Real Root Isolation for Exact and Approximate Polynomials Using Descartes' Rule of Signs

Dissertation zur Erlangung des Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

vorgelegt von

Arno Eigenwillig

Saarbrücken
2008

Tag des Kolloquiums: 15. Mai 2008

Dekan der Naturwissenschaftlich-Technischen Fakultät I:
Professor Dr. Joachim Weickert

Berichterstatter:
Professor Dr. Kurt Mehlhorn, Max-Planck-Institut für Informatik, Saarbrücken
Professor Dr. Raimund Seidel, FR Informatik, Universität des Saarlandes, Saarbrücken
Professor Chee K. Yap Ph.D., Courant Institute, New York University, New York, USA

Mitglieder des Prüfungsausschusses:
Professor Dr. Reinhard Wilhelm (Vorsitzender)
Professor Dr. Kurt Mehlhorn
Professor Dr. Raimund Seidel
Professor Chee K. Yap Ph.D.
Dr. Michael Sagraloff

Für Julia

Acknowledgements

First and foremost, I wish to express my gratitude to my thesis advisor Kurt Mehlhorn, who started the work on the bitstream Descartes method – not just as a general idea, but right into the crucial details. Moreover, it has been a great pleasure and a big advantage for me to be part of the extraordinary research environment that he has created in his group at the *Max-Planck-Institut für Informatik* in Saarbrücken, Germany.

At the institute, I have worked on various topics with my colleagues Eric Berberich, Michael Hemmer, Michael Kerber, Lutz Kettner, Joachim Reichel, Michael Sagraloff, Susanne Schmitt, and last, but certainly not least, Nicola Wolpert. I thank all of you, and our short-term visitor Werner Krandick, for the pleasant and stimulating collaboration.

My special thanks go to Vikram Sharma and Chee Yap for the fruitful cooperation on analyzing the subdivision tree of the Descartes method, and for being such superb hosts during my visit to the Courant Institute at New York University in March 2006.

Arnold Schönhage has sent me a thoughtful critique of the original publication on the bitstream Descartes algorithm [EKK$^+$05], which I gratefully acknowledge. I hope the improved presentation on the pages to follow can address some of the issues raised.

In the preparation of this thesis, I received help from several colleagues. Michael Kerber and Michael Sagraloff have read drafts, which was an invaluable service. Ralitsa Angelova and Annamária Kovács have been so kind to help me with handling the Bulgarian and Hungarian sources in Chapter 2. Marco Kuhlmann delighted me with a modification of Olaf Kummer's Doublestroke font for improved appearance on high-resolution printers. Thank you all!

The work on this thesis has been supported partially by the Research Training Group "Quality Guarantees for Computer Systems" of the German Research Foundation at Saarland University *(DFG-Graduiertenkolleg 623)*, coordinated by Raimund Seidel.

Zusammenfassung

Collins und Akritas (1976) haben das Descartes-Verfahren zur Einschließung der reellen Nullstellen eines ganzzahligen Polynoms in einer Veränderlichen angegeben. Das Verfahren unterteilt rekursiv ein Ausgangsintervall, bis die Descartes'sche Vorzeichenregel anzeigt, dass alle Nullstellen getrennt worden sind. Die partielle Umkehrung der Descartes'schen Regel nach Obreschkoff (1952) in Verbindung mit der Schranke von Mahler (1964) und Davenport (1985) führt uns auf eine asymptotisch fast scharfe Schranke für den sich ergebenden Unterteilungsbaum. Daraus folgen direkt die besten bekannten Komplexitätsschranken für die äquivalenten Formen des Descartes-Verfahrens in der Monom-Basis (Collins/Akritas, 1976), der Bernstein-Basis (Lane/Riesenfeld, 1981) und der skalierten Bernstein-Basis (Johnson, 1991), die hier vereinheitlicht dargestellt werden.

Ohne dass die Korrektheit der Ausgabe verloren geht, modifizieren wir das Descartes-Verfahren so, dass es mit „Bitstream"-Koeffizienten umgehen kann, die beliebig genau angenähert, aber nicht exakt bestimmt werden können. Wir analysieren die erforderliche Rechenzeit und Präzision. Das vom Verfasser mit Kerber/Wolpert (2007) und Kerber (2008) an anderer Stelle beschriebene Verfahren zur Bestimmung des Arrangements (der Schnittfigur) ebener algebraischer Kurven fußt wesentlich auf Varianten des Bitstream-Descartes-Verfahrens; wir analysieren einen zentralen Teil davon.

Diese Arbeit ist in englischer Sprache verfasst.

Abstract

Collins und Akritas (1976) have described the Descartes method for isolating the real roots of an integer polynomial in one variable. This method recursively subdivides an initial interval until Descartes' Rule of Signs indicates that all roots have been isolated. The partial converse of Descartes' Rule by Obreshkoff (1952) in conjunction with the bound of Mahler (1964) and Davenport (1985) leads us to an asymptotically almost tight bound for the resulting subdivision tree. It implies directly the best known complexity bounds for the equivalent forms of the Descartes method in the power basis (Collins/Akritas, 1976), the Bernstein basis (Lane/Riesenfeld, 1981) and the scaled Bernstein basis (Johnson, 1991), which are presented here in a unified fashion.

Without losing correctness of the output, we modify the Descartes method such that it can handle bitstream coefficients, which can be approximated arbitrarily well but cannot be determined exactly. We analyze the computing time and precision requirements. The method described elsewhere by the author together with Kerber/Wolpert (2007) and Kerber (2008) to determine the arrangement of plane algebraic curves rests in an essential way on variants of the bitstream Descartes algorithm; we analyze a central part of it.

This thesis is written in English.

Contents

Chapter 1

Introduction

Problem statement and motivation

This thesis is concerned with a family of algorithms that accept as input a sequence of real numbers $(a_0, \ldots, a_n)$, understood to be the coefficients of a polynomial

$$A(X) = a_n X^n + \cdots + a_2 X^2 + a_1 X + a_0, \qquad a_n \neq 0, \qquad n \geq 2, \tag{1.1}$$

and produce as output a sequence of pairwise disjoint intervals $(I_1, \ldots, I_r)$ such that r is the number of distinct real roots of $A(X)$ and each interval I_i, $1 \leq i \leq r$, contains exactly one real root of $A(X)$. This is commonly called real root isolation, and the intervals $I_1, \ldots, I_r$ are called isolating intervals for the real roots of $A(X)$.

All algorithms we consider follow the same method: A bounded initial interval I_0 is recursively subdivided, typically at interval midpoints, until Descartes' Rule of Signs (Theorem 2.2 below) indicates that each subinterval contains at most one root. The study of real root isolation by this method on digital computers starts with Collins and Akritas [CA76]. Following the contemporary research literature, we call it the Descartes method for real root isolation; references are given in §3.1.2, where we discuss this designation in more detail. Since Descartes' Rule counts real roots according to their multiplicities, it is a prerequisite for the Descartes method that the roots to be isolated are simple.

The Descartes method enjoys an excellent reputation regarding its practical performance [Joh91] [Joh98], especially when approximate arithmetic is used to accelerate the computations with the polynomial's coefficients [JK97] [CJK02] [RZ04]. However, all previous approaches to the use of approximate arithmetic needed exact arithmetic as a back-up for certain problematic inputs, see §3.3.1. This limitation is overcome by the bitstream Descartes algorithm, which has been developed by the author of this thesis in joint work with Mehlhorn et al. [EKK$^+$05], and which is presented here in a revised version. The bitstream Descartes algorithm is the first form of the Descartes method that can handle all inputs exclusively with approximate arithmetic, and thus the first that is applicable to inputs whose coefficients are "bitstreams", i.e., numbers only known through approximations with an arbitrarily small but positive absolute error. (The formal definition of a bitstream appears on page 76.)

The bitstream Descartes algorithm, in several variations, has been an important source of efficiency for the recent developments [EKW07] [EK08] [BK08] [BKS08] in Exacus,[1] a set of software libraries for exact yet efficient non-linear computational geometry [BEH$^+$05]. The use of approximate arithmetic to compute exact isolating intervals matches the Exact Geometric Computation (EGC) paradigm [Yap04b] followed by Exacus.

[1] `http://exacus.mpi-inf.mpg.de/`

Outline of this thesis

This thesis investigates the mathematical foundations and the computational complexity of the Descartes method in general, and the bitstream Descartes algorithm in particular. The Descartes method is sufficiently popular and successful in practice to justify this investigation, and to make a study of its extension to bitstream coefficients worthwhile.

Chapter 2 reviews Descartes' Rule of Signs and its generalization from $(0, \infty)$ to arbitrary open intervals. The systematic use of polar forms leads to a unified treatment of Descartes' Rule in the power and Bernstein basis. With the Bernstein basis comes de Casteljau's algorithm, whose variation-diminishing property is a powerful tool in reasoning about the Descartes method. In Appendix A.1, we use it to derive the Budan-Fourier Theorem as a corollary to Descartes' Rule, refining an argument of Schoenberg [Sch34].

The analysis of the Descartes method depends critically on partial converses of Descartes' Rule. We derive the partial converse of Obreshkoff [Obr52a] [Obr52b] in slightly generalized form. We also give an improved account of the partial converse by differentiation from [Eig07]. En route to its comparison with Obreshkoff's partial converse, we generalize a result of Dimitrov [Dim98] on the proximity of roots to roots of derivatives.

The chapter closes with a discussion of a family of bounds in terms of $|a_i/a_n|^{n-i}$ on the magnitudes of roots in the style of van der Sluis [vdS70]. Such bounds are needed for the choice of an initial interval for root isolation.

In **Chapter 3**, we introduce the general form of the Descartes method. The well-known algorithms of Collins/Akritas [CA76], Lane/Riesenfeld [LR81] and Johnson [Joh91, §4.2.2] are obtained as specializations of this general form by choosing specific bases for the representation of polynomials.

We give a new and almost tight bound on the size of the subdivision tree constructed by the Descartes method, based on the Davenport-Mahler bound. Our tree bound entails almost immediately the best known bit complexity statements for the aforementioned algorithms on polynomials of degree n with τ-bit integer coefficients, namely $O(n^5(\tau + \log n)^2)$ with classical and $O^{\sim}(n^4\tau^2)$ with asymptotically fast subdivision. These bounds on tree size and bit complexity originate from ideas by Vikram Sharma and Chee Yap that were worked out and published jointly with the author of this thesis in [ESY06], cf. [Sha07a, §2]. Here, we present a revised derivation of the tree bound.

The bitstream Descartes algorithm uses a randomized choice of subdivision points to escape from the numerically hard boundary cases that can arise for any fixed choice of subdivision points. When we allow one more level of subdivision than for exact coefficients, the magnitude of the polynomial's value, compared to the approximation precision, becomes an effective criterion for the feasibility of a subdivision point. The algorithm determines by exponential guessing a precision that suffices to make a large fraction of subdivision points feasible. We present a revised form of the algorithm that determines this precision directly, without the detour through an estimate of root separation and without the close coupling between precision and subdivision depth that existed in [EKK$^+$05].

The two key ingredients to the analysis of the bitstream Descartes algorithm are a bound on its subdivision tree, inherited from our treatment of the Descartes method for exact coefficients, and an estimate of the precision required. For the latter, we borrow a technique from Neumaier [Neu03] to remove the squarefreeness condition imposed by the original estimate in [EKK$^+$05]. The results of our analysis are stated in §3.3.8 and §3.3.9.

The last part of the thesis treats an application of the bitstream Descartes algorithm in the geometric analysis of a square-free algebraic curve $F \in \mathbb{Z}[X, Y]$ by the method of [EKW07]. A central task in this method is real root isolation on $F(\alpha, Y)$, where α is a real root of $R(X) := \mathrm{Res}(F, \mathrm{D}_Y F, Y)$, for the case that $F(\alpha, Y)$ has a unique multiple real root and that the number m of distinct real roots is known in advance. These conditions allow to adapt the Descartes method such that it can isolate the multiple real root, too. The bitstream Descartes algorithm so modified needs $O(n^9 \log n \cdot (\tau + \log n)^2)$ bit operations in total for root isolation on $F(\alpha, Y)$ at all real roots α of $R(X)$, if F has degree n and τ-bit integer coefficients. (This bound excludes the cost of coefficient approximation.) Our derivation of this result exemplifies the techniques necessary to analyze the bitstream Descartes algorithm for input polynomials with algebraic coefficients. It relies on a generalization of the Davenport-Mahler bound to non-square-free polynomials (see page 54), in which the discriminant is replaced by a suitable subdiscriminant.

Other methods for real root isolation

Solving polynomial equations is a fundamental task in symbolic computation as well as numerical analysis, and there are many different algorithms that compute their solutions, for various notions of what constitutes a solution. We mention selected examples.

The Descartes method belongs to a family of methods that compute solutions in the form of isolating intervals by recursive subdivision of an initial interval. Historically, the Descartes method has close links to the Continued Fractions method, whose study on digital computers also starts with Collins and Akritas [CA76], but whose origin is the work of Vincent [Vin36] from the 19th century. The Continued Fractions method combines Descartes' Rule with a different subdivision scheme, the modern forms of which are controlled by lower bounds on the positive real roots. We touch upon the Continued Fractions method again in §3.1.2, where further references are given.

Another relative of the Descartes method retains its subdivision scheme – recursive bisection of bounded intervals –, but employs a different termination criterion, namely Sturm's Theorem. The best known complexity bound for the resulting algorithm on polynomials of degree n with τ-bit integer coefficients is $O^{\sim}(n^4\tau^2)$, see [DSY07], the same as for the Descartes method. However, the simpler evaluation of Descartes' Rule makes the Descartes method more efficient in practice on a wide range of inputs [Joh91] [Joh98], even if the initial computation of the Sturm sequence is not taken into account.

The asymptotically efficient splitting-circle method[2] of Schönhage [Sch82] for numerical factorization in $\mathbb{C}[X]$ and the subsequent work by Pan [Pan02], as well as the high-performance complex rootfinder package MPSolve[3] of Bini and Fiorentino [BF00] exemplify a different notion of solution: Those methods accept as input a complex polynomial of degree n and produce an output that provides numerical approximations to the n complex roots, together with error bounds that can be made arbitrarily small. In more geometric terms, those algorithms deliver the solutions of a polynomial equation as circles in the complex plane, which are not necessarily disjoint. However, if the polynomial's coefficients are real and if all real roots are simple, then isolating intervals can be obtained from such an output, provided that the circles have radii small enough to satisfy the following

[2] `http://www.cs.uni-bonn.de/~schoe/tp/TPpage.html`
[3] `http://www.dm.unipi.it/cluster-pages/mpsolve/`

condition: All circles that intersect the real line are pairwise disjoint (to achieve isolation) and have images under complex conjugation that are disjoint to any other circle (to certify that they belong to a real root). In that case, the non-empty intersections of the circles with the real line are isolating intervals for the real roots.

An advantage of those methods is their ability to zoom in quickly on a cluster of several complex roots. By contrast, the interval boundaries in recursive bisection converge only linearly towards a cluster of roots before eventually separating them.

It is possible to generalize the concept of a "real root" beyond the field of real numbers through the theory of real closed fields [vdW93, Kap. 11] [BPR06, §2], which goes back to Artin and Schreier. A real closed field is an ordered field such that adjunction of $\sqrt{-1}$ yields an algebraically closed field, analogous to the relation between the real numbers and the complex numbers. Archimedean real closed fields are isomorphic to subfields of the real numbers and are, in that sense, covered by the setting discussed in this thesis.

On the other hand, a non-archimedean real closed field contains, by definition, infinitesimal elements. If a polynomial has two real roots that differ by an infinitesimal, recursive bisection of an interval with non-infinitesimal width cannot separate them. Hence the field of real numbers is the most general coefficient domain that makes sense for root isolation with rational intervals as in the Descartes method. In the non-archimedean setting, a Thom encoding [BPR06, §2.1, §10.4] can be used instead to represent real roots.

Notation

We write $\mathbb{N} := \{1, 2, 3, \ldots\}$ and $\mathbb{N}_0 := \{0, 1, 2, 3, \ldots\}$. The letters $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ denote the integer, rational, real and complex numbers, as usual. A complex number z is either real, $z \in \mathbb{R}$, or imaginary, $z \in \mathbb{C} \setminus \mathbb{R}$. A superscript asterisk denotes a ring's group of units; in particular, $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ and $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$.

For intervals, inclusion of boundary points is denoted by square brackets, as for the closed interval $[c, d]$, whereas parentheses denote exclusion of boundary points, as for the open interval (c, d).

We use the following forms of Gauss brackets: Given $x \in \mathbb{R}$, the floor function is written $\lfloor x \rfloor := \max\big(\mathbb{Z} \cap (-\infty, x]\big)$, and $\lceil x \rceil := \min\big(\mathbb{Z} \cap [x, +\infty)\big)$ stands for the ceiling function. In pseudocode, $\lfloor x \rceil$ denotes rounding to the nearest integer, with .5 rounded arbitrarily, whereas $[x]$ will be assigned a special meaning in Definition 3.35 on page 76.

The reader is alerted to the visual similarity between binomial coefficients $\binom{n}{k}$, which are scalars, and column vectors $\left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)$ with two components; unfortunately this cannot be avoided without deviating from standard notation, but the meaning is always clear from the context.

We write $\ln x$ for the natural logarithm of a real number $x > 0$. Its base is Euler's number $\mathrm{e} = \exp(1) = 2.71828\ldots$. The symbol $\log_b x = (\ln x)/(\ln b)$ denotes the logarithm of x with base $b > 0$. In case of the dyadic logarithm, we omit the subscript $b = 2$ for brevity and write $\log x$.

To compare the order of growth of non-negative functions $f(x)$ and $g(x)$ for $x \to +\infty$, up to constant factors, we employ the notations $f(x) = O(g(x))$, $f(x) = \Theta(g(x))$ and $f(x) = \Omega(g(x))$ advocated by Knuth [Knu76] [Knu97, §1.2.11.1], which have gained almost universal acceptance in computer science, despite justified concerns about "one-way equalities".

Chapter 2

Descartes' Rule of Signs, Some Extensions, and Other Foundations

This chapter reviews mathematical background material for the Descartes method. With a few, explicitly indicated exceptions, none of the results are new. However, some items have not previously been presented in relation to the Descartes method.

2.1 Descartes' Rule and Obreshkoff's extension

2.1.1 Descartes' Rule

Definition 2.1. Let $(a_0, \ldots, a_n)$ be a sequence of real numbers. We write $\operatorname{var}(a_0, \ldots, a_n)$ for the *number of sign variations* in the sequence, or formally

$$\operatorname{var}(a_0, \ldots, a_n) := \#\{(i,k) \in \{0, \ldots, n\}^2 \mid i < k,\ a_i a_k < 0,\ \forall i < j < k\colon a_j = 0\}. \quad (2.1)$$

Theorem 2.2 (Descartes' Rule of Signs). *Let $A(X) = \sum_{i=0}^{n} a_i X^i$ be a polynomial of degree n with real coefficients that has exactly p positive real roots, counted with multiplicities. Let $v = \operatorname{var}(a_0, \ldots, a_n)$ be the number of sign variations in its coefficient sequence. Then $v \geq p$ and $v \equiv p \pmod 2$. If all roots of $A(X)$ are real, then $v = p$.*

Proof. We first show $v \equiv p \pmod 2$. The multiplicity of a real root ϑ is odd if and only if $A(X)$ changes sign at ϑ. The number p is the sum of the multiplicities of the positive real roots, hence p is odd if and only if the signs of $A(x)$ for $x \to +\infty$ and for $x \to 0^+$ disagree. As these are the signs of the leading coefficient and the lowest non-zero coefficient, resp., they differ if and only if v is odd.

Let us now show $v \geq p$ by induction on the degree n. The base case $n = 1$ is trivial. For the inductive step, we consider the derivative A' of degree $n-1$ with its p' positive real roots and v' sign variations, and we assume $v' \geq p'$. By Rolle's Theorem, there is a real root of A' between any two adjacent real roots of A. If A has a k-fold real root ϑ, this is a $(k-1)$-fold root of A'; we think of this as $k-1$ roots of A' at ϑ "between" the k roots of A at ϑ. Altogether, there are at least $p-1$ positive roots of A' between the p positive roots of A, hence $p' \geq p-1$. In summary, we obtain $v \geq v' \geq p' \geq p-1$, where $v = p-1$ is excluded by congruence modulo 2, hence $v \geq p$, as desired.

It remains to show $v \leq p$ in case all roots of A are real. We argue by induction as above. According to our preceding considerations, there are $n-1$ real roots of A' between the n real roots of A. In particular, all $n-1$ roots of A' are real, so we may assume inductively that $v' \leq p'$. Of the $n-1$ roots of A', $n-p-1$ are between the $n-p$ non-positive roots of A, hence at most p roots of A' are positive. Thus we obtain $v - 1 \leq v' \leq p' \leq p$, where again equality is excluded, so $v \leq p$ as desired. □

Proofs of Descartes' Rule that employ Rolle's Theorem in such an inductive argument have been discovered repeatedly, for example by Jaccottet [Jac09] and Wang [Wan04].

2.1.2 Obreshkoff's extension

Descartes' Rule only gives an upper bound on the number of roots in a certain range. Under which conditions does it give an exact count? We shall present a particularly general answer due to Obreshkoff[1] that has previously been used by Alesina and Galuzzi [AG98] in connection with the Continued Fractions method, but appears to have been overlooked in previous work (including our own) on the Descartes method, the algorithm that we will present in Chapter 3. Obreshkoff's starting point is the following result from [Obr25, §1.III], concerning the inclusion of imaginary roots in the count of Descartes' Rule.

Lemma 2.3 (Obreshkoff (1925)). *Let $A(X) = \sum_{i=0}^{n} a_i X^i$ be a real polynomial of degree n. Let $B(X) = \sum_{i=0}^{n+2} b_i X^i = A(X) \cdot (X^2 - 2\rho\cos(\varphi)X + \rho^2)$ with $\rho > 0$. Let $v = \operatorname{var}(a_0, \ldots, a_n)$ and $v' = \operatorname{var}(b_0, \ldots, b_{n+2})$. If $0 \leq \varphi < \frac{\pi}{n+2-v}$, then $v' \geq v + 2$ and $v' \equiv v \pmod 2$.*

Obreshkoff's proof of the lemma also appears in [Obr03, §II.8] and [Obr63, §17]; a more recent textbook reference is [RS02, Lem. 10.3.2(ii)]. Once this lemma is established, the further argumentation is straightforward. Nevertheless, we carry it out to achieve a slight but helpful generalization.

Theorem 2.4 (Obreshkoff (1952)). *Consider the real polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$ of degree n and its complex roots, counted with multiplicities. Let $v = \operatorname{var}(a_0, \ldots, a_n)$. If $A(X)$ has at least p roots with arguments in the range $-\frac{\pi}{n+2-p} < \varphi < \frac{\pi}{n+2-p}$, then $v \geq p$.*

Proof. Let us call the roots in question $\vartheta_1, \ldots, \vartheta_p$. We may assume w.l.o.g. that imaginary roots among them occur in pairs of complex conjugates. (If not, we can add conjugates and increase p, this only widens the range of permissible arguments.) We choose indices such that the first $2c$ roots $\vartheta_1 = \overline{\vartheta_2}$, $\ldots$, $\vartheta_{2c-1} = \overline{\vartheta_{2c}}$ are pairs of complex conjugates and the remaining $r = p - 2c$ roots are real. Let ρ_i denote the magnitude and $\pm\varphi_i$, $0 < \varphi_i < \pi$, denote the arguments of ϑ_{2i-1} and ϑ_{2i} for any $1 \leq i \leq c$.

To each pair of complex-conjugate roots corresponds a quadratic factor

$$Q_i(X) = (X - \vartheta_{2i-1})(X - \vartheta_{2i}) = X^2 - 2\rho_i \cos(\varphi_i)X + \rho_i^2$$

of $A(X)$. Let $G_c(X) = A(X)$ and $G_{i-1}(X) = G_i(X)/Q_i(X)$ for $i = c, \ldots, 1$. Clearly, $\deg(G_i) = n - 2c + 2i$. We write v_i for the number of sign variations in the coefficient sequence of G_i. Let us prove the theorem inductively by showing that $v_i \geq r + 2i$.

For the base case $i = 0$, we observe that $G_0(X)$ has at least r positive real roots and apply Descartes' Rule.

For the inductive step from i to $i + 1$, we recall $p = r + 2c$ and observe

$$0 < \varphi_{i+1} < \frac{\pi}{n + 2 - (r + 2c)} = \frac{\pi}{(n - 2c + 2i) + 2 - (r + 2i)} \leq \frac{\pi}{\deg(G_i) + 2 - v_i}.$$

[1] Nikola Obreshkoff / Никола Обрешков (1896–1963), prominent Bulgarian mathematician, professor at the University of Sofia and member of the Bulgarian Academy of Sciences. His last name also occurs in the transcriptions "Obrechkoff" (French), "Obreschkoff" (German), "Obreshkov" and "Obreškov". *Serdica Mathematical Journal* **22** (1996), issue 4, commemorates his 100th birthday and contains a short biography (pp. ii–vi) by P. Russev.

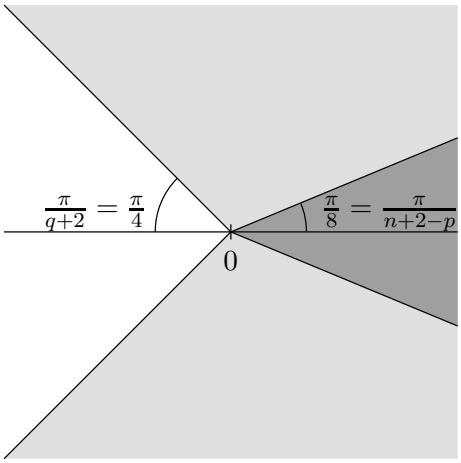


Figure 2.1: The argument ranges of Theorem 2.7 for $n = 8$ and $p = q = 2$.

Hence Lemma 2.3 applied to $G_i(X)$ and $Q_{i+1}(X)$ implies $v_{i+1} \geq v_i + 2 \geq r + 2(i+1)$, as desired. □

Corollary 2.5. *If $A(X)$ has exactly p roots with arguments in the range $-\frac{\pi}{n+2-p} < \varphi < \frac{\pi}{n+2-p}$, then $v \geq p$ and $v \equiv p \pmod 2$.*

Obreshkoff demonstrates that the argument range has to exclude its boundaries by means of the example $A(X) = X^n + 1$, which has a pair of conjugate roots with arguments $\pm\pi/n$ but no sign variations.

Lemma 2.6. *Let $(a_0, \ldots, a_n)$ be a sequence of n real numbers. Then $\operatorname{var}(a_0, \ldots, a_n) + \operatorname{var}(a_0, \ldots, (-1)^i a_i, \ldots, (-1)^n a_n) \leq n$.*

Proof. By induction on n. Let $v_n := \operatorname{var}(a_0, \ldots, a_n) + (a_0, \ldots, (-1)^i a_i, \ldots, (-1)^n a_n)$. The base case $n = 0$ is clear. Let us proceed from n to $n+1$. We distinguish two cases.

If $a_n \neq 0$, then at most one of (a_n, a_{n+1}) and $((-1)^n a_n, (-1)^{n+1} a_{n+1})$ exhibits a sign variation, hence $v_{n+1} \leq v_n + 1 \leq n + 1$.

If $a_n = 0$, then $v_n = v_{n-1} \leq n - 1$ and $v_{n+1} \leq v_n + 2 \leq n + 1$. □

Theorem 2.7 (Obreshkoff (1952)). *Consider the real polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$ of degree n and its complex roots, counted with multiplicities. Let v denote $\operatorname{var}(a_0, \ldots, a_n)$. If $A(X)$ has at least p roots with arguments in the range $-\frac{\pi}{n+2-p} < \varphi < \frac{\pi}{n+2-p}$, and at least $n - q$ roots with arguments in the range $\pi - \frac{\pi}{q+2} \leq \psi \leq \pi + \frac{\pi}{q+2}$, then $q \geq v \geq p$. If, in particular, $q = p$, then $A(X)$ has exactly p roots with arguments φ in the range given above and $v = p$.*

Notice that Theorem 2.4 is contained in this theorem as special case $q = n$.

Proof. By Theorem 2.4, $v \geq p$. It remains to show $v \leq q$. In a first step, we prove this under the stronger condition that $n - q$ roots have an argument in the *open* range $\pi - \frac{\pi}{q+2} < \psi < \pi + \frac{\pi}{q+2}$. This allows us to apply Theorem 2.4 to the number v' of sign variations in the coefficient sequence of $A(-X)$ and thus to obtain $v' \geq n - q$. Since $v + v' \leq n$ by the preceding lemma, we find $v \leq n - v' \leq q$.

Let us now extend the theorem to the case that $A(X)$ has $n - q$ roots with arguments in the closed range $\pi - \frac{\pi}{q+2} \leq \varphi \leq \pi + \frac{\pi}{q+2}$. As the coefficients of $A(X)$ depend continuously on

the roots, $A(X)$ is the coefficient-wise limit of a sequence $(A_0, A_1, A_2, \ldots)$ of polynomials $A_k(X) = \sum_{i=0}^{n} a_{ki} X^i$ that satisfy the stronger condition above and thus have at most q sign variations. Each A_k presents a certain sign pattern $(\operatorname{sgn}(a_{k0}), \operatorname{sgn}(a_{k1}), \ldots, \operatorname{sgn}(a_{kn}))$, of which there are only finitely many. Thus there exists a pattern $(\sigma_0, \ldots, \sigma_n)$ that is assumed by all elements of an infinite subsequence and has at most q sign variations. In the limit, each coefficient a_i retains the sign σ_i or vanishes. Hence $\operatorname{var}(a_0, \ldots, a_n) \leq \operatorname{var}(\sigma_0, \ldots, \sigma_n) \leq q$. □

Obreshkoff originally published the two preceding theorems 1952 both in Russian [Obr52a] and in Bulgarian with French summary [Obr52b]. More accessible sources are the respective sections in Obreshkoff's textbooks: §II.8 of the English book [Obr03] on (complex) *Zeros of Polynomials*, which originally appeared 1963 in Bulgarian language; and §17 of the German book [Obr63] from 1963 on *Distribution and Computation of the Zeros of Real Polynomials*. Obreshkoff's literal formulation of the first theorem is what we called Corollary 2.5. Obreshkoff's wording of Theorem 2.7 treats the particular case $p = q$ and excludes the boundaries in the condition on ψ. We have come to consider the case $p \neq q$ in response to a question posed by R. Seidel (personal communication, February 2007). While our formulation of the theorems are more general, we did not have to add anything new to Obreshkoff's proof technique to obtain them.

These two theorems were no isolated results at that time. Already in the 1920s, Obreshkoff found other extensions of the rules of Descartes and Budan-Fourier, which are also presented in his books. The recent book by Rahman and Schmeisser [RS02, §10] contains a wealth of such results; a further reference is Marden's classic monograph [Mar66].

2.1.3 Simpler proofs for special cases

We need only the special cases $p = q = 0$ and $p = q = 1$ of Theorem 2.7 in our analysis of the Descartes method (§3.1.5). For them, we can give short, complete proofs.

Proposition 2.8 (case $p = q = 0$). *If all n complex roots of $A(X) = \sum_{i=0}^{n} a_i X^i$ have a non-positive real part, then* $\operatorname{var}(a_0, \ldots, a_n) = 0$.

Proof. Each real root x of $A(X)$ corresponds to a linear factor $X - x$, $x \leq 0$. Each complex-conjugate pair of imaginary roots $z, \overline{z}$ of $A(X)$ corresponds to a quadratic factor $X^2 - (z + \overline{z})X + |z|^2$, $z + \overline{z} \leq 0$. Their product $A(X)/a_n$ has no negative coefficients. □

This argument is so simple that it seems hopeless to pinpoint its first occurrence. Lagrange argued in this fashion already in the late 18th century [AG98]. The "one-circle theorem" arising from Proposition 2.8 through any Möbius transformation $(0, \infty) \to (a, b)$ (Proposition 2.33 below) was already used by Vincent [Vin36, p. 345].

Proposition 2.9 (case $p = q = 1$). *If $A(X) = \sum_{i=0}^{n} a_i X^i$ has one simple positive real root, and all other complex roots have an argument φ in the range $\frac{2}{3}\pi \leq \varphi \leq \frac{4}{3}\pi$, then* $\operatorname{var}(a_0, \ldots, a_n) = 1$.

Notice that this is equivalent to Theorem 2.7 for $p = q = 1$, because a unique complex root with positive real part lacks a complex conjugate and thus is necessarily real.

Proof. Let us write $A(X)$ as $A(X) = A_0(X) \prod_{j=1}^{k} (X^2 - 2\xi_j X + \xi_j^2 + \eta_j^2)$ where all roots of $A_0(X)$ are real and $\xi_j \pm \mathrm{i}\eta_j$, $j = 1, \ldots, k$, are the pairs of complex-conjugate roots of $A(X)$. The condition on their arguments is equivalent to

$$\xi_j < 0 \quad \text{and} \quad \eta_j^2 \leq 3\xi_j^2, \tag{2.2}$$

because $\tan(\pi/3) = \sqrt{3}$. As $A_0(X)$ has only real roots, Descartes' Rule counts exactly, so $A_0(X)$ exhibits just one sign variation. For a proof by induction, it suffices to demonstrate:

(*) If the real polynomial $F(X)$ has exactly one sign variation, and if $X^2 - 2\xi_j X + \xi_j^2 + \eta_j^2$ satisfies the condition (2.2), then also their product has exactly one sign variation.

To show (*), we change coordinates by replacing X with $-2\xi_j X$. Since $-2\xi_j > 0$, this does not affect the numbers of sign variations, and it puts the quadratic factor into the simpler form $4\xi_j^2(X^2 + X + \lambda)$ where $\lambda = (\xi_j^2 + \eta_j^2)/(4\xi_j)$. Clearly, $\lambda \geq \frac{1}{4}$. The condition (2.2) implies $\lambda \leq 1$. Let $f_0, \ldots, f_m$ denote the coefficients of $F(X)$, and set $f_{-2} = f_{-1} = 0 = f_{m+1} = f_{m+2}$. We may assume without loss of generality that $f_m > 0$. The existence of a unique sign variation of $F(X)$ means that there exists an index j, $0 \leq j < m$, such that $f_0, \ldots, f_j \leq 0$ and $f_{j+1}, \ldots, f_m \geq 0$. The coefficients of the product $G(X) = (X^2 + X + \lambda)F(X)$ are

$$g_i = f_{i-2} + f_{i-1} + \lambda f_i, \quad i = 0, \ldots, m+2.$$

Clearly, $g_0, \ldots, g_j \leq 0$ and $g_{j+3}, \ldots, g_{m+2} \geq 0$. It remains to show $g_{j+1} \leq g_{j+2}$; this guarantees a unique sign variation of $G(X)$ irrespective of the actual signs of g_{j+1} and g_{j+2}. Indeed, we have

$$g_{j+1} = f_{j-1} + f_j + \lambda f_{j+1} \leq f_j + \lambda f_{j+1} \leq f_j + f_{j+1} \leq f_j + f_{j+1} + \lambda f_{j+2} = g_{j+2}.$$

The condition (2.2) enters the proof at the middle inequality in the form $\lambda \leq 1$. □

This proposition and a proof similar to the one above have been published 1941 by Stephan Lipka (Hungarian: Lipka István) both in German [Lip41b] and in Hungarian with German summary [Lip41a]. (Lipka is also known for other results in this area, e.g., [Lip42], see [RS02, §10.2].) Lipka's proof differs from the one above in that he invokes the famous general theorem of Schoenberg[2] [Sch30] on variation-diminishing linear transformations to treat the subsequence $(g_j, \ldots, g_{j+3})$. This saves our arguments regarding $g_{j+1} \leq g_{j+2}$ but necessitates a discussion of signs of minors of the transformation matrix and some extra arguments in case of vanishing coefficients. A direct proof with some similarity to the one above, but with more case distinctions, appears in [AG98, Cor. 8.2]. Ostrowski [Ost50] gave a different proof of this proposition based on his earlier results on normal power series, see also [KM06].

How should we refer to the results of this section? Theorem 2.7 subsumes all the other results. It deserves to be reported and remembered in full generality to avoid needless gaps in the presentation of the mathematical background of Descartes' Rule. The name that has to be attached to it is Obreshkoff, for two reasons:

- Obreshkoff [Obr52a] [Obr52b] was the first to prove this theorem in essentially full generality in 1952 (our additions are minor); even though Lipka [Lip41a] [Lip41b] had previously treated special cases, Proposition 2.9 and beyond. (The crucial argument (*) in Lipka's and our proof of Proposition 2.9, however, is in turn a special case of an older lemma by Obreshkoff [Obr25, §1.IV].)
- The core of the proof of Theorem 2.7 is Lemma 2.3, which appears in an article by Obreshkoff [Obr25, §1.III] received August 20th, 1923, clearly predating the contributions of Lipka and Ostrowski.

[2] Isaac Jacob Schoenberg; see footnote 2 on page 124.

Hence we refer to Theorem 2.7 as "Obreshkoff's extension of Descartes' Rule of Signs" in the sequel. We refer to the "$q \geq v$" part of the Theorem as "Obreshkoff's partial converse to Descartes' Rule of Signs", because it gives an upper bound on v in terms of roots, whereas Descartes' Rule gives a lower bound.

2.2 Descartes' Rule for arbitrary open intervals

Descartes' Rule, in its form discussed above, is concerned with the number of roots in the open interval $(0, \infty)$. From §2.2.3 onwards, we will generalize it to arbitrary open intervals. This requires some preparations.

2.2.1 The projective line

Let K stand for any of $\mathbb{R}$ and $\mathbb{C}$.

To get a handle on the right endpoint of the interval $(0, \infty)$, we wish to extend the affine line K by a unique *point at infinity* denoted ∞. To achieve this in terms of sets, we can simply define $\widehat{K} := K \cup \{\infty\}$. We can even turn this into a topological space by distinguishing a subset $U \subseteq \widehat{K}$ as *open* if $\forall x \in U\colon \exists \varepsilon > 0\colon U_\varepsilon(x) \subseteq U$, where we define the ε-neighbourhood $U_\varepsilon(x)$ as usual for $x \in K$ and as $U_\varepsilon(\infty) = \{x \in K \mid |x| > 1/\varepsilon\} \cup \{\infty\}$ for the point at infinity. For the obvious reasons, $\widehat{K}$ is called *one-point compactification* of K. Clearly, the subspace topology of K in $\widehat{K}$ is the usual topology of K. Convergence to ∞ as defined by the topology of $\widehat{K}$ is equivalent to the usual definition of proper divergence in K.

We want a model of $\widehat{K}$ that comes with a uniform coordinate system: the 1-dimensional projective space $\mathbb{P}^1(K)$, also called the projective line. Recall that $\mathbb{P}^1(K)$ is defined as the set of equivalence classes of $K^2 \setminus \{(0,0)\}$ under collinearity, which is the relation $(x, y) \sim (x', y') \;:\Leftrightarrow\; \exists \lambda \in K^*\colon \lambda(x, y) = (x', y')$. Let us write $\eta\colon (K^2 \setminus \{(0,0)\}) \to \mathbb{P}^1(K)$ for the canonical projection of representatives onto equivalence classes. The equivalence class $\eta((x, y))$ of $(x, y) \in K^2$ is also written $[x : y]$. If $y = 0$, then we have $[x : y] = [1 : 0]$. If $y \neq 0$, then $[x : y] = [x/y : 1]$. Hence $\mathbb{P}^1(K)$ is indeed the disjoint union of the point at infinity $[1 : 0]$ and a copy of K injected by $K \to \mathbb{P}^1(K)$, $x \mapsto [x : 1]$.

Consider the group $\mathrm{GL}_2(K)$ of invertible 2×2 matrices. The matrices $\left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right)$, $a \in K^*$, form a normal subgroup N. In the quotient $\mathrm{PGL}_2(K) := \mathrm{GL}_2(K)/N$, two matrices are equivalent iff they differ only by multiplication with $a \in K^*$. We write $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ for the equivalence class of $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. $\mathrm{PGL}_2(K)$ acts on $\mathbb{P}^1(K)$ by matrix multiplication of representatives, i.e., $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = \left[\begin{smallmatrix} ax+by \\ cx+dy \end{smallmatrix}\right]$. This is independent of the choice of representatives and thus well-defined.

Let us re-examine this group action in more familiar terms. We write x for $[x : 1]$ and ∞ for $[1 : 0]$. Let us first assume that $c \neq 0$. Then $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ acts as follows:

$$\begin{aligned} x &\mapsto \frac{ax+b}{cx+d} \qquad \text{if } x \neq -d/c, \\ -\frac{d}{c} &\mapsto \lim_{x \to -\frac{d}{c}} \frac{ax+b}{cx+d} = \infty, \\ \infty &\mapsto \lim_{x \to \infty} \frac{ax+b}{cx+d} = \frac{a}{c}. \end{aligned}$$

In the special case $c = 0$, we have $d \neq 0$ by invertibility, so we can choose $d = 1$. The action of $\left[\begin{smallmatrix} a & b \\ 0 & 1 \end{smallmatrix}\right]$ is an affine map: $x \mapsto ax + b$, $\infty \mapsto \infty$.

Suppose that $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right]$ act equally on $\mathrm{PGL}_2(K)$. Then $\left[\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right]^{-1}\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ acts as the identity. This is only true for $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, so we actually have $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] = \left[\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right]$. (In other words: The action we consider is *faithful*.) We can thus identify the elements of $\mathrm{PGL}_2(K)$ with the maps $\widehat{K} \to \widehat{K}$ of the form $x \mapsto \frac{ax+b}{cx+d}$, $ad - bc \neq 0$. These maps are called *Möbius transformations* of $\widehat{K}$. From the matrix representation, it is clear that they form a group and thus are bijective. From the fractional representation, it follows that they are continuous and hence even homeomorphisms.

Much of the material above is developed in greater detail in the initial chapters of Anderson's book [And99].

Let us restrict now to the case $K = \mathbb{R}$ and discuss affine and projective intervals. For the purposes of this thesis, neither the empty set nor the entire line shall be considered intervals; in other words, our intervals are always a non-empty proper subsets. The points on the boundary of an interval are its *endpoints*.

An *affine open interval* is a connected open subset of $\mathbb{R}$. It can be identified unambiguously by specifying its two endpoints, one of which may be the point at infinity. A *projective open interval* I is a connected open subset of $\widehat{\mathbb{R}} = \mathbb{P}^1(\mathbb{R})$. As in the affine case, every projective open interval has two distinct endpoints, but the correspondence of intervals to endpoints is not bijective.

Proposition 2.10. *Let* $\binom{c_1}{c_2}$, $\binom{d_1}{d_2}$ *be two linearly independent elements of* $\mathbb{R}^2$.

(i) There are exactly two projective open intervals that have endpoints $[c_1 : c_2]$ *and* $[d_1 : d_2]$*; these are the two connected components of* $\mathbb{P}^1(\mathbb{R}) \setminus \{[c_1 : c_2], [d_1 : d_2]\}$*. One of them, call it* I^+*, contains* $[c_1 + d_1 : c_2 + d_2]$ *but not* $[c_1 - d_1 : c_2 - d_2]$*; the other one, call it* I^-*, contains* $[c_1 - d_1 : c_2 - d_2]$ *but not* $[c_1 + d_1 : c_2 + d_2]$*.*

(ii) There is a uniquely determined projective open interval I *that has endpoints* $[c_1 : c_2]$ *and* $[d_1 : d_2]$*, and contains* $[c_1 + d_1 : c_2 + d_2]$*.*

(iii) Any projective open interval I *arises in this way. The representatives* $\binom{c_1}{c_2}$, $\binom{d_1}{d_2}$ *of its endpoints are determined up to order and multiplication by positive constants and a common sign* ± 1*.*

Intervals as in the pair (I^+, I^-) are called *complementary* to each other.

Proof. Ad (i). As $\{\binom{c_1}{c_2}, \binom{d_1}{d_2}\}$ is a basis of $\mathbb{R}^2$, every point of $\mathbb{R}^2$ has a unique expression of the form $\lambda\binom{c_1}{c_2} + \mu\binom{d_1}{d_2}$. This induces a partition of $\mathbb{R}^2$ as follows: The points with $\lambda\mu = 0$ lie on one of the lines through the origin spanned by $\binom{c_1}{c_2}$ and $\binom{d_1}{d_2}$. The remaining points form two open double cones: X^+ in which $\lambda\mu > 0$, and X^- in which $\lambda\mu < 0$. Their projections $I^+ := \eta(X^+)$ and $I^- := \eta(X^-)$ to $\mathbb{P}^1(\mathbb{R})$ are the desired intervals.

Ad (ii). Immediate from (i).

Ad (iii). Obvious. □

We remark that the point $[c_1 + d_1 : c_2 + d_2]$ is not special. Any point of I can be written in this way by choosing suitable representatives for the endpoints.

A projective open interval I is an affine open interval if and only if $\infty \notin I$. The image of an (affine or projective) open interval under a Möbius transformation M is a projective open interval, because M is a homeomorphism and thus preserves openness

and connectedness. However, the image of an affine open interval I is not necessarily an affine open interval, since M may take an element of I to ∞.

Proposition 2.11. *Let* $\binom{c_1}{c_2}$, $\binom{d_1}{d_2}$ *be two linearly independent elements of* $\mathbb{R}^2$. *Let* I *be the projective open interval that has endpoints* $[c_1 : c_2]$ *and* $[d_1 : d_2]$ *and contains* $[c_1 + d_1 : c_2 + d_2]$. *The matrices* $M \in \mathrm{GL}_2(\mathbb{R})$ *that represent Möbius transformations mapping* $(0, \infty)$ *to* I *are precisely those of the forms* $M = \left(\begin{smallmatrix} \lambda c_1 & \mu d_1 \\ \lambda c_2 & \mu d_2 \end{smallmatrix}\right)$ *and* $M = \left(\begin{smallmatrix} \mu d_1 & \lambda c_1 \\ \mu d_2 & \lambda c_2 \end{smallmatrix}\right)$ *with* $\lambda, \mu \in \mathbb{R}^*$, $\lambda\mu > 0$.

Proof. As Möbius transformations are homeomorphisms, endpoints are mapped to endpoints. Therefore, the matrices of said form with unconstrained signs of λ and μ are precisely those that map $(0, \infty)$ to I or the projective open interval J that is complementary to I. The point $[1 : 1] \in (0, \infty)$ is mapped to $[\lambda c_1 + \mu d_1 : \lambda c_2 + \mu d_2] = [c_1 + \mu/\lambda \cdot d_1 : c_2 + \mu/\lambda \cdot d_2]$. By Proposition 2.10(ii), this shows that $(0, \infty)$ is mapped to I for $\lambda\mu > 0$ and to J otherwise. □

2.2.2 Polar forms

Let K stand for any of $\mathbb{R}$ and $\mathbb{C}$.

Definition 2.12. Let $n \in \mathbb{N}$. A *homogeneous polar form* of degree n is an n-ary map $K^2 \times \cdots \times K^2 \to K$ which is multilinear (i.e., linear in each argument) and symmetric.

We write the evaluation of a polar form F on an n-tuple of vectors $(\binom{x_1}{y_1}, \binom{x_2}{y_2}, \ldots, \binom{x_n}{y_n})$ as $F\left[\binom{x_1}{y_1}\binom{x_2}{y_2}\cdots\binom{x_n}{y_n}\right]$ and denote a repetition of arguments by superscripts, as in $F\left[\binom{c_1}{c_2}^2\binom{d_1}{d_2}^3\right] = F\left[\binom{c_1}{c_2}\binom{c_1}{c_2}\binom{d_1}{d_2}\binom{d_1}{d_2}\binom{d_1}{d_2}\right]$.

Polar forms are classical objects. They were rediscovered and popularized under the name "blossoms" by Ramshaw [Ram87] [Ram89] for the study of parametric curves, see also [PBP02] [Far97]. Ramshaw explains how to turn our notational convention of juxtaposition of arguments and superscripts for repetitions into a proper mathematical construction (a tensor product).

We proceed to determine a closed-form expression for the value $F\left[\binom{X_1}{Y_1}\cdots\binom{X_n}{Y_n}\right]$ at indeterminate arguments. Using multilinearity, let us "multiply out" the arguments $\binom{X_i}{Y_i} = X_i\binom{1}{0} + Y_i\binom{0}{1}$ to obtain a polynomial in the X_i, Y_i whose coefficients are values of F at n-tuples of basis vectors $\binom{1}{0}$, $\binom{0}{1}$. By symmetry, we can reorder these n-tuples and find

$$F\left[\binom{X_1}{Y_1}\cdots\binom{X_n}{Y_n}\right] = \sum_{i=0}^{n} F\left[\binom{1}{0}^i\binom{0}{1}^{n-i}\right]\left(\sum_{\#J=i}\prod_{j\in J} X_j \prod_{j\notin J} Y_j\right). \tag{2.3}$$

The summation in the parentheses is over all subsets J of $\{1, \ldots, n\}$ with cardinality i. It follows that all polar forms of degree n have a unique expression of the form

$$F\left[\binom{X_1}{Y_1}\cdots\binom{X_n}{Y_n}\right] = \sum_{i=0}^{n} f_i \cdot \left(\sum_{\#J=i}\prod_{j\in J} X_j \prod_{j\notin J} Y_j\right), \quad f_0, \ldots, f_n \in K. \tag{2.4}$$

Conversely, any choice of the f_i in (2.4) yields a polar form F.

Let us now consider the *diagonal* of the polar form (2.4), that is the map $K^2 \to K$ obtained by substituting the same vector for all arguments. It is a homogeneous polynomial

$$F\left[\binom{X}{Y}^n\right] = \sum_{i=0}^{n} f_i \binom{n}{i} X^i Y^{n-i}, \quad f_i = F\left[\binom{1}{0}^i \binom{0}{1}^{n-i}\right]. \tag{2.5}$$

Vice versa, we see that any homogeneous polynomial is the diagonal of a uniquely determined polar form. Justified by this bijective correspondence, we will often use the same symbol, such as F, to refer interchangeably to a homogeneous polynomial $F(X,Y)$ and its polar form $F\left[\binom{X_1}{Y_1} \cdots \binom{X_n}{Y_n}\right]$.

The ability to obtain a polynomial's coefficients as values of the polar form immediately yields the following explicit expression for coefficients arising from a parameter transformation.

Lemma 2.13. *Consider a polar form F of degree n and $M \in \mathrm{GL}_2(K)$. The coefficient of $\binom{n}{i}X^iY^{n-i}$ in $G(X,Y) := F(M(X,Y))$ is $G\left[\binom{1}{0}^i\binom{0}{1}^{n-i}\right] = F\left[(M\binom{1}{0})^i(M\binom{0}{1})^{n-i}\right]$.*

Also, we can express a polynomial's derivative by evaluating its polar form suitably.

Lemma 2.14. *The derivative of $F\left[\binom{X}{Y}^n\right]$ w.r.t. X is $n \cdot F\left[\binom{X}{Y}^{n-1}\binom{1}{0}\right]$.*

Proof. Apply the product rule and differentiate the multilinear map in each argument, noting that the derivative of $\binom{X}{Y}$ is $\binom{1}{0}$, then use symmetry to reorder the arguments and attain n identical summands. □

This gives a different interpretation to the coefficients f_i in (2.5): Taylor expansion at the point $[0:1]$. We remark that fixing one argument of F at an arbitrary vector $\binom{x_0}{y_0}$ in place of $\binom{1}{0}$ elegantly implements the differential operator $x_0\frac{\partial}{\partial X} + y_0\frac{\partial}{\partial Y}$, which yields the *polar derivative* with pole $[x_0 : y_0]$; cf. [Far97, §4.7] [Mar66, §10] [RS02, §3.1].

