizing the results of Section 3 from single indecomposable factors to clusters (Lemma 4.7 and Lemma 3.36) shows that these cleanups are essentially unique up to scalings by ± 1 (Proposition 4.10).

The decomposition with $k = 0$ factors, whose clustering has $r = 0$ clusters and whose cleanup has no h_i and $L_0 = \mathrm{id}$, provides a conveniently trivial base case for proofs by induction on the number of clusters.

Remark 4.5.

(1) A decomposition with an unswappable factor does not admit a clustering. Dealing with such decompositions is much easier and does not require the machinery of this section.

(2) If every factor f_i of $\vec{f}$ is swappable, then $r := k$ and $a_j := j$ is a preclustering of $\vec{f}$.

(3) All these notions (cluster, preclustering, cleanup) only depend on the linear equivalence class of the decomposition $\vec{f}$.

(4) A nonempty subsequence of a cluster is a cluster, unless a C cluster loses all of its odd-degree factors. That is, if $\vec{f}_{[b,a)}$ is a cluster of $\vec{f}$ and $b \geq b' > a' \geq a$, and $\vec{f}_{[b',a')}$ is not a cluster, then $\vec{f}_{[b',a')}^{\circ}$ is linearly related to C_{2M} for some $M \geq 1$ and the original cluster $\vec{f}_{[b,a)}$ was a C cluster.

(5) If $t_i \star \vec{f}$ is defined, then (f_{i+1}, f_i) is a cluster.

The next lemma is used to achieve the last part of the definition of "cleanup."

LEMMA 4.6. *Any C-free cluster $\vec{f}_{[b,a)}$ is linearly equivalent to*

$$(B \circ \tilde{h}_b, \tilde{h}_{b-1}, \ldots, \tilde{h}_{a+2}, \tilde{h}_{a+1} \circ A)$$

for some Ritt polynomials $\tilde{h}_i$, some linear B, and some translation A such that $\tilde{h}_{a+1} \circ A$ is a Ritt polynomial only if $A = \mathrm{id}$.

Proof. Take linear A_0 and B_0 and Ritt polynomials h_i from the definition of C-free cluster. Write $A_0 := S_0 \circ A_1$ for a scaling $S_0 =: (\cdot \lambda)$ and a translation A_1.

Let $S_1 := (\cdot \lambda^{\deg(h)})$; by Remark 3.5, $\tilde{h}_{a+1} := \lambda * h_{a+1} = S_1^{-1} \circ h_{a+1} \circ S_0$ is a Ritt polynomial. Similarly, there is a scaling S_2 such that $\tilde{h}_{a+2} := S_2^{-1} \circ h_{a+2} \circ S_1$ is Ritt polynomial; and so on until we get $\vec{f}_{[b,a)}$ linearly equivalent to $(B_0 \circ S_{b-a} \circ \tilde{h}_b, \tilde{h}_{b-1}, \ldots, \tilde{h}_{a+2}, \tilde{h}_{a+1} \circ A_1)$. Set $B := B_0 \circ S_{b-a}$.

Finally, if $\tilde{h}_{a+1} \circ A_1$ is Ritt polynomial, replace $\tilde{h}_{a+1}$ by $\tilde{h}_{a+1} \circ A_1$, and let $A := \mathrm{id}$. Otherwise, leave $\tilde{h}_{a+1}$ as is, and let $A := A_1$. $\square$

The next lemma in some sense generalizes the results of Section 3 from single indecomposable factors to whole C-free clusters; for C clusters, this is already done in Lemma 3.36. The two are induction steps of the proof of uniqueness of cleanups in Lemma 4.10.

LEMMA 4.7. *Suppose that all h_i and $\tilde{h}_i$ are Ritt polynomials, not type C; that B and $\tilde{B}$ are linear, and A and $\tilde{A}$ are translations; and that*

$$(B \circ h_b, h_{b-1}, \ldots, h_{a+2}, h_{a+1} \circ A)$$

is linearly equivalent to

$$(\tilde{B} \circ \tilde{h}_b, \tilde{h}_{b-1}, \ldots, \tilde{h}_{a+2}, \tilde{h}_{a+1} \circ \tilde{A}).$$

Then $B = \tilde{B}$, and $h_i = \tilde{h}_i$ for all $i > a + 1$, and $\tilde{h}_{a+1} \circ \tilde{A} = h_{a+1} \circ A$. Furthermore, if h_{a+1} is not type A, then also $\tilde{h}_{a+1} = h_{a+1}$ and $A = \tilde{A}$.

Proof. Let $L_b, \ldots, L_{a+2}$ witness linear equivalence:

$$(B \circ h_b \circ L_b, L_b^{-1} \circ h_{b-1} \circ L_{b-1}, \ldots, L_{a+2} \circ h_{a+1} \circ A) = (\tilde{B} \circ \tilde{h}_b, \tilde{h}_{b-1}, \ldots, \tilde{h}_{a+1} \circ \tilde{A}).$$

We induct right-to-left from $i = a + 2$ to show that all $L_i = \mathrm{id}$, and for $i > a + 2$, all $\tilde{h}_{i-1} = h_{i-1}$.

For the base case $i = a + 2$, we have

$$L_{a+2}^{-1} \circ h_{a+1} \circ A = \tilde{h}_{a+1} \circ \tilde{A}.$$

Since the right-hand side of the equation is monic, L_{a+2} is a translation. Since h_{a+1} is not type C, $L_{a+2} = \mathrm{id}$, so $\tilde{h}_{a+1} \circ \tilde{A} = h_{a+1} \circ A$.

For the induction step $a + 1 < i < b$, we have $L_{i-1} = \mathrm{id}$ and $L_i^{-1} \circ h_{i-1} \circ L_{i-1} = \tilde{h}_{i-1}$ with both h_{i-1} and $\tilde{h}_{i-1}$ Ritt polynomials not type C, so $L_i = \mathrm{id}$ and $h_i = \tilde{h}_i$.

Finally, we have $L_b = \mathrm{id}$ and $B \circ h_b \circ L_b^{-1} = \tilde{B} \circ \tilde{h}_b$, which forces $B = \tilde{B}$ and $h_b = \tilde{h}_b$. $\square$

LEMMA 4.8. *Every preclustering admits a cleanup.*

Proof. Let $k = a_r > \cdots > a_1 > a_0 = 0$ be a preclustering of a decomposition $\vec{f}$, and induct on r the number of clusters. For the trivial base case, $L_0 := \mathrm{id}$ is a cleanup of the preclustering with zero clusters of the decomposition with no factors.

The preclustering $k' := a_{r-1} > \cdots > a_1 > a_0 = 0$ of $\vec{f}_{[k',0)}$ has one less cluster, so by induction it admits a cleanup $(\vec{h}_{[k',0)}; \vec{L}_{[k',0]})$. Let $\vec{h}_{[k,k')}$ and linear A_0 and B_0 witness that $\vec{f}_{[k,k')}$ is a cluster.

Now $\vec{f}$ is linearly equivalent to $(B \circ h_k, h_{k-1}, \ldots, h_{k'+1} \circ A \circ L_{k'}, h_{k'} \circ L_{k'-1}, \ldots, h_1 \circ L_0)$. (Here, A and B are A_0 and B_0 or their inverses, depending on kind of cluster.) Set $L_k := B$; replace $L_{k'}$ by $A \circ L_{k'}$; and, if $\vec{f}_{[k,k')}$ is a C-free cluster, apply Lemma 4.6 to $(B \circ h_k, h_{k-1}, \ldots, h_{k'+1} \circ L_{k'})$ to obtain the desired cleanup of $\vec{f}$. $\square$

The next lemma says that cleanups are unique, up to two minor variations arising from Lemmas 3.36 and 4.7: a translation to the right of a type A factor is not well defined, and scalings by ± 1 can appear and disappear as in the following remark.

Remark 4.9. Here is what scalings by -1 can do. Suppose that $(\vec{h}, \vec{L})$ is a cleanup of a preclustering $\vec{a}$ of a decomposition $\vec{f}$, and that $\vec{f}_{[a_{j+1}, a_j)}$ is a C cluster. A new cleanup $(\vec{g}, \vec{M})$ of $\vec{a}$ can be obtained by introducing a scaling by -1 into L_{a_j} and then pushing it left through the cleanup until it is swallowed by a factor of even degree or absorbed into the last linear factor of the cleanup. To state this precisely, let $b := a_j$ to lighten notation.

- For $i < b$, let $g_i := h_i$ and $M_i := L_i$.

- Let $g_b := h_b$ and $M_b := (\cdot - 1) \circ L_b$.
- For $i > b$, let $s_i := \deg(\vec{f}^{\circ}_{[i,b)})$. (We only care about its parity.)
- For $k > i > b$, let $M_i := (\cdot - 1)^{\circ s_i} \circ L_i \circ (\cdot - 1)^{\circ s_i}$ and let $g_i := (-1)^{s_i} * h_i$. Finally, let $M_k := L_k \circ (\cdot - 1)^{\circ s_k}$.

Of course, this can happen several times with different starting points b.

LEMMA 4.10. *Suppose that* $(\vec{h}, \vec{L})$ *and* $(\vec{g}, \vec{M})$ *are two different cleanups of the same preclustering* $\vec{a}$ *of the same decomposition* $\vec{f}$. *Then*

(1) $g_i = (\pm 1) * h_i$ *for all* i *except as in* (3) *below;*

(2) $M_k = L_k \circ (\cdot(\pm 1))$ *and* $M_i = (\cdot(\pm 1)) \circ L_i \circ (\cdot(\pm 1))$ *for all* i *except as in* (3) *below;*

(3) *if* f_{a_j+1} *is type* A, *then there is a translation* T *such that*

$$\tilde{g}_{a_j+1} := g_{a_j+1} \circ T = (\pm 1) * h_{a_j+1} \text{ and } \tilde{M}_{a_j} := T^{-1} \circ M_{a_j} = (\cdot(\pm 1)) \circ L_{a_j} \circ (\cdot(\pm 1)).$$

Remark 4.9 gives more detail about the scalings $(\cdot(\pm 1))$.

Proof. By definition of cleanup, both $(L_k \circ h_k \circ L_{k-1}, h_{k-1} \circ L_{k-2}, \ldots, h_1 \circ L_0)$ and $(M_k \circ g_k \circ M_{k-1}, g_{k-1} \circ M_{k-2}, \ldots, g_1 \circ M_0)$ are linearly equivalent to $\vec{f}$, so they are linearly equivalent to each other. Name the linear factors witnessing this, and then start from the right and induct leftward, exactly as in the proof of Lemma 4.8.

More formally, induct again on the number of clusters, with the same trivial base case of the unique cleanup $(; \mathrm{id})$ of the clusterless preclustering of a decomposition with zero factors. Once again, from the preclustering $k = a_r > \cdots > a_1 > a_0 = 0$ of the decomposition $\vec{f}$, we obtain a preclustering $k' := a_{r-1} > \cdots > a_1 > a_0 = 0$ of $\vec{f}_{[k',0)}$, with one less cluster.

Now $(\vec{g}; (E \circ M_{k'}, M_{k'-1}, \ldots, M_0))$ and $(\vec{h}; (D \circ L_{k'}, L_{k'-1}, \ldots, L_0))$ are both cleanups of this preclustering of $\vec{f}_{[k',0)}$ for some linear D and E. Applying the inductive hypothesis, we get the desired conclusion for h_i with $i \leq k'$ and L_i with $i < k'$, and we also get that

$$(13) \qquad\qquad E \circ M_{k'} = D \circ L_{k'} \circ (\cdot(\pm 1)).$$

Finally, the last cluster $f_{[k,k')}$ is now linearly equivalent to both $(L_k \circ h_k, h_{k-1}, \ldots, h_{k'+1} \circ D^{-1})$ and $(M_k \circ g_k, g_{k-1}, \ldots, g_{k'+1} \circ E^{-1})$.

If this last cluster is a C cluster $g_i = C_{\deg(f_i)} = h_i$ and Lemma 3.36 gives $M_k = L_k \circ (\cdot(\pm 1))$ and $E^{-1} = D^{-1} \circ (\cdot(\pm 1))$. This together with equation (13) gets us the desired conclusion for $M_{k'}$ and $L_{k'}$.

If this last cluster is a C-free cluster, then $M_{k'}$ and $L_{k'}$ are both translations, so in equation (13) the scaling parts of D and E are off by ± 1. That is, for some scaling S and some translations T and U, we have $E^{-1} = (\cdot \pm 1) \circ U \circ S$ and $D^{-1} = T \circ S$. So now we have $(L_k \circ h_k, h_{k-1}, \ldots, h_{k'+1} \circ T)$ linearly equivalent to $(M_k \circ g_k, g_{k-1}, \ldots, g_{k'+1} \circ (\cdot \pm 1) \circ U)$. Once we push the scaling by (± 1)

left through the factors g_i, Lemma 4.7 applies, giving us the desired conclusion for L_k and M_k, and for h_i and g_i for $i \neq k'+1$. If $h_{k'+1}$ is not type A, that lemma also gives the desired conclusion for $h_{k'+1}$ as well as $T = U$, which gives $E = (\cdot \pm 1)D$, and then the desired conclusion for $L_{k'}$ and $M_{k'}$. If $h_{k'+1}$ is type A, we fall into case (3) of the conclusion.   □

The next proposition says that, for a given preclustering and cleanup, any Ritt swap inside a cluster can be witnessed by identity linear factors relative to the factors of that cleanup and so can be accomplished without changing the linear factors of the cleanup.

PROPOSITION 4.11. *Suppose that $(\vec{h}, \vec{L})$ is a cleanup of a preclustering $\vec{a}$ of a decomposition $\vec{f}$; that $t_i \star \vec{f}$ is defined; and that i is not a cluster boundary of $\vec{a}$. Then $\vec{a}$ is also a preclustering of $t_i \star \vec{f}$, and it admits a cleanup $(\vec{g}, \vec{L})$ where $g_{i'} = h_{i'}$ for all $i' \neq i, i+1$, and where $g_{i+1} \circ g_i = h_{i+1} \circ h_i$ is a basic Ritt identity.*

Proof. Inside a C cluster, all factors are type C or degree two, and we have shown (Lemma 3.35) that the only Ritt swaps amongst these come from $C_m \circ C_n = C_n \circ C_m$, which clearly satisfy the conclusion of this proposition. Inside a C-free cluster, this follows immediately from Lemma 3.37.   □

If each decomposition admitted a unique preclustering with maximal clusters, all possible Ritt swaps would be completely described by Proposition 4.11 because of the observation (Remark 4.5(5)) that (f_{i+1}, f_i) is a cluster whenever $t_i \star \vec{f}$ is defined. The following example demonstrates how this can fail.

Remark 4.12. Consider $\vec{f} := (C_3, C_2, x^{17} \cdot (x^2 + 1))$. Clearly, $3 > 1 > 0$ is a preclustering of $\vec{f}$, with a cleanup given by $h_i = f_i$ and $L_i = \mathrm{id}$ for all i. But $3 > 2 > 0$ is also a preclustering, with a cleanup $(C_3, x^2, x^{17} \cdot (x^2 + 1); (\mathrm{id}, +(-2), \mathrm{id}))$. Both Ritt swaps are defined, but for each preclustering, only one of them is inside a cluster. It is also clear that $\vec{f}$ is not a single cluster.

How clusters might fuse and overlap can be read off easily from the linear factors in a cleanup. It is not hard to see that two adjacent clusters of the same kind (both C or both C-free) can be fused into one cluster if and only if the linear factor between them is identity in some cleanup. It is a good deal harder to show that a C cluster and a C-free cluster can only fuse when the C-free cluster is made up of a single factor of degree two. Along the way we show that two overlapping clusters fuse, unless at least one is a C cluster and the overlap is a single factor of degree two.

LEMMA 4.13. *Suppose that $(\vec{h}, \vec{L})$ is a cleanup of a preclustering $k > \cdots > c > b > a > \cdots > 0$ of a decomposition $\vec{f}$; that $\vec{f}_{[c,b)}$ and $\vec{f}_{[b,a)}$ are both C-free clusters; and their concatenation $\vec{f}_{[c,a)}$ is also a cluster. Then $L_b = \mathrm{id}$ and*

$(\vec{h}, \vec{L})$ is also a cleanup of the preclustering $k > \cdots > c > a > \cdots > 0$ with the two clusters fused.

Proof. Since $\vec{f}_{[a,c)}$ is a cluster, $k > \cdots > c > a > \cdots > 0$ is indeed a preclustering of $\vec{f}$, which admits a cleanup $(\vec{g}, \vec{M})$ with $M_b = \mathrm{id}$. Clearly, $(\vec{g}, \vec{M})$ is also a cleanup of the original preclustering $k > \cdots > c > b > a > \cdots > 0$. Apply Lemma 4.10 to compare the translations M_b and L_b sitting to the right of the C-free cluster $\vec{f}_{[a,b)}$ in these two cleanups of the original preclustering. If f_{b+1} is not type A, part (2) of that lemma immediately gives $L_b = M_b$. If f_{b+1} is type A, part (3) of that lemma gives a translation T such that

$$g_{b+1} \circ T = (\pm 1) * h_{b+1} \text{ and } T^{-1} \circ M_b = (\cdot(\pm 1)) \circ L_b \circ (\cdot(\pm 1)).$$

So $g_{b+1} = ((\pm 1) * h_{b+1}) \circ T^{-1} = ((\pm 1) * h_{b+1}) \circ ((\cdot(\pm 1)) \circ L_b \circ (\cdot(\pm 1)))$ is a Ritt polynomial. With the more detailed analysis of $(\cdot(\pm 1)))$ in Remark 4.9, it follows that $h_{b+1} \circ L_b$ is a Ritt polynomial, and $L_b = \mathrm{id}$. $\square$

LEMMA 4.14. *Suppose that $(\vec{h}, \vec{L})$ is a cleanup of a preclustering $k > \cdots > c > b > a > \cdots > 0$ of a decomposition $\vec{f}$; that $\vec{f}_{[c,b)}$ and $\vec{f}_{[b,a)}$ are both C clusters; and their concatenation $\vec{f}_{[c,a)}$ is also a cluster. Then $L_b = (\cdot \pm 1)$.*

If $L_b = \mathrm{id}$, then $(\vec{h}, \vec{L})$ is also a cleanup of the preclustering $k > \cdots > c > a > \cdots > 0$ with the two clusters fused. Otherwise, $L_b = (\cdot - 1)$ and a cleanup of this preclustering may be obtained by pushing the scaling $(\cdot - 1)$ left as in Remark 4.9.

Proof. As in Lemma 4.13, any cleanup of the new preclustering $k > \cdots > c > a > \cdots > 0$ is also a cleanup of the old preclustering, and applying Lemma 4.10 to compare the two cleanups of the old preclustering immediately gives $L_b = (\cdot \pm 1)$. $\square$

It is clear that a C cluster cannot merge with a C-free cluster unless all factors inside the C-free cluster are quadratic. The issue of quadratics is somewhat delicate. The intent of the next definition is that a quadratic needs a gate in the correct direction to get from one cluster to another.

Definition 4.15. Suppose that $(\vec{h}, \vec{L})$ is a cleanup of a preclustering $\vec{a}$ of a decomposition $\vec{f}$. Whether or not there are gates between two of the clusters depends on the kinds (C or C-free) of the two clusters and the linear factor between them. Fix $a_j \neq 0, k$, call $\vec{f}_{[a_{j+1}, a_j)}$ the *left cluster*, call $\vec{f}_{[a_j, a_{j-1})}$ the *right cluster*, and let $L_{a_j} =: L$.

- If the left cluster and the right cluster are both C clusters, this preclustering has
  - a *left-to-right gate* at j if L is a scaling; and
  - a *right-to-left gate* at j if $L = A_\lambda$ or $L = (\cdot - 1) \circ A_\lambda$ for some λ (see Remark 3.14 for the definition of A_λ).

- If the left cluster is a C-free cluster and the right cluster is a C cluster, this preclustering has
    - a *left-to-right gate* at j if $L = \mathrm{id}$ and h_{a_j+1} is a monomial or a Ritt polynomial with in-degree greater than 1; and
    - a *right-to-left gate* at j if $h_{a_j+1} \circ L \circ (-2)$ is a Ritt polynomial.
- If the left cluster is a C cluster and the right cluster is a C-free cluster, this preclustering has
    - a *left-to-right gate* at j if L is a scaling; and
    - a *right-to-left gate* at j if $L = B_\lambda$ or $L = (\cdot - 1) \circ B_\lambda$ for some λ. (See Remark 3.14 for the definition of B_λ.)
- If both are C-free clusters, then this preclustering has   a *two-way gate* at j if $L = \mathrm{id}$.

