Distinct distances on algebraic curves in the plane

János Pach[*] Frank de Zeeuw[†]

October 11, 2013

Abstract

Let P 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 P 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 [16].

1 Introduction

A famous conjecture of Erdős, first mentioned in [8], 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 [10], Guth and Katz established an almost complete solution, proving a 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 distinct distances occurring 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 [6], where they derived it from a more general result about polynomials, which they proved using a combination of combinatorial and algebraic methods. In [5], Elekes specialized these methods to Purdy's question, resulting in a lower bound of $\Omega(n^{5/4})$ on the number of distinct distances, if the two lines are not parallel or orthogonal. Recently, Sharir, Sheffer, and Solymosi improved this bound to $\Omega(n^{4/3})$ in [16], again using algebraic methods. In [14], 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 [16], 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 [16] from lines to arbitrary plane algebraic curves (see Section 2 for definitions). The results take several forms; perhaps the most interesting of them is the following.

[*] EPFL, Lausanne and Rényi Institute, Budapest. 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 and 20021-137574. Email: `pach@cims.nyu.edu`

[†] EPFL, Lausanne. Supported by Swiss National Science Foundation Grant 200021-513047. Email: `fdezeeuw@gmail.com`

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^{-19} for an absolute constant c. We have not tried to optimize it, but in a remark at the end of Section 3 we will suggest some improvements.

Theorem 1.1 is a simple 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 $c'_d \cdot \min\left\{m^{2/3}n^{2/3}, m^2, n^2\right\}$, 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: They should not both contain a line, such that the two lines are parallel or they are orthogonal, and they should not both contain a circle, such that the two circles are concentric.

While this manuscript was being finalized, several related preprints were posted on the arXiv. 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 [5] with analytic as well as algebraic tools, and even extends it to higher dimensions, with a more complicated set of exceptions. In [15], Sharir and Solymosi showed, using a method based on that of [16], that between three non-collinear points and n other points there are $\Omega(n^{6/11})$ distinct distances. In [17], Sheffer, Zahl, and De Zeeuw extend the method of [16] to the case where one set of points in $\mathbb{R}^2$ is constrained to a line, while the other is unconstrained.

We say a few words about our proofs compared to those of the similar results mentioned above. Both [5] and [16] derive their bound by constructing a set of new curves and applying an incidence bound to them. Their construction of these curves 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 define the new curves in an implicit and algebraic way (see in particular (1) on page 7).

To apply the incidence bound to the new curves, one needs to show that the curves have small intersection, and in particular that they are distinct. In [5] and [16], this was relatively easy because the curves had low degree. In [4], it was done using concepts from the theory of structural rigidity. We do this by observing that if some of the new curves have large intersection, this must be due to some kind of symmetry of the original curve. The only curves that have too much symmetry are lines and circles.

In Section 2, we introduce the incidence bound that we will use, we define algebraic curves, and we state several results from algebraic geometry. 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

In this section we provide the necessary background for the two main tools that we use, namely, an incidence theorem for points and curves with two degrees of freedom, and Bézout's inequality. We also prove two simple results about linear transformations that fix algebraic curves. The first-time reader is advised to skip to the next section, and refer back here when needed.

One of our main tools will be an incidence bound from combinatorial geometry, due to Pach and Sharir [12, 13]. We will use a version stated in [18].

Let $P \subset \mathbb{R}^D$ and let Γ be a set of subsets of $\mathbb{R}^D$. We define $I(P, \Gamma)$ to be the number of *incidences*, i.e., the number 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 number M (the *multiplicity* of the system) such that:

(1) any two sets in Γ intersect in at most M points of $\mathbb{R}^D$;

(2) any k points of P belong to at most M sets in Γ.

Note that our definition is slightly different from the usual one, since the condition that k points belong to at most M curves is only required to hold for points in P. We will use this definition both for algebraic curves and for continuous curves. We have stated it for $\mathbb{R}^D$, because we will also apply this term to curves in four dimensions, even though we will only use the following incidence bound in the plane (and only in the special case $k = 2$).

Theorem 2.1 (Pach-Sharir)**.** *Suppose a set of points $P \subset \mathbb{R}^2$ and a set of simple continuous curves Γ form a system with 2 degrees of freedom and multiplicity M. Then*

$$I(P, \Gamma) \leq C \cdot \max\{M^{2/3}|P|^{2/3}|\Gamma|^{2/3}, M|P|, |\Gamma|\},$$

where C is an absolute constant.

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

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

We define the *degree* of C to be the degree of a minimum-degree polynomial f such that $C = Z_{\mathbb{R}}(f)$. If a curve has degree 2, we call it a *conic*.

