INSCRIBED SQUARES IN PLANE CURVES

BY

R. P. JERRARD

Consider a simple, closed, plane curve $C$ which is a real-analytic image of the unit circle, and which is given by $x = x(t)$, $y = y(t)$. These are real analytic periodic functions with period $T$. In the following paper it is shown that in a certain definite sense, exactly an odd number of squares can be inscribed in every such curve which does not contain an infinite number of inscribed squares. This theorem is similar to the theorem of Kakutani [1] that there exists a circumscribing cube around any closed, bounded convex set in $E_3$. The latter theorem has been generalized by Yamabe and Yujobo [2], and Cairns [3] to show that in $E_n$ there are families of such cubes. Here, for the case of squares inscribed in plane curves, we remove the restriction to convexity and give certain other results.

A square inscribed in a curve $C$ means a square with its four corner points on the curve, though it may not lie entirely in the interior of $C$. Indeed, the spiral

$$r = k\theta, \quad 2\pi \leq \theta \leq 4\pi,$$

with the two endpoints connected by a straight line possesses only one inscribed square. The square has one corner point on the straight line segment, and does not lie entirely in the interior.

On $C$, from the point $P$ at $t_0$ to the point $Q$ at $t = t_0 + s$, we construct the chord, and upon the chord as a side erect a square in such a way that as $s$ approaches zero the square is inside $C$. As $s$ increases we consider the two free corner points of the square, $P_c$ and $Q_c$, adjacent to $P$ and $Q$ respectively. As $s$ approaches $T$ the square will be outside $C$ and therefore both $P_c$ and $Q_c$ must cross $C$ an odd number of times as $s$ varies from zero to $T$. The points may also touch $C$ without crossing.

Suppose $P_c$ crosses $C$ when $s = s_1, s_2, \ldots, s_n$. We now have certain squares with three corners on $C$. For any such square the middle corner of these will be called the vertex of the square and the corner not on the curve will be called the diagonal point of the square. Each point on $C$, as a vertex, may possess a finite number of corresponding diagonal points by the above construction.

To each paired vertex and diagonal point there corresponds a unique forward corner point, i.e., the corner on $C$ reached first by proceeding along $C$ from the vertex in the direction of increasing $t$. If the vertex is at $t_0$, and if the interior of $C$ is on the left as one moves in the direction of increasing $t$,

Received by the editors December 28, 1959 and, in revised form, June 29, 1960.

234
then every such corner can be found from the curve obtained by rotating \( C \) clockwise through 90° about the vertex. The set of intersections of \( C_{t_0} \), the rotated curve, with the original curve \( C \) consists of just the set of forward corner points on \( C \) corresponding to the vertex at \( t_0 \), plus the vertex itself. We note that two such curves \( C \) and \( C_t \) cannot coincide at more than a finite number of points; otherwise, being analytic, they would coincide at all points, which is impossible since they do not coincide near \( t_0 \).

With each vertex we associate certain numerical values, namely the set of positive differences in the parameter \( t \) between the vertex and its corresponding forward corner points. For the vertex at \( t_0 \), these values will be denoted by \( f(t_0) \). The function \( f(t) \) defined in this way is multi-valued.

We consider now the graph of the function \( f(t) \) on \([0, T] \times [0, T]\). We will refer to the plane of \( C \) and \( C_t \) as the \( C \)-plane and to the plane of the graph as the \( f \)-plane. The graph, as a set, may have a finite number of components. We will denote the values of \( f(t) \) on different components by \( f_1(t), f_2(t), \ldots \).

Each point with abscissa \( t \) on the graph represents an intersection between \( C \) and \( C_t \). There are two types of such intersections, depending essentially on whether the curves cross at the point of intersection. An ordinary point will be any point of intersection \( A \) such that in every neighborhood of \( A \) in the \( C \)-plane, \( C_t \) meets both the interior and the exterior of \( C \). Any other point of intersection between \( C \) and \( C_t \) will be called a tangent point. This terminology will also be applied to the corresponding points in the \( f \)-plane. We can now prove several lemmas.