In general, there is a *two-way gate* whenever there are both a right-to-left gate and a left-to-right gate. Otherwise, there is a *one-way gate*.

Gates are a property of a preclustering and decomposition together, but when $\vec{f}$ is understood, we often say "$\vec{a}$ has a such-and-such gate at j," and vice versa.

PROPOSITION 4.16. *If a cleanup of a preclustering has a two-way gate at* j, *then the* j^{th} *and the* $(j+1)^{st}$ *cluster are of the same kind (both* C *or both* C-free), *and the concatenation* $\vec{f}_{[a_{j+1}, a_{j-1}]}$ *of the two is itself a cluster.*

Proof. Between clusters of different kinds, two-way gates are not possible. If the C cluster is on the left, note that B_λ is never a scaling. If the C cluster is on the right, note that in order for both $h_{a_j+1} \circ L$ and $h_{a_j+1} \circ L \circ (-2)$ to be Ritt polynomials, they must be type A, but that is incompatible with the "monomial or nontrivial in-degree" part of the definition.

Between two C-free clusters, a two-way gate means that the linear factor L_{a_j} is identity by definition. Between two C clusters, a two-way gate means that L_{a_j} is both a scaling and $(\cdot(\pm 1)) \circ A_\lambda$ for some λ, which is only possible for $L_{a_j} = (\cdot(\pm 1))$, which can then be absorbed into or passed through the left C cluster. $\square$

Lemmas 4.13 and 4.14 give the converse of Proposition 4.16.

LEMMA 4.17.
(1) *Gates are properties of the preclustering, independent of the cleanup.*
(2) *In Proposition 4.11, the preclustering of the pre-swap decomposition and the preclustering of the post-swap decomposition obtained there have gates in the same places, in the same directions.*

Proof. For the first part, it is easy to see that the definition of gates is invariant under the few ways listed in Lemma 4.10 for two cleanups of the same decomposition to differ from each other.

For the second part, recall that the kinds (C or C-free) of clusters and the linear factors of the cleanup do not change in Proposition 4.11. Thus, continuing to use the notation from the definition of gates, the only case that needs any work is when the left cluster is a C-free cluster and the right cluster is a C cluster, and (f_{a_j+2}, f_{a_j+1}) are the factors involved in the Ritt-swap.

By Remark 3.33, the $(a_j+1)^{\text{st}}$ factors h_{a_j+1} of the pre-swap decomposition and g_{a_j+1} of the post-swap decomposition are not type A. So for fixed linear M, $h_{a_j+1} \circ M$ is a Ritt polynomial if and only if $M = \mathrm{id}$, if and only if $g_{a_j+1} \circ M$ is a Ritt polynomial. $\qquad\square$

LEMMA 4.18. *If $\vec{a}$ is preclustering of a decomposition $\vec{f}$, then for any linear M and N, $\vec{a}$ is also a preclustering of the decomposition $\vec{g} := (N \circ f_k, f_{k-1}, \ldots, f_2, f_1 \circ M)$, with the same kinds of gates in the same places.*

Proof. For any b and c, it follows immediately from the definition of "cluster" that $\vec{g}_{[c,b)}$ is a cluster if and only if $\vec{f}_{[c,b)}$ is a cluster, so $\vec{a}$ is also a preclustering of $\vec{g}$. It clearly suffices to prove that the gates remain the same in two special cases, when both M and N are scalings, and when both M and N are translations. Let $(\vec{h}, \vec{L})$ be a cleanup of $\vec{f}$.

If both M and N are translations, then one of $(\vec{h}; (N \circ L_k, L_{k-1}, \ldots L_1, L_0 \circ M)$ or $((h_n, \ldots h_2, h_1 \circ L_0 \circ M); (N \circ L_k, L_{k-1}, \ldots L_1, \mathrm{id}))$ is a cleanup of $\vec{g}$. Since the two outside linear factors L_k and L_0 do not contribute to gates in any way, $\vec{g}$ obviously has the same gates as $\vec{f}$.

If both $M := (\cdot \lambda)$ and N are scalings, then a cleanup of $\vec{g}$ is obtained by pushing M left through $(\vec{h}, \vec{L})$ until it hits a C cluster or the leftmost linear factor L_k, as in the proof of the existence of cleanups (Lemma 4.8). That is, the linear factors L_i' of the new cleanup of $\vec{g}$ will be given by $L_i' := \mu_i * L_i$ for $i < j_0$, $L_{j_0}' := L_{j_0} \circ (\cdot \mu_{j_0})$, and $L_i' := L_i$ for $i > j_0$, where j_0 is the index of the rightmost C cluster of $\vec{f}$, or $j_0 = k$ if $\vec{f}$ has no C clusters; and μ_i are integer powers of λ.

The cluster boundaries with $j > j_0$ are clearly unaffected. The cluster boundaries $j < j_0$ lie between two C-free clusters, so $\vec{f}$ has a gate at j if and only if $L_{a_j} = \mathrm{id}$, if and only if $\mu_{a_j} * L_{a_j} = \mathrm{id}$, if and only if $\vec{g}$ also has a gate at j. If j_0 is a cluster boundary, then it has a C cluster on the left and a C-free cluster on the right. Thus $\vec{f}$ has a left-to-right gate at j_0 if and only if $L_{a_{j_0}}$ is a scaling, if and only if $L_{a_{j_0}} \circ (\cdot \mu_{a_{j_0}})$ is a scaling, if and only if $\vec{g}$ also has a left-to-right gate at j. Similarly, $\vec{f}$ has a right-to-left gate at j_0 if and only if $L_{a_{j_0}} = B_\nu = (-2) \circ (\cdot \frac{1}{\nu^2})$ for some ν, if and only if $L_{a_{j_0}} \circ (\cdot \mu_{a_{j_0}}) = (-2) \circ (\cdot \frac{1}{\nu^2}) \circ (\cdot \mu_{a_{j_0}}) = B_{\nu'}$ for $\nu' := \frac{\nu}{\sqrt{\mu_{a_{j_0}}}}$, if and only if $\vec{g}$ also has a right-to-left gate at j. $\qquad\square$

Definition 4.19. If there is a left-to-right gate at j and f_{a_j+1} is quadratic, or if there is a right-to-left gate at j and f_{a_j} is quadratic, we say that this quadratic is a *wandering quadratic*.

If the wandering quadratic f_i is a whole cluster (that is, $a_j = i$ and $a_{j-1} = i - 1$ for some j), then f_i is a *fake wandering quadratic* of this preclustering. Otherwise, it is a *semi-persistent wandering quadratic*.

A one-way gate with no fake wandering quadratics next to it is a *semi-persistent one-way gate*.

Remark 4.20. Like gates, fake and semi-persistent wandering quadratics and semi-persistent one-way gates are properties of the preclustering, independent of the cleanup.

A straightforward exercise in matching Remark 3.14 with the definition of gates shows that the concatenation of a cluster $f_{[b,a)}$ and an adjacent quadratic factor f_a (respectively, f_{b+1}) is a cluster if and only if any preclustering of $\vec{f}$ with $a_{j+1} = b$ and $a_j = a$ has a right-to-left gate at j (respectively, a left-to-right gate at $j + 1$). This implies that the adjacent factor is a wandering quadratic. Almost conversely, if the adjacent factor is a semi-persistent wandering quadratic, then the concatenation is a cluster.

LEMMA 4.21. *If f_{a_j} (respectively, f_{a_j+1}) is a semi-persistent wandering quadratic of a preclustering $\vec{a}$ of a decomposition $\vec{f}$, then $\vec{b}$ given by $b_j := a_j - 1$ (respectively, $b_j := a_j + 1$) and $b_i = a_i$ for all $i \neq j$ is another preclustering of $\vec{f}$. For any cleanup $(\vec{h}, \vec{L})$ of $\vec{a}$, there is a cleanup $(\vec{g}, \vec{M})$ of $\vec{b}$ with $g_i = h_i$ for $i \neq j$ (respectively, $i \neq j + 1$) and with $M_i = L_i$ for $i \neq j - 1, j$ (respectively, for $i \neq j, j + 1$). In particular, $\vec{a}$ and $\vec{b}$ have the same gates at all $j' \neq j$.*

Proof. First, we show that all $\vec{f}_{[b_i, b_{i-1})}$ are clusters. One of these is the concatenation of a cluster of $\vec{a}$ and an adjacent semi-persistent wandering quadratic of $\vec{a}$, so it is a cluster by Remark 4.20. Another is a cluster of $\vec{a}$ that lost a semi-persistent wandering quadratic, so it is nonempty by semi-persistence and a cluster by Remark 4.5. The rest are clusters of $\vec{a}$.

The rest of the proof is the same straightforward exercise in matching Remark 3.14 with the definition of gates as in the second part of Remark 4.20. □

The next lemma somewhat justifies the terminology "semi-persistent."

LEMMA 4.22. *Suppose that two preclusterings $\vec{a}$ and $\vec{b}$ of the same decomposition $\vec{f}$ only differ at one place j and only by 1; that is,*

$$k = a_r = b_r > \cdots > a_{j+1} = b_{j+1} > b_j = a_j + 1 > a_j > a_{j-1}$$
$$= b_{j-1} > \cdots > a_0 = b_0 = 0.$$

Then $\vec{a}$ and $\vec{b}$ agree on whether the contested factor f_{b_j} is a wandering quadratic; and if it is, they also agree on whether the gate at j is one-way or two-way. Of course, if the gate is one-way, it goes in different directions for the two preclusterings.

Proof. Since $a_{j+1} = b_{j+1} > b_j = a_j + 1 > a_j$, the contested factor f_{b_j} is semi-persistent if it is a wandering quadratic. This keeps other nontrivial linear factors of the cleanup from interfering. This proof is another straightforward exercise in matching Remark 3.14 with the definition of gates.   □

We now return to the question of fusing and overlapping clusters. The next Lemma 4.23 serves two purposes. First, it describes a way for the concatenation of a C cluster and a C-free cluster to be itself a cluster. Lemma 4.24 asserts that this is the only way. Second, Lemma 4.23 states that if the concatenation of two clusters is not a cluster, but a quadratic factor can enter one cluster from the other, then a gate in the correct direction must be present at this cluster boundary in the original preclustering, and a gate in the other direction is present in the new preclustering; thus, no other quadratic can follow this one. A similar result is mentioned on page 4 of [21] but is not explicitly stated as a theorem in the text.

LEMMA 4.23. *Fix a preclustering $\vec{a}$ of a decomposition $\vec{f}$, and suppose that the concatenation of a C cluster and an adjacent factor f_i is a cluster. Then either f_i is a wandering quadratic, or it comes from another C cluster, and the concatenation of these two clusters is itself a cluster.*

Proof. To state the lemma more precisely and less readably, let $c := a_{j+1}$, $b := a_j$, and $a := a_{j-1}$ to lighten notation. The lemma then says:

If $[c, \vec{b})$ is a C cluster and the concatenation $\vec{f}_{[c,b-1)}$ of it and the next factor f_b is a cluster, then one of the following happens.

(1) The whole concatenation $\vec{f}_{[c,a)}$ of the two clusters if itself a cluster; the other cluster $[b, \vec{a})$ either is a C cluster, or has only one (quadratic) factor so $b = a + 1$.

(2) The factor f_b is a semi-persistent wandering quadratic, and Lemma 4.22 applies to the original preclustering and the preclustering $k > \cdots > c > b - 1 > a > \cdots > 0$.

Similarly, if $\vec{f}_{[b,a)}$ is a C cluster and the concatenation $\vec{f}_{[b+1,a)}$ of it and the next factor f_{b+1} is a cluster, then either the whole $\vec{f}_{[c,a)}$ is a cluster and the other cluster $\vec{f}_{[c,b)}$ is a C cluster or a single quadratic factor; or f_{b+1} is a wandering quadratic next to a one-way gate in both preclusterings.

The new factor f_b or f_{b+1} joining the C cluster must be type C or quadratic. A type C factor must come from a C cluster, and then the whole $\vec{f}_{[c,a)}$ is a cluster

by Lemma 4.14 and Proposition 4.16. For the rest of the proof, we assume that the new factor is quadratic, and a wandering quadratic by the first part of Remark 4.20.

If the gate next to this wandering quadratic is two-way, by Proposition 4.16 the other cluster must be a C cluster, and the whole $\vec{f}_{[c,a)}$ is a cluster.

If the new quadratic factor is a fake wandering quadratic (that is, a whole cluster), the whole $\vec{f}_{[c,a)}$ is precisely the thing assumed to be a cluster in the first place. □

LEMMA 4.24. *Fix a preclustering $\vec{a}$ of a decomposition $\vec{f}$, and suppose that the concatenation of a C cluster and more than one factor of a neighboring cluster is itself a cluster. Then the concatenation of these two clusters is itself a cluster, and the other cluster is a C cluster.*

Proof. Again, let $c := a_{j+1}$, $b := a_j$, and $a := a_{j-1}$ to lighten notation. Suppose the new factors $f_b, \ldots, f_{b'+1}$ are to the right of the C cluster $\vec{f}_{[c,b)}$, all inside the next cluster $\vec{f}_{[b,a)}$. The other case, when the new factors are to the left of the C cluster, is essentially identical.

Note that $\vec{f}_{[c,d)}$ is a C cluster for any d with $b \geq d \geq b'$, because it sits inside a C cluster and has a C cluster sitting inside it.

In particular, the concatenation $\vec{f}_{[c,b-1)}$ of the C cluster $\vec{f}_{[c,b)}$ and one factor f_b is a cluster, so Lemma 4.23 applies. If the other cluster $\vec{f}_{[b,a)}$ is a C cluster and the concatenation $\vec{f}_{[c,a)}$ is a cluster, we are done.

Otherwise, f_b is a wandering quadratic and we work toward a contradiction. Since more than one factor from $\vec{f}_{[b,a)}$ does something in the hypothesis of this lemma, f_b cannot be fake and must be a semi-persistent quadratic. Thus the new preclustering $k = a_r > \cdots > a_{j+1} > a_j - 1 > a_{j-1} > \cdots > a_0 = 0$ from Lemma 4.22 has a one-way left-to-right gate at j. More than one factor from $\vec{f}_{[b,a)}$ joins $\vec{f}_{[c,b)}$, so $\vec{f}_{[c,b-2)}$ must also be a cluster. If f_{b-1} is quadratic, then by Remark 4.20 there must be a right-to-left gate at j, a contradiction. If f_{b-1} is not quadratic, then Lemma 4.14 implies that there is a two-way gate at j, also a contradiction. □

LEMMA 4.25. *Fix a decomposition $\vec{f}$. If two clusters overlap, that is, $\vec{f}_{[d,b)}$ and $\vec{f}_{[c,a)}$ are both clusters with $d > c > b > a$, then either the whole $\vec{f}_{[d,a)}$ is a cluster, or one of the clusters is a C cluster and $c = b + 1$ and $f_{b+1} = f_c$ is a semi-persistent wandering quadratic of every preclustering $\cdots d > c > a > \cdots$ of $\vec{f}$.*

Proof. Consider the three pieces $\vec{f}_{[d,c)}$, $\vec{f}_{[c,b)}$, and $\vec{f}_{[b,a)}$. If both original clusters $\vec{f}_{[d,b)}$ and $\vec{f}_{[c,a)}$ are C-free clusters, then all three of those pieces are C-free clusters, and Lemma 4.13 forces the whole $\vec{f}_{[d,a)}$ to be a C-free cluster.

If both $\vec{f}_{[d,b)}$ and $\vec{f}_{[c,a)}$ are C clusters and the middle piece $\vec{f}_{[c,b)}$ is a C cluster, then the whole $\vec{f}_{[d,a)}$ is a cluster by Lemma 4.14.

The only possibility left is that at least one of $\vec{f}_{[d,b)}$ and $\vec{f}_{[c,a)}$ is a C cluster, but the middle piece $\vec{f}_{[c,b)}$ is not a C cluster. We handle the case when $\vec{f}_{[d,b)}$ is a C cluster; the other case is identical.

Since $\vec{f}_{[c,b)}$ sits inside a C cluster but is not a C-cluster itself, it must be linearly related to C_{2^M} for some M. Thus the leftover $f_{[d,c)}$ of the C cluster $f_{[d,b)}$ is still a C cluster. If $M \geq 2$, this C cluster $f_{[d,c)}$ absorbs more than one factor from the adjacent cluster $f_{[c,a)}$, so by Lemma 4.24, $f_{[c,a)}$ is again a C cluster and the whole $\vec{f}_{[d,a)}$ is a cluster. The case $M = 1$ is precisely the last option in the conclusion of this lemma. □

The next definition characterizes preclusterings that have as few clusters as possible. The next few results build up to show that this minimal number of clusters, as well as the presence or absence of a gate at each cluster boundary, are properties of a polynomial, independent of decomposition; and that cluster boundaries can only change by 1 between two clusterings, and then only because of quadratic factors.

Definition 4.26. A preclustering $\vec{a}$ of $\vec{f}$ is a *clustering* if

- the concatenation of any two adjacent clusters is not a cluster, and
- no cluster consists of only wandering quadratics.

(The second part is only relevant for clusters with exactly two factors.)

Remark 4.27. The second part of the definition of "clustering" can be replaced by

- No cluster with exactly two factors can be devoured by adjacent clusters: if $a_i = a_{i-1} + 2$, then $\vec{f}_{[a_{i+1},a_i-1)}$ and $\vec{f}_{[a_i-1,a_{i-2})}$ are not both clusters.

It follows from the first part of the definition of "clustering" that if this condition is violated, both factors of the violating cluster must be wandering quadratics. The advantage of this formulation is that it is entirely in terms of which $\vec{f}_{[c,b)}$ are clusters.

LEMMA 4.28. *If $\vec{a}$ is clustering of a decomposition $\vec{f}$, then for any linear M and N, $\vec{a}$ is also a clustering of the decomposition*

$$\vec{g} := (N \circ f_k, f_{k-1}, \ldots, f_2, f_1 \circ M).$$

Proof. By Lemma 4.18, $\vec{a}$ is a preclustering of $\vec{g}$. As noted in the proof of that lemma, for any b and c, $\vec{g}_{[c,b)}$ is a cluster if and only if $\vec{f}_{[c,b)}$ is a cluster. As noted in Remark 4.27, this suffices to show that $\vec{a}$ is a clustering of $\vec{g}$. □

Remark 4.29. By Proposition 4.16, there are no two-way gates in a clustering. Thus, between two C-free clusters of a clustering, there are neither gates nor wandering quadratics. There are no fake wandering quadratics in a clustering. Thus, all gates in a clustering are semi-persistent one-way gates.

LEMMA 4.30. *If all factors of $\vec{f}$ are swappable, then $\vec{f}$ admits a clustering.*

Proof. We already know that $\vec{f}$ admits a preclustering, so take one, and induct on the number of clusters in it: with Remark 4.27, it is clear how to rectify a violation of either of the two extra requirements of a clustering, and both decrease the number of clusters. □

LEMMA 4.31. *Suppose that $\vec{a}$ is a clustering of $\vec{f}$, that i is not a cluster boundary of $\vec{a}$, and that $\vec{g} := t_i \star \vec{f}$ is defined. Then $\vec{a}$ is also a clustering of $\vec{g}$.*

Proof. By Proposition 4.11, $\vec{a}$ is a preclustering of $\vec{g}$. The rest is trivial. □

LEMMA 4.32. *If f_{a_j} (respectively, f_{a_j+1}) is a wandering quadratic of a clustering $\vec{a}$ of a decomposition $\vec{f}$, then $\vec{b}$ given by $b_j := a_j - 1$ (respectively, $b_j := a_j + 1$) and $b_{j'} = a_{j'}$ for all $j' \neq j$ is another clustering of $\vec{f}$. For each j', either both clusterings have a gate at j', or neither has a gate at j'.*

Proof. We first show that $\vec{b}$ is a preclustering of $\vec{f}$, then that $\vec{b}$ has gates in the same places as $\vec{a}$, and then that $\vec{b}$ is a clustering.

By Remark 4.29, the clustering $\vec{a}$ has no two-way gates and no fake wandering quadratics, so f_{a_j} (respectively, f_{a_j+1}) is a semi-persistent wandering quadratic of $\vec{a}$. Then $\vec{b}$ is a preclustering of $\vec{f}$ by Lemma 4.21 and has the same gates as $\vec{a}$ by Lemmas 4.21 and 4.22. It remains to show that $\vec{b}$ is a clustering.

