Combinatorics, Probability and Computing (2017) 26, 99–117. © Cambridge University Press 2016
doi:10.1017/S0963548316000225

Distinct Distances on Algebraic Curves in the Plane

JÁNOS PACH[1†] and FRANK DE ZEEUW[2‡]

[1]Department of Mathematics, EPFL, Lausanne, Switzerland
and Rényi Institute, Budapest, Hungary
(e-mail: pach@cims.nyu.edu)
[2]Department of Mathematics, EPFL, Lausanne, Switzerland
(e-mail: fdezeeuw@gmail.com)

Received 20 April 2014; revised 23 April 2016; first published online 15 July 2016

Let S be a set of n points in $\mathbb{R}^2$ contained in an algebraic curve C of degree d. We prove that the number of distinct distances determined by S is at least $c_d n^{4/3}$, unless C contains a line or a circle.

We also prove the lower bound $c'_d \min\{m^{2/3}n^{2/3}, m^2, n^2\}$ for the number of distinct distances between m points on one irreducible plane algebraic curve and n points on another, unless the two curves are parallel lines, orthogonal lines, or concentric circles. This generalizes a result on distances between lines of Sharir, Sheffer and Solymosi in [19].

2010 *Mathematics subject classification*: Primary 52C10
Secondary 52C45

1. Introduction

A famous conjecture of Erdős, first mentioned in [9], states that any set of n points in $\mathbb{R}^2$ determines at least $\Omega(n/\sqrt{\log n})$ distinct distances. Over the years this has been a central problem in combinatorial geometry, with many successive improvements of the best-known lower bound (see [3, Section 5.3]). In [11], Guth and Katz established an almost complete solution, proving the lower bound $\Omega(n/\log n)$. A new element in their proof was the use of tools from algebraic geometry.

A related problem posed by Purdy (see [3, Section 5.5]) is to determine the least number of distances that can occur between two collinear point sets, say n points on a line l_1 and n points on a line l_2. If l_1 and l_2 are parallel or orthogonal, then $O(n)$ distances are possible, but otherwise there should be substantially more. This was proved by Elekes and Rónyai in [7],

† Supported by NSF grant CCF-08-30272, by OTKA under EuroGIGA projects GraDR and ComPoSe 10-EuroGIGA-OP-003, and by Swiss National Science Foundation grants 200020-144531, 200020-162884, and 200021-137574.
‡ Supported by Swiss National Science Foundation grants 200020-144531 and 200021-137574.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

where they derived it from a more general result about polynomials, which they proved using a combination of combinatorial and algebraic methods. In [6], Elekes specialized these methods to Purdy's question, resulting in a lower bound of $\Omega(n^{5/4})$ on the number of distances, if the two lines are not parallel or orthogonal. Recently, Sharir, Sheffer and Solymosi improved this bound to $\Omega(n^{4/3})$ in [19], again using algebraic methods. In [18], Schwartz, Solymosi and de Zeeuw extended the general result of Elekes and Rónyai in several ways, one of which resulted in an unbalanced version of Purdy's problem, where one line contains m points and the other n. This was also strengthened for Purdy's problem in [19], to a lower bound $\Omega(\min\{m^{2/3}n^{2/3}, m^2, n^2\})$.

The aim of this paper is to extend the result of [19] from lines to arbitrary plane algebraic curves (see Section 2.1 for definitions). The results take several forms; perhaps the most interesting of them is the following.

Theorem 1.1. *Let C be a plane algebraic curve of degree d that does not contain a line or a circle. Then any set of n points on C determines at least $c_d n^{4/3}$ distinct distances, for some $c_d > 0$ depending only on d.*

Note that if the curve is a line or a circle, $O(n)$ distances are possible for certain point sets, including any sequence of equidistant points. With the current proof, the constant c_d roughly comes out to cd^{-7} for an absolute constant c. We have not made a serious effort to optimize it, but in a remark at the end of Section 3 we suggest some improvements.

Theorem 1.1 is a direct consequence of the proof of the following Theorem.

Theorem 1.2. *Let C_1, C_2 be two irreducible plane algebraic curves of degree at most d which are not parallel lines, orthogonal lines, or concentric circles. Then for any m points on C_1 and n points on C_2, the number of distinct distances between the two sets is at least*

$$c'_d \cdot \min\{m^{2/3}n^{2/3}, m^2, n^2\},$$

for some $c'_d > 0$ depending only on d.

In the excluded cases, $O(n)$ distances are again possible for certain point sets. One can also deduce a result for two curves that are not necessarily irreducible, but it would be more inconvenient to state as follows: The curves should not both contain a line, such that the two lines are parallel or orthogonal, and they should not both contain a circle, such that the two circles are concentric.

During the finalization of this paper, a number of related results have been made public. In [4], Charalambides establishes a version of Theorem 1.1 with the weaker lower bound $c_d n^{5/4}$. He combines the technique of [6] with analytic as well as algebraic tools, and even extends it to higher dimensions, with a more complicated set of exceptions. In [20], Sharir and Solymosi show, using a method based on that of [19], that between three non-collinear points and n other points there are $\Omega(n^{6/11})$ distinct distances. In [21], Sheffer, Zahl and de Zeeuw extend the method of [19] to the case where one set of points in $\mathbb{R}^2$ is constrained to a line, while the other is unconstrained. Finally, in [17], Raz, Sharir and Solymosi use the approach of [19] to improve the bounds in the more general problem about polynomials of [7].

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

We say a few words about our proofs compared to those of the similar results mentioned above. Both [6] and [19] derive their bound by constructing a set of new curves and applying an incidence bound to them. The way these curves are constructed relies heavily on the fact that lines can be parametrized. This makes it possible to extend their methods to parametrizable curves, but makes it harder to extend to general algebraic curves, which are defined by an implicit equation. In [4], this is overcome using the Implicit Function Theorem, which allows implicit curves to be 'parametrized' analytically. One important new element of our proofs is that we construct the new curves in an implicit and algebraic way (see in particular (3.1) in Section 3), making parametrization unnecessary.

To apply the incidence bound to the constructed curves, one needs to show that the curves have small intersection, and in particular that they are distinct. In [6] and [19], this is relatively easy because the curves have low degree. In [4], it is done using concepts from the theory of rigidity. We observe instead that, if some of the constructed curves have large intersection, this must be due to some kind of symmetry of the original curve. The only curves that have too many symmetries are lines and circles, which is why these are the exceptions in Theorem 1.1. In Theorem 1.2, the exceptions are those pairs of curves that have too many symmetries *in common*.

In Section 2.1, we define algebraic curves and state several results from algebraic geometry, in a way that we hope is accessible to readers that are not familiar with algebraic geometry. In Section 2.2 we introduce the incidence bound that is central to our proof, and in Section 2.3 we prove two elementary results about symmetries of curves. In Section 3, we give the proof of our two main theorems, up to the more delicate proof of one lemma, which we give separately in Section 4.

2. Preliminaries

2.1. Definitions and tools from algebraic geometry

We define a set $C \subset \mathbb{R}^2$ to be a (*plane*) *algebraic curve* if there is a polynomial $f \in \mathbb{R}[x,y]\backslash\{0\}$ such that

$$C = Z_{\mathbb{R}}(f) = \{(a,b) \in \mathbb{R}^2 : f(a,b) = 0\}.$$

Note that in our definition, an algebraic curve can also be a *finite* set, *e.g.*, $Z_{\mathbb{R}}((x(x-1))^2+y^2)$. Fortunately, the size of this finite set is bounded in terms of the degree of the polynomial (by Theorem 2.2 below), so this is not a problem in our main theorem, since for finite sets of bounded size the statement is trivial. We define the *degree* of a curve C to be the degree of a minimum-degree polynomial f such that $C = Z_{\mathbb{R}}(f)$. We call a curve a *conic* if it has degree two, it is not a union of two lines, and it is not a finite set of points.

We say that a plane algebraic curve C is *irreducible* if we can write $C = Z_{\mathbb{R}}(f)$ with a polynomial $f \in \mathbb{R}[x,y]$ that is irreducible over $\mathbb{R}$. By an *irreducible component* of an algebraic curve $Z_{\mathbb{R}}(f)$ we mean an irreducible algebraic curve $Z_{\mathbb{R}}(h)$ for some non-constant $h \in \mathbb{R}[x,y]$ that divides f; it then follows that $Z_{\mathbb{R}}(h) \subset Z_{\mathbb{R}}(f)$. We say that two curves $Z_{\mathbb{R}}(f)$ and $Z_{\mathbb{R}}(g)$ *have a common component* if there is a non-constant polynomial $h \in \mathbb{R}[x,y]$ that divides f and g; it then follows that $Z_{\mathbb{R}}(h) \subset Z_{\mathbb{R}}(f) \cap Z_{\mathbb{R}}(g)$.