We say that a plane algebraic curve $C = Z_{\mathbb{R}}(f)$ is *irreducible* if the polynomial $f \in \mathbb{R}[x, y]$ 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 nonconstant $h \in \mathbb{R}[x, y]$ 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 nonconstant polynomial $h \in \mathbb{R}[x, y]$ that divides f and g; it then follows that $Z_{\mathbb{R}}(h) \subset Z_{\mathbb{R}}(f) \cap Z_{\mathbb{R}}(g)$.

Note that our definition of algebraic curve does not allow a *finite* set like $Z_{\mathbb{R}}((x(x-1))^2+y^2)$. Also note that, for a real polynomial with infinite zero set, irreducibility over $\mathbb{R}$ is equivalent to irreducibility over $\mathbb{C}$. Indeed, suppose $f \in \mathbb{R}[x, y]$ is irreducible over $\mathbb{R}$ but reducible over $\mathbb{C}$, so it has a factor $h_1 + ih_2$ with nonzero real polynomials h_1, h_2. Then f also has the factor $h_1 - ih_2$, hence it has the real factor $h_1^2 + h_2^2$, so in fact $f = c \cdot (h_1^2 + h_2^2)$ for some $c \in \mathbb{R}$. But then $Z_{\mathbb{R}}(f) = Z_{\mathbb{R}}(h_1) \cap Z_{\mathbb{R}}(h_2)$, which is finite by Theorem 2.2 below, contradicting our assumption.

We will 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 multiplicities of intersection points and works in the complex projective plane, but for us the inequality suffices. See [9], Lemma 14.4, for exactly this statement, or [11], Exercise 13.17, for the complex version.

Theorem 2.2 (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 will involve curves in higher dimensions. Specifically, we will encounter curves that are zero sets in $\mathbb{R}^4$ of three polynomials. To analyze these real curves, we will also consider the complex curves defined by the same equations.

Given polynomials $f_1, \dots, f_k \in \mathbb{R}[x_1, \dots, x_D]$, we define

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

For the definition of the *dimension* of a complex zero set we refer to Lecture 11 of [11], and for the definition of its *degree*, see Lecture 18. For a real zero set $Z_{\mathbb{R}}(f_1, \dots, f_k)$, we define its *complex dimension* to be the dimension of $Z_{\mathbb{C}}(f_1, \dots, f_k)$. If a complex zero set has dimension 1, then we will call it a *complex (algebraic) curve*. If a real zero set has complex dimension 1, we will call it a *real (algebraic) curve*. Note that with this definition a real curve could be zero-dimensional in the manifold sense (for instance $Z_{\mathbb{R}}(x^2 + y^2)$), but this will not be a problem in our theorems. In fact, we will only consider the complex dimension of real zero sets.

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 a common component* if there is a curve which is a subset of both.

We will need a statement in higher dimensions that is similar to Bézout's inequality. Over $\mathbb{C}$, there are far-reaching generalizations of Bézout's inequality, stating for instance that if the intersection of two varieties without a common component is finite, then its size is at most the product of the degrees of the varieties. But over $\mathbb{R}$ some such generalizations may fail: Take for instance in $\mathbb{R}^3$ the intersection of the plane $z = 0$ with the zero set of $(x(x-1)(x-2))^2 + (y(y-1)(y-2))^2$, 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 consider the corresponding complex zero sets, but we will instead rely on the following bound on the number of connected components of a real zero set, which we also need for other purposes. 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 2.3. *A zero set in $\mathbb{R}^D$ defined by polynomials of degree at most d has at most $(2d)^D$ connected components.*

This theorem is due to Oleinik-Petrovski, Milnor, and Thom. For an exposition see [19] or [2], Chapter 7.

We will also need bounds on the number of singularities and irreducible components of an algebraic curve. We define a *singularity* of a plane algebraic curve $C = Z_{\mathbb{R}}(f)$ or $C = Z_{\mathbb{C}}(f)$ to be a point $(a, b) \in C$ such that $\frac{\partial f}{\partial x}(a, b) = \frac{\partial f}{\partial y}(a, b) = 0$. For a definition of singularities in higher dimensions, see [11], Lecture 14.

Theorem 2.4. *An algebraic curve in $\mathbb{R}^2$ or $\mathbb{C}^2$ of degree d has at most d^2 singularities.*

We have stated the bound in a form that is simple but not tight. The bound in $\mathbb{R}^2$ follows from the same bound in $\mathbb{C}^2$, which can be found in [11], Lecture 20.

The following bound on the number of irreducible components of a complex curve is proved in [19].

Theorem 2.5. *A zero set in $\mathbb{C}^D$ defined by polynomials of degree at most d has at most d^D irreducible components.*

Finally, we need two simple 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 followed by a translation).

In Section 4, we will make use of the following bound on the number of symmetries of a plane algebraic curve. Note that we count the identity as a symmetry. As can be seen from the proof, the bound is certainly not tight, but it suffices for our purposes.