We first show that for every i, the factor f_i is a wandering quadratic of $\vec{a}$ if and only if it is a wandering quadratic of $\vec{b}$. For $i = a_j$ (respectively, $i = a_j + 1$), this is Lemma 4.22. For all other cluster boundaries, this is Lemma 4.21. It is clear that the two factors adjacent to f_{a_j} (respectively, f_{a_j+1}) are not wandering quadratics of $\vec{b}$, so it remains to show that they are not wandering quadratics of $\vec{a}$. By Remark 4.29, the clustering $\vec{a}$ has no two-way gates and no fake wandering quadratics, so f_{a_j+1} (respectively, f_{a_j}), the factor on the other side of the cluster boundary, is not a wandering quadratic of $\vec{a}$. Since the jth (respectively, $(j+1)^{\text{st}}$) cluster of $\vec{a}$ does not consist of two wandering quadratics, f_{a_j-1} (respectively, f_{a_j+2}) is also not a wandering quadratic of $\vec{a}$. From now on we say "wandering quadratic" without specifying $\vec{a}$ or $\vec{b}$.

Let us now verify the two parts of the definition of clustering for the two clusters of $\vec{b}$ that differ from those of $\vec{a}$. We handle the case when f_{a_j+1} is the wandering quadratic; the other case is analogous.

Two instances of the second part of the definition of clustering need to be verified.

- Since the jth cluster $\vec{f}_{[a_j,a_{j-1})}$ of $\vec{a}$ is not a single wandering quadratic, the jth cluster $\vec{f}_{[a_j+1,a_{j-1})}$ of $\vec{b}$ does not consist of two wandering quadratics.
- We have already shown that the rightmost factor $f_{b_j+1} = f_{a_j+2}$ of the $(j+1)^{\text{st}}$ cluster of $\vec{b}$ is not a wandering quadratic.

Three instances of the first part of the definition of clustering need to be verified.

- The concatenation $\vec{f}_{[a_j+1,a_{j-2})}$ of the jth and the $(j-1)^{\text{st}}$ clusters of $\vec{b}$ is not a cluster because it is the concatenation of the quadratic $f_{b_j} = f_{a_j+1}$ and $\vec{f}_{[a_j,a_{j-2})}$, and this $\vec{f}_{[a_j,a_{j-2})}$ is not a cluster because it is the concatenation of the corresponding clusters of $\vec{a}$.
- The concatenation of the $(j+1)^{\text{st}}$ and the jth clusters of $\vec{b}$ is exactly the same as the concatenation of the $(j+1)^{\text{st}}$ and the jth clusters of $\vec{a}$, so it is not a cluster.
- If the concatenation $\vec{f}_{[a_j+2,a_j+1)}$ of the $(j+2)^{\text{nd}}$ and the $(j+1)^{\text{st}}$ clusters of $\vec{b}$ is a cluster, then it overlaps the $(j+1)^{\text{st}}$ cluster $\vec{f}_{[a_j+1,a_j)}$ of $\vec{a}$ in $\vec{f}_{[a_j+1,a_j+1)}$. We have already shown that the rightmost factor of this overlap f_{a_j+2} is not a wandering quadratic, so by Lemma 4.25 the concatenation $\vec{f}_{[a_j+2,a_j)}$ of these overlapping clusters is itself a cluster. But that is also the concatenation of two clusters of $\vec{a}$, which cannot be a cluster. □

PROPOSITION 4.33. *Suppose that $\vec{a}$ is a clustering of $\vec{f}$. If (f_{a_j+1}, f_{a_j}) is a cluster and one of f_{a_j+1} and f_{a_j} is not quadratic, then the other one of f_{a_j+1} and f_{a_j} is a semi-persistent wandering quadratic, and one of the clusters $\vec{f}_{[a_j+1,a_j)}$ and $\vec{f}_{[a_j,a_{j-1})}$ is a C cluster.*

Proof. Lemma 4.25 applies to the clusters $\vec{f}_{[a_j+1,a_j)}$ and (f_{a_j+1}, f_{a_j}) that overlap in f_{a_j+1}. If f_{a_j+1} is a wandering quadratic and $\vec{f}_{[a_j+1,a_j)}$ is a C cluster, we are done. If f_{a_j+1} is a wandering quadratic and (f_{a_j+1}, f_{a_j}) is a C cluster, then f_{a_j} is type C, so $\vec{f}_{[a_j,a_{j-1})}$ is a C cluster, and we are done. Otherwise, the whole $\vec{f}_{[a_j+1,a_j-1)}$ is a cluster.

Now Lemma 4.25 applies to this $\vec{f}_{[a_j+1,a_j-1)}$ and $\vec{f}_{[a_j,a_{j-1})}$, which overlap in f_{a_j}. The whole $\vec{f}_{[a_j+1,a_j-1)}$ is the concatenation of two clusters of $\vec{a}$, so it cannot be a cluster. Thus f_{a_j} is a wandering quadratic and one of $\vec{f}_{[a_j+1,a_j-1)}$ and $\vec{f}_{[a_j,a_{j-1})}$ is a C cluster. It is easy to see that $\vec{f}_{[a_j+1,a_j)}$ is a C cluster whenever $\vec{f}_{[a_j+1,a_j-1)}$ is, so we are done in either case. □

LEMMA 4.34. *If $\vec{a}$ is a clustering of $\vec{f}$ and $t_{a_j} \star \vec{f} = \vec{g}$ is defined, then there is a clustering $\vec{b}$ of $\vec{g}$ with $b_{j'} = a_{j'}$ for all $j' \neq j$. At each $j' \neq j$, the clustering $\vec{a}$ of $\vec{f}$ and the clustering $\vec{b}$ of $\vec{g}$ have the same gate(s). As for b_j, either*

- *f_{a_j+1} is a wandering quadratic of the clustering $\vec{a}$ of $\vec{f}$, and $b_j := a_j + 1$, and the clustering $\vec{b}$ of $\vec{g}$ has a one-way right-to-left gate at b_j; or*
- *f_{a_j} is a wandering quadratic of the clustering $\vec{a}$ of $\vec{f}$, and $b_j := a_j - 1$, and the clustering $\vec{b}$ of $\vec{g}$ has a one-way left-to-right gate at b_j.*
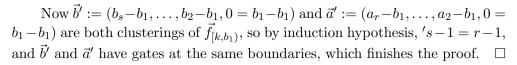
Proof. As noted in Remark 4.5, since $t_{a_j} \star \vec{f}$ is defined, (f_{a_i+1}, f_{a_i}) is a cluster. As tautological Ritt swaps are not permitted, f_{a_j+1} and f_{a_j} are not both quadratic. By Proposition 4.33 one of f_{a_i+1} and f_{a_i} is a semi-persistent wandering quadratic. By Lemma 4.32, $\vec{b}$ is a clustering of $\vec{f}$. Since a_j is not a cluster boundary of $\vec{b}$, by Lemma 4.31, $\vec{b}$ is also a clustering of $\vec{g}$. By the second part of Lemma 4.17, $\vec{f}$ and $\vec{g}$ have the same gates in the same places with respect to $\vec{b}$. □

With the next lemma and proposition we show that two clusterings of the same decomposition can only differ by putting wandering quadratics into different clusters.

LEMMA 4.35. *Any two clusterings $\vec{a}$ and $\vec{b}$ of the same decomposition $\vec{f}$ have the same number of clusters. For each j, either $b_j = a_j - 1$ (respectively, $b_j = a_j + 1$) and f_{a_j} (respectively, f_{a_j+1}) is a wandering quadratic of both, or $b_j = a_j$. At each j, either both clusterings have a gate at j, or neither has a gate at j.*

Proof. Let $\vec{a}$ of length r and $\vec{b}$ of length s be two clusterings of the same decomposition $\vec{f}$. Intuitively, we start from the right and match clusters of $\vec{b}$ with those of $\vec{a}$ one at a time. More formally, we induct on the number r of clusters in $\vec{a}$. For the base case of the induction, take $r = 1$, i.e., the whole decomposition is a single cluster; clearly, no other clustering is possible.

For the induction step, we first match the rightmost clusters of the two clusterings; that is, we show that $k = a_r > \cdots > a_2 > b_1 > b_0 = a_0 = 0$ is another clustering of $\vec{f}$, with gates at the same places as $\vec{a}$. If $a_1 = b_1$, we are done. Suppose that $b_1 > a_1$. Since the concatenation $\vec{f}_{[a_2,a_0)}$ of the first two clusters of $\vec{a}$ is not a cluster, and $\vec{f}_{[a_2,a_1)}$ is not a single wandering quadratic, this $\vec{f}_{[a_2,a_0)}$ cannot be contained in the first cluster of $\vec{b}$. That is, $a_2 > b_1$. Applying Lemma 4.25 to $\vec{f}_{[a_2,a_1)}$ and $\vec{f}_{[b_1,b_0)}$, we see that $b_1 = a_1 + 1$ and f_{b_1} is a wandering quadratic of $\vec{a}$. Identical reasoning shows that if $a_1 > b_1$, then $a_1 = b_1 + 1$ and f_{a_1} is a wandering quadratic of $\vec{a}$. In any case, Lemma 4.32 makes $k = a_r > \cdots > a_2 > b_1 > b_0 = a_0 = 0$ another clustering of $\vec{f}$, with gates at the same places as $\vec{a}$.

Now $\vec{b}' := (b_s - b_1, \dots, b_2 - b_1, 0 = b_1 - b_1)$ and $\vec{a}' := (a_r - b_1, \dots, a_2 - b_1, 0 = b_1 - b_1)$ are both clusterings of $\vec{f}_{[k,b_1)}$, so by induction hypothesis, $'s - 1 = r - 1$, and $\vec{b}'$ and $\vec{a}'$ have gates at the same boundaries, which finishes the proof. $\square$

PROPOSITION 4.36. *Given a clustering $\vec{a}$ of $\vec{f}$, another tuple $\vec{b}$ of the same length as $\vec{a}$ is a clustering of $\vec{f}$ if and only if for each j where $a_j \neq b_j$, there is a wandering quadratic f_i such that $\{a_j, b_j\} = \{i, i+1\}$.*

Proof. The "if" follows from applying Lemma 4.32 at each j where $a_j \neq b_j$. The "only if" follows immediately from Lemma 4.35. $\square$

THEOREM 4.37. *The number of clusters in a clustering, the kind (C or C-free) of each cluster, and the presence of a gate between the jth and the $(j+1)^{st}$ clusters are properties of the polynomial, independent of decomposition, clustering, and cleanup.*

Proof. It suffices to show that these are invariant under Ritt swaps, and we have in fact already done so. Suppose that $\vec{f}$ and $\vec{g} := t_i \star \vec{f}$ are two decompositions of the same polynomial and $\vec{a}$ is a clustering of $\vec{f}$.

If $i \neq a_j$ for any j, Lemma 4.31 shows that $\vec{a}$ is also a clustering of $\vec{g}$, and Lemma 4.17 shows it has all the same gates in the same places. Wandering quadratics are unchanged from $\vec{f}$ to $\vec{g}$ unless the swap brings a quadratic to a gate or moves a quadratic inside a cluster away from the gate.

If $i = a_j$ for some j, then Lemma 4.34 shows that Lemma 4.23 applies and gives a clustering of $\vec{g}$ with a_j moved left or right by one; in any case, gates in this new clustering are the same as in the old clustering, except that the one at j switches direction. $\square$

The next few results begin to apply the technical machinery around clusterings to the action of the Ritt monoid.
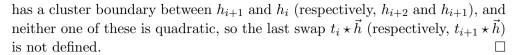
LEMMA 4.38. *If $t_{i+1} \star \vec{f}$ and $t_i \star \vec{f}$ are both defined, then either (f_{i+2}, f_{i+1}, f_i) is a cluster, or f_{i+1} is a wandering quadratic (in any clustering of $\vec{f}$).*

Proof. As remarked in Remark 4.5, both (f_{i+2}, f_{i+1}) and (f_{i+1}, f_i) must be clusters, and Lemma 4.25 finishes the proof. $\square$

LEMMA 4.39. *If $t_i t_{i+1} t_i \star \vec{f}$ (respectively, $t_{i+1} t_i t_{i+1} \star \vec{f}$) is defined, then (f_{i+2}, f_{i+1}, f_i) is a cluster.*

Proof. Note that in either sequence, each factor swaps with every other factor. So as long as at least one of the two is defined, all of f_i, f_{i+1}, and f_{i+2} are swappable, and no more than one is quadratic.

Lemma 4.38 applies to the first intermediate decomposition $\vec{g} := t_i \star \vec{f}$ (respectively, $\vec{g} := t_{i+1} \star \vec{f}$). If $\vec{g}$ is a cluster, we are done. Otherwise, the second intermediate decomposition $\vec{h} := t_{i+1} t_i \star \vec{f}$ (respectively, $\vec{h} := t_i t_{i+1} \star \vec{f}$)

has a cluster boundary between h_{i+1} and h_i (respectively, h_{i+2} and h_{i+1}), and neither one of these is quadratic, so the last swap $t_i \star \vec{h}$ (respectively, $t_{i+1} \star \vec{h}$) is not defined. $\square$

Our fundamental Theorem 2.52 is now an easy corollary.

Proof. (This is the proof of Theorem 2.52.) We need to show that $t_{i+1}t_it_{i+1} \star \vec{f}$ is defined if and only if $t_it_{i+1}t_i \star \vec{f}$ is defined, and they are equal when defined.

By Lemma 4.39, the whole (f_{i+2}, f_{i+1}, f_i) must be a cluster, so Ritt swaps can be witnessed by identity linear factors, and the result is immediate. $\square$

Let us prove two more statements of this flavour, with a view towards normal forms.

PROPOSITION 4.40. *Suppose that $\vec{h} := (t_{k-1}t_{k-2} \cdots t_1) \star \vec{f}$ is defined, and let $\vec{a}$ be a clustering of $\vec{f}$ with $r > 1$ clusters. Then f_1 is quadratic and there are (one-way) right-to-left gate at every $j \neq 0, r$. Furthermore, $\vec{b}$ given by $b_j = a_j - 1$ for all $j \neq 0, r$ is a clustering of $\vec{h}$, which has (one-way) left-to-right gates at every $j \neq 0, r$.*

Proof. If $k = 2$ and $t_1 \star \vec{f}$ is defined, then the whole $\vec{f}$ is a cluster by Remark 4.5, contradicting the hypothesis that $r > 1$. So $k - 1 \geq 2$.

For $0 \leq i < k$, let $\vec{f^i} := (t_it_{i-1} \ldots t_1) \star \vec{f}$. We inductively define clusterings $\vec{a^i}$ of $\vec{f^i}$, starting with $\vec{a^0} = \vec{a}$. As long as i is not a cluster boundary, Lemma 4.31 applies to $t_i \star \vec{f^{i-1}} = \vec{f^i}$, so $\vec{a^i} := \vec{a^{i-1}}$ works. In particular, $\vec{a}$ is still a clustering of $\vec{f}^{a_1-1}$.

Since both $t_{a_1} \star \vec{f}^{a_1}$ and $t_{a_1+1} \star \vec{f}^{a_1}$ are defined, by Lemma 4.38, either $(f^{a_1}_{a_1+2}, f^{a_1}_{a_1+1}, f^{a_1}_{a_1})$ is a cluster, or $f^{a_1}_{a_1+1}$ is a wandering quadratic. The first option is impossible, as it would force the first two clusters of $\vec{f}^{a_1}$ to fuse (Lemma 4.25).

Since f_1 becomes $f^{a_1}_{a_1+1}$ via $\vec{f}^{a_1} = t_{a_1} \ldots t_1 \star \vec{f}$ in the sense of Remark 2.46, f_1 must also be quadratic. Since f_1 becomes f^{i-1}_i via $\vec{f^i} := (t_it_{i-1} \ldots t_1) \star \vec{f}$ in the same sense, f^{i-1}_i are also quadratic for each i.

We now continue to inductively define clusterings $\vec{a^i}$ of $\vec{f^i}$. If i is not a cluster boundary of $\vec{a^{i-1}}$, then Lemma 4.31 applies to $t_i \star \vec{f^{i-1}} = \vec{f^i}$, so $\vec{a^i} := \vec{a^{i-1}}$ works. If $i = a^{i-1}_j$ is a cluster boundary of $\vec{a^{i-1}}$, then Lemma 4.34 applies to $t_i \star \vec{f^{i-1}} = \vec{f^i}$. We already know that f^{i-1}_i is quadratic, so the second option in Lemma 4.34 happens. (The first option cannot happen because at most one factor in a Ritt swap can be quadratic.) Thus, $\vec{a^i}$ with $a^i_j = a^{i-1}_j - 1$ and $a^i_{j'} = a^{i-1}_{j'}$ for all $j' \neq j$ is a clustering of $\vec{f^i}$. It is now clear that $\vec{b} := \vec{a^{k-1}}$ is as wanted. $\square$

LEMMA 4.41. *Suppose that $k \geq 3$, let $v_1 := t_{k-1}t_{k-2}\cdots t_2$ and $v_2 := t_{k-2}t_{k-3}\cdots t_1$, and suppose that $v_2v_1 \star \vec{f}$ is defined. Then the whole $\vec{f}$ is a cluster, unless $k = 3$ and f_3 is quadratic.*

Proof. Let $\vec{g} := v_1 \star \vec{f}$, and note that Proposition 4.40 applies to both (1) $v_i \star \vec{f}_{[k,1)} = \vec{g}_{[k,1)}$ and (2) $v_2 \star \vec{g}_{[k-1,0)}$. If $f_1 = g_1$ is not quadratic, (2) makes $\vec{g}_{[k-1,0)}$ a cluster. If f_2 is not quadratic, (1) makes both $\vec{g}_{[k,1)}$ and $\vec{h}_{[k,1)}$ a cluster. If either f_1 or f_2 is quadratic, then none of the other factors f_i with $i \geq 3$ are quadratic, since we do not allow tautological Ritt identities.

We now treat the four cases separately.

Case 1. If neither f_1 nor f_2 is quadratic, then Lemma 4.25 applies to the clusters $\vec{g}_{[k-1,0)}$ and $\vec{g}_{[k,1)}$. If $\vec{g}$ is a cluster, then so is $\vec{f}$. Otherwise, the overlap $\vec{g}_{[k-1,1)}$ is a single quadratic g_2, so $k = 3$, $v_1 = t_2$, and f_3 is quadratic.

Case 2. Suppose $f_1 = g_1$ is not quadratic but f_2, and therefore g_k, is. Now $(g_{k-1}, \ldots, g_2, g_1 = f_1)$ is a cluster. On the other hand, $t_{k-1}\vec{g} = t_{k-2}\cdots t_2 \star \vec{f}$ is defined, so (g_k, g_{k-1}) is also cluster, and g_{k-1} is not quadratic. Then by Lemma 4.25 the whole $\vec{g}$ (and, therefore, $\vec{f}$) is a cluster as desired.

Case 3. Suppose $f_1 = g_1$ is quadratic but f_2, and therefore g_k, is not. Now $(g_k, \ldots, g_2)$ is a cluster. But $v_2 \star \vec{g}$ is defined, so $t_1 \star \vec{g}$ is defined, so (g_2, g_1) is a cluster. Since g_2 is not quadratic, by Lemma 4.25 the whole $\vec{g}$ (and, therefore, $\vec{f}$) is a cluster as desired.

Case 4. Finally, suppose that both f_1 and f_2 are quadratic. If $(f_k, \ldots, f_3)$ is not a cluster, according to Proposition 4.40 there must be a right-to-left gate at every boundary between clusters inside there, which becomes a one-way left-to-right in the corresponding place in $(g_{k-1}, \ldots, g_2)$. That, according to the same proposition, makes it impossible for $v_2 \star \vec{g}$ to be defined. So $(f_k, \ldots, f_3)$ is a cluster. Since $t_2 \star \vec{f}$ is defined, (f_3, f_2) is also a cluster. Since f_3 is not quadratic, this makes $(f_k, \ldots, f_2)$ a cluster. Then $(g_k, \ldots, g_2)$ is also a cluster. Since $t_1 \star \vec{g}$ is defined, $(g_2, g_1))$ is a cluster. Since g_2 is not quadratic, this means the whole $\vec{g}$ is a cluster. So $\vec{f}$ is a cluster. $\square$

5. Canonical forms

Much of this section is devoted to using syntactic operations on words in the Ritt monoid RM_k that appear in Theorem 2.52 to show that any decomposition of a polynomial may be obtained from any other by a sequence of Ritt swaps *in a particular canonical order*. The words in the Ritt monoid RM_k corresponding to such sequences are said to be in a canonical form.

