The Descartes Method for Real Root Isolation

Let $A(x) = \sum_{i=0}^{n} a_i x^i$ denote a degree n polynomial with real coefficients. Given a sequence of $n+1$ real numbers $a_0, \ldots, a_n$, recall that $\mathrm{Var}(a_0, \ldots, a_n)$ is the number of sign changes, i.e., change from positive to negative and vice versa, in the sequence obtained from $a_0, \ldots, a_n$ after dropping all the zero entries. If a_i's are the coefficients of a polynomial, then we will use the succinct notation $\mathrm{Var}(A)$.

The famous Descartes's rule of signs states the following:

> THEOREM 1 (Descartes). *The number of positive real roots of a polynomial $A(x)$, where we count roots with their multiplicities, is smaller than the number of sign variation in its coefficients, $\mathrm{Var}(a_0, \ldots, a_n)$, by an even number. Moreover, if all the roots of A are real then the count is exact.*

As a corollary it follows that if there are zero or one sign variation, then there are no positive real roots or exactly one positive real root (resp.). But the rule gives us an estimate of the roots in the interval $[0, \infty)$. However, for our algorithm we need an estimate on the number of roots in any interval (a, b); let us denote it by $N_A(a, b)$. A "little observation" of Jacobi gives us the desired result: Define

$$B(x) := (x+1)^n A\left(\frac{ax+b}{x+1}\right) = \sum_{i=0}^{n} b_i x^i. \tag{1}$$

Then

$$N_A(a, b) = \mathrm{Var}(b_0, \ldots, b_n) - \text{ some even number.}$$

We will write $\mathrm{Var}(A; a, b)$ to stand for the sign variations in the coefficients of the polynomial B. The result follows since the roots of A in the interval (a, b) are mapped to the positive roots of B. Observe that the transformation $x \to (ax+b)/(x+1)$ maps $[0, \infty)$ to (a, b), i.e. transforming the domain of the polynomial A by ϕ to get B implies that the inverse map ϕ^{-1} maps the roots of A to the roots of B.

Based upon the Descartes's rule of signs and Jacobi's trick a straightforward algorithm can be described for real root isolation in any input interval I_0.

```
The Descartes Method
INPUT: Polynomial A(x) ∈ ℝ[x] and an interval I_0.
OUTPUT: A sequence of isolating intervals for the real roots of A in I_0.
1.  Initialize a queue Q ← I_0.
2.  While Q is not empty do
        Pop an interval I from Q.
        If Var(A; I) = 1 then output I.
        If Var(A; I) > 1 then subdivide I into two equal halves and push the two halves onto Q.
```

Does this algorithm terminate? Why will the sign-variations come down? Since the algorithm will clearly not terminate if there are roots in I_0 with multiplicity greater than one, so to simplify matters we will assume that A is square-free throughout. Note that if we want to isolate the positive roots of A then we should choose $I_0 := (0, 2\|A\|_\infty)$ (from Cauchy's bound, assuming $A \in \mathbb{Z}[x]$).

The termination criterion for the Descartes method is not clear, in contrast to the Sturm's method where we knew that once an interval contained at most one root, we will terminate immediately. We next study termination criteria for the Descartes's rule of signs, and try to understand why the rule counts with an excess of even number.

1 Termination – Obreshkoff's Results

In this section we will study criteria on the geometry of roots that guarantee exact count by the sign variations. We have already seen in Theorem 1 that when all roots are real the count is exact. What can we say in the presence of non-real roots? We start with a simple observation.

THEOREM 2. *If all the roots of B have negative real part then* $\mathrm{Var}(B) = 0$.

This can be proved using induction; the base case $x + \alpha$, $\alpha > 0$, is trivially true; clearly, multiplying a polynomial with all positive coefficients with a polynomial $(x + \alpha)$ yields a polynomial with all positive coefficients.

When can we say that $\mathrm{Var}(B) = 1$, i.e., B has exactly one positive real root? Where should the other roots of B be such that this is the case? From the theorem above, we can assume that all the other real roots are negative. But what about the non-real roots? Is it sufficient that they have negative real parts as in Theorem 3?

Let's suppose $B(x) = B_0(x)\prod_{j=1}^{k}(x^2 - 2i\alpha_j x + \alpha_j^2 + \beta_j^2)$, where all roots of $B_0(x)$ are real and exactly one of them is positive; from Theorem 1, we know that $\mathrm{Var}(B_0) = 1$. We want to impose constraints on α_j, β_j such that $\mathrm{Var}(B) = 1$, i.e., the product by the k quadratic factors does not increase $\mathrm{Var}(B_0)$. Consider the product of B_0 with one quadratic factor $(x^2 - 2i\alpha x + \alpha^2 + \beta^2)$. Considering Theorem 1, it helps to assume that $\alpha < 0$. Let

$$C(x) := B_0(x)(x^2 - 2i\alpha x + \alpha^2 + \beta^2).$$

We want to show that $\mathrm{Var}(C) = 1$. Since scaling the coefficients by a positive factor does not change the sign variations, we can consider the scaled polynomial

$$C(-2\alpha x) = B_0(-2\alpha x)(4\alpha^2 x^2 + 4\alpha^2 x + \alpha^2 + \beta^2) = B_0(-2\alpha x)4\alpha^2(x^2 + x + \lambda),$$

where $\lambda := (\alpha^2 + \beta^2)/4\alpha^2$, which has the same sign variations as C since $-2\alpha > 0$. The scaling helps us to simplify the quadratic factor slightly. Let $b_m, \ldots, b_0$ be the coefficients of $4\alpha^2 B_0(-2\alpha x)$. Since $4\alpha^2 B_0(-2\alpha x)$ and $B_0(x)$ have the same sign variation, let j be the index such that

$$b_0, \ldots, b_j \le 0 \text{ and } b_{j+1}, \ldots, b_m \ge 0. \tag{2}$$

Then the coefficients of $C(-2\alpha x)$ are of the form

$$c_k = b'_{k-2} + b'_{k-1} + \lambda b'_k, \quad k = 0, \ldots, m+2$$

where $b'_{-2} = b'_{-1} = b'_{m+1} = b'_{m+2} := 0$. From (3) it follows that

$$c_0, \ldots, c_j \le 0 \text{ and } c_{j+3}, \ldots, cm + 2 \ge 0.$$

Thus to show that $\mathrm{Var}(C) = 1$ it suffices to show that $c_{j+1} \le c_{j+2}$ (independent of their signs). Now

$$c_{j+1} = b'_{j-1} + b'_j + \lambda b'_{j+1} \le b'_j + \lambda b'_{j+1}$$

as $b_{j-1} \le 0$. Similarly,

$$c_{j+2} = b'_j + b'_{j+1} + \lambda b'_{j+2} \ge b'_j + b'_{j+1}$$

since $b'_{j+2} \ge 0$. Thus $c_{j+1} \le c_{j+2}$ if $\lambda \le 1$, i.e., if $\beta^2 \le 3\alpha^2$, which is the same as saying that argument of the two conjugate roots $\alpha \pm i\beta$ is in the range $\pi \pm \pi/3$. An inductive argument shows that multiplying all the remaining quadratic factors does not increase the sign variations as long as $\beta_j^2 \le 3\alpha_j^2$, $j = 1, \ldots, k$. Thus we have the following result.