With Lemma 2.14, we can generalize the well-known relation between the multiplicity of roots of polynomials and the vanishing of derivatives as follows.

Proposition 2.15. *Let S be a subset of K^2 that contains two linearly independent vectors. Let $k \in \mathbb{N}$. The point $[x:y]$ is a root of the homogeneous polynomial $F(X,Y)$ with multiplicity at least k if and only if $F\left[\binom{x_1}{y_1}\cdots\binom{x_{k-1}}{y_{k-1}}\binom{x}{y}^{n-k+1}\right] = 0$ for any choice of $\binom{x_1}{y_1}, \ldots, \binom{x_{k-1}}{y_{k-1}} \in S$.*

Proof. By a suitable change of coordinates in K^2, we attain $\binom{x}{y} = \binom{0}{1}$. The symmetric multilinear map $\binom{x_1}{y_1}\cdots\binom{x_{k-1}}{y_{k-1}} \mapsto F\left[\binom{x_1}{y_1}\cdots\binom{x_{k-1}}{y_{k-1}}\binom{0}{1}^{n-k+1}\right]$ yields 0 for all arguments taken from S iff it yields 0 for all arguments of the form $\binom{1}{0}^i\binom{0}{1}^{k-i-1}$. This is in turn equivalent to $F(0,1) = \frac{\mathrm{d}}{\mathrm{d}X}F(0,1) = \cdots = \frac{\mathrm{d}^{k-1}}{\mathrm{d}X^{k-1}}F(0,1) = 0$ by the preceding lemma. □

When we consider the polar form $F\left[\binom{X_1}{Y_1}\cdots\binom{X_n}{Y_n}\right]$ only at arguments of the kind $\binom{X_i}{1}$, the canonical representatives of the finite elements of $\mathbb{P}^1(K)$, we can substitute $Y_1 = \cdots = Y_n = 1$ throughout, analogous to the dehomogenization of a polynomial. Since $\lambda\binom{a}{1} + \mu\binom{b}{1} = \binom{\lambda a+\mu b}{1}$ iff $\lambda+\mu = 1$, the resulting map $(X_1,\ldots,X_n) \mapsto F\left[\binom{X_1}{1}\cdots\binom{X_n}{1}\right]$ is affine (not linear) in each argument.

Definition 2.16. Let $n \in \mathbb{N}$. An *affine polar form* of formal degree n is an n-ary map $K \times \cdots \times K \to K$ which is multiaffine (i.e., affine in each argument) and symmetric.

Homogeneous polar forms of degree n and affine polar forms of formal degree n correspond bijectively to each other by *dehomogenization* (as above) and *homogenization*

$$F\left[\binom{X_1}{Y_1}\cdots\binom{X_n}{Y_n}\right] := Y_1\cdots Y_n \cdot F[X_1/Y_1,\ldots,X_n/Y_n]. \tag{2.6}$$

Given an affine polar form $F[X_1,\ldots,X_n]$ of formal degree n, its diagonal $F[X,\ldots,X]$ is a polynomial $F(X)$ of degree at most n. Vice versa, a polynomial $A(X)$ of degree at most n corresponds to a homogeneous polynomial and thus to a homogeneous polar form of degree n and further to an affine polar form $F[X_1,\ldots,X_n]$ of formal degree n such that $F[X,\ldots,X] = A(X)$. As in the homogeneous case, we often identify affine polar forms and polynomials.

2.2.3 Generalization to arbitrary open intervals

We are now ready to generalize Descartes' Rule to arbitrary projective open intervals.

Theorem 2.17. *Let the real homogeneous polynomial $F(X,Y)$ have exactly p roots in the projective open interval I, counted with multiplicities. Let $M \in \mathrm{GL}_2(\mathbb{R})$ represent a Möbius transformation that maps $(0,\infty)$ to I. Let*

$$v = \mathrm{var}(g_0,\ldots,g_n) \quad \textit{where} \quad G(X,Y) = \sum_{i=0}^{n} g_i X^i Y^{n-i} = (F\circ M)(X,Y). \tag{2.7}$$

Then $v \geq p$ and $v \equiv p \pmod 2$. If all roots of F are real, then $v = p$.

Proof. Apply Descartes' Rule (Theorem 2.2) to $G(X,1)$. □

For the case of an affine interval $I = (c,d)$ and the specific transformation M defined by $M^{-1}(X) = (d-X)/(X-c)$, this "little observation" was enunciated by Jacobi [Jac35, IV.] and has therefore been called "Jacobi's rule of signs" in the literature [RS02, Cor. 10.1.13]. Jacobi's contemporary Vincent [Vin36], building on earlier work of Lagrange, has used such a combination of Descartes' Rule and Möbius transformations for root isolation at about the same time; see [AG98] for a modern account and for references to earlier versions of Vincent's work. We refrain from discussing these historical aspects further and instead proceed to translate this result into the language of polar forms.

Corollary 2.18. *With notation as in Theorem 2.17, let I have endpoints $[c_1:c_2]$, $[d_1:d_2]$ and contain the point $[c_1+d_1 : c_2+d_2]$. The number v of sign variations satisfies*

$$v = \mathrm{var}(F\left[\tbinom{c_1}{c_2}^n\right],\ldots,F\left[\tbinom{c_1}{c_2}^{n-i}\tbinom{d_1}{d_2}^i\right],\ldots,F\left[\tbinom{d_1}{d_2}^n\right]) \tag{2.8}$$

and is determined uniquely by F and I.

Proof. Proposition 2.11 describes the form of M in terms of $\binom{c_1}{c_2}$ and $\binom{d_1}{d_2}$. By symmetry, we can restrict to the case $M = \left(\begin{smallmatrix}\mu d_1 & \lambda c_1\\ \mu d_2 & \lambda c_2\end{smallmatrix}\right)$ with $\lambda\mu > 0$. We have $v = \mathrm{var}(g_0,\ldots,g_n)$, where, by Lemma 2.13,

$$g_i = \binom{n}{i} F\left[(M\tbinom{0}{1})^{n-i}(M\tbinom{1}{0})^i\right] = \binom{n}{i}\lambda^{n-i}\mu^i F\left[\tbinom{c_1}{c_2}^{n-i}\tbinom{d_1}{d_2}^i\right].$$

Thus, the sequence $(g_0,\ldots,g_n)$ defining v equals the sequence in (2.8) up to multiplication by factors $\binom{n}{0}\lambda^n$, $\binom{n}{1}\lambda^{n-1}\mu$, …, $\binom{n}{i}\lambda^{n-i}\mu^i$, …, $\binom{n}{n}\mu^n$. Since $\mu/\lambda > 0$, we see that all these factors have the same sign and therefore do not change the number of sign variations.

The various possible choices of $\binom{c_1}{c_2}$, $\binom{d_1}{d_2}$ and M lead to different factors λ and μ, but do not – by Propositions 2.10(iii) and 2.11 – change the sign of $\lambda\mu$. □

Replacing $\binom{d_1}{d_2}$ by $\binom{-d_1}{-d_2}$ (or equivalently, $\binom{c_1}{c_2}$ by $\binom{-c_1}{-c_2}$) switches from I to its complementary projective open interval and flips every other sign in (2.8). This is the generalization of using Descartes' Rule on $A(-X)$ to count negative roots of $A(X)$.

Let us summarize: We have seen an extension of Descartes' Rule to projective open intervals, with two equivalent ways of obtaining the sequence in which to count sign variations. Jacobi's approach (2.7) is more natural if you think of a polynomial as a coefficient sequence w.r.t. the fixed basis $1, X, X^2, \ldots, X^n$. In §3.2.4, we will encounter the original form of the Descartes method, which takes this point of view. The seemingly more complicated approach of (2.8), however, turns out to be more natural for the form of the Descartes method using the Bernstein basis, which we will meet in §3.2.5 and on which our extension to bitstream coefficients rests. We introduce this basis now.

2.2.4 The Bernstein basis

Consider a real polynomial $F(X)$ of degree n and its real roots, counted with multiplicities. Let (c, d) be a bounded open interval that contains exactly p of these roots. We can dehomogenize Theorem 2.17 and Corollary 2.18 to the following statement: *If*

$$v = \operatorname{var}(F[(c)^n], \ldots, F[(c)^{n-i}(d)^i], \ldots, F[(d)^n]), \tag{2.9}$$

then $v \geq p$ *and* $v \equiv p \pmod 2$*, and if all roots of* F *are real, then* $v = p$*.* The condition $(c + d)/2 \in (c, d)$ of Corollary 2.18 becomes vacuous in the affine setting, of course.

In order to make this form of Descartes' Rule easy to apply, we will now determine a basis for the vector space of polynomials with degree at most n such that the coefficient sequence of a polynomial F w.r.t. that basis is the sequence in (2.9). To do so, we "multiply out" the diagonal $F(X) = F[(X)^n]$ of the affine polar form after expressing the indeterminate point X on the affine line as affine combination of c and d:

$$F(X) = F[(\tfrac{d-X}{d-c}c + \tfrac{X-c}{d-c}d)^n] = \sum_{i=0}^{n} F[(c)^{n-i}(d)^i]\binom{n}{i}\frac{(X-c)^i(d-X)^{n-i}}{(d-c)^n}. \tag{2.10}$$

Definition 2.19. Let $c, d \in \mathbb{R}$, $c \neq d$, $n \in \mathbb{N}_0$ and $0 \leq i \leq n$. The polynomial

$$B_i^n[c, d](X) := \binom{n}{i}\frac{(X-c)^i(d-X)^{n-i}}{(d-c)^n} \tag{2.11}$$

is the ith *Bernstein polynomial* of degree n w.r.t. the interval $[c, d]$. $(B_0^n[c, d], \ldots, B_n^n[c, d])$ is the *Bernstein basis* of degree n w.r.t. $[c, d]$. We also write $B_i^n(X)$ for $B_i^n[0, 1](X)$.

We remark that $(B_0^n[c, d], \ldots, B_n^n[c, d])$ is indeed a basis of the vector space of polynomials of degree at most n, since it is a generating set of this vector space by (2.10), and it has the minimum cardinality $n + 1$.

Almost always, we will use Bernstein bases for $c < d$; in this case, $[c, d]$ is indeed an interval. If $c > d$, speaking of "the interval $[c, d]$" is an abuse of terminology; however, we wish to define $B_i^n[c, d](X)$ nevertheless to avoid case distinctions in situations where the boundaries c and d are given by expressions without a fixed order.

Proposition 2.20. *Let F be an affine polar form of degree n whose diagonal is the polynomial $F(X) = \sum_{i=0}^{n} b_i B_i^n[c,d](X)$.*

(i) The coefficient of $B_i^n[c,d](X)$ in $F(X)$ is $b_i = F[(c)^{n-i}(d)^i]$; we call b_i the ith Bernstein coefficient of F w.r.t. $[c,d]$.

(ii) In particular, the first and last Bernstein coefficients are values of the polynomial F: $b_0 = F(c)$ and $b_n = F(d)$.

(iii) The derivative of $F(X)$ is

$$F'(X) = \frac{n}{d-c} \sum_{i=0}^{n-1} \Delta b_i B_i^{n-1}[c,d](X), \quad \text{where } \Delta b_i = b_{i+1} - b_i. \tag{2.12}$$

Proof. Ad (i) & (ii). Immediate from (2.10).

Ad (iii). Passing to the homogeneous polar form F, we can apply Lemma 2.14 to obtain $F'(X) = n \cdot F\big[\binom{X}{1}^{n-1}\binom{1}{0}\big] = n/(d-c) \cdot \Big(F\big[\binom{X}{1}^{n-1}\binom{d}{1}\big] - F\big[\binom{X}{1}^{n-1}\binom{c}{1}\big]\Big)$, from which the claim follows using (i). □

As immediate consequence of Proposition 2.20(i), we get the following equivalence between an affine transformation of the indeterminate and a change of Bernstein basis interval.

Lemma 2.21. *Let $c \neq d$. Consider an affine polar form F of degree n and an affine transformation M. We write $M(c) = p$ and $M(d) = q$. Let $G(X) = F(M(X))$. The coefficient of $B_i^n[c,d](X)$ in $G(X)$ is*

$$G[(c)^{n-i}(d)^i] = F[(M(c))^{n-i}(M(d))^i] = F[(p)^{n-i}(q)^i] \tag{2.13}$$

and thus equal to the coefficient of $B_i^n[p,q](X)$ in $F(X)$.

By construction of the Bernstein basis, we attain the following form of Descartes' Rule.

Theorem 2.22 (Descartes' Rule, Bernstein form). *Consider a real polynomial $F(X) = \sum_{i=0}^{n} b_i B_i^n[c,d](X)$ and its real roots, counted with multiplicities. Let exactly p of them lie in the open interval (c,d). Let $v = \operatorname{var}(b_0, \ldots, b_n)$. Then $v \geq p$ and $v \equiv p \pmod 2$. If all roots of $A(X)$ are real, then $v = p$.*

This relation between Descartes' Rule and the Bernstein basis is not at all new. Pólya and Schoenberg [PS58, §7] (equivalently, [Sch59, §1]) used it in 1958 to show that Bernstein approximation of functions (not explained here, the original motivation of Bernstein[3] [Ber12] to consider the polynomials $B_i^n(X)$) is variation-diminishing.

The notion of a Bézier curve [Far97] [PBP02] provides a geometric interpretation for the Bernstein coefficients of a real polynomial $F(X) = \sum_{i=0}^{n} b_i B_i^n[c,d](X)$. We mention this here for completeness but will not rely on it; the reader will find more explanations and pictures in any of the textbooks [BPR06, §10.2] [Far97, §5.5] [PBP02, §2.8]. A *Bézier curve* with parameter interval $[c,d]$ is a parametric planar curve $\mathbf{b}(t) = \sum_{i=0}^{n} \mathbf{b}_i B_i^n[c,d](t)$ with $\mathbf{b}_0, \ldots, \mathbf{b}_n \in \mathbb{R}^2$. The graph $[c,d] \to \mathbb{R}^2$, $t \mapsto \binom{t}{F(t)}$ of $F(X)$ over $[c,b]$ can be expressed

[3] Sergei Natanovich Bernstein / Сергей Натанович Бернштейн (1880–1968), prominent Ukrainian mathematician, member of the Soviet Academy of Sciences, major contributor to approximation theory. His discoveries are so well-known under the name Bernstein that we refrain from using a systematic transliteration like Bernshteĭn. An English translation of the obituary by Aleksandrov et al. appeared in *Russian Mathematical Surveys* **24** (1969), pp. 169-176 (cited after Zbl 0197.26904).

as a Bézier curve by setting the second coordinate of each $\mathbf{b}_i$ to b_i and the first coordinate to the ith Bernstein coefficient of the identity map, which is $((n-i)d+ic)/n$. The *control polygon* of $\mathbf{b}(t)$ is the polyline $\overline{\mathbf{b}_0\mathbf{b}_1} \cup \overline{\mathbf{b}_1\mathbf{b}_2} \cup \cdots \cup \overline{\mathbf{b}_{n-1}\mathbf{b}_n}$. The curve $\mathbf{b}(t)$ loosely follows the control polygon. Theorem 2.22 states that the number p of intersections of $\mathbf{b}(t)$ and the X-axis, counted with multiplicities, does not exceed the number v of crossings of the control polygon over the X-axis.

In the sequel, we will refer uniformly to all the equivalent generalizations of Descartes' Rule to affine open intervals in the following way.

Definition 2.23. Given a real polynomial $F(X)$ and two real numbers $c < d$, the *Descartes test* for roots in the interval (c, d) is the number $v \in \mathbb{N}_0$ of sign variations counted for (c, d) in the equivalent formulations of Theorem 2.17, Corollary 2.18, and Theorem 2.22. We write $v = \mathit{DescartesTest}(F, (c, d))$.

We conclude our introduction of the Bernstein basis by expressing Bernstein coefficients in terms of roots.

Proposition 2.24. *Consider a polynomial* $F(X) = a_n \prod_{j=1}^{n}(X - \vartheta_j) = \sum_{i=0}^{n} b_i B_i^n[c, d](X)$ *with roots* $\vartheta_1, \ldots, \vartheta_n \in \mathbb{C}$. *Its Bernstein basis coefficients satisfy*

$$b_i = (-1)^{n-i} a_n \sum_{\#J=n-i} \prod_{j\in J}(\vartheta_j - c) \prod_{j\notin J}(d - \vartheta_j) \Big/ \binom{n}{i} \qquad \text{for} \quad 0 \le i \le n,$$

where the sum is taken over all subsets J *of* $\{1, \ldots, n\}$ *with cardinality* $n - i$.

Proof. We have $b_i = F[(c)^{n-i}(d)^i] = G\big[\binom{0}{1}^{n-i}\binom{1}{0}^{i}\big]$ with the polar form $G = F \circ M$ and the matrix $M = \left(\begin{smallmatrix} d & c \\ 1 & 1 \end{smallmatrix}\right)$ representing $M(X) = (dX + c)/(X + 1)$. Thus, b_i is the coefficient of $\binom{n}{i}X^i$ in the diagonal of G, which is

$$\begin{aligned} G(X) = (X+1)^n F(\frac{dX+c}{X+1}) &= a_n \prod_{j=1}^{n}((dX + c) - \vartheta_j(X + 1)) \\ &= a_n \prod_{j=1}^{n}((d - \vartheta_j)X - (\vartheta_j - c)). \end{aligned}$$

Multiplying out, one finds that the coefficient of X^i is the claimed value of b_i times $\binom{n}{i}$. □

2.2.5 De Casteljau's algorithm

In our algorithms, we will need a subroutine for the following task: Given a polynomial's coefficients $(F[(c)^{n-i}(d)^i])_i$ w.r.t. $(B_i^n[c, d])_i$ and a number $c < m < d$, compute the coefficients $(F[(c)^{n-i}(m)^i])_i$ and $(F[(m)^{n-i}(d)^i])_i$ w.r.t. $(B_i^n[c, m])_i$ and $(B_i^n[m, d])_i$, respectively. This is called *subdivision* at m. The number m is specified by a parameter $0 < \alpha < 1$ such that $m = (1 - \alpha)c + \alpha d$. De Casteljau's algorithm [BM99] [Far97, §3] [PBP02, §2.3] carries out this task by filling in a triangular array of numbers, labelled as follows:

$$
\begin{array}{ccccccccccc}
b_{0,0} & & b_{0,1} & & b_{0,2} & \dots & b_{0,n-2} & & b_{0,n-1} & & b_{0,n} \\
& b_{1,0} & & b_{1,1} & & \dots & & b_{1,n-2} & & b_{1,n-1} & \\
& & b_{2,0} & & b_{2,1} & \dots & b_{2,n-3} & & b_{2,n-2} & & \\
& & & b_{3,0} & & \dots & & b_{3,n-3} & & & \\
& & & & \ddots & \dots & \iddots & & & & \\
& & & & & b_{n,0} & & & & &
\end{array}
$$

1: **procedure** *DeCasteljau*$((b_0, \dots, b_n), \alpha)$
2: $(b_{0,0}, b_{0,1}, \dots, b_{0,n}) \leftarrow (b_0, \dots, b_n)$; // input goes to top side
3: **for** j **from** 1 **to** n **do**
4: **for** i **from** 0 **to** $n - j$ **do**
5: $b_{j,i} \leftarrow (1-\alpha)b_{j-1,i} + \alpha b_{j-1,i+1}$;
6: **od**;
7: **od**;
8: $(b'_0, b'_1, \dots, b'_n) \leftarrow (b_{0,0}, b_{1,0}, \dots, b_{n,0})$; // left side
9: $(b''_0, b''_1, \dots, b''_n) \leftarrow (b_{n,0}, b_{n-1,1}, \dots, b_{0,n})$; // right side
10: **return** $((b'_j)_{j=0}^n, (b''_i)_{i=0}^n)$;
11: **end procedure**;

An actual implementation will, of course, not store all $(n+2)(n+1)/2$ numbers simultaneously; storage space for $2n+2$ numbers (i.e., output size) suffices.

Proposition 2.25. *Let $c \neq d$ and $0 < \alpha < 1$ and $m = (1-\alpha)c + \alpha d$. Let F be an affine polar form of formal degree n. Consider the execution of de Casteljau's algorithm invoked as $((b'_j)_j, (b''_i)_i) \leftarrow$ DeCasteljau$((b_i)_i, \alpha)$ for $b_i = F[(c)^{n-i}(d)^i]$, $0 \le i \le n$.*

(i) We have $b_{j,i} = F[(c)^{n-(i+j)}(m)^j(d)^i]$ for $0 \le j \le n$ and $0 \le i \le n-j$.

(ii) We have $F(X) = \sum_{i=0}^n b'_i B_i^n[c,m](X) = \sum_{i=0}^n b''_i B_i^n[m,d](X)$.

(iii) $F(\alpha X + (1-\alpha)c) = \sum_{i=0}^n b'_i B_i^n[c,d](X)$ and $F((1-\alpha)X + \alpha d) = \sum_{i=0}^n b''_i B_i^n[c,d](X)$.

Proof. Ad (i). Follows by induction, using the multiaffinity of F.

Ad (ii). Immediate from (i) in conjunction with Proposition 2.20(i).

Ad (iii). Follows from (ii) using Lemma 2.21: The transformation $X \mapsto \alpha X + (1-\alpha)c$ takes $c \mapsto c$ and $d \mapsto m$. Likewise, $X \mapsto (1-\alpha)X + \alpha d$ takes $c \mapsto m$ and $d \mapsto d$. □

We use de Casteljau's algorithm to derive the following result that relates the bound v from Theorem 2.22 for the interval (c,d) to bounds for its parts (c,m), $\{m\}$, and (m,d).

Proposition 2.26 (Subdivision is variation-diminishing). *Consider a real affine polar form F of formal degree n. Let $c < m < d$. Let $k \ge 0$ denote the multiplicity of m as root of the polynomial $F(X)$. Then*

$$\operatorname{var}((F[(c)^{n-i}(d)^i])_{i=0}^n) \ge \operatorname{var}((F[(c)^{n-i}(m)^i])_{i=0}^n) + k + \operatorname{var}((F[(m)^{n-i}(d)^i])_{i=0}^n). \quad (2.14)$$

The difference between both sides is an even number.

We recall from Definition 2.23 that $\operatorname{var}((F[(a)^{n-i}(b)^i])_{i=0}^n) = \mathit{DescartesTest}(F,(a,b))$. The usual wording of this result, e.g., [BPR06, Prop. 10.41], does not contain the term k on the right-hand side.

Proof. Let us first treat the case $k = 0$, i.e., $F(m) \neq 0$. We follow the execution of de Casteljau's algorithm. Consider the rth partial de Casteljau triangle ("the rth de Casteljau trapezoid"?) consisting of rows $j = 0, \ldots, r$ for some $0 \leq r \leq n$ and the sequence $s_r = (b_{0,0}, \ldots, b_{r,0}, \ldots, b_{r,n-r}, \ldots, b_{0,n})$ comprising its left, lower, and right side. Notice that s_0 is the sequence on the left and that s_n is the concatenation of the sequences on the right in (2.14).

Let us show $\mathrm{var}(s_r) \geq \mathrm{var}(s_{r+1})$ and $\mathrm{var}(s_r) \equiv \mathrm{var}(s_{r+1}) \pmod 2$ for all $0 \leq r < n$. Think of the transformation $s_r \to s_{r+1}$ as happening in two stages: First insert between any two $b_{r,i}$, $b_{r,i+1}$ their linear combination $b_{r+1,i} = (1-\alpha)b_{r,i} + \alpha b_{r,i+1}$. Since $1-\alpha$ and α are positive, this does not change the number of sign variations. Then delete all entries $b_{r,i}$ with $0 < i < n-r$. This leaves the number of sign variations unchanged or decreases it by an even number.

In summary, we obtain $\mathrm{var}(s_0) \geq \mathrm{var}(s_n)$ and $\mathrm{var}(s_0) \equiv \mathrm{var}(s_n) \pmod 2$. Duplicating the entry $b_{n,0}$ and breaking up s_n between its two copies leaves the sum of the sign variations unchanged, as $b_{n,0} = F(m) \neq 0$.

Let us now turn to the case $k > 0$. By Proposition 2.25(i) in conjunction with Proposition 2.15 applied to $S = \{\binom{c}{1}, \binom{d}{1}\}$, the last k rows of the de Casteljau triangle (indices $j = n-k+1, \ldots, n$) consist entirely of zeros, whereas the preceding row $(b_{n-k,0}, b_{n-k,1}, \ldots, b_{n-k,k})$ does not. It is easily seen that $(b_{n-k,0}, b_{n-k,1}, \ldots, b_{n-k,k})$ consists of non-zero elements with alternating signs, so $\mathrm{var}(b_{n-k,0}, b_{n-k,1}, \ldots, b_{n-k,k}) = k$. The inductive argument from above shows $\mathrm{var}(s_0) \geq \mathrm{var}(s_{n-k}) = \mathrm{var}(b_{0,0}, \ldots, b_{n-k,0}) + k + \mathrm{var}(b_{n-k,k}, \ldots, b_{0,n})$ with an even difference, and the claim follows. □

When thinking about the graph of $F(X)$ as a Bézier curve, the preceding proposition corresponds to the fact that subdivision pulls the control polygon closer to the curve; in particular, pairs of crossings of the control polygon forth and back over the X-axis that do not correspond to intersections of the graph with the X-axis may disappear in the process.

For later reference, we record the following consequence.

Corollary 2.27. *Let $F(X)$ be a real polynomial. If the pairwise disjoint open intervals $J_1, \ldots, J_\ell$ are subsets of the open interval I, then*

$$\mathit{DescartesTest}(F, I) \geq \sum_{j=1}^{\ell} \mathit{DescartesTest}(F, J_j). \tag{2.15}$$

Proof. Let us write $I = (a, b)$ and $J_j = (c_j, d_j)$ for $1 \leq j \leq \ell$. By a suitable permutation of indices, we attain $a \leq c_1 < d_1 \leq c_2 < d_2 \leq \cdots \leq c_\ell < d_\ell \leq b$. We argue by induction on ℓ, using Proposition 2.26. For the base case $\ell = 1$, we subdivide (a, b) at c_1 and (c_1, b) at d_1 to get $\mathit{DescartesTest}(F, (a, b)) \geq \mathit{DescartesTest}(F, (c_1, d_1))$. In the inductive step, $\mathit{DescartesTest}(F, (c_\ell, d_{\ell+1})) \geq \mathit{DescartesTest}(F, (c_\ell, d_\ell)) + \mathit{DescartesTest}(F, (c_{\ell+1}, d_{\ell+1}))$ allows us to proceed from $(c_1, d_1), \ldots, (c_\ell, d_{\ell+1})$ to $(c_1, d_1), \ldots, (c_\ell, d_\ell), (c_{\ell+1}, d_{\ell+1})$ □

The essence of Proposition 2.26 can be traced back to work of Schoenberg [Sch34] on real root counting that clearly predates de Casteljau's algorithm and the notion of a Bézier curve. We come back to this in Appendix A.1.

2.2.6 Relation between Bernstein and power basis

For constructing the Bernstein basis w.r.t. interval (c, d), our starting point has been Theorem 2.17 in its reformulation (2.9) into the language of polar forms. Let us now return

to the original viewpoint of Theorem 2.17 and describe how the Bernstein coefficients of a polynomial $F(T)$ arise from a Möbius transformation $T = M(X)$ of the indeterminate.

The $[c,d]$-Bernstein coefficients of $F(T)$ are $b_i = F\big[\binom{c}{1}^{n-i}\binom{d}{1}^{i}\big]$ for $0 \le i \le n$. Letting $M = \left(\begin{smallmatrix} c & d \\ 1 & 1 \end{smallmatrix}\right)$, we can rewrite this as $b_{n-i} = F\big[\binom{c}{1}^{i}\binom{d}{1}^{n-i}\big] = F\big[(M\binom{1}{0})^i(M\binom{0}{1})^{n-i}\big]$. The Möbius transformation represented by M is $T = (cX+d)/(X+1)$. Thus, Lemma 2.13 states that b_{n-i} is the coefficient of $\binom{n}{i}X^i$ in $G(X) := (X+1)^n F((cX+d)/(X+1))$. The subsequent proposition records this correspondence and extends it to a correspondence between the polynomials created by subdivision of $F(T)$ at $T = m = (c+d)/2$ and matching transformations of $G(X)$.

Proposition 2.28. *Let $c \neq d$ and $m = (c+d)/2$. Considering subdivision at m, we let $F(T) = \sum_{i=0}^n b_i B_i^n[c,d](T) = \sum_{i=0}^n b'_i B_i^n[c,m](T) = \sum_{i=0}^n b''_i B_i^n[m,d](T)$. It holds that*

(i) $G(X) := (X+1)^n F((cX+d)/(X+1)) = \sum_{i=0}^n b_{n-i}\binom{n}{i}X^i$,

(ii) $C(G(X)) := 2^{-n}G(2X+1) = \sum_{i=0}^n b'_{n-i}\binom{n}{i}X^i$,

(iii) $RCR(G(X)) = \sum_{i=0}^n b''_{n-i}\binom{n}{i}X^i$ *where* $R(A(X)) := X^n A(1/X)$.

Proof. Recall from our discussion above that the homogeneous polar forms F and G satisfy $G = F \circ M$ with $M = \left(\begin{smallmatrix} c & d \\ 1 & 1 \end{smallmatrix}\right)$, and that this implies (i) by Lemma 2.13.

Regarding (ii), we observe that $\binom{m}{1} = \frac{1}{2}(\binom{c}{1} + \binom{d}{1}) = M\binom{1/2}{1/2}$, so that we have $b'_{n-i} = F\big[\binom{c}{1}^{i}\binom{m}{1}^{n-i}\big] = G\big[\binom{1}{0}^{i}\binom{1/2}{1/2}^{n-i}\big] = G\big[(M'\binom{1}{0})^i(M'\binom{0}{1})^{n-i}\big]$ for $M' = \left(\begin{smallmatrix} 1 & 1/2 \\ 0 & 1/2 \end{smallmatrix}\right)$. Using Lemma 2.13, we see that b'_{n-i} is the coefficient of $\binom{n}{i}X^i$ in $(1/2)^n G((X+1/2)/(1/2)) = 2^{-n}G(2X+1)$, as desired.

Finally, $b''_{n-i} = F\big[\binom{m}{1}^{i}\binom{d}{1}^{n-i}\big] = G\big[\binom{1/2}{1/2}^{i}\binom{0}{1}^{n-i}\big] = G\big[(M''\binom{1}{0})^i(M''\binom{0}{1})^{n-i}\big]$ with $M'' = \left(\begin{smallmatrix} 1/2 & 0 \\ 1/2 & 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) M' \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Now claim (iii) follows because the polynomial transformation R is effected by transforming the parameter range with $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. □

2.3 Partial converses for arbitrary open intervals

2.3.1 Circular regions in the complex plane

So far, we have used Möbius transformations of $\mathbb{P}^1(\mathbb{R})$ to generalize Descartes' Rule beyond the open interval $(0,\infty)$. To generalize Obreshkoff's extension of Descartes' Rule (Theorem 2.7) in the same fashion, we need to investigate in §2.3.2 what Möbius transformations of $\mathbb{P}^1(\mathbb{C})$ do to the argument ranges in Obreshkoff's result, and in particular to their boundary rays. In a preparatory step, we briefly review circular regions in the complex plane, cf. [And99].

Definition 2.29. A *projective circle* C is the set of points $[z:w] \in \mathbb{P}^1(\mathbb{C})$ that satisfies

$$\alpha z\overline{z} + cz\overline{w} + \overline{cz}w + \beta w\overline{w} = 0 \quad \text{for some } \alpha,\beta \in \mathbb{R}, c \in \mathbb{C} \text{ such that } c\overline{c} > \alpha\beta. \qquad (2.16)$$

An *open (closed) circular region* R is a set of points $[z:w] \in \mathbb{P}^1(\mathbb{C})$ for which the expression in (2.16) is positive (non-negative).

These definitions are independent of the choice of representatives; replacing $[z:w]$ by $[az:aw]$, $a \in \mathbb{C}^*$, multiplies (2.16) by the constant factor $a\overline{a} > 0$. Clearly, projective circles and closed circular regions are closed subsets, whereas open circular regions are open subsets of $\widehat{\mathbb{C}}$.

The notion of a projective circle C in $\widehat{\mathbb{C}}$ generalizes the usual notion of a *Euclidean circle* in $\mathbb{C}$. Let us consider the finite points $[x + \mathrm{i}y : 1]$ of $\widehat{\mathbb{C}}$ as elements (x, y) of $\mathbb{R}^2$. Equation (2.16) takes the form

$$\alpha(x^2 + y^2) + 2(c_1 x - c_2 y) + \beta = 0 \quad \text{with } \alpha, \beta \in \mathbb{R}, c = c_1 + \mathrm{i}c_2 \in \mathbb{C}, c\overline{c} > \alpha\beta.$$

If $\alpha \neq 0$, then C is a Euclidean circle with center $(-c_1, c_2)/\alpha$ and radius $\sqrt{c\overline{c} - \alpha\beta}\,/\,|\alpha|$. Any Euclidean circle occurs in this way. The condition $c\overline{c} > \alpha\beta$ is equivalent to C not being empty or a single point. If, however, $\alpha = 0$, then C is an affine line plus the point at infinity $[1 : 0]$. Any line occurs in this way. The condition $c\overline{c} > \alpha\beta = 0$ is equivalent to $c \neq 0$, so that C is neither the entire plane ($c = 0$, $\beta = 0$) nor empty ($c = 0$, $\beta \neq 0$). Thus, projective circles are precisely the Euclidean circles together with affine lines, the latter being the "circles through infinity".

The circular region R, regarded in the affine plane $\mathbb{C}$, consists of the points inside the circle C (in case $\alpha < 0$), outside the circle C (in case $\alpha > 0$), or on one side of the line C (in case $\alpha = 0$), respectively. If R is open (closed), then its boundary C is excluded from (included in) R.

The crucial property of projective circles, which is not shared by lines or Euclidean circles considered separately, is expressed by the following classical result.

Proposition 2.30. *The image of a projective circle C under a Möbius transformation M of $\mathbb{P}^1(\mathbb{C})$ is a projective circle.*

Proof. The equation of $M(C)$ is obtained by substituting $M^{-1}([z : w])$ for $[z : w]$ in (2.16). Multiplying out, one sees that the resulting expression is again of the form (2.16); the condition $c\overline{c} > \alpha\beta$ holds because $M(C)$, just like C, is neither empty, nor a single point, nor the entire plane. □

While the natural habitat of Möbius transformations and projective circles is $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, we will encounter them mostly in $\mathbb{C}$. In other words, we simply ignore the point at infinity. This creates a minor problem when we subject C to a Möbius transformation M that takes ∞ to a finite point: the image $M(C)$ may lack the point $M(\infty)$. To fix this, we regard M as taking C to the closure of its pointwise image $M(C)$.

The Möbius transformation $M(X) = (aX + b)/(cX + d)$, $c \neq 0$, is a map $\mathbb{C} \setminus \{-d/c\} \to \mathbb{C} \setminus \{a/c\}$ which is holomorphic (complex differentiable in a neighbourhood of any point in its domain). Therefore, M preserves oriented angles between tangent vectors to curves. (This is almost a tautology: Multiplication with a complex number (complex derivative) can express a linear map $\mathbb{R}^2 \to \mathbb{R}^2$ (real derivative) if and only if that map preserves oriented angles.) This property will help us to describe images of projective circles under Möbius transformations geometrically.

2.3.2 Obreshkoff's partial converse transformed

We are now ready to transfer Obreshkoff's extension of Descartes' Rule (Theorem 2.7) to arbitrary bounded open intervals (c, d). In fact, obvious modifications of the subsequent development allow a generalization to arbitrary projective open intervals, but we intend to keep the exposition as concrete as possible. An example of the following definitions is depicted in Figure 2.2.

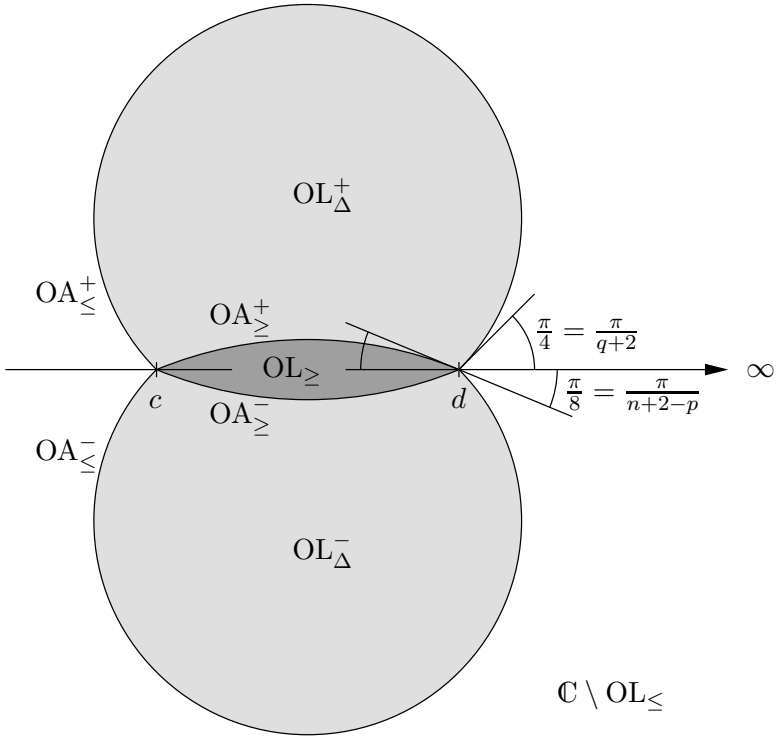


Figure 2.2: The (p, q, n)-Obreshkoff arcs and loci for $n = 8$ and $p = q = 2$.

Definition 2.31. Let $p, q, n \in \mathbb{N}_0$, $p \leq q \leq n$. Let $c, d \in \mathbb{R}$, $c < d$. We define the four (p, q, n)-*Obreshkoff arcs* w.r.t. interval (c, d) as follows: $(\mathrm{OA}^-_\geq, \mathrm{OA}^+_\geq)$ and $(\mathrm{OA}^-_\leq, \mathrm{OA}^+_\leq)$ are pairs of circular arcs in the complex plane. Within each pair, the arcs are symmetric to each other about the real axis. The arcs with superscript $+$ ($-$) run above (below) the real line.

(i) Each of the arcs $\mathrm{OA}^-_\geq$ and $\mathrm{OA}^+_\geq$ joins the points c and d and makes an angle of $\pi/(n + 2 - p)$ with the line segment $[c, d]$ at d.

(ii) Each of the arcs $\mathrm{OA}^-_\leq$ and $\mathrm{OA}^+_\leq$ joins the points c and d and makes an angle of $\pi/(q + 2)$ with the ray $[d, \infty)$ at d.

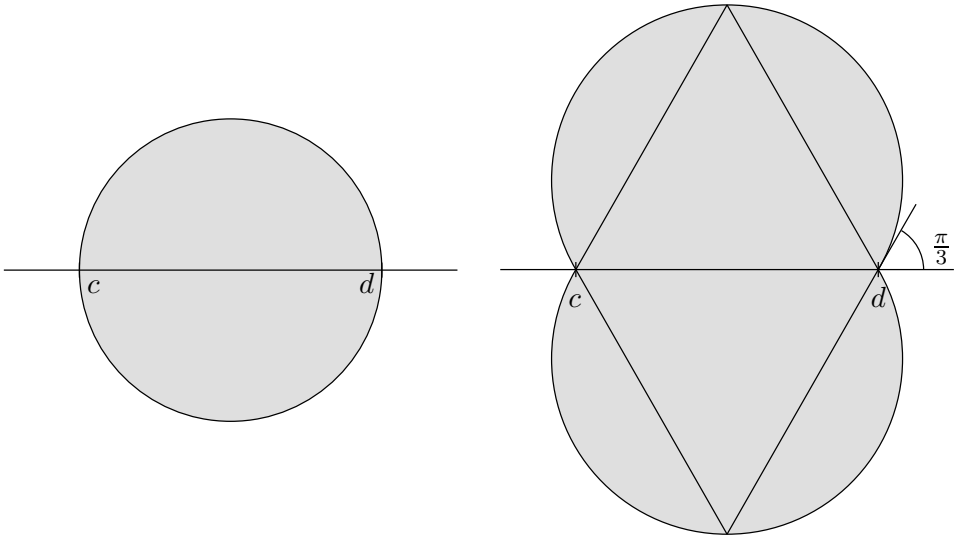
For each Obreshkoff arc $\mathrm{OA}^\pm_\bowtie$, we denote by $\mathrm{OC}^\pm_\bowtie$ its supporting circle and by $\mathrm{OD}^\pm_\bowtie$ the open disc within that circle. We proceed to define several (p, q, n)-*Obreshkoff loci* (OL) in terms of these discs.

(iii) The (p, q, n)-*Obreshkoff lens* is $\mathrm{OL}_\geq = \mathrm{OD}^-_\geq \cap \mathrm{OD}^+_\geq$; or equivalently, the open region that contains (c, d) and has the boundary $\mathrm{OA}^-_\geq \cup \mathrm{OA}^+_\geq$.

(iv) The (p, q, n)-*Obreshkoff range* is $\mathrm{OL}_\leq = \mathrm{OD}^-_\leq \cup \mathrm{OD}^+_\leq$; or equivalently, the open region that contains (c, d) and has the boundary $\mathrm{OA}^-_\leq \cup \mathrm{OA}^+_\leq$.

(v) The set difference $\mathrm{OL}_\Delta := \mathrm{OL}_\leq \setminus \mathrm{OL}_\geq$ consists of two connected components: the *upper* (p, q, n)-*Obreshkoff lune* OL^+_Δ with boundary $\mathrm{OA}^+_\geq \cup \mathrm{OA}^+_\leq$, including the relative interior of $\mathrm{OA}^+_\geq$ but excluding $\mathrm{OA}^+_\leq$ and the points c, d; and the *lower* (p, q, n)-*Obreshkoff lune* OL^-_Δ defined symmetrically in terms of $\mathrm{OA}^-_\geq$ and $\mathrm{OA}^-_\leq$.

With any of these symbols, we specify the parameters (p, q, n) and (c, d) in parentheses as in $\mathrm{OA}^\pm_\bowtie(p, q, n; (c, d))$ where the necessity arises.



(a) One circle $\mathrm{OC}^{\pm}_{\leq}(0,0,n)$. **(b)** Two circles $\mathrm{OC}^{\pm}_{\leq}(1,1,n)$.

Figure 2.3: The circles from Propositions 2.33 and 2.34.

By elementary geometry, a circular arc makes the same angle with its chord at both endpoints, hence the conditions above on angles at d apply mutatis mutandis at c, and we see that each of the four Obreshkoff arcs is symmetric about the perpendicular bisector of the chord $[c,d]$.

Theorem 2.32 (Obreshkoff). *Consider the real polynomial $F(X)$ of degree n and its roots in the complex plane, counted with multiplicities. Let $v = \mathit{DescartesTest}(F,I)$, where I is a bounded open interval. If $F(X)$ has at least p roots in the Obreshkoff lens $\mathrm{OL}_{\geq}(p,q,n;I)$ and at most q roots in the Obreshkoff range $\mathrm{OL}_{\leq}(p,q,n;I)$, then $v \geq p$ and $v \leq q$.*

If, in particular, $q = p$, meaning that $A(X)$ has exactly p roots in the Obreshkoff lens $\mathrm{OL}_{\geq}(p,q,n;I)$ and none in the Obreshkoff lunes $\mathrm{OL}_{\Delta}(p,q,n;I)$, then $v = p$.

In more vivid terms, we claim that the Descartes test does see everything in the lens, does not see anything beyond its range, and may see some complex-conjugate couples lingering in lunar twilight. Let us now demonstrate this nyctalopic myopia.

Proof. We take a viewpoint akin to Theorem 2.17 and consider an arbitrary real Möbius transformation M with $M((0,\infty)) = (c,d)$ and, w.l.o.g., $M(0) = d$. Let us inspect how M maps the argument ranges from Theorem 2.7. (See Figure 2.2, compare to Figure 2.1 on page 15.) The range for arguments φ is the intersection of two open half-spaces, both of which contain $(0,\infty)$, and its boundary rays make an angle of $\pi/(n+2-p)$ with $[0,\infty)$ at apex 0. Since M preserves angles, their images, which are arcs of projective circles that join c and d, make an angle of $\pi/(n+2-p)$ with $(c,d]$ at apex d, and we see that these are in fact the Obreshkoff arcs $\mathrm{OA}^{\pm}_{\geq}$. Arguing in the same way, we see that the boundary rays of the range for arguments ψ are mapped to $\mathrm{OA}^{\pm}_{\leq}$. Now the theorem follows at once from Theorem 2.7. □

The special cases $p = q = 0$ and $p = q = 1$ (cf. Propositions 2.8 and 2.9) will be of particular importance later on, so we formulate them separately. Figure 2.3 depicts the circles referred to.

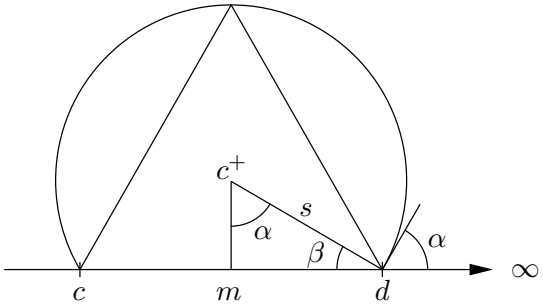


Figure 2.4: Proofs of Propositions 2.34 and 2.35.

Proposition 2.33 (case $p = q = 0$: "one-circle theorem"). *Let (c, d) be a bounded open interval, and let $D \subseteq \mathbb{C}$ be the open disc within the circumcircle C of the line segment $[c, d]$. For any $n \in \mathbb{N}$, it holds that $\mathrm{OL}_{\leq}(0, 0, n; (c, d)) = D$. In particular, if D does not contain any root of the real polynomial $F(X)$, then DescartesTest$(F, (c, d)) = 0$.*

Proof. The two $(0, 0, n)$-Obreshkoff arcs $\mathrm{OA}^-_{\leq}$, $\mathrm{OA}^+_{\leq}$ form an angle of $\pi/2$ with the real axis, so that $C = \mathrm{OA}^-_{\leq} \cup \mathrm{OA}^+_{\leq}$ and $D = \mathrm{OL}_{\leq}(0, 0, n)$. □

Proposition 2.34 (case $p = q = 1$: "two-circle theorem"). *Let (c, d) be a bounded open interval. Let Δ^+ and Δ^- be the equilateral triangles in the upper and lower half, resp., of the complex plane that have $[c, d]$ as one edge. Let $C^{\pm}$ be the circumcircle of $\Delta^{\pm}$, and let $D^{\pm}$ be the open disc within $C^{\pm}$. For any $n \in \mathbb{N}$, it holds that $\mathrm{OL}_{\leq}(1, 1, n; (c, d)) = D^+ \cup D^-$. In particular, if $D^+ \cup D^-$ contains exactly one simple root of the real polynomial $F(X)$, then DescartesTest$(F, (c, d)) = 1$.*

The root in question is necessarily real, as it lacks a complex conjugate.

Proof. The boundary of $D^+ \cup D^-$ consists of an arc A^+ of C^+ and an arc A^- of C^-, each of which joins the points c and d. Let us inspect the angle α (see Figure 2.4) formed by A^+ with the ray $[d, \infty)$ at apex d. The line segment s between d and the center c^+ of C^+ bisects the $\pi/3$ angle at d of the equilateral triangle Δ^+, so s forms an angle $\beta = \pi/6$ with the line segment $[c, d]$. On the other hand, the radius s is perpendicular to the tangent of C^+ at d, so $\alpha + \beta = \pi/2$, and we obtain $\alpha = \pi/3$. The symmetric argument holds for A^-. It follows that A^+ and A^- are the two $(1, 1, n)$-Obreshkoff arcs $\mathrm{OA}^+_{\leq}$ and $\mathrm{OA}^-_{\leq}$, so $D^+ \cup D^- = \mathrm{OL}_{\leq}(1, 1, n)$. □

Historically, the main challenge in obtaining these results was to come up with partial converses of Descartes' Rule for the untransformed interval $(0, \infty)$; we refer to §2.1.2 and §2.1.3 for a discussion. The statement of Proposition 2.33 was already used by Vincent [Vin36, p. 345]. Obreshkoff also transformed his extension of Descartes' Rule to arbitrary affine open intervals [Obr63, Satz 17.5]. In the context of analyzing the Descartes method, Proposition 2.34 has first been formulated and used by Krandick and Mehlhorn [KM06], who have obtained it from Proposition 2.9 (which they attribute to Ostrowski [Ost50]) by arguments similar to ours in the proof of Theorem 2.32. This potential use of Ostrowski's result is also mentioned but not carried out in the thesis of Batra [Bat99, p. 18].

Later on, we will need the following geometric facts to relate Obreshkoff's partial converse to the distances of roots.

Proposition 2.35. *Consider the (p,q,n)-Obreshkoff arcs for a bounded open interval I.*

(i) The diameter of the circles $\mathrm{OC}^{\pm}_{\leq}(p,q,n;I)$ is $D(q) = |I| \,/\, \sin(\pi/(q+2))$.

(ii) Every point of I has distance less than $D(q)$ to any point in $\mathrm{OL}_{\leq}(p,q,n;I)$.

(iii) $D(0) = |I|$ and $D(1) = |I| \cdot 2/\sqrt{3}$.

(iv) Asymptotically, $D(q) = |I| \cdot \Theta(q)$ for $q \to \infty$.

Proof. Let $I = (c,d)$ and $m = (c+d)/2$. Let $\alpha = \pi/(q+2)$. Let c^+ be the center of $\mathrm{OC}^+_{\leq}$ (see Figure 2.4). The part of the tangent to $\mathrm{OC}^+_{\leq}$ at d that extends upwards makes an angle α with the ray $[d,\infty)$ and an angle $\pi/2$ with the radius s between d and c^+. Hence the line segments from c^+ to m and d, resp., also make an angle α, so that $\sin(\alpha) = |d-m| \,/\, |d-c^+| = |I| \,/ D(q)$, from which (i) follows for $\mathrm{OC}^+_{\leq}$. The claim for $\mathrm{OC}^-_{\leq}$ follows by symmetry.

Claim (ii) is immediate, noting that I is inside the circles $\mathrm{OC}^{\pm}_{\leq}$, which have diameter $D(q)$. Claim (iii) follows from $\sin(\pi/2) = 1$ and $\sin(\pi/3) = \sqrt{3}/2$, claim (iv) from $\lim_{x\to 0} \sin(x)/x = 1$. □

In the analysis of root isolation algorithms, we are able to use the results of this section in the following form. For the standard Descartes method, of course only the case $k = 1$ is relevant, but we will also meet the general case $k \geq 1$ again, namely in §3.4.4.

Proposition 2.36. *Let the real polynomial F of degree n have a k-fold root α in the bounded open interval I, and let s be the minimum distance from α to any other root $\beta \neq \alpha$ of F. If $s \geq |I| \,/\, \sin(\pi/(k+2))$, then DescartesTest$(F,I) = k$.*

Proof. By Proposition 2.35(ii), no complex root $\beta \neq \alpha$ of F is contained in the (k,k,n)-Obreshkoff range of I, and the claim follows from Theorem 2.32. □

2.3.3 A partial converse by differentiation

> For reasons to be discussed in §2.3.5, we will not make use of the following results in later chapters, so the impatient reader with an exclusive interest in algorithms may skip ahead to §2.4 on page 40; the mathematically inclined reader is encouraged to keep reading, not least because of the result in §2.3.4, which is of independent interest and uses a nice proof technique.

Before becoming aware of Obreshkoff's result, the author discovered another partial converse of Descartes' Rule. It is presented here in a slightly more general way than in the original publication [Eig07], taking advantage of polar forms.