If one thinks of permuting factors as putting them in a particular order, then our first canonical form roughly corresponds to an insert-sort, and the

second one to a merge-sort. While it is well known that every permutation is represented by a sequence of transpositions of each of these forms, we could not find a reference in the literature for the more refined results taking into account the irreversibility of operation (1) in Remark 5.1 below.

For each of our two canonical forms, we show (see Propositions 5.11 and 5.15) that for any word $w \in \mathrm{RM}_k$, there is a word $\hat{w}$ of this canonical form such that whenever $w \star \vec{f}$ is defined, $\hat{w} \star \vec{f} = w \star \vec{f}$. For example, for $w = t_i t_i$, we set $\hat{w}$ to be the empty word.

While it is convenient to speak of the factors of $\vec{f}$ in the statements and proofs of intermediate results, the canonical word $\hat{w}$ ultimately only depends on w and works for all $\vec{f}$.

Remark 5.1. Recall the three syntactic operations on words in the Ritt monoid RM_k from Theorem 2.52.

(1) Delete subword $t_i t_i$.
(2) Replace subword $t_i t_j$ by $t_j t_i$ for nonconsecutive i and j.
(3) Replace subword $t_{i+1} t_i t_{i+1}$ by $t_i t_{i+1} t_i$, or vice versa.

Operations (2) and (3) are reversible, while (1) is not. Operation (1) decreases the length of the word, while (2) and (3) leave it the same.

Remark 5.2. If a word v is obtained from a word w by operations (1), (2) and (3) above, then they represent the same permutation (see Definition 2.51), 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.

If v and w also have the same length or, equivalently, if one was obtained from the other by operations (2) and (3) alone, then $v \star \vec{f} = w \star \vec{f}$ for all $\vec{f}$.

These observations motivate the following definitions.

Definition 5.3. If two words $v, w \in \mathrm{RM}_k$ can be obtained from each other by operations (2) and (3) above, we write $w \simeq v$. A word $w \in \mathrm{RM}_k$ *is length-minimal* if no strictly shorter word v may be obtained from w by operations (1), (2), and (3) above.

Remark 5.4. This notion of equivalent words, only used in this section, is stronger than $\approx$ in Definition 2.65. For example, they disagree on the pair $t_i t_j t_j t_j$ and $t_i t_i t_i t_j$ for $j \neq i - 1, i, i + 1$. By Theorem 2.52, $v \simeq w$ implies $v \approx w$.

Instead of inducting on the length of w, we begin most proofs in this section with replacing w by some length-minimal word that can be obtained from it by operations (1)–(3), and then we reach a contradiction every time we get a chance to cancel $t_i t_i$.

Remark 5.5. Let t, u, v be words in the Ritt monoid. If $w' = tuv$ is length-minimal, then u is length-minimal.

We use the same interval-subscript notation for long sequences of t_i as we did for long sequences of f_i. These intervals may be increasing or decreasing, and open or closed on either end.

Definition 5.6.

- If $a < b$, then $t_{(a,b]} = t_{[a+1,b]} = t_{(a,b+1)} = t_{[a+1,b+1)} := t_{a+1}t_{a+2}\cdots t_b$. A word of this form is a *left-to-right transit*.
- If $a > b$, then $t_{(a,b]} = t_{[a-1,b]} = t_{(a,b-1)} = t_{[a-1,b-1)} := t_{a-1}t_{a-2}\cdots t_b$. A word of this form is a *right-to-left transit*.
- If $a = b$, then $t_{[a,b]} := t_a = t_b$, while $t_{(a,b]} = t_{[a,b)} = t_{(a,b)}$ is the empty word.

Transits are so named because, for example, when $a > b$, in $t_{(a,b]} \star \vec{f} = \vec{g}$ one factor f_b "travels" left from its original b^{th} position to become (in the sense of Remark 2.46) the factor g_a in a^{th} position in $\vec{g}$.

The following generalizations of operation (3) are useful. The last one says that if two adjacent factors f_a and f_{a-1} travel some number of steps to the left and then switch places, they could just as well have switched places first and traveled later.

LEMMA 5.7.

- If $r + 1 > r \geq s$, then $t_r t_{[r+1,s]} \simeq t_{[r+1,s]}t_{r+1}$.
- If $p > r \geq s$, then $t_r t_{[p,s]} \simeq t_{[p,s]}t_{r+1}$.
- If $b \geq a$, then $t_{[b,a-1]}t_{[b,a]} \simeq t_{[b-1,a-1]}t_{[b,a]}t_{a-1}$.

Proof. For (1), $t_r t_{r+1} t_r \simeq t_{r+1} t_r t_{r+1}$, 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. We prove (3) by induction on $b - a$. The base case $b = a$ is exactly operation (3) in Remark 5.1. For the induction step, $b \gneq a$, so $b - 2 \geq a - 1$ and $b - 1 \geq a$ in the following computations:

$$t_{[b,a-1]}t_{[b,a]} = t_b t_{b-1} t_{[b-2,a-1]} t_b t_{[b-1,a]} \simeq t_b t_{b-1} t_b t_{[b-2,a-1]} t_{[b-1,a]}$$
$$\simeq t_{b-1} t_b t_{b-1} t_{[b-2,a-1]} t_{[b-1,a]} = t_{b-1} t_b t_{[b-1,a-1]} t_{[b-1,a]} =: u.$$

Applying the inductive hypothesis to $t_{[b-1,a-1]}t_{[b-1,a]}$, we get

$$u \simeq t_{b-1} t_b t_{[b-2,a-1]} t_{[b-1,a]} t_{a-1} \simeq t_{b-1} t_{[b-2,a-1]} t_b t_{[b-1,a]} t_{a-1}. \qquad \square$$

A sequence of Ritt swaps in the first canonical form is a sequence of right-to-left transits whose action resembles an insert-sort: having arranged f_k through f_{i+1} in the right order, this sequence *inserts* f_i in the required a_i^{th} place among f_k through f_{i+1} and then proceeds to deal with f_{i-1}, and so on, until all factors are arranged as wanted.

Definition 5.8. A word $w \in \mathrm{RM}_k$ is in *first canonical form* if it has the form $w = t_{(a_1,1]}t_{(a_2,2]} \cdots t_{(a_{k-1},k-1]}$ for some $a_1, a_2, \ldots, a_{k-1}$ such that $i \leq a_i \leq k$ for each i.

A word $w \in \mathrm{RM}_k$ is in *reverse first canonical form* if it has the form $w = t_{[a_k,k)}t_{[a_3,3)} \cdots t_{[a_2,2)}$ for some $a_1, a_2, \ldots, a_{k-1}$ such that $1 \leq a_i \leq i$ for each i.

For example, if $k = 4$, $a_2 = 2$, $a_3 = 1$, and $a_4 = 3$, then $w = (t_3)(t_1t_2)() = t_3t_1t_2$.

Remark 5.9. Omitting those transits that are empty words and reindexing gives an alternative formulation of first canonical form as $t_{[a_m,b_m]}t_{[a_{m-1},b_{m-1}]} \cdots t_{[a_1,b_1]}$ with $a_i \geq b_i$ for all i, and $b_m < b_{m-1} < \cdots < b_1$.

The three syntactic operations in Remark 5.1 are invariant under switching left and right, so anything that holds for the first canonical form also holds, mutatis mutandis, for the reverse first canonical form.

We now work towards taking an arbitrary word in RM_k and finding an equivalent word in first canonical form. The following lemma straightens out two right-to-left transits that occur in the wrong order.

LEMMA 5.10. *If $a \geq b$ and $c \geq d$ and $w' := t_{[a,b]}t_{[c,d]}$ is length-minimal, then it is equivalent to a single right-to-left transit, or to $t_{[a',b']}t_{[c',d']}$ for some a', b', c', d' such that $a' \geq b'$ and $c' \geq d'$ and $b' < d'$.*

Proof. 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 $\hat{w} = t_{[c,d]}t_{[a,b]}$ works, because in this case each t_i in the first transit of w' commutes with each t_j in the second transit, and $b > c + 1 > d$.
- If $b = c + 1$, then w' is already a single transit as wanted.
- If $b = c$, then operation (1) shortens the word w' contradicting length-minimality.
- This leaves the case where $c > b \geq d$ for which we use Lemma 5.7 and another case-out.

So $c > b \geq d$; compare a and c:

- If $a < c$, then Lemma 5.7 can be applied to each t_i for $a \geq i \geq b$ giving $w' \simeq t_{[c,d]}t_{[a+1,b+1]} =: \hat{w}$, of the desired form because $d \leq b$ implies $d < b+1$.
- If $a \geq c$, Lemma 5.7 can still be applied to each t_i for $c - 1 \geq i \geq b$ giving

$$w' = t_{[a,b]}t_{[c,d]}$$
$$= t_{[a,c]}t_{[c-1,b]}t_{[c,d]}$$
$$\simeq t_{[a,c]}t_{[c,d]}t_{[c,b-1]},$$

contradicting length-minimality, as t_ct_c sits in the middle of $t_{[a,c]}t_{[c,d]}$. $\square$

PROPOSITION 5.11. *For every $w \in \mathrm{RM}_k$, there exists a unique $\hat{w} \in \mathrm{RM}_k$ in first canonical form that represents the same permutation as w. This $\hat{w}$ can be 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. First, replace w by some length-minimal w' obtained from w by operations (1), (2), and (3) in Remark 5.1. Any word including w' can be written as a sequence of right-to-left transits $t_{[a_m,b_m]} t_{[a_{m-1},b_{m-1}]} \cdots t_{[a_1,b_1]}$ with $a_i \geq b_i$ for all i. To achieve the additional requirement that $b_m < b_{m-1} < \cdots < b_1$ in Remark 5.9, use Lemma 5.10 repeatedly to straighten out pairs of adjacent out-of-order b_i's. Clearly, this process terminates. □

COROLLARY 5.12. *If two words w and w' in the Ritt monoid RM_k 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. □

Another immediate consequence is a bound on the length of words and the number of (linear equivalence classes of) decompositions of a given polynomial; better bounds are obtained in [21].

COROLLARY 5.13. *For any polynomial f and decomposition $(f_k, \ldots, f_1)$ of f, there are at most $k!$ other decompositions $\vec{g}$ of f (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.*

Our main use of the first canonical form is to define and obtain our second canonical form.

It is sometimes natural and often useful to break a decomposition into chunks before analyzing it. For example, in analyzing the commutative diagram $\pi^\sigma \circ f = g \circ \pi$, it is natural to start with a decomposition of $g \circ \pi$ that is a decomposition of π followed by a decomposition of g. Clusterings in Section 4 are another example. Words in second canonical form (with respect to such a break-up into chunks) first shuffle factors within each chunk as much as necessary and only then move factors between chunks. That is, these words perform a merge-sort.

Definition 5.14. Given integers $k = c_r > c_{r-1} > \cdots > c_1 > c_0 = 0$, a word $w \in \mathrm{RM}_k$ is in *second canonical form with respect to $\vec{c}$* if it is of the form $w = v w_r w_{r-1} \cdots w_1$ and all of the following hold.

- All w_i and v are in first canonical form.
- For each j, only t_i with $c_{j+1} > i > c_j$ appear in w_j; that is, w_j only permutes factors in $\vec{f}_{[c_{j+1}, c_j)}$.

- For any two transits $t_{[a,b]}$ and $t_{[a',b']}$ in v (including empty transits in Definition 5.8!) with both $b, b' \in [c_{j+1}, c_j)$ for some j, $a < a'$ if and only if $b < b'$. That is, v does not change the order of two factors originating inside the same $\vec{f}_{[c_{j+1}, c_j)}$.

Note that w_i in the definition above act on disjoint sets of factors, and therefore commute with each other.

PROPOSITION 5.15. *For every word $w \in \mathrm{RM}_k$ and every tuple $\vec{c}$ of integers with $k = c_r > c_{r-1} > \cdots > c_1 > c_0 = 0$, there is a word $\hat{w} \in \mathrm{RM}_k$ in second canonical form with respect to $\vec{c}$ that represents the same permutation as w. This $\hat{w}$ can be obtained from w by operations (1), (2), and (3) in Remark 5.1, so for any decomposition $\vec{f}$ such that $w \star \vec{f}$ is defined, $\hat{w} \star \vec{f} = w \star \vec{f}$.*

We first prove a special case $r = 2$ of this proposition in the following lemma. We then prove the full proposition.

LEMMA 5.16 (Proposition 5.15 for $r = 2$). *Fix $k > e > 0$. For every $w \in \mathrm{RM}_k$, there are $v, w_G, w_H \in \mathrm{RM}_k$ such that*

- *$vw_G w_H$ is obtained from w by operations (1), (2), and (3) in Remark 5.1;*
- *only t_i with $i > e$ occur in w_H;*
- *only t_i with $i < e$ occur in w_G;*
- *$v = t_{[a_m, b_m]} t_{[a_{m-1}, b_{m-1}]} \cdots t_{[a_1, b_1]}$ is in first canonical form, and*
  $$b_1 = e, b_2 = e - 1, \ldots, b_m = e - m + 1 \text{ and } a_1 > a_2 > \cdots > a_m.$$

Proof. First reductions: Without loss of generality, we may assume that w is already in first canonical form; that is,
$$w = w_n w_{n-1} \cdots w_1 \text{ where each } w_i = t_{[c_i, d_i]}$$
with $d_1 > d_2 > \cdots > d_n$ and $c_i \geq d_i$ for each i.

We may further assume that $e \leq d_1$. Indeed, otherwise let $j \leq n$ be the greatest such that $d_j > e$, and let
$$\widetilde{w_H} := w_j w_{j-1} \cdots w_1 \text{ and } w' := w_n \cdots w_{j+1} \text{ so that } w = w' \widetilde{w_H}.$$
Since $\widetilde{w_H}$ only involves t_i with $i > e$, it clearly suffices to prove the proposition for w'.

So, we have
$$w = w_n w_{n-1} \cdots w_1 \text{ where each } w_i = t_{[c_i, d_i]}$$
in first canonical form with
$$e \geq d_1 > d_2 > \cdots > d_n \text{ and } c_i \geq d_i \text{ for each } i.$$
The first inequality above makes $d_n \leq e - n + 1$.

CLAIM. *There are v and w_G satisfying the last two requirements in the lemma with $m \leq n$, such that $vw_G \simeq w$.*

We obtain these v and w_G by induction on n.

Base case. When $n = 1$, consider the one and only transit $w_n = t_{[c_n,d_n]}$ of w. We know that $d_n \leq e$.

Case 1. If $c_n < e$, then $w_n \in \mathrm{RM}_e$, so $v := \emptyset$ and $w_G := w_n$ work.

Case 2. If $c_n \geq e$, then $w_n = t_{[c_n,e]}t_{[e-1,d_n]}$, and $v := t_{[c_n,e]}$ and $w_G := t_{[e-1,d_n]}$ work.

Induction step. We apply the induction hypothesis to the initial $(n-1)$ chunks of w to get
$$w_{n-1}\cdots w_1 \simeq \widetilde{v}\widetilde{w_G}$$
for some
$$\widetilde{v} = t_{[a_m,e-m+1]}t_{[a_{m-1},e-m+2]}\cdots t_{[a_1,e]}$$
with $m \leq n-1$ and $a_1 > a_2 \cdots > a_m$, and some $\widetilde{w_G} \in \mathrm{RM}_e$ (so $\widetilde{w_G}$ only involves t_i with $i < e$). So
$$w = w_n w_{n-1}\cdots w_1 \simeq w_n\widetilde{v}\widetilde{w_G}.$$
Let $w' := w_n\widetilde{v}$, so that $w = w'\widetilde{w_G}$.

SUBCLAIM. *There are words v and u_G such that $w' \simeq vu_G$, and $u_G \in \mathrm{RM}_e$ (so u_G only involves t_i with $i < e$), and v has the required shape.*

Once we prove this subclaim, v and $w_G := u_G\widetilde{w_G}$ will satisfy the claim.

Proof of the subclaim. We have
$$w' = t_{[c_n,d_n]}t_{[a_m,e-m+1]}t_{[a_{m-1},e-m+2]}\cdots t_{[a_1,e]}$$
and $m \leq n-1$ and $a_1 > a_2 \cdots > a_m$ and $d_n \leq e-n+1$. So $d_n \leq e-m$, and so
$$t_{[c_n,d_n]} = t_{[c_n,e-m]}t_{[e-m-1,d_n]}$$
and, since the least index appearing in $\widetilde{v}$ is $e-m+1$ and the greatest in $t_{[e-m-1,d_n]}$ is $e-m-1$,
$$w' = t_{[c_n,e-m]}t_{[e-m-1,d_n]}\widetilde{v} \simeq t_{[c_n,e-m]}\widetilde{v}t_{[e-m-1,d_n]}.$$
Since $t_{[e-m-1,d_n]}$ can be absorbed into u_G, it suffices to prove the subclaim for the special case where $d_n = e-m$. If now $c_n < a_m$, we may set $a_{m+1} := c_n$ and set $v := t_{[c_n,e-m]}\widetilde{v}$ and be done.

So, it suffices to prove the subclaim for
$$w' = t_{[c_n,e-m]}t_{[a_m,e-m+1]}t_{[a_{m-1},e-m+2]}\cdots t_{[a_1,e]},$$
where $a_1 > a_2 \cdots > a_m$ but $c_n \geq a_m$. So
$$w' = t_{[c_n,a_m+1]}t_{[a_m,e-m]}t_{[a_m,e-m+1]}t_{[a_{m-1},e-m+2]}\cdots t_{[a_1,e]}.$$
By [Lemma 5.7](), $t_{[a_m,e-m]}t_{[a_m,e-m+1]} \simeq t_{[a_m-1,e-m]}t_{[a_m,e-m+1]}t_{e-m}$, so
$$w' \simeq t_{[c_n,a_m+1]}t_{[a_m-1,e-m]}t_{[a_m,e-m+1]}\mathbf{t_{e-m}}t_{[a_{m-1},e-m+2]}\cdots t_{[a_1,e]}$$
$$\simeq t_{[c_n,a_m+1]}t_{[a_m-1,e-m]}t_{[a_m,e-m+1]}t_{[a_{m-1},e-m+2]}\cdots t_{[a_1,e]}\mathbf{t_{e-m}}$$
$$\simeq t_{[a_m-1,e-m]}\mathbf{t_{[c_n,e_m+1]}}t_{[a_m,e-m+1]}t_{[a_{m-1},e-m+2]}\cdots t_{[a_1,e]}t_{e-m}.$$

Now if $c_n < a_{m-1}$, we are done, because t_{m-e} may be absorbed into u_G, and the rest is already of the right form. Otherwise, we move $t_{[c_n, a_{m-1}+1]}$ one more step to the right in exactly the same way and then compare c_n to a_{m-2}. Since there is no requirement on a_1, this process ends in success after at most m steps. □

The full version of Proposition 5.15 now follows by an easy induction.

Proof. (This is the proof of Proposition 5.15.) We induct on the number of chunks r. For $r = 1$, this is just first canonical form. The case $r = 2$ is Lemma 5.16. For the induction step, we suppose that the proposition holds for $r = s$ and prove that it holds for $r = s + 1$. Fix w and $k = c_{s+1} > c_s > \cdots > c_1 > c_0 = 0$.

First, apply Lemma 5.16 to $\vec{d}$ where $k = d_2 > c_1 = d_1 > d_0 = 0$. That is, replace w by $v_Q \circ u_1 \circ u_2$, where u_1 only involves t_i with $i < c_1$, u_2 only involves t_i with $i > c_1$, and v_Q does what it is supposed to.

Then, apply the inductive hypothesis to u_2 and $k = c_{s+1} > c_s > \cdots > c_2 > c_0 = 0$ to get $u_2 \simeq v'w_2'w_3' \cdots w_{s+1}'$. So $w \simeq v_Q u_1 v' w_2' w_3' \cdots w_{s+1}' \simeq v_Q v' u_1 w_2' w_3' \cdots w_{s+1}'$, the second equivalence because u_1 and v' act on disjoint sets of factors. Finally, let $v := v_Q v'$, $w_1 := u_1$, and $w_i := w_i'$ for $i \geq 2$. □

When the second canonical form is applied to a clustering, the characterization of v can be substantially strengthened.