THEOREM 3 (Obreshkoff Special Case). *If B has only one positive real root and all the other roots have an argument in the range $\pi \pm \pi/3$ then* $\mathrm{Var}(B) = 1$. *Let $\mathcal{C}$ be the conical region of $\mathbb{C}$ containing points with argument in the range $\pi \pm \pi/3$.*

Let B be the polynomial obtained from by transforming A according to (2). Theorem 3 and Theorem 4 tell us constraints on the geometry of roots of a polynomial B in $\mathbb{C}$ such that $\mathrm{Var}(B) = 1$. However, to get the termination criteria for the Descartes method we have to transform these constraints to the roots of A How should the roots of A be such that the roots of B satisfy the constraints in Theorem 3 and Theorem 4? This reduces to how the map $\phi(z) := (az + b)/(z + 1)$ transforms the cone $\mathcal{C}$ and the negative half-plane of $\mathbb{C}$. This is illustrated in Figure 1, which shows the following mappings of ϕ: the imaginary axis is to the circle C_{ab} with $[a, b]$ as diameter; the upper ray of $\mathcal{C}$ is mapped to the circumscribing circle $\underline{C}_{ab}$ of the equilateral triangle with $[a, b]$ as base and above the real-axis; similarly, the lower ray of $\mathcal{C}$ is mapped to the circumscribing circle $\overline{C}_{ab}$ of the equilateral triangle with $[a, b]$ as base and lying below the real-axis; thus $\mathcal{C}$ is mapped to $\mathbb{C} \setminus \overline{C}_{ab} \cup \underline{C}_{ab}$. Note that $C_{ab} \subset \overline{C}_{ab} \cup \underline{C}_{ab}$. These observations, along with Theorem 3 and Theorem 4, give us the following termination criterion:

THEOREM 4. *If $\overline{C}_{ab} \cup \underline{C}_{ab}$ contains at most one root of A then* $\mathrm{Var}(A; a, b) \leq 1$.

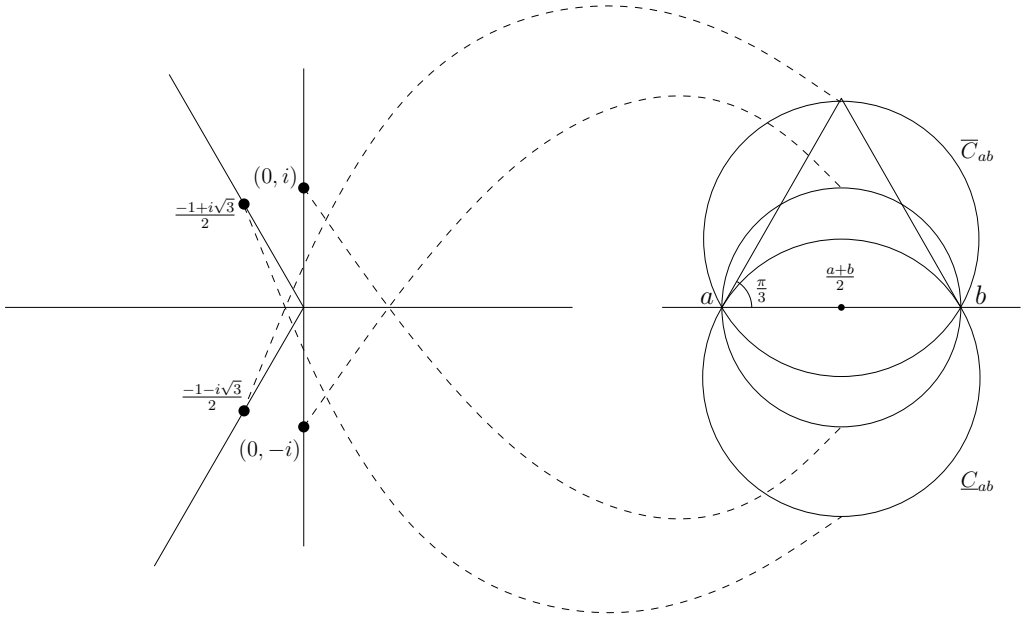


Figure 1

2 Size of the Subdivision Tree

Let $\mathcal{T}$ be the subdivision tree. We do the standard trick of pruning the leaves to obtain the tree $\mathcal{T}$. Let I be the interval associated with some leaf of $\mathcal{T}'$. Since I was non-terminal in $\mathcal{T}$, it it follows from Theorem 5 that $\overline{C}_I \cup \underline{C}_I$ contains a pair of roots α_I, β_I. Thus $2w(I) \geq |\alpha_I - \beta_I|$. Continuing the argument as we had done in earlier lectures, we obtain that

$$|\mathcal{T}'| \leq n \log w(I_0) - \log \prod_I |\alpha_I - \beta_I|.$$

We can again apply the DMM bound to derive an upper bound on $|\mathcal{T}'|$. But that is not sufficient. In Sturm's method we were sure that no two pairs corresponding to different intervals overlap. However, now the regions $\overline{C}_I \cup \underline{C}_I$ overlap, and depending on how many regions overlap, the pair α_I, β_I will have to account for all the intervals I that share this pair. So we have to derive an upper bound on how many intervals I can have their two-disc regions overlapping. Clearly, the two-discs corresponding to two neighboring intervals overlap, but can the regions of two non-neighboring discs overlap?