Lemma 2.37. *Let F be a real homogeneous polar form of degree n. Let $\binom{c_1}{c_2}, \binom{d_1}{d_2}, \binom{x_1}{y_1} \in \mathbb{R}^2$ such that $\binom{c_1}{c_2}$, $\binom{d_1}{d_2}$ are linearly independent and $\binom{x_1}{y_1} = \lambda\binom{c_1}{c_2} + \mu\binom{d_1}{d_2} \neq 0$ with $\lambda\mu \leq 0$. Then* $\mathrm{var}((F[\binom{c_1}{c_2}^{n-i}\binom{d_1}{d_2}^{i}])_{i=0}^{n}) \leq \mathrm{var}((F[\binom{x_1}{y_1}\binom{c_1}{c_2}^{n-1-i}\binom{d_1}{d_2}^{i}])_{i=0}^{n-1}) + 1$.

Proof. It is no restriction to assume $\lambda > 0$, $\mu \leq 0$, because we can exchange the roles of $\binom{c_1}{c_2}$ and $\binom{d_1}{d_2}$, and we can replace $\binom{x_1}{y_1}$ by $\binom{-x_1}{-y_1}$ without changing the statement. For brevity, we set $p_i = F[\binom{c_1}{c_2}^{n-i}\binom{d_1}{d_2}^{i}]$ and $q_i = F[\binom{x_1}{y_1}\binom{c_1}{c_2}^{n-1-i}\binom{d_1}{d_2}^{i}]$. We have $q_i = \lambda p_i + \mu p_{i+1}$. It may be helpful to visualize these two sequences arranged like this:

$$\begin{array}{ccccccccc} p_0 & & p_1 & & p_2 & \cdots & p_{n-1} & & p_n \\ & q_0 = \lambda p_0 + \mu p_1 & & q_1 = \lambda p_1 + \mu p_2 & & \cdots & & q_{n-1} = \lambda p_{n-1} + \mu p_n & \end{array}$$

Each sign variation in $(p_0, \dots, p_n)$ is an index pair $0 \le i < j \le n$ such that $p_i p_j < 0$ and $p_{i+1} = \cdots = p_{j-1} = 0$. Let there be exactly v such pairs $(i_1, j_1), \dots, (i_v, j_v)$ with indices $i_1 < j_1 \le i_2 < j_2 \le \cdots \le i_v < j_v$. Sign variations are either "positive to negative" ($p_{i_\ell} > 0$) or "negative to positive" ($p_{i_\ell} < 0$). Obviously, these types alternate. If $p_{i_\ell} > 0$, then $p_{i_\ell+1} \le 0$ and thus $q_{i_\ell} = \lambda p_{i_\ell} + \mu p_{i_\ell+1} > 0$. Similarly, if $p_{i_\ell} < 0$ then $q_{i_\ell} < 0$. Hence the sequence $(q_0, \dots, q_{n-1})$ contains an alternating subsequence $\operatorname{sgn}(q_{i_1}) \ne \operatorname{sgn}(q_{i_2}) \ne \cdots \ne \operatorname{sgn}(q_{i_v})$, so $(q_0, \dots, q_{n-1})$ has at least $v-1$ sign variations. □

Theorem 2.38. *Let $\binom{c_1}{c_2}, \binom{d_1}{d_2} \in \mathbb{R}^2$ be linearly independent. Let I be the projective open interval that has endpoints $[c_1 : c_2]$, $[d_1 : d_2]$ and contains $[c_1 + d_1 : c_2 + d_2]$. Let $\binom{x_1}{y_1}, \dots, \binom{x_r}{y_r} \in \mathbb{R}^2$ be non-zero vectors such that $[x_1 : y_1], \dots, [x_r : y_r] \notin I$. Consider a real homogeneous polar form F of degree n and the polar form G of degree $n-r$ defined as $G\left[\binom{X_{r+1}}{Y_{r+1}} \cdots \binom{X_n}{Y_n}\right] := F\left[\binom{x_1}{y_1} \cdots \binom{x_r}{y_r}\binom{X_{r+1}}{Y_{r+1}} \cdots \binom{X_n}{Y_n}\right]$. Write v_F and v_G for the number of sign variations counted by Corollary 2.18 for of F and G, respectively. Then $v_F \le v_G + r$.*

Proof. We saw in the proof of Proposition 2.10(i) that every representative of $[x_i : y_i] \notin I$ has the form $\lambda\binom{c_1}{c_2} + \mu\binom{d_1}{d_2}$ with $\lambda\mu \le 0$, so the theorem follows from the preceding lemma by induction. □

We will use only a special case of this result in the sequel. If I is an affine open interval, then, by definition, it does not contain the point $[1 : 0]$ at infinity. Hence we can set all $\binom{x_i}{y_i} = \binom{1}{0}$ and, by Lemma 2.14, G is the rth derivative of F with respect to X.

Corollary 2.39. *Consider the polynomial F of degree n, its rth derivative $F^{(r)}$, $r \le n$, and an affine open interval I. We have $\mathit{DescartesTest}(F, I) \le r + \mathit{DescartesTest}(F^{(r)}, I)$.*

How does Corollary 2.39 compare with Obreshkoff's partial converse, Theorem 2.32? Consider a real polynomial $F(X)$ and a bounded open interval I that contains exactly p roots, counted with multiplicities.

If all roots of F are real, then Theorem 2.32 reduces immediately to the optimal statement $\mathit{DescartesTest}(F, I) = p$, whereas the result above imposes significant additional conditions on roots of some derivative. Thus, Obreshkoff's partial converse is not implied by ours.

On the other hand, we can demonstrate by means of an explicit example that Corollary 2.39 is not implied by Obreshkoff's result. We aim for a minimal example, in the sense that we fix r at the lowest non-trivial value $r = 1$ and intend to choose the degree n as small as possible. Figure 2.5 shows an example for $n = 3$: The polynomial F has a real root in $I = (0, 1)$ and two further imaginary roots in $\mathrm{OL}_{\le}(1, 1, 3; I)$. The roots of F' are outside $\mathrm{OL}_{\le}(0, 0, 3; I)$, so $\mathit{DescartesTest}(F', I) = 0$ by Proposition 2.33, and Corollary 2.39 with $r = 1$ implies $\mathit{DescartesTest}(F, I) \le 1$ (the optimal upper bound). On the other hand, Obreshkoff's result for $q = 1$ (Proposition 2.34) is not applicable: the two imaginary roots of F are in the way.

We cannot give an example of smaller degree: By the preceding deliberations, we need imaginary roots. This rules out $n = 1$ right away. What about $n = 2$? Can we define F by a pair of complex-conjugate roots $x \pm \mathrm{i}y$ that are inside the circumcircle of I (its $(0, 0, n)$-Obreshkoff range) while the root of F' is not? No, because the root of F' is the real part x. (More generally, the roots of F' are contained in the convex hull of the roots of F by the Gauss-Lucas theorem [RS02, Thm. 2.1.1].) So the example above is indeed of minimal degree.

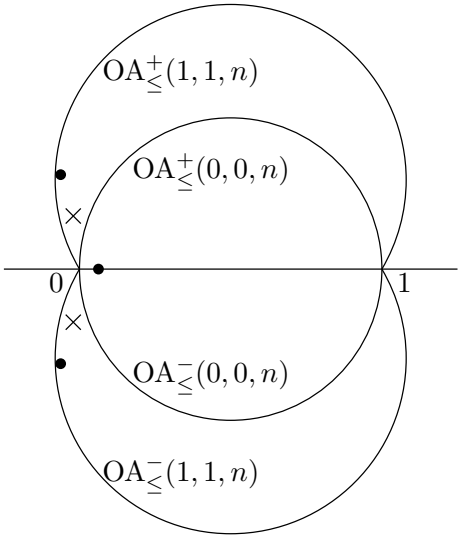


Figure 2.5: $F(X) = 2048X^3 + 128X^2 + 192X - 13$ has roots $1/16$, $(-1 + 5\mathrm{i})/16$, $(-1 - 5\mathrm{i})/16$ (shown as dots). Its derivative $F'(X)$ has roots $(-1 \pm \sqrt{-71})/48 \approx -0.0208 \pm 0.1755\,\mathrm{i}$ (crosses).

The statement of Corollary 2.39 is somewhat indirect: It bounds one Descartes test in terms of another one. By choosing r and imposing conditions that restrict the value of $\mathit{DescartesTest}(F^{(r)}, I)$, one can easily deduce various more concrete partial converses. We formulate some for later discussion.

Proposition 2.40. *Let the polynomial $F(X)$ have at least p roots, counted with multiplicities, in the bounded open interval I. If the open disc within the circumcircle of I does not contain any roots of $F^{(p)}$, then $\mathit{DescartesTest}(F, I) = p$.*

Of course, p is then the exact number of roots in I.

Proof. By Proposition 2.33, the condition on $F^{(p)}$ guarantees $\mathit{DescartesTest}(F^{(p)}, I) = 0$. The claim follows at once. □

If p is known a priori to be the exact number of roots in I, one can take advantage of the fact that the error in Descartes' Rule is an even number: $\mathit{DescartesTest}(F, I) \leq p + 1$ suffices to deduce $\mathit{DescartesTest}(F, I) = p$.

Proposition 2.41. *Let the polynomial $F(X)$ have exactly p roots, counted with multiplicities, in the bounded open interval I. If one of the following conditions holds, then $\mathit{DescartesTest}(F, I) = p$.*

(i) Proposition 2.33 applied to $F^{(p+1)}$ yields $\mathit{DescartesTest}(F^{(p+1)}, I) = 0$.
(ii) Proposition 2.33 or 2.34 applied to $F^{(p)}$ yields $\mathit{DescartesTest}(F^{(p)}, I) \leq 1$.

With a view towards the analysis of root isolation algorithms, we want to turn these results into a partial converse in terms of roots of $F(X)$ alone, and not its derivatives, analogous to Proposition 2.36. The next section paves the way for this.

2.3.4 Distance of roots to roots of derivatives

Throughout this section, we consider a real polynomial F of degree n and its k-fold root α, and we discuss the question: If the minimum distance of α to any other root of F is s,

and if the minimum distance of α to any root $\xi \neq \alpha$ of $F^{(r)}$ is σ for some fixed $r < n$, how can we bound the ratio σ/s from below?

We begin with a simple, negative result: There is no such lower bound for $r > k$ in general. A counterexample is provided by the polynomials

$$A_\varepsilon(X) = (X + 1 + \varepsilon)^\ell (X - 1)^\ell X^k \tag{2.17}$$

of degree $n = k + 2\ell$. For any $\varepsilon \geq 0$, the root $x = 0$ of $A_\varepsilon(X)$ has multiplicity k, and its distance to the nearest other root, namely $x = 1$, is 1. Let us first set $\varepsilon = 0$. Since $((X + 1)(X - 1))^\ell = (X^2 - 1)^\ell$, the coefficient of X^{k+i} in $A_\varepsilon(X)$ is zero for odd i. So if $r = k + 2j + 1 < n$, $j \in \mathbb{N}_0$, then $A_0^{(r)}(0) = 0$. Let us now take an arbitrarily small $\varepsilon > 0$. As complex roots depend continuously on the coefficients [RS02, §1.3], the polynomial $A_\varepsilon^{(r)}(X)$ has a root arbitrarily close to $x = 0$, but the coefficient of X^r in $A_\varepsilon(X)$ and thus the value $A_\varepsilon^{(r)}(0)$ is non-zero for sufficiently small $\varepsilon > 0$. In particular, the k-fold root $\alpha = 0$ of $F(X)$ can be arbitrarily close (measured in multiples of s) to the nearest root $\xi \neq \alpha$ of $F^{(k+1)}(X)$. This means that Proposition 2.41(i) cannot be turned into a partial converse in terms of s.

The main result of this section is a lower bound on σ/s for the case $r \leq k$. The relative position of roots of F and $F^{(r)}$ is invariant under translations, so it is no disadvantage for the generality of our considerations to assume $\alpha = 0$. In this situation, we define polynomials G and H that have exactly those roots of F and $F^{(r)}$, resp., different from α:

$$\begin{aligned} F(X) &= \sum_{i=k}^{n} f_i X^i = X^k G(X) && \text{with } G(X) = \sum_{i=0}^{n-k} g_i X^i, \\ F^{(r)}(X) &= \sum_{i=k-r}^{n-r} \frac{(i+r)!}{i!} f_{i+r} X^i = X^{k-r} H(X) && \text{with } H(X) = \sum_{i=0}^{n-k} h_i X^i. \end{aligned} \tag{2.18}$$

The coefficients of G and H are related by

$$h_i = \frac{(i+k)!}{(i+k-r)!} f_{i+k} = \frac{(i+k)!}{(i+k-r)!} g_i. \tag{2.19}$$

Generalizing an approach of Dimitrov [Dim98] from the special case $r = 1$ to the general case $1 \leq r \leq k$, we use the following theorem to track how multiplying the coefficients in (2.19) changes the roots of G into those of H.

Theorem 2.42 (Schur-Szegő composition theorem). *Consider $A(X) = \sum_{i=0}^n a_i \binom{n}{i} X^i$, $B(X) = \sum_{i=0}^n b_i \binom{n}{i} X^i$, and $C(X) = \sum_{i=0}^n a_i b_i \binom{n}{i} X^i$. Let K be a closed circular region in the complex plane containing all roots of A. If ξ is a root of C, there is an element $w \in K$ and a root β of B such that $\xi = -w\beta$.*

Szegő formulates and proves this theorem as "Satz 2" in the single-authored paper [Sze22] as consequence of a "Faltungssatz" (convolution theorem) but remarks on p. 35: "This formulation of the convolution theorem was pointed out to me by Mr. I. Schur." We follow Rahman/Schmeisser [RS02, Thm. 3.4.1d] and use both names; other authors, such as Obreshkoff [Obr63, §7] [Obr03, §VII.7] and Marden [Mar66, Thm. 16,1], drop Schur's name. A proof of the theorem can be found in any of these references. We remark that

the occurrence of binomial coefficients, reminiscent of our Equation (2.5), is no coincidence: Szegő's Faltungssatz is a theorem on the zeros of a polar form, which he calls "Faltungsgleichung" [Sze22, p. 30].

The polynomial $C(X)$ in Theorem 2.42 is called the *composition* of $A(X)$ and $B(X)$. According to (2.19), $H(X)$ is the composition of $G(X)$ and

$$T(X) = \sum_{i=0}^{n-k} \frac{(i+k)!}{(i+k-r)!} \binom{n-k}{i} X^i. \tag{2.20}$$

This polynomial is closely related to the rth derivative of the kth Bernstein polynomial:

$$\begin{aligned} B_k^n(-X) &= \binom{n}{k}(-X)^k(1+X)^{n-k} \\ &= (-1)^k \binom{n}{k} \sum_{i=0}^{n-k} \binom{n-k}{i} X^{i+k} \\ \frac{\mathrm{d}^r}{\mathrm{d}X^r} B_k^n(-X) &= (-1)^k \binom{n}{k} \sum_{i=0}^{n-k} \frac{(i+k)!}{(i+k-r)!} \binom{n-k}{i} X^{i+k-r} \\ &= (-1)^k \binom{n}{k} X^{k-r}\, T(X). \end{aligned} \tag{2.21}$$

Lemma 2.43. *The roots of $T(X)$ are those of $\frac{\mathrm{d}^r}{\mathrm{d}X^r} B_k^n(-X)$, with the same multiplicities; except $x = 0$, which is not a root of $T(X)$. In particular, they are all real and contained in the interval $[-1, 0)$.*

Proof. The roots of $B_k^n(-X)$ are 0, with multiplicity k, and -1, with multiplicity $n-k$; in particular, they are all real and contained in the interval $[-1, 0]$. When we differentiate $B_k^n(-X)$ for r times, a simple inductive argument using Rolle's Theorem (similar to our proof of Theorem 2.2) shows that the roots remain real and contained in the interval $[-1, 0]$. The multiplicity of the root 0 drops from k to $k-r$. The claim now follows from (2.21). □

Theorem 2.44. *Let α be a k-fold root of the polynomial F, and let $s > 0$ be the minimum distance of α to any other root $\beta \neq \alpha$ of F. Let $1 \leq r \leq k$. Write $\vartheta_{k,r}^n$ for the smallest positive root of $\frac{\mathrm{d}^r}{\mathrm{d}X^r} B_k^n(X)$. If $\xi \neq \alpha$ is a root of $F^{(r)}$, then $|\xi - \alpha| \geq s \cdot \vartheta_{k,r}^n$.*

We will also write ϑ_k^n for $\vartheta_{k,k}^n$.

Proof. As before, we assume $\alpha = 0$, and we define G and H as in (2.18). By the preceding lemma, all roots of T are contained in the interval $[-1, -\vartheta_{k,r}^n]$. Let K be the closed disc whose boundary is the circumcircle of $[-1, -\vartheta_{k,r}^n]$. This disc K is a closed circular region containing all roots of T. Theorem 2.42 implies for the composition H of T and G that its root ξ has the form $\xi = -w\beta$ with $w \in K$ and $G(\beta) = 0$. One has $|w| \geq \vartheta_{k,r}^n$ and $|\beta| \geq s$, so that $|\xi| \geq s \cdot \vartheta_{k,r}^n$. □

The tightness of this bound is seen immediately from the example $B_k^n(X)$, which has a k-fold root at 0 and its only other root at 1.

For $r = 1$, it is well-known and trivial to verify that $\vartheta_{k,1}^n = k/n$; this is also mentioned by Dimitrov [Dim98]. For $r > 1$, we are not aware of any result in the literature, but have discovered the following ad-hoc estimate.

Proposition 2.45. *In Theorem 2.44, we have* $\vartheta^n_{k,r} \geq (1+k-r)/((k+1)(n-k))$.

Proof. Deleting constant factors in (2.21), we see that $\vartheta^n_{k,r}$ is the smallest positive root of

$$S(X) = \sum_{i=0}^{n-k} \frac{(i+k)!}{i!\,(i+k-r)!\,(n-k-i)!}\,(-X)^i$$

We group the terms of S into pairs; if the degree is even (and thus the number of terms is odd), the leading term is not paired:

$$\underbrace{\frac{n!}{(n-k)!\,(n-r)!}X^{n-k}}_{\text{if } n-k \text{ is even}} + \sum_{j=0}^{\lfloor\frac{n-k-1}{2}\rfloor} \underbrace{\left(-\frac{(k+2j+1)!}{(2j+1)!\,(k-r+2j+1)!\,(n-k-2j-1)!}\,X + \frac{(k+2j)!}{(2j)!\,(k-r+2j)!\,(n-k-2j)!}\right)}_{=:P_j} X^{2j}$$

We seek a bound $B > 0$ such that every underbraced subexpression above is positive when substituting $x \in (0, B)$ for X. If a separate leading term is present at all, it is positive anyway. We turn to the pairs P_j. Each P_j is a polynomial in X of degree 1. Setting $P_j(x_j) = 0$, cancelling the common factors of the factorials and solving for x_j, we see that the unique root of P_j is

$$x_j = \frac{2j+1+k-r}{k+2j+1} \cdot \frac{2j+1}{n-k+1-(2j+1)}.$$

The pair $P_j(x)$ is positive for $x < x_j$. To make all pairs positive, we choose $B = \min_j x_j$. To determine this minimum, let us minimize both factors of x_j, seen as functions of $u := 2j+1$. The first factor is of the form $(u+(k-r))/(u+k)$; its derivative is $r/(u+k)^2 \geq 0$, so it is nondecreasing in u. The second factor is of the form $u/(a-u)$ where $0 < u < a = n-k+1$; its derivative is $a/(a-u)^2 > 0$, so it is increasing in u. Thus, $B = x_0 = (1+k-r)/((k+1)(n-k))$. Since $S(x) > 0$ for $x \in (0, B)$, and $\vartheta^n_{k,r}$ is a positive root of S, the claim follows. □

The author has previously published his generalization (essentially, Theorem 2.44) of Dimitrov's result [Dim98] in the notice [Eig07]. However, the relation (2.21) between $T(X)$ and $\frac{\mathrm{d}^r}{\mathrm{d}X^r}B^n_k(-X)$ is new in this thesis and simplifies the argumentation; furthermore, Proposition 2.45 improves considerably on the naive bound in [Eig07].

We draw the following conclusion from the combined results of §2.3.3 and §2.3.4.

Proposition 2.46. *Let the real polynomial F of degree n have a k-fold root α in the bounded open interval I, and let s be the minimum distance from α to any other root $\beta \neq \alpha$ of F.*

(i) If $s \geq |I|\,/\vartheta^n_k$, then DescartesTest$(F, I) = k$.

(ii) If $s \geq |I| \cdot (k+1)(n-k)$, then the condition in (i) is satisfied.

Proof. Ad (ii). We have $(k+1)(n-k) \geq 1/\vartheta^n_k$ by Proposition 2.45.

Ad (i). Consider any root ξ of $F^{(k)}$. Theorem 2.44 implies $|\xi - \alpha| \geq s \cdot \vartheta^n_k \geq |I|$. Hence no such ξ is inside the circumcircle of I, and Proposition 2.40 yields the claim. □

Let us discuss the role of Proposition 2.40 in this proof. It allows us to deduce DescartesTest$(F, I) = k$ from $|\xi - \alpha| \geq |I|$. Can we attain a larger, i.e., less restrictive, upper bound on $|I|$ by replacing Proposition 2.40 with condition (i) or (ii) of Proposition 2.41?

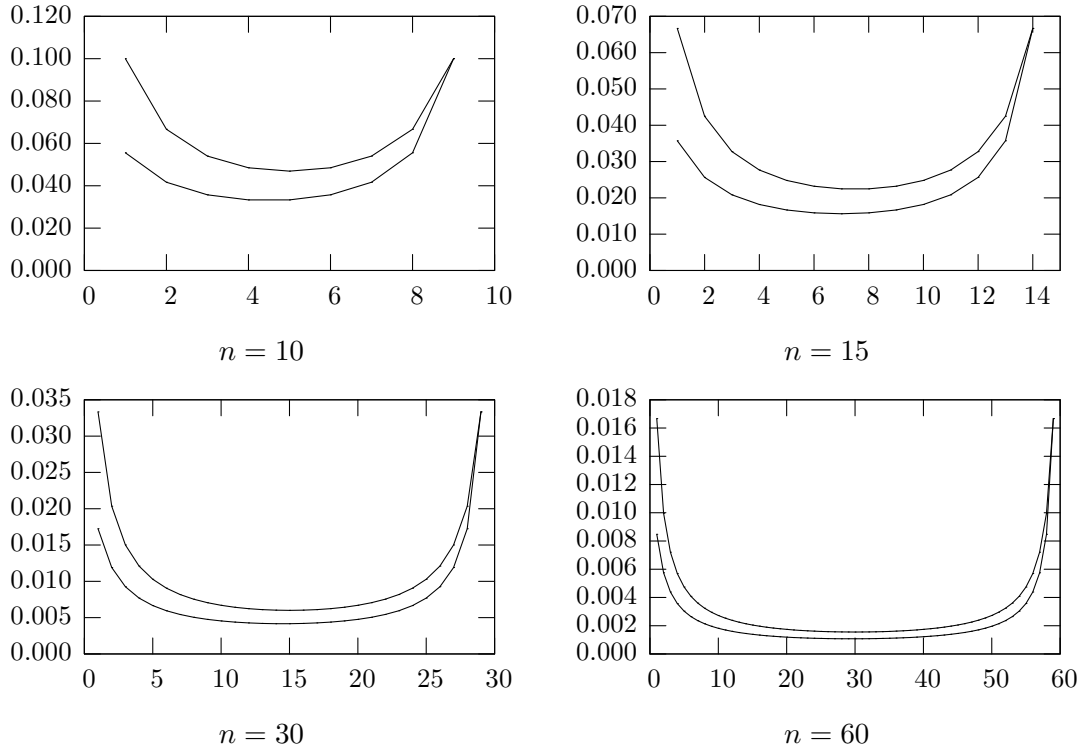


Figure 2.6: Empirical comparison of the value ϑ_k^n and its lower bound from Proposition 2.45 (on the vertical axis) as functions of $k = 1, \ldots, n-1$ (on the horizontal axis) for various choices of n. The graphs shown are interpolated linearly between successive values of k.

For condition (i), we demonstrated the impossibility with the counterexample (2.17) above. Condition (ii) would give us the liberty to ignore the root ξ_1 of $F^{(k)}$ nearest to α and instead argue that the distance to the second-nearest root ξ_2 is larger than the diameter $|I| \cdot 2/\sqrt{3}$ of $\mathrm{OC}_{\leq}^{\pm}(1, 1, n; I)$. However, the resulting upper bound $|\xi_2 - \alpha| \cdot \sqrt{3}/2$ on $|I|$ need not be larger than $|\xi_1 - \alpha|$ in general; it may even be smaller. Hence we regard Proposition 2.46(i) as the "right" consequence of Theorem 2.38.

Since we have no explicit expression for ϑ_k^n, we have to use part (ii) in applications of Proposition 2.46. Of course, that prompts the question: How good is the estimate $\vartheta_k^n \geq 1/((k+1)(n-k))$? We do not have a formal proof for its tightness, but a few computational experiments displayed in Figure 2.6 indicate that it reflects ϑ_k^n well (up to a small constant factor), including the bitonic behaviour that makes it smallest for $k \approx n/2$ and largest for the extreme values $k = 1$ and $k = n - 1$.

2.3.5 Comparison of the partial converses

Which of the two partial converses, as summarized in Propositions 2.36 and 2.46, should we use for the analysis of the Descartes method? While the two results are logically independent (as shown in §2.3.3), Obreshkoff's result is quantitatively superior.

Let us first consider the case that F has a simple root α in the bounded open interval I. Again, we write s for the minimum distance of α to any other root of F. Which upper bound on $|I|$ is sufficient for $\mathit{DescartesTest}(F, I) = 1$? Obreshkoff's approach gives the

bound $|I| \le s \cdot \sqrt{3}/2 \approx 0.886\, s$, see Proposition 2.36, whereas the approach of §2.3.3 yields the condition $|I| \le s/n$, because $\vartheta_1^n = 1/n$ in Proposition 2.44. Clearly, Obreshkoff's approach works better.

Let us now consider the case that α is a k-fold root of F, for any $1 \le k \le n$. Obreshkoff's approach yields a condition $|I| \le s/\Theta(k)$, whereas Proposition 2.46(ii) leads to $|I| \le s/\Theta(k(n-k))$. Again, Obreshkoff's approach gives the better result, except for highly degenerate situations where k is close to n.

Consequently, we will use Obreshkoff's partial converse to Descartes' Rule in the sequel. However, we point out that the inferior asymptotics of our partial converse would not damage the final complexity results, because they are dominated by other quantities, as we will see in §3.4.4. Lipka [Lip42] [RS02, Cor. 10.2.5] gave another partial converse leading to $|I| \le s/\Theta(\sqrt{n})$ irrespective of k; this would not change our final results either.

2.4 Bounds on the magnitude of roots

Bounds for the magnitude of roots of polynomials $A(X) = \sum_{i=0}^{n} a_i X^i$ are a classical mathematical topic; we refer to the book of Rahman and Schmeisser [RS02, §8] for a multitude of results and sources. We will not attempt a comparative study of root bounds here and point to van der Sluis[4] [vdS70] and Batra [Bat04] (also [Bat99, §1]) for that. The purpose of this section is to present one family of closely related bounds expressed in terms of $\sqrt[n-i]{|a_i/a_n|}$ for $0 \le i < n$ that is commonly used in conjunction with the Descartes method and is favoured by the above-mentioned comparisons. We include basic results from these sources regarding the quality of these bounds.

2.4.1 Bounds on all complex roots

In what follows, $A(X) = \sum_{i=0}^{n} a_i X^i$ denotes an arbitrary polynomial of degree $n > 0$. The statements of this section apply to the class of all non-constant polynomials A, no matter whether the coefficients $a_0, \ldots, a_n$ are taken to range over the real or complex numbers.

Definition 2.47. A *root bound functional* is a map R that takes a non-constant polynomial A to a real number $R(A) \ge 0$ such that $A(\alpha) = 0 \;\Rightarrow\; |\alpha| \le R(A)$ for all $\alpha \in \mathbb{C}$.

The optimal root bound functional is obviously the *complex root radius*

$$\mathrm{RR}(A) := \max\{|\alpha| \mid \alpha \in \mathbb{C},\, A(\alpha) = 0\} \in \mathbb{R}_{\ge 0}. \tag{2.22}$$

Let c denote an arbitrary positive real number. The following statements on a polynomial $A(X)$ and a complex number α are equivalent: (i) α is a root of $A(X)$; (ii) α is a root of $A(X)/\ell(A)$; (iii) $c\alpha$ is a root of $A(X/c)$. Hence RR has the following properties.

Definition 2.48. Consider a root bound functional R.

(i) If $R(A(X)) = R(A(X)/\ell(A))$ for all A, then R is *invariant under multiples*.

(ii) If $R(A(X/c)) = cR(A(X))$ for all A and all $c > 0$, then R is *homogeneous*.

Since RR has these properties, it is natural to demand that any root bound functional shall have them as well. If not, there are degrees of freedom left for optimization.

[4] Abraham van der Sluis (1928–2004), Dutch mathematician and numerical computing pioneer, professor at the University of Utrecht. An obituary by H. van der Vorst (in Dutch) appeared in *Nieuw Archief voor Wiskunde, 5th Ser.* **6** (2005), pp. 17–19, available from `http://www.math.leidenuniv.nl/~naw/`

The standard approach to root bounds is to consider polynomials in the power basis $(1, X, X^2, \ldots)$ and to use only the absolute values of their coefficients (but see [RS02, §8.2] for refined approaches). This is formalized as follows.

Definition 2.49. A root bound functional R is *absolute* if $R(A)$ for an indeterminate polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$ is a function of $|a_0|, \ldots, |a_n|$.

The following fundamental result is due to Cauchy and appeared 1821 in his *Cours d'Analyse* [Cau21][5]; another source is [Cau29, §5].

Theorem 2.50 (Cauchy (1821)). *Consider the map* $\mathrm{RB}_{\mathrm{Cp}}$ *that assigns to any polynomial* $A(X) = \sum_{i=0}^{n} a_i X^i$ *of degree* $n > 0$ *the unique positive real root of the Cauchy polynomial* $A^{\mathrm{C}}(X) := |a_n| X^n - \sum_{i=0}^{n-1} |a_i| X^i$, *or* 0, *if all coefficients other than* $|a_n|$ *are zero.* $\mathrm{RB}_{\mathrm{Cp}}$ *is a homogeneous absolute root bound functional that is invariant under multiples.*

Proof. We only need to discuss the non-trivial case $\{a_0, \ldots, a_{n-1}\} \neq \{0\}$. The existence of a unique positive real root ρ of $A^{\mathrm{C}}(X)$ follows from Descartes' Rule (Theorem 2.2). If $|x| > \rho$, then $A^{\mathrm{C}}(|x|) > 0$ and thus

$$|A(x)| \geq |a_n x^n| - |\sum_{i=0}^{n-1} a_i x^i| \geq |a_n|\,|x|^n - \sum_{i=0}^{n-1} |a_i|\,|x|^i = A^{\mathrm{C}}(|x|) > 0,$$

so $\mathrm{RB}_{\mathrm{Cp}}$ is indeed a root bound functional. By construction, it is absolute, homogeneous, and invariant under multiples. □

$\mathrm{RB}_{\mathrm{Cp}}$ is the optimal absolute root bound functional: The set of all polynomials of degree n with the given absolute values $|a_i|$ of coefficients contains $A^{\mathrm{C}}(X)$, so the positive number $\mathrm{RB}_{\mathrm{Cp}}(A)$ is itself one of the roots to be bounded, namely one that maximizes magnitude. Consequently, it is equivalent to think of an absolute root bound functional $R(A)$ as bounding the roots of a polynomial with the given absolute values of coefficients, or as bounding $\mathrm{RB}_{\mathrm{Cp}}(A)$.

Let us investigate how much we have lost by restricting attention to the absolute values of coefficients.

Proposition 2.51. *For any non-constant polynomial* $A(X)$, *whose degree we denote by* n, *it holds that* $\mathrm{RB}_{\mathrm{Cp}}(A) \leq 1/(\sqrt[n]{2} - 1) \cdot \mathrm{RR}(A) < n/\ln(2) \cdot \mathrm{RR}(A)$, *where* $n/\ln(2) \approx 1.44\,n$. *The first inequality is sharp: If* $A(X) = (X+1)^n$, *then* $\mathrm{RB}_{\mathrm{Cp}}(A) = 1/(\sqrt[n]{2} - 1)$.

Proof. On the first inequality, see [RS02, Thm. 8.1.4] or [vdS70, Thm. 3.8(e)]. For the second, truncate the series $\sqrt[n]{2} - 1 = \exp(\ln(2)/n) - 1 = \ln(2)/n + \frac{1}{2}(\ln(2)/n)^2 + \cdots$ after the first term. □

As $\mathrm{RB}_{\mathrm{Cp}}(A)$ is hard to compute exactly, let us now address the problem of obtaining simple upper bounds for it. Fujiwara[6] [Fuj16] gave a general form of such a bound, which is universal in a certain sense [vdS70, Thm. 2.2].

[5] The result appears in *Note III: Sur la résolution numérique des équations.* In the pagination of the *Œuvres*, the exact location is 392–393.

[6] Matsusaburô Fujiwara (1881–1946), professor of mathematics at Tohoku Imperial University, Sendai, Japan. (Tohoku is the north-east part of Honshu island.) An obituary note by T. Kubota appeared in *Tohoku Math. J. (2nd Ser.)* **1** (1949), pp. 1–2.

Proposition 2.52 (Fujiwara (1916)). *Consider a polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$ of degree $n > 0$. Let $\lambda_1, \ldots, \lambda_n$ be positive real numbers such that $\sum_{i=1}^{n} 1/\lambda_i \leq 1$. Then*

$$\mathrm{RR}(A) \leq \max\Big\{ \Big(\lambda_{n-i} \Big| \frac{a_i}{a_n} \Big| \Big)^{\frac{1}{n-i}} \;\Big|\; 0 \leq i < n \Big\}. \tag{2.23}$$

Proof. Write R for the right-hand side of (2.23). Consider an arbitrary $x > R$. We have $x^{n-i} > \lambda_{n-i} \, |a_i/a_n|$ and so $\lambda_{n-i}^{-1} > x^{i-n} \, |a_i/a_n|$ for all i. From $1 \geq \sum_{i=0}^{n-1} \lambda_{n-i}^{-1}$ we obtain

$$|a_n| \, x^n \geq \sum_{i=0}^{n-1} \lambda_{n-i}^{-1} \, |a_n| \, x^n > \sum_{i=0}^{n-1} |a_i| \, x^i,$$

hence $A^{\mathrm{C}}(x) > 0$ and $x > \mathrm{RB}_{\mathrm{Cp}}(A)$. It follows that $R \geq \mathrm{RB}_{\mathrm{Cp}}(A) \geq \mathrm{RR}(A)$. □

For any permissible choice of constants λ_i, the right-hand side of (2.23) is an absolute root bound functional. It is homogeneous and invariant under multiples, because already the subexpressions $\sqrt[n-i]{|a_i/a_n|}$ have these properties. (In fact, these properties make it natural to seek a bound in terms of $\sqrt[n-i]{|a_i/a_n|}$.)

Fujiwara discusses various choices of λ_i, including $\lambda_i = 2^i$ for $i = 1, \ldots, n$, which makes $\sum_{i=1}^{n} 1/\lambda_i = \sum_{i=1}^{n} 1/2^i = 1 - 2^{-n} < 1$. This yields the **folklore complex root bound**[7]

$$\mathrm{RR}(A) \leq \mathrm{RB}_{\mathrm{folk}}(A) := 2 \max\Big\{ \Big| \frac{a_{n-1}}{a_n} \Big|, \Big| \frac{a_{n-2}}{a_n} \Big|^{\frac{1}{2}}, \ldots, \Big| \frac{a_1}{a_n} \Big|^{\frac{1}{n-1}}, \Big| \frac{a_0}{a_n} \Big|^{\frac{1}{n}} \Big\}. \tag{2.24}$$

But there is room for improvement, since the constraint $\sum_{i=1}^{n} 1/\lambda_i \leq 1$ is not tight. We mention three possibilities.

The first (see, e.g., [RS02, Cor. 8.1.8]) consists in retaining the ansatz $\lambda_i = \sigma_n^i$; the optimal choice of σ_n is the unique positive root of $X^n - \sum_{i=0}^{n-1} X^i$. It is not hard to see that the sequence $(\sigma_n)_n$ is increasing and contained in the interval (3/2, 2) for $n \geq 2$; from $\sigma_n^n = (1 - \sigma_n^n)/(1 - \sigma_n)$ we get $2 - \sigma_n = \sigma_n^{-n}$ and see that $(\sigma_n)_n$ converges rather fast to 2. The resulting bound

$$\mathrm{RR}(A) \leq \sigma_n \max\Big\{ \Big| \frac{a_i}{a_n} \Big|^{\frac{1}{n-i}} \;\Big|\; 0 \leq i < n \Big\}, \qquad \text{with } 1 \leq \sigma_n < 2, \quad \sigma_n^n = \sum_{i=0}^{n-1} \sigma_n^i, \tag{2.25}$$

is interesting in so far as it explicitly shows the gap between $\mathrm{RR}(A)$ and the right-hand side in (2.24); we see that (2.24) is not sharp for any fixed degree n. Algorithmically, however, the non-constant, irrational quantity σ_n is inconvenient to handle.

The second improvement, favoured by van der Sluis [vdS70], keeps things simple for computers with binary numbers and redefines $\lambda_n = 2^{n-1}$ while retaining $\lambda_i = 2^i$ for $i < n$ to make the constraint $\sum_{i=1}^{n} 1/\lambda_i \leq 1$ satisfied with equality. Van der Sluis [vdS70, p. 252] refers to Marden [Mar66, §30 ex. 5] as a textbook source for this particular choice of parameters, but does not attach a name to the resulting bound. Fujiwara [Fuj16] has not spelled out this improvement, but it is so close to his presentation that it can hardly be counted as a discovery in its own right. Therefore, we follow Batra [Bat99, §1.1] and

[7] Knuth states (2.24) as Exercise 20 in [Knu69, §4.6.2] without an attribution (consistent with other authors). The exercise has been replaced in later editions of [Knu69].

name this bound after Fujiwara. However, to indicate the specific "base 2" choice of the parameters λ_i, we call it the **dyadic Fujiwara complex root bound**:

$$\mathrm{RR}(A) \leq \mathrm{RB}_{\mathrm{dF}}(A) := 2\max\Big\{ \left|\frac{a_{n-1}}{a_n}\right|, \left|\frac{a_{n-2}}{a_n}\right|^{\frac{1}{2}}, \ldots, \left|\frac{a_1}{a_n}\right|^{\frac{1}{n-1}}, \left|\frac{a_0}{2a_n}\right|^{\frac{1}{n}} \Big\}. \tag{2.26}$$

We summarize the findings of [vdS70] on $\mathrm{RB}_{\mathrm{dF}}$ as follows.

Proposition 2.53 (van der Sluis (1970)). *For $A(X) = \sum_{i=0}^{n} a_i X^i$ as above, it holds that*

(i) $\mathrm{RB}_{\mathrm{dF}}$ *is sharp:* $A(X) = X^n - X^{n-1} - \cdots - X - 2 \;\Rightarrow\; \mathrm{RR}(A) = 2 = \mathrm{RB}_{\mathrm{dF}}(A)$*;*

(ii) $\mathrm{RB}_{\mathrm{dF}}(A) \leq 2 \cdot \mathrm{RB}_{\mathrm{Cp}}(A)$*, with equality for* $A(X) = X^n - X^{n-1}$*;*

(iii) $\mathrm{RB}_{\mathrm{dF}}(A) \leq 2n \cdot \mathrm{RR}(A)$*, with equality for* $A(X) = (X+1)^n$*.*

The reader is invited to observe that the statements (ii) and (iii) together with their proofs apply as well to the folklore bound (2.24), and – with σ_n in place of the factor 2 – also to the bound (2.25).

Proof. Ad (i). $A(X) = (X-2)(X^n-1)/(X-1)$, so $\mathrm{RR}(A) = 2 = \mathrm{RB}_{\mathrm{dF}}(A)$.

Ad (ii). If $\mathrm{RB}_{\mathrm{Cp}}(A) = 0$, then $\mathrm{RB}_{\mathrm{dF}}(A) = 0$. Otherwise, we can use homogeneity to scale the indeterminate without changing the ratio $\mathrm{RB}_{\mathrm{dF}}(A)/\mathrm{RB}_{\mathrm{Cp}}(A)$ to attain $\mathrm{RB}_{\mathrm{Cp}}(A) = 1$. From $A^{\mathrm{C}}(1) = 0$ it follows that $1 = \sum_{i=0}^{n-1} |a_i/a_n|$, hence $|a_i/a_n| \leq 1$ for all $i < n$, and thus $\mathrm{RB}_{\mathrm{dF}}(A) \leq 2$.

Ad (iii). Again, we can restrict to the case $\mathrm{RR}(A) = 1$ by homogeneity. By invariance under multiples, it is not a restriction either to assume $a_n = 1$. Consider the set of all polynomials $P(X) = \prod_{j=1}^{n}(X - \vartheta_j) = \sum_{i=0}^{n} p_i X^i$ with roots $|\vartheta_j| \leq 1$. The coefficients satisfy $|p_{n-i}| = |\sum_{j_1<\cdots<j_i} \vartheta_{j_1} \cdots \vartheta_{j_i}| \leq \binom{n}{i}$, and the upper bound is attained for $\vartheta_1 = \cdots = \vartheta_n = -1$. As $\mathrm{RB}_{\mathrm{dF}}(P)$ is a non-decreasing function of each p_i, this choice $P(X) = (X+1)^n$ also maximizes $\mathrm{RB}_{\mathrm{dF}}(P)$. The value of the maximum is $\mathrm{RB}_{\mathrm{dF}}(P) = 2n$, since $\binom{n}{i}^{1/i} \leq n$, with equality for $i = 1$. It follows that $\mathrm{RB}_{\mathrm{dF}}(A) \leq 2n$, as desired. □

Independently of Fujiwara, Westerfield [Wes31] has described his own parametric approach to bounds on $\mathrm{RB}_{\mathrm{Cp}}$ and derives the following result [op. cit., Eq. (e)]. Instead of taking twice the largest number in (2.24), it suffices to take the sum of the largest and the second-largest, or formally:

$$\mathrm{RR}(A) \leq \mathrm{RB}_{\mathrm{Lgr}}(A) := \max\Big\{ \left|\frac{a_i}{a_n}\right|^{\frac{1}{n-i}} + \left|\frac{a_j}{a_n}\right|^{\frac{1}{n-j}} \;\Big|\; 0 \leq i < j < n \Big\}. \tag{2.27}$$

We comment on the history of this bound in §2.4.2 near Equation (2.31). At this point, let us discuss the quality of this bound in the style of van der Sluis.

Proposition 2.54. *For $A(X) = \sum_{i=0}^{n} a_i X^i$ as above, it holds that*

(i) $\mathrm{RB}_{\mathrm{Lgr}}$ *is sharp:* $A(X) = X^n - 1 \;\Rightarrow\; \mathrm{RR}(A) = 1 = \mathrm{RB}_{\mathrm{Lgr}}(A)$*;*

(ii) $\mathrm{RB}_{\mathrm{Lgr}}(A) < 2 \cdot \mathrm{RB}_{\mathrm{Cp}}(A)$*, and no constant factor smaller than 2 is possible;*

(iii) $\mathrm{RB}_{\mathrm{Lgr}}(A) \leq (n + \sqrt{n(n-1)/2}) \cdot \mathrm{RR}(A)$ *with equality for* $A(X) = (X+1)^n$*, where* $n + \sqrt{n(n-1)/2} < (1 + 1/\sqrt{2}) \cdot n \approx 1.71\, n$*.*

Proof. Ad (i). This is obvious.

Ad (ii). As in the proof of Proposition 2.53(ii), we may assume $\mathrm{RB}_{\mathrm{Cp}}(A) = 1$. From $A^{\mathrm{C}}(1) = 0$ it follows that $1 = \sum_{i=0}^{n-1} |a_i/a_n|$. If only one summand is non-zero, then $\mathrm{RB}_{\mathrm{Lgr}}(A) = 1 < 2$. If at least two summands are non-zero, then $|a_i/a_n| < 1$ for all $i < n$, and thus $\mathrm{RB}_{\mathrm{Lgr}}(A) < 2$.

For asymptotic tightness, consider the polynomials $A_n(X) = X^n - X - 1$ for $n \geq 2$. Clearly, $\mathrm{RB}_{\mathrm{Lgr}}(A_n) = 2$. On the other hand, $A_n(1 + \frac{2}{n}) > (1 + n\frac{2}{n}) - (1 + \frac{2}{n}) - 1 \geq 0$, so $\mathrm{RB}_{\mathrm{Cp}}(A_n) < 1 + \frac{2}{n}$. Hence $2 > \mathrm{RB}_{\mathrm{Lgr}}(A_n)/\mathrm{RB}_{\mathrm{Cp}}(A_n) > 2n/(n+2) \to 2$ for $n \to \infty$.

Ad (iii). By the same arguments as in the proof of Proposition 2.53(iii), we can restrict to the case $\mathrm{RR}(A) = 1$ and obtain that $\mathrm{RB}_{\mathrm{Lgr}}(A)$ is maximized for $A(X) = (X+1)^n$. Its coefficients are $a_{n-i} = \binom{n}{i}$, $0 \leq i \leq n$, and the two largest elements of $(\binom{n}{i}^{1/i})_{i=1}^n$ are those with indices $i = 1$ and $i = 2$. The claim follows. □

2.4.2 Bounds on positive real roots

Let us now restrict to real polynomials $A(X)$ and seek bounds for the *positive root radius*

$$\mathrm{RR}^+(A) := \max\left(\{\alpha \mid \alpha \in \mathbb{R}_{>0},\, A(\alpha) = 0\} \cup \{0\}\right) \in \mathbb{R}_{\geq 0}. \tag{2.28}$$

We take the same approach as Kioustelidis [Kio86]. We begin by constructing an analogue to the Cauchy polynomial from Theorem 2.50.

Theorem 2.55. *Let $A(X) = \sum_{i=0}^n a_i X^i$ be a polynomial of degree $n > 0$ with real coefficients. Let $I = \{i \in \{0, \ldots, n-1\} \mid a_i/a_n < 0\}$. Let $\mathrm{RB}^+_{\mathrm{Cp}}(A)$ denote the unique positive real root of $A^{\mathrm{C}+}(X) := |a_n|\, X^n - \sum_{i \in I} |a_i|\, X^i$, or 0, if all coefficients other than $|a_n|$ are zero. It holds that $\mathrm{RR}^+(A) \leq \mathrm{RB}^+_{\mathrm{Cp}}(A)$.*

Proof. We assume w.l.o.g. that $a_n > 0$. The polynomial $A^{\mathrm{C}+}(X)$ consists of the leading term and the negative terms of $A(X)$. For all $x > 0$, it holds that $A(x) - A^{\mathrm{C}+}(x) \geq 0$, as all remaining terms are positive. For all $x > \mathrm{RB}^+_{\mathrm{Cp}}(A)$, it holds that $A^{\mathrm{C}+}(x) > 0$. Taken together, this implies $A(x) > 0$ for all $x > \mathrm{RB}^+_{\mathrm{Cp}}(A)$, as desired. □

We note that Rahman and Schmeisser [RS02, Thm. 8.2.4] found an elegant generalization to roots on arbitrary rays $re^{i\varphi}$ ($r > 0$ varying, $\varphi \in \mathbb{R}$ fixed) in the complex plane.

What can we say about the quality of this bound? As in the case of Theorem 2.50, the bound $\mathrm{RB}^+_{\mathrm{Cp}}(A)$ applies to its own defining polynomial, so it is sharp by construction. Also, it is easy to see that $\mathrm{RB}^+_{\mathrm{Cp}}(A) \leq \mathrm{RB}_{\mathrm{Cp}}(A)$, so the maximum overestimation factor from Proposition 2.51 in relation to all complex roots carries over immediately. However, no continuous functional R that bounds RR^+ has a maximum overestimation factor in relation to the positive real roots alone: Think of a real polynomial $A(X)$ with one double positive real root α that is much larger in magnitude than the next largest positive real root β; i.e., $\alpha/\beta \gg 1$. Obviously, there exist examples for arbitrarily large values of α/β. Of course, we have $R(A) \geq \alpha$, so $q := R(A)/\beta$ can be made arbitrarily large as well. Now let us study small perturbations of A. Every neighbourhood of A in the space of fixed-degree real polynomials contains elements $\tilde{A}$ for which α splits into a pair of imaginary roots while the other roots remain fixed. The coefficients of $\tilde{A}$ and thus $R(\tilde{A})$ have changed only by an arbitrarily small amount, so the new overestimation factor $R(\tilde{A})/\beta$ is close to q, that is, arbitrarily large.

In the preceding section, we studied simple bounds on $\mathrm{RB}_{\mathrm{Cp}}$. Let us now transfer these results to $\mathrm{RB}^+_{\mathrm{Cp}}$, cf. [Kio86]. The key observation is this: The polynomial $A^{\mathrm{C}+}(X)$ from Theorem 2.55 applied to $A(X)$ is the Cauchy polynomial $A^{\mathrm{C}}_*(X)$ from Theorem 2.50 applied to a modified polynomial $A_*(X)$. The polynomial $A_*(X)$ is obtained from $A(X)$ by setting those coefficients a_i, $i < n$, to zero that agree with a_n in sign. Thus, we

can transfer the simpler bounds from §2.4.1 by propagating this annihilation of certain coefficients into their defining expressions. This is conveniently done with the symbol $|x|_- := |\min\{0, x\}|$ denoting the magnitude of a negative number x and zero otherwise. We observe that the resulting bounds are equal to or less than their counterparts for all complex roots, because we effectively form the maximum of a subset.

We obtain the **folklore positive root bound**

$$\mathrm{RR}^+(A) \le \mathrm{RB}^+_{\mathrm{folk}}(A) := 2\max\Big\{ \left|\frac{a_{n-1}}{a_n}\right|_-, \left|\frac{a_{n-2}}{a_n}\right|_-^{\frac{1}{2}}, \ldots, \left|\frac{a_1}{a_n}\right|_-^{\frac{1}{n-1}}, \left|\frac{a_0}{a_n}\right|_-^{\frac{1}{n}} \Big\}, \quad (2.29)$$

the **dyadic Fujiwara positive root bound**

$$\mathrm{RR}^+(A) \le \mathrm{RB}^+_{\mathrm{dF}}(A) := 2\max\Big\{ \left|\frac{a_{n-1}}{a_n}\right|_-, \left|\frac{a_{n-2}}{a_n}\right|_-^{\frac{1}{2}}, \ldots, \left|\frac{a_1}{a_n}\right|_-^{\frac{1}{n-1}}, \left|\frac{a_0}{2a_n}\right|_-^{\frac{1}{n}} \Big\}, \quad (2.30)$$

and **Lagrange's positive root bound**

$$\mathrm{RR}^+(A) \le \mathrm{RB}^+_{\mathrm{Lgr}}(A) := \max\Big\{ \left|\frac{a_i}{a_n}\right|_-^{\frac{1}{n-i}} + \left|\frac{a_j}{a_n}\right|_-^{\frac{1}{n-j}} \mid 0 \le i < j < n \Big\}. \quad (2.31)$$

Lagrange stated the bound (2.31) in [Lag69, 12.] (also in [Lag08, p. 32]); this source from 1769 is older than any other literature reference we know for the bounds presented here. Lagrange specifically restricted his considerations to positive real roots; however, using Theorem 2.50 and the Cauchy polynomial, it immediately entails the bound (2.27) on all complex roots. Therefore, we have to regard (2.27) as already known at the time when Westerfield [Wes31] derived it, and we use the symbol $\mathrm{RB}_{\mathrm{Lgr}}$ for it. The folklore bounds (2.24) and (2.29) follow from Lagrange's result (2.31) but are strictly weaker. Therefore, we refrain from attaching Lagrange's name to them.