Lemma 2.6. *An irreducible plane algebraic curve C of degree d can have at most $5d$ symmetries, unless it is a line or a circle.*

Proof. Suppose 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 2.2, this implies that C equals l.

Suppose C has two rotation symmetries R_a, R_b with distinct centers 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 center 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 center 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 is not a line and has a rotation symmetry with center c, then every other rotation symmetry has the same center 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 2.2, 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 has 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 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 $k - 1$ others; this will give $k - 1$ distinct rotation symmetries, proving that $k \leq 2d + 1$.

Finally, suppose C has a glide reflection symmetry G which is not a reflection. Then $G \circ G$ is a nontrivial 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 $2d + 1$ reflection symmetries, which together with the identity give $4d + 2 \leq 5d$ symmetries. □

Next we consider affine transformations and conics. By an *affine transformation* of the plane we mean any transformation of the form $T(x, y) = (ax+by+c, dx+ey+f)$ with $a, b, c, d, e, f \in \mathbb{R}$, i.e., a linear transformation followed by a translation.

The following lemma describes which irreducible conics are fixed by an affine transformation. We see that unlike for symmetries of conics other than circles, there can be infinitely many affine transformations that fix the curve. These can be viewed as "rotations" along curves other than circles. We state the cases in a somewhat technical form that is convenient for our purposes in Section 4.

Lemma 2.7. *Let T be an affine transformation that fixes an irreducible 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 $r \neq 0$

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

(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 wil 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. □

3 Proof of Theorems 1.1 and 1.2

The idea of both proofs is the following. 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 Theorem 2.1 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)

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

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 will 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 center is not in S_2 (resp. S_1);*
(4) *If C_1 (resp. C_2) is a circle, a concentric circle contains at most one point of S_2 (resp. S_1);*
(5) *If C_1 (resp. C_2) is a line, any parallel line contains at most one point of S_2 (resp. S_1);*
(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_s and q_t be points on C_2 (not necessarily in S_2), written as

$$q_s = (x_s, y_s), \quad q_t = (x_t, y_t).$$

We think of q_s and q_t as varying along C_2, while p_i and p_j are kept fixed on C_1. Let P be the set of points $(x_s, y_s, x_t, y_t) \in \mathbb{R}^4$ for $1 \le s, t \le n$.

For $1 \le i, j \le m$, we define C_{ij} to be the algebraic curve in $\mathbb{R}^4$ consisting of all points (x_s, y_s, x_t, y_t) satisfying

$$f_2(x_s, y_s) = 0, \quad f_2(x_t, y_t) = 0, \quad (x_s - a_i)^2 + (y_s - b_i)^2 = (x_t - a_j)^2 + (y_t - b_j)^2. \qquad (1)$$

In Lemma 3.3 we will prove that C_{ij} has complex dimension 1, which implies that it is indeed a real algebraic curve (by our definition).

Note that $(q_s, q_t) = (x_s, y_s, x_t, y_t)$ 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 and q_t on C_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 $d(p_i, q_s) = d(p_j, q_t)$.

We let Γ be the set of curves C_{ij} for $1 \le i, j \le m$. 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. To overcome this obstacle, we will analyze 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 relatively long proof of this lemma to the next section, and first 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 Γ_0 of Γ of at most $5dm$ curves C_{ij} 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 2.3 on the number of connected components.

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 $(x_s, y_s, x_t, y_t) \in \mathbb{R}^4$ for which $f_2(x_s, y_s) = 0$ and $f_2(x_t, y_t) = 0$. It is in fact the Cartesian product of two copies of C_2, which implies that it is indeed two-dimensional.

Lemma 3.3. *Each curve C_{ij} has complex dimension 1 and at most $4d^4$ singularities. If two curves C_{ij} and C_{kl} have finite intersection, then $|C_{ij} \cap C_{kl}| \le 16d^4$. For any curve $C_{ij} \in \Gamma\backslash\Gamma_0$, there are at most d^4 curves $C_{kl} \in \Gamma\backslash\Gamma_0$ such that $|C_{ij} \cap C_{kl}|$ is infinite.*

Proof. Let $C_{ij}^{\mathbb{C}}$ (resp. $S^{\mathbb{C}}$) be the complex zero set defined by the same equations as C_{ij} (resp. S). Note that $C_{ij}^{\mathbb{C}} = S^{\mathbb{C}} \cap Z_{\mathbb{C}}(F)$ for $F = (x_s - a_i)^2 + (y_s - b_i)^2 - (x_t - a_j)^2 - (y_t - b_j)^2$. To prove that $C_{ij}^{\mathbb{C}}$ has dimension 1, we will use the following fact (see [11], 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 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 [11], Exercise 5.9 and the remark before Theorem 11.12). Then all we have to show is that F does not vanish on all of $S^{\mathbb{C}}$. But this would imply that every point q_s is at the same distance from p_i, which is not the case by Assumption 3.1.3.