Let I and J be the intervals associated with two nodes in the subdivision tree, such that $w(I) \geq w(J)$; thus J is deeper in the tree than I; see Figure 2 for an illustration of the subdivision tree. If $J \subseteq I$ or J and I share an endpoint then it is clear that their two-circles overlap. We claim that these are the only cases when this can happen. For sake of simplicity let's assume that J is to the right of I. The interval between I and J, is partitioned by the intervals J' associated with the leaves appearing to the right of I and to the left of J. If J' is at a smaller depth than J then we subdivide J' till we partition it into intervals J'' of the same width as J; if, however, J' is deeper than J, then we go to J's ancestor J'' that is at the same depth as J, i.e., collapse the subdivision tree rooted at J''. What this subdivision and collapsing ensures is that the interval between I and J is partitioned by intervals of the form J'' all of them having the same width as J. Let I'' be the neighbour of I to the right in this partitioning. Then from Figure 2 it is clear that the two-circle figure corresponding to I is to the left of the equilateral triangles corresponding to I'', whereas the two-circle figure for J is to the right of these equilateral triangles, and hence the two two-circle figures cannot intersect. This implies that a pair of roots α_I, β_I corresponding to a leaf can be shared by at most one of its neighbours. Thus attributing two intervals to each pair of roots in the worst case, we have the following bound on the size of the tree

$$|\mathcal{T}'| \leq n \log w(I_0) - 2 \log \prod_I |\alpha_I - \beta_I|.$$

Now, we can apply the DMM bound on the RHS to get the following bound in the case of A is an integer polynomial with L-bit coefficients:

$$|\mathcal{T}'| = O(nL + n^2).$$

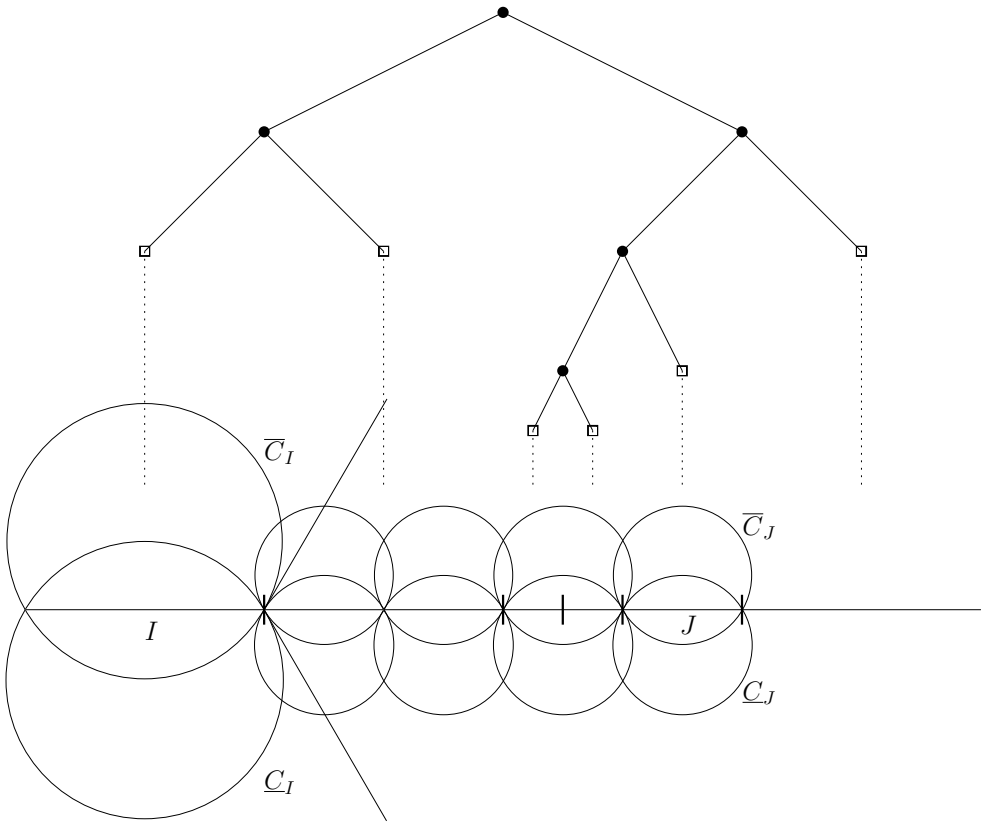


Figure 2

3 Bernstein Basis Version of the Descartes Method

A more geometric interpretation of the Descartes method is obtained by choosing the Bernstein basis instead of the monomial basis for representing A. The **Bernstein basis w.r.t. an interval** $I = [a, b]$ is defined as

$$B_i^n(x; a, b) := \binom{n}{i} \left(\frac{(x-a)}{(b-a)}\right)^i \left(\frac{(b-x)}{(b-a)}\right)^{n-i}, \quad i = 0, \ldots, n. \tag{3}$$

The **control polygon**, $\mathcal{P}$, of a polynomial $A = \sum_{i=0}^n a_i' B_i^n(x; a, b)$ is the piecewise linear polygon obtained by joining the points $(a_i, a + (b-a)i/n)$, $i = 0, \ldots, n$, by straight line segments. Given the "local nature" of the representation, a Bernstein basis representation of a polynomial has many nice geometric properties.

Prop. 1. $A(a) = a_0$ and $A(b) = a_n$.

Prop. 2.

$$A'(x; a, b) = \frac{n}{b-a} \sum_{i=0}^{n-1} (a_{i+1}' - a_i') B_i^{n-1}(x; a, b).$$

Prop. 3. The polynomial is contained in the convex hull of the control points. An even tighter estimate on the neighborhood of the control polygon that contains the polynomial is given by the following bound:

$$\max_{x \in [a,b]} |A(x) - \mathcal{P}(x)| \leq \frac{d}{8}(b-a)^2 \max_{0<i<n} |a_{i+1}' - 2a_i' + a_{i-1}'|.$$

More surprising is the fact that the number of intersections of the control polygon with the interval $[a, b]$ is an upper bound on the number of real roots of A in $[a, b]$. This easily follows if we substitute Jacobi's little observation (2) into the Bernstein representation of A:

$$\begin{aligned} B := (x+1)^n A\left(\frac{ax+b}{x+1}\right) &= (x+1)^n \sum_{i=0}^n a_i \binom{n}{i} (b-a)^{-n} \left(\frac{ax+b}{x+1} - a\right)^i \left(b - \frac{ax+b}{x+1}\right)^{n-i} \\ &= (x+1)^n \sum_{i=0}^n a_i \binom{n}{i} (b-a)^{-n} \left(\frac{(b-a)}{x+1}\right)^i \left(\frac{(b-a)x}{x+1}\right)^{n-i} \\ &= \sum_{i=0}^n a_i \binom{n}{i} x^{n-i}. \end{aligned}$$