With exactly the same arguments as for Propositions 2.53 and 2.54, we attain the following quality guarantees.

Proposition 2.56. *Let $A(X) = \sum_{i=0}^{n} a_i X^i$ be an arbitrary real polynomial of degree $n > 0$. It holds that*

(i) $\mathrm{RB}^+_{\mathrm{dF}}$ *is sharp:* $A(X) = X^n - X^{n-1} - \cdots - X - 2 \;\Rightarrow\; \mathrm{RR}(A) = 2 = \mathrm{RB}_{\mathrm{dF}}(A)$;

(ii) $\mathrm{RB}^+_{\mathrm{dF}}(A) \le 2 \cdot \mathrm{RB}^+_{\mathrm{Cp}}(A)$, *with equality for* $A(X) = X^n - X^{n-1}$.

Item (ii) also holds for $\mathrm{RB}^+_{\mathrm{folk}}$.

Proposition 2.57. *For $A(X) = \sum_{i=0}^{n} a_i X^i$ as above, it holds that*

(i) $\mathrm{RB}^+_{\mathrm{Lgr}}$ *is sharp:* $A(X) = X^n - 1 \;\Rightarrow\; \mathrm{RR}(A) = 1 = \mathrm{RB}^+_{\mathrm{Lgr}}(A)$;

(ii) $\mathrm{RB}^+_{\mathrm{Lgr}}(A) < 2 \cdot \mathrm{RB}^+_{\mathrm{Cp}}(A)$, *and no constant factor smaller than 2 is possible.*

As discussed above, there no longer is a maximum overestimation factor.

Chapter 3

The Descartes Method for Real Root Isolation

We present the general form of the Descartes method and give a new and almost tight bound on its subdivision tree (§3.1). We review the Descartes method for exact integer coefficients (§3.2); the tree bound lets us derive the best known bit complexity statements quickly. These bounds on tree size and bit complexity originate from ideas by Vikram Sharma and Chee Yap that were worked out and published jointly with the author of this thesis in [ESY06], cf. [Sha07a, §2]. We present a revised derivation of the tree bound.

The main contribution of this chapter is an extension of the Descartes method to polynomials with bitstream coefficients (§3.3) based on joint work with Kurt Mehlhorn et al. [EKK+05]. We give an improved version of the algorithm and a refined complexity analysis based on our tree bound. We discuss a geometric application of the bitstream Descartes algorithm (§3.4) from joint work with Michael Kerber and Nicola Wolpert [EKW07].

3.1 The Descartes method and its subdivision tree

3.1.1 General form of the Descartes method

Throughout this chapter, we consider the following task: Given a non-constant real polynomial $A_{\text{in}}(X)$ and two real numbers $c_0 < d_0$, compute isolating intervals $I_1, \ldots, I_r$ for those real roots of $A_{\text{in}}(X)$ that lie within the initial interval $I_0 = (c_0, d_0)$. We say that the intervals $I_1, \ldots, I_r$ are *isolating* for some subset S of the real roots of A_{in} if (i) the intervals are pairwise disjoint, (ii) their union contains the roots S, and (iii) each interval contains one root from S and no other root of A_{in}. Root isolation differs from root approximation in that isolating intervals need not be small, and – depending on the application – sometimes *should not* be small but rather have endpoints with short representations.

The *Descartes method* solves the real root isolation problem under the condition that all roots to be isolated are simple. Its approach is *recursive subdivision*: A sequence $\mathcal{P}_0, \mathcal{P}_1, \ldots$ of partitions of I_0 is constructed, starting from the trivial partition $\mathcal{P}_0 = \{I_0\}$. To construct $\mathcal{P}_{t+1}$ from $\mathcal{P}_t$, one checks whether $\mathcal{P}_t$ contains an interval $I = (c, d)$ with $\mathit{DescartesTest}(A_{\text{in}}, I) \geq 2$. If such an interval I exists, a point $m \in (c, d)$ is chosen and the refined partition $\mathcal{P}_{t+1} = (\mathcal{P}_t \setminus \{I\}) \cup \{(c, m), \{m\}, (m, d)\}$ is constructed by subdividing I at m.[1] If no such I exists, the sequence $\mathcal{P}_0, \mathcal{P}_1, \ldots$ terminates at $t_{\max} := t$. If the sequence terminates, the final partition $\mathcal{P}_{t_{\max}}$ consists of singleton intervals $[m, m]$ and open intervals

[1] Virtually all previous work on the Descartes method uses bisection, i.e., subdivision at the interval midpoint $m = (c+d)/2$. However, the randomization technique to be introduced in §3.3 will need some freedom in choosing the subdivision point m.

(c, d) containing at most one root each. Those elements of $\mathcal{P}_{t_{\max}}$ that contain a root (as evident from $A_{\text{in}}(m) = 0$ or $\mathit{DescartesTest}(A_{\text{in}}, (c, d)) = 1$, resp.) are reported as isolating intervals.

We regard recursive subdivision as constructing an ordered binary tree, the *subdivision tree* $\mathcal{T}$. Its node set is the set of all open intervals in $\bigcup_t \mathcal{P}_t$, which is finite iff the sequence $(\mathcal{P}_t)_t$ terminates. To each open interval (c, d) that has been subdivided, we associate its subintervals (c, m) as left child and (m, d) as right child. The root of $\mathcal{T}$ is the initial interval I_0. The leaves of $\mathcal{T}$ are the intervals I that have $\mathit{DescartesTest}(A_{\text{in}}, I) \leq 1$. The singleton sets $\{m\}$ in $\bigcup_t \mathcal{P}_t$ correspond bijectively to the internal nodes of $\mathcal{T}$: each such m was chosen as subdivision point for one subdivided interval. If a final partition $\mathcal{P}_{t_{\max}}$ is reached, we thus have a bijective correspondence between its elements and the nodes of $\mathcal{T}$.

Let us now cast this approach into pseudocode. Our procedure *Descartes* below maintains a sequence P and a set Q. Just before the main loop begins for the $(t + 1)$st time, the contents of P and Q reflect the current partition $\mathcal{P}_t$ in the following fashion: The sequence P consists of the singletons of $\mathcal{P}_t$ that contain a root and the open intervals of $\mathcal{P}_t$ that have a positive Descartes test. The elements of P are sorted in their natural order on the real line. The set Q records the open intervals of $\mathcal{P}_t$ that have a Descartes test larger than one and thus require further subdivision.

For the purposes of this piece of pseudocode, assignment of polynomials is understood up to multiplication by non-zero constants; what these constants are is not specified at this point. We write $A \sim B$ for two polynomials that are equal up to multiplication by a non-zero constant and recall that the Descartes test is invariant under multiplication by non-zero constants.

---

procedure *Descartes*$(A_{\text{in}}, (c_0, d_0))$
  $P \leftarrow ()$; $Q \leftarrow \{\}$;
  $A_0(X) \leftarrow A_{\text{in}}((d_0 - c_0)X + c_0)$;
  $v_0 \leftarrow \mathit{DescartesTest}(A_0, (1, 0))$; // i.e., $v_0 = \mathit{DescartesTest}(A_{\text{in}}, (c_0, d_0))$;
  if $v_0 \geq 1$ **then** $P \leftarrow ((c_0, d_0))$; **fi**;
  if $v_0 \geq 2$ **then** $Q \leftarrow \{((c_0, d_0), A_0)\}$; **fi**;
  while $Q \neq \{\}$ **do**
    // **Invariant:** $Q = \{((c, d), A) \mid$
    // $(c, d) \in P,\ \mathit{DescartesTest}(A_{\text{in}}, (c, d)) \geq 2,\ A(X) \sim A_{\text{in}}((d - c)X + c))\}$;
    choose an element $((c, d), A) \in Q$;
    choose $\alpha \in [\frac{1}{4}, \frac{3}{4}]$; $m \leftarrow (1 - \alpha)c + \alpha d$;
    $I_L \leftarrow (c, m)$; $I_M \leftarrow [m, m]$; $I_R \leftarrow (m, d)$;
    $A_L(X) \leftarrow A(\alpha X)$; $A_R(X) \leftarrow A((1 - \alpha)X + \alpha)$;
    $v_L \leftarrow \mathit{DescartesTest}(A_L, (1, 0))$; // i.e., $v_L = \mathit{DescartesTest}(A_{\text{in}}, I_L)$
    $v_M \leftarrow \max\{k \mid X^k \text{ divides } A_R\}$; // ≥ 1 iff $A_{\text{in}}(m) = 0$
    $v_R \leftarrow \mathit{DescartesTest}(A_R, (1, 0))$; // i.e., $v_R = \mathit{DescartesTest}(A_{\text{in}}, I_R)$
    in P, replace entry (c, d) by subsequence $(I_i \mid i \in (L, M, R), v_i \geq 1)$;
    in Q, replace element $((c, d), A)$ by elements $\{(I_i, A_i) \mid i \in \{L, R\}, v_i \geq 2\}$;
  od;
  report sequence P of isolating intervals;
end procedure;

For each subinterval with Descartes test larger than one, the set Q contains a pair $((c,d),A)$, in which $A(X) \sim A_{\text{in}}((d-c)X+c)$ is a transformed version of the input polynomial A_{in}. It is easy to check by substitution that $A_L(X) \sim A_{\text{in}}((m-c)X+c)$ and $A_R(X) \sim A_{\text{in}}((d-m)X+m)$. This, together with the derivation of the Descartes test in §2.2, makes clear that $v_L = \mathit{DescartesTest}(A_{\text{in}},(c,m))$ and $v_R = \mathit{DescartesTest}(A_{\text{in}},(m,d))$. Thus, the invariant claimed in line 8 does indeed hold.

What about termination? We observe that subdivision of an interval I creates subintervals with lengths at most $3/4 \cdot |I|$. Due to the condition that all roots of $A_{\text{in}}(X)$ in I_0 have multiplicity $k = 1$, Proposition 2.36 (on page 33) or Proposition 2.46 (on page 38) guarantee that subdivision stops as soon as the interval length is sufficiently small. We will see a more thorough version of this argument in §3.1.5, where we derive a bound on the size of the subdivision tree (Theorem 3.19).

3.1.2 Remark on sources and names

The Descartes method goes back to an algorithm of Collins and Akritas [CA76], which we will discuss in §3.2.4. They called it the "Modified Uspensky Algorithm"; the reference to Uspensky rather than Vincent was later criticized heavily by Akritas [Akr86]. Johnson [Joh91] [Joh98] has used the name "coefficient sign variation method". Current work of Collins, Johnson and Krandick speaks of "the Descartes method" (e.g., [CJK02] [KM06] [JKL$^+$06]), carefully avoiding the possessive "Descartes' method" – after all, Descartes himself did not envision this algorithm.[2] We follow their terminology, as it is widely (although not universally) established in the field.

The Descartes method has to be distinguished from the Continued Fractions method for real root isolation, which is also based on Descartes' Rule. Its basic form, Vincent's method [Vin36] [AG98], is the unmodified "Uspensky's algorithm" of [CA76]. Akritas has pursued this approach (not the "modified" one); we refer to [AS05] [ASV07] for the latest results of Akritas et al. and references to older work. A complexity analysis that copes with the the lack of a maximum overestimation factor for positive root bounds (see §2.4.2) has been undertaken by Sharma [Sha07a, §3] [Sha07b].

Both the Descartes method and the Continued Fractions method can be seen as methods for root isolation by recursive subdivision with the Descartes test as termination criterion. In this respect, they share a lot of the underlying mathematics and auxiliary algorithms. However, when we abstract from the use of this specific test for roots and consider the overall search strategy, efficient forms of the Continued Fractions method are markedly different from the Descartes method, because they critically depend on a further ingredient (positive root bound) to control subdivision, whereas the Descartes method follows the same search strategy as, say, Sturm's method, namely unguided subdivision starting from a bounded interval. Therefore, the relation of the Descartes method and the Continued Fractions method should not be overplayed.

[2] The author conjectures that René Descartes (1596–1650) would indeed assert that he has developed a method that should bear his name – not for the humble problem of isolating roots, but for seeking the truth through scientific reasoning. This is the subject of his influential *Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences* (1637). Ironically, the *Discours de la méthode* is indeed the link between René Descartes and "the Descartes method", as its appendix *La Géométrie* contains the remark that gave rise to Descartes' Rule of Signs, see [Kra95, §3.3] [KM06].

3.1.3 Details on the sets of subintervals

Various aspects of the sequence P and the set Q in procedure *Descartes* call for further explanations.

I. Let us begin by elucidating the basic roles of P and Q. In a straightforward implementation of procedure *Descartes*, both P and Q can be represented by linked lists, where the elements of Q are augmented with pointers to their counterparts in P, such that locating them in line 17 is easy.

All significant operations of procedure *Descartes* are controlled by the contents of Q and the manner of choosing elements from it in line 10. The sequence P, by contrast, is merely a repository for the resulting isolating intervals; the intervals in P that are not yet isolating serve as placeholders for isolating intervals arising from them in later iterations. The point here is that reporting the isolating intervals in sorted order is easy, even if the procedure may not discover them in that order. Also, a sorted sequence of isolating intervals often is the convenient form for the result of a root isolation procedure.

If sorted order of isolating intervals is not an issue, or if it is achieved by other means, an implementation can replace line 17 by

if $v_L = 1$ **then** report I_L; **fi**;
if $v_M \geq 1$ **then** report I_M; **fi**;
if $v_R = 1$ **then** report I_R; **fi**;

and thus effectively eliminate P.

II. Let us now discuss our way of recording intervals in Q. Following [CA76], we insist that Q shall not simply store intervals (c, d) but also transformed polynomials $A(X) \sim A_{\text{in}}((d-c)X + c)$. This, together with the transformations in line 13, allows us to perform Descartes tests only with respect to the fixed interval $(0, 1)$. The necessary transformation of A_{in} for testing w.r.t. (c, d) is performed incrementally in relatively cheap steps. When we discuss concrete implementations, it will become clear that this is more efficient than transforming A_{in} afresh for each Descartes test.

Depending on the application, it may be important to deliver the result of root isolation in a way that retains the internal state of the Descartes method and allows further subdivision. In such a situation, it may be useful to implement *Descartes* with a "fat" form of P that stores pairs $((c, d), A)$ and a "thin" form of Q that stores only pointers into P.

III. At the beginning of procedure *Descartes*, we have $v_0 \leq n := \deg(A_{\text{in}})$. Suppose $v_0 \geq 2$, so that the main loop is executed. Consider the subdivision of an interval (c, d) with $v = \mathit{DescartesTest}(A_{\text{in}}, (c, d))$. The variation-diminishing property (Proposition 2.26 on page 26) implies $v \geq v_L + v_M + v_R$. In other words, we can think of each element in P as carrying a positive, integral charge (its v value). Subdivision distributes charge onto subintervals; some charge may be lost. This has two important consequences.

At any time, the sum of all charges is at most n. Consequently, there are at most n entries in P (as they all have charge at least 1) and at most $n/2$ entries in Q (as they all have charge at least 2) at any stage of the algorithm.

Let us call the total charge of Q minus the cardinality of Q the *excess charge* of Q. This is a non-negative integer, which is at most $n - 1$ initially, can never grow, and reaches zero when and only when the algorithm terminates. If subdivision leads to $v_L, v_R \geq 1$,

the excess charge of Q drops by at least 1. If subdivision at m hits a root of $A_{\text{in}}(X)$, the excess charge drops by at least $v_M \geq 1$. If subdivision verifies $v > v_L + v_M + v_R$, the excess charge drops at least by the difference, which is no less than 1. Thus, with at most $n-1$ exceptions, the outcome of subdivision is $(v_L, v_M, v_R) \in \{(v, 0, 0), (0, 0, v)\}$. The total number of subdivisions can be much larger than $n-1$, as we will see in §3.1.5 and §3.2.2. Thus we expect the most frequent result of subdividing an interval recorded in Q to be simply one smaller interval in P and Q, with no change in the number of intervals or the Descartes test values. The implementation of the main loop should therefore be optimized to handle this easy case fast (e.g., avoid freeing and allocating list nodes when replacing (c, d) by a single subinterval in P).

IV. An important reason for describing the set Q as separate from the sequence P is the resulting freedom for choosing an entry from Q in line 10. Different policies for choosing from Q induce different traversal orders of the subdivision tree $\mathcal{T}$.

- Suppose Q is maintained as a sorted sequence, the first element of Q is chosen in line 10, and it is replaced in line 18 by new elements (if any) inserted at the beginning of Q in their natural order on the real line. Then Q is essentially a stack, and $\mathcal{T}$ is traversed depth first from left to right. Consequently, isolating open intervals are discovered in their natural order on the real line. (This does not hold for singleton intervals; they are discovered before further subdivision of the left subinterval.)

  This is the same order as for a recursive formulation of the Descartes method that first examines the subdivision point m and then makes recursive calls for the left and right subinterval, provided they have a Descartes test of at least 2. However, a straightforward implementation of this recursive formulation would use a number of stack frames proportional to the height of $\mathcal{T}$, which can be much larger than $n/2$, the maximum size of Q. Instead, an implementation of this recursive formulation should eliminate tail recursion as much as possible and perform a recursive call only if there are two subintervals that warrant further subdivision. In a recursive implementation so optimized, the machine stack plays the role of Q.

- A seemingly minor variation of the preceding policy can reduce the size of Q dramatically: order two new elements in line 18 such that the one with smaller v value comes first in Q, with arbitrary order in case of equality. This also leads to a depth-first traversal, but it is, in general, not from left to right any more. We show now that Q has at most $\log n$ elements at all times.

  Let I be the interval recorded in the first element of Q. Consider the path in $\mathcal{T}$ from the root down to I; we regard this path as a sequence of nodes. Let us now restrict to the subsequence of those nodes whose sibling is recorded in Q; we call these nodes $(N_1, \ldots, N_k)$ and their siblings $(S_1, \ldots, S_k)$. We observe that $Q = (I, S_k, \ldots, S_1)$. Let F_i denote the father of N_i and S_i. Let N_0 denote the root. Writing $v(\,\cdot\,)$ for the v value (Descartes test) of a node, our policy together with the variation-diminishing property entails $2v(N_i) \leq v(N_i) + v(S_i) \leq v(F_i) \leq v(N_{i-1})$ and thus $v(N_i) \leq v(N_{i-1})/2$ for all $1 \leq i \leq k$. It follows that $v(N_k) \leq v(N_0)/2^k \leq n/2^k$. Since $v(N_k) \geq 2$, this implies $k \leq \log n - 1$, so Q has at most $\log n$ elements.

  This trick is mentioned by Akritas/Strzeboński [AS05, §1.1]. In terms of a recursive implementation, it is akin to eliminating tail recursion for the smaller of two subproblems, which was already described by Hoare [Hoa62, p. 11].

- Suppose Q is maintained as a sorted sequence, the first element of Q is chosen in line 10, and it is replaced in line 18 by new elements (if any) that are appended in sorted order at the *end* of Q. In other words, Q is used as a queue. Then $\mathcal{T}$ is traversed breadth first from left to right.

  For breadth-first traversal, a bound of n (as opposed to $n/2$) on the size of Q was already known to Krandick [Kra95, Satz 21] [KM06, Thm. 28(1)] independently of Proposition 2.26. We will meet breadth-first traversal again in §3.4.

Before us, Rouillier and Zimmermann [RZ04] have presented a "generic Descartes algorithm" for a unifying treatment of various techniques to navigate the subdivision tree $\mathcal{T}$. They consider subdivision only at $\alpha = 1/2$, but they do not restrict tree traversal to following edges and backtracking (as we have done) and thus attain a variant of the Descartes method that stores only one polynomial at any time. However, in consideration of our algorithm for approximate coefficients in §3.3, we prefer the more restrictive setting in which polynomials are only transformed downwards along tree edges and thus content ourselves with limiting memory usage to $\log n$ polynomials as described above, if memory usage should turn out to be a concern.

3.1.4 A generalized Davenport-Mahler bound

With the goal of a complexity analysis in mind, we devote this section to a lower bound on the distances between certain pairs of roots in terms of degree and coefficients, i.e., natural parameters of input size. The next section will then make the link from distances between suitably chosen pairs of roots to the number of subdivisions performed by *Descartes*. But first we need to introduce some scalar quantities associated with a polynomial.

Definition 3.1. Let $F(X) = f_n \prod_{j=1}^{n}(X - \vartheta_j)$ be a complex polynomial of degree $n \geq 1$. The *Mahler measure* of F is $\mathrm{Mea}(F) := |f_n| \prod_{j=1}^{n} \max\{1, |\vartheta_j|\} > 0$.

We point out two basic properties that we will need later on.

Lemma 3.2. *If* $|f_n| \geq 1$, *in particular if* $F \in \mathbb{Z}[X]$, *then* $\mathrm{Mea}(F) \geq 1$.

Lemma 3.3. *If* $F, G \in \mathbb{C}[X] \setminus \mathbb{C}$, *then* $\mathrm{Mea}(FG) = \mathrm{Mea}(F) \cdot \mathrm{Mea}(G)$.

To make a link between the Mahler measure and the polynomial's coefficients, we recall the usual definition of a polynomial's p-norm: $\|\sum_i f_i X^i\|_p := \sqrt[p]{\sum_i |f_i|^p}$ for $1 \leq p < \infty$ and $\|\sum_i f_i X^i\|_\infty = \max_i |f_i|$.

Proposition 3.4. *Let* $F \in \mathbb{C}[X]$. *Then* $\mathrm{Mea}(F) \leq \|F\|_2 \leq \sqrt{\deg(F) + 1}\, \|F\|_\infty$.

The second inequality is trivial. The first is known as Landau's inequality (Landau, 1905); it is contained in the stronger inequality of Vicente Gonçalves (1950). Elementary proofs appear in the textbooks [Yap00, Lem. 4.14] [BPR06, Prop. 10.9]. We refer to [RS02, §9.1, §9.6] for citations of the original sources and a more detailed treatment that includes the viewpoint of complex analysis.

The form of the Davenport-Mahler bound usually found in the literature (e.g., [BPR06, Prop. 10.23] [Yap00, Thm. 6.28]) relates the Mahler measure to the discriminant of a square-free polynomial. We state a generalization beyond the square-free case in which the discriminant is replaced by a suitable subdiscriminant.

Definition 3.5. Let $F(X) = \sum_{i=0}^{n} f_i X^i$ be a polynomial of degree $n \geq 1$ with coefficients f_i in an integral domain R, and let $1 \leq r \leq n$.

(i) The $(n-r)$th *subdiscriminant* of F is the element $\mathrm{sDisc}_{n-r}(F) \in R$ given by

$$\mathrm{sDisc}_{n-r}(F) := \frac{1}{f_n} \begin{vmatrix} f_n & f_{n-1} & \cdots & f_i & \cdots & \cdots & \cdots & \cdots \\ & \ddots & \ddots & & \ddots & & & \\ & & f_n & f_{n-1} & \cdots & f_i & \cdots & \cdots \\ & & & nf_n & \cdots & \cdots & if_i & \cdots \\ & & \iddots & & & \iddots & & \\ & \iddots & & & \iddots & & & \\ nf_n & \cdots & \cdots & if_i & \cdots & \cdots & \cdots & \cdots \end{vmatrix}. \tag{3.1}$$

In the $(2r-1) \times (2r-1)$ determinant on the right, the first $r-1$ rows hold the coefficients of F arranged in echelon form and padded with zeros or cut off at the right as necessary to fill $2r-1$ columns. The remaining r rows hold the coefficients of F', arranged in echelon form upwards from the bottom row and likewise padded or cut off at the right.

As evident from the first column, the division by f_n is indeed possible within R.

(ii) The *discriminant* of F is $\mathrm{Discr}(F) = \mathrm{sDisc}_0(F)$, the 0th subdiscriminant ($r = n$).

The fundamental property of subdiscriminants is their relation to root differences, which was already known to Sylvester [Syl39]; see also [BPR06, §4.1, §4.2.2].

Theorem 3.6. *If $F(X) = f_n \prod_{j=1}^{n} (X - \vartheta_j)$, $f_n \neq 0$, then*

$$\mathrm{sDisc}_{n-r}(F) = f_n^{2r-2} \sum_{\#I=r} \prod_{\substack{(i,j) \in I^2 \\ i>j}} (\vartheta_i - \vartheta_j)^2, \tag{3.2}$$

with summation over all r-element subsets I of $\{1, \ldots, n\}$.

Clearly, $\mathrm{sDisc}_{n-r}(F) = 0$ if F has less than r distinct complex roots.

Proposition 3.7. *Let $F(X) = f_n \prod_{j=1}^{r} (X - \eta_j)^{m_j}$ be a complex polynomial of degree $n \geq 2$ with exactly r distinct complex roots $\eta_1, \ldots, \eta_r$. Then*

$$\mathrm{sDisc}_{n-r}(F) = f_n^{2r-2} \cdot \prod_{j=1}^{r} m_j \cdot \prod_{1 \leq i < j \leq r} (\eta_i - \eta_j)^2 \neq 0. \tag{3.3}$$

Proof. We consider an arbitrary summand in (3.2) for an index set I of size r. If there are distinct indices $i, j \in I$ such that $\vartheta_i = \vartheta_j$, then this summand vanishes. Conversely, if the ϑ_i, $i \in I$, are pairwise distinct, then these r numbers are exactly the r distinct complex roots $\eta_1, \ldots, \eta_r$, so that this summand is equal to $\prod_{i<j} (\eta_i - \eta_j)^2$. This is the case for exactly $\prod_{j=1}^{r} m_j$ distinct subsets I of $\{1, \ldots, n\}$. □

Lemma 3.8. *If $m_1, \ldots, m_r \in \mathbb{N}$ and $\sum_{i=1}^{r} m_i = n$, then $\prod_{i=1}^{r} m_i \leq 3^{\min\{n, 2n-2r\}/3}$.*

Proof. We begin by considering the function $f(x) = 3^{x/3} - x$ for $x \in (0, \infty)$ and showing that $f(m) \geq 0$ for all $m \in \mathbb{N}$, which is equivalent to $m \leq 3^{m/3}$.

The derivative of $f(x) = \exp((\ln 3)/3 \cdot x) - x$ is $f'(x) = (\ln 3)/3 \cdot \exp((\ln 3)/3 \cdot x) - 1$. Its unique zero is $x = 3 - 3(\ln \ln 3)/\ln 3 < 3$, so that $f(x)$ is strictly increasing for $x \geq 3$.

From $f(3) = 0$ we can thus conclude $f(x) > 0$ for $x > 3$. For the remaining cases $m = 1$ and $m = 2$, it is easy to check that $m^3 \le 3^m$.

We now turn to the product $\prod_{i=1}^r m_i$. How many factors $m_i \ge 2$ occur in this product? Their number is at most $\sum_{i=1}^r (m_i - 1) = n - r$.

In case $n \ge 2r$, this bound is vacuous, as there are only r factors at all. We conclude

$$\prod_{i=1}^r m_i \le \prod_{i=1}^r 3^{m_i/3} = 3^{\sum_{i=1}^r m_i/3} = 3^{n/3} \le 3^{(2n-2r)/3}.$$

In case $n < 2r$, at most $n - r < r$ factors are different from 1, and we may assume that these are $m_1, \ldots, m_{n-r}$. We obtain

$$\prod_{i=1}^r m_i = \prod_{i=1}^{n-r} m_i \le 3^{\sum_{i=1}^{n-r} m_i/3} = 3^{(\sum_{i=1}^r m_i - (r-(n-r)))/3} = 3^{(2n-2r)/3} < 3^{n/3}. \qquad \square$$

We are now ready for the main result of this section.

Theorem 3.9 (Generalized Davenport-Mahler bound). *Let $F(X)$ be a complex polynomial of degree $n \ge 2$ that has exactly $r \le n$ distinct complex roots $V = \{\eta_1, \ldots, \eta_r\}$. Let $\mathcal{G} = (V, E)$ be a directed graph on the roots such that*

(i) every edge $(\alpha, \beta) \in E$ satisfies $|\alpha| \le |\beta|$,

(ii) $\mathcal{G}$ is acyclic, and

(iii) the in-degree of any node is at most 1.

Then

$$\prod_{(\alpha,\beta)\in E} |\alpha - \beta| \ge \frac{\sqrt{|\mathrm{sDisc}_{n-r}(F)|}}{\mathrm{Mea}(F)^{r-1}} \cdot \Big(\frac{\sqrt{3}}{r}\Big)^{\#E} \cdot \Big(\frac{1}{r}\Big)^{r/2} \cdot \Big(\frac{1}{\sqrt{3}}\Big)^{\min\{n,2n-2r\}/3}. \tag{3.4}$$

We note that both sides of (3.4) are invariant under replacing $F(X)$ by $\lambda F(X)$, $\lambda \in \mathbb{C}^*$.

Proof. In the light of the preceding remark, we may assume that F is monic. According to Proposition 3.7,

$$\sqrt{|\mathrm{sDisc}_{n-r}(F)|} = \Big(\prod_{j=1}^r m_j\Big)^{1/2} \cdot \Big|\prod_{i<j} (\eta_j - \eta_i)\Big|. \tag{3.5}$$

The product of differences on the right is the determinant of an $r \times r$ Vandermonde matrix [BPR06, Lem. 4.11] [Yap00, Lem. 6.26]:

$$\prod_{i<j} (\eta_j - \eta_i) = \det W \qquad \text{where} \qquad W = \begin{pmatrix} 1 & \eta_1 & \eta_1^2 & \cdots & \eta_1^{r-1} \\ 1 & \eta_2 & \eta_2^2 & \cdots & \eta_2^{r-1} \\ \vdots & & & & \vdots \\ 1 & \eta_r & \eta_r^2 & \cdots & \eta_r^{r-1} \end{pmatrix}. \tag{3.6}$$

Let us now consider an arbitrary edge $(\alpha, \beta) \in E$. The vertices α and β are roots of F, each of them gives rise to one row of W. Without changing $\det W$, we can subtract the row of α from the row of β, which then takes the form

$$\begin{pmatrix} 0 & \beta - \alpha & \beta^2 - \alpha^2 & \cdots & \beta^j - \alpha^j & \cdots & \beta^{r-1} - \alpha^{r-1} \end{pmatrix}, \qquad 0 \le j < r.$$

Using the equality $\beta^j - \alpha^j = (\beta - \alpha) \sum_{\nu=0}^{j-1} \alpha^\nu \beta^{j-1-\nu}$, we extract a factor of $\beta - \alpha$ and obtain a modified row

$$\begin{pmatrix} 0 & 1 & \beta + \alpha & \cdots & \sum_{\nu=0}^{j-1} \alpha^\nu \beta^{j-1-\nu} & \cdots & \sum_{\nu=0}^{r-2} \alpha^\nu \beta^{r-2-\nu} \end{pmatrix}, \qquad 0 \le j < r.$$

Let W' denote the matrix so modified. It satisfies $\det W = (\beta - \alpha) \det W'$. The theorem is proved by performing a chain $W \rightsquigarrow W' \rightsquigarrow W'' \rightsquigarrow \cdots \rightsquigarrow W^{(\#E)} =: W^*$ of such transformations, one for each edge, to obtain

$$\det W = \prod_E (\beta - \alpha) \cdot \det W^*, \tag{3.7}$$

where the product is over all edges $(\alpha, \beta) \in E$. We can thus interpret the existence of an edge $(\alpha, \beta) \in E$ as saying "row α modifies row β". The conditions (ii) and (iii) guarantee that there is an order of edges in which all these modifications are possible:

Condition (iii) guarantees that every row is modified at most once.

Condition (ii) guarantees that there exists a topological ordering of $\mathcal{G}$, according to which a row is modified itself only after all modifications of other rows by this row have taken place.

Thus, there is a chain of transformations producing the matrix W^* as in (3.7). We proceed to bound $|\det W^*|$. Let w_i be the ith row of W^*. Its Euclidean norm is denoted by $\|w_i\|_2$. Hadamard's inequality states $|\det W^*| \le \prod_{i=1}^r \|w_i\|_2$. (In geometrical terms: The volume of a parallelepiped with edge lengths $\|w_i\|_2$ is maximized by orthogonal edges.) It remains to estimate these norms. The untransformed rows of W^* have the form

$$w_i = \begin{pmatrix} 1 & \eta_i & \eta_i^2 & \cdots & \eta_i^j & \cdots & \eta_i^{r-1} \end{pmatrix}, \qquad 0 \le j < r,$$

so that

$$\|w_i\|_2^2 = \sum_{j=0}^{r-1} |\eta_i|^{2j} \le r \max\{1, |\eta_i|\}^{2(r-1)}.$$

The transformed rows of W^* have the form

$$w_i = \begin{pmatrix} 0 & 1 & \eta_i + \alpha & \cdots & \sum_{\nu=0}^{j-1} \alpha^\nu \eta_i^{j-1-\nu} & \cdots & \sum_{\nu=0}^{r-2} \alpha^\nu \eta_i^{r-2-\nu} \end{pmatrix}, \qquad 0 \le j < r,$$

with another root α. Condition (i) states $|\alpha| \le |\eta_i|$, so we can bound the jth entry by

$$\left|\sum_{\nu=0}^{j-1} \alpha^\nu \eta_i^{j-1-\nu}\right| \le \sum_{\nu=0}^{j-1} |\alpha|^\nu \, |\eta_i|^{j-1-\nu} \le j \, |\eta_i|^{j-1}.$$

Using $\sum_{j=0}^m j^2 = (2m^3 + 3m^2 + m)/6 < (m+1)^3/3$, which is easily verified by induction, it follows that

$$\|w_i\|_2^2 \le \sum_{j=0}^{r-1} j^2 \, |\eta_i|^{2(j-1)} \le r^3/3 \cdot \max\{1, |\eta_i|\}^{2(r-2)}.$$

Thus, Hadamard's inequality yields

$$|\det W^*| \le \prod_{i=1}^r \|w_i\|_2 \le \left(\frac{r}{\sqrt{3}}\right)^{\#E} \cdot r^{r/2} \cdot \prod_{i=1}^r \max\{1, |\eta_i|\}^{r-1}. \tag{3.8}$$

For the last product, we observe

$$\prod_{i=1}^r \max\{1, |\eta_i|\} \le \prod_{i=1}^r \max\{1, |\eta_i|\}^{m_i} = \mathrm{Mea}(F). \tag{3.9}$$

In combination, Equations (3.5), (3.6), (3.7), (3.8), (3.9) and Lemma 3.8 show

$$\begin{aligned}\sqrt{|\mathrm{sDisc}_{n-r}(F)|} &= \Big(\prod_{j=1}^{r} m_j\Big)^{1/2} \cdot \Big(\prod_{E} |\alpha-\beta|\Big) \cdot |\det W^*| \\ &\le \sqrt{3}^{\min\{n,2n-2r\}/3} \cdot \Big(\prod_{E} |\alpha-\beta|\Big) \cdot \Big(\frac{r}{\sqrt{3}}\Big)^{\#E} \cdot r^{r/2} \cdot \mathrm{Mea}(F)^{r-1}.\end{aligned}$$

The claim follows by rearranging terms. □

This theorem has come about by a series of generalizations. Its proof technique originates with Mahler[3] [Mah64, Thm. 2], who treated the special case of a square-free polynomial ($r = n$) and a single pair of roots ($\#E = 1$). Davenport [Dav85, Prop. I.5.8] took advantage of the fact that Mahler's proof technique extends to products of several root distances. He gave the bound (3.4) for $r = n$, with the last factor reducing to 1. Davenport's crucial achievement is that the dominant factor on the right-hand side, namely the first, is independent of $k = \#E$; if we form $k \ge 2$ permissible pairs of roots, the resulting lower bound on the product of their distances is much better (i.e., larger) than the kth power of Mahler's bound on the distance of an arbitrary pair of roots. The resulting advantage for the analysis of root isolation will become apparent in §3.2.2.

Davenport [Dav85] applied his bound to the number of subdivisions in root isolation with Sturm's theorem, similar to our subsequent considerations, but with an exclusive interest in real roots, he restricted the formulation of his bound to all pairs of adjacent real roots. Johnson [Joh91, Thm. 11] [Joh98, Thm. 10], working on the Descartes method, lifted this restriction and formulated the broadest condition on the pairing of roots supported by the proof technique. Du, Sharma and Yap (2005) replaced Johnson's somewhat technical indexing condition by the equivalent but more intuitive formulation in terms of the graph $\mathcal{G}$ that we have seen above. The final version of their work appeared as [DSY07]. Their formulation is also used in [BPR06, Prop. 10.23].

All these previous forms of the theorem treat the case $r = n$ of a square-free polynomial, using its discriminant. The formulation above, using the $(n-r)$th subdiscriminant for the general case $r \le n$, appears to be new. For polynomials $F(X) \in \mathbb{Z}[X]$ with an unconstrained value of r anywhere between 2 and n, this generalization is not particularly useful, because other techniques apply for the reduction to the square-free case, as we shall see in the proof of Corollary 3.11. On the other hand, our study of polynomials with algebraic coefficients in §3.4.3 will benefit considerably from this generalization.

A bound similar to the traditional case $r = n$ of Theorem 3.9 appears in [Mig95, Thm. 1]. Instead of $\mathrm{Mea}(F)^{n-1}$, it uses a product of root magnitudes with varying exponents of $n-1$ or less.

Let us now focus on the case of integral coefficients.

Corollary 3.10. *If $F \in \mathbb{Z}[X]$ in Theorem 3.9, then $|\mathrm{sDisc}_{n-r}(F)| \ge 1$.*

Proof. The choice of r in Theorem 3.9 satisfies the conditions of Proposition 3.7, so $\mathrm{sDisc}_{n-r}(F) \ne 0$. On the other hand, $\mathrm{sDisc}_{n-r}(F)$ is an integer by (3.1). □

[3] Kurt Mahler (1903–1988), prominent number theorist of German-Jewish origin, emigrated to escape the national socialists. According to the obituaries by Cassels (*Acta Arith.* **58** (1991), pp. 215–228; also *Bull. LMS* **24** (1992), pp. 381–397) and van der Poorten (*J. Autral. Math. Soc. Ser. A* **51** (1991), pp. 343–380), the measure $|f_n| \prod_j \max\{1, |\vartheta_j|\}$ occurs before Mahler (cf. Proposition 3.4), but he was the first to study it systematically.

Corollary 3.11. *Let $F(X) = \sum_{i=0}^{n} f_i X^i$ be a polynomial of degree $n \geq 2$ with integer coefficients such that $|f_i| \leq 2^\tau$ for $0 \leq i \leq n$. Let V be the set of distinct complex roots of F. If a directed graph $\mathcal{G} = (V, E)$ satisfies conditions (i–iii) in Theorem 3.9, then*

$$\prod_{(\alpha,\beta)\in E} |\alpha - \beta| \geq \frac{1}{((n+1)^{1/2}\, 2^\tau)^{n-1}} \cdot \left(\frac{\sqrt{3}}{n}\right)^{\#E} \cdot \left(\frac{1}{n}\right)^{n/2}. \tag{3.10}$$

Notice that the corollary applies to any number $r \leq n$ of distinct roots. The proof below follows the standard approach to reduce this claim to the traditional case $r = n$ of Theorem 3.9, even if F has multiple roots. We intentionally ignore our generalization to $r \leq n$ for a moment and follow that approach. This highlights which special properties of $\mathbb{Z}$ it requires.

Proof. The ring $\mathbb{Z}$ of integers is a unique factorization domain. Standard arguments based on Gauss' Lemma (primitive polynomials have a primitive product) show that whenever $F \in \mathbb{Z}[X]$ has a factorization $F = F_1 F_2$ with factors $F_1, F_2 \in \mathbb{Q}[X]$, there exist constant multiples $\overline{F}_1, \overline{F}_2 \in \mathbb{Z}[X]$ of F_1 and F_2 such that $F = \overline{F}_1 \overline{F}_2$, see, e.g., [BPR06, Lem. 10.17] [vdW93, §30]. Therefore, when we factor $F = F_1 \cdot \gcd(F, F')$ in $\mathbb{Q}[X]$ to produce the *square-free part* F_1 of F, we can take F_1 to be a divisor of F in $\mathbb{Z}[X]$. In particular, the leading coefficient of F_1 divides the leading coefficient of F, and since $a|b$ implies $|a| \leq |b|$ for integers a and $b \neq 0$, we obtain $\mathrm{Mea}(F_1) \leq \mathrm{Mea}(F) \leq (n+1)^{1/2}\, 2^\tau$. Now we apply Theorem 3.9 to F_1. Denoting its degree by n_1, we obtain a lower bound of

$$\frac{\sqrt{|\mathrm{Discr}(F_1)|}}{\mathrm{Mea}(F_1)^{n_1-1}} \cdot \left(\frac{\sqrt{3}}{n_1}\right)^{\#E} \cdot \left(\frac{1}{n_1}\right)^{n_1/2} \geq \frac{1}{((n+1)^{1/2}\, 2^\tau)^{n_1-1}} \cdot \left(\frac{\sqrt{3}}{n_1}\right)^{\#E} \cdot \left(\frac{1}{n_1}\right)^{n_1/2}$$

The right-hand side is a decreasing function of n_1, so we may substitute the estimate $n_1 \leq n$ to establish the claim. □

3.1.5 Size of the subdivision tree

Consider the subdivision tree $\mathcal{T}$ generated by executing *Descartes*(A_{in}, I_0) on a real polynomial A_{in} of degree n. The purpose of this section is to derive an upper bound on the number of internal nodes of $\mathcal{T}$. This is the number of subdivisions performed by *Descartes* and thus a natural object of study in a complexity analysis; in particular, it is finite iff *Descartes* terminates. If the binary tree $\mathcal{T}$ has a finite number T of internal nodes, it has $T + 1$ leaves and $2T + 1$ nodes in total.

We begin with a notion that quantifies the effect of subdivision on the width of intervals.

Definition 3.12. We say that the real number $\rho > 1$ is a *subdivision ratio bound* for the subdivision tree $\mathcal{T}$ if for any non-root node I and its parent J in $\mathcal{T}$ we have $|J| \,/\, |I| \geq \rho$.

By choice of α in line 11 of procedure *Descartes*, $\rho = 4/3$ is always a subdivision ratio bound for $\mathcal{T}$. As every internal node J has two children I_1, I_2 and $|J| = |I_1| + |I_2|$, the best (i.e., largest) possible subdivision ratio bound is $\rho = 2$. This is attained if the choice of α is fixed to $\alpha = 1/2$, meaning that intervals are bisected evenly.

The internal nodes of $\mathcal{T}$ satisfy the following tetrachotomy:

Proposition 3.13. *If an interval $I = (c, d)$ satisfies $v = \mathit{DescartesTest}(A_{\mathrm{in}}, I) \geq 2$, then one and only one of the following conditions holds:*

(i) I contains no real root of A_{in}, and $\mathrm{OL}_{\leq}(0, 0, n; I)$ contains a complex-conjugate pair of imaginary roots $\xi \pm \mathrm{i}\eta$, $\eta > 0$, with distance $0 < |(\xi + \mathrm{i}\eta) - (\xi - \mathrm{i}\eta)| < |I|$.

(ii) I contains exactly one real root ϑ of A_{in}, which is simple, and $\mathrm{OL}_{\leq}(1, 1, n; I)$ contains a complex-conjugate pair of imaginary roots $\xi \pm \mathrm{i}\eta$, $\eta > 0$; for both of them, the distance to ϑ is $0 < |\vartheta - (\xi \pm \mathrm{i}\eta)| < 2/\sqrt{3} \cdot |I|$.

(iii) I contains two distinct real roots ϑ, ϑ' of A_{in}, and their distance is $0 < |\vartheta - \vartheta'| < |I|$.

(iv) I contains exactly one real root ϑ of A_{in}, and its multiplicity is at least 2.

Proof. Let p be the number of real roots of A_{in} in I, counted with multiplicities.

Case $p = 0$. Clearly, (ii–iv) do not hold. To prove (i) by contradiction, suppose $\mathrm{OL}_{\leq}(0, 0, n; I)$ contains no roots of A_{in}. Then Theorem 2.32 for $p = q = 0$ (a.k.a. "one-circle theorem", Proposition 2.33) implies $v = 0$ in contradiction to the hypothesis $v \geq 2$. Thus, $\xi \pm \mathrm{i}\eta$ exist, and Proposition 2.35 (on page 33) yields the bound on their distance.

Case $p = 1$. Clearly, (i), (iii), and (iv) do not hold. To prove (ii) by contradiction, suppose $\mathrm{OL}_{\leq}(1, 1, n; I)$ contains no further roots of A_{in}. Theorem 2.32 for $p = q = 1$ (a.k.a. "two-circle theorem", Proposition 2.34) implies $v = 1$ in contradiction to the hypothesis $v \geq 2$. Thus, $\xi \pm \mathrm{i}\eta$ exist, and Proposition 2.35 yields the bound on their distance to ϑ.

Case $p \geq 2$. Clearly, either (iii) or (iv) holds, and all other items are false. □

This proposition justifies the subsequent definition, in which we declare one pair (α, β) of roots to be responsible for the subdivision occurring at each internal node of $\mathcal{T}$. In case (α, β) is not determined uniquely, an arbitrary choice is made.

Definition 3.14. To each internal node $I = (c, d)$ of the subdivision tree $\mathcal{T}$, we assign one ordered pair (α, β) of complex roots of A_{in} as *responsible for subdivision of I*. According to the cases (i–iv) distinguished in Proposition 3.13 and the respective existence statements, (α, β) is chosen as follows:

(i) If $\xi > (c + d)/2$, let $(\alpha, \beta) = (\xi - \mathrm{i}\eta,\ \xi + \mathrm{i}\eta)$, else let $(\alpha, \beta) = (\xi + \mathrm{i}\eta,\ \xi - \mathrm{i}\eta)$. (Either way, we have $|\alpha| = |\beta|$.)

(ii) If $\xi > (c + d)/2$, let $\{\alpha, \beta\} = \{\vartheta,\ \xi + \mathrm{i}\eta\}$, else let $\{\alpha, \beta\} = \{\vartheta,\ \xi - \mathrm{i}\eta\}$; then choose an order such that the pair (α, β) satisfies $|\alpha| \leq |\beta|$.

(iii) Let $\{\alpha, \beta\} = \{\vartheta, \vartheta'\}$ and choose an order such that the pair (α, β) satisfies $|\alpha| \leq |\beta|$.

(iv) Let $(\alpha, \beta) = (\vartheta, \vartheta)$. (Clearly, $|\alpha| = |\beta|$.)

In cases (i–iii), we say the internal node I is *regular*; in case (iv), it is *singular*.

Proposition 3.15. *Let $I_1, \ldots, I_k$ be regular internal nodes of the subdivision tree $\mathcal{T}$ such that any two of these intervals are disjoint. For $1 \leq i \leq k$, let (α_i, β_i) be the pair of roots responsible for subdivision of I_i. Then $(\alpha_i, \beta_i) \neq (\alpha_j, \beta_j)$ for $i \neq j$, and a graph $\mathcal{G}$ with edge set $E = \{(\alpha_i, \beta_i) \mid 1 \leq i \leq k\}$ satisfies conditions (i–iii) of Theorem 3.9.*

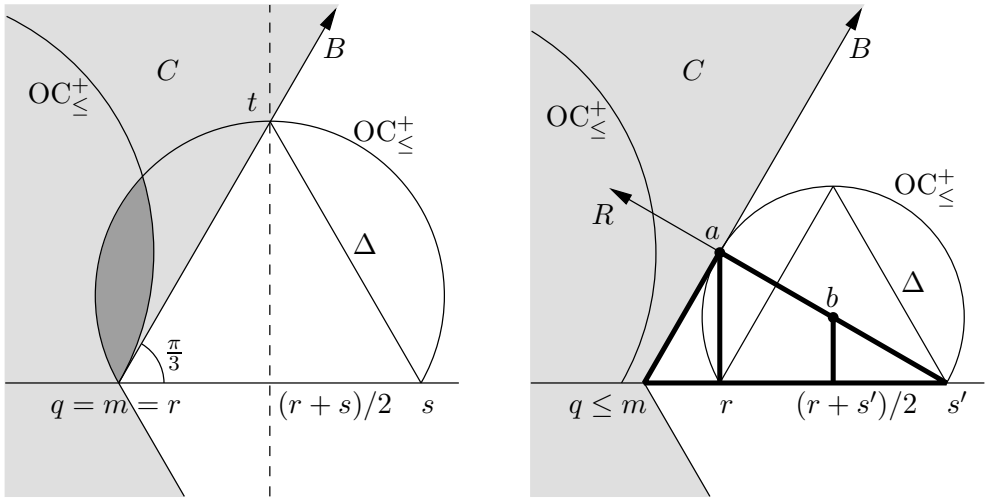
Recall that $I_i \subseteq \mathrm{OL}_{\leq}(0, 0, n; I_i) \subseteq \mathrm{OL}_{\leq}(1, 1, n; I_i)$ and so $\alpha_i, \beta_i \in \mathrm{OL}_{\leq}(1, 1, n; I_i)$. As the first step of a proof, we investigate how much the sets $\mathrm{OL}_{\leq}(1, 1, n; I_i)$, $1 \leq i \leq k$, overlap.

Lemma 3.16. *Let (p, q) and (r, s) be two nodes of $\mathcal{T}$ such that $q \leq r$.*

(i) If $q = r$, then $\mathrm{OL}_{\leq}(1, 1, n;\ (p, q)) \cap \mathrm{OL}_{\leq}(1, 1, n;\ (r, s)) \subseteq ((p + q)/2,\ (r + s)/2) + \mathrm{i}\mathbb{R}$.

(ii) If $q < r$, then $\mathrm{OL}_{\leq}(1, 1, n;\ (p, q)) \cap \mathrm{OL}_{\leq}(1, 1, n;\ (r, s)) = \emptyset$.

Proof of Lemma 3.16. Let (c, d) be the deepest common ancestor of (p, q) and (r, s) in $\mathcal{T}$. Suppose the children of (c, d) are (c, m) and (m, d). By choice of (c, d), (p, q) is a descendant of, or equal to, (c, m), and (r, s) is a descendant of, or equal to, (m, d), so $q \leq m \leq r$.



(a) Case (i). **(b)** Case (ii).

Figure 3.1: Proof of Lemma 3.16.

The open cone $C = \{\lambda e^{i\varphi} + m \mid \lambda > 0,\ \pi/3 < \varphi < 5\pi/3\}$ in the complex plane contains $\mathrm{OL}_{\leq}(1,1,n;(c,m))$, because its bounding rays are the tangents to $\mathrm{OA}^{\pm}_{\leq}(1,1,n;(c,m))$ at m. Since $(p,q) \subseteq (c,m)$, also $\mathrm{OL}_{\leq}(1,1,n;(p,q)) \subseteq C$. See Figure 3.1 for what follows.

Ad (i). Since $q = r$, we have $q = m = r$. By Proposition 2.34, $\mathrm{OC}^+_{\leq}(1,1,n;(r,s))$ is the circumcircle of the equilateral triangle Δ with base $[r,s]$ in the closed upper half-plane. The left edge of Δ lies on the upper boundary ray B of C, hence B leaves $\mathrm{OC}^+_{\leq}(1,1,n;(r,s))$ at the tip t of Δ, whose real part, by symmetry, is $(r+s)/2$. Thus, $\mathrm{OD}^+_{\leq}(1,1,n;(p,q)) \cap \mathrm{OD}^+_{\leq}(1,1,n;(r,s)) \subseteq C \cap \mathrm{OD}^+_{\leq}(1,1,n;(r,s)) \subseteq (-\infty,\,(r+s)/2) + i\mathbb{R}$. By horizontal and vertical symmetry, $\mathrm{OD}^{\pm}_{\leq}(1,1,n;(p,q)) \cap \mathrm{OD}^{\pm}_{\leq}(1,1,n;(r,s)) \subseteq ((p+q)/2,\,(r+s)/2) + i\mathbb{R}$.