We frequently use Bézout's inequality in the plane, which is an upper bound on the number of intersection points of algebraic curves. It is in fact an equality (Bézout's theorem) if one defines

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

multiplicities of intersection points and works in the complex projective plane, but for us the inequality suffices. See [10, Lemma 14.4] for exactly the statement below, or [5, Theorem 8.7.10] for the complex version.

Theorem 2.1 (Bézout's inequality). *Two algebraic curves in $\mathbb{R}^2$ with degrees d_1 and d_2 have at most $d_1 \cdot d_2$ intersection points, unless they have a common component.*

Although our objects of study are curves in the plane, our proofs involve curves in higher dimensions. Specifically, we encounter curves in $\mathbb{R}^4$ that are zero sets of three polynomials. To analyse these real curves, we also consider complex zero sets.

Given polynomials $f_1,\dots,f_k \in \mathbb{R}[x_1,\dots,x_D]$, we define the (real) *zero set*

$$Z_{\mathbb{R}}(f_1,\dots,f_k) = \{p \in \mathbb{R}^D : \forall i \; f_i(p) = 0\},$$

and given polynomials $f_1,\dots,f_k \in \mathbb{C}[x_1,\dots,x_D]$, we define the (complex) zero set

$$Z_{\mathbb{C}}(f_1,\dots,f_k) = \{p \in \mathbb{C}^D : \forall i \; f_i(p) = 0\}.$$

We also refer to zero sets as *varieties*. Given a real zero set $Z_{\mathbb{R}}(f_1,\dots,f_k)$, we may consider the corresponding complex zero set $Z_{\mathbb{C}}(f_1,\dots,f_k)$; note that this complex zero set depends on the choice of polynomials defining the real zero set. For a complex zero set $X \subset \mathbb{C}^D$ we define its *real part* to be $X \cap \mathbb{R}^D$ (where we view $\mathbb{R}^D$ as a subset of $\mathbb{C}^D$ in the usual way).

For the definitions of the *dimension* and *degree* of a complex zero set we refer to [12]. For real zero sets, these notions can be problematic, so we will avoid them by relying on the complex versions (except in $\mathbb{R}^2$). For a real zero set $Z_{\mathbb{R}}(f_1,\dots,f_k)$ and a choice of defining polynomials $f_1,\dots,f_k$, we define the *complex dimension* to be the dimension of $Z_{\mathbb{C}}(f_1,\dots,f_k)$. If a complex zero set has dimension one, then we call it a *complex* (*algebraic*) *curve*; note that a complex curve may contain zero-dimensional components (a complex zero set may have components of different dimension, and its dimension is the maximum). If a real zero set with a given choice of defining polynomials has complex dimension one, we call it a *real* (*algebraic*) *curve*. Note that a real curve may be a finite set. In $\mathbb{R}^2$, the zero set of a single (non-zero) polynomial always has complex dimension one, so our definitions of plane curves and real curves coincide.

The degree of a complex zero set $Z_{\mathbb{C}}(f_1,\dots,f_k)$ is bounded above by the product of the degrees of the polynomials f_i, by the generalized Bézout's inequality (see for instance [13]). By our definition, a complex curve C in $\mathbb{C}^2$ of degree d can be written as $Z_{\mathbb{C}}(f) \cup S$ for a single $f \in \mathbb{C}[x,y]$ of degree at most d, and a finite set S of size at most d. We claim that the real part $C \cap \mathbb{R}^2$ is then a real curve of degree at most $4d$. To prove this, we use that the complex polynomial f can be split into real polynomials $\mathrm{Re}\, f$ and $\mathrm{Im}\, f$ of degree d as in

$$f(x_1 + ix_2, y_1 + iy_2) = (\mathrm{Re}\, f)(x_1,x_2,y_1,y_2) + i\,(\mathrm{Im}\, f)(x_1,x_2,y_1,y_2).$$

If $C = Z_{\mathbb{C}}(f) \cup S$, then $Z_{\mathbb{C}}(f) \cap \mathbb{R}^2$ is the zero set of $(\mathrm{Re}\, f(x_1,0,y_1,0))^2 + (\mathrm{Im}\, f(x_1,0,y_1,0))^2$, and $S \cap \mathbb{R}^2$ is the zero set of at most d polynomials of the form $(x-x_0)^2 + (y-y_0)^2$. Thus $C \cap \mathbb{R}^2$ is the zero set of a polynomial of degree at most $4d$.

A complex curve in $\mathbb{C}^D$ is *irreducible* if it is not the union of two proper subsets which are curves; an *irreducible component* is a subset which is an irreducible curve; and two curves *have*

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

a common component if there is a curve which is a subset of both. The number of irreducible components of a curve is bounded above by its degree.

We also need a version of Bézout's inequality for $\mathbb{R}^D$. Over $\mathbb{C}$, there are far-reaching generalizations of Bézout's inequality, like the one from [13] mentioned above. But over $\mathbb{R}$ some such generalizations may fail. For instance, when a real zero set is finite, the size of the set need not be bounded above by the product of the degrees of the polynomials defining it: take in $\mathbb{R}^3$ the intersection of the plane $z = 0$ with the zero set of the polynomial $(x(x-1)(x-2))^2 + (y(y-1)(y-2))^2$, which is a set of 9 points, while the product of the degrees of the polynomials is 6.

To overcome this complication, one could carefully use the complex version of Bézout's inequality, but we instead rely on a bound on the number of connected components of a real zero set. The bound that we use is due to Oleinik and Petrovskii, Milnor, and Thom; for an exposition see [2, Chapter 7]. A *connected component* of an algebraic curve in $\mathbb{R}^D$ is a connected component in the Euclidean topology on $\mathbb{R}^D$. Note that this is not the same as an irreducible component; for instance, the curve $y^2 = x^3 - x$ in $\mathbb{R}^2$ has one irreducible component, but two connected components.

Theorem 2.2. *A zero set in $\mathbb{R}^D$ defined by polynomials of degree at most d has at most $(2d)^D$ connected components.*

2.2. Incidence bound

We will use an incidence bound from combinatorial geometry, due to Pach and Sharir [15, 16]; the version in Theorem 2.3 below is proved in [14].

Let $P \subset \mathbb{R}^D$ and let Γ be a set of curves $\mathbb{R}^D$. We define $I(P,\Gamma)$ to be the set of *incidences*, that is, the set of pairs $(p,\gamma) \in P \times \Gamma$ such that $p \in \gamma$. We say that P and Γ form a *system with k degrees of freedom* if there is a *multiplicity M* such that any two curves in Γ intersect in at most M points of P, and any k points of P belong to at most M curves in Γ.

Theorem 2.3 (Pach–Sharir). *If a set P of points in $\mathbb{R}^2$ and a set Γ of algebraic curves in $\mathbb{R}^2$ of degree at most d form a system with two degrees of freedom and multiplicity M, then*

$$|I(P,\Gamma)| \leqslant C_{d,M} \cdot \max\{|P|^{2/3}|\Gamma|^{2/3}, |P|, |\Gamma|\},$$

where $C_{d,M}$ is a constant depending only on d and M.

We will deduce a version for curves in higher dimensions, using a 'generic projection trick' as used by Solymosi and Tao in [22, Section 5.1].

Corollary 2.4. *If a set P of points in $\mathbb{R}^D$ and a set Γ of real algebraic curves in $\mathbb{R}^D$, each defined by e polynomials of degree at most d, form a system with two degrees of freedom and multiplicity M, then*

$$|I(P,\Gamma)| \leqslant C_{D,e,d,M} \cdot \max\{|P|^{2/3}|\Gamma|^{2/3}, |P|, |\Gamma|\},$$

where $C_{D,e,d,M}$ is a constant depending only on D, e, d and M.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

Proof. We will use a generic projection from $\mathbb{R}^D$ to $\mathbb{R}^2$ and then apply Theorem 2.3. We have to face the technical complication that the projection of a real algebraic curve need not be a real algebraic curve; for instance, a circle may be projected down to a line segment. To deal with this, we partly have to work with the complex versions of the curves. Given $\gamma \in \Gamma$ and its defining polynomials $f_1, \ldots, f_e$, we set $\gamma^{\mathbb{C}} = Z_{\mathbb{C}}(f_1, \ldots, f_e)$, which is a complex curve of degree at most d^e (here we use the fact that by our definition, the real curves in Γ have complex dimension one with respect to the given defining polynomials). We write $\Gamma^{\mathbb{C}}$ for the set $\{\gamma^{\mathbb{C}} : \gamma \in \Gamma\}$. We consider $P \subset \mathbb{R}^D$ as a subset of $\mathbb{C}^D$ in the natural way, and for any real linear transformation we consider the corresponding complex linear transformation (*i.e.*, the one given by the same real matrix); we abuse notation slightly and use the same symbol for the real transformation and the complex transformation.