**Lemma 1.** In some neighborhood in the \( f \)-plane of any ordinary point of the graph, the function \( f \) is a single-valued, continuous function.

**Proof.** We first show that the function is single-valued in some neighborhood. With the vertex at \( t_0 \) in the \( C \)-plane we assume that \( t_1 = t_0 + f_1(t_0) \) is the parametric location on \( C \) of an ordinary intersection \( Q \) between \( C \) and \( C_{t_0} \). In the \( f \)-plane the coordinates of the corresponding point are \((t_0, f_1(t_0))\). We know that in the \( C \)-plane both \( C \) and \( C_{t_0} \) are analytic. In the \( C \)-plane we construct a set of rectangular Cartesian coordinates \( u, v \) with the origin at \( Q \) and such that both \( C \) and \( C_{t_0} \) have finite slope at \( Q \). Near \( Q \), both curves can be represented by analytic functions of \( u \). In a neighborhood of \( Q \) the difference between these functions is also a single-valued, analytic function of \( u \). Furthermore, one can find a neighborhood of \( Q \) in which the difference function is monotone, for since it is analytic it can have only a finite number of extrema in any interval. Now, to find \( f_i(t_0 + \epsilon) \), one needs the intersection of \( C \) and \( C_{t_0+\epsilon} \) near \( Q \). But \( C_{t_0+\epsilon} \) is just the curve \( C_{t_0} \) translated without rotation through a small arc, for \( C_t \) is always obtained by rotating \( C \) through exactly 90°. The arc is itself a segment of an analytic curve. Thus if \( \epsilon \) is sufficiently small, there can be only one intersection of \( C \) and \( C_{t_0+\epsilon} \) near \( Q \), for if there were more than one intersection for every \( \epsilon \) then the difference between \( C \)
and \( C_{t_0} \) near \( Q \) would not be a monotone function. Therefore, \( f_i \) is single-valued near \( Q \). It is also seen that

\[
\lim_{\epsilon \to 0} f_i(t_0 + \epsilon) = f_i(t_0),
\]

since the change from \( C_{t_0} \) to \( C_{t_0 + \epsilon} \) is accomplished by a continuous translation. Thus \( f_i \) is also continuous at \( t_0 \), and in a neighborhood of \( t_0 \) which does not contain a tangent point.

We turn now to the set of tangent points on the graph. This set must consist of isolated points and closed intervals. The fact that there can not be any limit points of the set except in closed intervals follows from the argument used in Lemma 1, namely, that near any tangent point in the \( C \)-plane the curves \( C \) and \( C_i \) are analytic, and therefore the difference between them must be a monotone function in some neighborhood on either side of the tangent point. This prevents the occurrence of an infinite sequence of isolated tangent points.

**Lemma 2.** In some neighborhood of an isolated tangent point in the \( f \)-plane, say \((t_i, f_j(t_i))\), the function \( f_j \) is either double-valued or has no values defined, except at the tangent point itself, where it is single-valued.

**Proof.** A tangent point \( Q \) in the \( C \)-plane occurs when \( C \) and \( C_{t_i} \) are tangent to one another. A continuous change in \( t \) through an amount \( \epsilon \) results in a translation along an analytic arc of the curve \( C_{t_i} \). There are three possibilities: (a) \( C_{t_i} \) remains tangent to \( C \) as it is translated; (b) \( C_{t_i} \) moves away from \( C \) and does not intersect it at all for \( t_i < t \leq t_i + \epsilon \); (c) \( C_{t_i} \) cuts across \( C \) and there are two ordinary intersections for every \( t \) in \( t_i < t \leq t_i + \epsilon \). The first possibility results in a closed interval of tangent points in the \( f \)-plane, the end points of which fall into category (b) or (c). In the second category the function \( f_j \) has no values defined in a neighborhood \((t_i, t_i + \epsilon)\). In the third category the function is double-valued in this interval. The same remarks apply to an interval on the other side of \( t_i \). Again, the analyticity of the two curves guarantee that such intervals exist. In the neighborhood of an end point of an interval of tangent points in the \( f \)-plane the function is two-valued or no-valued on one side, and is a single-valued function consisting entirely of tangent points on the other side.