Next we consider singularities. The degree of $C_{ij}^{\mathbb{C}}$ is at most $2d^2$, which follows from Bézout's inequality in higher dimensions (specifically, Corollary 18.5 in [11], using the fact that C_{ij} has dimension 1). We apply a projection $\pi : \mathbb{C}^4 \to \mathbb{C}^2$ to $C_{ij}^{\mathbb{C}}$, chosen "generically" in the sense that the image of every singularity of $C_{ij}^{\mathbb{C}}$ is a singularity of $\pi(C_{ij}^{\mathbb{C}})$. Then $\pi(C_{ij}^{\mathbb{C}})$ is an algebraic curve in $\mathbb{C}^2$ of degree at most $2d^2$, and $\pi(C_{ij}^{\mathbb{C}})$ has at least as many singularities as $C_{ij}^{\mathbb{C}}$. By Theorem 2.4, $\pi(C_{ij}^{\mathbb{C}})$ has at most $4d^4$ singularities, so the same bound holds for $C_{ij}^{\mathbb{C}}$. Since every singularity of C_{ij} is also a singularity of $C_{ij}^{\mathbb{C}}$, this proves the second claim of the lemma.

The intersection points $(x_s, y_s, x_t, y_t) \in C_{ij} \cap C_{kl}$ satisfy the four equations

$$\begin{aligned} (x_s - a_i)^2 + (y_s - b_i)^2 &= (x_t - a_j)^2 + (y_t - b_j)^2 = 0, \\ (x_s - a_k)^2 + (y_s - b_k)^2 &= (x_t - a_l)^2 + (y_t - b_l)^2 = 0, \\ f_2(x_s, y_s) = 0, \qquad & \qquad f_2(x_t, y_t) = 0, \end{aligned}$$

which have degree at most d. By Theorem 2.3, 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 third 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 the corresponding $C_{ij}^{\mathbb{C}}$ do not share a component with infinitely many real points. Fix a curve $C_{ij} \in \Gamma \backslash \Gamma_0$. By Theorem 2.5, $C_{ij}^{\mathbb{C}}$ has at most d^4 irreducible components. It follows that at most d^4 curves $C_{kl}^{\mathbb{C}}$ share with $C_{ij}^{\mathbb{C}}$ a component with infinitely many real points, which implies that there are at most d^4 curves C_{kl} with which C_{ij} has infinite 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 $L = d^4 + 1$. There are partitions of P into $P_0, \ldots, P_L$ and Γ into $\Gamma_0, \ldots, \Gamma_L$ such that $|\Gamma_0| \le 5dm$ and $|P_0| \le 5dn$, and such that for all $1 \le \alpha, \beta \le L$, 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 $|\Gamma_0| \le 5dm$. 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 $d^4 = L - 1$ 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 $\Gamma_1, \ldots, \Gamma_L$ 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 establish the condition 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 (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

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

By the statement analogous to Lemma 3.2, there is a subset $\widetilde{\Gamma}_0$ of $5dn$ 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 $\widetilde{\Gamma}_0$.

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 $P_1, \ldots, P_L$ so that for any two points $(q_s, q_t), (q_{s'}, q_{t'})$ in the same P_α with $\alpha \ge 1$, 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 (q_s, q_t) and $(q_{s'}, q_{t'})$ from the same P_α.

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

Lemma 3.5. *For all $1 \le \alpha, \beta \le L$ we have, for some constant A_d,*

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

Proof. We split C_2 into $6d^2$ graphical pieces B_λ, i.e., for $1 \le \lambda \le 6d^2$ there is a function g_λ such that B_λ can be parametrized as $(x, g_\lambda(x))$ for x in some closed interval I_λ. To do this, we "cut" C_2 at every point with vertical tangent line, and at every singularity. By Theorem 2.4, the number of singularities is at most d^2. The number of points with vertical tangent line is also bounded by d^2, since they are intersection points of C_2 with the curve defined by $\partial f_2 / \partial y = 0$. By Assumption 3.1.1, C_2 is not a vertical line, so $\partial f_2 / \partial y$ is not identically zero and does indeed describe a curve. It has degree at most $d - 1$, so the bound d^2 follows by Theorem 2.2 and the fact that we assumed f_2 to be of minimum degree, which implies that f_2 and $\partial f_2/\partial y$ do not have a common factor. Since C_2 has at most $4d^2$ connected components by Theorem 2.3, this results in at most $6d^2$ graphical pieces.

Every pair B_λ, B_μ, defines a patch $S_{\lambda\mu} = B_\lambda \times B_\mu \subset \mathbb{R}^4$ of the surface S, which can be parametrized as $(x_s, g_\lambda(x_s), x_t, g_\mu(x_t))$ for $(x_s, x_t) \in I_\lambda \times I_\mu$. There are at most $36d^4$ patches in total, and every incidence between P and Γ occurs on at least one patch (an incidence that occurs on a boundary will occur on multiple patches).