Let us make explicit what we mean by 'generic'. We first apply a suitable linear transformation φ to $\mathbb{R}^D$, and then we apply the standard projection $\psi : \mathbb{R}^D \to \mathbb{R}^2$ defined by $(x_1, \ldots, x_D) \mapsto (x_1, x_2)$. We claim that φ can be chosen so that $\pi = \psi \circ \varphi$ has the two properties below. If this is true, then we will be able to apply Theorem 2.3 to the projected points and curves, which proves the required bound.

(1) The map π is bijective on P and induces a bijection from $I(P, \Gamma^{\mathbb{C}})$ to $I(\pi(P), \pi(\Gamma^{\mathbb{C}}))$.
(2) For distinct $\gamma, \gamma' \in \Gamma$, $\pi(\gamma^{\mathbb{C}})$ and $\pi(\gamma'^{\mathbb{C}})$ are distinct algebraic curves in $\mathbb{C}^2$ of degree at most d.

We show that the *complex* linear transformations φ for which π fails to have these properties lie in a lower-dimensional subvariety of the D^2-dimensional space $M_D(\mathbb{C})$ of (matrices of) linear transformations of $\mathbb{C}^D$. This will then imply that there is a *real* matrix with these properties, because $M_D(\mathbb{R})$ cannot lie within a proper subvariety of $M_D(\mathbb{C})$.[1] Thus our final map π will be a real transformation, even though we argue via complex transformations.

First we treat property (1). For $p \neq p' \in P$, the complex linear transformations for which $\pi(p) = \pi(p')$ are in a lower-dimensional subspace of $M_D(\mathbb{C})$, since the entries of their matrices satisfy a common linear equation. Taking the union of these subspaces for all pairs of distinct points in P, we still get a lower-dimensional subspace.

For the second part of property (1), note that if $(p, \gamma^{\mathbb{C}}) \in I(P, \Gamma^{\mathbb{C}})$, then obviously

$$(\pi(p), \pi(\gamma^{\mathbb{C}})) \in I(\pi(P), \pi(\Gamma^{\mathbb{C}})).$$

Furthermore, we can choose φ so that if $p \in P$ and $p \notin \gamma$, then $\pi(p) \notin \pi(\gamma)$. To see this, we note that $W = \psi^{-1}(\pi(p))$ is a linear subspace of dimension $D-2$ in $\mathbb{C}^D$. For fixed points $q \in \gamma^{\mathbb{C}}$ and $r \in W$, the $\varphi \in M_D(\mathbb{C})$ such that $\varphi(q) = r$ lie in a subspace of codimension D. Letting r vary in W gives a subspace of codimension two, and letting q vary in $\gamma^{\mathbb{C}}$ gives a subvariety of codimension one. In other words, for any φ outside this subvariety of $M_D(\mathbb{C})$ we have $\varphi(\gamma^{\mathbb{C}}) \cap W = \emptyset$, which means that $\pi(p) \notin \pi(\gamma^{\mathbb{C}})$.

For property (2), we first ensure that $\pi(\gamma^{\mathbb{C}})$ is a complex algebraic curve in $\mathbb{C}^2$. To prove that this is possible we use the Extension Theorem of [5, Chapter 3, Theorem 1.3 and Corollary 1.4], which says that if every defining polynomial of $\gamma^{\mathbb{C}}$ has cx_D^N as its leading term in x_D, then the projection onto the first $D-1$ coordinates is an algebraic curve. The $\varphi \in M_D(\mathbb{C})$ that do not

[1] This follows from the fact that the only subvariety of $\mathbb{C}^E$ that contains $\mathbb{R}^E$ is $\mathbb{C}^E$ itself.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

put the polynomials of $\gamma^{\mathbb{C}}$ into this form lie in a lower-dimensional subvariety. Repeating the same argument $D-3$ times for the projection that removes the last coordinate, each time further restricting the choice of φ, we prove the claim for the projection onto the first two coordinates. The fact that the degree of $\pi(\gamma^{\mathbb{C}})$ is at most the degree of $\gamma^{\mathbb{C}}$ follows from [13, Lemma 2].

By the same argument used for property (1), we can also ensure that distinct curves are mapped to distinct curves, thereby finishing property (2). Indeed, given distinct $\gamma^{\mathbb{C}}, \gamma'^{\mathbb{C}}$, we can choose any point $p \in \gamma^{\mathbb{C}} \backslash \gamma'^{\mathbb{C}}$, and ensure that $\pi(p) \notin \gamma'^{\mathbb{C}}$, which implies that $\pi(\gamma^{\mathbb{C}})$ and $\pi(\gamma'^{\mathbb{C}})$ are distinct; doing this for all pairs of curves from $\Gamma^{\mathbb{C}}$ ensures that no two image curves coincide.

We have established that there is a $\pi : \mathbb{R}^D \to \mathbb{R}^2$ satisfying properties (1) and (2). Let $\Gamma' = \{\gamma^{\mathbb{C}} \cap \mathbb{R}^2 : \gamma \in \Gamma\}$, which is a set of distinct real algebraic curves of degree at most $4d$ (as observed in Section 2.1). We have established bijections between $I(\pi(P), \Gamma')$, $I(\pi(P), \pi(\gamma^{\mathbb{C}}))$, $I(P, \Gamma^{\mathbb{C}})$, and $I(P, \Gamma)$, which implies that $\pi(P)$ and Γ' have two degrees of freedom. Therefore, we can apply Theorem 2.3 to them, which finishes the proof. □

2.3. Symmetries of curves

We need two elementary results about linear transformations that fix plane algebraic curves. Given a set $S \subset \mathbb{R}^2$ and a transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$, we say that T *fixes* S if $T(S) = S$. We say that a transformation T is a *symmetry* of a plane algebraic curve C if T is an isometry of $\mathbb{R}^2$ and fixes C. Recall that an isometry of $\mathbb{R}^2$ is either a rotation, a translation, or a glide reflection (a reflection combined with a translation).

In Section 4, we will use the following bound on the number of symmetries of a plane algebraic curve.

Lemma 2.5. *An irreducible plane algebraic curve of degree d has at most $4d$ symmetries, unless it is a line or a circle.*

Proof. Suppose the curve C has a translation symmetry T. Let l be a line in the direction of T that contains some point p of C. Then $l \cap C$ must contain the entire trajectory under T of p, which consists of infinitely many points on l. By Theorem 2.1, this implies that C equals l.

Suppose C has two rotation symmetries R_a, R_b with distinct centres a, b and rotation angles α, β. We claim that then C must also have a translation symmetry, hence equals a line. Indeed, consider the composition $R_b \circ R_a$, and note that a composition of two rotations is either a translation or a rotation. If $R_b \circ R_a$ is a translation, then we are done; otherwise it is a rotation R_c with a centre c distinct from a and b, and with angle $\alpha + \beta$. Similarly, $R_b^{-1} \circ R_a^{-1}$ is a rotation around a distinct centre with angle $-\alpha - \beta$. It follows that $R_b^{-1} \circ R_a^{-1} \circ R_b \circ R_a$ is a translation, because it cannot be a rotation. Indeed, it would have angle 0, so equal the identity, but it is easily checked that, for instance, it does not fix a.

Hence, if C has a rotation symmetry with centre c, then every other rotation symmetry has the same centre c. Let p be any point on C that is not c. The image of p under any rotation symmetry then lies on a circle around c, and no two rotation symmetries give the same point. By Theorem 2.1, either C equals this circle, or it intersects it in at most $2d$ points. Therefore, C is a circle or has at most $2d$ rotation symmetries.

If C had two reflection symmetries with parallel axes of symmetry, then C would have a translation symmetry, hence would be a line. If C has two reflection symmetries with axes intersecting

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

in c, then it has a rotation symmetry around c, so by the above all axes of reflection symmetries must intersect in the same point. Suppose C has k such reflection symmetries. Pick one of them and combine it with each of the k reflections; this gives k distinct rotation symmetries, including the identity. This proves that $k \leqslant 2d$.

Finally, suppose C has a glide reflection symmetry G which is not a reflection. Then $G \circ G$ is a non-trivial translation, so C must be a line.