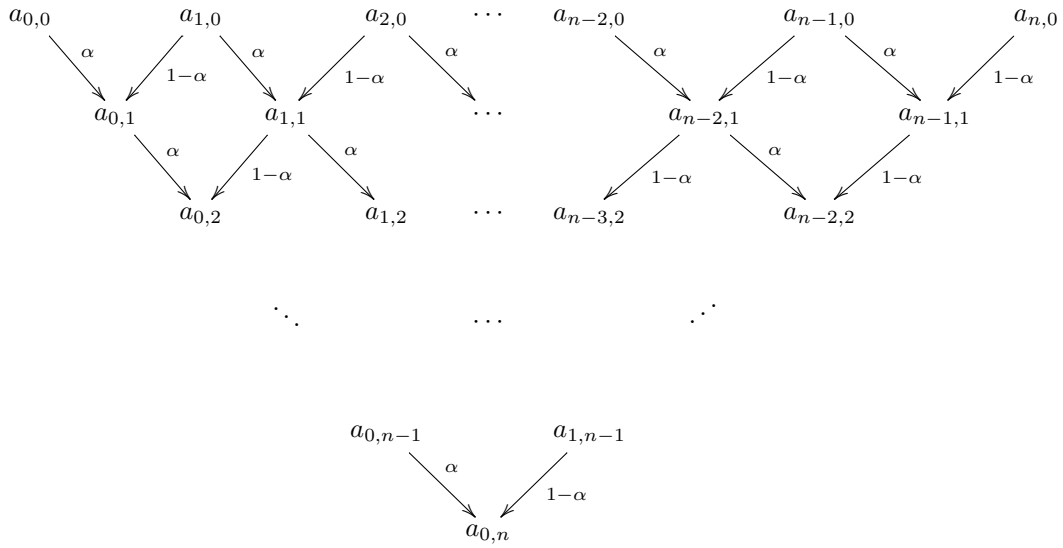
Thus $\mathrm{Var}(B)$ is nothing but the number of times the control polygon intersects the interval $[a, b]$. Since $\mathrm{Var}(B) \geq N_A(a, b)$, exceeding by an even number, we get the following:

> THEOREM 5 (Descartes's rule of Signs in Bernstein Basis). *The sign changes in the Bernstein basis of A w.r.t. an interval I exceeds the number of roots of A in I by an even number, where the roots are counted with multiplicities.*

Now that we have this nice observation, how do we implement the Descartes method in this basis? One approach is to switch to monomial basis and carry out the algorithm there. However, that is not necessary, and in some sense ugly as the conversion destroys the geometric interpretation. To carry out the method, what we need is a proceudre that takes as unput the bernstein coefficients of A w.r.t. $[a, b]$ and outputs the bernsteing coefficients of A w.r.t. to the two intervals $[a, m]$ and $[m, b]$, where $m = (a+b)/2$. The procedure we describe, in fact, computes the bernstein coefficients for any choice of $m \in [a, b]$.

`de Casteljau's Algorithm`
INPUT: Bernstien coefficients of A, $a_0, \ldots, a_n$, w.r.t. $[a, b]$, and $m \in [a, b]$.
OUTPUT
1. Define $\alpha := (m-a)/(b-a)$. $\triangleleft$ *For the midpoint, $\alpha = 1/2$.*
2. $(a_{0,0}, a_{1,0}, \ldots, a_{n,0}) \leftarrow a_0, \ldots, a_n$.
3. For i from 1 to n do
4. For j from 0 to $n-i$ do
$a_{j,i} \leftarrow \alpha a_{j,i-1} + (1-\alpha) a_{j+1,i-1}$.
5. Output the sequences $\{a_{0,i}\}$ and $\{a_{n-i,i}\}$, where $i = 0, \ldots, n$, as the Bernstein coefficients for the intervals $[a, m]$ and $[b, m]$ resp.

A usual way to depict the algorithm above is the following pictorical way: the coefficients on the left edge are the Bernstein coefficients w.r.t. $[a, m]$ and the coefficients on the right-edge are the bernstein coefficients w.r.t. $[m, b]$.



The correctness of the algorithm follows from ....Note that $a_{0,n} = A(m)$.

de Casteljau's algorithm gives us a better understanding of the effect of subdivision on the sign variations.

THEOREM 6 (Variation Diminishing Property). *Given an interval $[a, b]$ and a point $m \in [a, b]$,*

$$\mathrm{Var}(A; a, b) = \mathrm{Var}(A; a, m) + \mathrm{Var}(A; m, b) + \textit{ even number.}$$

The even number contains in it the multiplicity of m as a root of A.

Proof. The proof is by induction. Consider change in the sign variations between the row $a_{i,0}$, $i = 0, \ldots, n$, and the interleaved sequence

$$S_0 := (a_{0,0}, a_{0,1}, a_{1,0}a_{1,1}, a_{2,0}, \ldots, a_{n-1,1}, a_{n-1,0}a_{n,0}) .$$

What happens we start dropping elements of the first row from S? We claim that the sign variation of the trapezoidal sequence

$$S_1 := (a_{0,0}, a_{0,1}, a_{1,1}, , \ldots, a_{n-1,1}, a_{n,0})$$

so obtained is smaller by an even number. Suppose we have three numbers $a, b, c \in \mathbb{R}$, we claim that removing b can only drop the sign variation by two. There are three cases to consider: if $\mathrm{Var}(a, b, c) = 0$ then removing b doesn't change the sign variations; if $\mathrm{Var}(a, b, c) = 1$, then $a \cdot c < 0$, so removing b doesn't change the sign variations again; if $\mathrm{Var}(a, b, c) = 2$, then $a \cdot c > 0$ and so removing b drops the sign variations by two. **Q.E.D.**

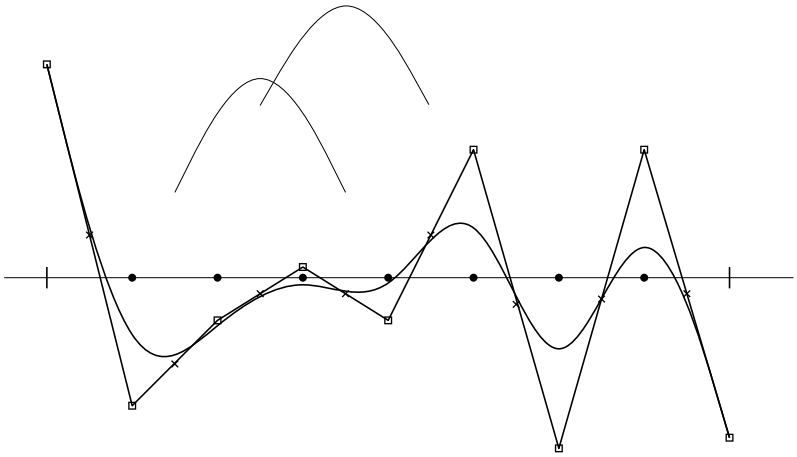


Figure 3: Variation Diminishing Property and the Bezier control polygons after subdivision.