Now fix α and β. We set $P_\alpha^{\lambda\mu} = P_\alpha \cap S_{\lambda\mu}$ and $C_{ij}^{\lambda\mu} = C_{ij} \cap S_{\lambda\mu}$, and we let $\Gamma_\beta^{\lambda\mu}$ be the set of curves $C_{ij}^{\lambda\mu}$ for $C_{ij} \in \Gamma_\beta$. Let π be the projection $(x_s, y_s, x_t, y_t) \mapsto (x_s, x_t)$. Then for each λ, μ we have by construction that π is injective on $S_{\lambda\mu}$, which implies that $\pi(P_\alpha^{\lambda\mu})$ and $\pi(\Gamma_\beta^{\lambda\mu})$ satisfy the conditions for a system with two degrees of freedom in the plane. However, the curves in $\pi(\Gamma_\beta^{\lambda\mu})$ need not be continuous or simple.

Each $\pi(C_{ij}^{\lambda\mu}) \subset I_\lambda \times I_\mu$ is a union of at most $32d^4$ continuous curves: $\pi(C_{ij})$ has at most $16d^4$ connected components, and it is cut by the boundary of a rectangle at most $16d^4$ times. The first bound follows by applying Theorem 2.3 to the three polynomials defining C_{ij}. The second bound follows by applying Theorem 2.3 to the three polynomials defining C_{ij} together with the polynomial defining the set of points that are projected to the boundary of $I_\lambda \times I_\mu$, which is a product of four linear polynomials.

To make these continuous curves simple, we also cut them at each self-intersection. Because C_{ij} has at most d^4 singularities by Lemma 3.3, and because π injective on it, $\pi(C_{ij}^{\lambda\mu})$ also has at most d^4 self-intersections. As a result a single curve $\pi(C_{ij}^{\lambda\mu})$ is split into at most $33d^4$ simple continuous curves. We let $G_\beta^{\lambda\mu}$ be the set of these simple continuous curves, so $\pi(P_\alpha^{\lambda\mu})$ and $G_\beta^{\lambda\mu}$ form a system with two degrees of freedom.

Therefore, we can apply Theorem 2.1 with $k = 2$ and $M = 16d^4$ to get

$$I(P_\alpha^{\lambda\mu}, \Gamma_\beta^{\lambda\mu}) \le I(\pi(P_\alpha^{\lambda\mu}), G_\beta^{\lambda\mu}) \le C \cdot \max\left\{(16d^4)^{2/3}(m^2)^{2/3}(33d^4n^2)^{2/3}, m^2, 33d^4n^2\right\}.$$

Thus

$$I(P_\alpha, \Gamma_\beta) \le \sum_{\lambda,\mu} I(P_\alpha^{\lambda\mu}, \Gamma_\beta^{\lambda\mu}) \le C \cdot (16d^4)^{2/3} \cdot 36d^4 \cdot 33d^4 \cdot \max\left\{m^{4/3}n^{4/3}, m^2, n^2\right\}. \qquad \square$$

Lemma 3.6. *We have*

$$I(P, \Gamma_0) \le 10d^2mn \qquad \text{and} \qquad I(P_0, \Gamma) \le 10d^2mn.$$

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 2.2, unless C_2 equals that circle, which cannot happen by Assumption 3.1.3. Therefore, we have $I(P, \Gamma_0) \le 2dn \cdot 5dm = 10d^2mn$. The second inequality follows by applying the same argument to the curves $\widetilde{C}_{st}$ defined in the proof of Lemma 3.4. $\square$

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 transformation as introduced in [7] and used in [10, 16]. Let Q be the set of quadruples (p_i, p_j, q_s, q_t), with $1 \le i, j \le m$ and $1 \le s, t \le n$, such that $d(p_i, q_s) = d(p_j, q_t)$, and let $D = D(S_1, S_2)$ be the set of distances between S_1 and S_2.

Lemma 3.7. *We have*

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

Proof. Write $E_d = \{(p,q) \in S_1 \times S_2 : |pq| = d\}$ for $d \in D$. Using the Cauchy-Schwarz inequality, we obtain

$$|Q| \geq \sum_{i=1}^{|D|} |E_d|^2 \geq \frac{1}{|D|}\left(\sum_{i=1}^{|D|} |E_d|\right)^2 = \frac{(mn)^2}{|D|}. \qquad \square$$

Proof of Theorem 1.2. We first justify Assumption 3.1. We rotate the coordinate axes so that neither C_1 nor C_2 is a vertical line. We remove at most d^2+2 points so that S_1 and S_2 are disjoint, and so that if one of C_1, C_2 is a circle, then its center 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 $|S_1|/(2d)$ points. We do the same for S_2. In case C_1 or C_2 is a line, we do an analogous removal frome every parallel or orthogonal line, leaving at least a fraction $1/d^2$ of the points, so that the fifth and sixth parts of the assumption will be satisfied. Altogether these steps leave at least m/d^2 points in S_1 and n/d^2 in S_2. Now redefine S_1 and S_2 to be the point sets after these modifications.