Altogether, if C is not a line or a circle, then it has at most $2d$ rotation symmetries and at most $2d$ reflection symmetries. □

Next we consider the same question for affine transformations (linear transformations combined with translations), but only in the case where the curve is a conic.

Lemma 2.6. *Let T be an affine transformation that fixes a conic C. Then, up to a rotation or a translation, the only possibilities are the following.*

(1) *C is a hyperbola of the form $y^2 + sxy = t$, with $s,t \neq 0$, and for some real $r \neq 0$*

$$T(x,y) = \left(rx + \frac{r^2-1}{rs}y,\ \frac{1}{r}y\right) \quad \textit{or}$$
$$T(x,y) = \left(-rx - \frac{r^2-1}{rs}y,\ rsx + ry\right).$$

(2) *C is an ellipse of the form $s^2x^2 + t^2y^2 = 1$, with $s,t \neq 0$, and for some $\theta \in [0,2\pi)$*

$$T(x,y) = \left((\cos\theta)x \pm \frac{t}{s}(\sin\theta)y,\ \frac{s}{t}(\sin\theta)x \mp (\cos\theta)y\right).$$

(3) *C is a parabola of the form $y = sx^2$, with $s \neq 0$, and for some $c \in \mathbb{R}$*

$$T(x,y) = (\pm x + c, \pm 2scx + y + sc^2).$$

Proof. Suppose C is a hyperbola. After a rotation or a translation we can assume that one of the asymptotes is the x-axis, and the other asymptote goes through the origin, so the hyperbola is of the form $y^2 + sxy = t$. Applying the shear transformation $T_1(x,y) = (sx + y, y)$ turns it into a hyperbola of the form $xy = t$. Suppose $T_2(x,y) = (ax + by + c, dx + ey + f)$ fixes $xy = t$. Then the equation of the image, $t = (ax + by + c)(dx + ey + f)$, should be the same equation (or a scalar multiple, but the constant term excludes that). This gives six equations, which one can solve to get either $T_2(x,y) = (rx, y/r)$, or $T_2(x,y) = (y/r, rx)$. Then it follows that the only affine transformations fixing the original hyperbola are of the form $T_1^{-1} \circ T_2 \circ T_1$, which gives the two forms in the lemma.

We leave it to the reader to check the other two cases in detail. For C an ellipse, we similarly apply a rotation to put it in the given form, then apply an expansion $T_1(x,y) = (sx, ty)$ to make it a circle. Then we check that rotations around the origin, possibly combined with a reflection in a line through the origin, are the only affine transformations that fix a circle around the origin. For C a parabola, a rotation puts it in the given form, and then one can check directly from the equations that the two given forms are the only ones. □

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

3. Proof of Theorems 1.1 and 1.2

We focus on the proof of Theorem 1.2, since Theorem 1.1 will follow directly by noting that the proof of Theorem 1.2 allows the curves C_1 and C_2 to be identical. Note that, as observed in Section 2.1, by our definition these curves may be finite sets. In that case, the size of this set is bounded in terms of d by Theorem 2.2, and then our theorems follow trivially by choosing c_d and c'_d large enough. Thus, we can assume throughout that C_1 and C_2 are infinite.

First we define suitable sets of points and curves, and we prove several lemmas about them (one of which, Lemma 3.2, is more involved, and we defer its proof to the next section). Together these lemmas will enable us to conclude that the points and curves essentially form a system with two degrees of freedom, so that the incidence bound from Corollary 2.4 can be applied to them. This leads to an upper bound on the number of certain quadruples of points. On the other hand, a standard argument due to Elekes gives a lower bound on the same quantity, inversely proportional to the number of distinct distances. Comparing these two bounds at the end of the section, we obtain the lower bounds on the number of distinct distances stated in Theorems 1.1 and 1.2.

We have irreducible plane algebraic curves C_1 and C_2 of degree at most d, given by polynomial equations (of minimum degree)

$$C_1 : f_1(x,y) = 0, \quad C_2 : f_2(x,y) = 0.$$

We also have sets S_1 on C_1 and S_2 on C_2 with $|S_1| = m$ and $|S_2| = n$; we write $S_1 = \{p_1, \dots, p_m\}$ and $S_2 = \{q_1, \dots, q_n\}$. We allow C_1 and C_2 to be the same curve, a possibility that will be crucial to the proof of Theorem 1.1. We make the following assumptions, which will be justified later.

Assumption 3.1. *We assume that the following hold.*

(1) *Neither C_1 nor C_2 is a vertical line.*

(2) *The sets S_1 and S_2 are disjoint.*

(3) *If C_1 (resp. C_2) is a circle, then its centre is not in S_2 (resp. S_1).*

(4) *If C_1 (resp. C_2) is a circle, any concentric circle contains at most one point of S_2 (resp. S_1).*

(5) *If C_1 (resp. C_2) is a line, there is at most one point of S_2 (resp. S_1) on any union of a line parallel to C_1 (resp. C_2) with its reflection in C_1 (resp. C_2).*

(6) *If C_1 (resp. C_2) is a line, any orthogonal line contains at most one point of S_2 (resp. S_1).*

We will define a new curve C_{ij} in $\mathbb{R}^4$ for each pair of points $p_i, p_j \in S_1$, written as

$$p_i = (a_i, b_i), \quad p_j = (a_j, b_j).$$

Let q and q' be points on C_2 (not necessarily in S_2), written as

$$q = (x,y), \quad q' = (x',y').$$

We think of q and q' as varying along C_2, while p_i and p_j are kept fixed on C_1. For $1 \leqslant i, j \leqslant m$, we define C_{ij} to be the algebraic curve in $\mathbb{R}^4$ consisting of all points $(q,q') = (x,y,x',y')$ satisfying

$$\begin{gathered} f_2(x,y) = 0, \quad f_2(x',y') = 0, \\ (x-a_i)^2 + (y-b_i)^2 = (x'-a_j)^2 + (y'-b_j)^2. \end{gathered} \tag{3.1}$$

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

In Lemma 3.3 we will prove that C_{ij}, with the given defining polynomials, has complex dimension one, which implies that it is indeed a real algebraic curve (by our definition, see Section 2.1).

Let P be the set of points $(q_s,q_t)=(x_s,y_s,x_t,y_t)\in\mathbb{R}^4$ for any $q_s,q_t\in S_2$. Note that $(q_s,q_t)\in P$ lies on C_{ij} if and only if

$$d(p_i,q_s)=d(p_j,q_t),$$

so a point on C_{ij} corresponds to points $q_s,q_t\in S_2$ that are equidistant from p_i and p_j, respectively. Therefore, an incidence of C_{ij} with P corresponds to a quadruple (p_i,p_j,q_s,q_t) such that $d(p_i,q_s)=d(p_j,q_t)$. These are the quadruples that we will find upper and lower bounds for.

We let Γ be the set of curves C_{ij} for $1\leqslant i,j\leqslant m$. Some pairs of these curves may coincide as sets of points, but we will consider them as different curves, so $|\Gamma|=m^2$. We would like P and Γ to form a system with two degrees of freedom, but this is false if some pairs of curves have a common component, which would mean they have infinite intersection. This can in fact occur for certain curves C_1 and C_2, even if they are not lines or circles. To overcome this obstacle, we will analyse when exactly the curves C_{ij} can have infinite intersection, which leads to the following lemma, stating that this obstacle is relatively rare. We will defer the longer proof of this lemma to the next section and the appendix, and first use it to complete the proof of Theorems 1.2 and 1.1.

Lemma 3.2. *If C_1 and C_2 are not parallel lines, orthogonal lines, or concentric circles, then there is a subset $\Gamma_0\subset\Gamma$ with $|\Gamma_0|\leqslant 4dm$ such that no three curves in $\Gamma\backslash\Gamma_0$ have infinite intersection.*

Next we show that when two curves have finite intersection, the number of their intersection points is bounded in terms of d. This essentially follows from Bézout's inequality in $\mathbb{C}^4$, but we will deduce it from the bound in Theorem 2.2 on the number of connected components of a real zero set.

In the proof we make use of the fact that the curves C_{ij} have two defining equations in common, or in other words, they lie on a common surface. We define S to be this surface, *i.e.*, the set of $(x,y,x',y')\in\mathbb{R}^4$ for which $f_2(x,y)=0$ and $f_2(x',y')=0$. It is in fact the Cartesian product of two copies of C_2, which implies that it does indeed have complex dimension two (with respect to the given defining polynomials).