Ad (ii). We discuss the case $q \leq m < r$; the excluded case $q < m = r$ is symmetric. Let (m,u) be the deepest ancestor of (r,s) that has m as its left endpoint. Suppose (m,u) is subdivided into (m,t) and (t,u). Since $m < r$, the interval (r,s) is a descendant of (t,u), that is, $(r,s) \subseteq (t,u)$. From $t = (1-\alpha)m + \alpha u$ with $1/4 \leq \alpha \leq 3/4$ it follows that $u - t \leq 3(t-m)$. Hence $s - r \leq u - t \leq 3(t-m) \leq 3(r-m)$. Geometrically, this says that the width of (r,s) is at most three times its distance to the apex m of C.

We will now deduce that $C \cap \mathrm{OL}_{\leq}(1,1,n;\,(r,s)) = \emptyset$. Since $\mathrm{OL}_{\leq}(1,1,n;\,\cdot\,)$ is inclusion-monotonic, if suffices to treat the interval $(r,s') \supseteq (r,s)$ with $s' = r + 3(r-m)$ that realizes the maximal interval width $s' - r = 3(r-m)$. We will show that the upper boundary ray B of C is tangent to $\mathrm{OC}^+_{\leq}(1,1,n;\,(r,s'))$. Together with the symmetric statement for the lower half-plane, this implies $C \cap \mathrm{OL}_{\leq}(1,1,n;\,(r,s')) = \emptyset$ and establishes (ii).

Consider the ray R in the upper half-plane that makes an angle $\pi/6$ with $(-\infty,s')$ at its origin s'. The ray R intersects B at a point a. The triangle with vertices m, s', a has angles $\pi/3$, $\pi/6$, $\pi/2$, so by elementary trigonometry, the perpendicular from a onto the hypotenuse $[m,s']$ divides the latter with ratio $1:3$, that is, at point r. As $\mathrm{OC}^+_{\leq}(1,1,n;\,(r,s'))$ is the circumcircle of the equilateral triangle with base $[r,s']$ by Proposition 2.34, its center b lies on R covertical to the midpoint of the segment $[r,s']$. Thus, b is the midpoint of the segment $[a,s']$. It follows that $[a,s']$ is a diameter of $\mathrm{OC}^+_{\leq}(1,1,n;\,(r,s'))$. As $[a,s']$ is perpendicular to B at a, it follows that B is tangent to $\mathrm{OC}^+_{\leq}(1,1,n;\,(r,s'))$ at a. □

Proof of Proposition 3.15. As $I_1, \dots, I_k$ are disjoint, they have a well-defined order on the real line, and we can choose indices such that $I_1 < I_2 < \cdots < I_k$. Let us now verify that conditions (i–iii) of Theorem 3.9 are satisfied.

Condition (i) is immediate: $|\alpha_i| \le |\beta_i|$ holds by construction in Definition 3.14.

Let us show that $\beta_i \neq \beta_j$ for $i < j$ to establish condition (iii) and to obtain $(\alpha_i, \beta_i) \neq (\alpha_j, \beta_j)$. If β_i or β_j is real, $\beta_i \neq \beta_j$ is immediate from $I_i \cap I_j = \emptyset$; it remains to consider the case that both are imaginary. For a contradiction, suppose $\beta := \beta_i = \beta_j$. Since $\beta \in \mathrm{OL}_{\le}(1, 1, n; I_i) \cap \mathrm{OL}_{\le}(1, 1, n; I_j)$, Lemma 3.16 yields $I_i = (a, m)$, $I_j = (m, b)$ and $(a+m)/2 < \operatorname{Re}\beta < (m+b)/2$. But the construction in cases (i) and (ii) of Definition 3.14 implies $\operatorname{Im}\beta > 0$ for $\beta = \beta_i$ and $\operatorname{Im}\beta < 0$ for $\beta = \beta_j$ – a contradiction to equality.

Finally, let us demonstrate acyclicity, which is condition (ii). As singular nodes are excluded, there are no cycles $\alpha_i = \beta_i$ of length 1. Let ℓ_i be the perpendicular bisector of I_i for $1 \le i \le k$. The vertical lines $\ell_1, \dots, \ell_k$ cut the complex plane into $k+1$ vertical open stripes. By Lemma 3.16, the endpoints of the edge (α_i, β_i) lie in the union of ℓ_i and the two stripes adjacent to it. When we embed the edges of $\mathcal{G}$ as straight-line segments, this implies that every line ℓ_i is intersected by at most one edge. Consequently, a cycle in $\mathcal{G}$, if any, lies entirely within one stripe. As each stripe contains endpoints of at most two edges, cycles of length 3 or more are impossible. What about a cycle of length 2, i.e., $(\alpha_i,\ \beta_i = \alpha_{i+1},\ \beta_{i+1} = \alpha_i)$? Since a real root in the pair (α_i, β_i) is an element of I_i, which is disjoint from I_{i+1}, both nodes must be imaginary roots, i.e., both I_i and I_{i+1} realize case (i) of Proposition 3.13 and Definition 3.14. This, however, is absurd, since $\mathrm{OL}_{\le}(0, 0, n; I_i) \cap \mathrm{OL}_{\le}(0, 0, n; I_{i+1}) = \emptyset$. □

We are now ready to bound subdivision depth in terms of root distances.

Lemma 3.17. *Let ρ be a subdivision ratio bound for the subdivision tree $\mathcal{T}$. Consider an internal node I of $\mathcal{T}$ at depth $k \ge 0$.*

(i) If (α, β) is responsible for subdivision of I, then $|\alpha - \beta| < 2/\sqrt{3} \cdot |I| \le 2/\sqrt{3} \cdot |I_0| / \rho^k$.

(ii) If I is regular, then $|\alpha - \beta| > 0$ and $k < \log_\rho(2/\sqrt{3} \cdot |I_0| / |\alpha - \beta|)$.

Proof. Immediate from Definition 3.12 and Proposition 3.13. □

Definition 3.18. An internal node of $\mathcal{T}$ is called *terminal* if its children are leaves. A path in $\mathcal{T}$ from the root to a regular terminal node is called an *rt-path*. We denote by $\mathcal{G}(\mathcal{T})$ the directed graph whose vertices are the distinct roots of A_{in} and whose edge set E consists of all pairs (α, β) that are responsible for subdivision of a regular terminal node of $\mathcal{T}$.

Theorem 3.19. *Consider the subdivision tree $\mathcal{T}$ generated by executing Descartes(A_{in}, I_0) on a real polynomial A_{in} of degree n. Let ρ be a subdivision ratio bound for $\mathcal{T}$.*

(i) The graph $\mathcal{G}(\mathcal{T})$ satisfies conditions (i–iii) of Theorem 3.9 and has at most $n/2$ edges.

(ii) If all roots of A_{in} in I_0 are simple, then $\mathcal{T}$ is finite, and all its internal nodes lie on an rt-path.

(iii) The sum P of the lengths of all rt-paths of $\mathcal{T}$ satisfies

$$P \le \log_\rho \Big(\prod_{(\alpha,\beta)} \frac{|I_0|}{|\alpha - \beta|} \Big) + \frac{n}{2} \log_\rho \Big(\frac{2}{\sqrt{3}} \Big), \tag{3.11}$$

with (α, β) ranging over the edges of $\mathcal{G}(\mathcal{T})$.

(iv) The number of all internal nodes of $\mathcal{T}$ that lie on an rt-path is at most $P + 1$.

Proof. Ad (i). No terminal node is a descendant of another, so any two terminal nodes are disjoint intervals, and the first claim reduces to Proposition 3.15. Concerning the second, Corollary 2.27 yields $n \geq \mathit{DescartesTest}(A_{\text{in}}, I_0) \geq \sum_I \mathit{DescartesTest}(A_{\text{in}}, I)$, where I ranges over all terminal nodes. Each terminal node I has $\mathit{DescartesTest}(A_{\text{in}}, I) \geq 2$, so there are at most $n/2$ of them.

Ad (ii). If all roots of A_{in} in I_0 are simple, then all internal nodes of $\mathcal{T}$ are regular, and the distance of any pair of roots responsible for a subdivision is positive. Since there are only finitely many pairs of roots, there is a minimum distance $s > 0$ between any two roots responsible for a subdivision and thus, by Lemma 3.17(ii), a maximum depth $\lfloor \log_\rho(2/\sqrt{3} \cdot |I_0| / s) \rfloor$ of internal nodes in the binary tree $\mathcal{T}$, hence $\mathcal{T}$ is finite. It follows that all internal nodes lie on a path from the root to a terminal node, which is also regular. This proves claim (ii).

Ad (iii). Consider any path $(I_0, \ldots, I_k)$ from the root node I_0 to a regular terminal node I_k. Its length k is the depth of I_k, which is bounded by Lemma 3.17(ii) in terms of the pair (α, β) responsible for subdivision of I_k as $k < \log_\rho(|I_0| \,/\, |\alpha - \beta|) + \log_\rho(2/\sqrt{3})$. Summing over all regular terminal nodes, of which there are at most $n/2$, we attain the claimed bound.

Ad (iv). To each non-root node I on an rt-path, we associate the edge to the parent of I; this map is injective, and each edge in its image is counted at least once in P. Adding 1 to account for the root node, we obtain the claimed bound. □

With this theorem in our hands, we can now bound the size of $\mathcal{T}$ with the Davenport-Mahler bound. We will do that in a moment in §3.2.2 for the case of integer coefficients and $\rho = 2$. We will do it again several times in §3.3 and §3.4 for bitstream coefficients, but with details depending on the origin of A_{in}. An alternative pairing of roots for the Davenport-Mahler bound is presented in §3.4.4

3.2 The Descartes method for exact integer coefficients

3.2.1 Generalities

Let us now discuss a more specific form of Descartes method: The input polynomial $A_{\text{in}}(X)$ and all polynomials obtained from it are represented with exact integer coefficients. Furthermore, intervals (c, d) are subdivided by bisection, i.e., at the midpoint $m = (c + d)/2$. In terms of our generic procedure *Descartes*, this means that the subdivision parameter α in line 11 is fixed as $\alpha = 1/2$. This is the setting considered originally by Collins and Akritas [CA76]. We will first study general aspects of this *integer Descartes method* and then discuss several concrete algorithms implementing it.

3.2.2 Size of the subdivision tree

Theorem 3.20. *Consider a polynomial $A_{\text{in}} \in \mathbb{Z}[X]$ and an open interval I_0 such that all roots of $A_{\text{in}}(X)$ in I_0 are simple. Let $\mathcal{T}$ be the subdivision tree produced by the integer Descartes method invoked for A_{in} and I_0. If A_{in} has degree $n \geq 2$ and each coefficient is an integer of magnitude at most 2^τ, then the number T of internal nodes in $\mathcal{T}$ satisfies*

$$T < (n-1)\tau + n/2 \cdot \max\{0, \log|I_0|\} + 3n/2 \cdot \log n + 1. \tag{3.12}$$

The proof will use the following technical lemma.

Lemma 3.21. *If $n \in \mathbb{N}$, $n \geq 2$, then $n/2 \cdot \log n > (n-1)/2 \cdot \log(n+1) + (n/2)(1 - \log 3)$.*

Proof of Lemma 3.21. Since $1 - \log 3 < 0$, it suffices to show $n \log n > (n-1)\log(n+1)$. W.l.o.g., $n \geq 3$. Consider the function $f\colon [0,1] \to \mathbb{R}$, $x \mapsto (n-x)\ln(n+x)$. It has the derivatives $f'(x) = -\ln(n+x) + (n-x)/(n+x)$ and $f''(x) = -(3n+x)/(n+x)^2$. As $f'(0) = 1 - \ln n < 0$ and $f''(x) < 0$ for all $x \in [0,1]$, the function f is strictly decreasing. It follows that $n \log n = f(0)/\ln 2 > f(1)/\ln 2 = (n-1)\log(n+1)$, as desired. □

Proof of Theorem 3.20. Theorem 3.19(iii/iv) yields

$$T \leq k \log |I_0| + \log \prod (1/|\alpha - \beta|) + n/2 \cdot \log(2/\sqrt{3}) + 1,$$

where k is the number of factors in the product over the root pairs. Since $0 \leq k \leq n/2$, the first term is bounded by $n/2 \cdot \max\{0, \log |I_0|\}$. For the rest, we obtain the following estimate from Corollary 3.11 to the Davenport-Mahler bound:

$$\begin{aligned} \ldots &\leq \big((n-1)(1/2 \cdot \log(n+1) + \tau) + n \log n - n/4 \cdot \log 3\big) + n/2 - n/4 \cdot \log 3 + 1 \\ &= (n-1)\tau + n \log n + (n-1)/2 \cdot \log(n+1) + (n/2)(1 - \log 3) + 1 \\ &< (n-1)\tau + 3n/2 \cdot \log n + 1, \end{aligned}$$

where the last inequality comes from Lemma 3.21. □

Corollary 3.22 (Eigenwillig-Sharma-Yap (2006)). *If in Theorem 3.20, the initial interval I_0 is such that $|I_0| \leq 2^{O(\tau)}$, then $T = O(n \cdot (\tau + \log n))$. This holds in particular for $I_0 = [-R, +R]$, where R is any of the bounds from §2.4 on the magnitudes of roots of A_{in}.*

This bound was first established by V. Sharma and C. Yap in joint work [ESY06, Cor. 3.5] with the author of this thesis. Krandick [Kra95, Sätze 21&47], also in revised form with Mehlhorn [KM06, Thm. 29], has previously given bounds on the subdivision tree $\mathcal{T}$ that lead to the strictly weaker estimate $T = O(n \log n \cdot (\tau + \log n))$ (Krandick and Mehlhorn, personal communication).

To demonstrate the quality of this bound, we will now construct a family of example input polynomials that force the Descartes method (or any other root isolation method that uses recursive subdivision) to subdivide deeply.

Theorem 3.23 (Generalized Mignotte polynomials). *Take any integers $a \geq 2$, $n \geq 3$, and $p \geq 1$. Consider the polynomial $A_p(X) = X^n - p(aX-1)^2$.*

(i) If p is prime, then $A_p(X)$ is irreducible and square-free.

(ii) It holds that $A_p(0) < 0$ and $A_p(1) \leq 0$. If n is even, then $A_p(x) \to +\infty$ for $x \to \pm\infty$.

(iii) Let $h = a^{-n/2-1}$. It holds that $A_p(a^{-1}) > 0$ and $A_p(a^{-1} - h) < 0$.
It also holds that $A_p(a^{-1} + h) < 0$ if $p = 2$ and $a \geq 3$, or if $p \geq 3$ and $a \geq 2$.

The proof of the last estimate requires the following technicality.

Lemma 3.24. *Let $p > 1$. The function $f\colon (0,\infty) \to (0,\infty)$, $x \mapsto (x/\ln p)^{1/x}$ attains its global maximum at $x = \mathrm{e} \ln p$, its value is $f(\mathrm{e}\ln p) = \mathrm{e}^{1/(\mathrm{e}\ln p)}$.*

Proof. As the natural logarithm is strictly increasing, it suffices to seek the maximum of $g(x) = \ln(f(x))$. We have

$$\begin{aligned} g(x) &= x^{-1}(\ln x - \ln\ln p), \\ g'(x) &= -x^{-2}(\ln x - \ln\ln p - 1), \\ g''(x) &= x^{-3}(2\ln x - 2\ln\ln p - 3). \end{aligned}$$

We observe $g'(x) = 0 \;\Leftrightarrow\; \ln x = \ln\ln p + 1 \;\Leftrightarrow\; x = \mathrm{e}\ln p$. At this unique zero of g', we have $g''(\mathrm{e}\ln p) = -(\mathrm{e}\ln p)^{-3} < 0$, since $\ln p > 0$. Hence g is increasing for $x < \mathrm{e}\ln p$ and decreasing for $x > \mathrm{e}\ln p$, so that indeed at $x = \mathrm{e}\ln p$ the global maximum is attained. □

Proof of Theorem 3.23. Ad (i). $A_p(X)$ is irreducible by Eisenstein's criterion [vdW93, §31] with prime number p. In particular, $A_p(X)$ is coprime to its derivative and thus square-free.

Ad (ii). We have $A_p(0) = -p < 0$ and $A_p(1) = 1 - p(a-1)^2 \le 0$. If n is even, the positivity of the leading coefficient implies $A_p(x) \to +\infty$ for $x \to \pm\infty$.

Ad (iii). It is clear that $A_p(a^{-1}) = a^{-n} > 0$ and $A_p(a^{-1} - h) = (a^{-1} - h)^n - pa^{-n} < (a^{-1})^n - pa^{-n} \le 0$. Let us now consider $A_p(a^{-1} + h) = (a^{-1} + h)^n - pa^{-n}$ with $p \ge 2$.

We have $A_p(a^{-1}+h) < 0 \;\Leftrightarrow\; (a^{-1}+h)^n < pa^{-n} \;\Leftrightarrow\; (1+ah)^n < p \;\Leftrightarrow\; n\ln(1+ah) < \ln p$. As $\ln x$ is a concave function, we can bound it from above by its tangent at $x = 1$, which is $y = x - 1$. It follows that $\ln(1+ah) < ah = a^{-n/2}$, so it is sufficient for $A_p(a^{-1}+h) < 0$ to have $na^{-n/2} < \ln p$, which is equivalent to $a > (n/\ln p)^{2/n}$. Now, in turn, it is sufficient for a to be larger than the maximum of the right-hand side over all $n \in (0, \infty)$, and by the preceding lemma, that means $a > \mathrm{e}^{2/(\mathrm{e}\ln p)}$. This lower bound on a is a decreasing function of p. Thus, it suffices for $p \ge 3$ to have $a > \mathrm{e}^{2/(\mathrm{e}\ln 3)} = 1.95...$, and for $p = 2$, to have $a > \mathrm{e}^{2/(\mathrm{e}\ln 2)} = 2.89...$. □

Our proof of (iii) extends an argument given by Krandick [Kra95, Satz 37][4] from $p = 2$ to the simpler cases $p \ge 3$. Mignotte [Mig81] discovered the polynomials A_2 as examples that exhibit a very small root separation (less than $2h$) for the given coefficient sizes and degree. For $p > 2$, the coefficients get longer while the bound $2h$ on root separation remains the same. In this regard, the generalization from $p = 2$ to $p \ge 2$ is not so interesting. However, here is a useful application.

We construct adverse input polynomials for the Descartes method that have a cluster of *three* close roots, enforcing a very small isolating interval for the middle one, even if subdivision is not restricted to $\alpha = 1/2$. (By contrast, two close roots could be separated in one subdivision with a lucky choice of a subdivision point.)

Theorem 3.25. *Take an integer $a \ge 2$ and an even integer $n \ge 4$. Consider the product $Q(X) = A_2(X) \cdot A_3(X)$ of polynomials from Theorem 3.23.*

(i) $Q(X)$ is a square-free integer polynomial of degree $2n$. Its longest coefficient has length $\tau := \lfloor \log \|Q\|_\infty \rfloor + 1 = \Theta(\log a)$.

(ii) An interval containing all real roots of $Q(X)$ is a superset of $(0, 1)$.

(iii) $Q(X)$ has three distinct roots in the interval $(a^{-1} - h,\, a^{-1} + h)$ with $h = a^{-n/2-1}$.

(iv) The Descartes method executed for $Q(X)$ and an initial interval $I_0 \supseteq (0, 1)$ constructs a subdivision tree $\mathcal{T}$ containing a path from the root down to a leaf of length $\Omega(n\tau)$.

Proof. The statements (i–iii) follow from the corresponding statements of Theorem 3.23:

Ad (i). The square-free polynomials A_2 and A_3 are coprime, hence their product is also square-free. The rest is obvious.

Ad (ii). Each of A_2 and A_3 has a root $x < 0$ and a root $x \ge 1$.

Ad (iii). Each of A_2 and A_3 has a root in $(a^{-1} - h,\, a^{-1})$, and the two are distinct, since A_2 and A_3 are coprime. Furthermore, A_3 has a root in $(a^{-1},\, a^{-1} + h)$. (If $a \ne 2$, the same holds for A_2.)

[4] Krandick [loc. cit.] shows (iii) also for the case $p = 2$, $a = 2$, $n \ge 7$, which we have omitted for brevity.

Ad (iv). By (iii), the interval I_0 contains three roots $\alpha < \beta < \gamma$ of Q with $\gamma - \alpha < 2h$. Let I denote the isolating interval computed for β. Clearly, $|I| < 2h$. Let $(I_0, \ldots, I_k)$ be the path in $\mathcal{T}$ from the root I_0 down to $I = I_k$. By choice of the subdivision parameter in line 11 of procedure *Descartes*, we have $|I_i| \,/\, |I_{i+1}| \leq 4$ for all $0 \leq i < k$, so that $2^{2k} \geq |I_0| \,/\, |I_k| \geq 1/(2h)$ and $k \geq (-\log h - 1)/2 = ((n/2+1)\log a - 1)/2 = \Omega(n\tau)$. □

Corollary 3.22 and Theorem 3.25 give upper and lower bounds $O(n\tau + n \log n)$ and $\Omega(n\tau)$, resp., for the size of $\mathcal{T}$. The upper bound exceeds the lower bound only by a logarithmic factor, so our upper bound is "almost tight" in the sense of orders of growth and the O-notation. The bounds coincide if we impose the side condition $\log n = O(\tau)$, which says that "the length of the degree shall be no more than a constant multiple of the maximal coefficient length". For many applications, $\log n$ is indeed much smaller than τ.

While the polynomials from Theorems 3.23 and 3.25 have sufficiently small minimum root distance to provide this asymptotic lower bound, they are not particularly strong examples of minimum root separation when one abandons the rather coarse viewpoint of examining its logarithm up to constant factors. Stronger examples have been constructed by Bugeaud/Mignotte [BM04] and Schönhage [Sch06].

It is worthwhile to remember that Theorem 3.25(iv) exhibits a subdivision tree whose size bound $O(n\tau + n \log n)$ is almost reached by the length $\Omega(n\tau)$ of a *single* rt-path. This means that no upper bound on the length of one path in a subdivision tree for a polynomial of degree n with τ-bit coefficients can be less than $O(n\tau)$. We can now see that it is essential for the quality of our upper bound on tree size (and of the complexity bounds to be derived from it) that we have followed Davenport's approach to bound the sum of all rt-path lengths at once: If we had considered separate bounds on the maximal length of one rt-path, $O(n\tau)$ or more, and on the maximal number of rt-paths, $O(n)$, then we would have arrived at a tree size bound $O(n^2\tau)$ or worse, one power of n too big.

Intuitively, Davenport's bound takes advantage of a balancing within the discriminant. In his own words, his result can be seen "as saying that there is only a certain amount of 'closeness', which can either be concentrated on one pair of roots, or spread between several pairs" [Dav85, p. 13]. In terms of rt-paths, this means that there can be one very long path, or several moderately long paths, but it is not possible that many rt-paths simultaneously realize the worst-case length possible for a single rt-path.

3.2.3 On interval boundaries and the initial interval

We will now discuss the representation of interval boundaries and the choice of an initial interval. Regarding these matters, it is instructive to review the application of the Descartes method considered by Collins and Akritas [CA76], namely isolation of all positive real roots of A_{in}. Using a bound on the positive real roots in terms of coefficients, they start from an initial interval $(0, 2^k)$, $k \in \mathbb{Z}$, that contains all positive real roots. All intervals constructed are *standard intervals*, that is to say, have the form $(a2^{k-b}, (a+1)2^{k-b})$ with $a \in \mathbb{Z}$ and $b \in \mathbb{N}_0$, where b is the number of bisections that created this interval, cf. [Joh91, Def. 31]. In our case, $a \geq 0$. Such an interval can be represented compactly as a string of b bits: Its length gives b, its value as a b-bit integer gives a (and thus implicitly $a+1$), and the succession of 0 and 1 bits encodes the left/right branches along a path in the subdivision tree $\mathcal{T}$ from the root down to the interval. Thus, we can view the integer Descartes method as a simultaneous computation of prefixes of the binary expansions of

real roots up to the point where they become distinct. This observation illustrates the difference between the Descartes method and the algorithms of Vincent and Akritas (see §3.1.2), which isolate real roots by computing prefixes of continued fraction expansions. It also shows that binary fractions are the natural form of interval boundaries in the integer Descartes method, and why it is desirable to use an upper bound on the roots that is a power of two.

Using the results of §2.4, let us now discuss how to determine from the coefficients of a polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$ an exponent $k \in \mathbb{Z}$ such that the interval $(0, 2^k)$ contains all positive real roots of $A(X)$ (with the possible exception of the boundary point $x = 2^k$). We will first present a solution as cheap as possible, which uses only the bit lengths $\lfloor \log |a_i| \rfloor + 1$, $0 \le i \le n$, of the coefficients, which are typically available from their machine representation in constant time[5], as opposed to $\lceil \log |a_i| \rceil$, which requires a loop over all bits to check for 1's other than the leading. Then we discuss how much this bound could be improved.

The basis of our discussion is the polynomial $A^{\mathrm{C+}}(X)$ from Theorem 2.55 (page 44). We leave aside the trivial case that $A^{\mathrm{C+}}(X)$ is reduced to a monomial and may thus take for granted that $A^{\mathrm{C+}}(X)$ has a unique positive real root $\mathrm{RB}^{+}_{\mathrm{Cp}}(A)$, which is the best bound on the positive roots of $A(X)$ among all bounds discussed in §2.4. Let j be the exponent of the lowest-order term of $A^{\mathrm{C+}}(X)$ with non-zero coefficient, or formally, $j = \min\{\, i \mid 0 \le i < n,\ \mathrm{sgn}(a_i) = -\mathrm{sgn}(a_n)\}$. We can replace $A^{\mathrm{C+}}(X)$ by $A^{\mathrm{C+}}(X)/X^j$ without changing its positive root $\mathrm{RB}^{+}_{\mathrm{Cp}}(A)$. We obtain an upper bound for $\mathrm{RB}^{+}_{\mathrm{Cp}}(A)$ by applying the dyadic Fujiwara root bound (2.26) to $A^{\mathrm{C+}}(X)/X^j$. Taking logarithms, we attain

$$\log(\mathrm{RB}_{\mathrm{dF}}(A^{\mathrm{C+}}(X)/X^j)) = 1 + \max\{\frac{l_i - l_n}{n-i} \mid j \le i < n,\ \mathrm{sgn}(a_i) = -\mathrm{sgn}(a_n)\}$$

where $l_i = \log |a_i|$ for $j < i \le n$, and $l_j = \log |a_j| - 1$. Consideration of $A^{\mathrm{C+}}(X)/X^j$ has allowed us to decrease l_j by 1 even if $j > 0$. We approximate this bound in terms of $\lfloor \log |a_i| \rfloor$ as

$$k := 1 + \max\{\left\lceil \frac{\lfloor l_i \rfloor + 1 - \lfloor l_n \rfloor}{n-i} \right\rceil \mid j \le i < n,\ \mathrm{sgn}(a_i) = -\mathrm{sgn}(a_n)\} \tag{3.13}$$

where $\lfloor l_i \rfloor = \lfloor \log |a_i| \rfloor$ for $j < i \le n$, and $\lfloor l_j \rfloor = \lfloor \log |a_j| \rfloor - 1$. Iterating over the coefficients $a_0, \ldots, a_{n-1}$, it is easy to determine j and then from a_j onwards compute the maximum. Under very reasonable assumptions on the relation of coefficient lengths to the maximum value of a machine word, each number encountered in (3.13) fits into one machine word, so (3.13) requires $O(n)$ operations on machine words. But how good is the resulting bound?

Lemma 3.26. *For k as above, $0 < k - \log(\mathrm{RB}_{\mathrm{dF}}(A^{\mathrm{C+}}(X)/X^j)) < 2$.*

Proof. We show that

$$0 < \left\lceil \frac{\lfloor l_i \rfloor + 1 - \lfloor l_n \rfloor}{n-i} \right\rceil - \frac{l_i - l_n}{n-i} < 2$$

[5] We regard the machine word size as constant, so this can include a logarithmic search for the leading 1 in the highest-order word.

for all i. For the lower bound, we observe

$$\left\lceil \frac{\lfloor l_i \rfloor + 1 - \lfloor l_n \rfloor}{n-i} \right\rceil - \frac{l_i - l_n}{n-i} \geq \frac{(\lfloor l_i \rfloor - l_i + 1) + (l_n - \lfloor l_n \rfloor)}{n-i} > \frac{(-1+1)+0}{n-i} = 0.$$

For the upper bound, distinguish two cases. If $n-i=1$, then there are no fractions and the outer Gauss bracket is redundant, so the expression reduces to $(\lfloor l_i \rfloor - l_i + 1) + (l_n - \lfloor l_n \rfloor) < 2$. On the other hand, if $n-i \geq 2$, then

$$\left\lceil \frac{\lfloor l_i \rfloor + 1 - \lfloor l_n \rfloor}{n-i} \right\rceil - \frac{l_i - l_n}{n-i} < 1 + \frac{(\lfloor l_i \rfloor - l_i + 1) + (l_n - \lfloor l_n \rfloor)}{n-i} < 1 + \frac{2}{n-i} \leq 2. \qquad \square$$

Lemma 3.27. *For k as above, $0 < k - \log(\mathrm{RB}^+_{\mathrm{Cp}}(A)) < 3$ and $0 \leq k - \lceil \log(\mathrm{RB}^+_{\mathrm{Cp}}(A)) \rceil \leq 2$.*

Proof. In §2.4.1, we saw $\mathrm{RB}_{\mathrm{Cp}} \leq \mathrm{RB}_{\mathrm{dF}} \leq 2 \cdot \mathrm{RB}_{\mathrm{Cp}}$ (see Proposition 2.53(ii)). From above, we recall $\mathrm{RB}_{\mathrm{Cp}}(A^{\mathrm{C}+}(X)/X^j) = \mathrm{RB}^+_{\mathrm{Cp}}(A)$. Thus, the first claim is immediate from the preceding lemma. For the second claim, we use the integrality of k to conclude that $k - \lceil \log(\mathrm{RB}^+_{\mathrm{Cp}}(A)) \rceil = \lfloor k - \log(\mathrm{RB}^+_{\mathrm{Cp}}(A)) \rfloor$. Now the second claim follows at once. $\square$

Let us summarize these findings. We have determined $k \in \mathbb{Z}$ such that all positive real roots of $A(X)$ are less than 2^k. Equality is not possible, since we overestimated $\log |a_i|$ as $\lfloor \log |a_i| \rfloor + 1$ for other reasons; but as a side effect, we can now indeed take the open interval $(0, 2^k)$ to enclose all real roots of A. Of all the bounds we saw for positive real roots, $\mathrm{RB}^+_{\mathrm{Cp}}(A)$ is the best. We have overestimated the bound $2^{\lceil \log(\mathrm{RB}^+_{\mathrm{Cp}}(A)) \rceil}$ resulting from it by a factor of at most $2^2 = 4$, which is not much compared to the maximum overestimation factors inherent in our whole approach to root bounds (cf. §2.4, esp. Theorem 2.51).

If it should nevertheless be desired to determine $\lceil \log(\mathrm{RB}^+_{\mathrm{Cp}}(A)) \rceil$ exactly, then we can take advantage of the preceding lemma, which tells us that there are only two other candidates besides k, namely $k-1$ and $k-2$. One can simply evaluate $A^{\mathrm{C}+}$ to compare the resulting bounds with the unique positive root $\mathrm{RB}^+_{\mathrm{Cp}}(A)$: If $A^{\mathrm{C}+}(2^{k-1}) > 0$, one can replace k by $k-1$, same for $k-2$. We point out that Horner's scheme for these one or two evaluations can be implemented with bit shifts instead of multiplication of long integers.

Why have we used the dyadic Fujiwara root bound? Compared to the folklore root bound, it reduces k by 1 in some boundary cases, which is not much, but the resulting implementation is just as simple. Lagrange's bound (2.31) does not translate easily to the logarithmic domain; improving k using $A^{\mathrm{C}+}(X)$ as above would be simpler, and the result is always at least as good.

Typical applications of the Descartes method require isolation of all real roots, not just the positive roots. Obviously, we can first run the integer Descartes method on $(0, +2^k)$ with k as in (3.13) for $A(X) = A_{\mathrm{in}}(X)$, and then again on $(-2^{k'}, 0)$ with k' as in (3.13) for $A(X) = A_{\mathrm{in}}(-X)$, and additionally test whether $A(0) = 0$; this is done in [CA76].

There is also a second option: choosing $r \in \mathbb{Z}$ such that $(-2^r, +2^r)$ contains all real roots of A_{in}, for example, $r = \max\{k, k'\}$. We insist that both boundaries have the same exponent, so that after the first bisection, all intervals are standard intervals. Executing the integer Descartes method with this initial interval will certainly do the job. Let us call the resulting subdivision tree $\mathcal{T}$. Suppose $(-2^{k'}, 0)$ and $(0, +2^k)$ are subintervals of $(-2^r, +2^r)$. If $\mathit{DescartesTest}(A_{\mathrm{in}}, (0, +2^k)) \geq 2$, then $(0, +2^k)$ appears as an internal node of $\mathcal{T}$, and the subtree of $\mathcal{T}$ rooted at $(0, +2^k)$ is the subdivision tree created by the integer Descartes method run on $(0, +2^k)$. The analogous statement holds for $(-2^{k'}, 0)$.

To keep the presentation simple and consistent with §3.3, we will follow the second option subsequently and consider one execution of the Descartes method starting from an initial interval $(-2^r, +2^r)$. However, we recommend to choose the first options for actual implementations. According to our argument above on corresponding subtrees of $\mathcal{T}$, we can regard this as an optimization of tree traversal that is completely orthogonal to the other implementation choices discussed below.

For later reference, we summarize the results of this section as follows.

Proposition 3.28. *Consider a polynomial $A_{\text{in}}(X) = \sum_{i=0}^{n} a_i X^i$ with integer coefficients of magnitude less than 2^τ. We can compute $r \in \mathbb{Z}$ with $|r| = O(\tau)$ such that the open interval $I_0 = (-2^r, +2^r)$ encloses all real roots of $A_{\text{in}}(X)$, with the quality guarantee $r - \lceil \log(\max\{\text{RB}^+_{\text{Cp}}(A(X), \text{RB}^+_{\text{Cp}}(A(-X))\})\rceil \leq 2$. Assuming that coefficient lengths are known and fit into a machine word, this computation needs $O(n)$ operations on machine words.*

3.2.4 The algorithm of Collins and Akritas (1976)

The original form of the Descartes method described by Collins and Akritas [CA76] represents all polynomials $A(X) = \sum_{i=0}^{n} a_i X^i$, where $n = \deg(A_{\text{in}})$, by coefficient vectors $(a_0, \ldots, a_n) \in \mathbb{Z}^{n+1}$ with respect to the power basis $1, X, \ldots, X^n$. Its operations on polynomials are usually described in terms of the following three linear transformations of coefficient vectors induced by transformations of the indeterminate:

1. The *Taylor shift* is induced by a translation of the indeterminate by $c \in \mathbb{Z}$. We will only need it for $c = \pm 1$. In the frequent case $c = 1$, we omit the subscript 1.

$$T_c(A(X)) := A(X + c) = \sum_{j=0}^{n} \Big(\sum_{i=j}^{n} \binom{i}{j} c^{i-j} a_i \Big) X^j. \tag{3.14}$$

2. *Coefficient reversal* is induced by exchanging the roles of X and its homogenizing variable (called Y before).

$$R(A(X)) := X^n A(1/X) = \sum_{i=0}^{n} a_{n-i} X^i. \tag{3.15}$$

3. The *homothetic transformation* multiplies the indeterminate with a scaling factor. We only need it for factors of the form $\sigma 2^{\pm k}$ with $\sigma \in \{-1, +1\}$ and $k \in \mathbb{N}_0$, and we define it in a way that preserves integrality of coefficients.

$$\begin{aligned} H_{\sigma 2^{+k}}(A(X)) &:= \quad A(\sigma 2^{+k} X) = \sum_{i=0}^{n} \sigma^i 2^{ik}\, a_i\, X^i, \\ H_{\sigma 2^{-k}}(A(X)) &:= 2^{nk} A(\sigma 2^{-k} X) = \sum_{i=0}^{n} \sigma^i 2^{(n-i)k}\, a_i\, X^i. \end{aligned} \tag{3.16}$$

With these transformations, we can now formulate the *power basis variant* of the Descartes method as specialization of procedure *Descartes* from page 48. We call the procedure *DescartesCA76*, because it captures the essence of the algorithm proposed by Collins and Akritas (1976). In the following pseudocode, the name of a polynomial stands for its coefficient vector w.r.t. the power basis.

```
 1: procedure DescartesCA76(A_in, (−2^r, +2^r)) // A_in ∈ ℤ[X], r ∈ ℤ
 2:     P ← ();   Q ← {};
 3:     A_0 ← H_2 T_{−1} H_{2^r}(A_in); // i.e., A_0(X) ∼ A_in(2^r(2X − 1)) = A_in(2^{r+1}X − 2^r)
 4:     v_0 ← var(TR(A_0)); // i.e., v_0 = DescartesTest(A_in, (−2^r, +2^r))
 5:     if v_0 ≥ 1 then P ← ((−2^r, +2^r)); fi;
 6:     if v_0 ≥ 2 then Q ← {((−2^r, +2^r), A_0)}; fi;
 7:     while Q ≠ {} do
 8:         // Invariant: Q = {((c, d), A) |
 9:         //       (c, d) ∈ P, DescartesTest(A_in, (c, d)) ≥ 2, A(X) ∼ A_in((d − c)X + c))};
10:         choose an element ((c, d), A) ∈ Q;
11:         m ← (c + d)/2;   // implicitly, α = 1/2
12:         I_L ← (c, m);   I_M ← [m, m];   I_R ← (m, d);
13:         A_L ← H_{1/2}(A);   A_R ← T(A_L); // i.e., A_R = T H_{1/2}(A)
14:         v_L ← var(TR(A_L)); // i.e., v_L = DescartesTest(A_in, I_L)
15:         v_M ← number of trailing zero coefficients in A_R; // i.e., vanishing order at m
16:         v_R ← var(TR(A_R)); // i.e., v_R = DescartesTest(A_in, I_R)
17:         in P, replace entry (c, d) by subsequence (I_i | i ∈ (L, M, R), v_i ≥ 1);
18:         in Q, replace element ((c, d), A) by elements {(I_i, A_i) | i ∈ {L, R}, v_i ≥ 2};
19:     od;
20:     report sequence P of isolating intervals;
21: end procedure;
```

The composite transformation $TR(\,\cdot\,)$ of a polynomial effects the Möbius transformation $X \mapsto 1/(X+1)$ of its indeterminate, mapping $(0,\infty)$ to $(0,1)$. Thus, $\mathrm{var}(TR(\,\cdot\,))$ implements the Descartes test w.r.t. interval $(0,1)$ with the method of Theorem 2.17 (page 22).

The transformations $H_{1/2}$ and R are easy to implement with a linear number of bit operations. (As R is only applied right before T, one can eliminate it completely by using a version of T that reads its input backwards.)

The Taylor shift T is the crucial operation in this algorithm. Since $T_{-1} = H_{-1}T_1H_{-1}$, it suffices to discuss $T = T_1$. It is treated extensively by Krandick [Kra95, §3.7]. Johnson, Krandick and Ruslanov [JKR05] have investigated efficient implementations that exploit characteristics of modern CPU architectures. From any of these sources or Johnson [Joh91, §2.4, §3.2], we get the following well-known result:

Theorem 3.29 (Classical Taylor shift). *Let $A(X) = \sum_{i=0}^{n} a_i X^i$ have integer coefficients of magnitude less than 2^τ. Let $T(A(X)) = \sum_{i=0}^{n} a'_i X^i$. One can compute $(a'_i)_{i=0}^n$ from $(a_i)_{i=0}^n$ using $n(n+1)/2$ additions, and all intermediate results, including the a'_i, are integers of magnitude less than $2^{n+\tau}$. This computation requires $O(n^2(n+\tau))$ bit operations.*

Combined with our results on the size of the subdivision tree and the choice of an initial interval, this leads to the following complexity statement.

Theorem 3.30. *Consider a polynomial $A_{\mathrm{in}}(X) = \sum_{i=0}^{n} a_i X^i$ with integer coefficients of magnitude less than 2^τ, all of whose real roots are simple. An interval $I_0 = (-2^r, +2^r)$, $r \in \mathbb{Z}$, as in Proposition 3.28 encloses all real roots of A_{in} and satisfies $|r| = O(\tau)$. Execution of DescartesCA76(A_{in}, I_0) using the classical Taylor shift from Theorem 3.29 isolates the real roots of $A_{\mathrm{in}}(X)$ with $O(n^5(\tau + \log n)^2)$ bit operations.*

Proof. The polynomial $A_0 = H_2T_{-1}H_{2^r}(A_{\text{in}})$ has coefficients of length $O(\tau n)$ each, because $H_{2^r}(A)$ has coefficients of length $O(\tau + |r|\, n) = O(\tau n)$, and the transformations H_2T_{-1} add only $O(n)$ further bits to each coefficient.

Subdivision replaces a polynomial A by $H(A)$ and $TH(A)$; this increases the coefficient lengths by $O(n)$ bits. By Corollary 3.22, the size and thus the height of the subdivision tree is bounded by $O(n(\tau + \log n))$. It follows that at any internal node I of the subdivision tree, the coefficients of the polynomial A constructed for it have lengths bounded by $O(n^2(\tau + \log n))$. The computational cost spent for node I is dominated by the three Taylor shifts for that node, whose cost is bounded by Theorem 3.29 as $O(n^4(\tau + \log n))$ bit operations. As there are $O(n(\tau + \log n))$ internal nodes, the total cost is $O(n^5(\tau + \log n)^2)$ bit operations. □

The complexity bound originally asserted by Collins and Akritas [CA76, Thm. 2] translates to $O(n^6(\tau + \log n)^2)$ in our notation. The present improvement by a factor n is due to our use of the Davenport-Mahler bound. Davenport [Dav85, p. 18] mentioned this relation of his bound (originally used for Sturm's method, cf. [DSY07]) to the Descartes method but did not work it out. This was undertaken by Johnson [Joh91, Thm. 53] [Joh98, Thm. 13]. A gap in Johnson's proof was filled by Krandick [Kra95, §3.10]; this revised argument was improved further by Krandick and Mehlhorn [KM06]. However, their arguments consider the subdivision tree level by level in a somewhat involved manner. Our proof above is conceptually simpler: it just multiplies tree size with cost per node.

This simplicity of our proof makes it obvious how to replace the complexity bound for the classical Taylor shift by 1 from Theorem 3.29 with that of an asymptotically fast algorithm. Gerhard [Ger04, §4.1] (also in [vzGG97, §2]) discusses and compares three such techniques and proposes a fourth. As usual, we write $f(n) = O^{\sim}(g(n))$ for positive functions f and g if there exists $k \in \mathbb{N}_0$ such that $f(n) = O(g(n) \log^k(3 + g(n)))$. (Adding 3 makes sure the logarithm is larger than 1.)

Theorem 3.31 (Asymptotically fast Taylor shift). *Let $A(X) = \sum_{i=0}^{n} a_i X^i$ have integer coefficients of magnitude less than 2^τ. Let $T(A(X)) = \sum_{i=0}^{n} a'_i X^i$. One can compute $(a'_i)_{i=0}^n$ from $(a_i)_{i=0}^n$ with $O^{\sim}(n \cdot (n + \tau))$ bit operations.*

This bound holds for the divide&conquer method of von zur Gathen [vzG90, Fact 2.1(iv)] (for a proof, see there or case E in [Ger04, Thm. 4.5]), for the convolution method going back to Aho et al. [ASU75] (for a proof, see case F in [Ger04, Thm. 4.5] [vzGG97, Thm. 2.4]), and for the modular method of Gerhard [Ger04, Thm. 4.8].

Up to the logarithmic factors suppressed by the $O^{\sim}$-notation, this complexity matches the output size bound from Theorem 3.29 ($n + 1$ coefficients of up to $n + \tau$ bits).

Theorem 3.32. *Consider a polynomial $A_{\text{in}}(X) = \sum_{i=0}^{n} a_i X^i$ with integer coefficients of magnitude less than 2^τ, all of whose real roots are simple. An interval $I_0 = (-2^r, +2^r)$, $r \in \mathbb{Z}$, as in Proposition 3.28 encloses all real roots of A_{in} and satisfies $|r| = O(\tau)$. Execution of DescartesCA76(A_{in}, I_0) using an asymptotically fast Taylor shift such as in Theorem 3.31 isolates the real roots of $A_{\text{in}}(X)$ with $O^{\sim}(n^4\tau^2)$ bit operations.*

Proof. Take the proof of Theorem 3.30 and replace the complexity bound from Theorem 3.29 with that from Theorem 3.31. □

This complexity bound $O^{\sim}(n^4\tau^2)$ is the same as the best known bound for asymptotically efficient forms of Sturm's method on square-free polynomials [DSY07, Cor. 8]. However, the practical use of this theoretical improvement appears to be limited, see §3.2.7.

The roles of the three Taylor shifts in *DescartesCA76* are different: the result of the Taylor shift in line 13 is essential, as it implements subdivision; the results of the two Taylor shifts in lines 14 and 16 are only needed temporarily to compute the numbers v_L and v_R of sign variations. Krandick [Kra95, §3.7.1] pointed out how the algorithm can be accelerated in practice by performing these Taylor shifts only as far as necessary to decide whether there are zero, one, or at least two sign variations. Also, if all coefficients of A_R have the same sign, it is clear that $\mathrm{var}(TR(A_L)) = 0$, so we need not carry out this transformation [Kra95, §3.7.4]. Johnson et al. [JKL$^+$06, §2.4] describe further optimizations of this kind found in the implementation of the Descartes method for QI[6] by G. Hanrot, F. Rouillier, P. Zimmermann, and S. Petitjean.

Using the variation-diminishing property of subdivision, we arrive at two further shortcuts of this kind.

1. Without costly transformations, we can determine $w := \mathrm{var}(A_L) - \mathrm{var}(A_R)$, the Budan-Fourier bound on the number of roots of A_L in $(0, 1]$, see Appendix A.1. Corollary A.4 tells us that $w - v_M \geq v_L$, with an even difference between both sides. Thus, if $w - v_M$ is 0 or 1, it is equal to v_L, and we do not have to compute the transformation $TR(A_L)$.
2. Suppose we have knowledge of $v = \mathit{DescartesTest}(A_{\mathrm{in}}, (c, d))$; for example, because we store it as an additional component in the entry for (c, d) in Q. Proposition 2.26 tells us that $v - v_M \geq v_L + v_R$, with an even difference between both sides. Suppose further we just compute one of v_L, v_R. If it is equal to the upper bound $v - v_M$, we know the other summand is zero. If it is equal to $v - v_M - 1$, the other summand is one. According to our deliberations in §3.1.3.III, subdivision almost always results in $(v_L, v_M, v_R) \in \{(v, 0, 0), (0, 0, v)\}$, with at most $n-1$ exceptions. This justifies the following heuristic: Compute one of v_L, v_R chosen at random. With a probability of almost $1/2$, it is equal to the upper bound $v - v_M$, and we can save the third Taylor shift. In the rare case that we miss the upper bound by just 1, we may conclude that the other number of sign variations is exactly 1 and also save the third Taylor shift.

3.2.5 The algorithm of Lane and Riesenfeld (1981)

The algorithm *DescartesCA76* described above requires costly transformations of A_L and A_R to obtain coefficient sequences in which sign variations can be counted to implement the Descartes test for roots in $(0, 1)$. As discussed in §2.2.4, this can be avoided by replacing the power basis $1, X, \ldots, X^n$ with the $[0, 1]$-Bernstein basis $B_0^n(X), \ldots, B_n^n(X)$: the Descartes test is then simply the number of sign variations in the Bernstein coefficient sequence.

The input polynomial $A_{\mathrm{in}}(X)$ is given by coefficients w.r.t. the power basis. In the pseudocode below, A_{in} denotes this coefficient vector. As in *DescartesCA76*, we transform the initial interval $(-2^r, +2^r)$ to $(0, 1)$ by computing $A_{\mathrm{pow0}}(X) \leftarrow H_2 T_{-1} H_{2^r}(A_{\mathrm{in}}(X))$; we let A_{pow0} denote this polynomial's coefficient vector w.r.t. the power basis. We then have to convert $A_{\mathrm{pow0}}(X) = \sum_i a_i X^i$ to the Bernstein basis representation $A_0(X) = \sum_i b_i B_i^n(X)$. In fact, A_{pow0} and A_0 are equal as polynomials. but in pseudocode, we let the different symbol A_0 denote the vector of Bernstein coefficients. Proposition 2.28(i) shows us that we

[6] QI: Quadric Intersection. `http://www.loria.fr/equipes/vegas/qi/`

can obtain $(\binom{n}{i} b_i)_i$ in reversed order as power basis coefficients of $TR(A_0)$. This reduces the conversion into Bernstein basis to transformations that we already know, except that we still need to scale the coefficients by inverses of the binomial coefficients. To achieve that without producing fractions, we introduce one more linear transformation of a polynomial $A(X) = \sum_{i=0}^{n} a_i X^i$; this one is not induced by a transformation of the indeterminate.

$$\beta(A(X)) := n! \sum_{i=0}^{n} \binom{n}{i}^{-1} a_i \, X^i = \sum_{i=0}^{n} i! \, (n-i)! \, a_i \, X^i. \tag{3.17}$$

We thus arrive at the *Bernstein basis variant* of the Descartes method described in the following pseudocode. Except A_{in} and A_{pow0}, all polynomials are represented in the $[0,1]$-Bernstein basis, and their names denote the respective vectors of Bernstein coefficients. Correspondingly, $\text{var}(\,\cdot\,)$ denotes the number of sign variations in the Bernstein coefficients; by Theorem 2.22 (page 24), this is the Descartes test for interval $(0,1)$.

1: **procedure** *DescartesLR81*$(A_{\text{in}}, (-2^r, +2^r))$ // $A_{\text{in}} \in \mathbb{Z}[X]$, $r \in \mathbb{Z}$
2: $P \leftarrow ()$; $Q \leftarrow \{\}$;
3: $A_{\text{pow0}} \leftarrow H_2 T_{-1} H_{2^r}(A_{\text{in}})$; $A_0 \leftarrow \beta RTR(A_{\text{pow0}})$; // Bernstein conversion
4: $v_0 \leftarrow \text{var}(A_0)$; // i.e., $v_0 = \mathit{DescartesTest}(A_{\text{in}}, (-2^r, +2^r))$
5: **if** $v_0 \geq 1$ **then** $P \leftarrow ((-2^r, +2^r))$; **fi**;
6: **if** $v_0 \geq 2$ **then** $Q \leftarrow \{((-2^r, +2^r), A_0)\}$; **fi**;
7: **while** $Q \neq \{\}$ **do**
8: // **Invariant:** $Q = \{((c,d), A) \mid$
9: // $(c,d) \in P,\ \mathit{DescartesTest}(A_{\text{in}}, (c,d)) \geq 2,\ A(X) \sim A_{\text{in}}((d-c)X + c))\}$;
10: choose an element $((c,d), A) \in Q$;
11: $m \leftarrow (c+d)/2$; // implicitly, $\alpha = 1/2$
12: $I_L \leftarrow (c, m)$; $I_M \leftarrow [m, m]$; $I_R \leftarrow (m, d)$;
13: $(A_L, A_R) \leftarrow \mathit{deCasteljau}(2^{\deg A_{\text{in}}} A,\ 1/2)$;
14: $v_L \leftarrow \text{var}(A_L)$; // i.e., $v_L = \mathit{DescartesTest}(A_{\text{in}}, I_L)$
15: $v_M \leftarrow$ number of leading zero coefficients in A_R; // i.e., vanishing order at m
16: $v_R \leftarrow \text{var}(A_R)$; // i.e., $v_R = \mathit{DescartesTest}(A_{\text{in}}, I_R)$
17: in P, replace entry (c,d) by subsequence $(I_i \mid i \in (L, M, R), v_i \geq 1)$;
18: in Q, replace element $((c,d), A)$ by elements $\{(I_i, A_i) \mid i \in \{L, R\}, v_i \geq 2\}$;
19: **od**;
20: report sequence P of isolating intervals;
21: **end procedure**;

In line 13, we apply de Casteljau's algorithm to $2^n A$ so that all intermediate results, including the coefficients of A_L and A_R, remain integral. (Of course, an implementation will simply compute $b_{j,i} = b_{j-1,i} + b_{j-1,i+1}$ and perform bit shifts to post-multiply the output by the missing powers of 2, that is, $b'_j := 2^{n-j} b_{j,0}$ and $b''_i := 2^i b_{n-i,i}$.) Proposition 2.25(iii) (page 26) shows that this does indeed implement the polynomial transformations prescribed in line 13 of the generic form of the Descartes method (page 48).