Combining the bounds from Lemma 3.5 and 3.6, we obtain

$$\begin{aligned} I(P,\Gamma) &\leq I(P_0,\Gamma) + I(P,\Gamma_0) + \sum_{\alpha,\beta\geq 1} I(P_\alpha, \Gamma_\beta) \\ &\leq 20d^2mn + \sum_{\alpha,\beta\geq 1} A_d \cdot \max\left\{m^{4/3}n^{4/3}, m^2, n^2\right\} \\ &\leq B_d \cdot \max\left\{m^{4/3}n^{4/3}, m^2, n^2\right\}, \end{aligned}$$

for the constant $B_d = 3d^8A_d$, noting that the number of terms in the sum is at most $L^2 \leq 2d^8$.

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

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

This implies $|D| \geq c'_d \cdot \min\left\{m^{2/3}n^{2/3}, m^2, n^2\right\}$ for the constant $c'_d = 1/(d^4B_d)$, 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 this component C^* and set $S^* = S \cap C^*$. Now set $C_1 = C^*$, $C_2 = C^*$, and arbitrarily split S^* into two disjoint sets S_1^*, S_2^* of size roughy $n/2$. Then Assumption 3.1 holds. Therefore, the proof above gives (with a constant c_d different from c'_d)

$$|D(S)| \geq |D(S_1^*, S_2^*)| \geq c_dn^{4/3}. \qquad \square$$

Remark. *(Dependence on d)*
With the proof above, the constant c_d in Theorem 1.1 would come out to be $c_d = cd^{-56/3}$ for some absolute constant c. Roughly speaking, we get a factor $d^{8/3}$ from the application of Theorem 2.1, a factor d^8 from splitting up P and Γ in Lemma 3.4, a factor d^4 from cutting up the surface S into graphical pieces, and a factor d^4 from splitting up the projected curves into simple continuous curves. For c_d' in Theorem 1.2, we would get another factor d^4 to account for the removed points (in case C_1 or C_2 is a circle or a line).

Several of these factors would be easy to improve somewhat by being more careful. More significantly, we could replace Theorem 2.3 by a refined bound due to Barone and Basu [1], which takes into account the fact that the defining polynomials may have different degrees. With this we could in several places improve a factor d^4 to d^2. There are various other ways to bound the incidences between P and Γ, which may result in a better dependence on d. Finally, if we could replace 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 a factor d^8.

With the current proof it seems hard to improve the constant beyond $d^{-4/3}$, because of the factor $M^{2/3}$ in Theorem 2.1, and the fact that d^2 appears to be the right order of magnitude for the size of the intersections of the curves C_{ij}, and thus also for M. 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 arbitary points determine $\Omega(n)$ distances, unless the points lie on parallel lines or concentric circles.

4 Proof of Lemma 3.2

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

We are going to analyze 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 infinite intersection implies the existence of a symmetry of C_2. This is a real possibility, but cannot happen too often, as will become clear in the proof of Lemma 4.1.

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.

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 $(x_s, y_s, x_t, y_t) = (q_s, q_t) \in C_{ij} \cap C_{kl}$ corresponds to a pair of points $q_s, q_t \in C_2$ such that

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

It follows that

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

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

The idea of the proof is to reconstruct points of C_2 from the points p_i, p_k using the distance pairs from D. 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:

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

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 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 $S_1 = E_1 \cap E_1^*$ is an infinite subset of C_2 such that $T(S_1) \subset C_2$;

(2) If $|E_1 \cap E_1^*|$ is not infinite, then $S_2 = |E_1 \cap E_2^*|$ must be infinite. Then $S_2 = E_1 \cap E_2^*$ is an infinite subset of C_2 such that $(T \circ M)(S_2) \subset C_2$, since M maps S_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 S of an irreducible algebraic curve C such that $T(S) \subset C$, then $T(C) = C$, i.e., T is a symmetry of C. This holds because $T(C)$ is also an irreducible plane algebraic curve, so by Theorem 2.2 it either has finite intersection with C, or equals it.