With the above results we can make the following remarks about the graph of \( f \). First, for any value of \( t \) for which all values of \( f(t) \) are ordinary points the number of values of \( f(t) \) must be odd. For it is clear that the total number of ordinary intersections of \( C \) and \( C_i \) must be even (otherwise, starting in the interior of \( C \), \( C_i \) could not finally return to the interior), and the center of rotation at \( t \) is the argument of the function, not a value. Therefore, for any value of \( t \) the number of values of \( f(t) \) is equal to the (finite) number of tangent points corresponding to the argument \( t \) plus an odd number.
Definition. The number of ordinary values of the function \( f(t) \) at \( t \) will be called its multiplicity at \( t \).

Lemma 3. The graph of \( f \) has at least one component whose support is the entire interval \([0, T]\).

Proof. We suppose not. Then every component of the graph of \( f \) must be defined over a bounded sub-interval. Suppose \( f_a \) is defined in the sub-interval \( t_a \leq t \leq t_{a+1} \). Now \((t_{a+1}, f_a(t_{a+1}))\) and \((t_a, f_a(t_a))\) must both be tangent points on the \( n \)th component in the \( f \)-plane; otherwise by Lemma 1 the component would extend beyond these points. Further, we see by Lemma 2 that the multiplicity of \( f \) can only change at a tangent point, and at such a point can only change by an even integer. Thus the multiplicity of \( f_a(t) \) for a given \( t \) must be an even number. This is true of all components which have such a bounded support. But this is a contradiction, for we know that the multiplicity of \( f(t) \) is odd for every \( t \).

We have shown that the graph of \( f \) contains at least one component whose inverse is the entire interval \([0, T]\), and whose multiplicity is odd. There must be an odd number of such components, which will be called complete components. The remaining (incomplete) components all have an even number of ordinary points at any argument, and are defined only on a proper sub-interval of \([0, T]\).

We must now show that on some component of the graph there exist two points for which the corresponding diagonal points in the \( C \)-plane are on opposite sides of \( C \). We again consider a fixed point \( P \) at \( t_0 \) and a variable point \( Q \) at \( t_0 + s \) on \( C \). We erect a square with \( PQ \) as a side and with free corners \( P_c \) and \( Q_c \) adjacent to \( P \) and \( Q \) respectively. As \( s \) varies from zero to \( T \), the values of \( s \) for which \( P_c \) and \( Q_c \) cross \( C \) will be denoted by \((s_1, s_2, \ldots, s_n)\) and \((s'_1, s'_2, \ldots, s'_m)\) respectively. We have

\[
f(t_0) = s_1, s_2, \ldots, s_n, \text{ plus tangent points.}
\]

These \( s \)-values are just the ordinary values of \( f(t_0) \).

Lemma 4. The values \((s'_1, s'_2, \ldots, s'_m)\) are the ordinary values at \( t_0 \) of a multivalued function \( g(t) \) which has components corresponding to those of \( f(t) \).

Proof. We first define a function \( b(t) \) as follows: given the set of squares such that each has three corners on \( C \) and vertex at \( t \), \( b(t) \) is the corresponding set of positive parametric differences between \( t \) and the backward corner points. The functions \( f \) and \( b \) have exactly the same multiplicity at every argument \( t \). Now with \( P \) fixed at \( t_0 \), \( s'_i \)-values occur when the corner \( Q_c \) crosses \( C \), and are among the values of \( s \) such that \( s = b(t_0 + s) \). The roots of this equation are just the ordinates of the intersections of the graph of \( b \) with a straight line of unit slope through \( t_0 \) in the \( b \)-plane (the plane of the graph of \( b \)). We define these values as \( g(t_0) \), and define \( g(t) \) in the same way for each \( t \). Thus
we obtain $g(t)$ by introducing an oblique $g(t)$-axis in the $b$-plane. The function $g$ has ordinary and tangent points according to whether the line of unit slope in the $b$-plane crosses or is tangent to the graph of $b$, and the numbers $s_i$ are the ordinary values of $g(t_0)$. The graph of $g$ has components, including an odd number of complete components, corresponding to the components of $b$ and hence to the components of $f$.