Looking at de Casteljau's algorithm in this way – the input is transformed into two new polynomials, all in the same basis – matches the interpretation of the transformations in *DescartesCA76* and highlights the equivalence of the power and Bernstein basis variants. Alternatively, one can think of *DescartesLR81* as keeping the input polynomial fixed

(up to repeated multiplication by 2^n), but transforming it into many different Bernstein bases, one for each interval. This is the viewpoint of Proposition 2.25(ii); it matches the intuition of the polynomial's graph as a Bézier curve that is approximated increasingly well by repeated subdivision of its control polygon.

The original source of *DescartesLR81* is an article by Lane and Riesenfeld [LR81], which describes an algorithm for isolating the real roots (and the maxima and minima) of a polynomial by recursive bisection, using the sign variations in the Bernstein coefficients to test for roots and de Casteljau's algorithm for subdivision. The authors explain how "Descartes' Rule of Signs 'carries over' to the Bernstein form of the polynomial" [op. cit., p. 113] and point out the similarity of their algorithm to that of Collins and Akritas [CA76]. The fact that Lane and Riesenfeld implemented their algorithm with floating-point arithmetic instead of exact integer arithmetic does not matter at this level of abstraction.

It appears that the contribution of Lane and Riesenfeld and the link between the Descartes method and the Bernstein basis went unnoticed for a long time in the symbolic computation community. The article [MVY02, §2.1], the first edition (2003) of the book [BPR06, §10.2] and the survey [MRR05] describe essentially *DescartesLR81* and have drawn attention to the Descartes method implemented in the Bernstein basis, but they do not point out clearly the original contributions of Collins/Akritas and Lane/Riesenfeld.

Let us now analyze the bit complexity of *DescartesLR81*.

Theorem 3.33. *Consider a polynomial $A_{\text{in}}(X) = \sum_{i=0}^{n} a_i X^i$ with integer coefficients of magnitude less than 2^τ, all of whose real roots are simple. An interval $I_0 = (-2^r, +2^r)$, $r \in \mathbb{Z}$, as in Proposition 3.28 encloses all real roots of A_{in} and satisfies $|r| = O(\tau)$. Execution of DescartesLR81(A_{in}, I_0) isolates the real roots of $A_{\text{in}}(X)$ with $O(n^5(\tau+\log n)^2)$ bit operations.*

Proof. The polynomial A_{pow0} has power basis coefficients of length $O(n\tau)$. During the composite transformation βRTR, the coefficient length grows by $O(n)$ due to T and by $O(n \log n)$ due to β. Thus, A_0 has Bernstein coefficients of length $O(n \cdot (\tau + \log n))$.

De Casteljau's algorithm, invoked in line 13, increases the length of coefficients by at most n bits. By Corollary 3.22, the size and thus the height of the subdivision tree is bounded by $O(n \cdot (\tau + \log n))$. It follows that at any internal node I of the subdivision tree, the Bernstein coefficients of the polynomial A constructed for it have lengths bounded by $O(n^2(\tau + \log n))$. The computational effort for node I is dominated by de Casteljau's algorithm. It performs $O(n^2)$ integer additions and bit shifts; together, this requires $O(n^4(\tau + \log n))$ bit operations. As there are $O(n(\tau + \log n))$ internal nodes, the total cost is $O(n^5(\tau + \log n)^2)$ bit operations. □

This theorem improves upon the bound stated in [MRR05] and the first edition (2003) of [BPR06] by removing a factor of n through the use of the Davenport-Mahler bound.

Can we improve the bound further by replacing de Casteljau's algorithm and its $O(n^2)$ arithmetic operations with an asymptotically faster method of subdivision? Proposition 2.28(ii/iii) shows how one execution of de Casteljau's algorithm that maps $(b_i)_i$ to $((b'_i)_i, (b''_i)_i)$ can be exchanged for two asymptotically fast Taylor shifts (see Theorem 3.31) plus several linear-time scaling and reversal transformations. In particular, the first step is to multiply Bernstein coefficients with binomial coefficients to obtain $(\binom{n}{i} b_i)_i$, and the last step is to divide binomial coefficients out of the result $((\binom{n}{i} b'_i)_i, (\binom{n}{i} b''_i)_i)$. Emiris et al. [EMT06] have given an improved bound of $O^\sim(n^4\tau^2)$ for *DescartesLR81* analo-

gously to the improvement for *DescartesCA76* leading to Theorem 3.32. However, the necessity to put in and take out binomial coefficients shows that fast subdivision using Proposition 2.28 and Theorem 3.31 and thus this $O^{\sim}(n^4\tau^2)$ complexity bound are more naturally associated to a different basis. That is the subject of the next section.

3.2.6 The "dual" algorithm of Johnson (1991)

Johnson [Joh91, §4.2.2] formulated a variant of *DescartesCA76* that saves one third of all Taylor shifts by replacing the polynomial $A(X) \sim A_{\text{in}}((d-c)X+c)$ with $TR(A(X))$ and computing $TR(A_L(X))$ and $TR(A_R(X))$ directly from $TR(A(X))$. To distinguish it from the "primal" power basis Descartes method that performs subdivision in the initial affine chart of $\widehat{\mathbb{R}}$, he called it the *dual algorithm* or, in SACLIB parlance, IPRICSD.

The original description of Johnson's dual algorithm does not fit our general form of the Descartes method from §3.1.1, which normalizes the interval of interest to $(0,1)$ but leaves the basis chosen for representing polynomials unspecified. Using Proposition 2.28, we can translate Johnson's dual algorithm to our setting. The resulting coefficient vectors and operations on them are exactly the same, we just describe them with respect to a transformed indeterminate $T = 1/(X+1)$; this replaces the power basis $(X^i)_{i=0}^n$ used by Johnson with the basis $(T^{n-i}(1-T)^i)_{i=0}^n$. Since $\binom{n}{i}T^{n-i}(1-T)^i = B_{n-i}^n(T)$, this is the $[0,1]$-Bernstein basis up to reversed order and multiplication by binomial coefficients. This method has therefore been called the *scaled Bernstein basis variant* of the Descartes method. For this basis, de Casteljau's algorithm is replaced by the method from Proposition 2.28(ii/iii).

We represent A_{in} and A_{pow0} in the power basis and all other polynomials in the basis $(T^{n-i}(1-T)^i)_{i=0}^n$. In pseudocode, the names of polynomials denote their respective coefficient vectors.

```
 1: procedure DescartesJ91d(A_in, (−2^r, +2^r)) // A_in ∈ ℤ[T], r ∈ ℤ
 2:     P ← ();   Q ← {};
 3:     A_pow0 ← H_2 T_{−1} H_{2^r}(A_in);   A_0 ← TR(A_pow0); // basis conversion
 4:     v_0 ← var(A_0); // i.e., v_0 = DescartesTest(A_in, (−2^r, +2^r))
 5:     if v_0 ≥ 1 then P ← ((−2^r, +2^r)); fi;
 6:     if v_0 ≥ 2 then Q ← {((−2^r, +2^r), A_0)}; fi;
 7:     while Q ≠ {} do
 8:         // Invariant: Q = {((c,d), A) |
 9:         //       (c,d) ∈ P, DescartesTest(A_in, (c,d)) ≥ 2, A(X) ~ A_in((d−c)X+c))};
10:         choose an element ((c,d), A) ∈ Q;
11:         m ← (c+d)/2;   // implicitly, α = 1/2
12:         I_L ← (c,m);   I_M ← [m,m];   I_R ← (m,d);
13:         A_L ← H_2 T(A);   A_R ← R H_2 TR(A);
14:         v_L ← var(A_L); // i.e., v_L = DescartesTest(A_in, I_L)
15:         v_M ← number of leading zero coefficients in A_R; // i.e., vanishing order at m
16:         v_R ← var(A_R); // i.e., v_R = DescartesTest(A_in, I_R)
17:         in P, replace entry (c,d) by subsequence (I_i | i ∈ (L,M,R), v_i ≥ 1);
18:         in Q, replace element ((c,d), A) by elements {(I_i, A_i) | i ∈ {L,R}, v_i ≥ 2};
19:     od;
20:     report sequence P of isolating intervals;
21: end procedure;
```

Theorem 3.34. *Consider a polynomial $A_{\text{in}}(X) = \sum_{i=0}^{n} a_i X^i$ with integer coefficients of magnitude less than 2^τ, all of whose real roots are simple. An interval $I_0 = (-2^r, +2^r)$, $r \in \mathbb{Z}$, as in Proposition 3.28 encloses all real roots of A_{in} and satisfies $|r| = O(\tau)$. Execution of DescartesJ91d(A_{in}, I_0) isolates the real roots of $A_{\text{in}}(X)$ with a number of bit operations bounded by $O(n^5(\tau + \log n)^2)$, if the classical Taylor shift from Theorem 3.29 is used, or bounded by $O^\sim(n^4\tau^2)$, if a Taylor shift as in Theorem 3.31 is used.*

Proof. Similar to the proofs for in the preceding sections, we find that all transformed polynomials A have coefficient lengths bounded by $O(n^2(\tau + \log n))$. Substituting this into Theorem 3.29 or 3.31, resp., yields the cost per subdivision, and multiplying with the size $O(n \cdot (\tau + \log n))$ of the subdivision tree yields the claimed complexity bounds. □

3.2.7 Comparison of the exact integer algorithms

We have now seen three algorithms implementing the Descartes method for polynomials with integer coefficients. They were obtained from the general procedure *Descartes* (§3.1.1) by choosing a basis for representing the polynomials and supplying the matching sub-algorithms for subdivision and evaluating the Descartes test. The power basis $(X^i)_{i=0}^n$ chosen for *DescartesCA76* has the advantage that inputs are typically provided in it, so that no conversion is necessary. The basis $(X^{n-i}(1-X)^i)_{i=0}^n$ implicitly used by Johnson allows to perform a Descartes test w.r.t. $(0, 1)$ immediately. For both of these bases, the Taylor shift is the fundamental operation. By contrast, for the Bernstein basis $(\binom{n}{i} X^i(1-X)^{n-i})_{i=0}^n$, the fundamental operation is de Casteljau's algorithm. Like one classical Taylor shift, it needs $(n+1)n/2$ additions, but it provides *two* useful results at once. As discussed in Appendix A.1, there is an analogue of de Casteljau's algorithm that implements subdivision of $(0, \infty)$ at $m \in (0, \infty)$, and it belongs to the basis $(\binom{n}{i} X^i)_{i=0}^n$. However, this particular pattern of subdivision fits the Continued Fractions method (see §3.1.2) better than the Descartes method.

We are now faced with a choice between three bases and, for two of them, between a classical and an asymptotically fast Taylor shift. In terms of practical performance, these choices have been studied by Johnson [Joh91, §4.5] and Johnson et al. [JKL+06].

Comparing the best implementation of the classical Taylor shift and the best implementation of an asymptotically fast Taylor shift at the authors' avail, [JKL+06, Fig. 3] reports that the classical Taylor shift is clearly superior up to degree 1000 (speed-up around 5 or more; much more for small degrees) and remains competitive at least up to degree 10 000. These degrees are far beyond our geometric applications, so the asymptotically fast variants are of no interest to us, and we only discuss the classical implementations of Taylor shift and de Casteljau subdivision.

Johnson [Joh91, §4.5] compares *DescartesCA76* (called IPRICS in SACLIB parlance) and *DescartesJ91d* (called IPRICSD) on polynomials of degree up to 100 with coefficients chosen at random [loc. cit., Tbl. 13] and on polynomials of degree 20 with real and complex roots chosen at random [loc. cit., Tbl. 19]. *DescartesCA76* is consistently faster, albeit only by small factors (less than 1.5), presumably because the increased length of coefficients outweighs the elision of the third Taylor shift.

Johnson et al. [JKL+06] compare, inter alia, the implementations of *DescartesCA76* and *DescartesLR81* from SACLIB for four classes of polynomials with degrees of several hundreds and on several modern CPU architectures. *DescartesLR81* is consistently faster; typical speed-up factors are in the range of 2 to 4. In particular, the reduction of three

Taylor shifts to one de Casteljau subdivision apparently more than compensates the increase in coefficient length caused by the fraction-free conversion to Bernstein basis. Since both implementations share the same infrastructure, i.e., SACLIB, we take this – with all due caution – as an indication of the inherent strength of the Bernstein basis variant. We conclude that *DescartesLR81* is a good way to implement the Descartes method.

Johnson et al. [JKL+06] proceed to compare the SACLIB methods further to several advanced implementations of the Descartes method. Under the influence of different CPU architectures and different implementations of integer addition, the picture becomes less clear, but that is beyond the scope of this thesis.

3.3 The Descartes method for bitstream coefficients

3.3.1 Introduction

The preceding discussion of the Descartes method has taken for granted that the coefficients of the input polynomial $A_{\text{in}}(X)$ come from a subring R of $\mathbb{R}$ whose elements can be represented in a way that allows us to carry out arithmetic operations and to determine the sign of any element (or equivalently, to decide equality and inequality relations between elements). We will speak of the *exact Descartes method* in the sequel to highlight this assumption where necessary. Such exact arithmetic is possible and well-understood for algebraic numbers (e.g., [Loo83] [Coh93, §4.2] [Yap04a]), but it becomes a major problem even for elementary classes of transcendental numbers (e.g., [Ric97] [Yap04a]). Moreover, where possible at all, exact arithmetic tends to be expensive. Previous work [JK97] [CJK02] [RZ04] reports significant practical accelerations of the Descartes method achieved by using approximate arithmetic instead of exact arithmetic with long integers or especially with algebraic numbers. However, these previous approaches have to fall back to exact arithmetic for certain problematic inputs;[7] Collins et al. [CJK02, p. 152] give an explicit example. As we shall see, the boundary cases that necessitate exact arithmetic for such problematic inputs are artifacts of the discrete grid imposed by recursive bisection, they are not inherent in the problem to be solved: Since we have made a restriction to polynomials A_{in} with *simple* real roots, a set of isolating open intervals for the real roots of A_{in} is a set of isolating intervals for all polynomials in any sufficiently small neighbourhood of A_{in} in $\mathbb{R}[X]$, because complex roots vary continuously with the coefficients [RS02, §1.3] and imaginary roots occur only in complex-conjugate pairs.

A particularly general interface to the coefficients $a_0, \ldots, a_n$ of $A_{\text{in}}(X) = \sum_{i=0}^{n} a_i X^i$ that implements this notion of "sufficiently small neighbourhood", with "sufficiently small" to be quantified by the algorithm itself, is the following. It matches the common practice, which we follow, of approximating real numbers by (finite) binary fractions $m2^{-p}$ with a significand $m \in \mathbb{Z}$ and an exponent $p \in \mathbb{Z}$.

[7] Rouillier and Zimmermann [RZ04] consider the setting with integer coefficients approximated by intervals whose boundaries are binary numbers of a certain precision. They remark that "when the precision grows, at some point [any transformed polynomial] will be represented in an exact way" [op. cit., p. 45], namely when intervals collapse to single points. This also constitutes a fall-back to exact arithmetic, because the two identical interval boundaries *are* exact representations of the numbers in question. This implicit way of falling back to exact arithmetic is a side effect of the restriction to integer coefficients; it cannot occur for coefficients that do not possess a representation as binary fractions of finite length.

Definition 3.35. A *bitstream* representing a real number r is a procedure that accepts an integral precision parameter p and returns an integer m such that $|m - r2^p| \leq 1$. In pseudocode, we write $[r2^p]$ for m (Gauss brackets with unspecified rounding direction).

We can regard a bitstream as providing approximations $\widetilde{r} = m2^{-p}$ subject to the error bound $|\widetilde{r} - r| \leq 2^{-p}$. For any real number r, there exist sequences $(m_p)_{p=0}^{\infty}$ of integers such that $\forall p\colon\ |m_p 2^{-p} - r| \leq 2^{-p}$, so this interface per se can accommodate all real numbers, even though only countably many of them possess a procedure to *compute* such a sequence. We refer to [Yap04a] for general investigations and more background on this and related models for computing with real numbers.

We have chosen the name "bitstream", because it concisely captures the basic intuition that we can let more and more bits flow into the significand of $\widetilde{r}$. However, a bitstream does not necessarily behave like a fixed sequence of bits that is read incrementally: by its definition, it is perfectly valid for a bitstream representing the number 1 to provide the binary approximations 1.0_2, 1.00_2, 0.11_2, 1.01_2, 0.111_2, resp., in successive invocations for $p = 1, 2, 2, 2, 3$.

We have not required a strict inequality in the approximation guarantee of a bitstream, because error bounds obtained from interval arithmetic naturally come with boundaries included.[8] Within the error bound $|m - r2^p| \leq 1$, there are always two possible values of m approximating $r2^p$ (three in the special case that $r2^p$ is itself an integer). It might seem tempting to insist on the tighter error bound $|m - r2^p| \leq 1/2$, on the grounds that rounding $r2^p$ to the nearest integer could achieve that; however, the ensuing discontinuity at the midpoint between two successive integers would preclude approximate computation within the procedure providing the bitstream.

Let us now start to discuss how we can apply the Descartes method to a polynomial A_{in} with bitstream coefficients. Our goal is an algorithm that operates essentially like the exact Descartes method, even though it only knows approximate values for the coefficients of A_{in} and its transformations. In §3.2, we have met three choices of bases to represent polynomials and the sub-algorithms arising from them for the necessary transformations of A_{in}. If we choose the Bernstein basis, the only transformation needed in the main loop of the Descartes method is de Casteljau's algorithm. It is particularly well-suited for approximate computation, because it consists entirely of convex combinations, which are numerically very stable. For this reason, and because the Bernstein basis has already demonstrated a good practical performance for exact integer coefficients (see §3.2.7), we will use the Bernstein basis representation for all polynomials, except the input. We discuss in §3.3.2 how to convert the input polynomial from the power basis to an approximate Bernstein basis representation w.r.t. a suitable initial interval, and in §3.3.3 how to execute de Casteljau's algorithm for approximate coefficients. We use approximate Bernstein representations of the form $(b_0 2^{-q}, \ldots, b_n 2^{-q})$ with significands $b_i \in \mathbb{Z}$ and a common exponent $q \in \mathbb{Z}$, and we perform fixed-point arithmetic on them. Our algorithm maintains a global precision parameter $p \in \mathbb{Z}$ and chooses q large enough such that all Bernstein coefficients that occur have an absolute error bounded by $\varepsilon = 2^{-p}$. One can think of p as the "payload precision" that indicates how good the available coefficient approximations are, and of $q > p$ as the "working precision" necessary to achieve that in the presence of accumulating arithmetical error.

[8] The image of a compact interval under a continuous function is a compact interval; the image of a bounded open interval may be neither open nor bounded.

While the good numerical properties of the Bernstein basis will allow us to keep q close to p, a fundamental problem persists. The Descartes method needs to make decisions based on the signs of coefficients, namely to check whether the subdivision point is itself a root and to count the number of sign variations in the coefficient sequences arising from subdivision (cf. lines 14–16 of procedure *DescartesLR81* on page 71). However, an approximation $\widetilde{a}$ of a real number a satisfying $|\widetilde{a} - a| \leq \varepsilon$ for some error bound $\varepsilon > 0$ does not uniquely determine $\operatorname{sgn}(a)$ unless $|\widetilde{a}| > \varepsilon$. Using $\widetilde{a}$, we can merely determine the *ε-approximate sign* of a:

$$\operatorname{sgn}_\varepsilon(\widetilde{a}) = \begin{cases} + & \text{if } \widetilde{a} > +\varepsilon, \\ - & \text{if } \widetilde{a} < -\varepsilon, \\ ? & \text{if } |\widetilde{a}| \leq \varepsilon. \end{cases} \tag{3.18}$$

If $|\widetilde{a}| > \varepsilon$, it holds that $\operatorname{sgn}_\varepsilon(\widetilde{a}) = \operatorname{sgn}(a)$, and we say that $\widetilde{a}$ is *(sign-)determinate*; otherwise, we say that $\widetilde{a}$ is *(sign-)indeterminate*.

If a is very close to zero, determining its sign from an approximation requires a very good (and thus computationally expensive) approximation. If a is in fact equal to 0, it is impossible to determine its sign from an ε-approximation alone, no matter how small the positive number ε is. Therefore, the crucial issue in adapting the Descartes method to polynomials with bitstream coefficients is to avoid the necessity of sign determination for coefficients that are "too small" in magnitude.

It is relatively obvious how to do this for the leftmost Bernstein coefficient $A_R(0)$ of A_R, inspected to determine whether the subdivision point is a root of A_{in}: the subdivision point m has to be chosen sufficiently far away from any complex root of A_{in}, then $|A_R(0)| = |A_{\text{in}}(m)|$ is large, because a polynomial can only be small in magnitude close to one of its roots. (A precise form of this statement is given as Theorem 3.53 on page 94.)

It is less obvious how to make sure that the ε-approximate signs of the Bernstein coefficients inspected in a Descartes test allow to distinguish between 0, 1, or more than 1 sign variation. In §3.3.4, we will meet two crucial lemmas that give a sufficient condition in terms of $|A_{\text{in}}(m)|$ for the choice of subdivision points m to guarantee that we can eventually make this distinction, but possibly at the price of an additional subdivision step in case of uncertainty.

Our algorithm, which we will outline in §3.3.5, is designed around these lemmas. The algorithm (as opposed to its analysis) is straightforward and uses exponential guessing to determine a sufficient precision such that good subdivision points can be found quickly by randomization. With good subdivision points, the algorithm can then mimic the exact Descartes method and produce a set of isolating open intervals for the real roots of A_{in}.

After resolving in §3.3.6 a technicality concerning the accumulation of arithmetical error, we are ready to fully specify our algorithm in §3.3.7. In §3.3.8, we estimate which value of the precision parameter p is sufficient for a given input A_{in}, and based on this, we analyze the algorithm's bit complexity in §3.3.9. We conclude with a description of some possible variants of the algorithm in §3.3.10 and a discussion of our results in §3.3.11.

3.3.2 On the initial interval and conversion to Bernstein basis

The bitstream Descartes algorithm receives as input a real polynomial

$$A_{\text{in}}(X) = \sum_{i=0}^{n} a_i X^i, \qquad n = \deg(A_{\text{in}}) \geq 2, \tag{3.19}$$

given by bitstreams representing the coefficients $a_0, \ldots, a_n$. The input is subject to the condition that all real roots of A_{in} shall be simple. This has two immediate consequences: A_{in} has more than one complex root (because a unique complex root would be both real and multiple), and A_{in} has a non-zero coefficient besides a_n (since $a_0 = a_1 = 0$ would make $x = 0$ a double real root).

The bitstream Descartes algorithm needs an initial interval

$$I_0 := (c_0, d_0) := (-2^{r+1}, +2^{r+1}), \qquad r \in \mathbb{Z}, \tag{3.20}$$

such that

$$A_{\text{in}}(\zeta) = 0 \implies |\zeta| \leq 2^r \qquad \text{for any} \quad \zeta \in \mathbb{C}. \tag{3.21}$$

Notice that the circumcircle of I_0 includes the imaginary roots as well and has a radius overestimating the complex root magnitudes by a factor of two; the motivation for this will become apparent in the sequel.

A straightforward way to determine r is to evaluate the dyadic Fujiwara complex root bound (2.26) from page 43 for approximations of $\log|a_i|$. The trick from §3.2.3 about locating the lowest non-zero coefficient is, of course, inapplicable in the bitstream setting. In addition, the quantities $\lfloor \log|a_i| \rfloor$ or $\lceil \log|a_i| \rceil$ are no longer accessible: to avoid discontinuities near integral logarithms, we need to allow an error margin. Specifically, we require for the leading coefficient a_n of A_{in} a lower bound

$$l_n^- \in \mathbb{Z} \quad \text{such that} \quad l_n^- \leq \log|a_n| < l_n^- + 2. \tag{3.22}$$

For the moment, let us assume that we have also have upper bounds

$$l_0^+, \ldots, l_{n-1}^+ \in \mathbb{Z} \cup \{-\infty\} \quad \text{such that} \quad l_i^+ \geq \log|a_i| > l_i^+ - 2 \quad \text{for } 0 \leq i < n. \tag{3.23}$$

We set

$$r := 1 + \max\Big\{ \left\lceil \frac{l_{n-1}^+ - l_n^-}{1} \right\rceil, \left\lceil \frac{l_{n-2}^+ - l_n^-}{2} \right\rceil, \ldots, \left\lceil \frac{l_1^+ - l_n^-}{n-1} \right\rceil, \left\lceil \frac{l_0^+ - l_n^- - 1}{n} \right\rceil \Big\}. \tag{3.24}$$

The approximation error we make here is bounded as follows.

Lemma 3.36. *For r as in (3.24), it holds that $0 \leq r - \lceil \log \mathrm{RB}_{\mathrm{dF}}(A_{\text{in}}) \rceil \leq 3$.*

Proof. Let us write $l_i = \log|a_i|$. Analogous to the proof of Lemma 3.26, we find that

$$0 \leq \left\lceil \frac{l_i^+ - l_n^-}{n-i} \right\rceil - \frac{l_i - l_n}{n-i} < 4, \qquad 0 \leq \left\lceil \frac{l_0^+ - l_n^- - 1}{n} \right\rceil - \frac{l_0 - l_n - 1}{n} < 4,$$

so that $0 \leq r - \log \mathrm{RB}_{\mathrm{dF}}(A_{\text{in}}) < 4$ and $0 \leq \lfloor r - \log \mathrm{RB}_{\mathrm{dF}}(A_{\text{in}}) \rfloor \leq 3$. □

However, there is a problem: One should not – and in some cases, cannot – evaluate (3.24) by first computing all upper bounds l_i^+ and the values resulting from them and then picking the maximum. The reason is that the magnitudes $|a_i|$, $0 \le i < n$, may vary a lot, and approximating their logarithms in accordance to (3.23) may be quite expensive for those which are small. In the extreme case $a_i = 0$, meaning $\log|a_i| = -\infty$, this is not even possible in the bitstream setting, since no approximation of a_i is good enough to determine $\log|a_i| = -\infty$.

To overcome this, we postulate a mechanism in addition to the bitstream interface from Definition 3.35 that provides successively improving upper bounds $l_i^+ \ge \log|a_i|$ together with an indication whether they can still be improved (i.e., decreased); if not, they must satisfy $\log|a_i| > l_i^+ - 2$. We can thus maintain tentative values for the expressions in (3.24) in a priority queue and repeatedly improve the tentative maximum until the true value of (3.24) is found.

To keep our subsequent presentation of the bitstream Descartes algorithm separate from these matters, which have more to do with the representation of the coefficients behind the bitstream interface than with the algorithm itself, we describe the algorithm in terms of additional input parameters r, chosen to satisfy (3.21), and l_n^-, chosen to satisfy (3.22).

Let us now discuss the conversion from power to Bernstein basis. We are interested in the $[0,1]$-Bernstein coefficients of $A_{\text{in}}((d_0-c_0)X+c_0) = A_{\text{in}}(2^{r+2}X - 2^{r+1})$, or equivalently (see Lemma 2.21), the $[-1,1]$-Bernstein coefficients of $A_{\text{in}}(2^{r+1}X) = \sum_{j=0}^n a_j 2^{j(r+1)} X^j$.

Proposition 3.37. *If* $F(X) = \sum_{j=0}^n a_j X^j = \sum_{i=0}^n b_i B_i^n[-1,1](X)$, *then*

$$b_i = \sum_{j=0}^{n} \left(\binom{n}{i}^{-1} \sum_{\nu} (-1)^{j-\nu} \binom{j}{\nu} \binom{n-j}{i-\nu} \right) a_j \tag{3.25}$$

with summation over all ν *such that* $\max\{0, i+j-n\} \le \nu \le \min\{i,j\}$. *The coefficients of* $a_0, \ldots, a_n$ *have magnitudes at most 1.*

Proof. Recall that $b_i = F\big[\binom{-1}{1}^{n-i}\binom{1}{1}^{i}\big]$ (Proposition 2.20(i)). With $M = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ we find $b_i = (F \circ M)\big[\binom{0}{1}^{n-i}\binom{1}{0}^{i}\big]$, cf. Lemma 2.13, so b_i is the coefficient of $\binom{n}{i}X^i$ in

$$\begin{aligned}
(X+1)^n F\Big(\frac{X-1}{X+1}\Big) &= \sum_{j=0}^{n} a_j (X-1)^j (X+1)^{n-j} \\
&= \sum_{j=0}^{n} a_j \sum_{\mu,\nu} (-1)^{j-\nu} \binom{j}{\nu} \binom{n-j}{\mu} X^{\mu+\nu} \\
&= \sum_{i=0}^{n} \sum_{j=0}^{n} a_j \sum_{\nu} (-1)^{j-\nu} \binom{j}{\nu} \binom{n-j}{i-\nu} \binom{n}{i}^{-1} \binom{n}{i} X^i \qquad \text{where} \quad i = \mu+\nu.
\end{aligned}$$

The summation range for ν is limited to $0 \le \nu \le j$ by the first binomial coefficient and to $0 \le i-\nu \le n-j \;\Leftrightarrow\; -i \le -\nu \le n-i-j$ by the second.

The claim on magnitudes is an immediate consequence of Vandermonde's convolution formula $\binom{n}{i} = \sum_\nu \binom{j}{\nu}\binom{n-j}{i-\nu}$, see, e.g., [Knu97, §1.2.6]. □

In general, the coefficients in (3.25) are fractional, due to the division by $\binom{n}{i}$. To avoid fractions, we multiply by $n!$; of course, $n!\binom{n}{i}^{-1} = i!(n-i)!$.

Corollary 3.38. *If* $A(X) = \sum_{j=0}^{n} a_j X^j$, *then* $n!A(2^{r+2}X - 2^{r+1}) = \sum_{i=0}^{n} b_i B_i^n[0,1](X)$ *with*

$$b_i = \sum_{j=0}^{n} m_{ij}^{(n)} 2^{j(r+1)} a_j, \qquad m_{ij}^{(n)} = i!(n-i)! \sum_{\nu} (-1)^{j-\nu} \binom{j}{\nu} \binom{n-j}{i-\nu}. \tag{3.26}$$

The coefficients $m_{ij}^{(n)}$ *are integers of magnitude up to* $n!$.

Lemma 3.39. *Given* $n \in \mathbb{N}$, *the numbers* $m_{ij}^{(n)}$, $0 \le i, j \le n$, *can be computed with* $O(n^4 \log n)$ *bit operations.*

Proof. The columns of the matrix $(m_{ij}^{(n)})_{i,j}$ are the images of the basis polynomials $1, X, X^2, \ldots, X^n$ under the conversion to the Bernstein basis w.r.t. $[-1,1]$, followed by multiplication with $n!$. As discussed in §3.2.5 for the Bernstein basis variant of the integer Descartes method, this transformation can be implemented as the composition $\beta RTRH_2T_{-1}$ of two Taylor shifts, a homothetic transformation, two coefficient reversals, and the final multiplication by $i!(n-i)!$. The coefficients of X^j have length 1. The chain of transformations $RTRH_2T_{-1}(X^j)$ produces intermediate results with coefficients of length $O(n)$ and requires $O(n^3)$ bit operations. The final multiplications with factors $i!(n-i)!$ of lengths bounded by $\log n! = O(n \log n)$ requires $O(n^3 \log n)$ bit operations. Thus, transforming all $n+1$ basis polynomials needs $O(n^4 \log n)$ bit operations in total. □

This $O(n^4 \log n)$ bound seems surprisingly large compared to other steps of our algorithm. As we need to compute the $m_{ij}^{(n)}$ only once, at the very beginning of the algorithm, this straightforward computation does not form a bottleneck, though (neither practically nor theoretically), and we contend ourselves with it for the purposes of this thesis.[9]

Let us now determine which multiple of $A_{\text{in}}(2^{r+2}X - 2^{r+1})$ we will actually compute. As a point of reference for precision management, we wish to choose a multiple such that the leading coefficient (i.e., coefficient of X^n) has a magnitude of at least 1, but not much more. Already above, we have postulated the availability of a lower bound $l_n^- \le \log|a_n| < l_n^- + 2$ for the leading coefficient a_n of A_{in}. Based on this, we choose the multiple

$$A_0(X) := 2^{-l} n! A_{\text{in}}(2^{r+2}X - 2^{r+1}) \quad \text{with} \quad l := l_n^- + \lfloor \log n! \rfloor + n(r+2), \tag{3.27}$$

which has a leading coefficient $|a_n^{(0)}| = |2^{-l} n! 2^{n(r+2)} a_n| = n!/2^{\lfloor \log n! \rfloor} \cdot |a_n| / 2^{l_n^-} \in [1, 8)$, meeting our needs. Hence we let our algorithm compute approximations to the following Bernstein coefficients:

$$A_0(X) = \sum_{i=0}^{n} \beta_i^{(0)} B_i^n[0,1](X), \qquad \beta_i^{(0)} = \sum_{j=0}^{n} m_{ij}^{(n)} 2^{j(r+1)-l} a_j. \tag{3.28}$$

Proposition 3.40. *The Bernstein coefficients of* $A_0(X)$ *satisfy*

$$|\beta_0^{(0)}|, |\beta_n^{(0)}| \ge (1/4)^n, \qquad |\beta_i^{(0)}| < 8 \cdot (3/4)^n \quad \text{for} \quad 0 \le i \le n. \tag{3.29}$$

[9] Nevertheless, the matrices $(m_{ij}^{(n)})_{i,j}$ exhibit an interesting structure, and one might speculate that more sophisticated algorithms could take advantage of it.

Proof. The roots $\vartheta_1, \ldots, \vartheta_n$ of A_0 satisfy $|\vartheta_j - 1/2| \le 1/4$, so $1/4 \le |\vartheta_j| \le 3/4$ and $1/4 \le |1-\vartheta_j| \le 3/4$ for all j. The leading coefficient $a_n^{(0)}$ of A_0 has magnitude $1 \le |a_n^{(0)}| < 8$. The claims now follow with Proposition 2.24 (page 25):

$$|\beta_0^{(0)}/a_n^{(0)}| = \prod_{j=1}^{n} |\vartheta_j| \ge (1/4)^n, \qquad |\beta_n^{(0)}/a_n^{(0)}| = \prod_{j=1}^{n} |1-\vartheta_j| \ge (1/4)^n,$$
$$|\beta_i^{(0)}/a_n^{(0)}| \le \tbinom{n}{i}^{-1} \sum_{\#J=n-i} \prod_{j\in J} |\vartheta_j| \prod_{j\notin J} |1-\vartheta_j| \le (3/4)^n \quad \text{for} \quad 0 \le i \le n. \qquad \square$$

Whenever the bitstream Descartes algorithm chooses a precision parameter p, we have to compute approximations of $\beta_i^{(0)}$ with absolute errors bounded by 2^{-p-1}. (The use of $-p-1$ instead of $-p$ is explained in the next section.) To do so, we compute at precision $q := p + \lceil \log n! \rceil + \lceil \log(n+1) \rceil + 2$. We extract significands $c_j \leftarrow [2^{j(r+1)-l+q} a_j]$ out of the bitstreams a_j. They satisfy

$$|c_j - 2^{j(r+1)-l+q} a_j| \le 1, \qquad \text{or equivalently} \quad |c_j 2^{-q} - 2^{j(r+1)-l} a_j| \le 2^{-q}. \tag{3.30}$$

Then we set

$$b_i = \left\lfloor \Big(\sum_{j=0}^{n} m_{ij}^{(n)} c_j \Big) / 2^{q-p-1} \right\rceil \qquad \text{for} \quad 0 \le i \le n, \tag{3.31}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer (arbitrarily for .5) and introduces an error of at most 1/2.

Proposition 3.41. *The computation (3.31) yields approximations $|b_i 2^{-p-1} - \beta_i^{(0)}| \le 2^{-p-1}$ to the $[0,1]$-Bernstein coefficients of $A_0(X)$ with integers $b_0, \ldots, b_n$. Each non-zero b_i has a bit length bounded by $\log |b_i| < p+4$. This computation requires $O(n^3 \log n \cdot (p + \log n))$ bit operations, assuming the factors $m_{ij}^{(n)}$ have already been constructed.*

Proof. Regarding the error bound, we deduce from (3.31), (3.28) and (3.30) that

$$|b_i 2^{-p-1} - \beta_i^{(0)}| \le 2^{-p-2} + |\sum_{j=0}^{n} m_{ij}^{(n)} (c_j 2^{-q} - 2^{j(r+1)-l} a_j)| \le 2^{-p-2} + (n+1) n! 2^{-q},$$

using that $|m_{ij}^{(n)}| \le n!$ for all $0 \le i, j \le n$. By choice of q, this error is at most 2^{-p-1}.

Concerning bit length, we observe that $(|b_i| - 1) 2^{-p-1} \le |\beta_i^{(0)}| < 8 \cdot (3/4)^n$ by Proposition 3.40, so $|b_i| < 1 + 2^{p+4-n\log(4/3)} < 2^{p+4}$.

Let us now examine the computational effort. Since all roots of A_{in} have magnitude 2^r or less, we have $|a_j| \le \binom{n}{j} 2^{r(n-j)} |a_n| \le 2^{r(n-j)+n-1} |a_n|$. Also, $|c_j| \le 1 + 2^{j(r+1)-l+q} |a_j|$. We have $-l + q < -\log |a_n| - n(r+2) + \lceil \log(n+1) \rceil + p + 5$. For $|c_j| \ge 2$, this entails $\log(|c_j| - 1) \le (j(r+1) - l + q) + (r(n-j) + n - 1) + \log |a_n| < p - (n-j) + \lceil \log(n+1) \rceil + 4$. Thus, all c_j are integers of $p + O(\log n)$ bits. We multiply them with weights $0 \le m_{ij}^{(n)} \le n!$, where $\log n! = O(n \log n)$. These $(n+1)^2$ multiplications and the subsequent additions require $O(n^2 \cdot n \log n \cdot (p + \log n))$ bit operations.[10] $\square$

[10] If r is chosen according to the dyadic Fujiwara complex root bound as in (3.24), then one even has $2^{r-1} \ge |a_j/a_n|^{1/(n-j)} \Leftrightarrow |a_j| \le 2^{(r-1)(n-j)} |a_n|$, and can replace $p + O(\log n)$ by $p + O(1)$. However, we do not want to tie our analysis to this specific choice of r.

3.3.3 De Casteljau's algorithm in fixed precision

Let us consider a sequence of real numbers $(\beta_i)_{i=0}^n$ that are Bernstein coefficients of some polynomial, and a subdivision parameter $\alpha \in (0,1)$ of the form $\alpha = u/2^k$ for integers k and $0 < u < 2^k$. Suppose we have an approximation $(b_i 2^{-q})_{i=0}^n$ of the coefficients with integers $b_0, \ldots, b_n$ and q such that $\forall i\colon |b_i 2^{-q} - \beta_i| \leq \varepsilon$ for some $\varepsilon > 0$ moderately larger than 2^{-q}. Our goal is to approximate the exact result of de Casteljau's algorithm (see §2.2.5) when invoked conceptually as $((\beta'_i)_i, (\beta''_i)_i) \leftarrow \mathit{DeCasteljau}((\beta_i)_i, \alpha)$.

Let us begin with a thought experiment: We run de Casteljau's algorithm in exact arithmetic on the approximations $(b_i 2^{-q})_i$ by letting $b_{0,i} = b_i$ and $b_{j,i} = (2^k - u)b_{j-1,i} + ub_{j-1,i+1}$ for $j > 0$. The number $b_{j,i}2^{-q-jk}$ approximates its counterpart $\beta_{j,i}$ in the idealized computation. It holds that $|b_{j,i}2^{-q-jk} - \beta_{j,i}| \leq \varepsilon$, because inductively $|b_{j,i}2^{-q-jk} - \beta_{j,i}| = |(1-u/2^k)(b_{j-1,i}2^{-q-(j-1)k} - \beta_{j-1,i}) + (u/2^k)(b_{j-1,i+1}2^{-q-(j-1)k} - \beta_{j-1,i+1})| \leq (1-\alpha)\varepsilon + \alpha\varepsilon = \varepsilon$. Have we now reached our goal? Yes, but in a very costly way: The inputs have binary representations with exponent $-q$. Each convex combination adds k additional bits, so the entries in row j of the de Casteljau triangle have exponents $-(q+jk)$, whereas the error bound still is ε, so all these extra bits are useless for approximating $\beta_{j,i}$. Therefore, it is preferable to run de Casteljau's algorithm in approximate arithmetic: We set $b_{j,i} = \lfloor ((2^k - u)b_{j-1,i} + ub_{j-1,i+1})/2^k \rceil$, where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer (arbitrarily for .5). In this way, all significands $b_{j,i}$ are understood w.r.t. the same exponent, namely $-q$. We can leave out the factor 2^{-q} entirely and arrive at the following algorithm (cf. §2.2.5).

```
 1: procedure DeCasteljauApprox((b_0, ..., b_n), u, k)
 2:     (b_{0,0}, b_{0,1}, ..., b_{0,n}) ← (b_0, ..., b_n);
 3:     for j from 1 to n do
 4:         for i from 0 to n − j do
 5:             b_{j,i} ← ⌊((2^k − u)b_{j−1,i} + ub_{j−1,i+1})/2^k⌉;
 6:         od;
 7:     od;
 8:     (b'_0, b'_1, ..., b'_n) ← (b_{0,0}, b_{1,0}, ..., b_{n,0});
 9:     (b''_0, b''_1, ..., b''_n) ← (b_{n,0}, b_{n−1,1}, ..., b_{0,n});
10:     return ((b'_j)_{j=0}^n, (b''_i)_{i=0}^n);
11: end procedure;
```

Proposition 3.42. *Consider a sequence of Bernstein coefficients $(\beta_0, \ldots, \beta_n)$ and an approximation in terms of integers $b_0, \ldots, b_n$ and q such that $|b_i 2^{-q} - \beta_i| \leq \varepsilon$ and $|b_i| < 2^\tau$ for all $0 \leq i \leq n$. Let $u, k \in \mathbb{N}$ such that $0 < u < 2^k$. Consider exact and approximate subdivision at $\alpha = u/2^k$ via*

$$((\beta'_i)_i, (\beta''_i)_i) \leftarrow \mathit{DeCasteljau}((\beta_i)_i, u/2^k),$$
$$((b'_i)_i, (b''_i)_i) \leftarrow \mathit{DeCasteljauApprox}((b_i)_i, u, k).$$

It holds that $|b'_i 2^{-q} - \beta'_i| \leq \varepsilon + n2^{-q-1}$ and $|b'_i| < 2^\tau$ for all $0 \leq i \leq n$; the analogous bounds hold for b''_i. Executing $\mathit{DeCasteljauApprox}((b_i)_i, u, k)$ takes $O(n^2 k\tau)$ bit operations.

Proof. We show by induction on j that $|b_{j,i}2^{-q} - \beta'_{j,i}| \leq j2^{-q-1} + \varepsilon$ and $|b_{j,i}| \leq 2^\tau - 1$ for all $0 \leq j \leq n$ and $0 \leq i \leq n - j$. The base case $j = 0$ is immediate from the condition on

the inputs. For the inductive step from $j-1$ to j concerning the first claim, we observe

$$\begin{aligned}
|b_{j,i}2^{-q} - \beta_{j,i}| &= \left|\left\lfloor((2^k-u)b_{j-1,i} + ub_{j-1,i+1})/2^k\right\rceil 2^{-q} - ((1-\alpha)\beta_{j-1,i} + \alpha\beta_{j-1,i+1})\right| \\
&\le 2^{-q-1} + \left|((1-\alpha)b_{j-1,i}2^{-q} + \alpha b_{j-1,i+1}2^{-q}) - ((1-\alpha)\beta_{j-1,i} + \alpha\beta_{j-1,i+1})\right| \\
&\le 2^{-q-1} + (1-\alpha)\left|b_{j-1,i}2^{-q} - \beta_{j-1,i}\right| + \alpha\left|b_{j-1,i+1}2^{-q} - \beta_{j-1,i+1}\right| \\
&\le j2^{-q-1} + \varepsilon.
\end{aligned}$$

Concerning the second claim, we observe that $2^\tau - 1$ is an integer upper bound for the convex combination $(1-\alpha)b_{j-1,i} + \alpha b_{j-1,i+1}$, and rounding does not go beyond this upper bound, so $b_{j,i} = \lfloor(1-\alpha)b_{j-1,i} + \alpha b_{j-1,i+1}\rceil \le 2^\tau - 1$. Symmetrically, $b_{j,i} \ge -(2^\tau - 1)$.

The algorithm performs $n(n+1)/2$ convex combinations. In each convex combination, the multiplication of a k-bit weight with a τ-bit coefficient needs $O(k\tau)$ bit operations. The results have lengths $k+\tau$ and can be added with $O(k+\tau)$ operations. Since the weights add up to 2^k, the sum has again length $k+\tau$. Its rounded division by a power of two can also be performed with $O(k+\tau)$ operations. □

We will also have occasion to perform approximate subdivision at $\alpha = u/v \in (0,1)$ where $u, v \in \mathbb{N}$ but v is not necessarily a power of two. The following approximate de Casteljau algorithm is similar to the preceding one. In fact, it obeys the same bounds on bit complexity and approximation error.

```
1: procedure DeCasteljauRatApprox((b_0, ..., b_n), u, v)
2:     (b_{0,0}, b_{0,1}, ..., b_{0,n}) ← (b_0, ..., b_n);
3:     for j from 1 to n do
4:         for i from 0 to n − j do
5:             b_{j,i} ← ⌊((v − u)b_{j−1,i} + ub_{j−1,i+1})/v⌉;
6:         od;
7:     od;
8:     (b'_0, b'_1, ..., b'_n) ← (b_{0,0}, b_{1,0}, ..., b_{n,0});
9:     (b''_0, b''_1, ..., b''_n) ← (b_{n,0}, b_{n−1,1}, ..., b_{0,n});
10:    return ((b'_j)_{j=0}^n, (b''_i)_{i=0}^n);
11: end procedure;
```

Proposition 3.43. *Consider a sequence of Bernstein coefficients $(\beta_0, \ldots, \beta_n)$ and an approximation in terms of integers $b_0, \ldots, b_n$ and q such that $|b_i 2^{-q} - \beta_i| \le \varepsilon$ and $|b_i| < 2^\tau$ for all $0 \le i \le n$. Let $u, v, k \in \mathbb{N}$ such that $0 < u < v < 2^k$. Consider exact and approximate subdivision at $\alpha = u/v$ via*

$$\begin{aligned}
((\beta'_i)_i, (\beta''_i)_i) &\leftarrow \mathit{DeCasteljau}((\beta_i)_i, u/v), \\
((b'_i)_i, (b''_i)_i) &\leftarrow \mathit{DeCasteljauRatApprox}((b_i)_i, u, v).
\end{aligned}$$

It holds that $|b'_i 2^{-q} - \beta'_i| \le \varepsilon + n2^{-q-1}$ and $|b'_i| < 2^\tau$ for all $0 \le i \le n$; the analogous bounds hold for b''_i. Executing $\mathit{DeCasteljauRatApprox}((b_i)_i, u, v)$ takes $O(n^2 k\tau)$ bit operations.

Proof. The proof is mostly identical to Proposition 3.42. It just remains to show that each convex combination $b_{j,i} \leftarrow \lfloor(v-u)b_{j-1,i} + ub_{j-1,i+1})/v\rceil$ can be performed with $O(k\tau)$ bit operations. The two multiplications of a k-bit weight and a τ-bit coefficient are certainly covered by this bound, as is the subsequent addition of two $(k+\tau)$-bit numbers.

The final integer division of a $(k+\tau)$-bit dividend by a k-bit divisor produces a τ-bit quotient, so ordinary long division needs $O(k\tau)$ bit operations as well [Knu69, §4.3.1]. □

Now we consider the chains of de Casteljau subdivisions by which the Descartes method produces its transformed polynomials. It is our goal to provide the Bernstein coefficients of all transformed polynomials with an absolute error bounded by $\varepsilon = 2^{-p}$, where the precision parameter p is chosen by the algorithm; this needs to cover both the approximation error in the coefficients we start from and the arithmetical error accumulating during subdivision. We start from an approximation $(b_i^{(0)}2^{-q})_i$ of initial Bernstein coefficients $(\beta_i^{(0)})_i$ with errors $|b_i^{(0)}2^{-q} - \beta_i^{(0)}| \le 2^{-p-1}$ and perform a sequence of subdivisions with subdivision parameters $(\alpha_j)_{j\ge1}$ in $(0,1)$, such that after the jth subdivision, either the left or right part of the result is taken as input for the $(j+1)$st subdivision. Conceptually, these subdivisions are carried out exactly by procedure *DeCasteljau*, inducing a sequence $((\beta_i^{(j)})_i)_{j\ge1}$ of exact transformations of $(\beta_i^{(0)})_i$. What we actually compute, though, is a sequence of approximations $((b_i^{(j)}2^{-q})_i)_j$ produced by invocations of *DeCasteljauApprox* and *DeCasteljauRatApprox*. Inductive application of the preceding propositions shows $|b_i^{(j)}2^{-q} - \beta_i^{(j)}| \le 2^{-p-1} + jn2^{-q-1}$. How do we choose q in order to restrict this to 2^{-p} or less?

Let us assume for the time being that we know an a priori bound d_{bd} (a power of two) on the length of any chain, that is, the maximal subdivision depth. Then it suffices to have $d_{\mathrm{bd}}n2^{-q-1} \le 2^{-p-1}$, which is equivalent to $q \ge p + \log(d_{\mathrm{bd}}n)$ and easily achieved by setting $q := p + \lceil \log n \rceil + \log d_{\mathrm{bd}}$. Until we return to the issue in §3.3.6, we stick to this choice of q.

3.3.4 Sign variations from approximate coefficients

This section presents two lemmas demonstrating that uncertainty in the distinction between 0, 1, and more than 1 sign variation in a sequence of approximate Bernstein coefficients disappears after one further subdivision step, provided that the first and last Bernstein coefficient have a magnitude exceeding the error bound $\varepsilon = 2^{-p}$ by a certain factor C. For concreteness, we formulate the lemmas with particular signs of the relevant coefficients, but by the linearity and symmetry of de Casteljau's algorithm, the extension to the remaining cases is sufficiently obvious not to formulate it explicitly. We begin by formalizing the ambiguity in the number of sign variations of approximate numbers.