Remark 5.17. Suppose that $\vec{a}$ is a clustering of a decomposition $\vec{f}$, and fix $w \in \mathrm{RM}_k$ for which $w \star \vec{f}$ is defined. If $w = vw_1w_2 \cdots w_t$ is in second canonical form with respect to $\vec{a}$, then each w_j only permutes factors within the jth cluster, and v only moves factors left from cluster to cluster but does not change the order of those originating in the same cluster. Since only quadratics can move between clusters, the Ritt swaps in v can only move quadratics. These quadratics can go left or right, but cannot overtake each other because of gates, and cannot collide with each other because tautological Ritt swaps are not allowed.

By Proposition 4.11, all Ritt swaps in w_j may be witnessed by identity linear factors. Now $\vec{a}$ is still a clustering of $\vec{g} := w_1 w_2 \cdots w_t \star \vec{f}$. To illustrate what can happen in $v \star \vec{g}$, we describe in detail an initial chunk of v.

The rightmost symbol in v is t_{a_j} for some j, since v does not swap factors inside a cluster. By Lemma 4.34, exactly one of g_{a_j} and g_{a_j+1} must be a wandering quadratic of $\vec{a}$ and $\vec{g}$.

- If g_{a_j} is a wandering quadratic, v moves it to the left, leaves a one-way left-to-right gate at j, and then only permutes factors further right. That is, $v = v' t_{[b,a_j]}$ for some $b \geq a_j$ and $t_{[b,a_j]} \star \vec{g}$ has a one-way left-to-right gate between the jth and the $(j+1)^{\text{st}}$ clusters, and v' is a word in $\{t_i \mid j < a_j - 1\}$.

- If g_{a_j+1} is a wandering quadratic, v moves it right to a new position, and the rest of v cannot move other quadratics left past that new position. That is, $v = v't_bt_{b+1}\cdots t_{a_j}$ and v' is a word in $\{t_i \mid i < b-1\}$.

Applying the same analysis to the remaining v' gives an inductive characterization of v as a concatenation of (left-to-right or right-to-left) transits, each of which moves a wandering quadratic of $\vec{g}$ some number of steps (right or left).

Our first application of canonical forms is Proposition 5.23 characterizing those (f,g)-skew-invariant curves that have nothing to do with skew-twists. More precisely, we consider 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. We continue to only consider disintegrated polynomials f and g, that is f and g that are not skew-conjugate to monomials, Chebyshev polynomials, and negative Chebyshev polynomials. We begin by defining the terminology in the conclusion of Proposition 5.23.

Definition 5.18. A decomposition $\vec{f}$ has *in-degree* (*respectively, out-degree*) *divisible by* n if no f_i is linearly related to P_p for any p that divides n, and every nonmonomial factor of the decomposition is monic and has in-degree (respectively, out-degree) divisible by n. In particular, this forces all f_i to be Ritt polynomials.

We first show (Lemma 5.21) that a skew-conjugacy class of decompositions of disintegrated polynomials has at most one decomposition with nontrivial in- or out-degree, up to skew-conjugating by scalings.

Remark 5.19. In- and out-degrees make sense for decomposable polynomials (see Definition 3.17), and if $\vec{f}$ has in-degree divisible by p, then indeed the in-degree of $\vec{f}^\circ$ is divisible by p. A converse requires additional hypotheses and is proved in [21].

Remark 5.20. It is clear that if $\vec{h}$ is skew-linearly equivalent to $\vec{f}$ via scalings, then $\vec{f}$ and $\vec{h}$ have the same in- and out-degrees.

LEMMA 5.21. *If $\vec{f}$ is a decomposition of a disintegrated polynomial, $\vec{g}$ is skew-linearly equivalent to $\vec{f}$ via translations, and each of the two has nontrivial in- or out-degree, then $\vec{g} = \vec{f}$.*

Proof. We actually prove the following generalization. Suppose that $\vec{f}$ is a decomposition of a disintegrated polynomial and that $\vec{g}$ is linearly equivalent via translations to $M_L \circ \vec{f} \circ M_R$ for some translations M_L and M_R. If both $\vec{f}$ and $\vec{g}$ have nontrivial in- or out-degree, then $M_L = \mathrm{id}$.

Let $\vec{a}$ be a clustering of $\vec{f}$; it is then also a clustering of $\vec{g}$. Consider the leftmost clusters of the two decompositions. For some translation U, $\vec{g}_{[k,a_{r-1})}$ is

linearly equivalent to $(M_L \circ f_k, f_{k-1}, \ldots, f_{a_{r-1}+2}, f_{a_{r-1}+1} \circ U)$. If this is a C-free cluster, then by Lemma 4.7, $M_L = \mathrm{id}$. So suppose that this is a C cluster.

Recall that the only Ritt polynomials linearly related to Chebyshev polynomials are $C_{p,\lambda}$ and $\widehat{C}_{p,\lambda}$, whose in- and out-degrees are 2. Thus, in order for the whole decomposition to have nontrivial in- or out-degree, there must be no quadratic factors, so all factors inside C clusters have odd degree. Furthermore,

$$\vec{f}^{\circ}_{[k,a_{r-1})} = (\cdot\lambda_f^{-N}) \circ T_f^{-1} \circ C_N \circ T_f \circ (\cdot\lambda_f)$$

and

$$\vec{g}^{\circ}_{[k,a_{r-1})} = (\cdot\lambda_g^{-N}) \circ T_g^{-1} \circ C_N \circ T_g \circ (\cdot\lambda_g)$$

for some nonzero scalars λ_f and λ_g and translations T_f and T_g by 0 or by -2. So,

$$M_L \circ (\cdot\lambda_f^{-N}) \circ T_f^{-1} \circ C_N \circ T_f \circ (\cdot\lambda_f) \circ U = (\cdot\lambda_g^{-N}) \circ T_g^{-1} \circ C_N \circ T_g \circ (\cdot\lambda_g).$$

Now by Lemma 3.36, each of $T_g \circ (\cdot\lambda_g^N) \circ M_L \circ (\cdot\lambda_f^{-N}) \circ T_f^{-1}$ and $T_f \circ (\cdot\lambda_f) \circ U \circ (\cdot\lambda_g)^{-1} \circ T_g^{-1}$ must be scalings by ± 1. So $\lambda_f = \pm\lambda_g$, and $T_g \circ \tilde{M}_L \circ T_f^{-1} = T_f \circ \tilde{U} \circ T_g^{-1} = \mathrm{id}$ where $\tilde{U} = U^{\pm 1}$ and $\tilde{M}_L = M_L^{\pm 1}$. Thus, $\tilde{M}_L = T_g^{-1} \circ T_f$ and $\tilde{U} = T_f^{-1} \circ T_g = (\tilde{M}_L)^{-1}$. Therefore, $U = \mathrm{id}$ if and only if $M_L = \mathrm{id}$.

Since $\vec{f}$ is a decomposition of a disintegrated polynomial, and each f_i is a Ritt polynomial, it cannot consist of a single C cluster. So, consider now the second leftmost cluster $\vec{f}_{[a_{r-1},a_{r-2})}$. If it is a C-free cluster, U must be identity as before. If it also a C cluster, a straightforward computation, which we leave to the reader, shows that either U is identity, or the concatenation $\vec{f}_{[k,a_{r-2})}$ is itself a cluster, contradicting the fact that $\vec{a}$ is a clustering. $\qquad\square$

PROPOSITION 5.22. *Suppose that $\vec{f}$ is a decomposition of a disintegrated polynomial, $\vec{g}$ is skew-linearly equivalent to $\vec{f}$, and each has nontrivial in- or out-degree. Then there is some λ such that $g_i = \lambda^{m_i} * f_i$, where $m_i := \deg(\vec{f}^{\circ}_{(i,1]})$. In particular, their in- and out-degrees are the same.*

Proof. As usual, we can deal with scalings and translations separately. Separating scalings from translations in the linear factors witnessing skew-linear equivalence, we find an intermediate decomposition $\vec{h}$ that is skew-linearly equivalent to $\vec{f}$ via scalings, and skew-linearly equivalent to $\vec{g}$ via translations. By Remark 5.20, $\vec{h}$ has the same in- and out-degrees as $\vec{f}$. By Lemma 5.21, $\vec{h} = \vec{g}$. $\qquad\square$

The last proposition essentially says that for decompositions of disintegrated polynomials, in- and out-degrees are invariant under skew-linear equivalence.

PROPOSITION 5.23. *If two disintegrated 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 whose degree divides the in-degree of some decomposition of $(M^\sigma)^{-1} \circ g \circ M$ and the out-degree of some decomposition of $L^\sigma \circ f \circ L^{-1}$.

The following slight weakening of this proposition, which does not refer to decompositions, is an immediate consequence of Lemma 2.8 of [21], and the full version follows from other results in that paper.

COROLLARY 5.24. *If two nonlinear 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, Chebyshev polynomials, or negative 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.*

Our slightly stronger statement is the one we use to characterize skew-invariant curves. The rest of this section constitutes the proof of Proposition 5.23. The next proposition translates it into the language of decompositions and canonical forms.

PROPOSITION 5.25 (Translating Proposition 5.23). *Suppose that polynomials f, g, and π satisfy $g \circ \pi = \pi^\sigma \circ f$, and suppose 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*

$$(t_{[\ell,1]} \cdots t_{[\ell+m-2,m-1]} t_{[l+m-1,m]}) \star \vec{g}\vec{\pi} = \vec{\rho}\vec{f}.$$

Proof. Let $(\pi_m, \ldots, \pi_1)$ be a decomposition of π and $(g_l, \ldots, g_1)$ 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

$$v = t_{[a_k,b_k]} t_{[a_{k-1},b_{k-1}]} \cdots t_{[a_1,b_1]}$$

with $a_i \geq b_i - 1$ for all i ($a_i = b_i - 1$ means that the word $(t_{a_i}, \cdots t_{b_i})$ is empty); $b_i = \text{length}(\vec{\pi}) + 1 - i$, and $a_k < \cdots < a_2 < a_1$; 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 $v = t_{[\ell,t_1]} \cdots t_{[\ell+m-2,m-1]} t_{[\ell+m-1,m]}$ as wanted. $\qquad\square$

We have $v := v_1 v_2 \cdots v_m$ where $v_i := t_{[\ell+1-i,i]}$, and we have $v \star \vec{g}\vec{\pi}$ defined. The next lemma shows that $\vec{g}\vec{\pi}$ is a cluster or one of π and g has degree 2. The following two lemmas handle these two cases.

LEMMA 5.26. *Suppose that the conclusion of* Proposition 5.25 *holds, but* $\vec{g}\vec{\pi}$ *is not a cluster. Then one of* π *and* g *is quadratic.*

Proof. Suppose that $\vec{g}\vec{\pi}$ is not a cluster, and fix a clustering $\vec{a}$ of it.

If the leftmost cluster $(g_k, \ldots, g_e)$ of $\vec{a}$ does not contain all of $\vec{g}$, then $(g_k, \ldots, g_1, \pi_m)$ is not a cluster, so Proposition 4.40 applied to $v_m \star \vec{g}\vec{\pi} =:$ $(\pi'_m, \vec{g}', \pi_{m-1}, \pi_{m-2}, \ldots, \pi_1)$ makes π_m quadratic and leaves one-way left-to-right gate between g'_e and g'_{e-1}. This prevents π_{m-1} from crossing into the leftmost cluster, contradicting the conclusion of Proposition 5.25, so there must be no π_{m-1}, so the whole of $\pi = \pi_m$ is quadratic.

Symmetrically, if the rightmost cluster of $\vec{a}$ does not contain all of π, then g must be quadratic.

So suppose that the two clusters of $\vec{a}$ are exactly $\vec{g}$ and $\vec{\pi}$. Since (g_1, π_m) is swappable, one of them must be a wandering quadratic of this clustering. If g_1 is the wandering quadratic, then π_m cannot move further left, so there must be no more left for it to go, i.e., $g = g_1$ is quadratic. If π_m is the wandering quadratic, then it leaves a one-way left-to-right gate between the two clusters, so π_{m-1} cannot enter the left cluster, so there must be no π_{m-1}, so the whole of $\pi = \pi_m$ is quadratic. $\qquad\square$

LEMMA 5.27. Proposition 5.23 *holds when one of* π *and* g *has degree* 2.

Proof. If π is quadratic, then skew-conjugate f and g to make $\pi = P_2$. Let $\vec{a}$ be a clustering of $\vec{g}$. Applying Proposition 4.40 to $v_m \star (\vec{g}P_2)$ shows that there must be a right-to-left gate between any two clusters of $\vec{a}$. In order for $\pi = P_2$ to enter, $\vec{a}$ must also have a right-to-left gate at 0. At the far left end, $\vec{a}$ must also have a right-to-left gate at k, since after all the Ritt swaps are performed, the now leftmost quadratic factor must exit as $\pi^\sigma = P_2$ with no additional linear factors. It is now routine to verify that having all these gates is sufficient for the conclusion of Proposition 5.23.

If g is quadratic, the in/out degrees of the factors of π that are not type C are irreversably changed by the traversing quadratic factor, so C-free clusters in π must be purely monomial. Similarly, if π is not a single cluster, then the gates in π (before the quadratic gets across to become f) and in π^σ (after the quadratic has gotten across) do not match up. Thus we can skew-conjugate f and g to make π a monomial or a Chebyshev polynomial. It is now routine

to verify that this forces f and g to be monomials or Chebyshev polynomials, contradicting disintegratedness. □

LEMMA 5.28. *Proposition* 5.23 *holds when $\vec{g}\vec{\pi}$ is a cluster.*

Proof. We may and do assume that neither g nor π are quadratic, since Lemma 5.27 takes care of those cases. The statement of Proposition 5.23 is invariant under skew-conjugating f and g (and changing π accordingly), so we may do so during the proof without loss of generality.

First reductions. Since $\vec{g}\vec{\pi}$ is a single cluster, there are linear A and B and Ritt polynomials $\tilde{g}_i$ and $\tilde{\pi}_i$ such that $\vec{g}\vec{\pi}$ is linearly equivalent to $(A \circ \tilde{g}_\ell, \tilde{g}_{\ell-1}, \ldots, \tilde{g}_1, \tilde{\pi}_m, \ldots, \tilde{\pi}_2, \tilde{\pi}_1 \circ B)$. Skew-conjugating f by B, we may assume that $B = \mathrm{id}$.

Now there is a linear C such that $(C^{-1} \circ \tilde{\pi}_m, \ldots, \tilde{\pi}_2, \tilde{\pi}_1)$ is linearly equivalent to $\vec{\pi}$ and $(A \circ \tilde{g}_\ell, \tilde{g}_{\ell-1}, \ldots, \tilde{g}_1 \circ C)$ is linearly equivalent to $\vec{g}$. Skew-conjugating g by C, that is replacing π by $C \circ \pi$ and replacing g by $C^\sigma \circ g \circ C^{-1}$, and replacing A by $C^\sigma \circ A$, we may assume that $C = \mathrm{id}$.

Thus, replacing $\vec{\pi}$ and $\vec{g}$ by linearly equivalent decompositions, we may assume that all π_i are Ritt polynomials $\tilde{\pi}$, that g_i for $i \neq \ell$ are Ritt polynomials $\tilde{g}_i$, and that $g_\ell = A \circ \tilde{g}_\ell$ for the linear A and Ritt polynomial $\tilde{g}_\ell$.

Case 1. If this is a C cluster, then all $\tilde{g}_i$ and π_i are Chebyshev polynomials, so $\pi = \pi^\sigma$ is a Chebyshev polynomial and $g = A \circ C_{\deg(g)}$. Now $v \star \vec{g}\vec{\pi} = (A \circ \pi_m, \pi_{m-1}, \ldots, \pi_1, g_l, \ldots, g_2, g_1)$ is linearly equivalent to $\vec{\rho}\vec{f}$ for some decompositions $\vec{f}$ of f and some decomposition ρ of the Chebyshev polynomial π^σ of degree greater than 2. In particular, there is a linear D such that $A \circ C_{\deg(\pi)} \circ D = C_{\deg(\pi)}$, so by Lemma 3.36, $A = (\cdot \pm 1)$, contradicting the hypothesis that g is disintegrated.

Case 2. If $\vec{g}\vec{\pi}$ is a C-free cluster, at least one of g and π must be linearly related to a monomial.

Suppose towards contradiction that π is not monomial, so at least one factor π_i is not monomial, nor type C. In $v \star \vec{g}\vec{\pi} = \vec{\rho}\vec{f}$, each such factor $\pi_i(x) = x^{k_i} \cdot u(x^{\ell_i})^{n_i}$ becomes, in the sense of Remark 2.46, $\tilde{\rho}_i$ with new in-degree $\ell_i \cdot \deg(f)$ and new out-degree $\frac{n_i}{\deg(f)}$. For some linear D, the decomposition $\vec{\rho}$ of π^σ is linearly equivalent to $(\tilde{\rho}_m, \ldots, \tilde{\rho}_2, \tilde{\rho}_1 \circ D)$. (Recall that the action $\star$ is only defined up to linear equivalence.) Since $\sigma(\vec{\pi})$ is another decompositions of π^σ, it must be possible to obtain $\sigma(\vec{\pi})$ from $(\tilde{\rho}_m, \ldots, \tilde{\rho}_2, \tilde{\rho}_1 \circ D)$ by a sequence of Ritt swaps. Recall that in order for $v \star \vec{g}\vec{\pi}$ to be defined, the degrees of the monomial factors of π must be relatively prime to $\deg(f)$, so Ritt swaps within $\vec{\rho}$ cannot undo the changes to in- and out-degrees. None of ρ_i are type C, and all have nontrivial in-degree, so inserting linear factors also cannot undo those changes. This is a contradiction.

Thus, all π_i and the whole π are monomials, so

$$v \star \vec{g}\vec{\pi} = (A \circ \pi_m, \pi_{m-1}, \ldots, \pi_1, f_l, \ldots, f_2, f_1).$$

Thus, there is a linear D such that $A \circ P_{\deg(\pi)} \circ D = P_{\deg(\pi)}$, so A must be a scaling. As we are working over a difference-closed field, we can get rid of A by skew-conjugating g by an appropriate scaling. All Ritt swaps within a cluster can be witnessed by identity linear factors, so f_i have the requisite in- and out-degrees. $\square$

Having finished the proof of Proposition 5.23, we note two consequences of it.

COROLLARY 5.29. *If f, g, and π are as in Proposition 5.23, then there are indecomposable π_i for $i \leq m$, and polynomials $f = f_0, f_1, \ldots, f_m = g$ such that $\vec{\pi}^\circ = \pi$, and for each i, f_{i-1}, π_i, and f_i are also as in Proposition 5.23.*

COROLLARY 5.30. *If f, g, and π are as in Proposition 5.23, then the degree of π is bounded by the degree of any indecomposable factor of f, so a fortiori bounded by the degree of f.*

6. Classification of skew-invariant curves from clusterings

In this last technical Section 6, we bring together clusterings, the action by the skew-twist monoid ST_k on skew-linear-equivalence classes of decompositions, and the characterization in Proposition 5.23 of indecomposable curves that do not come from skew-twists in order to finally state and prove our classification of irreducible plane curves that are (f, g)-skew-invariant for a given pair (f, g) of disintegrated polynomials.

In order to describe how correspondences not coming from skew-twists interact with skew-twists, we bring them into our monoid-action formalism via new monoid generators. To characterize correspondences coming from skew-twists, we describe the interaction between clusterings and skew-twists. Finally, we put it all together to write out the final characterization, and we then state a few special cases and more readable weakenings.

6.1. *Augmented skew-twist monoid.* By Corollary 5.29, skew-invariant correspondences arising from Proposition 5.23 (rather than from skew-twists) can also be broken down into indecomposable factors that are graphs of monomials of prime degree p. For each prime p, the graph P_p and its converse relation will be encoded by the new generators δ_p and ε_p, respectively. Most of the time, the action of these new generators will be undefined, since Proposition 5.23 forces the decomposition to have a nontrivial in- or out-degree, which is rare. Proposition 5.22 essentially shows that the action is well defined on skew-linear equivalence classes.

We now do for correspondences coming from Proposition 5.23 what we did for skew-twists in Section 2.5: we define a new monoid with more generators, its action on decompositions, witnessing sequences, encoded correspondences, its action on skew-linear equivalence classes of decompositions, and a notion of equivalence for words in the new monoid. As for skew-twists, we then show that the correspondence encoded is essentially well-defined, and prove enough equivalences for words to get enough control on degrees to characterize invariant curves.

Definition 6.1. The free monoid generated by the generators of ST_k together with countably many new symbols ε_p and δ_p as p ranges through the primes is denoted by ST_k^+.