We now return to the values

$$f(t_0) = s_1, s_2, \ldots, \bar{s}_p, \ldots, s_n, \text{ plus tangent points},$$

$$g(t_0) = s'_1, s'_2, \ldots, \bar{s}'_q, \ldots, s'_m, \text{ plus tangent points},$$

where we have distinguished the least $s$-value on a complete component with a bar. We define certain closed intervals on the $s$-axis between zero and $T$. An $I$-interval will be defined by two adjacent $s_i$-values belonging to the same component of $f$. We start with the two largest such values, continue with the next largest pair, etc. If the component is incomplete all the $s_i$-values on it will be exhausted; if it is complete there will be one remaining value, say $s_p$, which we reserve. We define a set of $I'$-intervals in the same way, pairing $s'_i$-values which belong to the same component of $g$. Finally we define an odd number of $\bar{I}$-intervals, each having for its endpoints the least values $\bar{s}_i$ and $\bar{s}'_q$ belonging to a corresponding complete component of $f$ and $g$.

**Lemma 5.** On some component of $f$ or $g$ there exist two points whose corresponding diagonal points in the C-plane lie on opposite sides of $C$.

**Proof.** We write $D(s_i)=0$ or 1 according to whether the diagonal point corresponding to $s_i$ is outside or inside $C$, respectively. Suppose the lemma is false; then if any interval defined above is $(c, d)$ we must have $D(c) = D(d)$. This requirement yields the following properties of the intervals: (a) Each $I$-interval must contain an even number of $s'_i$-values. (b) Each $I'$-interval must contain an even number of $s_i$-values. Property (a) follows from the fact that the $s'_i$-values represent the crossings of the diagonal point $Q_c$ corresponding to $P$, and there must be an even number of such crossings in any $I$-interval if its endpoints are to have their corresponding diagonal points on the same side of $C$. A similar argument holds for property (b).

Suppose now that there is just one $\bar{I}$-interval $(\bar{s}_p, \bar{s}'_q)$. We can assume without loss of generality that $\bar{s}_p < \bar{s}'_q$. There are two possibilities. (A) If $D(\bar{s}_p) = 1$, then $D(\bar{s}'_q) = 1$. But then $\bar{s}'_q$ must be contained in an $I$-interval, because the diagonal point $P$ passes outside $C$ when $s = \bar{s}_p$, and $D(\bar{s}'_q) = 1$ implies $P$ is inside $C$ when $s = \bar{s}'_q$. Now, except for $\bar{s}'_q$, the $s'_i$-values are divided into pairs corresponding to the $I'$-intervals. Then using property (a) we see that some $I'$-interval, say $(a, b)$, must span an endpoint of an $I$-interval. Then $D(a) \neq D(b)$. (B) If $D(\bar{s}_p) = 0$, then $\bar{s}_p$ is contained in an $I'$-interval. For if not, since the diagonal point $Q_c$ passes outside $C$ at $s = \bar{s}_p$, $Q_c$ would be inside $C$ at $s = \bar{s}_p$ contrary to hypothesis. Then by using property (b) we see that some $I = (c, d)$ must span the endpoint of an $I'$-interval, and $D(c) \neq D(d)$. 

---

License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use
If there is more than one complete component the arguments still hold with slight modifications, for there must be an odd number of such components and a corresponding odd number of $T$-intervals. The omission of tangent point values of $f(t_0)$ and $g(t_0)$ does not affect the proof, for the points $P_e$ and $Q_e$ do not cross $C$ at these omitted values of $s$.

**Theorem 1.** A square can be inscribed in $C$.

**Proof.** The result follows immediately from Lemma 5. There exist two points on some component of $f$ or $g$, say on $f_n$, whose corresponding diagonal points are on opposite sides of $C$. But the position of the diagonal point of the vertex at $t$ depends continuously on $f_n(t)$. Then there exists $t_1$ such that the diagonal point corresponding to $f_n(t_1)$ is on $C$. Since the other three corners of the square are on $C$ by construction, a square is inscribed in $C$.