If T is a rotation, then in case (1) C_2 has a rotation symmetry, while in case (2) it has a reflection symmetry. If T is a translation, then in case (1) it has a translation symmetry, while in case (2) it has a glide reflection symmetry (a glance at the proof of Lemma 2.6 would now make clear that T can only be a translation if C_2 is a line). □

Lemma 4.2. *Suppose that $p_i, p_j, p_k, p_l, p_q, p_r \in S_1$ satisfy $d(p_i, p_k) \neq d(p_j, p_l)$, $d(p_i, p_q) \neq d(p_j, p_r)$, and $d(p_k, p_q) \neq d(p_l, p_r)$. If*

$$|C_{ij} \cap C_{kl} \cap C_{qr}| \geq 2d + 1,$$

then C_2 is a conic or a line.

Proof. A point in $|C_{ij} \cap C_{kl} \cap C_{qr}|$ corresponds to two points (x, y) and (u, v) on C_2 such that the distances from p_i, p_k, p_q to (x, y) are respectively equal to those from p_j, p_l, p_r to (u, v). We will show that the set of such points (u, v) (or (x, y)) is forced to lie on a conic or a line, so by Theorem 2.2 C_2 contains at most $2d$ of them, unless C_2 is a conic or a line.

We can assume that $p_i = (0, 0)$ and $p_k = (1, 0)$. We can also assume that $p_j = (0, 0)$ and $p_l = (L, 0)$, with $L \neq 0, 1$, since the property of lying on a conic is preserved under a rotation, as are distances. Finally, write $p_q = (a, b)$ and $p_r = (c, d)$.

Consider the points (x, y) and (u, v) that have the same distances d_1, d_2, d_3 from respectively p_i, p_k, p_q and p_j, p_l, p_r. They satisfy the equations

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

$$(x - 1)^2 + y^2 = (u - L)^2 + v^2, \tag{3}$$

$$(x - a)^2 + (y - b)^2 = (u - c)^2 + (v - d)^2. \tag{4}$$

Subtracting (2) from (3) gives

$$x = Lu + \frac{1}{2}(1 - L^2). \tag{5}$$

Subtracting (2) from (4), and plugging (5) into the result, leads to

$$by = (c - aL)u + dv + \frac{1}{2}(a^2 + b^2 - c^2 - d^2 + aL^2 - a). \tag{6}$$

Plugging the linear equations (5) and (6) into (2) leads to

$$(b^2L^2 + (c - aL)^2 - b^2)u^2 + (d^2 - b^2)v^2 + 2d(c - aL)uv + l(u, v) = 0,$$

where $l(u, v)$ is a linear function of u and v. This shows that (u, v) must lie on a conic or a line, and finishes the proof, unless this equation is identically zero.

We show that in that case C_2 must be a line, using the coefficients of the quadratic terms. We would have $d(c - aL) = 0$ from the term uv, so either $d = 0$, or $c - aL = 0$. If $c - aL = 0$, then the coefficient of the term u^2 would give $b^2(L^2 - 1) = 0$. Since $L \neq 1$, we would get $b = 0$, and then the term v^2 would give $d = 0$. So in all cases we have $d = 0$. But then we see from (5) and (6) that the linear transformation does not depend on v, which implies that its image is a line, hence C_2 must be a line. □

Lemma 4.3. *Suppose that $p_i, p_j, p_k, p_l, p_q, p_r \in S_1$ satisfy $d(p_i, p_k) \neq d(p_j, p_l)$, $d(p_i, p_q) \neq d(p_j, p_r)$, and $d(p_k, p_q) \neq d(p_l, p_r)$.*
If C_2 is a conic then

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

Proof. Suppose that $|C_{ij} \cap C_{kl} \cap C_{qr}| \geq 5$. Then we have three equations of the form

$$(x - a_\alpha)^2 + (y - b_\alpha)^2 = (u - c_\alpha)^2 + (v - d_\alpha)^2, \tag{7}$$

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, as in the previous lemma, we can view as an affine transformation T sending (x, y) to (u, v). 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 2.7 then tells us which form 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). Then by Lemma 2.7, T must have the form $(u, v) = T(x, y) = (rx + (r^2 - 1)y/r, y/r)$ (or the second form, which we will leave to the reader). Plugging this into (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 (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 $s^2x^2 + t^2y^2 = 1$. Then by Lemma 2.7, T must have the form $(u,v) = T(x,y) = ((\cos\theta)x \pm \frac{t}{s}(\sin\theta)y, \frac{s}{t}(\sin\theta)x \mp (\cos\theta)y)$. Plugging it into (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 the term 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 $s = \pm t$. That 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

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

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

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

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, again contradicting the assumption of the lemma. □

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

$$|C_{ij} \cap C_{kl}| \geq 3,$$

then p_i and p_k lie on a line orthogonal to C_2.