If $\vec{f}$ has nontrivial in-degree divisible by p, then $\varepsilon_p \star \vec{f} := \vec{g}$ where $g_i := f_i$ whenever f_i is a monomial, and for all other i there are monic nonconstant polynomials u_i and integers $k_i \geq 1$ such that $f_i(x) = x^{k_i} \cdot u_i(x^{p\ell_i})^{n_i}$ and $g_i := x^{k_i} \cdot u_i(x^{\ell_i})^{pn_i}$.

To undo what ε_p does, we define $\delta_p \star \vec{g} := \vec{f}$ as above when $\vec{g}$ has nontrivial out-degree divisible by p.

LEMMA 6.2. *Suppose that $\vec{f}$ and $\vec{h}$ are skew-linearly equivalent decompositions of disintegrated polynomials, both have nontrivial in-degree divisible by p, and $\vec{g} = \varepsilon_p \star \vec{f}$ and $\vec{\tilde{g}} = \varepsilon \star \vec{h}$. Then there is some λ such that $h_i = \lambda^{m_i} * f_i$, where $m_i := \deg(\vec{f}^{\circ}_{(i,1]})$, and $\tilde{g}_i = \mu^{m_i} * g_i$ for $\mu = \lambda^p$. The corresponding result holds for δ in place of ε.*

Proof. The first conclusion is Proposition 5.22, and the second follows immediately. □

Definition 6.3. Suppose that $\vec{f}$ is a decomposition of a polynomial f and $w := w_n \cdots w_2 w_1 \in \mathrm{ST}_k^+$, where each w_j is a single generator: a Ritt swap t_i, a single skew-twist ϕ or β, or one of the new generators ε_{p_j} or δ_{p_j} for some prime p_j.

A sequence of decompositions $\vec{f} = \vec{f}^0, \vec{f}^1, \dots, \vec{f}^n$ is a *witnessing sequence* for $w \star \vec{f}$ if for each j,

- if w_j is t_i, ϕ, or β, see Definition 2.56;
- if $w_j = \varepsilon_p$, then $\vec{f}^j$ is skew-linearly equivalent to some $\vec{h}$ that has nontrivial in-degree divisible by p_j, and $\vec{f}^{j+1} = \varepsilon_p \star \vec{h}$;
- if $w_j = \delta_p$, switch the roles of $\vec{f}^j$ and $\vec{f}^{j+1}$ above.

If such a sequence exists, we write $w \star [[\vec{f}]] = [[\vec{f}^n]]$; otherwise, $w \star [[\vec{f}]] = \infty$.

The correspondence $\mathcal{A}$ encoded by this witnessing sequence is again the composite of curves $\mathcal{B}_j$; for $1 \leq j \leq n$,

- if w_j is t_i, ϕ, or β, the curve $\mathcal{B}_j$ is exactly as in Definition 2.56;
- if $w_j = \varepsilon_p$, C_j is the graph of $P_p \circ T_j$, where T_j is the outside linear factor witnessing that $\varepsilon_p \star [[\vec{f}^j]]$ is defined;
- if $w_j = \delta_p$, then C_j is the converse of this graph.

We also say that $\mathcal{A}$ is *a correspondence encoded by* $w \star \vec{f}$.

Remark 6.4. Note that it may well be that $\varepsilon \star [[\vec{f}]]$ is defined but $\varepsilon \star \vec{f}$ is not. By Lemma 6.2, the witnessing sequence is well defined up to skew-linear equivalence, so $w \star [[\vec{f}]]$ is well defined and gives an action of ST_k^+ on skew-linear equivalence classes of decompositions.

Further, by Lemma 6.2 and the fact that $\lambda * P_p = P_p$, the curve encoded by $\varepsilon \star [[\vec{f}]] = [[\vec{g}]]$ does not depend on the choice of the witnessing $\vec{h}$, up to a terminal linear factor exactly as Lemma 2.61. The same holds for δ, so Lemma 2.61 holds for words in ST_k^+, along with Remark 2.57, which says that concatenation of words corresponds to composition of encoded correspondences.

We define equivalence for words in ST_k^+ exactly as in Definition 2.65 for words in ST_k.

Definition 6.5. Fix $v, w \in \mathrm{ST}_k^+$ and a decomposition $\vec{f} = (f_k, \ldots, f_1)$. We say that v and w are *equivalent with respect to* $\vec{f}$ and write $v \approx_{\vec{f}} w$ if $v \star [[\vec{f}]] = w \star [[\vec{f}]]$ and there are witnessing sequences $(\vec{g}^j)$ and $(\vec{h}^j)$ for $v \star \vec{f}$ and $w \star \vec{f}$, respectively, so that the final $\vec{g}^n$ and $\vec{h}^n$ are decompositions of the same polynomial g, and $(\mathcal{A}_v)_{\mathrm{inv}} = (\mathcal{A}_w)_{\mathrm{inv}}$ for the curves $\mathcal{A}_v$ and $\mathcal{A}_w$ encoded by v (respectively, w) via $(\vec{g}^j)$ (respectively, $(\vec{h}^j)$).

When $v \approx_{\vec{f}} w$ for all $\vec{f}$, we write $v \approx w$ and say that the two words are *equivalent*.

Remark 6.6. It is clear that this equivalence again respects concatenation: if $u \approx u'$ and $v \approx v'$, then $uv \approx u'v'$.

LEMMA 6.7.

- *For any of the old generators* $x = t_i, \phi, \psi$ *and any prime* p, $\varepsilon_p x \approx x \varepsilon_p$ *and* $\delta_p x \approx x \delta_p$.
- *For any* $p \neq q$ *and any* $x, y \in \{\varepsilon, \delta\}$, $x_p y_q \approx y_q x_p$.
- $\varepsilon_p \delta_p \approx \mathrm{id}$.

Proof. It is clear that single skew-twists do not change the in- and out-degrees of a decomposition. To see that Ritt swaps also do not change them, recall that a decomposition with nontrivial in- or out-degree divisible by p cannot have the monomial P_p among its factors; when $p = 2$, this precludes wandering quadratics and makes clusterings completely rigid. This also takes care of the second part. The last part is obvious. $\square$

Remark 6.8. The (f, g)-(skew-)invariant curve encoded by $\delta_p \varepsilon_p \star f = g$, defined by $x^p = y^p$, is the union of p lines whose slopes are p^{th} roots of unity. Because P_p is not a compositional factor of f, components other than the diagonal may be skew-periodic, unlike in the case of skew-twists, so it is not true that $\delta_p \varepsilon_p \approx \text{id}$. However, composing the curve defined by $x^p = y^p$ with the one defined by $y^p = z^p$ does not give anything new, so $\delta_p \varepsilon_p \delta_p \varepsilon_p \approx \delta_p \varepsilon_p$.

The next corollary together with Corollary 5.30 bounds the degrees of the correspondence coming from Theorem 5.23.

COROLLARY 6.9. *For any word w consisting entirely of ε_p and δ_p for various p, there are words u and v such that $w \approx uv$, and u consists entirely of δ_p for various p, and v consists entirely of ε_p for various p. Thus, the degrees of the two monomials encoded by u and v are bounded by $\deg(f)$.*

6.2. *Clusterings and skew-twists.* The interaction between clusterings and skew-twists is the key to finishing the characterization of curves encoded by words in ST_k. Recall (Remark 2.71) that in the context of skew-twists one should imagine the factors of a decomposition standing in a circle, rather than in a line, with only a faint marker between the "first" and "last" factors to remind one to act by σ or σ^{-1} on factors passing the marker. In this vein, recall that the skew-twist monoid ST_k acts on skew-linear equivalence classes of decompositions, that is on decompositions up to linear equivalence and skew-conjugacy.

When a preclustering is skew-twisted, one of the clusters might end up cut in two by that faint marker, some of its factors on one end of the decomposition and some on the others. The following definition addresses this issue.

Definition 6.10. A sequence $k \geq a_r > \cdots > a_1 > 0$ with $r \geq 2$ is a *skew-preclustering* of a decomposition $\vec{f}$ if $f_{[a_i, a_{i-1})}$ is a cluster for each $r \geq i > 1$, and $(f_{a_1}^\sigma, f_{a_1-1}^\sigma, \ldots, f_1^\sigma, f_k, \ldots, f_{a_r+1})$ is also a cluster. When $a_r = k$, that last cluster is simply $(f_{a_1}^\sigma, \ldots f_1^\sigma)$. For example, if $\vec{f}$ is itself a cluster, then the sequence (k) is a skew-preclustering of $\vec{f}$.

A skew-preclustering is a *skew-clustering* if $r = 1$; or if $r \geq 2$ and no cluster, including
$$(f_{a_1}^\sigma, f_{a_1-1}^\sigma, \ldots, f_1^\sigma, f_k, \ldots, f_{a_r+1}),$$
consists of two wandering quadratics, and the concatenation of two adjacent clusters is never a cluster, including the concatenations
$$(f_{a_2}^\sigma, f_{a_2-1}^\sigma, \ldots, f_1^\sigma, f_k, \ldots, f_{a_r+1}) \text{ and } (f_{a_1}^\sigma, f_{a_i-1}^\sigma, \ldots, f_1^\sigma, f_k, \ldots, f_{a_{r-1}+1})$$
that wrap around the end of the polynomial.

The next lemma collects a number of immediate observations that connect skew-clusterings to clusterings and uses the new freedom of skew-conjugacy to improve cleanups.

LEMMA 6.11.

(1) *Suppose that $\vec{a}$ is a skew-clustering of a decomposition $\vec{f}$. Then there is a skew-clustering $\vec{b}$ of the decomposition $\beta^m \star \vec{f}$ such that i is a cluster boundary of $\vec{a}$ if and only if $(i + m) \mod k$, that is, the remainder of $(i + m)$ upon division by k, is a cluster boundary of $\vec{b}$. Similarly, there is a skew-clustering $\vec{c}$ of the decomposition $\phi^m \star \vec{f}$ such that i is a cluster boundary of $\vec{a}$ if and only if $(i - m) \mod k$ is a cluster boundary of $\vec{c}$. The same is true for skew-preclusterings.*

(2) *If $\vec{a}$ is a skew-clustering of a decomposition $\vec{f}$ and $a_r = k$, then $(a_r, \ldots, a_1, 0)$ is a clustering of $\vec{f}$. In this case, we call $(a_r, \ldots, a_1)$ a robust skew-clustering of $\vec{f}$ and $(a_r, \ldots, a_1, 0)$ a robust clustering of $\vec{f}$.*

(3) *For any preclustering $(a_r, \ldots, a_1, a_0)$ of any decomposition, $(a_r, \ldots, a_1)$ is a skew-preclustering of the same decomposition.*

(4) *Any skew-preclustering can be refined to a skew-clustering; in particular, every decomposition with no unswappable factors admits a skew-clustering.*

(5) *Every decomposition with no unswappable factors has a plain skew-twist that has a robust clustering.*

(6) *Every robust clustering admits (up to skew-conjugacy!) a cleanup with $L_k = \mathrm{id}$.*

(7) *If $(\vec{L}, \vec{h})$ is a cleanup of a robust clustering $\vec{a}$ of $\vec{f}$ with $L_k = \mathrm{id}$, then the clustering of $\phi^{a_i} \star \vec{f}$ obtained in part (1) is also robust, and reindexing L_i and h_i and applying σ as necessary produces a cleanup of it.*

Remark 6.12. The notion of robust clustering is necessary in that parts (2) and (7) become much more complicated without this extra hypothesis, because a plain skew-twist might break a C cluster into two pieces, and things become complicated if one of the pieces does not contain an odd-degree factor, and particularly complicated if that piece has degree two.

Proof. For part (4), induct on the number of clusters exactly as in the proof of the existence of clusterings, Lemma 4.30.

Only part (6) merits detailed explanation. Take some cleanup $(\vec{L}, \vec{h})$ of a robust skew-clustering. Skew-conjugating, we may absorb the translation part of L_k into L_0 and assume without loss of generality that L_k is a scaling.

If there is a C cluster, skew-conjugate to move scaling L_k into L_0 and then move it left as in the proof of the existence of cleanups until it sits to the right of a C cluster, where it may stay without violating the definition of "cleanup."

If there are no C clusters, skew-conjugating by $\cdot \frac{1}{\lambda}$ and moving the new scaling left as in the proof of the existence of cleanups replaces L_k by $L_k \circ (\cdot \frac{\lambda^\sigma}{\lambda^{\deg(f)}})$. Here again, because we are working over a difference closed field, there is no problem to find λ such that $L_k \circ (\cdot \frac{\sigma(\lambda)}{\lambda^{\deg(f)}}) = \mathrm{id}$. If one is interested

only in the case of an algebraic dynamical system, then it suffices to find a $(\deg(f) - 1)^{\text{st}}$ root of the leading coefficient of f inside the fixed field of σ. $\quad\square$

For a robust skew-clustering and a cleanup with $L_k = \text{id}$, it is clear what "gate at k" should mean, except maybe when $\vec{f}$ is itself a cluster. Recall that the factors are now standing in a circle, so a gate at k and a gate at 0 are intuitively the same thing.

Definition 6.13. Let $\vec{a}$ be a robust skew-clustering of a decomposition $\vec{f}$, and let $(\vec{L}, \vec{h})$ be a cleanup of it with $L_k = \text{id}$.

If $\vec{a}$ has more than one cluster, then $\vec{f}$ with $\vec{a}$ has *a (left-to-right or right-to-left) gate at k* if and only if the clustering and cleanup of $\phi^{a_i} \star \vec{f} = (f^\sigma_{a_1}, f^\sigma_{a_1-1}, \ldots, f^\sigma_1, f_k, f_{k-1}, \ldots, f_{a_1+1})$ obtained in (7) above has a gate in that direction between f^σ_1 and f_k.

If $\vec{a} = (k)$ has exactly one cluster, then $\vec{f}$ has *a (left-to-right or right-to-left) gate at k* if and only if the preclustering $(2k, k, 0)$ of $\vec{f}^\sigma \vec{f}$ has a gate in that direction at k.

Recall that two adjacent clusters with a two-way gate between them can be fused into a single cluster; this has the following convenient consequence.

LEMMA 6.14. *Suppose that $\vec{f}$ is a decomposition of a disintegrated polynomial f and that a skew-clustering $(a_r, \ldots, a_1)$ of $\vec{f}$ has a two-way gate at some j. Then $\vec{f}$ is a single C-free cluster with at least one nonmonomial factor and admits a cleanup with no linear factors.*

Proof. Lemma 4.16 forbids two-way gates between distinct clusters of a clustering, so $j = r$ and $a_r = k$. If $\vec{f}$ has more than one cluster, this makes $(f^\sigma_{a_1}, f^\sigma_{a_i-1}, \ldots, f^\sigma_1, f_k, \ldots, f_{a_{r-1}+1})$ into a cluster, contradicting the definition of skew-clustering. If $\vec{f}$ is a single cluster, then the two-way gate at k means that (up to skew-conjugacy) $\vec{f}$ admits a cleanup with $L_k = \text{id}$ and $L_0 = (\cdot \pm 1)$. Since f is disintegrated, $\vec{f}$ cannot consist of a single C cluster or a single C-free cluster with only monomial factors. $\quad\square$

The next remark gathers the results we have proved about the interaction of (skew-)clusterings, Ritt swaps, and skew-twists.

Remark 6.15. Since robust skew-clusterings correspond precisely to robust clusterings, different robust clusterings of skew-linearly equivalent decompositions obey Proposition 4.36: they have the same number of clusters, the same cluster boundaries with gates (possibly in different directions), and cluster boundaries may only differ by one, and then only by one wandering quadratic. (This is the case when the gate changes direction.) It is clear that the two robust skew-clusterings have the same gates at k.

Skew-twists act on skew-clusterings (and their gates) via the first part of Lemma 6.11: given a skew-clustering $\vec{a}$ of $\vec{f}$, the corresponding skew-clustering $\vec{b}$ of $\phi^i \star \vec{f}$ has cluster boundaries at $(a_j - i) \mod k$, and similarly for β_i. This $\vec{b}$ is robust if and only if $i = a_j$ for some j. It is clear that $\vec{f}$ has a gate at the cluster boundary at a_j if and only if $\phi^i \star \vec{f}$ has a gate at the corresponding cluster boundary at $(a_j - i) \mod k$.

As noted in Lemmas 4.17 and 4.31, a Ritt swap inside a cluster (that is, t_i for some $i \neq a_j$ for all j) does not affect the clustering or its gates. By skew-twisting until the boundary is inside, it is clear that the new notion of gates at k for a skew-clustering is also unaffected.

Recall (Lemma 4.34) that a Ritt swap across clusters (that is, t_{a_j} for some j) always involves a wandering quadratic that leaves one cluster and enters the other, changing the direction of the gate at the boundary.

It should now be clear that the number of clusters in a robust skew-clustering, and the presence of a gate between particular clusters, are invariant under skew-linear equivalence, Ritt swaps, and plain skew-twists by whole clusters, though the indexing of the clusters changes in this last case.

The conclusion of Lemma 6.16 is used in Lemma 6.17 to bound the number of consecutive ψ's or γ's in a word from the border guard monoid (see Definition 2.69) acting on $\vec{g}$.

LEMMA 6.16. *Every decomposition $\vec{f}$ of a disintegrated polynomial has a plain skew-twist $\vec{g} := \phi^i \star \vec{f}$ that has a robust clustering $\vec{a}$ with one of the following properties*:

(1) *$\vec{a}$ has no gates at k;*

(2) *$\vec{a}$ has a one-way gate at k;*

(3) *$\vec{g}$ is a single C-free cluster, $\vec{a}$ has a two-way gate at k, and g_k is not a monomial.*

Proof. We know that $\vec{f}$ has a skew-clustering $\vec{b}$. If some b_j satisfies one of the first two items in the conclusion, let $i := b_j$. Otherwise, Lemma 6.14 says that any plain skew-twist of f is a single cluster and one of the factors f_i is not a monomial. In any case, the plain skew-twist $\phi^i \star \vec{f}$ or $\beta^{k-i} \star \vec{f}$ with the corresponding skew-clustering works. □

LEMMA 6.17. *If $\vec{g}$ and $\vec{a}$ satisfy one of the three conclusions of Lemma 6.16, and if $w \in \mathrm{BG}_k$ such that $w \star \vec{g} =: \vec{h}$ is defined, then $\vec{h}$ satisfies the same conclusion. If, furthermore, w contains no instances of β (respectively, γ), then the number of instances of γ (respectively, β) in w is bounded by 0 in the first case of Lemma 6.16, by 1 in the second case, and by the degree of g_k in the last case.*

These bounds are useful because in any word in the border guard monoid, the ψ's can be separated from the β's in the following sense.

LEMMA 6.18. *For any word* $w \in \mathrm{BG}_k$, *there are* $w_i \in \mathrm{BG}_k$ *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 $w \in \mathrm{BG}_k$, we find an equivalent word w' that has no substrings of the form $\psi u \gamma$ for some $u \in \mathrm{RM}_{k-1}$. Clearly, w' is the desired word. To construct w', we prove a

CLAIM. *For any* $u \in \mathrm{RM}_{k-1}$, *there are words* $v', v'' \in \mathrm{RM}_{k-1}$ *such that* $\psi u \gamma \approx v'$ *or* $\psi u \gamma \approx v'' \gamma t_{k-2} \psi v'$.

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 $u \in \mathrm{RM}_{k-1}$ is in first canonical form; i.e., either $u = v$ or $u = t_b t_{b-1} \cdots t_1 v$, where t_1 does not appear in $v \in \mathrm{RM}_{k-1}$. Now $v\gamma \approx \gamma v'$ for some word $v' \in \mathrm{RM}_{k-1}$ by part (4) of [Lemma 2.66](#). In the first case that $u = v$, this makes $\psi u \gamma \approx v'$. For the second case, by part (4) of [Lemma 2.66](#), $\psi t_b \ldots t_2 \approx v'' \psi$ where $v'' := t_{b-1} \ldots t_1$. Thus, $\psi u \gamma \approx v'' \psi t_1 \gamma v'$. To see that $\psi t_1 \gamma \approx \gamma t_{k-2} \beta$, recall that $t_{k-1} t_{k-2} t_{k-1} \approx t_{k-2} t_{k-1} t_{k-2}$ and $\beta t_{k-1} \phi \approx \phi t_1 \beta$, and use part (4) of [Lemma 2.66](#) again.   □

6.3. *Characterization of skew-invariant curves.* Finally, we show that every skew-invariant curve is encoded by some word in the augmented skew-twist monoid ST_k^+, give a normal form for such words, and thereby obtain a normal form for the skew-invariant curves.