Lemma 3.3. *Each zero set C_{ij}, with the three defining polynomials given in* (3.1)*, has complex dimension one. If $|C_{ij}\cap C_{kl}|$ is finite, then $|C_{ij}\cap C_{kl}|\leqslant 16d^4$. For any curve $C_{ij}\in\Gamma\backslash\Gamma_0$, there are at most $2d^2$ curves $C_{kl}\in\Gamma\backslash\Gamma_0$ such that $|C_{ij}\cap C_{kl}|$ is infinite.*

Proof. Let $S^{\mathbb{C}}$ be the complex variety defined by the two defining polynomials of S, and let $C_{ij}^{\mathbb{C}}$ be the complex variety defined by the three polynomials in (3.1). Note that $C_{ij}^{\mathbb{C}}=S^{\mathbb{C}}\cap Z_{\mathbb{C}}(F)$ for $F=(x-a_i)^2+(y-b_i)^2-(x'-a_j)^2-(y'-b_j)^2$. To prove that $C_{ij}^{\mathbb{C}}$ has dimension one, we use the following fact (see [12, Exercise 11.6]): If X is an irreducible variety in $\mathbb{C}^n$, and F is any polynomial in $\mathbb{C}[x_1,\ldots,x_n]$ that does not vanish on all of X, then $\dim(X\cap Z_{\mathbb{C}}(F))=\dim(X)-1$. We observe that $S^{\mathbb{C}}$ is two-dimensional and irreducible because it is a product of two one-dimensional irreducible varieties (see [12, Exercise 5.9] and the remark before [12, Theorem 11.12]). Then all we have to show is that F does not vanish on all of $S^{\mathbb{C}}$. But if

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

it did, then every point q_s would be at the same distance from p_i, which is not the case by Assumption 3.1.3. This proves the first claim of the lemma.

The intersection points $(x,y,x',y') \in C_{ij} \cap C_{kl}$ satisfy the four equations

$$(x-a_i)^2+(y-b_i)^2 = (x'-a_j)^2+(y'-b_j)^2,$$
$$(x-a_k)^2+(y-b_k)^2 = (x'-a_l)^2+(y'-b_l)^2,$$
$$f_2(x,y)=0, \qquad f_2(x',y')=0,$$

each of which has degree at most d. By Theorem 2.2, it follows that $C_{ij} \cap C_{kl}$ has at most $(2d)^4$ connected components. If this intersection is finite, every point is a connected component, so the number of points is at most $16d^4$, proving the second claim.

For the last claim, observe that if $|C_{ij} \cap C_{kl}|$ is infinite, then so is $|C_{ij}^{\mathbb{C}} \cap C_{kl}^{\mathbb{C}}|$, which implies that $C_{ij}^{\mathbb{C}}$ and $C_{kl}^{\mathbb{C}}$ have a common component. No three curves $C_{ij} \in \Gamma \backslash \Gamma_0$ have infinite real intersection by Lemma 3.2, so no three of the corresponding $C_{ij}^{\mathbb{C}}$ share a component that contains infinitely many real points. Fix a curve $C_{ij} \in \Gamma \backslash \Gamma_0$. Then an irreducible component of $C_{ij}^{\mathbb{C}}$ that has infinitely many real points is shared with at most one other $C_{kl}^{\mathbb{C}}$. The curve $C_{ij}^{\mathbb{C}}$ has degree at most $2d^2$ and thus at most $2d^2$ irreducible components. It follows that at most d^2 curves $C_{kl}^{\mathbb{C}}$ share with $C_{ij}^{\mathbb{C}}$ a component that contains infinitely many real points, which implies that there are at most $2d^2$ curves C_{kl} with which C_{ij} has infinite real intersection. □

The two lemmas above let us conclude that, although P and Γ need not have two degrees of freedom, we can partition them into subsets that do. For each of these subsets we can then bound the number of incidences.

Lemma 3.4. *Let $L = 2d^2+1$. There are partitions of P into $P_0,\dots,P_L$ and Γ into $\Gamma_0,\dots,\Gamma_L$ such that $|\Gamma_0| \leqslant 4dm$ and $|P_0| \leqslant 4dn$, and such that for all $1 \leqslant \alpha,\beta \leqslant L$, the pair P_α,Γ_β forms a system with two degrees of freedom, with multiplicity $M = 16d^4$.*

Proof. Let Γ_0 be the subset given by Lemma 3.2, so $|\Gamma_0| \leqslant 4dm$. We define a graph G with vertex set $\Gamma \backslash \Gamma_0$, connecting two vertices by an edge if the corresponding curves have infinite intersection. By Lemma 3.3, a curve in $\Gamma \backslash \Gamma_0$ has infinite intersection with at most $2d^2 = L-1$ other curves, so the graph has maximum degree $L-1$. It follows that the chromatic number of G is bounded by L, which means that we can partition the vertices into L independent sets. In other words, we can partition $\Gamma \backslash \Gamma_0$ into L subsets $\Gamma_1,\dots,\Gamma_L$ so that no two curves in the same Γ_β have infinite intersection. Lemma 3.3 then implies that they intersect in at most $16d^4$ points.

To show that a bounded number of curves passes through two points, we can reverse the roles of C_1 and C_2. We let $\widetilde{C}_{st}$ be the resulting curves in $\mathbb{R}^4$, defined analogously to equation (3.1). So, given $(x_s,y_s),(x_t,y_t) \in C_2$, $\widetilde{C}_{st}$ is the set of all points (a_i,b_i,a_j,b_j) satisfying

$$f_1(a_i,b_i)=0, \quad f_1(a_j,b_j)=0,$$
$$(x_s-a_i)^2+(y_s-b_i)^2 = (x_t-a_j)^2+(y_t-b_j)^2.$$

By the statement analogous to Lemma 3.2, there is a subset $\widetilde{\Gamma}_0$ of $4dn$ of these curves $\widetilde{C}_{st}$ such that in the remainder no three curves have infinite intersection. Let P_0 be the set of points $(q_s,q_t) \in \mathbb{R}^4$ corresponding to the curves $\widetilde{C}_{st}$ in $\widetilde{\Gamma}_0$.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

We define a graph H with vertex set $P\backslash P_0$, connecting two points $(q_s,q_t),(q_{s'},q_{t'})$ if the corresponding curves $\widetilde{C}_{st}$ and $\widetilde{C}_{s't'}$ have infinite intersection. As in the case of G, we can partition $P\backslash P_0$ into subsets $P_1,\dots,P_L$ so that for any two points $(q_s,q_t),(q_{s'},q_{t'})$ in the same P_α with $\alpha \geqslant 1$, the curves $\widetilde{C}_{st}$ and $\widetilde{C}_{s't'}$ do not have infinite intersection. It follows that there are at most $16d^4$ curves from Γ passing through any two points from the same P_α.

This establishes, for all $\alpha,\beta \geqslant 1$, that P_α and Γ_β form a system with two degrees of freedom, with $M = 16d^4$. □

Applying Corollary 2.4 now gives an incidence bound for each combination of a point set and a curve set from the partitions.

Lemma 3.5. *For all $1 \leqslant \alpha,\beta \leqslant L$ we have*

$$|I(P_\alpha,\Gamma_\beta)| \leqslant A_d \cdot \max\{m^{4/3}n^{4/3}, m^2, n^2\}.$$

For the relatively few curves and points that we placed in Γ_0 and P_0, we can easily establish the following incidence bound.

Lemma 3.6. *We have*

$$|I(P,\Gamma_0)| \leqslant 8d^2mn \quad \textit{and} \quad |I(P_0,\Gamma)| \leqslant 8d^2mn.$$

Proof. Each $C_{ij} \in \Gamma_0$ has at most $2dn$ incidences with a point $(q_s,q_t) \in P$. This follows from the fact that there are n choices of $q_s \in S_2$, and for each of those, the corresponding $q_t \in S_2$ can be found by intersecting C_2 with a circle around p_j of radius $d(p_i,q_s)$. This gives at most $2d$ solutions by Theorem 2.1, unless C_2 equals that circle, which cannot happen by Assumption 3.1.3. Therefore, we have $|I(P,\Gamma_0)| \leqslant 2dn \cdot 4dm = 8d^2mn$. The second inequality follows by applying the same argument to the curves $\widetilde{C}_{st}$ defined in the proof of Lemma 3.4. □

Before finally proving the main theorems, we need the following observation about a certain set of quadruples. This observation is a key element in the 'Elekes–Sharir transformation', as introduced in [8] and used in [11, 19]. Let Q be the set of quadruples (p_i,p_j,q_s,q_t), with $1 \leqslant i,j \leqslant m$ and $1 \leqslant s,t \leqslant n$, such that $d(p_i,q_s) = d(p_j,q_t)$, and let D be the set of distances between S_1 and S_2.