4 Generalizing Descartes's rule of Signs – Functions

We have seen that both the monomial and Bernstein basis have the following property: given a linear combination of the basis with the vector $(a_0, \ldots, a_n)$ there is an associated interval ($(0, \infty)$ for monomial basis and the underlying interval for the Bernstein basis) such that the number of real roots, counting with multiplicities, of the linear combination in the associated interval doesn't exceed $\mathrm{Var}(a_0, \ldots, a_n)$, and differs by an even number. In general, we can ask what properties should a sequence of functions $(\phi_0(x), \ldots, \phi_n(x))$ satisfy such that there is a suitable Descartes's rule of signs for them. More precisely, we say that the **sequence of functions** $\phi_0, \ldots, \phi_n$ **satisfy Descartes's rule of sign (drs) w.r.t an interval** (a, b) if for all linear combinations $f := \sum_{i=0}^{n} a_i \phi_i$, where some a_i are non-zero, the number of roots of f in the *open* interval (a, b) is smaller than $\mathrm{Var}(a_0, \ldots, a_n)$ (note that we are not imposing the condition that the excess should be an even number). In short, we call the sequence $(\phi_0, \ldots, \phi_n)$ and I a **Descartes System**. We start with simple cases and try to reverse engineer the properties, before answering the question in full generality.

Suppose $n = 0$ and $\phi(x)$ is our function. Clearly, the choices of f are just scalings of $\phi(x)$ by some constant. As we have only one function, the drs states that $a\phi(x)$, where $a \neq 0$, has zero roots in the interval of interest. In particular, $\phi(x)$ should have no roots in the interval of interest. Thus for $n = 0$, it is necessary that if $\phi(x)$ has no roots in an interval I. We claim that this is sufficient as well, i.e., if $\phi(x)$ has no roots in I then $\{\phi(x)\}$ satisfies the drs wrt I. This is easy to see since $a\phi(x)$, for $a \neq 0$, has no roots and no sign variations. Let's see what happens when $n = 1$.

Suppose we are given ϕ_0, ϕ_1, and an interval I and we want to figure out the conditions on them such that the sequence (ϕ_0, ϕ_1) satisfies the drs wrt I. Let us assume that they satisfy the drs wrt I. Suppose $f = a\phi_0 + b\phi_1$, where $a, b \in \mathbb{R}$. If either a or b is zero then we have no sign variation and hence both ϕ_0 and ϕ_1 should have no roots in I. Moreover, they should have the same sign on I, as otherwise the linear combination $a\phi_0 + b\phi_1$, where $ab > 0$, can have roots whereas we have no sign variations. Thus our first property is that ϕ_0 and ϕ_1 do not vanish on I and in fact have the same sign. With this condition it is clear that the only interesting linear combinations are $a\phi_0 - b\phi_1$, where $ab > 0$; since $a, b \in \mathbb{R}_{\neq 0}$, we can wlog assume that $a = 1$. Now the drs states that for all $b \in \mathbb{R}$, $\phi_0(x) = b\phi_1(x)$ has at most one solution $x \in I$. Or in other words, for all $b \in \mathbb{R}$, the function $g(x) := \phi_0(x)/\phi_1(x)$ equals b at most one $x \in I$; note that as the denominator does not vanish on I the function g is well-defined and continuous on I. This is true iff $g(x)$ is monotone on I, i.e., $g'(x) \neq 0$ for all $x \in I$, which is equivalent to the statement that $(\phi_0\phi_1' - \phi_0'\phi_1)(x)$ has no root in I. Thus we have shown that if the sequence (ϕ_0, ϕ_1) satisfies the drs wrt I then the following properties are necessary:

P1. ϕ_0 and ϕ_1 do not vanish on I and in fact have the same sign, and

P2. $(\phi_0\phi_1' - \phi_0'\phi_1)(x)$ has no root in I.

The converse is also immediate from the argument above. To obtain the correct generalization to higher dimensions, we first massage the second property. The second property is equivalent to the statement that

$$\det \begin{bmatrix} \phi_0 & \phi_1 \\ \phi_0' & \phi_1' \end{bmatrix}$$

does not vanish on I. But this is equivalent to saying that for all $\alpha \in I$ there is no linear combination f such that $f(\alpha) = f'(\alpha) = 0$, i.e., α is a double root of f. This last property can be easily generalized to any n, and gives us the motivation for the next result.

A result of Pólya and Szegö gives a characterization, but to describe that result we first need the following definition: Given $(n+1)$ functions $\phi_0, \ldots, \phi_n$ that have continuous derivatives of order n on an interval I, their **Wronskian** is defiend as follows:

$$W(\phi_0, \ldots, \phi_n) := \det \begin{bmatrix} \phi_0 & \phi_1 & \phi_2 & \cdots & \phi_n \\ \phi_0^{(1)} & \phi_1^{(1)} & \phi_2^{(1)} & \cdots & \phi_n^{(1)} \\ \phi_0^{(2)} & \phi_1^{(2)} & \phi_2^{(2)} & \cdots & \phi_n^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \phi_0^{(n)} & \phi_1^{(n)} & \phi_2^{(n)} & \cdots & \phi_n^{(n)} \end{bmatrix}. \tag{4}$$

Given k, $0 \le k \le n$, and indices $0 \le i_0 < \cdots < i_k \le n$, define $W(\phi_{i_0}, \ldots, \phi_{i_k})$ as the determinant of the $(k+1) \times (k+1)$ matrix formed by picking the first $k+1$ rows and the columns $i_0, \ldots, i_k$. A sequence of n-times differentiable functions $(\phi_0, \ldots, \phi_n)$ and an interval I is said to form a **Wronskian System** if the following conditions are met:

1. For all $k \in \{0, \ldots, n\}$ and integers $i_0, i_1, \ldots, i_k$, such that $0 \le i_0 \le i_1 < i_2 < \cdots < i_k \le n$, the Wronskian $W(\phi_{i_0}, \phi_{i_1}, \ldots, \phi_{i_k})$ is not zero for all $x \in (a, b)$, and
2. Wronskians with the same number of rows have the same sign.

In the next theorem we need the following property of determinants. Let A be an $n \times n$ matrix. Then from Laplace expansion along the ith row we know that

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} M_{ij} a_{ij},$$

where M_{ij} is the (i,j)th minor, i.e., the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting the ith row and jth column. Suppose we replace the ith row by any other row k in A to obtain a new matrix A'. Then as A' has a repeated row its determinant is zero, but applying the Laplacian expansion along the ith row of A' we obtain that

$$\sum_{j=1}^{n} (-1)^{i+j} M_{ij} a_{kj} = 0, \tag{5}$$

where $k \neq i$. In other words, the kth row is orthogonal to the vector of cofactors along any other row.