Definition 3.44. Let $\varepsilon > 0$. For a sequence $(\widetilde{a}_0, \ldots, \widetilde{a}_n)$ of real numbers, its *set of ε-approximate numbers of sign variations* is

$$\mathrm{var}_\varepsilon(\widetilde{a}_0, \ldots, \widetilde{a}_n) := \big\{ \mathrm{var}(a_0, \ldots, a_n) \;\big|\; |\widetilde{a}_i - a_i| \le \varepsilon \text{ for all } 0 \le i \le n \big\}. \tag{3.32}$$

Given ε-approximations $\widetilde{a}_0, \ldots, \widetilde{a}_n$ of $a_0, \ldots, a_n$, clearly $\mathrm{var}(a_0, \ldots, a_n) \in \mathrm{var}_\varepsilon(\widetilde{a}_0, \ldots, \widetilde{a}_n)$.

Lemma 3.45. *Let $\varepsilon = 2^{-p}$ and $q \ge p + \log n$ and $C \ge 4^{n+1}$. Let $(b_i 2^{-q})_{i=0}^n$ be a vector of approximate Bernstein coefficients such that $b_0 2^{-q}, b_n 2^{-q} > C\varepsilon$. Consider the execution of $((b'_i)_i, (b''_i)_i) \leftarrow$ DeCasteljauApprox$((b_i)_i, u, k)$ for some $\alpha = u/2^k \in [1/4, 3/4]$. If $0 \in \mathrm{var}_\varepsilon((b_i 2^{-q})_i)$, then $\mathrm{var}_\varepsilon((b'_i 2^{-q})_i) = \mathrm{var}_\varepsilon((b''_i 2^{-q})_i) = \{0\}$.*

Proof. We argue using an idealized de Casteljau triangle $c_{j,i}$, which is computed exactly from inputs c_i modified as follows: If $|b_i 2^{-q}| \le \varepsilon$, we set $c_i = 0$; otherwise, we set $c_i = b_i 2^{-q}$.

By Proposition 3.42, corresponding entries in the idealized triangle $c_{j,i}$ and the actually computed triangle $b_{j,i}2^{-q}$ differ by at most $\varepsilon + n2^{-q-1} \leq 2^{-p} + 2^{-p-1} = 3/2 \cdot \varepsilon$. Whenever we can show $|c_{j,i}| > 5/2 \cdot \varepsilon$, we may conclude $\operatorname{sgn}(c_{j,i}) = \operatorname{sgn}_\varepsilon(b_{j,i}2^{-q}) \neq ?$.

Let us now inspect the idealized de Casteljau triangle. All modified inputs c_i are non-negative. Due to the contribution of c_0 or c_n, resp., any element in the idealized outputs $(c'_i)_i$ and $(c''_i)_i$ is at least $4^{-n}C\varepsilon \geq 4\varepsilon$. Thus any element of the actual outputs $(b'_i 2^{-q})_i$ and $(b''_i 2^{-q})_i$ is sign-determinate and positive. □

Lemma 3.46. *Let $\varepsilon = 2^{-p}$ and $q \geq p + \log n$ and $C \geq 16^n$. Let $(b_i 2^{-q})_{i=0}^n$ be a vector of approximate Bernstein coefficients such that $b_0 2^{-q} > C\varepsilon$ and $b_n 2^{-q} < -C\varepsilon$. Consider the execution of $((b'_i)_i, (b''_i)_i) \leftarrow$ DeCasteljauApprox$((b_i)_i, u, k)$ for $\alpha = u/2^k \in [1/4, 3/4]$. If $1 \in \operatorname{var}_\varepsilon((b_i 2^{-q})_i)$ and if, at the tip of the de Casteljau triangle, $b'_n 2^{-q} = b''_0 2^{-q} < -C\varepsilon$, then $\operatorname{var}_\varepsilon((b'_i 2^{-q})_i) = \{1\}$ and $\operatorname{var}_\varepsilon((b''_i 2^{-q})_i) = \{0\}$.*

Proof. We argue using a modified de Casteljau triangle $c_{j,i}$ as in the proof of Lemma 3.45. By its construction (see above), the modified input sequence $(c_i)_i$ consists of non-negative followed by non-positive numbers. It is easy to see inductively that all rows of the modified de Casteljau triangle consist of zero or more non-negative elements followed by one or more non-positive elements. Once some row consists entirely of non-positive elements, the same holds for all further rows.

We first prove the claim about b''. The element $c_{n,0}$ at the tip of the modified triangle is less than $-(C - 3/2)\varepsilon$. An element $c_{j,i}$ in row $j \geq 1$ cannot be less than the minimum of its parents $c_{j-1,i}$ and $c_{j-1,i+1}$, so there is a path $\mathcal{P}$ of elements less than $-(C - 3/2)\varepsilon$ from row 0 to row n. The elements right of $\mathcal{P}$ are non-positive.

Now consider the rightmost element c''_{n-i} in row i of the triangle, for arbitrary $0 \leq i < n$. Go up $0 \leq k \leq i$ times to the left parent until you reach an element of $\mathcal{P}$ in row $i - k$ or end up in row 0 right of the path (with $k = i$). In either case, the last $k + 1$ elements of row $i - k$ are non-positive, one of them, say c_*, is less than $-(C - 3/2)\varepsilon$ (namely the path element or c_n), and c''_{n-i} is a convex combination of them. Due to the contribution of c_*, we have $c''_{n-i} < -4^{-k}(C - 3/2)\varepsilon \ll -3\varepsilon$, and thus all $b''_{n-i}2^{-q}$ are sign-determinate and negative. Consequently, $\operatorname{var}_\varepsilon((b''_i 2^{-q})_i) = \{0\}$.

We turn to b'. Its modified counterpart begins with $c'_0 > C\varepsilon$ and ends at the tip of the triangle with $c'_n < -(C - 3/2)\varepsilon$. We will demonstrate the existence of a unique ε-approximate sign variation in the sequence $(b'_i 2^{-q})_i$ near index

$$i = \min\left\{ i \in \{1, \ldots, n\} \;\middle|\; c'_i \leq 0 \text{ or } |c'_i| \leq |c'_{i-1}|/16 \right\}.$$

Since c'_n is negative, i exists. By minimality of i, we have for all $0 < j < i$ that $c'_j > 0$ and $c'_j > c'_{j-1}/16 > c'_0/16^j > (C/16^{n-1})\varepsilon > 16\varepsilon$. Thus $c'_0, c'_1, \ldots, c'_{i-1} > 16\varepsilon$.

Now we consider the remaining elements. We have chosen i such that there is a sharp decrease from c'_{i-1} to c'_i; so sharp in fact that $c_{i-1,1}$ must be much less than zero, and that forces $c'_{i+1}, \ldots, c'_n$ to negativity as well. More precisely, we will show $c'_{i+1}, \ldots, c'_n < -3\varepsilon$. For c'_n, this is already known, so we may assume $i \leq n-2$. By choice of i, we have $c'_{i-1} > 0$ and $c'_i \leq c'_{i-1}/16$. From $c'_i = (1 - \alpha)c'_{i-1} + \alpha c_{i-1,1}$ follows then

$$c_{i-1,1} = (c'_i - (1 - \alpha)c'_{i-1})/\alpha \leq (1/16 - (1 - \alpha))/\alpha \cdot c'_{i-1} = (16\alpha - 15)/(16\alpha) \cdot c'_{i-1}.$$

This upper bound is negative for any $\alpha \in [1/4,\, 3/4]$, so $c_{i-1,1} < 0$ and all successive entries in same row are not positive either; in particular, $c_{i-1,2} \leq 0$. Now we observe

$$\begin{aligned} c'_{i+1} &= (1-\alpha)^2 c'_{i-1} + 2(1-\alpha)\alpha c_{i-1,1} + \alpha^2 c_{i-1,2} \\ &\leq (1-\alpha)^2 c'_{i-1} + 2(1-\alpha)\alpha c_{i-1,1} \\ &\leq (1-\alpha)^2 c'_{i-1} + (1-\alpha)(16\alpha - 15)/8 \cdot c'_{i-1} \\ &= \underbrace{(-\alpha^2 + 15\alpha/8 - 7/8)}_{\leq -1/32 \text{ for } 1/4 \leq \alpha \leq 3/4} c'_{i-1} \\ &< (-1/32)(c'_0/16^{i-1}) \\ &< -1/2^{4i+1} \cdot C\varepsilon. \end{aligned}$$

All entries in rows $i+1$ to n are non-positive, and the entries c'_j for $j > i+1$ receive a fraction $(1-\alpha)^{j-(i+1)}$ of c'_{i+1}. Thus, $c'_j \leq c'_{i+1}/4^{j-(i+1)} < -C/2^{2i+2j-1} \cdot \varepsilon < -3\varepsilon$ for all $i+1 \leq j \leq n-1$.

In summary, we see that $b'_0 2^{-q}, \ldots, b'_{i-1} 2^{-q}$ are sign-determinate and positive, whereas $b'_{i+1} 2^{-q}, \ldots, b'_n 2^{-q}$ are sign-determinate and negative, and so $\mathrm{var}_\varepsilon((b'_i 2^{-q})_i) = \{1\}$. □

The two preceding lemmas were discovered by K. Mehlhorn (personal communication, December 2004) for the case of exact arithmetic and bisection at the midpoint ($\alpha = 1/2$).

3.3.5 The bitstream Descartes algorithm: outline

We are now ready to outline our variant of the Descartes method designed for polynomials with bitstream coefficients, which we call the *bitstream Descartes algorithm* for short. For simplicity, we still uphold the assumption made in §3.3.3, namely that we know an a priori bound d_{bd} on the maximal subdivision depth and can thus determine a global exponent $-q$ for our fixed-point arithmetic. The actual bitstream Descartes algorithm presented in §3.3.7 is slightly more complicated, owing to the removal of this assumption. Further variants of the algorithm are discussed in §3.3.10.

With reference to the preprocessing steps described in §3.3.2, we formulate our algorithm in terms of the initial interval I_0 from (3.20) and the polynomial $A_0(X)$ from (3.27). We recall that all real roots of A_0 (or equivalently, of A_{in}) must be simple. With reference to Lemmas 3.45 and 3.46, we define

$$C := 2^{4n}, \tag{3.33}$$

satisfying the premises of both.

The algorithm maintains a state comprising the following items.

- A precision parameter p, initialized to $p \leftarrow p_0 := 6n+1$, as well as counters $N_{\mathrm{try}}, N_{\mathrm{fail}}$, initially set to 0. The parameter p determines the global error bound

  $$\varepsilon := 2^{-p} \tag{3.34}$$

  to be observed by all approximate Bernstein coefficients. Furthermore, p determines the exponent $q_0 > p$ to be used in the representation of approximate Bernstein coefficients, see §3.3.3.

  When the algorithm decides to increase the precision, it sets $p \leftarrow 2p$; thus, the value of p after μ increments is $p_\mu = 2^\mu p_0$.

- An approximation of $A_0(X)$ represented by integers $b_0^{(0)}, \ldots, b_n^{(0)}$ such that

$$|b_i^{(0)} 2^{-q_0} - \beta_i^{(0)}| \leq 2^{-p-1} = \varepsilon/2 \qquad \text{for} \quad 0 \leq i \leq n. \tag{3.35}$$

  To obtain $b_0^{(0)}, \ldots, b_n^{(0)}$, we perform the computation described by Equation (3.31) and Proposition 3.41, and then change the exponent from $-p-1$ to $-q_0$ by setting $b_i^{(0)} = b_i 2^{q_0-p-1}$.
- A sequence P recording the current partition of the initial interval as in the general form of the Descartes method (§3.1.1).
- A set Q with entries $((c,d),(b_0,\ldots,b_n),V)$ to record intervals (c,d) for further subdivision. The numbers $(b_i 2^{-q})_i$ are approximations of the $[0,1]$-Bernstein coefficients of $A(X) = A_0((d-c)X + c)$ with absolute error at most ε. For the moment, we assume the exponent q is equal to the global value q_0 derived from p. The entry $V = \mathrm{var}_\varepsilon((b_i 2^{-q})_i)$ is the set of ε-approximate numbers of sign variations in the coefficient sequence; it may be represented compactly by the two integers $\min V$ and $\max V$.

The algorithm maintains the following invariant:

$$((c,d),(b_0,\ldots,b_n),\ldots) \in Q \implies |b_0 2^{-q}|, |b_n 2^{-q}| > C\varepsilon. \tag{3.36}$$

In other words, the first and last coefficient of each transformed polynomial stored in Q have sufficient magnitude to satisfy the respective conditions of Lemmas 3.45 and 3.46. For the initial interval, the invariant is satisfied by construction of A_0 and choice of p_0: The true leftmost Bernstein coefficient of A_0 satisfies $|\beta_0^{(0)}| > 2^{-2n}$ by Proposition 3.40, so we have for its approximation that $|b_0^{(0)} 2^{-q}| \geq 2^{-2n} - 2^{-6n-1} > 2^{-2n-1} = C2^{-p_0}$; likewise $|b_n^{(0)} 2^{-q}| > C2^{-p_0}$ for the rightmost Bernstein coefficient.

The main loop of the bitstream Descartes algorithm operates as follows. While Q is non-empty, an entry $((c,d),(b_i)_i,V)$ is extracted for subdivision. (Our complexity analysis will impose a mild condition on how to choose from Q in Proposition 3.60.) The algorithm guesses a subdivision parameter $\alpha = u/K$ by choosing u uniformly at random from

$$u \in \{K/4,\ K/4+1,\ \ldots,\ 3K/4\}, \qquad K := 2^k, \quad k := 4 + \lceil \log n \rceil. \tag{3.37}$$

This corresponds to choosing the subdivision point $m = (1-\alpha)c + \alpha d$. The subdivision at m is performed tentatively by executing $((b_j^L)_j,\ (b_j^R)_j) \leftarrow \mathit{DeCasteljauApprox}((b_j)_j, u, k)$, and the counter N_{try} is incremented. The algorithm inspects the magnitude of $b_0^R 2^{-q} = b_n^L 2^{-q}$ to check whether the invariant (3.36) is violated by m (case a) or satisfied by m (case b).

(a) Let us first suppose that a tentative subdivision at m cannot be used because $|b_0^R 2^{-q}| \leq C\varepsilon$. We say that such a subdivision has *failed*. Whenever a subdivision has failed, the algorithm increments N_{fail}. If the error bound ε is small enough, subdivisions should fail rarely.

If $N_{\mathrm{fail}} < 2$ or $N_{\mathrm{fail}} < N_{\mathrm{try}}/2$, we conclude that failing subdivisions are indeed rare, and so we go back to making a new random choice of α and continue as above.

If $N_{\mathrm{fail}} \geq 2$ and $N_{\mathrm{fail}} \geq N_{\mathrm{try}}/2$, however, failing subdivisions appear to be relatively frequent, so we are led to believe that ε is not yet small enough. Thus, we increase the precision by setting $p \leftarrow 2p$, and we reinitialize A_0 according to this increased precision.

From A_0, we recompute the coefficients in each entry of Q by performing two subdivision steps $[c_0, d_0] \rightsquigarrow [c, d_0] \rightsquigarrow [c, d]$. It is straightforward to verify that this reinitialization preserves the invariant (3.36): The decrease in the reduced bound $C\varepsilon$ on the magnitude is much larger than the potential decrease of magnitude in the improved approximation. For the new coefficients, the set V of ε-approximate sign variations is recomputed. The new set V may be a proper subset of the previous set V; if $\max V < 2$, the entry is removed from Q.

In the analysis (§3.3.8), we will show that for any input A_{in} there is a sufficient precision such that the probability of each further precision increment beyond this threshold is no more than 1/10. Thus, with probability 1, the algorithm settles at some maximal precision and explores its entire subdivision tree $\mathcal{T}'$.

(b) Let us now discuss the case that a tentative subdivision at m yields $|b_0^R 2^{-q}| > C\varepsilon$ and thus satisfies the invariant (3.36). In this case, we say that subdivision at m has *succeeded*, and the algorithm commits to this subdivision. For both subintervals (c, m) and (m, d), the algorithm checks the set of ε-approximate numbers of sign variations: Let V denote $\text{var}_\varepsilon((b_j^L)_j)$ or $\text{var}_\varepsilon((b_j^R)_j)$, resp., and let v denote the true value of *DescartesTest*$(A_{\text{in}}, \cdot)$ for the interval considered. Of course, $V \supseteq \{v\}$. The elements of V are all even or all odd, because the first and last coefficient are sign-determinate. The following five possibilities remain:

(D0) $V = \{0\}$ — definitely $v = 0$,

(D1) $V = \{1\}$ — definitely $v = 1$,

(D2) $V \cap \{0, 1\} = \emptyset$ — definitely $v \geq 2$,

(M0) $V \supsetneq \{0\}$ — maybe $v = 0$, maybe $v \geq 2$,

(M1) $V \supsetneq \{1\}$ — maybe $v = 1$, maybe $v \geq 2$.

The algorithm behaves as follows.

If a newly constructed interval I falls into case (D0) or (M0), it is discarded, as in case $v = 0$ in the exact Descartes method. As a consequence, intervals with a true Descartes test value $v = 0$ are never recorded in Q. But why is this the right action, even in case (M0)? Lemma 3.45 shows that any further subdivision of I with $1/4 \leq \alpha \leq 3/4$ would produce two subintervals of type (D0), so we may conclude right away that I does not contain any root.

In case (D1), the newly constructed interval I is processed as in case $v = 1$ in the exact Descartes method: it is retained as isolating interval in the sequence P but not recorded for further subdivision in the set Q.

In cases (D2) and (M1), the newly constructed interval I is recorded in P and put back into Q as if $v \geq 2$, with the transformed coefficients as computed in subdivision. If indeed $v \geq 2$, this is certainly the right action. But what happens to an interval I with true Descartes test $v = 1$ if we put it back into Q? The algorithm will attempt to subdivide it further. Potentially, this may trigger precision increments. Ultimately, one of two alternatives occurs: One is that a precision increment removes the sign-indeterminacy of some coefficients and promotes I into case (D1); then I remains un-subdivided. The other alternative is that the algorithm succeeds in subdividing I; if so, Lemma 3.46 combined with the invariant (3.36) guarantees that the subintervals of I fall into the definite cases (D0) and (D1), and thus no further subdivision takes place.

Either way, the bitstream Descartes algorithm subdivides at most one level deeper than the exact Descartes method. We record this for later reference.

Lemma 3.47. *Consider the subdivision tree $\mathcal{T}'$ constructed by the bitstream Descartes algorithm and the subdivision tree $\mathcal{T}$ constructed by the exact Descartes method when executed with the same inputs and choices of subdivision points. If an interval occurs as a non-root node in $\mathcal{T}'$, then its parent in $\mathcal{T}'$ occurs as a node in $\mathcal{T}$.*

Since we have required all real roots of A_{in} to be simple, $\mathcal{T}$ and therefore also $\mathcal{T}'$ is finite (see Theorem 3.19(ii) on page 60). With probability 1, the algorithm settles at a maximal precision p, explores the finite tree $\mathcal{T}'$ entirely, and terminates.

There is a sharp asymmetry between the cases (M0) and (M1): For (M0), Lemma 3.45 allows us to draw a conclusion from a hypothetical subdivision that we do not actually carry out. For (M1), however, we need to perform this subdivision, because we need to check that there actually is a subdivision point satisfying the additional hypothesis of Lemma 3.46 on the coefficient at the tip of the de Casteljau triangle.[11]

The sequence P and the set Q can be represented efficiently by linked lists as discussed for the exact Descartes method in §3.1.3; the explanations there on the relation of ordering Q and the induced traversal order of the subdivision tree carry over.

3.3.6 Adaptive choice of working precision

As we have chosen to carry out the transformations of A_0 in fixed-point arithmetic with least significant digit 2^{-q}, a succession of d subdivisions leads to accumulated arithmetical error up to $dn2^{-q-1}$ on top of the initial approximation error bound 2^{-p-1} (see §3.3.3). Up to this point, we have made the assumption that we know an a priori bound on d and can thus choose q large enough to make $2^{-p-1} + dn2^{-q-1} \leq 2^{-p}$. We will now remove this assumption.

We make an initial estimate d_0 of subdivision depth (a power of two) and choose an initial precision $q_0 = p + \lceil \log n \rceil + \log d_0 + 1$. Subdivision up to depth d_0 incurs a cumulative arithmetical error of at most 2^{-p-2}, half of the available error margin. Having reached subdivision depth d_ν (where $\nu \geq 0$), with an error margin of $2^{-p-\nu-2}$ remaining, we extend the significand lengths of the approximations from q_ν to $q_{\nu+1} = q_\nu + 2 = p + \lceil \log n \rceil + \log d_0 + 2(\nu+1) + 1$ and allow $d_{\nu+1} = 2d_\nu = 2^{\nu+1} d_0$ further levels of subdivision. They introduce an additional error of $2^{\nu+1} d_0 n 2^{-q_{\nu+1}-1} \leq 2^{-p-\nu-3}$, so an error margin of $2^{-p-\nu-2} - 2^{-p-\nu-3} = 2^{-p-(\nu+1)-2}$ remains, and the process can repeat. The margin left for later arithmetical errors keeps shrinking to 2^{-p-3}, 2^{-p-4}, 2^{-p-5}, ..., but never reaches zero. With reference to a certain dispute in theoretical mechanics, we call this adaptive strategy after Zeno of Elea (5th century BC) the *Zeno trap* for arithmetical error.

How many extra bits does this adaptive strategy require, compared to the previously discussed static choice of q, assuming that we had a priori knowledge on the true subdivision depth? If the true subdivision depth is small, this depends in the obvious way on the choice of the initial estimate d_0. But suppose the true subdivision depth d is large, say $d = d_\nu + 1 = d_0 2^\nu + 1$, one more than the νth estimate. Had we known this bound in advance, we would have used precision $q_{\text{opt}} = p + \lceil \log(d_\nu + 1) n \rceil \approx p + \log n + \log d_0 + \nu$ through-

[11] In §3.3.8, we will prove the existence of suitable subdivision points, provided that the precision is sufficiently high. But that is only a tool in the analysis, it is not guaranteed that the algorithm ever reaches this sufficient precision; therefore, we cannot invoke this result here.

out. With our strategy, we have used precisions up to $q_{\nu+1} \approx p + \log n + \log d_0 + 2\nu + 2$, so $q_{\nu+1} - q_{\text{opt}} \approx \nu + 2$. However, ν is essentially the logarithm of subdivision depth d and thus bounded *doubly* logarithmically in the root separation of A_{in}. In the analysis of the bitstream Descartes algorithm, we will learn that p grows up to a value which, in expectancy, obeys a bound singly logarithmic in root separation. Therefore, we can hope that neither the $\log(nd)$ significand bits exceeding the "payload precision" p for the optimal choice of approximate precision nor the additional $\nu + 2$ bits incurred by our adaptive strategy have a substantial impact on significand lengths.

To incorporate the Zeno trap into our algorithm, we make the following changes to the algorithm as outlined in §3.3.5. We retain the global quantity q_0, but now define it as $q_0 = p + \lceil \log n \rceil + \log d_0 + 1$ (see above); this is the exponent for the approximate coefficients $b_i^{(0)} 2^{-q_0}$ of A_0. The entries of Q become sixtuples $((c, d), (b_i)_i, q, d, d_{\text{bd}}, V)$: The polynomial transformed for the interval (c, d) is given by the approximate coefficients $(b_i 2^{-q})_i$, whose exponent q (initially set to q_0) is now stored locally in each entry of Q. As before, V is the set of ε-approximate numbers of sign variations in the coefficients. The number d records the current subdivision depth; the number d_{bd} is the current bound on subdivision depth. If d meets d_{bd}, we make the necessary adjustments to q and $(b_i)_i$, then we continue with d_{bd} doubled and d reset to 0. The constant d_0 leaves some freedom for fine-tuning the Zeno trap. In our implementation, we have fixed it arbitrarily at $2^6 = 64$.

Let us summarize what we have achieved. Given a precision parameter p and approximations of Bernstein coefficients with absolute error bounded by 2^{-p-1}, we can provide approximations of all Bernstein coefficients arising from subdivision up to depth d with absolute error bounded by 2^{-p} by executing de Casteljau's algorithm in fixed-point arithmetic with precisions q somewhere between $p + \lceil \log n \rceil + 1$ and $p + \lceil \log n \rceil + 2\lceil \log d \rceil + 1$, where d does not have to be specified in advance.

3.3.7 The bitstream Descartes algorithm: pseudocode

In this section, we describe the bitstream Descartes algorithm in pseudocode. The following procedure combines the initialization steps from §3.3.2 with the algorithm as outlined in §3.3.5 and refined in §3.3.6. It accepts as input the real polynomial $A_{\text{in}}(X) = \sum_{j=0}^{n} a_j X^j$ of degree $n \geq 2$, whose coefficients are bitstreams and whose real roots must all be simple, and the integers r and l_n^- chosen to satisfy (3.21) and (3.22), respectively. The procedure reports isolating intervals for the real roots of $A_{\text{in}}(X)$ when it reaches line 46.

While the procedure is correct for any order of choosing from Q in line 18, the complexity analysis in §3.3.9 will only apply under some mild conditions, see Proposition 3.60.

1: **procedure** *DescartesE08basic*$((a_0, \ldots, a_n), r, l_n^-)$
2: $C \leftarrow 2^{4n}$; $d_0 \leftarrow 64$; $k \leftarrow 4 + \lceil \log n \rceil$; $K \leftarrow 2^k$; // global constants
3: initialize $(m_{ij}^{(n)})_{i,j=0}^{n}$; // basis conversion matrix computed per Lemma 3.39
4: $(c_0, d_0) \leftarrow (-2^{r+1}, +2^{r+1})$; // initial interval, (3.20)
5: $l \leftarrow l_n^- + \lfloor \log n! \rfloor + n(r+2)$; // offset to normalize leading coefficient, (3.27)
6: $p \leftarrow 6n + 1$; $\varepsilon \leftarrow 2^{-p}$; // precision parameter
7: $q' \leftarrow p + \lceil \log n! \rceil + \lceil \log(n+1) \rceil + 2$; // working precision for initialization
8: $q_0 \leftarrow p + \lceil \log n \rceil + \log d_0 + 1$; // initial working precision for subdivision
9: $N_{\text{try}} \leftarrow 0$; $N_{\text{fail}} \leftarrow 0$;

10: **for** j **from** 0 **to** n **do** $c_j \leftarrow [2^{j(r+1)-l+q'}a_j]$; **od**; // (3.30)

11: **for** i **from** 0 **to** n **do** $b_i^{(0)} \leftarrow \left\lfloor \left(\sum_{j=0}^{n} m_{ij}^{(n)} c_j \right) / 2^{q'-p-1} \right\rceil \cdot 2^{q_0-p-1}$; **od**; // (3.31)

12: $V_0 \leftarrow \mathrm{var}_\varepsilon(b_0^{(0)}2^{-q_0}, \ldots, b_n^{(0)}2^{-q_0})$; // set of ε-approx. numbers of sign variations

13: $P \leftarrow ()$; $Q \leftarrow \{\}$;

14: **if** $\min V_0 \geq 1$ **then** $P \leftarrow ((c_0, d_0))$; **fi**;

15: **if** $\min V_0 \geq 1 \wedge \max V_0 \geq 2$ **then** $Q \leftarrow \{((c_0, d_0),\ (b_i^{(0)})_i,\ q_0,\ 0,\ d_0,\ V_0)\}$; **fi**;

16: **while** $Q \neq \{\}$ **do**

17: // **Invariant:** (i) Q consists of sixtuples $((c, d),\ (b_i)_i,\ q,\ d,\ d_{\mathrm{bd}},\ V)$ such that $\sum_{i=0}^{n} b_i 2^{-q} B_i^n(X)$ approximates $2^{-l} n! A_{\mathrm{in}}((d-c)X + c)$ with absolute error up to $2^{-p-1}(2 - 2^{-\log(d_{\mathrm{bd}}/d_0)}) + dn2^{-q-1} < \varepsilon$ in each coefficient; such that $V = \mathrm{var}_\varepsilon((b_i 2^{-q})_i)$; and such that $|b_0 2^{-q}|, |b_n 2^{-q}| > C\varepsilon$.
(ii) Let $(c, d) \in P$ and $v = \mathit{DescartesTest}(A_{\mathrm{in}}, (c, d))$. It holds that $v \geq 1$. The number of entries $((c, d), \ldots) \in Q$ is at most one, exactly one if $v \geq 2$.

18: choose $((c, d),\ (b_i)_i,\ q,\ d,\ d_{\mathrm{bd}},\ V) \in Q$; // cf. Proposition 3.60 on page 98

19: **if** $d = d_{\mathrm{bd}}$ **then** // Zeno trap, §3.3.6

20: **for** i **from** 0 **to** n **do** $b_i \leftarrow 4b_i$; **od**;

21: $q \leftarrow q + 2$; $d \leftarrow 0$; $d_{\mathrm{bd}} \leftarrow 2d_{\mathrm{bd}}$;

22: **fi**;

23: *done* $\leftarrow$ false;

24: **while** $\neg$*done* **do** // repeatedly attempt subdivision

25: choose $u \in \{K/4,\ K/4 + 1,\ \ldots,\ 3K/4\}$ uniformly at random; // (3.37)

26: $((b_i^L)_i,\ (b_i^R)_i) \leftarrow \mathit{DeCasteljauApprox}((b_i)_i,\ u,\ k)$; // $\alpha = u/K$

27: $N_{\mathrm{try}} \leftarrow N_{\mathrm{try}} + 1$;

28: **if** $|b_0^R 2^{-q}| > C\varepsilon$ **then** // subdivision has succeeded

29: $m \leftarrow ((K-u)c + ud)/K$; $I_L \leftarrow (c, m)$; $I_R \leftarrow (m, d)$;

30: $V_L \leftarrow \mathrm{var}_\varepsilon((b_i^L 2^{-q})_i)$; $V_R \leftarrow \mathrm{var}_\varepsilon((b_i^R 2^{-q})_i)$;

31: in P, replace entry (c, d) by subsequence $(I_s \mid s \in (L, R),\ \min V_s \geq 1)$;

32: in Q, replace element $((c, d), \ldots)$ by elements
$\{(I_s,\ (b_i^s)_i,\ q,\ d+1,\ d_{\mathrm{bd}},\ V_s) \mid s \in \{L, R\},\ \min V_s \geq 1,\ \max V_s \geq 2\}$;

33: *done* $\leftarrow$ true;

34: **else** // subdivision has failed

35: $N_{\mathrm{fail}} \leftarrow N_{\mathrm{fail}} + 1$;

36: **if** $N_{\mathrm{fail}} \geq 2 \wedge N_{\mathrm{fail}} \geq N_{\mathrm{try}}/2$ **then** // switch to higher precision

37: $p \leftarrow 2p$;

38: ⟨reinitialize Q from new p⟩

39: **if** $\neg$*done* **then**

40: re-fetch current entry $((c, d),\ (b_i)_i,\ q,\ d,\ d_{\mathrm{bd}},\ V)$ from Q;

41: **fi**;

42: **fi**;

43: **fi**;

44: **od**;

45: **od**;

46: report sequence P of isolating intervals;

47: **end procedure**;

After a higher precision p has been chosen, the elements of Q are reinitialized as follows.

$\langle$reinitialize Q from new $p\rangle \equiv$

1: $\varepsilon \leftarrow 2^{-p}$;
2: $q' \leftarrow p + \lceil \log n! \rceil + \lceil \log(n+1) \rceil + 2$; // as above
3: $q_0 \leftarrow p + \lceil \log n \rceil + \log d_0 + 1$; // as above
4: $N_{\text{try}} \leftarrow 0$; $N_{\text{fail}} \leftarrow 0$;
5: **for** j **from** 0 **to** n **do** $c_j \leftarrow \lfloor 2^{j(r+1)-l+q'} a_j \rfloor$; **od**; // as above
6: **for** i **from** 0 **to** n **do** $b_i^{(0)} \leftarrow \left\lfloor \left(\sum_{j=0}^{n} m_{ij}^{(n)} c_j \right) / 2^{q'-p-1} \right\rfloor \cdot 2^{q_0-p-1}$; **od**; // as above
7: **for each** $((c,d), (b_i)_i, q, d, d_{\text{bd}}, V)$ **in** Q **do**
8: let $c = c'2^{-h}$, $d = d'2^{-h}$ with $c', d', h \in \mathbb{Z}$; // significands and common exponent
9: $(*, (b_i^{(1)})_i) \leftarrow \mathit{DeCasteljauApprox}((b_i^{(0)})_i, c' + 2^{h+r+1}, h + r + 2)$; // $\alpha = \frac{c-c_0}{d_0-c_0}$
10: $((b_i)_i, *) \leftarrow \mathit{DeCasteljauRatApprox}((b_i^{(1)})_i, d' - c', 2^{h+r+1} - c')$; // $\alpha = \frac{d-c}{d_0-c}$
11: $q \leftarrow q_0$; $d \leftarrow 2$; $d_{\text{bd}} \leftarrow d_0$; // reset Zeno trap
12: $V \leftarrow \text{var}_\varepsilon((b_i 2^{-q})_i)$;
13: **if** $\max V \geq 2$ **then**
14: write modified entry $((c,d), (b_i)_i, q, d, d_{\text{bd}}, V)$ back to Q;
15: **else** // $V = \{1\}$
16: discard entry $((c,d), \ldots)$ in Q; // (c,d) is retained in P as isolating interval
17: *done* $\leftarrow$ true;
18: **fi**;
19: **od**;

3.3.8 On sufficient precision

This entire section is devoted to the proof of the following theorem.

Theorem 3.48. *Let the polynomial A_{in} of degree n and the initial interval I_0 be as above. Let m denote the maximum multiplicity of any complex root of A_{in}. Let $0 < s < |I_0|$ be a lower bound on the distance between any two distinct complex roots of A_{in}. Let $0 < w \leq s$ be a lower bound on the length of any interval subdivided by the bitstream Descartes algorithm. If the precision parameter p satisfies*

$$p \geq p_{\text{ok}} := \log \frac{|I_0|^n}{w^m s^{n-m}} + 2m \log n + 4n + 6m + 1 = O(n \cdot (\log \frac{|I_0|}{w} + \log n)), \quad (3.38)$$

then the probability of the algorithm choosing a higher precision is no more than 1/10.

Before embarking on the proof, we record a simple bound in terms of p_{ok} for later reference.

Lemma 3.49. *Consider the subdivision tree $\mathcal{T}'$ constructed by an execution of the bitstream Descartes algorithm, and let h denote its height. It holds that $nh = O(p_{\text{ok}})$.*

Proof. $\mathcal{T}'$ contains an internal node I at depth $h-1$ whose two children are leaves. We have $w \leq |I| \leq (3/4)^{h-1} |I_0|$, so $h \leq \log(|I_0|/w)/\log(4/3) + 1 = O(\log(|I_0|/w))$. □

As a first step towards the proof of Theorem 3.48, we formulate a general theorem on the distance of roots of two polynomials and a consequence concerning the proximity of roots to values of small magnitude.

Theorem 3.50. *Let $F(X) = \sum_{i=0}^{n} f_i X^i$ be a complex polynomial of degree $n > 0$. Let $G(X) = g_n \prod_{i=1}^{k} (X - z_i)^{m_i}$ be another complex polynomial of degree n with pairwise distinct roots $z_1, \ldots, z_k$. Let $H(X) = F(X)/G(X) = f_n/g_n + \sum_{i=1}^{k} \sum_{j=1}^{m_i} p_{ij}(X - z_i)^{-j}$. If $F(\zeta) = 0$, then there exists a pair (i, j) with $1 \le i \le k$, $1 \le j \le m_i$ such that*

$$|\zeta - z_i| \le \sqrt[j]{\left| \frac{n g_n p_{ij}}{f_n} \right|}. \tag{3.39}$$

In case $j = 1$, it even holds that

$$|\zeta - (z_i - \frac{n g_n p_{i1}}{2 f_n})| \le \left| \frac{n g_n p_{i1}}{2 f_n} \right|. \tag{3.40}$$

The proof of the theorem uses the following lemma.

Lemma 3.51. *For $\alpha > 0$ and $u, v \in \mathbb{C}$, $v \neq 0$, it holds that*

$$\operatorname{Re}\left(\alpha + \frac{u}{v}\right) \le 0 \iff |v + \frac{u}{2\alpha}| \le \left| \frac{u}{2\alpha} \right|. \tag{3.41}$$

Proof. We have $\overline{v} v \operatorname{Re}(\alpha + u/v) = \overline{v} v \cdot (\alpha + u/(2v) + \overline{u}/(2\overline{v})) = \alpha \overline{v} v + \overline{v} u/2 + \overline{u} v/2 = \alpha \cdot (|v + u/(2\alpha)|^2 - |u/(2\alpha)|^2)$, so $\operatorname{Re}(\alpha + u/v) \le 0 \Leftrightarrow |v + u/(2\alpha)| \le |u/(2\alpha)|$. □

Proof of Theorem 3.50. If $\zeta = z_i$ for some $1 \le i \le k$, there is nothing to be shown. Otherwise, we have $H(\zeta) = 0$ and can conclude

$$0 = \operatorname{Re}\left(\frac{n g_n}{f_n} H(\zeta)\right) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \operatorname{Re}\left(1 + \frac{n g_n p_{ij}}{f_n (\zeta - z_i)^j}\right).$$

There is an index pair (i, j) for which the summand on the right-hand side is non-positive. By Lemma 3.51, it follows that $|(\zeta - z_i)^j + n g_n p_{ij}/(2 f_n)| \le |n g_n p_{ij}/(2 f_n)|$. This yields the claim for $j = 1$. For the general case $j \ge 1$, we observe $|\zeta - z_i|^j \le |n g_n p_{ij}/f_n|$. □

This proof technique and the statement for $j = 1$ are due to Neumaier [Neu03, Thm. 3.2], who used it as a first step in computing an a posteriori bound on the error made in approximating the complex roots of $F(X)$ by the distinct numbers $z_1, \ldots, z_n$. Neumaier also makes the usual homotopy argument to show that the number of discs forming a connected component is the number of roots in that component.

We proceed to give a more explicit form of the coefficients p_{ij} in the partial fraction decomposition of $H(X)$. This is well-known; a proof appears, e.g., in [Hen74, p. 555].

Proposition 3.52. *With F and G as in Theorem 3.50, let $Q_i(X) = g_n \prod_{\ell \neq i} (X - z_\ell)^{m_\ell}$ and*

$$p_{ij} = \frac{1}{(m_i - j)!} \left[\frac{F(X)}{Q_i(X)} \right]_{X = z_i}^{(m_i - j)} \qquad \text{for} \quad 1 \le i \le k, \quad 1 \le j \le m_i. \tag{3.42}$$

Then $F(X)/G(X) = f_n/g_n + \sum_{i=1}^{k} \sum_{j=1}^{m_i} p_{ij}(X - z_i)^{-j}$.

The symbol $[\,\cdot\,]_{X=x_0}^{(i)}$ denotes the ith derivative evaluated at $X = x_0$.

If $m_i = 1$, then (3.42) reduces to $p_{i1} = F(z_i)/G'(z_i)$, and (3.39) turns into a well-known bound that is usually derived from Gershgorin's theorem, we refer to [Neu03] for discussion and references.

We only need a particular consequence of Theorem 3.50, namely a quantitative version of the notion that a polynomial's values are "small" in magnitude only at points "close" to a root.

Theorem 3.53. *Let $G(X) = g_n \prod_{i=1}^{k} (X - z_i)^{m_i}$ be a complex polynomial of degree n with pairwise distinct roots $z_1, \ldots, z_k$. For every $\zeta \in \mathbb{C}$ there exists a pair (i, j) with $1 \le i \le k$, $1 \le j \le m_i$ such that*

$$|\zeta - z_i| \le \sqrt[j]{n \binom{n-1-j}{n-1-m_i} \frac{|G(\zeta)|}{|g_n|\, s_i^{n-j}}} \qquad \text{where} \quad s_i = \min_{\ell \ne i} |z_i - z_\ell| \,. \tag{3.43}$$

Proof. Theorem 3.50 and Proposition 3.52 applied to $F(X) := G(X) - G(\zeta)$ yield a pair (i, j) such that

$$|\zeta - z_i| \le \sqrt[j]{\frac{n}{(m_i - j)!} \left| \left[\frac{F(X)}{Q_i(X)} \right]_{X=z_i}^{(m_i-j)} \right|}. \tag{3.44}$$

With the Leibniz rule, we find

$$\begin{aligned} \left[F(X) \cdot \frac{1}{Q_i(X)} \right]_{X=z_i}^{(m_i-j)} &= \sum_{\nu=0}^{m_i-j} \binom{m_i - j}{\nu} [F(X)]_{X=z_i}^{(\nu)} \left[\frac{1}{Q_i(X)} \right]_{X=z_i}^{(m_i-j-\nu)} \\ &= F(z_i) \left[\frac{1}{Q_i(X)} \right]_{X=z_i}^{(m_i-j)} = -G(\zeta) \left[\frac{1}{Q_i(X)} \right]_{X=z_i}^{(m_i-j)}, \end{aligned} \tag{3.45}$$

where the second equality holds because the m_i-fold root z_i of G is also a root of $F' = G'$, $F'' = G''$, …, $F^{(m_i-1)} = G^{(m_i-1)}$.

The function $1/Q_i(X)$ is of the form $1/(g_n \prod_\nu (X - z_{\ell_\nu}))$ with an index sequence $(\ell_\nu)_\nu$ of length $n - m_i$ such that $\ell_\nu \ne i$ for any ν. The derivative of $1/Q_i(X)$ is

$$\left[\frac{1}{g_n \prod_\nu (X - z_{\ell_\nu})} \right]' = \frac{-g_n \sum_\mu \prod_{\nu \ne \mu} (X - z_{\ell_\nu})}{g_n^2 \prod_\nu (X - z_{\ell_\nu})^2} = -\sum_\mu \frac{1}{g_n (X - z_{\ell_\mu}) \prod_\nu (X - z_{\ell_\nu})},$$

that is, the derivative is (up to a minus sign) the sum of $n - m_i$ functions of the same form as $1/Q_i(X)$, with index sequences of length $n - m_i + 1$. Continuing up to the derivative of order $m_i - j$, we obtain a sum of $(n-m_i)(n-m_i+1)\cdots(n-j-1) = (n-1-j)!/(n-1-m_i)!$ fractions, each of which has $n - j$ factors $X - z_\ell$, $\ell \ne i$, in the denominator. The value at $X = z_i$ is therefore bounded in terms of the minimum distance to the other z_ℓ as follows:

$$\left| \left[\frac{1}{Q_i(X)} \right]_{X=z_i}^{(m_i-j)} \right| \le \frac{(n-1-j)!}{(n-1-m_i)!} \cdot \frac{1}{|g_n|\, s_i^{n-j}}.$$

Equation (3.44) combined with (3.45) and this estimate yields the claim. □

The second step in our proof of Theorem 3.48 employs Theorem 3.53 to bound the probability that a subdivision parameter $\alpha = u/K$ chosen at random as in (3.37) fails.

Proposition 3.54. *Under the conditions of Theorem 3.48, subdivision of an interval I at the point m fails only if there is complex root of A_{in} within the open disc around m of diameter w/K.*

Proof. We recall from §3.3.2 that *DescartesE08basic* does not operate directly on A_{in}, but on the polynomial A_0, which has been constructed from A_{in} by transforming the initial interval I_0 to $(0,1)$ and approximately normalizing the leading coefficient $a_n^{(0)}$; in particular, $|a_n^{(0)}| \geq 1$. Suppose $z_1, \ldots, z_k$ are the distinct complex roots of $A_0(X)$. Let us write $\hat{m}$ for the point m transformed to the argument space of A_0. Subdivision at m fails only if the approximate value of A_0 at $\hat{m}$, known as an ε-approximate Bernstein coefficient, is $C\varepsilon$ or less, which in turn is possible only if the exact value $A_0(\hat{m})$ is $(C+1)\varepsilon$ or less in magnitude. To prove the proposition, we will show for any $\zeta \in \mathbb{C}$ that $|A_0(\zeta)| \leq (C+1)\varepsilon$ implies $|\zeta - z_i| < \hat{w}/(2K)$ for some root z_i, where $\hat{w} := w/|I_0|$.

We write $\hat{s}_i$ for the minimum distance from the ith complex root of $A_0(X)$ to any other. The minimum distance $\hat{s}$ between any two distinct complex roots of $A_0(X)$ is $\hat{s} = \min_i \hat{s}_i = s/|I_0|$. By Theorem 3.53 applied to $G(X) = A_0(X)$, all points ζ for which $|G(\zeta)| \leq (C+1)\varepsilon$ lie in closed discs around the roots z_i of A_0 with radii

$$\begin{aligned} R_{ij} &\leq \sqrt[j]{n\binom{n-1-j}{n-1-m_i}\frac{(C+1)\varepsilon}{\hat{s}_i^{n-j}}} \qquad \text{for} \quad 1 \leq j \leq m_i \\ &\leq \sqrt[j]{n\binom{n-2}{m_i-1}\frac{(C+1)\varepsilon}{\hat{s}_i^{n-j}}} \qquad \begin{array}{l}\text{where } n-2 \geq n-1-j \\ \text{and } m_i - 1 = (n-2) - (n-1-m_i).\end{array} \end{aligned}$$

We will now choose p sufficiently large in order for any such radius to be less than $\hat{w}/(2K)$. We observe

$$\begin{aligned} &\sqrt[j]{n\binom{n-2}{m_i-1}\frac{(C+1)\varepsilon}{\hat{s}_i^{n-j}}} < \frac{\hat{w}}{2K} \\ \iff & n\binom{n-2}{m_i-1}\frac{(2^{4n}+1)2^{-p}}{\hat{s}_i^{n-j}} < \left(\frac{\hat{w}}{2K}\right)^j \\ \iff & 2^{-p} < \frac{\hat{w}^j \hat{s}_i^{n-j}}{(2^{4n}+1)\cdot 2^{j(5+\lceil \log n \rceil)} \cdot n\binom{n-2}{m_i-1}} \\ \iff & p > \log \frac{1}{\hat{w}^j \hat{s}_i^{n-j}} + \log(2^{4n}+1) + j(5+\lceil \log n \rceil) + \log(n\tbinom{n-2}{m_i-1}) \\ \impliedby & p \geq \log \frac{1}{\hat{w}^j \hat{s}_i^{n-j}} + 2m_i \log n + 4n + 6m_i + 1 \\ \impliedby & p \geq \log \frac{1}{\hat{w}^m \hat{s}^{n-m}} + 2m \log n + 4n + 6m + 1, \end{aligned}$$

where the last line uses $\hat{w} \leq \hat{s} \leq \hat{s}_i$ and $m \geq m_i$. Since $\hat{w} = w/|I_0|$ and $\hat{s} = s/|I_0|$, we have arrived precisely at the condition (3.38) imposed on p in Theorem 3.48. Thus, under the conditions of Theorem 3.48, the claimed proximity statement holds. □

Corollary 3.55. *Under the conditions of Theorem 3.48, the probability of any one attempted subdivision to fail is less than* 1/8.

Proof. When the procedure *DescartesE08basic* attempts to subdivide an interval I, it chooses a subdivision point m uniformly at random from $K/2 + 1 > 8n$ candidates on an evenly spaced grid of width $|I|/K$. The open discs with diameter $w/K \leq |I|/K$ around the candidate points are pairwise disjoint, so each complex root of A_{in} can fulfill

the condition of the preceding proposition for at most one candidate. Thus, at most n of more than $8n$ possible choices lead to a failure. □

We have now bounded the failure probability of a single subdivision attempt at precision $p \geq p_{\text{ok}}$. For the third and last step towards Theorem 3.48, it remains to bound the probability that despite $p \geq p_{\text{ok}}$, subdivision attempts fail so often that we increase the precision further. This step is entirely stochastic and has nothing to do with the primary subject matter of our thesis; an impatient reader might skip over it.

As a stochastic model of precision management in the bitstream Descartes algorithm, we consider an infinite sequence of independent coin tosses, each of which produces a "failure" or a "success"; the failure probability of each coin is bounded by a constant $\rho < 1/2$. (Later on, we will of course set $\rho = 1/8$ with reference to Corollary 3.55.)

For any $n \in \mathbb{N}$, let k_n denote the number of failures among the first n coin tosses. We say that the sequence *dies at position* n if $n = \min\{m \in \mathbb{N} \mid k_m \geq 2 \text{ and } 2k_m \geq m\}$, see line 36 in procedure *DescartesE08basic*. We say that the sequence *dies* if it dies at position n for some n.

Proposition 3.56. *The probability of dying is bounded by*

$$\frac{8\rho^3(1-\rho)^3}{(1-2\rho)^2} + 6\rho^2 - 8\rho^3 + 3\rho^4.$$

The proof of this proposition rests on the following two lemmas.

Lemma 3.57. *Let $n \geq 5$. If a sequence dies at position n, then n is even and $k_n = n/2$ and the coin tosses n and $n-1$ were both failures.*

Proof. Suppose the sequence dies at position $n \geq 5$.

We begin by showing that $2k_n = n$ to establish the first two claims. Clearly, we have $2k_n \geq n$, so that $k_n \geq \lceil n/2 \rceil \geq 3$. By minimality of n, the coin toss n was a failure. Hence we have $k_{n-1} = k_n - 1 \geq 2$. This and the minimality of n imply $2k_n - 2 = 2k_{n-1} < n - 1$ and thus $2k_n \leq n$, as desired.

It remains to show that not only coin toss n but also toss $n-1$ failed. If toss $n-1$ was a success, then $k_{n-2} = k_n - 1 \geq 2$ and $2k_{n-2} = 2k_n - 2 \geq n - 2$, which is a contradiction to the minimality of n. □

Lemma 3.58. *If a sequence dies, then it satisfies at least one of the following conditions:*

(a) There are at least two failures among the first four coin tosses.

(b) There exists $\ell \in \mathbb{N}$, $\ell \geq 3$, such that there are exactly $\ell - 2$ failures and ℓ successes among the first $2\ell - 2$ coin tosses, and the tosses $2\ell - 1$ and 2ℓ are both failures.

Proof. Let the sequence in question die at position n. Clearly, $n \geq 2$. We distinguish two cases. If $n \leq 4$, then condition (a) holds. If $n \geq 5$, then, by Lemma 3.57, n is even and thus at least 6, and condition (b) holds with $\ell = n/2$. □

We can now prove the proposition by bounding the probability of the sequences that satisfy (a) or (b).

Proof of Proposition 3.56. Distinguishing the cases of exactly two, three or four failures among the first four coin tosses, we find that the probability of (a) is bounded by

$$\binom{4}{2}\rho^2(1-\rho)^2 + \binom{4}{3}\rho^3(1-\rho) + \binom{4}{4}\rho^4 = 6\rho^2 - 8\rho^3 + 3\rho^4.$$

The probability of (b) for a fixed $\ell \geq 3$ is at most

$$\binom{2\ell-2}{\ell-2}\rho^{\ell-2}(1-\rho)^{\ell}\cdot\rho^2 \leq 2^{2\ell-3}\rho^{\ell}(1-\rho)^{\ell} = \frac{1}{8}\big(4\rho(1-\rho)\big)^{\ell},$$

where the inequality uses the estimate $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \leq 2^{n-1}$. Hence the probability of (b) being satisfied for any $\ell \geq 3$ is bounded by

$$\sum_{\ell=3}^{\infty}\frac{1}{8}\big(4\rho(1-\rho)\big)^{\ell} = \frac{1}{8}\big(4\rho(1-\rho)\big)^{3}\sum_{\ell=0}^{\infty}\big(4\rho(1-\rho)\big)^{\ell} = \frac{8\rho^3(1-\rho)^3}{1-4\rho(1-\rho)}.$$

The geometric series converges indeed as claimed, because $4\rho(1-\rho) > 0$ and

$$\rho \neq 1/2 \implies 0 < (1-2\rho)^2 = 1-4\rho(1-\rho) \implies 4\rho(1-\rho) < 1.$$

The claim follows by adding the bounds for (a) and (b) and simplifying the denominator according to the preceding equality. □

We can now combine the results of the three steps and prove the result announced at the beginning of this section.