Lemma 3.7. *We have*

$$|Q| \geqslant \frac{m^2n^2}{|D|}.$$

Proof. Write $E_d = \{(p,q) \in S_1 \times S_2 : |pq| = d\}$ for $d \in D$. Then we have

$$|Q| \geqslant \sum_{d\in D} |E_d|^2 \geqslant \frac{1}{|D|}\left(\sum_{d\in D} |E_d|\right)^2 = \frac{(mn)^2}{|D|}$$

by the Cauchy–Schwarz inequality. □

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

Proof of Theorem 1.2. First we establish Assumption 3.1. We rotate the coordinate axes so that neither C_1 nor C_2 is a vertical line. We make S_1 and S_2 disjoint by arbitrarily removing at most half the points of each. We remove at most two more points so that if one of C_1, C_2 is a circle, then its centre is not in the other set. For the fourth part of the assumption, if C_2 is a circle, we observe that since C_1 is not a circle concentric with C_2, S_1 can contain at most $2d$ points of any concentric circle. We remove at most $2d-1$ points from S_1 from every concentric circle, which leaves at least a $1/(2d)$ fraction of the points. We do the same for S_2. If C_1 or C_2 is a line, we do an analogous removal from every orthogonal line, and from every union of a parallel line and its reflection. This leaves at least a fraction $1/(2d^2)$ of the points, so that the fifth and sixth parts of the assumption are satisfied. Altogether these steps leave at least $m/(4d^2)$ points in S_1 and $n/(4d^2)$ in S_2. Now we redefine S_1 and S_2 to be the point sets resulting from these modifications.

Combining the bounds from Lemmas 3.5 and 3.6, we obtain

$$\begin{aligned}|I(P,\Gamma)| &\leqslant |I(P_0,\Gamma)| + |I(P,\Gamma_0)| + \sum_{\alpha,\beta\geqslant 1} |I(P_\alpha,\Gamma_\beta)| \\ &\leqslant 16d^2mn + \sum_{\alpha,\beta\geqslant 1} A_d \cdot \max\{m^{4/3}n^{4/3}, m^2, n^2\} \\ &\leqslant B_d \cdot \max\{m^{4/3}n^{4/3}, m^2, n^2\},\end{aligned}$$

for the constant $B_d = 6d^4A_d$, noting that the sum has at most $L^2 \leqslant 5d^4$ terms.

On the other hand, by our definitions, an incidence of a curve in Γ with a point in P corresponds exactly to a quadruple (p_i, p_j, q_s, q_t) satisfying $d(p_i,q_s) = d(p_j,q_t)$. Combined with Lemma 3.7, this gives

$$\frac{m^2n^2}{|D|} \leqslant |Q| = |I(P,\Gamma)| \leqslant B_d \cdot \max\{m^{4/3}n^{4/3}, m^2, n^2\}.$$

This implies $|D| \geqslant c'_d \cdot \min\{m^{2/3}n^{2/3}, m^2, n^2\}$ for the constant $c'_d = 1/(16d^4B_d)$, which also accounts for the points removed at the start of this proof. □

Proof of Theorem 1.1. We have a curve C of degree d, not containing a line or a circle, with a set S of n points on it. It has a defining polynomial of degree d, which has at most d factors, so the curve has at most d irreducible components. Then there must be a component with at least n/d points; call it C^* and set $S^* = S \cap C^*$. Applying the proof above to $C_1 = C^*$, $C_2 = C^*$ shows that S determines at least $c_dn^{4/3}$ distinct distances. □

Remark on the dependence of c_d on d. With the proof above, the constant c_d in Theorem 1.1 would come out to be $c_d = cd^{-20/3}$ for some absolute constant c. Roughly speaking, we get a factor $d^{8/3}$ from the application of Corollary 2.4 (using the more precise $C_{\delta,M} = C \cdot \delta^{2/3} \cdot M^{1/3}$ in Theorem 2.3, $M = 16d^4$, and the fact that the projected curves have degree $\delta \leqslant 2d^2$), and a factor d^4 from splitting up P and Γ in Lemma 3.4. This gives $c_d = c \cdot (d^{8/3} \cdot d^4)^{-1} = c \cdot d^{-20/3}$. For c'_d in Theorem 1.2, we would get another factor d^4, to account for the removed points when C_1 or C_2 is a circle or a line.

To improve the first factor $d^{8/3}$, we could replace Theorem 2.2 by a refined bound due to Barone and Basu [1], which takes into account the fact that the defining polynomials may have different degrees. This would replace the factor $d^{8/3}$ by d^2. Furthermore, if we could replace

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

Lemma 3.2 by a similar statement for double rather than triple intersections (which we expect to be true), it would make it unnecessary to partition P and Γ as in Lemma 3.4, removing the factor d^4. Together these two improvements would give $c_d = c \cdot d^{-2}$.

Note that, given an arbitrary set of n points in $\mathbb{R}^2$, one can pass an algebraic curve of degree roughly $\sqrt{n}$ through these points. Therefore, a constant c_d on the order of $d^{-2/3}$ would be the best one could hope for, because this would imply that n arbitrary points determine $\Omega(n)$ distances, unless many of the points lie on parallel lines or concentric circles.

4. Proof of Lemma 3.2

Our proof of Lemma 3.2 requires four further lemmas that are established in this section. They will be combined at the very end of the section to deduce Lemma 3.2.

We are going to analyse how two curves C_{ij} and C_{kl} could have infinite intersection. The most clear-cut case is when $d(p_i,p_k) = d(p_j,p_l)$, because then, by Lemma 4.1, infinite intersection implies the existence of a symmetry of C_2. This is a real possibility, as will become clear in the proof of Lemma 4.1, but it cannot happen too often, because the number of symmetries is bounded if C_2 is not a line or circle.

On the other hand, when $d(p_i,p_k) \neq d(p_j,p_l)$, we expect that C_{ij} and C_{kl} cannot have infinite intersection. However, we were only able to prove the weaker statement that no three curves C_{ij}, C_{kl}, and C_{qr} can have infinite intersection in this case, which suffices for our purposes. We prove this in Lemma 4.2 when C_2 has degree at least 3, and in Lemma 4.3 for C_2 a conic. When C_2 is a line, we prove a stronger statement in Lemma 4.4.

It is worth pointing out where we use that our curves are not parallel lines, orthogonal lines, or concentric circles. The fact that the curves are not parallel lines allows us to make Assumption 3.1.5, which lets us bound the number of translation symmetries. Similarly, not having concentric circles allows Assumption 3.1.4, letting us bound the number of rotation symmetries. Together, these observations imply that an infinite intersection like in Lemma 4.1 cannot occur too often. Finally, because our curves are not orthogonal lines, we can make Assumption 3.1.6, which, together with Assumption 3.1.5, lets us exclude large intersections in Lemma 4.4.

Lemma 4.1. *If $d(p_i,p_k) = d(p_j,p_l)$ and C_{ij} and C_{kl} have infinite intersection, then C_2 has a symmetry that maps p_i to p_j and p_k to p_l.*

Proof. A point $(x,y,x',y') = (q,q') \in C_{ij} \cap C_{kl}$ corresponds to a pair of points $q,q' \in C_2$ such that

$$d(p_i,q) = d(p_j,q') \quad \text{and} \quad d(p_k,q) = d(p_l,q').$$

It follows that

$$\begin{aligned}&\{(d_1,d_2) : \exists (q,q') \in C_{ij} \cap C_{kl} \text{ such that } d_1 = d(p_i,q), d_2 = d(p_k,q)\}\\ &= \{(d_1,d_2) : \exists (q,q') \in C_{ij} \cap C_{kl} \text{ such that } d_1 = d(p_j,q'), d_2 = d(p_l,q')\}.\end{aligned}$$

Call this set of pairs of distances D. Since C_{ij} and C_{kl} have infinite intersection, D must be an infinite set.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

The idea of the proof is to reconstruct the points of C_2 using the distance pairs from D, by finding the points that respectively have those distances from the points p_i and p_k. The resulting set of points should consist of an infinite subset of C_2, together with its reflection in the line p_ip_k. The image of this set under the rotation that maps p_i, p_k to p_j, p_l should again have infinite intersection with C_2, because C_2 should have points at the same distance pairs from p_j, p_l. We will see that this implies that C_2 has a symmetry.