We will also need the following observation: given two differentiable functions, g, h, the formal derivative

$$\left(\frac{g}{h}\right)^{(j)} = \sum_{k=0}^{j} \binom{j}{k} (-1)^k \frac{g^{(k)}}{h} \Delta^k(h), \tag{6}$$

where $\Delta^k(h)$ is defined as follows:

$$\Delta^k(h) = \sum_{\pi \in \Pi(n)} s_\pi \frac{\prod_{\ell=1}^{|\pi|} h^{(\pi_\ell)}}{h^{|\pi|}}.$$

Here $\Pi(n)$ is the set of all partitions of n, and $s_\pi = \pm 1$ depending on the partition. So Δ^1 has only one term, namely h'/h; Δ^2 has two terms corresponding to the two partitions $(1,1)$ and 2, namely $(h'/h)^2$ and

h''/h; and Δ^3 has three terms corresponding to the partitions $(1,1,1),(1,2),(3)$. For instance, see these two formulas

$$\left(\frac{g}{h}\right)^{(3)} = \frac{g'''}{h} - 3\frac{g''}{h}\frac{h'}{h} + 3\frac{g'}{h}\left(2\left(\frac{h'}{h}\right)^2 - \frac{h''}{h}\right) - 3\frac{g}{h}\left(2\left(\frac{h'}{h}\right)^3 - \frac{h'h''}{h^2} - \frac{h'''}{h}\right)$$

$$\left(\frac{g}{h}\right)^{(2)} = \frac{g''}{h} - 2\frac{g'}{h}\frac{h'}{h} + \frac{g}{h}\left(\left(\frac{h'}{h}\right)^2 - \frac{h''}{h}\right).$$

The crucial interpretation of (8) for us is that $g^{(j)}/h$ can be expressed as a suitable linear combinatios of $g^{(j-1)}/h, \ldots, g/h$.

THEOREM 7. *A sequence of functions $\phi_0, \ldots, \phi_n$ and an interval I forms a Descartes system iff they form a Wronskian system.*

Proof. One direction is relatively easy to show, namely if the drs holds for the system wrt I then the two conditions must hold. The first condition should hold because if $W(\phi_{i_0}, \phi_{i_1}, \ldots, \phi_{i_k}) = 0$, then there exists $a_{i_0}, a_{i_1}, \ldots, a_{i_k}$, not all zero, such that $f := \sum_{j=0}^{k} a_{i_j}\phi_{i_j}$ has a root of multiplicity $k+1$, whereas it can only have at most k sign variations, giving us a contradiction.

As for the second property, we start by showing that all the $n \times n$ determinants, i.e., those corresponding to choosing $k = n-1$, have the same sign. For an $\alpha \in I$, let D_i be the determinant obtained by dropping the $(i+1)$th column, $i = 0, \ldots, n$, and let $W := W(\phi_0, \ldots, \phi_n)$. Then applying (7), where $i = (n+1)$ is the last row, we obtain that

$$\begin{bmatrix} \phi_0 & \phi_1 & \phi_2 & \cdots & \phi_n \\ \phi_0^{(1)} & \phi_1^{(1)} & \phi_2^{(1)} & \cdots & \phi_n^{(1)} \\ \phi_0^{(2)} & \phi_1^{(2)} & \phi_2^{(2)} & \cdots & \phi_n^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \phi_0^{(n)} & \phi_1^{(n)} & \phi_2^{(n)} & \cdots & \phi_n^{(n)} \end{bmatrix} \cdot \begin{bmatrix} D_0 \\ -D_1 \\ D_2 \\ \vdots \\ (-1)^n D_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ (-1)^n W \end{bmatrix}.$$

This implies that α is a root of multiplicity n of

$$f = D_0\phi_0(x) - D_1\phi_1(x) + D_2\phi_2(x) - \cdots + (-1)^n D_n\phi_n(x),$$

and hence all the D_j's have the same sign, as otherwise drs would be violated. To show the second property for $k = n-2$ we apply the claim inductively, and use the fact that the matrices corresponding to D_i's overlap considerably.

We now show the claim of sufficiency. This proof is similar to the proof of the drs by induction, except the induction is on n, the number of functions in the linear combination. We have already seen the proof of the base cases when $n = 0, 1$, so assume that the claim holds for $< n$, i.e., any Wronskain system with fewer than n functions is also a Descartes system. Suppose

$$f = a_0\phi_0 + \cdots + a_n\phi_n$$

and i is the first index where a sign change occurs; we can assume that all the a_i's are non-zero, otherwise the claim follows from the induction hypothesis; also, if there is no sign change in the a_i's, then as ϕ_i's have the same sign on I, the function f has no roots as well. Let μ be the number of roots of f and σ be the number of sign changes in $a_0, \ldots, a_n$. Consider the function

$$\frac{f}{\phi_i} = a_0\frac{\phi_0}{\phi_i} + \cdots + a_{i-1}\frac{\phi_{i-1}}{\phi_i} + a_i + a_{i+1}\frac{\phi_{i+1}}{\phi_{i+1}} + \cdots + a_n\frac{\phi_n}{\phi_i},$$

which is well-defined on I. Differentiating both sides we get that

$$\left(\frac{f}{\phi_i}\right)' = a_0\left(\frac{\phi_0}{\phi_i}\right)' + \cdots + a_{i-1}\left(\frac{\phi_{i-1}}{\phi_i}\right)' + a_{i+1}\left(\frac{\phi_{i+1}}{\phi_i}\right) + \cdots + a_n\left(\frac{\phi_n}{\phi_i}\right)'.$$

Define

$$\psi_j := \begin{cases} (\phi_j/\phi_i)', \text{ for } i = 0, \ldots, j-1, \\ (\phi_{j+1}/\phi_i)', \text{ for } i = j, \ldots, n-1. \end{cases}$$

We claim that $-\psi_0, \ldots, -\psi_{i-1}, \psi_i, \ldots, \psi_{n-1}$ and I forms a Wronskian system as well. More precisely, we claim that for $0 \le k \le n-1$, and indices $0 \le i_0 < \cdots < i_k \le n$

$$W(\psi_{i_0}, \ldots, \psi_{i_k}) = \begin{cases} (-1)^{k+1} W(\phi_{i_0}, \ldots, \phi_{i_k}, \phi_i), \text{ if } i_k < i, \\ W(\phi_i, \phi_{i_0+1}, \ldots, \phi_{i_k+1}), \text{ if } i \le i_0, \text{ and} \\ (-1)^{j+1} W(\phi_{i_0}, \ldots, \phi_{i_{j-1}}, \phi_i, \phi_{i_{j+1}+1}, \ldots, \phi_{i_k+1}), \text{ if } i_0 < i = i_j \le i_k. \end{cases} \tag{7}$$

To prove this equation, pick any rhs and reduce the ith row wrt to the rows above it based on the relation given in (8); this should be done starting from the bottom most row going down to the first row. The column corresponding to ϕ_i is transformed to a one followed by zeros. Then do Laplace's expansion of the determinant along this column. The power of (-1) comes from this expansion, and matches the number of columns to the left of ϕ_i. Therefore, it can be distributed over the appropriate ψ's on the lhs. Thus all the $k \times k$ wroskians of the sequence $\psi_0, \ldots, \psi_{n-1}$ have the same sign, namely the sign of the corresponding $(k+1) \times (k+1)$ wronskian for $\phi_0, \ldots, \phi_n$ where a column corresponding to ϕ_i always appears.