Proof. We can assume that $p_i = (0,0)$ and $p_k = (1,0)$. We can also apply a rotation so that $p_j = (0,0)$ and $p_l = (L,0)$, with $L > 1$ (possibly after changing the order of the points). Then a point in $C_{ij} \cap C_{kl}$ gives points (x,y) and (u,v) such that

$$x^2 + y^2 = u^2 + v^2 \tag{8}$$
$$(x-1)^2 + y^2 = (u-L)^2 + v^2 \tag{9}$$

with $L > 1$. Suppose that C_2 is a line satisfying $y = ax + b$. Then, as in the proof of Lemma 4.2, we have

$$x = Lu + \frac{1}{2}(1 - L^2), \quad y = aLu + \frac{1}{2}a(1 - L^2) + b.$$

Plugging into (8) and rearranging gives

$$(L^2(1+a^2)-1)u^2 - v^2 + l(u,v) = 0.$$

Since $L > 1$, the coefficient of u^2 is positive, so this equation describes either a hyperbola, or a union of two lines (if $l(u,v)$ is identically zero). In the first case, at most two points of the line C_2 can satisfy it. In the second case, C_2 must be a line of the form $v = \pm L\sqrt{1+a^2} \cdot u + c$. But then we must have $a = \pm L\sqrt{1+a^2}$, so $L^2 = a^2/(1+a^2) < 1$, contrary to our assumption. This proves the lemma for lines of the form $y = ax + b$.

The remaining possibility is that C_2 has the form $x = c$, which is the exception in the statement of the lemma. □

Proof of Lemma 3.2. If there is a symmetry T of C_2 that maps p_i to p_j, we will say that C_{ij} *comes from* T. Suppose that the curves C_{ij} and C_{kl} have infinite intersection and $d(p_i,p_k) = d(p_j,p_l)$. Then by Lemma 4.1, there is a symmetry T of C_2 such that C_{ij} and C_{kl} come from T.

In case C_2 is not a line or a circle, it has at most $5d$ symmetries, by Lemma 2.6. Given a fixed symmetry T, each p_i is sent to a unique point p_j, so there are at most m curves C_{ij} that come from T. Therefore, there are in total at most $5dm$ curves C_{ij} that come from some symmetry. We let Γ_0, the set to be excluded, contain all curves C_{ij} that come from some symmetry of C_2. Then $|\Gamma_0| \leq 5dm$.

In case 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. If p_i and p_j were distinct, they would have to lie on a concentric circle (see the proof of Lemma 2.6), which is excluded by Assumption 3.1.4. If $p_i = p_j$, then the symmetry would have to be a rotation around p_i, which would imply that C_2 is 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.

If $C_{ij}, C_{kl} \in \Gamma\backslash\Gamma_0$ have infinite intersection, then it follows from Lemma 4.1 that $d(p_i,p_k) \neq d(p_j,p_l)$. This allows us to conclude that there are no three curves in $\Gamma\backslash\Gamma_0$ that have infinite intersection, by Lemmas 4.2 and 4.3 if C_2 is not a line, and otherwise by Lemma 4.4 and Assumption 3.1.6. This finishes the proof of Lemma 3.2. □

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

References

[1] S. Barone, S. Basu, *On a real analogue of Bezout inequality and the number of connected components of sign conditions*, in arXiv:1303.1577.

[2] S. Basu, R. Pollack, and M. Roy, *Algorithms in real algebraic geometry*, Springer-Verlag, Berlin, 2006.

[3] P. Brass, W. Moser, and J. Pach, *Research Problems in Discrete Geometry*, Springer-Verlag, New York, 2005.

[4] M. Charalambides, *Exponent gaps on curves via rigidity*, in arXiv:1307.0870.

[5] G. Elekes, *A note on the number of distinct distances*, Period. Math. Hung., 38 (1999), 173–177.

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

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

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

[9] C.G. Gibson, *Elementary Geometry of Algebraic Curves*, Cambridge University Press, Cambridge, 1998.

[10] L. Guth and N. H. Katz, *On the Erdős distinct distances problem in the plane*, arxiv.

[11] J. Harris, *Algebraic Geometry: A First Course*, Springer-Verlag, New York, 1992.

[12] J. Pach and M. Sharir, *Repeated angles in the plane and related problems*, Journal of Combinatorial Theory, Series A, 59 (1990), 12–22.

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

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

[15] J. Solymosi and M. Sharir, *Distinct distances from three points*, in arXiv:1308.0814.

[16] M. Sharir, A. Sheffer, and J. Solymosi, *Distinct distances on two lines*, Journal of Combinatorial Theory, Series A, 120:1732–1736, 2013.

[17] A. Sheffer, J. Zahl, and F. de Zeeuw, *Few distinct distances implies no heavy lines or circles*, in arXiv:1308.5620.

[18] L. Székely, *Crossing Numbers and Hard Erdős Problems in Discrete Geometry*, Combinat. Probab. Comput., 6 (1997), 353–358.

[19] N. Wallach, *On a Theorem of Milnor and Thom*, Progress in Nonlinear Differential Equations, 20 (1996), 331–348.