PROPOSITION 6.19. *For any disintegrated polynomials* f *and* g *and any irreducible* (f, g)-*skew-invariant plane curve* $\mathcal{B}$, *there are a word* $w \in \mathrm{ST}_k^+$, *a decomposition* $(f_k, \ldots, f_1)$ *of* f, *and a curve* $\mathcal{A}$ *encoded by* $w \star \vec{f}$ *such that* $\mathcal{B} \subset \mathcal{A}$.

Proof. By [Proposition 2.34](#), $\mathcal{B} = \rho \circ (\pi^{-1})$ for some polynomials π and ρ; here π^{-1} denotes the converse relation to the graph of π and $\circ$ is curve composition in the sense of [Definition 2.13](#).

By successively factoring out single skew-twists from both sides of π, it is clearly possible to write $\pi := \pi_3 \circ \pi_2 \circ \pi_1$ so that π_1 is a skew-twist from h to some polynomial h_f, and π_3 is a skew-twist from some polynomial f_h to f, and the middle diagram $\pi_2^g \circ h_f = f_h \circ \pi_2$ satisfies the hypotheses of [Theorem 5.23](#); that is, h_f and π_2 share no initial compositional factors, and f_h and π_2^g share no terminal compositional factors. In a like manner, we may write $\rho = \rho_3 \circ \rho_2 \circ \rho_1$.

Now the graphs of ρ_i and the converses of graphs of π_i are encoded by words in ST_k^+, so the composition $\mathcal{B} = \rho_3 \circ \rho_2 \circ \rho_1 \circ (\pi_1^{-1}) \circ (\pi_2^{-1}) \circ (\pi_3^{-1})$ is encoded by the concatenation of these words. $\qquad\square$

PROPOSITION 6.20. *For any word $w \in \mathrm{ST}_k^+$ and any decomposition $\vec{f}$, there are integers $M, N \in \mathbb{N}$ with $M < k$, words $w_2, w_1 \in \mathrm{BG}_k$ with no instances of γ in w_1 and no instances of ψ in w_2, and a word $\tilde{w}$ consisting only of ε_p and δ_p for various p such that $w \approx_{\vec{f}} \phi^N \tilde{w} w_2 w_1 \phi^M$ or $w \approx_{\vec{f}} \beta^N \tilde{w} w_2 w_1 \phi^M$.*

Proof. By Lemma 6.16, we find $M < k$ such that $\vec{g} := \phi^M \star \vec{f}$ satisfies one of the three possible conclusions of that lemma. By Lemma 2.66, $w \approx w\beta^M\phi^M$. By Lemma 6.7, there are $\hat{w} \in \mathrm{ST}_k$ and $\tilde{w}$ consisting only of ε_p and δ_p for various p such that $w\beta^M \approx \tilde{w}\hat{w}$, and by Remark 6.6, $w \approx \tilde{w}\hat{w}\phi^M$. By Proposition 2.70 and Lemma 6.18, there are N and w_i as desired such that $\hat{w} \approx \phi^N w_2 w_1$ or $\hat{w} \approx \beta^N w_2 w_1$. Again by Remark 6.6, $w \approx \tilde{w}\phi^N w_2 w_1 \phi^M$ or $w \approx \tilde{w}\beta^N w_2 w_1 \phi^M$.

One last application of Lemma 6.7 finishes the proof. $\qquad\square$

Definition 6.21. A *monomial curve* is plane curve $C \subseteq \mathbb{A}^2$ defined by $x^n = y^m$ for some $m, n \in \mathbb{N}$.

Recall (Remark 2.62) that curves encoded by words in ST_k^+ are only defined up to a linear terminal compositional factor L that must be added manually.

THEOREM 6.22. *For any disintegrated polynomials f and g, any irreducible (f, g)-skew-invariant plane curve $\mathcal{B}$ is an irreducible component of $\hat{g} \circ \tilde{g} \circ \mathcal{A} \circ \mathcal{D} \circ \tilde{f}$ where*

- *$\tilde{f}$ is (the graph of) an initial compositional factor of f or linear L;*
- *$\mathcal{D}$ is a monomial curve encoded by a word in the border guard monoid whose degrees are bounded by Lemma 6.17, and in any case by the degree of f;*
- *$\mathcal{A}$ is a monomial curve encoded by a word in ε_p and δ_p whose degrees are bounded the product of in- and out-degrees of a compositional factor of f, and in any case by the degree of f;*
- *$\hat{g}$ and $\tilde{g}$ are one of the following for some $N \in \mathbb{N}$:*
  - *$\tilde{g}$ is the converse of the graph of an initial compositional factor of g^{σ^N} or linear, and $\hat{g}$ is the converse of the graph of $g^{\Diamond N}$;*
  - *$\tilde{g}$ is the graph of an initial compositional factor of $g^{\sigma^{-N}}$ or linear, and $\hat{g}$ is the graph of $(g^{(\sigma^{-N})})^{\Diamond N}$.*

Remark 6.23. If some factor of $\vec{f}$ is unswappable, or if some robust clustering of $\vec{f}$ has no gate at some cluster boundary, then $\mathcal{D}$ and $\mathcal{A}$ above must be diagonals. Then one of $\hat{g} \circ \tilde{g}$ and $\tilde{f}$ cancels with part of the other, and the whole (f, g)-skew-invariant plane curve $\mathcal{B}$ is the graph of something like $\hat{g} \circ \tilde{g}$. More generally, by Remark 5.19 amd Proposition 5.23, $\mathcal{A}$ is the diagonal unless

$\vec{f}$ is skew-conjugate to a polynomial of the form $x^k \cdot u(x^\ell)^n$ for some integers $k \geq 1$ and ℓ, n such that $\ell n \geq 2$.

The characterization of (f, g)-*invariant* curves becomes particularly simple when the two polynomials are the same.

THEOREM 6.24. *Fix an algebraic dynamical system* $(\mathbb{A}^2, (h, h))$ *for a disintegrated polynomial h. Any irreducible (h, h)-invariant plane curve $\mathcal{B}$ is the graph, or the converse of the graph, of $L \circ \tilde{h}^\ell$ for some linear linear L that commutes with some compositional power of h and some $\tilde{h}$ such that $\tilde{h}^{\circ r} = h$ for some r.*

Proof. Let $f := g := h$, and let $\hat{g}$, $\tilde{g}$, $\mathcal{A}$, $\mathcal{D}$, and $\tilde{f}$ be as in the conclusion of that theorem so that $\mathcal{B}$ is an irreducible component of $\hat{g} \circ \tilde{g} \circ \mathcal{A} \circ \mathcal{D} \circ \tilde{f}$.

Nontrivial $\mathcal{A}$ irreparably damages in- and out-degrees of factors f_i in a decomposition of $f = h$. This cannot be fixed by $\mathcal{D}$ or by Ritt swaps inside the decomposition because the monomials of $\mathcal{A}$ are not among the f_i (see Definition 5.18). If the same monomial occurs in both directions in $\mathcal{A}$, then $\mathcal{A}$ is reducible and its factors are given by replacing $x^p = y^p$ in its definition by $x = \xi y$ for various p^{th} roots of unity ξ.

Nontrivial $\mathcal{D}$ irreparably damages the gates of a clustering of $\vec{h}$ in the second case of Lemma 6.16 and irreparably damages the in- and out-degrees of the factor f_i guarding the border in the third case of Lemma 6.16.

Now, as in Remark 6.23, one of $\hat{g} \circ \tilde{g}$ and $\tilde{f}$ cancels part of the other, leaving behind the graph of a "fractional compositional power of h," since $f = g = h$ is defined over the fixed field of σ. That is, $\mathcal{B}$ is the graph (or the converse of the graph) of $h_0 \circ h^{\circ s}$ for some s, where h_0 is linear or the identity $h \circ h_0 = h_0 \circ h$ is a plain skew-twist. That is, for some h_1, both $h_0 \circ h_1 = h$ and $h_1 \circ h_0 = h$.

The theorem now follows from Ritt's theorem on commuting rational functions [16]. $\square$

Recall (see Corollary 2.35) that for every f and g, there is some h so that all (f, g)-invariant curves can be understood in terms of (h, h)-invariant curves.

LEMMA 6.25. *Suppose that f and g are disintegrated polynomials and there is an irreducible (f, g)-invariant curve. Then there are polynomials π, ρ, and h such that $\pi \circ h = f \circ \pi$ and $\rho \circ h = g \circ \rho$, and any irreducible (f, g)-invariant curve $\mathcal{A}$ is of the form $(\pi, \rho)(\mathcal{B}_0)$ for some irreducible (h, h)-periodic curve $\mathcal{B}_0$.*

Proof. Proposition 2.34 produces polynomials π, ρ, and h as required. Proposition 2.6 applied to the map of algebraic dynamical systems (π, ρ) : $(\mathbb{A}^2, (h, h)) \to (\mathbb{A}^2, (f, g))$ finishes the proof: any irreducible component $\mathcal{B}_0$ of $((\pi, \rho)^{-1}(\mathcal{A}))_{\text{inv}}$ works. $\square$

An (h, h)-periodic curve is $(h^{\circ m}, h^{\circ m})$-invariant for some m, so Theorem 6.24 almost applies to the conclusion of Lemma 6.25.

THEOREM 6.26. *Suppose that f and g are disintegrated polynomials and there is an irreducible (f, g)-invariant curve. Then there are polynomials π, ρ, and h such that $\pi \circ h = f \circ \pi$ and $\rho \circ h = g \circ \rho$, and any irreducible (f, g)-invariant curve $\mathcal{A}$ is of the form $(\pi, \rho)(\mathcal{B}_0)$, where $\mathcal{B}_0$ is the graph, or the converse of the graph, of $L \circ \tilde{h}^\ell$ for some linear L that commutes with some compositional power of h and some $\tilde{h}$ such that $\tilde{h}^{\circ r} = h^{\circ m}$ for some r and m.*

7. Applications

In this section we use our characterization of skew-invariant varieties to answer some open questions about the model theory of difference fields and the arithmetic of algebraic dynamical systems.

7.1. *Disintegrated minimal sets in* ACFA. In this subsection we address some fine structural questions about minimal sets in difference closed fields of characteristic zero. Specifically, we consider minimal sets of the form $(\mathbb{A}^1, f)^\sharp$. We show that if the isomorphism class of $(\mathbb{A}^1, f)$ is defined over the fixed field of some power of the distinguished automorphism, then nonorthogonality to $(\mathbb{A}^1, f)^\sharp$ is definable. Conversely, whenever the moduli point of $(\mathbb{A}^1, f)$ is transcendental over the fixed field, it is undefinable. We close out this section by showing that when f is disintegrated, $(\mathbb{A}^1, f)^\sharp$ has Morley rank one.

Notation 7.1. We fix a difference closed field $(\mathbb{U}, \sigma)$ of characteristic zero. All of the objects we discuss, such as polynomials, varieties, definable sets, etc., are defined over $\mathbb{U}$. Sometimes, we abuse notation writing expressions like "$a \in (X, f)^\sharp$" to mean that X is an algebraic variety over $\mathbb{U}$, $f : X \to X^\sigma$ is a dominant regular map, and $a \in (X, f)^\sharp(\mathbb{U}, \sigma)$ is a $(\mathbb{U}, \sigma)$-rational point of the σ-variety (X, f).

Notation 7.2. When we speak of properties of polynomials being definable, we are considering the polynomial ring as an ind-definable set. More concretely, we say that some class $\mathcal{K}$ of n-tuples of polynomials is definable if there is some natural number d and definable set $\mathsf{K} \subseteq M_{(d+1) \times n}(\mathbb{U})$ for which

$$\mathcal{K} = \Big\{ \Big(\sum_{i=0}^{d} a_{i,1} x^i, \dots, \sum_{i=0}^{d} a_{i,n} x^i \Big) : (a_{i,j}) \in \mathsf{K} \Big\}.$$

LEMMA 7.3. *For any given natural number d, the relation that two polynomials of degree d are skew-conjugate is definable.*

Proof. The action of the group of linear polynomials by skew-conjugation is definable with respect to our presentation of the space of degree d polynomials as a constructible subset of $\mathbb{A}^{d+1}(\mathbb{U})$. $\qquad \square$

LEMMA 7.4. *If f is a polynomial that is skew-conjugate to f^σ, then there is a polynomial g that is skew-conjugate to f and satisfies $g^\sigma = g$.*

Proof. By hypothesis, there is some linear λ with $f^\sigma = \lambda^\sigma \circ f \circ \lambda^{-1}$. As pre-composition with λ^{-1} defines an automorphism of the space of degree d polynomials, from the geometric axiom for difference closed fields, there is some linear μ satisfying $\mu^\sigma \circ \lambda = \mu$. Set $g := \mu^\sigma \circ f \circ \mu^{-1}$. Then $g^\sigma = \mu^{\sigma^2} \circ f^\sigma \circ (\mu^\sigma)^{-1} = \mu^{\sigma^2} \circ \lambda^\sigma \circ f \circ \lambda^{-1} \circ (\mu^\sigma)^{-1} = (\mu^\sigma \circ \lambda)^\sigma \circ f \circ (\mu^\sigma \circ \lambda)^{-1} = \mu^\sigma \circ f \circ \mu^{-1} = g$. □

Remark 7.5. The above lemmata hold more generally. For example, if $X = X^\sigma$ is an algebraic variety that descends to the fixed field, $\mathrm{Aut}(X)$ is represented by a connected algebraic group, and (X, f) is a σ-variety on X that is isomorphic as a σ-variety to (X, f^σ), then there is a map $g : X \to X$ for which $g^\sigma = g$ and (X, g) is isomorphic to (X, f).

With these observations in place, let us prove a theorem on definability of nonorthogonality.

THEOREM 7.6. *For a nonconstant polynomial f, the set of polynomials g with $(\mathbb{A}^1, g) \not\perp (\mathbb{A}^1, f)$ is definable if and only if f is not skew-conjugate to f^{σ^n} for every positive integer $n \in \mathbb{Z}_+$.*

Proof. If f is linear, then the set of polynomials orthogonal to f is precisely the set of linear polynomials, which is clearly definable. Likewise, if f is skew-conjugate to P_n or C_n or $-C_n$ for $n = \deg(f)$, then $(\mathbb{A}^1, f) \not\perp (\mathbb{A}^1, g)$ if and only if g is skew-conjugate to P_n or C_n or $-C_n$. As this class is also definable, we may restrict to the study of disintegrated f. By Proposition 6.19, g is nonorthogonal to f just in case the skew-conjugacy class of some decomposition of g is in the image of some (any) decomposition of f under the action of the augmented skew-twist monoid ST_k^+. As this monoid is countable, there are at most countably many skew-conjugacy classes of polynomials nonorthogonal to f. The map $f^{\Diamond n} : (\mathbb{A}^1, f) \to (\mathbb{A}^1, f^{\sigma^n})$ witnesses nonorthogonality between f and f^{σ^n}. Thus, if f is not skew-conjugate to any of its images under σ^n, we see that there are exactly $\aleph_0$ skew-conjugacy classes of polynomials nonorthogonal to f. As every infinite definable set in a sufficiently saturated model must be uncountable, we conclude that nonorthogonality to such an f is not definable.

Finally, consider the case when f is disintegrated and is skew-conjugate to some f^{σ^n}. Then by Lemma 7.4, we my assume that f itself is equal to f^{σ^n}. Considering Proposition 6.19 again we see that if g were nonorthogonal to f, then this nonorthogonality would be witnessed by the composition of plain skew-twists, monomial correspondences (with degrees bounded by $\deg(f)$), another plain skew-twist, and graphs of skew-composites $f^{\Diamond m}$ with $m < n$. As the set of such curves is finite (up to isomorphism), the set of polynomials skew-conjugate to f is definable. □

Remark 7.7. Curiously, if f is a polynomial defined over some small difference subfield K of $\mathbb{U}$, then model theoretic algebraic closure defines a locally finite closure operator on $(\mathbb{A}^1, f)^\sharp \setminus \mathrm{acl}(K)$ just in case the skew-conjugacy class of f is transcendental over the fixed field. If we further assume that K is finitely generated as a difference field, then it is not unreasonable to guess that $(\mathbb{A}^1, f)^\sharp \cap \mathrm{acl}(K)$ is finite. (This is plainly false when f is skew-conjugate to f^{σ^n} for some $n \in \mathbb{Z}_+$.)

Because quantifier elimination fails for ACFA, definable sets of D-rank one need not have Morley rank one. Indeed, it is easy to see that the fixed field is minimal, but its induced structure is unstable. More sophisticated examples of stable minimal sets of infinite multiplicity constructed from Hecke correspondences appear in [4]. Here we show that if f is a disintegrated polynomial, then $(\mathbb{A}^1, f)^\sharp$ has Morley rank one and is "usually" strongly minimal.

LEMMA 7.8. *Let X be an algebraic variety for which $X = X^\sigma$. Let $f : X \to X$ and $g : X \to X$ be two self-maps. Define $\phi := f \circ g$ and $\psi := g^\sigma \circ f$. Then $g : (X, \phi)^\sharp \to (X, \psi)^\sharp$ is onto:*

$$
\begin{array}{ccccc}
X & \xrightarrow{\;g\;} & X & \xrightarrow{\;f\;} & X \\
\phi \downarrow & & \downarrow \psi & & \downarrow \phi^\sigma \\
X & \xrightarrow{\;g^\sigma\;} & X & \xrightarrow{\;f^\sigma\;} & X.
\end{array}
$$

Remark 7.9. This proof works for any inversive difference field. It is not necessary to work with a difference closed field.

Proof. Let $P \in (X, \psi)^\sharp$. Set $Q := \sigma^{-1} f(P)$. Let us note that the identity $\phi^\sigma \circ f = f^\sigma \circ g^\sigma \circ f = f^\sigma \circ \psi$ shows that $f : (X, \psi) \to (X, \phi^\sigma)$ is a map of σ-varieties so that $f(P) \in (X, \phi^\sigma)^\sharp$. Applying σ^{-1}, we have $Q \in (X, \phi)^\sharp$. We compute $g(Q) = g\sigma^{-1} f(P) = \sigma^{-1} g^\sigma f(P) = \sigma^{-1} \psi(P) = \sigma^{-1} \sigma(P) = P$. $\square$

LEMMA 7.10. *If (X, f) is any σ-variety and $n \in \mathbb{Z}_+$, then the map $f^{\Diamond n} : (X, f)^\sharp \to (X^{\sigma^n}, f^{\sigma^n})$ is onto.*

Proof. Working by induction on n, it suffices to consider the case of $n = 1$. Let $a \in (X^\sigma, f^\sigma)^\sharp$. Set $b := \sigma^{-1}(a)$. Then $f(b) = f(\sigma^{-1}(a)) = \sigma^{-1} f^\sigma(a) = \sigma^{-1} \sigma(a) = a$. $\square$

With the next lemma we characterize the image of a power map.

LEMMA 7.11. *Given a nonconstant polynomial u, positive integer k, and a prime ℓ, we set $f(x) := x^k u(x^\ell)$ and $g(x) := x^k u(x)^\ell$. In general, $(\mathbb{A}^1, g)^\sharp$ is the image of P_ℓ on $\bigcup_{\zeta \in \mu_\ell} (\mathbb{A}^1, \zeta f)^\sharp$. If σ does not act on μ_ℓ, the group of ℓth roots of unity, by raising to the kth power, then $P_\ell : (\mathbb{A}^1, f)^\sharp \to (\mathbb{A}^1, g)^\sharp$ is surjective.*

Proof. Let $a \in (\mathbb{A}^1, g)^\sharp$. Let $b \in \mathbb{A}^1(\mathbb{U})$ be any solution to $P_\ell(b) = a$. From the equation $P_\ell \circ f = g \circ P_\ell$, we see that $P_\ell(f(b)) = g(a) = \sigma(a)$ while we also know that $P_\ell(\sigma(b)) = \sigma(a)$. Hence, there is some $\xi \in \mu_\ell$ for which $\xi f(b) = \sigma(b)$.

We assume now that σ does not act by raising the kth power on μ_ℓ. Thus, the map $\mu_\ell \to \mu_\ell$ given by $\zeta \mapsto \sigma(\zeta)/\zeta^k$ is onto. Thus, we may choose $\zeta \in \mu_\ell$ with $\sigma(\zeta)/\zeta^k = \xi$. We compute $f(\zeta b) = (\zeta b)^k u((\zeta b)^\ell) = \zeta^k b^k u(b^\ell) = \zeta^k f(b) = \zeta^k \xi \sigma(b) = \zeta^k \xi \sigma(\zeta)^{-1} \sigma(\zeta b) = \sigma(\zeta b)$. Thus, $\zeta b \in (\mathbb{A}^1, f)^\sharp$ and $P_\ell(\zeta b) = a$. $\square$

Combining the above lemmata, we conclude that disintegrated sets of the form $(\mathbb{A}^1, f)^\sharp$ have Morley rank one.