From the induction hypothesis, we have that the sequence $(-\psi_0, \ldots, -\psi_{i-1}, \psi_i, \ldots, \psi_{n-1})$ forms a Descartes system wrt I. Therefore, applying drs to the function

$$\frac{d}{dx}\frac{f}{\phi_i} = -a_0(-\psi_0) + \cdots + -a_{i-1}(-\psi_{i-1}) + a_{i+1}\psi_i + \cdots + a_n\psi_{n-1}$$

we obtain that its number of roots μ' does not exceed its number of variations σ'. The roots of $(f/\phi_i)'$ are the critical points of f/ϕ_i, but the roots of latter are the same as roots of f. Since between two consecutive roots of f/ϕ_i there is at least one critical point, the number of critical points are at least $\mu - 1$. Furthermore, the sign variation σ' in the sequence $(-a_0, \ldots, -a_{i-1}, a_{i+1}, \ldots, a_n)$ is the same as that in $(-a_0, \ldots, -a_{i-1}, a_i, a_{i+1}, \ldots, a_n)$ (as $a_{i-1}a_i < 0$ and $a_0, \ldots, a_{i-1}$ have the same sign), and hence one less than that in the original coefficient sequence, hence $\sigma' = \sigma - 1$. Applying the induction hypothesis to $(f/\phi_i)'$, we obtain that $\mu - 1 \le \mu' \le \sigma' = \sigma - 1$, which implies $\mu \le \sigma$ as desired.

Q.E.D.

As a consequence of the theorem above, we can check that the drs works for the set $\left\{e^{\lambda_1 x}, \ldots, e^{\lambda_n x}\right\}$, for distinct λ_i's.

5 Generalizing Descartes's rule of Signs – Complex Plane

One way to interpret drs is that the number of real roots of a polynomial is bounded by the number of non-zero monomials in the polynomial and is independent of the degree. In particular, this implies that for sparse polynomials the number of real roots is linear in the input size. Recall that for a sparse polynomial $f = \sum_{i=0}^{k} a_i x^{e_i}$, where $e_i \in \mathbb{N}$ and $a_i \in \mathbb{R}$, the size of the input is $k \max\{\log|a_i|, \log e_i\}$. But is this property restricted to $\mathbb{R}$, or can we say something about the non-real roots as well? In this section, we see such a generalization due to Hayman. The results roughly state that the *arguments* of the roots of a sparse polynomial are almost uniformly distributed in the interval $[0, 2\pi]$. What do we mean by uniform? Consider the polynomial $x^n - 1$. We know that the roots of the polynomial are $\exp(2\pi ik/n)$, where $k = 0, \ldots, n-1$. So the number of roots with arguments in the range $[\theta, \omega]$ should be $\lfloor n(\omega - \theta)/2\pi \rfloor$. The next result that we see claims that for a sparse polynomial the number of roots with arguments in the range $[\theta, \omega]$ is not much larger than this quantity, or the discrepancy is bounded as some nice function of the input.

To describe this result, we need the following generalization of the result earlier for boxes.

THEOREM 8. *Let $\gamma : [a, b] \to \mathbb{C}$ be a simple (not self-intersecting) arc. If f is a polynomial having no roots along γ, then define*

$$F(z) := \frac{\mathtt{Im}(f \circ \gamma(z))}{\mathtt{Re}(f \circ \gamma(z))}.$$

Then the change in the argument of f along γ satisfies the following relation:

$$|\Delta_\gamma(f) - \pi \cdot I_a^b F| < \pi,$$

where the RHS is the Cauchy index of F on the interval $[a,b]$, and we assume that $\gamma(a), \gamma(b)$ are not poles of F.

Proof. Let $a < t_1 < \cdots < t_k < b$ be the poles of F, then the change in argument is

$$\Delta_\gamma(\arg f) = \Delta_{\gamma[a,t_1]}(\arg f) + \pi \sum_{i=1}^{k-1} \frac{\texttt{sign}(F(t_{i+1}^-)) - \texttt{sign}(F(t_i^+))}{2} + \Delta_{\gamma[t_k,b]}(\arg f). \tag{8}$$

The intermediate sum on the rhs is almost the cauchy index, except it is missing the term $\texttt{sign}(F(t_1^-))$ and $\texttt{sign}(F(t_k^+))$. We claim that

$$\Delta_{\gamma[a,t_1]}(\arg f) = \pi \frac{\texttt{sign}(F(t_1^-))}{2}(1 \pm \epsilon), \tag{9}$$

where $0 \le \epsilon < 1$. There are two cases to consider:

1. if $\texttt{sign}(F(t_1^-)) \cdot \texttt{sign}(F(a)) > 0$, then both $F(t_1^-)$ and $F(a)$ are in the same quadrant and hence the change in argument is $(\pi/2)\texttt{sign}(F(t_1^-))(1-\epsilon)$, $0 \le \epsilon \le 1$;
2. if $\texttt{sign}(F(t_1^-)) \cdot \texttt{sign}(F(a)) < 0$, then $F(t_1^-)$ and $F(a)$ are in quadrants with opposite signs (but on the same side of the imaginary axis), and hence the change in argument is

$$\frac{\pi}{2}\texttt{sign}(F(t_1^-)) - \epsilon\frac{\pi}{2}\texttt{sign}(F(a)) = \frac{\pi}{2}\texttt{sign}(F(t_1^-))(1+\epsilon).$$

A similar argument shows that

$$\Delta_{\gamma[t_k,b]}(\arg f) = -\pi \frac{\texttt{sign}(F(t_k^+))}{2}(1 \pm \epsilon'), \tag{10}$$

where $0 \le \epsilon' < 1$. Substituting (11) and (12) in (10), along with the definition of Cauchy index, we obtain that

$$\Delta_\gamma(\arg f) = \pi I_a^b F \pm \frac{\pi}{2}(\epsilon + \epsilon'),$$

which implies the desired inequality as $(\epsilon + \epsilon') < 2$. [1] **Q.E.D.**

Remark: We can let $a = -\infty$ and $b = +\infty$ in the argument above, as long as we have finitely many poles of F in the interval $[-\infty, \infty]$, and $F(\gamma(t))$, for $t \to \pm\infty$, does not approach the imaginary axis.