We now consider the effect upon an inscribed square in $C$ of a deformation of the curve $C$. We will show that as the curve is deformed, a continuously changing square is inscribed in it. For this purpose we suppose that an isotopy $H_r (0 \leq r \leq 1)$ is defined by a continuously varying homeomorphism of an analytic curve $C_0$ such that $H_r(C_0) = C_r$ is again a simple, closed, plane analytic curve. Each such curve $C_r$ will be parameterized by $t$ ($0 \leq t \leq T$) so that when $t$ is fixed, the point on curve $C_r$ denoted by $t$ will vary continuously with $r$. For each curve $C_r$ in the course of this isotopy we can construct the functions $f_r$ and $g_r$, which we will now call $f$ and $g_r$. We denote by $F_r$ and $G_r$ the graphs of $f_r$ and $g_r$ with all tangent points which are not limit points of ordinary points of $f_r$ and $g_r$ omitted. This deletion simplifies the succeeding argument without altering it.

We can now define a new multiple-valued function $h: I \rightarrow \mathbb{R}$ (with graph $H$) as follows: $h(r)$ is the set of values of $t$ at which $F_r$ and $G_r$ intersect, but including only one, say the least, of the four such numbers corresponding to any inscribed square in $C_r$. The points of $H$ thus correspond to inscribed squares in the family of curves $C_r$. Due to the analyticity of $C_r$, $F_r$ and $G_r$ are differentiable, and we define as an ordinary point of $H$ any point corresponding to an intersection at which $F_r$ and $G_r$ cross one another. All other points of $H$ will be called tangent points.

**Lemma 6.** The multiplicity of $h$ is either zero or an odd number.

**Proof.** With certain exceptions we can use exactly the arguments of Lemmas 1 and 2. These arguments show that in the neighborhood of an ordinary point $h$ is single-valued and continuous, and that the multiplicity of $h$ can change only at a tangent point and then only by an even integer. One exception is provided by the case where $F_r$ and $G_r$ are tangent to one another at a point where they cross. At such an ordinary point of $H$ a small change in $r$ may correspond locally to a rotation of $F_r$ with respect to $G_r$, resulting in the appearance of an odd number of ordinary points. Indeed, such a point of $H$ can be the limit of a sequence of tangent points of $H$. Nonetheless, at such
an ordinary point the property that the multiplicity of \( h \) changes only by an even integer as \( r \) changes is preserved.

A second modification arises from the possibility that \( F_r \) and \( G_r \) may coincide. If \( F_r \) and \( G_r \) coincide over some interval, or if they have an infinite number of intersections, there must be a family of inscribed squares in \( C_r \) whose four corner points cover four convergent sequences of points on \( C_r \). But then because of the analyticity of \( C_r \) this family of squares can be specified by analytic functions which represent the position of the center of the square, its angular orientation, and its side length. These functions are determined everywhere by the sequences on which they are defined. The analytic curve described by any corner point of this family of squares must coincide with \( C_r \) at every point of a sequence, and must therefore coincide with \( C_r \) everywhere. Then the complete components of \( F_r \) and \( G_r \) must coincide. A curve with this property will be called a tangent curve. When \( C_r \) is a tangent curve, the multiplicity of \( h(r) \), that is, the number of ordinary points of \( h(r) \), may be zero.

To complete the proof, in the isotopy \( H \), we may take \( C_0 \) to be an ellipse and \( C_1 \) to be any analytic mapping of the unit circle. The isotopy can be defined by placing anywhere in the interior of \( C_1 \) an ellipse similar to \( C_0 \), and constructing the harmonic function \( u \) that vanishes on \( C_0 \) and takes the value 1 on \( C_1 \). The level curve \( u = r \) is the curve \( C_r \). The multiplicity of \( h(0) \) is 1 (there is one square inscribed in any ellipse). Since the multiplicity of \( h \) can change only by an even integer, \( h(1) \) has odd multiplicity unless it is a tangent curve, when \( h(1) \) may have zero multiplicity. The possibility of an intermediate tangent curve may be avoided by changing the initial ellipse \( C_0 \). Thus for an arbitrary analytic curve there are either an odd number of ordinary values of \( h \) or no such values, and the lemma is proved.