To make this more precise, let E be the set of all points that arise in this way from a pair of distances in D:

$$E = \{q \in \mathbb{R}^2 : (d(p_i,q), d(p_k,q)) \in D\}.$$

Let M be the reflection in the line p_ip_k. Set $E_1 = E \cap C_2$ and $E_2 = M(E_1)$; because D is infinite, so are E_1 and E_2. Let T be the rotation that maps p_i, p_k to p_j, p_l, if it exists; otherwise there is a translation or glide reflection that maps p_i, p_k to p_j, p_l, and we call that T. Then T must place an infinite subset of E onto C_2; call this subset E_1^*, and set $E_2^* = M(E_1^*)$. We distinguish two cases.

(1) If $|E_1 \cap E_1^*|$ is infinite, then $F_1 = E_1 \cap E_1^*$ is an infinite subset of C_2 such that $T(F_1) \subset C_2$.

(2) If $|E_1 \cap E_1^*|$ is not infinite, then $F_2 = |E_1 \cap E_2^*|$ must be infinite. Then $F_2 = E_1 \cap E_2^*$ is an infinite subset of C_2 such that $(T \circ M)(F_2) \subset C_2$, since M maps F_2 into E_1^*, which T maps into C_2.

In each case we use the following observation to deduce that C_2 has a symmetry. If we have an isometry T of the plane and an infinite subset A of an irreducible algebraic curve C such that $T(A) \subset C$, then $T(C) = C$, that is, T is a symmetry of C. This holds because $T(C)$ is also an irreducible plane algebraic curve, so by Theorem 2.1 it either has finite intersection with C, or equals it.

If T is a rotation, then in case (1) C_2 has the rotation symmetry T, while in case (2) it has the reflection symmetry $T \circ M$. If T is a translation, then in case (1) it has the translation symmetry T, while in case (2) it has the glide reflection symmetry $T \circ M$. □

Lemma 4.2. *Suppose we have $p_i, p_j, p_k, p_l, p_q, p_r \in S_1$ satisfying*

$$d(p_i,p_k) \neq d(p_j,p_l), \quad d(p_i,p_q) \neq d(p_j,p_r), \quad \textit{and} \quad d(p_k,p_q) \neq d(p_l,p_r).$$

Then

$$|C_{ij} \cap C_{kl} \cap C_{qr}| \leqslant 2d,$$

unless C_2 is a conic or a line.

Proof. A point in $|C_{ij} \cap C_{kl} \cap C_{qr}|$ corresponds to two points $q_s, q_t \in C_2$ such that the distances from p_i, p_k, p_q to q_s are equal to those from p_j, p_l, p_r to q_t. We will show that the set of such points q_t (or q_t) is forced to lie on a conic or a line, so by Theorem 2.1 C_2 contains at most $2d$ of them, unless C_2 is a conic or a line.

We can assume after a rotation that $p_i = (0,0)$ and $p_k = (1,0)$; we think of these points as lying in the xy-plane, and write $p_q = (a,b)$. We think of the points p_j, p_l as lying in a separate uv-plane, and there we can assume after a rotation that $p_j = (0,0)$ and $p_l = (L,0)$, with $L \neq 0,1$, and $p_r = (c,d)$.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

Consider the points (x,y) that have distances d_1,d_2,d_3 from respectively p_i,p_k,p_q in the xy-plane, and the points (u,v) that have the same distances from p_j,p_l,p_r in the uv-plane. Then we have the equations

$$x^2+y^2=u^2+v^2, \tag{4.1}$$

$$(x-1)^2+y^2=(u-L)^2+v^2, \tag{4.2}$$

$$(x-a)^2+(y-b)^2=(u-c)^2+(v-d)^2. \tag{4.3}$$

Subtracting (4.1) from (4.2) gives

$$x=Lu+\frac{1}{2}(1-L^2). \tag{4.4}$$

Subtracting (4.1) from (4.3), and plugging (4.4) into the result, leads to

$$by=(c-aL)u+dv+\frac{1}{2}(a^2+b^2-c^2-d^2+aL^2-a). \tag{4.5}$$

Plugging the linear equations (4.4) and (4.5) into (4.1) leads to

$$(b^2L^2+(c-aL)^2-b^2)u^2+(d^2-b^2)v^2+2d(c-aL)uv+l(u,v)=0, \tag{4.6}$$

where $l(u,v)$ is a linear function of u and v. If this equation is not identically zero, this shows that (u,v) must lie on a conic or a line, which implies that the original point q_t lies on a conic or line. Since C_2 is irreducible and not a conic or line, it follows that there are at most $2d$ such q_t.

This leaves us with the case where (4.6) holds identically. In that case we can see from the coefficients of the quadratic terms that $b=d=0$ and $c=aL$. Plugging these back into (4.3) easily leads to a contradiction. □

Lemma 4.3. *Suppose we have $p_i,p_j,p_k,p_l,p_q,p_r\in S_1$ satisfying*

$$d(p_i,p_k)\neq d(p_j,p_l),\quad d(p_i,p_q)\neq d(p_j,p_r),\quad \text{and}\quad d(p_k,p_q)\neq d(p_l,p_r).$$

If C_2 is a conic then

$$|C_{ij}\cap C_{kl}\cap C_{qr}|\leqslant 4.$$

Proof. Suppose that $|C_{ij}\cap C_{kl}\cap C_{qr}|\geqslant 5$. Then, similarly to the previous proof, we have three equations of the form

$$(x-a_\alpha)^2+(y-b_\alpha)^2=(u-c_\alpha)^2+(v-d_\alpha)^2, \tag{4.7}$$

satisfied by at least five pairs of points $(x,y),(u,v)$ on C_2. Subtracting the first equation from the second and third gives two linear equations, which we can view as an affine transformation T sending (u,v) to (x,y). Because T sends five points on C_2 to five points on C_2, it must fix C_2, since the image of C_2 must be a conic, which could only intersect C_2 four times if it was a different conic. Lemma 2.6 then tells us which forms T could have. We will show that in each case we get a contradiction.

Suppose that C_2 is a hyperbola. We can apply a rotation to make it of the form $y^2+sxy=t$ (note that the rotation moves the points (a_α,b_α) and (c_α,d_α), but does not change the form of the equations, or the condition of the lemma). By Lemma 2.6, T must have the form $(u,v)=$

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

$T(x,y) = (rx + (r^2-1)y/r, y/r)$ (or the second form, which we will leave to the reader). Plugging this into (4.7) gives

$$(x-a_\alpha)^2 + (y-b_\alpha)^2 = \left(rx + \frac{r^2-1}{r}y - c_\alpha\right)^2 + \left(\frac{1}{r}y - d_\alpha\right)^2.$$

This equation has a term x^2 with coefficient r^2-1. If $r \neq \pm 1$, then this equation describes a different hyperbola than C_2, so cannot be satisfied by more than four points of C_2. If $r=1$, then T is the identity, which would mean that we can put $u=x, v=y$ in (4.7). That would lead to $a_\alpha = c_\alpha, b_\alpha = d_\alpha$ for each α, contradicting the assumption of the lemma on the distances between the points. Finally, if $r=-1$, we could similarly put $u=-x, v=-y$, leading to $a_\alpha = -c_\alpha, b_\alpha = -d_\alpha$ for each α, contradicting the same assumption.

Suppose now that C_2 is an ellipse; without loss of generality we can assume that it is of the form $s^2x^2 + t^2y^2 = 1$. By Lemma 2.6, T must have the form

$$(u,v) = T(x,y) = \left((\cos\theta)x \pm \frac{t}{s}(\sin\theta)y, \frac{s}{t}(\sin\theta)x \mp (\cos\theta)y\right).$$

Plugging this into (4.7) gives

$$(x-a_\alpha)^2 + (y-b_\alpha)^2 = \left((\cos\theta)x \pm \frac{t}{s}(\sin\theta)y - c_\alpha\right)^2 + \left(\frac{s}{t}(\sin\theta)x \mp (\cos\theta)y - d_\alpha\right)^2,$$

which rearranges to

$$\left(\frac{s^2}{t^2}\sin^2\theta + \cos^2\theta - 1\right)\cdot x^2 + \left(\frac{t^2}{s^2}\sin^2\theta + \cos^2\theta - 1\right)\cdot y^2$$
$$\pm 2\sin\theta\cos\theta\left(\frac{t}{s} - \frac{s}{t}\right)\cdot xy + l(x,y) = 0.$$

For this to be an ellipse, the coefficient of xy must be zero, so $(t/s - s/t)\sin\theta\cos\theta = 0$. If $\cos\theta = 0$, then the equation takes the form $(s^2/t^2 - 1)x^2 + (t^2/s^2 - 1)y^2 + l(x,y) = 0$. Unless $s = \pm t$ (a case we will consider separately), the x^2 and y^2 terms have opposite signs, so this cannot be the equation of an ellipse. If $\sin\theta = 0$, then T is the identity, which leads to a contradiction as in the hyperbola case. It follows that we must have $s = \pm t$. This implies that the coefficients of x^2 and y^2 are also zero, so in fact the polynomial must vanish identically. The coefficients of the linear terms then give, after some rearranging, that for each α

$$a_\alpha = (\cos\theta)c_\alpha + (\sin\theta)d_\alpha, \quad b_\alpha = \pm(\sin\theta)c_\alpha \mp (\cos\theta)d_\alpha.$$

This says exactly that each (a_α, b_α) is the image of (c_α, d_α) under a rotation, or a rotation and a reflection. Both are isometries, so the distances between the points are preserved, again contradicting the assumption of the lemma.