In our case, we want to count the roots of f in a sector of the complex plane: Given angles $\theta, \omega \in [0, 2\pi]$, were $\theta < \omega$, the **sector** $S(\theta, \omega)$ is defined as

$$S(\theta, \omega) = \{z \in \mathbb{C} : \theta < \arg(z) < \omega\}. \tag{11}$$

So the number of roots of f with arguments in a certain range is the same as the number of roots of f within a certain sector. Therefore, we want to bound the number of roots of a polynomial within a given sector. By definition the sector is unbounded, but we know that the roots of f are in some bounded disc $D(0,R)$, which means that we can restrict our attention to $S_R(\theta,\omega) := S(\theta,\omega) \cap D(0,R)$. The advantage of considering a bounded region is that we can use the argument principle to find the number of roots in $S_R(\theta,\omega)$. Wlog, let us assume that $a_0 := f(0) \neq 0$, and in fact $\texttt{Re}(a_0) = 0$ (if not then multiply f with a suitable constant c such that $\texttt{Re}(ca_0) = 0$). We will say that a sector $S_R(\theta,\omega)$ is **regular wrt a polynomial** f, if there are no roots of f on the boundary of $S_R(\theta,\omega)$ and

$$\texttt{Re}(a_n e^{in\theta}) \neq 0 \text{ and } \texttt{Re}(a_n e^{in\omega}) \neq 0. \tag{12}$$

[1] Our definition of Cauchy index is negative of the standard definition in the literature. For us $I_a^b F$ of F on the interval $[a,b]$ is the number of poles across which the sign jumps from $+\infty$ to $-\infty$ minus the number of poles across which the sign jumps from $-\infty$ to $+\infty$.

The significance of these constraints will become clear later on, but basically we need them so that we can apply theorem 11 to count the roots of f in $S_R(\theta,\omega)$. To apply Theorem 11, we break the boundary of $S_R(\theta,\omega)$ as follows. Define

$$\gamma(t) := \begin{cases} -te^{i\omega} & t \in [-R, 0], \\ te^{i\theta} & t \in [0, R]. \end{cases} \tag{13}$$

So as $t \in [-1, 1]$, $\gamma(t)$ traces a counter clockwise direction along the two edges of the sector. For the arc, define $\delta(t) := Re^{it}$, $t \in [\theta, \omega]$. Thus $\partial\Omega$ is the union of $\gamma(t)$ with $\delta(t)$ as shown in Figure **??**. The change in argument along $\gamma(t)$ is given by Theorem 11:

$$\Delta_\gamma(\arg f) = \pi(I^0_{-R}F_\omega + I^R_0 F_\theta) \pm \pi$$

where

$$F_\omega := \frac{\texttt{Im} f(-e^{i\omega}t)}{\texttt{Re} f(-e^{i\omega}t)}, \text{ and } F_\theta := \frac{\texttt{Im} f(e^{i\theta}t)}{\texttt{Re} f(e^{i\theta}t)}.$$

What about the change in argument across $\delta(t)$? Let us express

$$f(z) = a_n z^n \left(1 + \sum_{i=0}^{n-1} \frac{a_i}{a_n z^{n-i}}\right).$$

Then across $\delta(t)$ we have

$$\Delta_\delta(\arg f) = n(\omega - \theta) + \epsilon(R)$$

where

$$\epsilon(R) := \Delta_\delta \arg \left(1 + \sum_{i=0}^{n-1} \frac{a_i}{a_n z^{n-i}}\right).$$

However, note that as $R \to \infty$ the contribution of $\epsilon(R)$ tends to zero. From these equations we obtain

$$\begin{aligned} \Delta_\Omega(\arg f) &= \Delta_\gamma(\arg f) + \Delta_\delta(\arg f) \\ &= \pi(I^0_{-R}F_\omega + I^R_0 F_\theta) + n(\omega - \theta) + \epsilon(R) \pm \pi. \end{aligned}$$

We now let $R \to \infty$, and note that we can still apply Theorem 11, since (14) implies that the denominators of F_θ and F_ω are polynomials of degree n, and hence there are only finitely many poles on $f(\gamma(t))$. Therefore, as $R \to \infty$ we have

$$\frac{1}{2\pi}\Delta_\Omega(\arg f) = \frac{1}{2}(I^0_{-\infty}F_\omega + I^\infty_0 F_\theta) + n\frac{(\omega - \theta)}{2\pi} \pm \frac{1}{2}.$$

As the error term $\pm 1/2 < 1/2$, we have the following result:

THEOREM 9.

$$\#(\textit{number of roots of } f \textit{ in } S(\theta,\omega)) = \frac{1}{2}(I^0_{-\infty}F_\omega + I^\infty_0 F_\theta) + \left\lfloor n\frac{(\omega - \theta)}{2\pi} \right\rfloor$$

For the special case where f is a sparse polynomial with k non-zero monomials, we know that both $\texttt{Re}(f(e^{i\theta}x))$ and $\texttt{Re}(f(-e^{i\omega}x))$ also have at most k non-zero monomials. Moreover, the cauchy index $I^0_{-\infty}F_\omega$ is bounded by the number of negative roots of $\texttt{Re}(f(-e^{i\omega}x))$, which by the drs is atmost k; similarly, the cauchy index $I^\infty_0 F_\theta$ is bounded by the number of positive roots of $\texttt{Re}(f(e^{i\theta}x))$, which are again at most k by drs. Therefore,

$$\#(\text{number of roots of } f \text{ in } S(\theta,\omega)) \leq k + \left\lfloor n\frac{(\omega - \theta)}{2\pi} \right\rfloor, \tag{14}$$

which means that the arguments of a sparse polynomial are almost uniformly distributed around the origin. This result generalizes drs to the complex plane, since if θ, ω are very close and $S(\theta,\omega)$ contains the $+ve$ x-axis then we have the standard drs. This result also gives us Obreskoff's result, namely f has at most k roots in the sector $(-\pi/n, \pi/n)$.

References

[1] A. Eigenwillig. *Real Root Isolation for Exact and Approximate Polynomials Using Descartes Rule of Signs*. Ph.D. thesis, University of Saarland, Saarbruecken, Germany, May 2008.

[2] Q. I. Rahman and G. Schmeisser. *Analytic Theory of Polynomials*. Oxford University Press, 2002.