We can make use of Lemma 6 after the definition: A double square is an inscribed square such that the corresponding point of \( H \) is a tangent point. This means that \( F_r \) and \( G_r \) are tangent without crossing at their points of intersection corresponding to the vertices of the square, and that an arbitrarily small deformation of \( C_r \) causes the square either to disappear or to split into an even number of squares. Since values of \( h(r) \) correspond to inscribed squares in \( C_r \) we have immediately from Lemma 6

**Theorem 2.** If double squares are counted twice, there are an odd number of inscribed squares in every simple, closed, analytic curve which has a finite number of inscribed squares; otherwise there is a continuous family of inscribed squares.

The statement and proof of Lemma 3 also apply to the function \( h \). The exceptional cases discussed above do not essentially alter the proof, which again rests upon the fact that the multiplicity of \( h \) can change only by an even integer. Therefore, \( H \) has a continuous component covering \( I \); a complete component. We note that the graph \( H_r \) of the complete component may contain vertical straight-line segments corresponding to values of \( r \) for which
$C_r$ is a tangent curve. The points of $h(r)$ on this component are in 1:1 correspondence with a family of squares $S_r$ inscribed in the curves $C_r$.

The graph $H_c$ can be described by a continuous mapping of the unit interval $I$ into the $H$-plane such that the points of $H_c$ are parameterized by a parameter $q$, $0 \leq q \leq 1$. This mapping can be used to define a continuous function $r(q)(r(0) = 0, r(1) = 1)$ as the $r$-value of the curve $C_r$ in which the square corresponding to the point of $H_c$ given by $q$ is inscribed. Returning to the isotopy $H_r: C_0 \rightarrow C_1$ we see that the function $r(q)$ enables us to define a new isotopy $H_{r(q)}: C_0 \rightarrow C_1$. This is essentially a mapping from points of the unit interval (identified by the number $q$) to the corresponding curves $C_{r(q)}$. As under $H_r$, the curve $C_0$ is continuously deformed into $C_1$ under the isotopy $H_{r(q)}$, but certain reversals of the course of the original isotopy are now introduced.

**Theorem 3.** Given an isotopy $H_r: C_0 \rightarrow C_1$ on analytic curves

$$(H_r(C_0) = C_r, 0 \leq r \leq 1)$$

there exists a continuous family $S_r$ of squares inscribed in the curves $C_r$ and a continuous function $r(q)(r(0) = 0, r(1) = 1)$ such that the squares $S_{r(q)}$ vary continuously with $q$ in the new isotopy $H_{r(q)}: C_0 \rightarrow C_1$; the function $r$ maps the unit interval onto itself, and $H_{r(q)}(C_0) = C_{r(q)}$.

**Proof.** There is a 1:1 correspondence between the inscribed squares $S_{r(q)}$ and the points of the complete component $H_c$, which is a continuous image of the unit interval. Indeed, considering only the complete component of $h$, the function $h(r(q))$ is continuous, and its value is just the parametric location on $C_{r(q)}$ of one corner of the inscribed square $S_{r(q)}$. This corner identifies the square, so the family $S_{r(q)}$ is continuous in $q$.

Some other remarks can be made. (1) Theorem 1 can be extended quite easily to apply to any simple, plane, closed differentiable curve. (2) A triangle similar to any given triangle can be inscribed in any Jordan curve. This follows from the construction used to obtain squares with three corners on a curve. (3) It is not possible to inscribe a regular $n$-gon, $n \geq 4$, in every convex curve. A counter-example is a semi-circle with its endpoints connected by a diameter. The circular segment must contain at least three corners of the $n$-gon, but then all other corners must lie on the same circle.

**References**


**University of Illinois, Urbana, Illinois**