Finally, if C_2 is parabola $y = cx^2$ and $T(x,y) = (\pm x + c, \pm 2scx + y + sc^2)$, we get

$$(x-a_\alpha)^2 + (y-b_\alpha)^2 = (\pm x + c - c_\alpha)^2 + (\pm 2scx + y + sc^2 - d_\alpha)^2.$$

This equation has an xy term with coefficient $\pm 4sc$, which implies $c = 0$, leaving only $T(x,y) = (-x,y)$. This is an isometry, which again contradicts the assumption of the lemma. □

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

The remaining case, where C_2 is a line, is considerably easier. The proof is reminiscent of the proofs in [19] and [21].

Lemma 4.4. *Suppose that $d(p_i,p_k) \neq d(p_j,p_l)$. If C_2 is a line, then*

$$|C_{ij} \cap C_{kl}| \leqslant 4.$$

Proof. We can assume within the proof of this lemma that C_2 is the x-axis. Then $C_2 \times C_2 \subset \mathbb{R}^4$ is the plane consisting of points of the form $(x,0,x',0)$, which we can think of as the xx'-plane. In that plane, C_{ij} is the curve defined by the equation

$$(x-a_i)^2 - (x'-a_j)^2 = b_j^2 - b_i^2.$$

Because $d(p_i,p_k) \neq d(p_j,p_l)$, we do not have both $i=j$ and $k=l$, and by Assumption 3.1.5, we have $b_j^2 - b_i^2 \neq 0$ when $i \neq j$ and $b_l^2 - b_k^2 \neq 0$ when $k \neq l$. This implies that at least one of C_{ij} and C_{kl} is a non-degenerate hyperbola. By Assumption 3.1.6, we have $a_i \neq a_k$, which implies that C_{ij} and C_{kl} are distinct. It follows that they intersect in at most four points. □

Finally, we put together the four lemmas in this section to obtain Lemma 3.2.

Proof of Lemma 3.2. If there is a symmetry T of C_2 that maps p_i to p_j, we will say (just within this proof) that T *respects* C_{ij}. Suppose that the curves C_{ij} and C_{kl} have infinite intersection and $d(p_i,p_k) = d(p_j,p_l)$. By Lemma 4.1, there is a symmetry T of C_2 that respects C_{ij} and C_{kl}.

If C_2 is not a line or a circle, it has at most $4d$ symmetries, by Lemma 2.5. Given a fixed symmetry T, each p_i is sent to a unique point p_j, so T respects at most m curves C_{ij}. Therefore, there are in total at most $4dm$ curves C_{ij} that are respected by some symmetry. We let Γ_0, the set to be excluded, contain all curves C_{ij} that are respected by some symmetry of C_2. Then $|\Gamma_0| \leqslant 4dm$.

If C_2 is a line or a circle, it does have many symmetries, but by Assumption 3.1, there are no $p_i, p_j \in S_1$ such that such a symmetry maps p_i to p_j as in Lemma 4.1, so we can take Γ_0 to be the empty set. Indeed, suppose C_2 is a circle and T is a symmetry of C_2 with $T(p_i) = p_j$. If $p_i \neq p_j$, then they would have to lie on a concentric circle (see the proof of Lemma 2.5), which is excluded by Assumption 3.1.4. If $p_i = p_j$, then C_2 would have to be the circle around p_i, contradicting Assumption 3.1.3. A similar argument applies if C_2 is a line, using Assumption 3.1.5.

With Γ_0 chosen as above, it follows that if $C_{ij}, C_{kl} \in \Gamma \backslash \Gamma_0$ have infinite intersection, then $d(p_i,p_k) \neq d(p_j,p_l)$. Then Lemmas 4.2, 4.3, and 4.4 (together with Assumptions 3.1.5 and 3.1.6) allow us to conclude that there are no three curves in $\Gamma \backslash \Gamma_0$ that have infinite intersection.

Applying the argument above to the dual curves $\widetilde{C}_{st}$ (defined in the proof of Lemma 3.3) gives the set P_0. This finishes the proof of Lemma 3.2. □

Acknowledgements

The authors would like to thank Filip Morić for some interesting discussions that led to this project, and Natan Rubin, Adam Sheffer, Joshua Zahl and an anonymous referee for many helpful comments.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press

References

[1] Barone, S. and Basu, S. (2016) On a real analogue of Bezout inequality and the number of connected components of sign conditions. *Proc. London Math. Soc.* **112** 115–145.

[2] Basu, S., Pollack, P. and Roy, M.-F. (2006) *Algorithms in Real Algebraic Geometry*, Springer.

[3] Brass, P., Moser, W. and Pach, J. (2005) *Research Problems in Discrete Geometry*, Springer.

[4] Charalambides, M. (2014) Exponent gaps on curves via rigidity. *Discrete Comput. Geom.* **51** 666–701.

[5] Cox, D., Little, J. and O'Shea, D. (2007) *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Springer.

[6] Elekes, G. (1999) A note on the number of distinct distances. *Period. Math. Hung.* **38** 173–301.

[7] Elekes, G. and Rónyai, L. (2000) A combinatorial problem on polynomials and rational functions. *J. Combin. Theory Ser. A* **89** 1–20.

[8] Elekes, G and Sharir, M. (2011) Incidences in three dimensions and distinct distances in the plane. *Combin. Probab. Comput.* **20** 571–608.

[9] Erdős, P. (1946) On sets of distances of *n* points. *Amer. Math. Monthly* **53** 248–250.

[10] Gibson, C. (1998) *Elementary Geometry of Algebraic Curves*, Cambridge University Press.

[11] Guth, L. and Katz, N. H. (2015) On the Erdős distinct distances problem in the plane. *Ann. of Math.* **181** 155–190.

[12] Harris, J. (1992) *Algebraic Geometry: A First Course*, Springer.

[13] Heintz, J. (1983) Definability and fast quantifier elimination in algebraically closed fields. *Theoret. Comput. Sci.* **24** 239–277.

[14] Kaplan, H., Matoušek, J. and Sharir, M. (2012) Simple proofs of classical theorems in discrete geometry via the Guth–Katz polynomial partitioning technique. *Discrete Comput. Geom.* **48** 499–517.

[15] Pach, J. and Sharir, M. (1992) Repeated angles in the plane and related problems. *J. Combin. Theory Ser. A* **59** 12–22.

[16] Pach, J. and Sharir, M. (1998) On the number of incidences between points and curves. *Combin. Probab. Comput.* **7** 121–127.

[17] Raz, O. E., Sharir, M. and Solymosi, J. (2014) Polynomials vanishing on grids: The Elekes–Rónyai problem revisited. In *SOCG'14: Proceedings of the 30th Annual Symposium on Computational Geometry*, pp. 198–205.

[18] Schwartz, R., Solymosi, J. and de Zeeuw, F. (2013) Extensions of a result of Elekes and Rónyai. *J. Combin. Theory Ser. A* **120** 1695–1713.

[19] Sharir, M., Sheffer, A. and Solymosi, J. (2013) Distinct distances on two lines. *J. Combin. Theory Ser. A* **120** 1732–1736.

[20] Sharir, M. and Solymosi, J. (2016) Distinct distances from three points. *Combin. Probab. Comput.* **25** 623–632.

[21] Sheffer, A., Zahl, J. and de Zeeuw, F. (2014) Few distinct distances implies no heavy lines or circles. *Combinatorica*, to appear. `doi:10.1007/s00493-014-3180-6`

[22] Solymosi, J. and Tao, T. (2012) An incidence theorem in higher dimensions. *Discrete Comput. Geom.* **48** 255–280.

https://doi.org/10.1017/S0963548316000225 Published online by Cambridge University Press