THEOREM 7.12. *If f is a disintegrated polynomial, then $(\mathbb{A}^1, f)^\sharp$ has Morley rank one.*

Proof. The quantifier-elimination to bounded existential quantifiers for ACFA together with the work around finite σ-stable extensions in [4] imply that every infinite definable subset of $(\mathbb{A}^1, f)^\sharp$ is (up to a finite sets) a finite union of sets of the form $h(\mathbb{A}^1, g)^\sharp$, where $h : (\mathbb{A}^1, g) \to (\mathbb{A}^1, f)$ is a map of σ-varieties.

By Theorem 6.22, h may be expressed as a composite of a sequence of skew-twists, power maps of degree bounded by $\deg(f)$, and maps of the form $k^{\diamond n}$. By Lemmata 7.8 and 7.10, the maps of the first and third type are always onto. By Lemma 7.11, there are at most $\deg(f)$ many distinct sets arising from the power maps. $\square$

7.2. *Density of dynamical orbits.* In this subsection we apply Theorem 6.24 to deduce a version of a conjecture of Zhang on the density of dynamical orbits. Let us recall Zhang's conjecture.

CONJECTURE 7.13 (Conjecture 4.1.6 of [20]). *Let $f : X \to X$ be a polarizable dynamical system over a number field k. Then there is point $a \in X(k^{\text{alg}})$ 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 7.13 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 7.13 due to Amerik and Campana [2], we propose a more general conjecture on the density of dynamical orbits.

CONJECTURE 7.14. *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 do not exist a positive dimensional algebraic variety Y and a 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.*

Remark 7.15. In [1], Amerik, Bogomolov, and Ravinsky prove some instances of Conjecture 7.14, without imposing any polarizability hypotheses, but instead arguing from the local behaviour of the dynamical system.

We shall prove the instance of Conjecture 7.14 in which X is affine space and Φ is given by a sequence of univariate polynomials.

THEOREM 7.16. *Let K be a field of characteristic zero, $f_1, \ldots, f_n \in K[x]$ nonconstant polynomials over K in one variable. Suppose that the linear polynomials amongst the f_i's are independent in the sense Definition 2.26. Let $\Phi : \mathbb{A}_K^n \to \mathbb{A}_K^n$ be given by $(x_1, \ldots, x_n) \mapsto (f_1(x_1), \ldots, f_n(x_n))$. Then there is a point $a \in \mathbb{A}^n(K)$ for which $\mathcal{O}_\Phi(a)$ is Zariski dense.*

Remark 7.17. 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 7.18. As the reader will see, the notion of *independence* is exactly what is required so that there is no dominant map from $(\mathbb{A}^n, \Phi)$ to a positive dimensional trivial algebraic dynamical system. 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 7.19. Theorem 7.16 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 σ-subvariety of $(\mathbb{A}^n, \Phi)$ when K is treated as a difference field with $\sigma = \mathrm{id}_K$. In fact, we prove Theorem 7.16 by explicitly describing the irreducible σ-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 σ-subvarieties.

We prove Theorem 7.16 as a consequence of a number of simple lemmata.

LEMMA 7.20. *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 $\overline{\mathcal{O}_f(a)} \subseteq Y \cup \{f^{\circ i}(a) : 0 \leq i \leq m\} \subsetneq X$. $\square$

LEMMA 7.21. *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 $Z_{j+1 \mod m}$, we must have $X = Z_j$ for all j. In particular, $X = Z_0 = \overline{\mathcal{O}_{f^{\circ m}}(a)}$. $\square$

LEMMA 7.22. *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 $(f, g)(\mathcal{O}_{(f,g)}(a, b)) \subseteq \mathcal{O}_{(f,g)}(a, b)$, possibly after removing finitely many points the variety Z is (f, g)-invariant. As $(X, f) \perp (Y, g)$, Z_{inv} must be a finite union of varieties of the form $A \times B$ where $A \subseteq X$ is f-invariant and $B \subseteq Y$ is g-invariant. Let $A \times B$ be a component containing (a, b). By Lemma 7.21, $X = \overline{\mathcal{O}_f(a)} \subseteq A \subseteq X$ and $Y = \overline{\mathcal{O}_g(b)} \subseteq B \subseteq Y$. Hence, $X \times Y = \overline{\mathcal{O}_{(f,g)}(a, b)}$. $\square$

LEMMA 7.23. *Let K be a field of characteristic zero and f and g two disintegrated polynomials over K. Then there is a point $(a, b) \in \mathbb{A}^2(K)$ for which $\mathcal{O}_{(f,g)}(a, b)$ is Zariski dense in $\mathbb{A}_K^2$.*

Proof. If f and g are orthogonal, Lemma 7.22 shows that (a, b) works whenever a is not f-preperiodic and b is not g-preperiodic, which can always be arranged. Thus we may assume that f and g are nonorthogonal. By Theorem 2.35 there are a natural number m, a polynomial h, and dominant maps of dynamical systems $\rho : (\mathbb{A}^1, h) \to (\mathbb{A}^1, f^{\circ m})$ and $\pi : (\mathbb{A}^1, h) \to (\mathbb{A}^1, g^{\circ m})$. It follows from Ritt's theorem on polynomials with common iterates [14] that there is a maximal k for which we may write $h = L^{-1} \circ (\zeta \tilde{h}^{\circ k}) \circ L$ for some linear L and some root of unity ζ. Let R be a finitely generated subring of K over which f, g, $\tilde{h}$, π, ρ, and all symmetries of $\tilde{h}$ are defined and which contains the multiplicative inverse of the leading coefficient of each of these polynomials. Let $\widetilde{R}$ be the integral closure of R in its field of fractions, regarded as a subfield of K. Let $\hat{a} \in \widetilde{R}$ and $\hat{b} \in K \setminus \widetilde{R}$ so that neither $\hat{a}$ nor $\hat{b}$ is h-pre-periodic. Set $a := \pi(\hat{a})$ and $b := \rho(\hat{b})$.

We claim that there is no (possibly reducible) weakly (h, h)-invariant curve C with $(\hat{a}, \hat{b}) \in C(K)$. As neither $\hat{a}$ nor $\hat{b}$ is preperiodic, we see that we may assume that each component of C projects dominantly in both directions. If $(\hat{a}, \hat{b}) \in C(K)$, then for $n \gg 0$, we would have $(h^{\circ n}(\hat{a}), h^{\circ n}(\hat{b})) \in C_{\mathrm{inv}}(K)$ so that $(h^{\circ n}(\hat{a}), h^{\circ n}(\hat{b}))$ would lie on an irreducible $(h^{\circ nN}, h^{\circ nN})$-invariant curve for some $N \gg 0$. By Theorem 6.24, such a curve is defined by $y = L \circ \tilde{h}^\ell(x)$ or $x = L \circ \tilde{h}^\ell(y)$. Neither such curve can contain a K-rational point of the form (c, d) with $c \in \widetilde{R}$ and $d \notin \widetilde{R}$, which is exactly the form of $(h^{\circ n}(\hat{a}), h^{\circ n}(\hat{b}))$ as the polynomial h maps R to R and cannot map a nonintegral point to an integral point as its leading coefficient is a unit.

It follows that $(a, b) = (f, g)(\hat{a}, \hat{b})$ cannot lie on any weakly (f, g)-invariant curve as the pullback of such a curve would be weakly (h, h)-invariant. Thus, $\overline{\mathcal{O}_{(f,g)}(a, b)} = \mathbb{A}^2$. $\qquad\square$

LEMMA 7.24. *Let K be a field of characteristic zero, and let $f, \ldots, f_r \in K[x]$ be a a sequence of nonconstant polynomials over K. We assume that each f_i has degree at least two and is not conjugate to a monomial, Chebyshev polynomial, or negative Chebyshev polynomial. Then there is a rational point $a = (a_1, \ldots, a_n) \in \mathbb{A}^n(K)$ with a dense $(f_1, \ldots, f_n)$-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 $(a_1, \ldots, a_i) \in \mathbb{A}^i(B)$ with $\mathcal{O}_{(f_1, \ldots, f_i)}(a)$ 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 $\lim_{m \to \infty} f_1^{\circ m}(a) = \infty$ so that, in particular, a is not preperiodic.)

In the inductive case, we have $(a_1, \ldots, a_i) \in \mathbb{A}^i(B)$ with a Zariski dense $(f_1, \ldots, f_i)$-orbit. Let $a_{i+1} \in K$ be any element of K that is not integral over B. Then for every m, $f^{\circ m}(a_{n+1})$ is also nonintegral, so by the second half of the proof of Lemma 7.23, $(f^{\circ m}(a_j), f^{\circ m}(a_{i+1}))$ does not belong to any $(f_j^{\circ m}, f_{i+1}^{\circ m})$-invariant curve. By disintegratedness, it follows that $(f_1^{\circ m}(a), \ldots, f_{i+1}^{\circ m}(a))$ does not belong to any $(f_1^{\circ m}, \ldots, f_{i+1}^{\circ m})$-invariant variety. $\qquad\square$

Let us now combine these results to complete the proof of Theorem 7.16.

Proof. Reordering the indices if need be, we may express $(\mathbb{A}^n, \Phi)$ as a product $(\mathbb{A}^{n_\ell}, \lambda) \times (\mathbb{A}^{n_G}, \gamma) \times (\mathbb{A}^{n_t}, \tau)$, where λ is given by a sequence of univariate linear polynomials, γ is given by a sequence of polynomials of degree at least two each conjugate to a monomial, Chebyshev polynomial, or negative Chebyshev polynomial, and τ is given by a sequence of disintegrated polynomials. By Proposition 2.27 there is some $a \in \mathbb{A}^{n_\ell}(K)$ with $\mathcal{O}_\lambda(a)$ Zariski dense

in $\mathbb{A}^{n_\ell}$, by Proposition 2.28 there is some $b \in \mathbb{A}^{n_G}(K)$ with $\mathcal{O}_\gamma(b)$ Zariski dense in $\mathbb{A}^{n_G}$, and by Lemma 7.24 there is some $c \in \mathbb{A}^{n_t})(K)$ with $\mathcal{O}_\tau(x)$ Zariski dense in $\mathbb{A}^{n_t}$. By Lemma 7.22, $\mathcal{O}_\Phi((a, b, c))$ is Zariski dense in $\mathbb{A}^n$. $\square$

7.3. *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 amongst the periodic points of such dynamical systems.

Notation 7.25. 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.

Notation 7.26. 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 7.27 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 7.27. *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 $f : U \to \mathbb{A}_R^m$ where $\widetilde{a} \in U(R)$, $U \subseteq X$ is an affine open subset and $f(\widetilde{a}) = \mathbf{0}$. Note that $f^\rho : U^\rho \to \mathbb{A}_R^m$ gives analytic coordinates on X^ρ near $\rho(\widetilde{a})$.

As $\Gamma \to X$ is étale, the set $(f, f^\rho)(\Gamma(R) \cap \pi^{-1}\{a\} \times (\pi^\rho)^{-1}\{F(a)\})$ is the graph of an analytic function $g : \mathfrak{m}^m \to \mathfrak{m}^m$ where $g(x_1, \ldots, x_m) \equiv (x_1^q, \ldots, x_m^q)$ mod $\mathfrak{m} \cdot R[[x_1, \ldots, x_m]]$. That we can find a solution to $g(x) = \rho(x)$ follows from Newton's method. (See [18] in this context.)

That is, if for some $\gamma > 0$ we have a solution to $g(x) \equiv \rho(x) \mod I_\gamma$ where $I_\gamma := \{z \in R : v(z) \geq \gamma\}$, we can find some x' with $x \equiv x' \mod I_\gamma$ but $g(x) \equiv \rho(x) \mod I_{\gamma+} := \{z \in R : v(z) > \gamma\}$, and then taking limits we find a true solution within the given neighborhood. In our case, we already know that $g(\mathbf{0}) \equiv \mathbf{0} \mod \mathfrak{m} = I_{0+}$. Given an approximate solution x, suppose that $g(x) \equiv \rho(x) \mod I_\gamma$ with $\gamma > 0$. Let $\varepsilon \in R$ with $v(\varepsilon) = \gamma$. We seek to find $x' = x + c\varepsilon$ with $c = (c_1, \ldots, c_m)$ and $v(c_i) \geq 0$ for each i. We have $g(x + c\varepsilon) = g(x) + \sum_{i=1}^m \frac{\partial g}{\partial X_i}(x)c\varepsilon + \varepsilon^2 * \equiv g(x) \mod I_{\gamma+}$ as $\frac{\partial g}{\partial X_i} \equiv qX_i^{q-1}$ mod $\mathfrak{m}R[[X_1, \ldots, X_m]]$. On the other hand, $\rho(x + c\varepsilon) = \rho(x) + \rho(c)\rho(\varepsilon) \equiv \rho(x) + (c_1^q, \ldots, c_m^q)\rho(\varepsilon) \mod I_{\gamma+}$. Subtracting, we need only solve $\rho(\varepsilon)(c_1^q, \ldots, c_m^q) \equiv g(x) - \rho(x) \mod I_{\gamma+}$. By hypothesis, each component of $g(x) - \rho(x)$ has valuation at least $\gamma = v(\rho(\varepsilon))$. 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 $c = (c_1, \ldots, c_m)$ is uniquely determined modulo $\mathfrak{m}$. Since we know the residue of the solution, this shows that the reduction map is injective. $\qquad \square$

COROLLARY 7.28. *With X and Γ as in Theorem 7.27, 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 hypotheses of Theorem 7.27 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. $\qquad \square$

Specializing Γ somewhat, we may use Theorem 7.27 to find a difference equation for periodic points.

THEOREM 7.29. *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 7.28. $\qquad \square$

Remark 7.30. Theorem 7.29 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 [17] for more details.)

Remark 7.31. If in Theorem 7.29 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 7.32. 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 [10], [13]. 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 7.33. *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* $i \le n$. *We suppose that for some* $m > 0$, *each* $f_i = f_i^{\sigma^m}$ *for each* i. *If* $X \subseteq \mathbb{A}_K^n$ *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 Theorem 6.26. 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 7.29, 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 7.34. Further specializing Theorem 7.33, one obtains statements about algebraic relations amongst the periodic points of a polynomial without

reference to valuations as announced in the introduction. 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 or a negative Chebyshev polynomial. Then every irreducible variety $X \subseteq \mathbb{A}^n_{\mathbb{C}}$ that contains a Zariski dense set of n-tuples of f-periodic points is defined by a sequence of equations of the form $g(x_i) = x_j$ or $g(x_\ell) = a$ for a some fixed f-periodic point and g a polynomial that commutes with f.

References

[1] E. AMERIK, F. BOGOMOLOV, and M. ROVINSKY, Remarks on endomorphisms and rational points, *Compos. Math.* **147** (2011), 1819–1842. MR 2862064. Zbl 1231.14014. http://dx.doi.org/10.1112/S0010437X11005537.

[2] E. AMERIK and F. CAMPANA, Fibrations méromorphes sur certaines variétés à fibré canonique trivial, *Pure Appl. Math. Q.* **4** (2008), 509–545. MR 2400885. Zbl 1143.14035. http://dx.doi.org/10.4310/PAMQ.2008.v4.n2.a9.

[3] Z. CHATZIDAKIS, Groups definable in ACFA, in *Algebraic Model Theory* (Toronto, ON, 1996), *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **496**, Kluwer Acad. Publ., Dordrecht, 1997, pp. 25–52. MR 1481438. Zbl 0884.03038. http://dx.doi.org/10.1007/978-94-015-8923-9_2.

[4] Z. CHATZIDAKIS and E. HRUSHOVSKI, Model theory of difference fields, *Trans. Amer. Math. Soc.* **351** (1999), 2997–3071. MR 1652269. Zbl 0922.03054. http://dx.doi.org/10.1090/S0002-9947-99-02498-8.

[5] Z. CHATZIDAKIS and E. HRUSHOVSKI, Difference fields and descent in algebraic dynamics. I, *J. Inst. Math. Jussieu* **7** (2008), 653–686. MR 2469450. Zbl 1165.03014. http://dx.doi.org/10.1017/S1474748008000273.

[6] Z. CHATZIDAKIS, E. HRUSHOVSKI, and Y. PETERZIL, Model theory of difference fields. II. Periodic ideals and the trichotomy in all characteristics, *Proc. London Math. Soc.* **85** (2002), 257–311. MR 1912052. Zbl 1025.03026. http://dx.doi.org/10.1112/S0024611502013576.

[7] D. GHIOCA, T. J. TUCKER, and S. ZHANG, Towards a dynamical Manin-Mumford conjecture, *Int. Math. Res. Not.* **2011** (2011), 5109–5122. MR 2854724. Zbl 1267.37110. http://dx.doi.org/10.1093/imrn/rnq283.

[8] M. HINDRY, Autour d'une conjecture de Serge Lang, *Invent. Math.* **94** (1988), 575–603. MR 0969244. Zbl 0638.14026. http://dx.doi.org/10.1007/BF01394276.

[9] E. HRUSHOVSKI and M. ITAI, On model complete differential fields, *Trans. Amer. Math. Soc.* **355** (2003), 4267–4296. MR 1990753. Zbl 1021.03024. http://dx.doi.org/10.1090/S0002-9947-03-03264-1.

[10] E. HRUSHOVSKI, The Manin-Mumford conjecture and the model theory of difference fields, *Ann. Pure Appl. Logic* **112** (2001), 43–115. MR 1854232. Zbl 0987.03036. http://dx.doi.org/10.1016/S0168-0072(01)00096-3.

[11] A. MEDVEDEV, *Minimal sets in ACFA*, ProQuest LLC, Ann Arbor, MI, 2007, Thesis (Ph.D.)–University of California, Berkeley. MR 2710587.

Zbl 1235.03069. Available at http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3275517.

[12] A. MEDVEDEV and T. SCANLON, Polynomial dynamics, 2009, 68 pages. arXiv 0901.2352v2.

[13] R. PINK and D. ROESSLER, On ψ-invariant subvarieties of semiabelian varieties and the Manin-Mumford conjecture, *J. Algebraic Geom.* **13** (2004), 771–798. MR 2073195. Zbl 1072.14054. http://dx.doi.org/10.1090/S1056-3911-04-00368-6.

[14] J. F. RITT, On the iteration of rational functions, *Trans. Amer. Math. Soc.* **21** (1920), 348–356. MR 1501149. JFM 47.0312.01. http://dx.doi.org/10.2307/1988936.

[15] J. F. RITT, Prime and composite polynomials, *Trans. Amer. Math. Soc.* **23** (1922), 51–66. MR 1501189. JFM 48.0079.01. http://dx.doi.org/10.2307/1988911.

[16] J. F. RITT, Permutable rational functions, *Trans. Amer. Math. Soc.* **25** (1923), 399–448. MR 1501252. JFM 49.0712.02. http://dx.doi.org/10.2307/1989297.

[17] T. SCANLON, Local André-Oort conjecture for the universal abelian variety, *Invent. Math.* **163** (2006), 191–211. MR 2208421. Zbl 1086.14020. http://dx.doi.org/10.1007/s00222-005-0460-1.

[18] T. SCANLON, Analytic difference rings, in *International Congress of Mathematicians. Vol. II* (Madrid, August 22–30, 2006), Eur. Math. Soc., Zürich, 2007, pp. 71–92. MR 2275589. Zbl 1104.03029. http://dx.doi.org/10.4171/022-2/4.

[19] J. H. SILVERMAN, *The Arithmetic of Dynamical Systems*, *Grad. Texts in Math.* **241**, Springer-Verlag, New York, 2007. MR 2316407. Zbl 1130.37001. http://dx.doi.org/10.1007/978-0-387-69904-2.

[20] S.-W. ZHANG, Distributions in algebraic dynamics, in *Surveys in Differential Geometry. Vol. X*, *Surv. Differ. Geom.* **10**, Int. Press, Somerville, MA, 2006, pp. 381–430. MR 2408228. Zbl 1207.37057.

[21] M. ZIEVE and P. MÜLLER, On Ritt's polynomial decomposition theorems, 2008, 38 pages. arXiv 0807.3578v1.

THE CITY COLLEGE OF NEW YORK, NEW YORK, NY
E-mail: amedvedev@ccny.cuny.edu

UNIVERSITY OF CALIFORNIA, BERKELEY, CA
E-mail: scanlon@math.berkeley.edu