Proof of Theorem 3.48. We suppose that the precision parameter p satisfies (3.38). By Corollary 3.55, the failure probability of each subdivision attempt is less than $\rho := 1/8$. The bitstream Descartes algorithm chooses a higher precision precisely if $N_{\text{fail}} \geq 2$ and $N_{\text{fail}} \geq N_{\text{try}}/2$. Substituting ρ into Proposition 3.56, we find that the probability of this condition is no more than $3593/36864 = 0.097466... < 1/10$. □

3.3.9 Complexity analysis

In this section, we analyze the computing time and coefficient precision needed by the bitstream Descartes algorithm.

What does the algorithm do? It extracts approximations of the input coefficients out of the bitstreams representing them, converts them to the Bernstein basis, and transforms the Bernstein coefficients by repeated de Casteljau subdivision. The essential data processed by the algorithm are the approximate transformed coefficients. By contrast, the parameters r and l_n^- of procedure *DescartesE08basic* only play a role at the interface of the algorithm to the outside world: they determine offsets for the precision parameters passed to the coefficient bitstreams, and r also determines relative to which exponent the endpoints of constructed subintervals are to be understood. We could have elided these parameters and formulated the algorithm for the special case $l_n^- = r = 0$, requiring the caller to normalize the input such that $|a_n| \in [1, 4)$ and $I_0 = (-2, +2)$. This, however, would be inconvenient for users of an implementation, so we have refrained from such a normalization requirement. In the complexity analysis, however, we need to make a weak form of this requirement: We assume that the bit length of r and l_n^- is sufficiently short such that the cost of handling them is dominated by the operations on coefficients. This is certainly true for actual implementations, where r and l_n^- can be represented by machine-word integers for all realistic inputs. Without this condition, an adversary could enforce an arbitrarily high running time for any input, simply by scaling the coefficients, $A_{\text{in}}(X) \rightsquigarrow \lambda A_{\text{in}}(X)$, or by scaling the indeterminate, $A_{\text{in}}(X) \rightsquigarrow \lambda^n A_{\text{in}}(X/\lambda)$, with a factor $\lambda = 2^k$, $k \in \mathbb{Z}$, such that the integer l_n^- or r becomes sufficiently long.

In a similar spirit, we make the usual assumption that we can indeed, at insignificant cost, make independent random choices of numbers u of length $O(\log n)$ whenever we determine a random subdivision point in line 25 of procedure *DescartesE08basic*.

For our complexity analysis, we need again a size measure of the subdivision tree. The following definition captures the properties that we rely upon in the subsequent analysis. If we were discussing the exact Descartes method, then the sum P of the length of all rt-paths from Theorem 3.19(iii) on page 60 would do, but here we will need to look a bit closer.

Definition 3.59. The number P' is a *tsqd-bound* for the bitstream Descartes algorithm on a polynomial A_{in} of degree n and an initial interval I_0 with a specific policy of choosing from Q, if it holds for any possible choice of subdivision points that:

(i) The subdivision tree $\mathcal{T}'$ constructed has at most $P' + 1$ internal nodes.

(ii) If $I_1, \ldots, I_q$ are the intervals recorded in Q at any stage of the algorithm, then the sum of the depths of $I_1, \ldots, I_q$ in $\mathcal{T}'$ is at most P'.

(iii) $P' \geq n/2$.

Condition (iii) is imposed for technical reasons and is no substantial restriction: If A_{in} has n real roots, then the binary tree $\mathcal{T}'$ has at least n leaves and $n - 1$ internal nodes.

In order to get a tsqd-bound P' that satisfies $P' = O(P)$, we need a traversal order of the subdivision tree that avoids larger accumulations of (M1)-intervals in Q. The following proposition gives two sufficient conditions on traversal orders to achieve this, but it is not particularly ambitious about the constant factors involved.

Proposition 3.60. *Let A_{in} be a real polynomial with degree $n \geq 2$, all of whose real roots are simple, and let I_0 be an initial interval for the bitstream Descartes algorithm on A_{in}. Let $P \geq n/4$ be an upper bound on the sum of the lengths of all rt-paths in a subdivision tree constructed by the exact Descartes method on A_{in} and I_0 with any possible choice of subdivision points. P gives rise to tsqd-bounds for the bitstream Descartes algorithm on A_{in} and I_0 as follows.*

(i) If entries are chosen from Q with priority given to those of type (M1), but otherwise in arbitrary order, then $P' = 3P + 2 = O(P)$ is a tsqd-bound.

(ii) If entries are chosen from Q in first-in–first-out order (i.e., Q is a queue), then $P' = 4P + 4n = O(P)$ is a tsqd-bound.

Proof. Consider any of the possible subdivision trees $\mathcal{T}'$ constructed by the bitstream Descartes algorithm. Let $\mathcal{T}$ denote the subdivision tree of the exact Descartes algorithm when making the same choices of subdivision points.

The binary tree $\mathcal{T}$ has no more than $P + 1$ internal nodes, cf. Theorem 3.19(ii/iv), and thus no more than $2P + 3$ nodes in total. Every internal node I of $\mathcal{T}'$ is another node's parent and thus occurs as a node in $\mathcal{T}$ by Lemma 3.47 from page 89. Therefore, the numbers P' in both claims satisfy condition (i) from Definition 3.59. It is immediate that they satisfy condition (iii) as well. It remains to establish condition (ii). Here, we need to argue separately for claim (i) and claim (ii).

Ad (i). If priority is given to subdivision of intervals of type (M1), then Q has at most two (M1)-entries at any time: To see this inductively, we note that subdivision of an (M1)-entry produces at most one new (M1)-entry, because two subintervals with odd Descartes test can only arise from an interval with even Descartes test by the variation-diminishing

property, Proposition 2.26 on page 26. On the other hand, subdivision of a (D2)-entry may produce two (M1)-entries, but occurs only if no (M1)-entries are present yet.

We are now ready to bound the sum of depths in Q. Let us first discuss those entries of Q that have type (D2). They occur as internal nodes in $\mathcal{T}$, because their true Descartes test value is at least 2. As any two of them are disjoint, each of them sits on a separate rt-path, and the sum of their depths is bounded by P. Now we turn to an entry I of type (M1), if any. By Lemma 3.46, its parent J occurs as an internal node in $\mathcal{T}$ and thus has a depth of at most P. Therefore, the depth of I is at most $P + 1$. Since there are at most two entries I of type (M1), the sum of all depths is bounded by $3P + 2 = P'$, as needed to be shown.

Ad (ii). Maintaining Q as a queue leads to breadth-first traversal of the subdivision tree. Let us inspect Q at a point of time when all its entries have the same depth d. The parents of all entries occur as internal nodes in $\mathcal{T}$ on level $d - 1$, hence they are at most $n/2$ in number by the variation-diminishing property, and the sum of their depths is at most P. Thus, Q has at most n entries, and the sum of their depths is no more than $2P+n$. When subdividing entries in Q to explore the next level of the subdivision tree, we replace each entry with at most two new entries, each of depth one larger. It follows that during the exploration of level $d+1$ in the subdivision tree, the sum of depths of all entries in Q is at most $4P + 4n = O(P)$. (This estimate is, of course, overly pessimistic, since it allows for two additional levels of subdivision, but that is still good enough to attain $P' = O(P)$.) □

If the conditions imposed on Q in this proposition are met, then the bound on P provided by Equation (3.11) in Theorem 3.19 on page 60 for the exact Descartes method immediately translates to a tsqd-bound P' for the bitstream Descartes algorithm with the same order of growth. Since we invoke Theorem 3.19 with the subdivision ratio bound $\rho = 4/3$, the condition $P \geq n/4$ is satisfied automatically.

With a tsqd-bound in our hands, we can now proceed to the main result.

Theorem 3.61. *Consider a real polynomial A_{in} of degree $n \geq 2$, all of whose real roots are simple, and an initial interval I_0 as in (3.20) and (3.21). Let $0 < s < |I_0|$ be a lower bound for the distance between any two distinct complex roots of A_{in}. Let $P' > 0$ be a tsqd-bound for the bitstream Descartes algorithm on A_{in} and I_0. The expected value of the number of bit operations performed by the randomized algorithm DescartesE08basic invoked for A_{in} and I_0 is*

$$O(n^3 \log n \cdot P' \cdot (\log \frac{|I_0|}{s} + \log n)). \tag{3.46}$$

In preparation of the proof, we derive two lemmas.

Lemma 3.62. *There exists a lower bound $0 < w \leq s$ on the length of any interval subdivided by DescartesE08basic such that $\log(|I_0|/w) = O(\log(|I_0|/s))$.*

Proof. We show that any interval I subdivided has a length $|I| > w := \sqrt{3}/8 \cdot s$. For $I = I_0$, this is immediate, so we may assume that I has a parent J in the subdivision tree $\mathcal{T}'$ of the bitstream Descartes algorithm. Clearly, $|I| \geq |J|/4$. Lemma 3.47 tells us that the exact Descartes method would subdivide J. By Proposition 3.13 on page 58, this entails $|J| > \sqrt{3}/2 \cdot s$, and the claim follows. □

Lemma 3.63. *For a given value of the precision parameter p, consider an approximate Bernstein coefficient sequence $(b_i 2^{-q})_i$ that has been produced by a succession of d de Cas-*

teljau subdivisions since the last initialization at a new precision. The significands $b_i \neq 0$ *have bit lengths bounded as* $\log |b_i| \leq p + 2\log d + O(1)$.

Proof. By Proposition 3.40, the exact coefficients of A_0 satisfy $|\beta_i^{(0)}| < 8 \cdot (3/4)^n$. The exact coefficients β_i that are approximated by $b_i 2^{-q}$ arise from $(\beta_i^{(0)})_i$ by a succession of convex combinations, so $|\beta_i| < 8 \cdot (3/4)^n$ as well. Consider an approximate coefficient $b_i 2^{-q}$ with $|b_i| \geq 3$. It satisfies $(|b_i|-1)2^{-q} \leq |\beta_i| < 2^{-n\log(4/3)+3}$ and $\log(|b_i|-1) < q - n\log(4/3) + 3$. According to the discussion at the end of §3.3.6, we have $q < p + \log n + 2\log d + 4$. Combining these two bounds and dropping the terms $\log n - n\log(4/3)$ yields the claim. □

Proof of Theorem 3.61. We consider one possible execution of *DescartesE08basic* for A_{in} and I_0, resulting in a subdivision tree $\mathcal{T}'$.

The precision parameter p of the algorithm takes successive values $p_\mu = 2^\mu p_0$, $\mu \in \mathbb{N}_0$. Theorem 3.48 combined with Lemma 3.62 gives a threshold $p_{\text{ok}} = O(n \cdot (\log(|I_0|/s) + \log n))$ beyond which precision increments are unlikely. More specifically, let $\mu_0 \in \mathbb{N}_0$ be the smallest index such that $p_{\mu_0} \geq p_{\text{ok}}$. The probability of p having reached a value p_μ with $\mu > \mu_0$ is less than $(1/10)^{\mu-\mu_0}$, so with probability 1, the precision p has reached a maximal value p_{μ_1} and the algorithm has terminated. We define $p_{\max} = \max\{p_{\mu_0}, p_{\mu_1}\}$. Clearly, it holds that $p_{\max} \geq p_{\text{ok}}$.

Let us show now that the algorithm has performed no more than

$$O(n^2 \log n \cdot P' \cdot p_{\max}) \tag{3.47}$$

bit operations. For this purpose, we divide the work it has done into three parts:

- *Initialization.* This covers the initializations in lines 2 to 5 and lines 13 to 15 of procedure *DescartesE08basic*, in particular computing the basis conversion matrix $(m_{ij}^{(n)})_{ij}$, but it excludes the computation of the $b_i^{(0)}$.
- *New precision.* This accounts for the effort of introducing the initial precision in lines 6 to 12 of procedure *DescartesE08basic*, and for switching to a higher precision in the body of the **if**-statement at line 36, including the whole block ⟨reinitialize Q from new p⟩.
- *Subdivision.* This comprises the entire main loop (starting in line 16), in particular, the invocations of de Casteljau's algorithm in line 26, but excludes switching to a higher precision.

We assume that the sequence P and the set Q are implemented in a straightforward way, such as discussed in §3.1.3 and §3.3.5, so that access to them has negligible cost.

We begin with the initialization. Lemma 3.39 specifies the cost of computing the $m_{ij}^{(n)}$ as $O(n^4 \log n)$; since $P' = \Omega(n)$ and $p_{\max} = \Omega(n)$, this is covered by (3.47), as are the other quantities whose bit lengths depend on n. (Regarding l_n^- and r, we refer to our discussion at the beginning of this section.)

Let us now consider the introduction of a new precision $p_\mu = 2^\mu p_0 \leq p_{\max}$ for $\mu \in \mathbb{N}_0$. This has two parts: updating A_0 and, if $\mu \geq 1$, updating the entries of Q. We will show that both have used no more than

$$O(n^2 \log n \cdot P' \cdot (p_\mu + \log(nh))) \tag{3.48}$$

bit operations, where h denotes height of the subdivision tree $\mathcal{T}'$. (The occurrence of h may appear spurious, but it gives us some leeway for a charging argument later on.)

Since $P' = \Omega(n)$, the bound $O(n^3 \log n \cdot (p_\mu + \log n))$ from Proposition 3.41 for computing A_0 at precision p_μ is covered by (3.48). What about the cost of updating Q? Consider the intervals $I_1, \ldots, I_q$ recorded in Q, and let δ_i, $1 \le i \le q$, denote their respective depths in the subdivision tree $\mathcal{T}'$. By condition (ii) in Definition 3.59, $\sum_i \delta_i \le P'$. For each interval I_i recorded in Q, we perform one call of *DeCasteljauApprox* and one call of *DeCasteljauRatApprox* to update it. By Lemma 3.63, the length of the coefficient significands is $O(p_\mu)$, but how long is the subdivision parameter? In each of the δ_i subdivision steps that have created I_i, the interval boundaries had their significands enlarged by k bits and their exponents correspondingly decreased by k, where $k = O(\log n)$, see (3.37). Thus, the subdivision parameters have lengths $O(\delta_i \log n)$, and the cost of these two subdivisions is $O(n^2 \cdot \delta_i \log n \cdot p_\mu)$ according to Propositions 3.42 and 3.43. Summing over all elements of Q, we arrive at a cost of updating Q to precision p_μ of $\sum_i \delta_i \cdot O(n^2 \log n \cdot p_\mu) = O(n^2 \log n \cdot P' \cdot p_\mu)$, which is covered by (3.48), as claimed.

The algorithm performs no more than $\log(p_{\max}/p_0)$ increments of precision. Summing up the bound (3.48) for all precisions $p_\mu = 2^\mu p_0$, $0 \le \mu \le \log(p_{\max}/p_0)$, we obtain the following bound on the total cost of introducing p_0 and all higher precisions:

$$\begin{aligned}
&\textstyle\sum_{\mu=0}^{\log(p_{\max}/p_0)} O(n^2 \log n \cdot P' \cdot (p_\mu + \log(nh))) \\
&= \textstyle\sum_{\mu=0}^{\log(p_{\max}/p_0)} \big(2^\mu \cdot O(n^2 \log n \cdot P' \cdot p_0) + O(n^2 \log n \cdot P' \cdot \log(nh))\big) \\
&= O(n^2 \log n \cdot P' \cdot (p_{\max} + \log(p_{\max}/p_0) \log(nh))),
\end{aligned} \tag{3.49}$$

using $\sum_{\mu=0}^{M} 2^\mu = O(2^M)$. This cost reduces to the claimed bound (3.47), since Lemma 3.49 implies $\log(p_{\max}/p_0) \log(nh) = O(\log^2 p_{\max})$, so that the last factor is $O(p_{\max})$.

Now we turn to the cost of subdivision. It is dominated by the invocations of de Casteljau's algorithm in line 26; auxiliary operations on coefficient sequences such as reading off their signs or adjusting them in the Zeno trap at lines 19ff. can be charged to the cost of producing the coefficients in the first place.

At precision p_μ, each call to de Casteljau's algorithm needs $O(n^2 \log n \cdot (p_\mu + \log h))$ bit operations, because the subdivision parameter has length $k = O(\log n)$ and, by Lemma 3.63, the coefficients have length $O(p_\mu + \log h)$. We distinguish three kinds of subdivisions.

First, there are successful subdivisions, and since they give rise to new nodes in the subdivision tree, their number is at most $P' + 1$ by condition (i) in Definition 3.59. Since $p_\mu \le p_{\max}$ and $\log h = O(p_{\max})$, see Lemma 3.49, their total cost is $O(n^2 \log n \cdot P' \cdot p_{\max})$, so they are covered by (3.47). The cost of computing the subdivision point m by a single convex combination in line 29 is insignificant: The interval boundaries have lengths $O(h \log n) = O(p_{\max})$ and the weights have lengths $O(\log n)$.

Second, there are failed subdivisions at values of p for which a successful subdivision has occurred previously. Whenever there is at least one successful subdivision at precision p, the ratio of failed over successful subdivisions at that precision is at most $2 : 1$. Consequently, we can charge them to the successful subdivisions at the same precision.

Third, for any value p_μ of p, there may be up to two subdivisions that fail before any subdivision at $p = p_\mu$ has succeeded. We can subsume their cost $O(n^2 \log n \cdot (p_\mu + \log h))$ into the cost of introducing p_μ, because it is covered by the bound (3.48).

In summary, we have shown that the bound (3.47) also covers the cost of all subdivisions. Thus, it is indeed a bound on the number of bit operations that the whole algorithm has performed, but it depends on $p_{\max}$, which is an a posteriori parameter.

Let us now make the transition to an a priori bound on the *expected* number of bit operations. The bound (3.47) is linear in $p_{\max}$. By linearity of expectation, we can compute its expected value by substituting the expected value of $p_{\max}$. According to Theorem 3.48 and the discussion at the beginning of the proof, we have

$$\begin{aligned}\mathrm{E}[p_{\max}] &\leq \frac{9}{10}p_{\mu_0} + \frac{1}{10}\Big(\frac{9}{10}\cdot 2p_{\mu_0} + \frac{1}{10}\big(\frac{9}{10}\cdot 2^2 p_{\mu_0} + \frac{1}{10}(\ldots)\big)\Big) \\ &= \frac{9}{10}p_{\mu_0}\sum_{i=0}^{\infty}\Big(\frac{2}{10}\Big)^i = \frac{9}{8}p_{\mu_0} < \frac{9}{4}p_{\mathrm{ok}} \\ &= O(n\cdot(\log\frac{|I_0|}{s} + \log n)).\end{aligned} \tag{3.50}$$

Thus, we arrive at the bound (3.46) and the theorem is established. □

We will now derive more specific complexity statements from Theorem 3.61 by supplying tsqd-bounds P', or equivalently, as justified by Proposition 3.60, bounds P on the sum of all rt-path lengths in the exact Descartes method. We recall the notion of a subdivision ratio bound from Definition 3.12 on page 57 and the fact that subdivision with a parameter $\alpha \in [1/4,\, 3/4]$ leads to a subdivision ratio bound of $\rho = 4/3$. Now Theorem 3.19(iii) provides a bound on P in terms of certain pairs of roots, suitable for a subsequent application of the Davenport-Mahler bound (Theorem 3.9). An essential factor in the Davenport-Mahler bound is a discriminant, or, in our generalization, a subdiscriminant. This is nice for integer coefficients, because a non-zero integral (sub)discriminant has magnitude at least 1. One can also cope with this (sub)discriminant in case of algebraic coefficients for which some parameters are known; we will see an example in §3.4.3.

But how should we bound this (sub)discriminant for a polynomial with arbitrary real coefficients? Conversely, if we cannot bound it, should we accept it as a first-class parameter for the analysis of our root searching problem? Admittedly, $\log|\mathrm{Discr}(F)|$ does have a meaning in terms of the relative position of the roots in the complex plane, as it is essentially their average logarithmic distance. But the Davenport-Mahler bound as a whole eludes such a geometric interpretation, because it is incompatible with coordinate changes $X \mapsto \lambda X$, $\lambda \in \mathbb{R}^*$, to which our algorithm is insensitive (at least if $\lambda = 2^k$, $k \in \mathbb{Z}$). The author therefore likes to think of the Davenport-Mahler bound as, ultimately, a non-geometric result pertaining to algebraic number theory, which is not natural for polynomials with real coefficients in full generality. Therefore, we retain the root separation s as parameter of our analysis.

Corollary 3.64. *Consider the situation of Theorem 3.61. If $I_0 = (-2^{r+1}, +2^{r+1})$ has been chosen using the approximate dyadic Fujiwara complex root bound (3.24), then $\log|I_0| \leq \log \mathrm{RR}(A_{\mathrm{in}}) + \log n + O(1)$, and if furthermore the order of choosing from Q satisfies one of the conditions from Proposition 3.60, then the expected number of bit operations is*

$$O(n^4 \log n \cdot (\log\frac{\mathrm{RR}(A_{\mathrm{in}})}{s} + \log n)^2). \tag{3.51}$$

The complex root radius RR has been defined in Equation (2.22) on page 40. The ratio $\mathrm{RR}(A_{\mathrm{in}})/s$ has an immediate geometric interpretation in terms of the relative position of the roots and the origin in the complex plane.

Proof. Proposition 2.53(iii) on page 43 implies $\log \mathrm{RB}_{\mathrm{dF}}(A_{\mathrm{in}}) \le \log \mathrm{RR}(A_{\mathrm{in}}) + \log n + 1$. According to Lemma 3.36, the radius 2^{r+1} of $|I_0|$ exceeds $\mathrm{RB}_{\mathrm{dF}}(A_{\mathrm{in}})$ by no more than a factor of 16. This establishes $\log |I_0| \le \log \mathrm{RR}(A_{\mathrm{in}}) + \log n + O(1)$.

Theorem 3.19(iii) yields the bound $P \le \log_\rho(\prod |I_0| / |\alpha - \beta|) + O(n)$ for the sum of all rt-path lengths, where the product is over at most $n/2$ pairs (α, β) of roots, each of which satisfies $|\alpha - \beta| \ge s$. Thus, $P = O(n \cdot (\log(|I_0| / s) + 1))$, and Theorem 3.61 in combination with Proposition 3.60 yields a bound of $O(n^4 \log n \cdot (\log(|I_0| / s) + 1) \cdot (\log(|I_0| / s) + \log n))$. Substituting the preceding estimate of $\log |I_0|$, the claim follows. □

We now turn to the case of integer coefficients, for which we can profitably invoke the Davenport-Mahler bound.

Corollary 3.65. *Consider a polynomial $A_{\mathrm{in}}(X) = \sum_{i=0}^{n} a_i X^i$ of degree $n \ge 2$ with integer coefficients of magnitude less than 2^τ, all of whose real roots are simple. If $I_0 = (-2^{r+1}, +2^{r+1})$ has been chosen using the approximate dyadic Fujiwara complex root bound (3.24), then $\log |I_0| \le \tau + O(1)$, and if furthermore the order of choosing from Q satisfies one of the conditions from Proposition 3.60, then the expected value of the number of bit operations performed by the randomized algorithm DescartesE08basic invoked for A_{in} and I_0 is*

$$O(n^5 \log n \cdot (\tau + \log n)^2). \tag{3.52}$$

Proof. The claim on I_0 follows from (3.24) using the estimates $l_i^+ < \tau + 2$ and $l_n^- > -2$.

Theorem 3.19 in conjunction with Corollary 3.11 to the Davenport-Mahler bound shows that the sum of all rt-path lengths for the exact Descartes method is $P = O(n \cdot (\tau + \log n))$, and by Proposition 3.60, there is a tsqd-bound P' for the bitstream Descartes algorithm with the same asymptotics. Corollary 3.11 applied to a closest pair of distinct complex roots yields $\log(|I_0| / s) = O(n \cdot (\tau + \log n))$ again. Substituting these estimates into the bound from Theorem 3.61 proves the claim. □

Compared to the exact integer algorithms (see Theorems 3.30, 3.33, and 3.34), this bound contains an additional factor of $\log n$, partly due to the occurrence of subdivision parameters α with lengths $O(\log n)$, as opposed to the fixed value $\alpha = 1/2$.

The bitstream Descartes algorithm has the advantage that coefficient lengths do not grow during subdivision; however, the worst-case bound on expected maximal precision, see (3.38) and (3.50), is essentially n times the bound on subdivision depth, which matches the coefficient growth of the exact integer algorithms. An attempt to improve the complexity bound for the bitstream Descartes algorithm would presumably have to improve upon our analysis of sufficient precision, since the other limiting factors, namely tree size and de Casteljau's algorithm are essentially tight. (We do not discuss asymptotically fast subdivision in the bitstream setting.) It appears that this would necessitate an analysis in parameters other than s, as we will see below.

Having analyzed time consumption, we turn to the precision consumption of our bitstream Descartes algorithm. This is an independent resource consumed by the algorithm; the computing time analysis above does not include the effort on the user's side to produce the necessary coefficient approximations. So the question is: How can we analyze the maximum precision extracted out of each coefficient a_j of A_{in}? By construction, the algorithm does not choose this precision absolutely, but relative to the leading coefficient and the initial interval, so we bound the precision relative to these quantities. It is easy to check

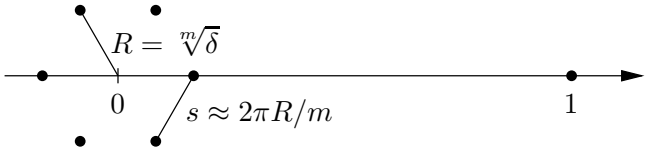


Figure 3.2: The roots of $A_\delta(X) = (X^m - \delta)(X - 1)$ for $m = 6$.

that the following bound, like the algorithm itself, is invariant under scaling of the coefficients, $A_{\text{in}}(X) \rightsquigarrow \lambda A_{\text{in}}(X)$, and scaling of the indeterminate, $A_{\text{in}}(X) \rightsquigarrow \lambda^n A_{\text{in}}(X/\lambda)$, with factors $\lambda = 2^k$, $k \in \mathbb{Z}$, assuming that the radius 2^{r+1} of the initial interval is transformed accordingly (cf. Definition 2.48 on page 40).

Proposition 3.66. *The randomized procedure DescartesE08basic$((a_0, \ldots, a_n), r, l_n^-)$ accesses the coefficients a_j through calls of the form $[2^{p'-(n-j)(r+1)-l_n^-} a_j]$, where p' is independent of j. The expected maximal value of p' is bounded by $O(n \cdot (\log(|I_0|/s) + \log n))$, with notation as in Theorem 3.61.*

Proof. Depending on the current precision parameter p, the coefficients of A_{in} are extracted out of their representing bitstreams in line 10 of procedure *DescartesE08basic* and in line 5 of the block ⟨reinitialize Q from new p⟩ as $[2^{j(r+1)-l+q'} a_j]$ with a precision

$$\begin{aligned} & j(r+1) - l + q' \\ & \quad = j(r+1) - (l_n^- + \lfloor \log n! \rfloor + n(r+2)) + (p + \lceil \log n! \rceil + \lceil \log(n+1) \rceil + 2) \\ & \quad \le (j-n)(r+1) - l_n^- + p - n + \lceil \log(n+1) \rceil + 3 \\ & \quad \le (j-n)(r+1) - l_n^- + p + 3. \end{aligned}$$

The claim follows from the bound (3.50) on the expected maximal value of p. □

How good is this bound? By scaling the indeterminate X suitably, we may restrict our subsequent considerations to the case $r = 0$ without a disadvantage to their generality. The preceding proposition, with logarithmic factors suppressed, quantifies the precision consumption as essentially $O^\sim(n \log(1/s))$. It is natural that the required precision depends on $\log(1/s)$. We intend to justify the coefficient n.

Infinitesimal perturbations of a polynomial's coefficients with magnitude ε introduce an error in an m-fold root on the order of $\sqrt[m]{\varepsilon}$, see [Wil63, §2.7]. Conversely, if a polynomial has a cluster of m close roots with minimum distance $0 < s \ll 1$, it may be necessary to approximate its coefficients to within an error on the order of s^m in order to distinguish a cluster of roots from one multiple root. Figure 3.2 shows a concrete example: The polynomial $A_\delta(X) = (X^m - \delta)(X - 1) = X^{m+1} - X^m - \delta X + \delta$ with $m \ge 2$ and $0 < \delta \ll 1$ has roots 1 and $\sqrt[m]{\delta} \cdot \zeta_m^i$, $1 \le i \le m$, where ζ_m is a primitive mth root of unity. The root $X = 1$ makes sure that $r = 0$ is a valid choice for the initial interval. The m other roots are evenly spread on a circle with radius $R = \sqrt[m]{\delta}$ and have minimum distance s slightly less than $2\pi R/m$. This implies $\delta = R^m \approx (ms/2\pi)^m$.

Suppose that m is even. Then $A_\delta(X)$ has two real roots $\pm\sqrt[m]{\delta}$ that coalesce for $\delta = 0$ and become imaginary for $\delta < 0$. To isolate the real roots of $A_\delta(X)$ from approximate coefficients, their approximation error ε has to be less than δ; otherwise the number of real roots is not well-defined. Using the estimate of δ from above, this asks for a precision of $\log(1/\varepsilon) > m \cdot (\log(1/s) + \log(2\pi) - \log m)$. The degree of $A_\delta(X)$ is $n = m + 1$, so we see that the coefficient of $\log(1/s)$ in the precision necessary for hard inputs is indeed $\Omega(n)$.

3.3.10 Variants of the algorithm

In this section, we discuss potential improvements of the bitstream Descartes algorithm. The first improvement is one that we definitely recommend, even though this does not make a difference in terms of the complexity analysis:

We should limit the adverse effects of subdivision with $\alpha \neq 1/2$. This is more costly than bisection, because it requires the multiplication with non-trivial weights in de Casteljau's algorithm. Furthermore, the algorithm's output deteriorates: The low-order bits in the boundaries of isolating intervals have a value much less than the width of the interval (the opposite of the good situation discussed in §3.2.3). To counteract this, we propose to attempt subdivision at $\alpha = 1/2$ before using the prescribed denominator. Of course, if these attempts fail, they must not be counted as a failures of the current precision.

We call the form of the bitstream Descartes algorithm resulting from this improvement *DescartesE08*.

We have intentionally chosen the radius of the initial interval I_0 in §3.3.2 as an overapproximation of the magnitudes of all complex roots; this has provided an important point of reference for precision management, cf. Proposition 3.40 and Lemma 3.63. However, if a smaller interval I_1 is known that contains all real roots and that satisfies the invariant (3.36) at some suitable initial precision, one can directly "zoom in" from I_0 on I_1 with two subdivision steps, analogously to the reinitialization of Q at a higher precision.

As described, the bitstream Descartes algorithm maintains a global precision parameter p together with the counters N_{try} and N_{fail} for subdivision attempts. A large number of failures to subdivide a single interval can trigger a global increase of precision. A possible alternative would be to keep these decisions local to each interval (with subdivision passing the relevant state on to the subintervals), matching the intuition that in some areas, the root separation and thus the acceptable approximation error in the coefficients may be much smaller than in others.[12] On the other hand, the precision extracted out of the bitstream coefficients is determined by the hardest subproblem anyway, and maintaining one global precision parameter and counters makes the algorithm behave more conservatively in its precision consumption. This is preferable in a setting where handling long coefficient approximations inside the bitstream Descartes algorithm is cheap compared to producing them on the other side of the bitstream interface; as M. Kerber reports (personal communication, December 2007), this is the case for our geometric application to be discussed in §3.4.

The initial precision parameter $p_0 = 6n + 1$ is barely large enough to make sure that the invariant (3.36) is satisfied by the initial interval. This is likely to necessitate a precision increment, and thus a reinitialization of Q, rather soon. It may therefore be preferable in practice to start at a moderately larger initial precision, such as $6n + 20$.

Further variations of the algorithm are possible.

Our original publication [EKK+05] proposed to restart from scratch each time a higher precision is needed. The obvious disadvantage is the loss of all partial information gained up to that point; however, by the usual argument that the last term dominates a geometric sum, cf. (3.49), this does not affect the asymptotic computing time. (Actually, the analysis

[12] An analysis of this scheme requires a sub-constant failure probability for a subdivision attempted in sufficient precision.

becomes simpler, because this eliminates the reinitialization of Q and thus the need for a bound on the sum of depths of its entries.) The preference of bisections over uneven subdivisions may become more effective in combination with this approach, because the algorithm gets a chance to retry previously failed bisections at higher precisions.

We have not explored the idea to use denominators other than powers of two for subdivision parameters. When the interval width is much larger than the positional value of the lowest-order bit in the endpoints, this would allow to choose subdivision points on the grid defined by the positional value of the lowest-order bit, thereby avoiding a further increase the bit length of the endpoints. On the other hand, this necessitates divisions of long integers.

For the exponential growth of the precision parameter p, we have chosen the customary base 2. It is possible to tune the performance of the algorithm by varying this base, although the asymptotics remain the same. Schönhage [Sch03] discusses this exponential raising and more advanced strategies with regard to worst-case and average-case bounds on the competitive factor of their cost compared to the unknown optimal choice. These results apply directly to algorithms that incur the full cost for each attempted parameter value; our situation with partial information potentially being gained cheaper at small parameter values is less clear. Also, any choice of a base would have to take into account the order of growth of the computing time as a function of the precision parameter p. The time for the bitstream Descartes algorithm itself is linear in this parameter, implying that base 2 is actually not a bad choice in terms of the worst-case competitive factor [Sch03, p. 611], but the time spent on the other side of the bitstream interface to produce coefficient approximations may have higher orders of growth, precluding a universal choice of an optimal base.

The bitstream Descartes algorithm uses randomization for a single purpose, namely the choice of a subdivision point among $K/2+1 \approx 8n$ candidates, of which at most n are bad, once the coefficient precision is sufficiently high. Thus, the algorithm can be derandomized in the trivial way – trying all possible choices – at the expense of an additional factor n in the complexity bound. This does not seem to be particularly useful from a practical point of view.

3.3.11 Discussion

The goal stated in §3.3.1 has been reached: We have formulated and analyzed a form of the Descartes method that produces correct isolating intervals for a polynomial A_{in} whose coefficients are only known through approximations. The necessity of an (explicit or implicit) fall-back to exact arithmetic present in earlier work has been overcome by the randomized choice of subdivision points. In particular, our bitstream Descartes algorithm works for all real polynomials with simple real roots; there is no need to insist on coefficients to be algebraic or otherwise amenable to exact arithmetic.

The development and analysis of the bitstream Descartes algorithm in this thesis improves upon the original publication [EKK$^+$05] in several ways. The treatment of case (M0) identical to (D0) is new. The computing time analysis can now take full advantage of the Davenport-Mahler bound (Theorem 3.9) and the resulting bound on the subdivision tree (Theorem 3.19), if applicable. The estimate of sufficient precision given here (Theorem 3.48) is a substantial generalization that works in the presence of multiple complex roots, which will be crucial for the application in §3.4.

Most importantly, though, the role of this bound on sufficient precision has changed. Originally, the algorithm was formulated to guess an estimate s of separation (minimum root distance) and to derive from it the necessary precision and a bound on subdivision depth. When this bound on subdivision depth was reached, a new estimate of separation and thus a higher precision were introduced. Consequently, there was little chance for the algorithm to work with a precision lower than the one prescribed by the analysis. Here, we have formulated the algorithm to choose the precision directly and to work with it as long as possible. (The Zeno trap from §3.3.6 has obviated the need for a bound on subdivision depth.) Thus, our estimate of sufficient precision no longer occurs as an artifact within the algorithm; it only appears in the analysis, and the algorithm's actual performance is not tied to it.

3.4 An application to algebraic curves

3.4.1 Overview

A *real plane algebraic curve* is the vanishing set of a non-constant polynomial $F \in \mathbb{R}[X, Y]$ in $\mathbb{R}^2$. For brevity, we will often refer to this vanishing set simply as the *algebraic curve* F. There are numerous books on the geometry of algebraic curves; we mention Walker's classic [Wal50] and Gibson's excellently readable introduction [Gib98].

We are interested in analyzing the geometry of an algebraic curve $F \in \mathbb{Z}[X, Y]$ in the following fashion: Consider a vertical line $\ell_x \colon X - x = 0$ that moves continuously and monotonically over the plane from left to right. At each position $x \in \mathbb{R}$, the vertical line intersects F in the points (x, y) whose Y-coordinates are the real roots of $F(x, Y)$. For brevity, we assume here that the partial degree in Y remains constant: $\deg(F(x, Y)) = \deg(F(X, Y)) > 0$ for all $x \in \mathbb{R}$. (This rules out vertical asymptotes and vertical line components.) How do the points of F on ℓ_x vary with x? We can take F to be square-free without changing its vanishing set; then F and its partial derivative $\mathrm{D}_Y F$ with respect to Y are coprime, so that by Bézout's theorem (see [Gib98, Thm. 14.7] [Wal50, Thm. III.3.1], cf. Lemma 3.70(i) below), the curves F and $\mathrm{D}_Y F$ have only finitely many points in common, which we call the *critical points* of F. Except where ℓ_x hits a critical point, the points of F on ℓ_x retain their relative position and vary smoothly according to the implicit function theorem, tracing out the X-monotone arcs of F. It is at critical points only that an arc can begin, end, or intersect, or that isolated points of F can occur. Our task is to determine the critical points, the arcs, and their adjacency relation.

In algorithmic terms, such an analysis is done by computing a cylindrical algebraic decomposition (c.a.d.) of the plane adapted to F, augmented with adjacencies: First, in the *projection phase*, one computes (a finite superset of) the set of *critical X-coordinates*, i.e., the X-coordinates of critical points. This partitions the X-axis into 0-cells (said coordinates) and 1-cells (the open intervals forming their complement). Second, in the *lifting phase*, one determines the shape of F over each cell by choosing a sample point α and computing the real roots of $F(\alpha, Y)$. Finally, one computes the adjacency relation between arcs over 1-cells and the points over neighbouring 0-cells.

The notion of a cylindrical algebraic decomposition was introduced by Collins [Col75] for his decision procedure for first-order formulae over real closed fields, see also [ACM84a] [ACM84b]; a textbook reference is [BPR06]. Our interest in this kind of curve analysis comes from research on a fundamental task in computational geometry: computing the

arrangement [AS00] [Hal04] induced by a finite set of algebraic curves. The basic geometric operations needed for arrangement computation can be reduced to analyzing the geometry of single curves and pairs of curves as sketched above [EKSW06, §5] [BK07].

Where does the bitstream Descartes algorithm enter the picture? In computing a c.a.d., real root isolation is needed for various polynomials. The polynomials $F(\alpha, Y)$ for critical X-coordinates α are arguably the hardest: α and thus the coefficients of $F(\alpha, Y)$ are, in general, irrational algebraic numbers; so as discussed in §3.3.1 and the previous work cited there, real root isolation with approximate coefficients is highly desirable. However, by choice of α, the polynomial $F(\alpha, Y)$ has a multiple root, and in the interesting cases actually a multiple *real* root, so that the bitstream Descartes algorithm as presented in §3.3 would not terminate – additional techniques are required.

Michael Kerber,[13] in joint work with Nicola Wolpert and the author of this thesis, has developed an algorithm for curve analysis that works for an arbitrary input curve F and performs lifting over critical X-coordinates α by isolating the real roots of $F(\alpha, Y)$ with variants of the bitstream Descartes algorithm, using additional information to terminate on intervals containing a multiple real root. The outstanding feature of this algorithm is the complete avoidance of factoring $F(\alpha, Y)$ or falling back to exact arithmetic. We refer to [EKW07] for a description of the algorithm and a comparison to related work. Subsequent joint work [EK08] describes the analysis of curve pairs in the same style and the resulting algorithm for arrangement computation.

A complete account of these algorithms, which comprise a number of case distinctions, can literally fill a thesis of its own, namely [Ker0X]. In the present thesis, we ignore all parts of these algorithms other than the use of the bitstream Descartes algorithm, and regarding that, we focus on the "backbone case" of the algorithm for a single curve, to be analyzed in §3.4.3. Its analysis is paradigmatic for the other uses of the bitstream Descartes algorithm in [EKW07] and [EK08].

3.4.2 Reminder on resultants

We briefly recall the notion of a resultant and some basic properties needed in the sequel.

Definition 3.67. Let $F(X) = \sum_{i=0}^{n} f_i X^i$ and $G(X) = \sum_{i=0}^{m} g_i X^i$ be polynomials with coefficients in an integral domain R. The *resultant* of F and G with respect to formal degrees n and m is

$$\mathrm{Res}(F,G) = \begin{vmatrix} f_n & \cdots & \cdots & f_0 & & \\ & \ddots & & & \ddots & \\ & & f_n & \cdots & \cdots & f_0 \\ g_m & \cdots & \cdots & g_0 & & \\ & \ddots & & & \ddots & \\ & & g_m & \cdots & \cdots & g_0 \end{vmatrix} \in R. \tag{3.53}$$

On the right is the determinant of the *Sylvester matrix*, which has size $(m+n) \times (m+n)$ and consists of m rows holding the coefficients of F and n rows holding the coefficients of G in echelon form; all other entries are zero.

[13] This Michael Kerber is a computer scientist from U Saarland and MPI-I in Saarbrücken, Germany; not to be confused with the mathematician from TU Kaiserslautern or the physicist from TU Vienna of the same name.

A few remarks are in order.

If $g_m = \cdots = g_{m-k+1} = 0$, the first k columns have upper triangular form, and the resultant w.r.t. formal degree m is the resultant w.r.t. formal degree $m-k$ times f_n^k. Analogous observations apply to the case $f_n = 0$, since $\operatorname{Res}(F,G) = (-1)^{mn}\operatorname{Res}(G,F)$.

Unless specifically indicated otherwise, we let the symbol $\operatorname{Res}(F,G)$ denote the resultant of F and G with respect to their actual degrees.

A comparison of Definition 3.67 above with Definition 3.5 on page 53 shows that $\operatorname{Res}(F,F') = (-1)^{n(n-1)/2} f_n \operatorname{Discr}(F)$. This correspondence is extended to jth subdiscriminants for $j > 0$ by the notion of subresultants, see [BPR06, §4.2.2].

The coefficient domain R may contain indeterminates; if it is necessary to specify the indeterminate eliminated by the resultant, we give it as a third argument as in $\operatorname{Res}(F,G,X)$. Likewise, we write $\operatorname{sDisc}_j(F,X)$ for a subdiscriminant of F with respect to the indeterminate X.

The fundamental property of resultants is the following relation to root differences and the equivalent relation of $\operatorname{Res}(F,G)$ to the values of one polynomial at the roots of the other.

Proposition 3.68. *If $F(X) = f_n \prod_{i=1}^{n}(X-\alpha_i)$ and $G(X) = g_m \prod_{j=1}^{m}(X-\beta_j)$, then*

$$\operatorname{Res}(F,G) = f_n^m g_m^n \prod_{i=1}^{n}\prod_{j=1}^{m}(\alpha_i - \beta_j) = f_n^m \prod_{i=1}^{n} G(\alpha_i) = (-1)^{mn} g_m^n \prod_{j=1}^{m} F(\beta_i). \tag{3.54}$$

For proofs see, e.g., [BPR06, Thm. 4.16] [vdW93, §35].

Corollary 3.69. *If $F, G, H \in \mathbb{C}[X]$, then $\operatorname{Res}(F,GH) = \operatorname{Res}(F,G)\cdot\operatorname{Res}(F,H)$.*

We now turn to quantitative aspects of the resultant and similar determinants. We consider two polynomials

$$\begin{aligned} F(U,V,X) &= \sum_{i=0}^{n} f_i(U,V)X^i, \quad f_i(U,V) = \sum_{j=0}^{n-i} a_{ij}U^jV^{n-i-j}, \\ G(U,V,X) &= \sum_{i=0}^{m} g_i(U,V)X^i, \quad g_i(U,V) = \sum_{j=0}^{m-i} b_{ij}U^jV^{m-i-j}, \end{aligned} \tag{3.55}$$

and an $(m+n-2j)\times(m+n-2j)$ subdeterminant

$$S_j(U,V) = \begin{vmatrix} f_n & f_{n-1} & f_{n-2} & \cdots & \cdots \\ & f_n & f_{n-1} & f_{n-2} & \cdots \\ & & \ddots & \ddots & \ddots \\ g_m & g_{m-1} & g_{m-2} & \cdots & \cdots \\ & g_m & g_{m-1} & g_{m-2} & \cdots \\ & & \ddots & \ddots & \ddots \end{vmatrix} \tag{3.56}$$

of the Sylvester matrix, constructed by taking the first $m+n-2j$ columns of the first $m-j$ rows holding the coefficients f_i and the first $n-j$ rows holding the coefficients g_i. For $j=0$, this is the Sylvester determinant defining $\operatorname{Res}(F,G,X)$. For $G = F'$, this is the determinant defining $\operatorname{sDisc}_j(F,X)$ up to sign (see page 53).

Lemma 3.70. *Consider the polynomials F, G and the determinant S_j as above.*

(i) $S_j(U,V) = \sum_{i=0}^{d} s_i U^i V^{d-i}$ *is homogeneous of degree* $d = (m-j)(n-j)$.

(ii) If $S_j(U,1) \neq 0$, *then* $\deg_U(S_j(U,1)) \leq (m-j)(n-j)$.

(iii) If all coefficients a_{ij}, b_{ij} *are complex numbers such that* $\log|a_{ij}| \leq \tau$ *and* $\log|b_{ij}| \leq \sigma$, *then* $\log \sum_{i=0}^{d} |s_i| \leq (m-j)\tau + (n-j)\sigma + O((m+n-2j)\log(m+n))$.

In (iii), we use the convention $\log 0 = -\infty < x$ for any $x \in \mathbb{R}$.

Proof. We abbreviate the numbers of rows to $n' = n - j$ and $m' = m - j$.

Ad (i). We have to show $S_j(\lambda U, \lambda V) = \lambda^{m'n'} S_j(U,V)$ for an indeterminate λ. We can express $S_j(\lambda U, \lambda V)$ as a determinant of the form (3.56) with f_{n-i} replaced by $\lambda^i f_{n-i}$ and g_{m-i} replaced by $\lambda^i g_{m-i}$. We modify this determinant further by multiplying the first F-row and G-row by λ, the second F-row and G-row by λ^2, and so on, which multiplies the determinant by λ raised to the power of $m'(m'+1)/2 + n'(n'+1)/2$. This has produced the determinant

$$\begin{vmatrix} \lambda^{1+0} f_n & \lambda^{1+1} f_{n-1} & \lambda^{1+2} f_{n-2} & \cdots \\ & \lambda^{2+0} f_n & \lambda^{2+1} f_{n-1} & \cdots \\ & & \ddots & \ddots \\ \lambda^{1+0} g_m & \lambda^{1+1} g_{m-1} & \lambda^{1+2} g_{m-2} & \cdots \\ & \lambda^{2+0} g_m & \lambda^{2+1} g_{m-1} & \cdots \\ & & \ddots & \ddots \end{vmatrix} \tag{3.57}$$

in which we can extract a factor of λ from the first column, a factor of λ^2 from the second column, and so on, altogether we extract λ raised to the power of $(m'+n')(m'+n'+1)/2$ The claimed degree now results from $(m'+n')(m'+n'+1) - (m'(m'+1)+n'(n'+1)) = 2m'n'$. (This trick is common in elementary proofs of Bézout's theorem for algebraic curves, see, e.g., [Gib98, Lem. 14.3] [Wal50, §III.3.1].)

Ad (ii). Immediate from (i).

Ad (iii). The $(m'+n') \times (m'+n')$ determinant S_j is a sum of no more than $(m'+n')!$ terms, each of which has the form $\prod_{\mu=1}^{m'} f_{i_\mu}(U,V) \cdot \prod_{\nu=1}^{n'} g_{j_\nu}(U,V)$. Multiplying out one term yields $\prod_{\mu=1}^{m'}(n - i_\mu + 1) \cdot \prod_{\nu=1}^{n'}(m - j_\nu + 1) \leq (n+1)^{m'}(m+1)^{n'}$ monomials with coefficients of magnitudes at most $2^{m'\tau + n'\sigma}$. Altogether, at most $(m'+n')!\,(n+1)^{m'}(m+1)^{n'}$ monomials are produced; the logarithm of this number is $O((m'+n')\log(m+n))$. □

3.4.3 Lifting with the bitstream (m,k)-Descartes algorithm

Let $F \in \mathbb{Z}[X,Y]$ be polynomial of total degree $n \geq 2$ that is *Y-regular*, i.e., its partial degree in Y is also n. Hence F has the form

$$F(X,Y) = \sum_{i=0}^{n} f_i(X) Y^i, \qquad f_i(X) = \sum_{j=0}^{n-i} a_{ij} X^j, \qquad f_n = a_{n0} \neq 0. \tag{3.58}$$

We assume that F is square-free and thus coprime to its partial derivative $\mathrm{D}_Y F$, hence we have a non-zero resultant

$$R(X) := \mathrm{Res}(F,\, \mathrm{D}_Y F,\, Y) = \prod_{\mu=1}^{M} R_\mu^{e_\mu}(X), \tag{3.59}$$

which is essentially $\mathrm{Discr}(F, Y)$, and which decomposes into $\mathbb{Q}$-irreducible and pairwise coprime integer polynomials

$$R_\mu(X) = \sum_{i=0}^{h_\mu} r_{\mu i} X^i = \ell_\mu \prod_{\nu=1}^{h_\mu} (X - \xi_{\mu\nu}) \in \mathbb{Z}[X] \quad \text{with} \quad \ell_\mu = r_{\mu h_\mu}, \quad \xi_{\mu\nu} \in \mathbb{C}. \tag{3.60}$$

The existence of a $\mathbb{Q}$-irreducible factorization over $\mathbb{Z}$ is a consequence of Gauss' Lemma (primitive polynomials have a primitive product), see [BPR06, Lem. 10.17] [vdW93, §30], and depends on unique factorization in $\mathbb{Z}$.

We set

$$F_{\mu\nu}(Y) := F(\xi_{\mu\nu}, Y) \in \mathbb{Z}[\xi_{\mu\nu}][Y] \subseteq \mathbb{C}[Y]. \tag{3.61}$$

By the Y-regularity of F, we have $\deg F_{\mu\nu} = \deg F = n$. The degree of $\gcd(F_{\mu\nu}, F'_{\mu\nu})$ is the smallest index j such that $\mathrm{sDisc}_j(F_{\mu\nu}) \neq 0$ and thus, by the irreducibility of R_μ, the same for all $1 \leq \nu \leq h_\mu$. So we can define

$$k_\mu := \deg(\gcd(F_{\mu\nu}, F'_{\mu\nu})) \leq n - 1. \tag{3.62}$$

Each polynomial $F_{\mu\nu}(Y)$ has exactly $n - k_\mu$ distinct complex roots.

We are now ready to state our objective. To analyze the curve F, it is necessary to isolate the real roots of $F(\alpha, Y)$ for all $\alpha \in \{\xi_{\mu\nu}\}_{\mu,\nu} \cap \mathbb{R}$. To do that with the Descartes method, we need to modify it such that it can terminate in the presence of a multiple real root. M. Kerber's (m, k)-*Descartes algorithm* [EKW07, §5] is a partial solution to this problem, using the following additional information provided by the signs of the subdiscriminants of $F_{\mu\nu}(Y)$: the number m of distinct real roots of $F_{\mu\nu}(Y)$, see [BPR06, Thm. 4.33], and the gcd-degree $k = k_\mu$ from (3.62). For $m \leq 1$, nothing needs to be done. For $m \geq 2$, the (m, k)-Descartes algorithm consists in traversing the subdivision tree of the Descartes method in breadth-first order until one of two termination conditions is verified:

(S) If the current partition of the initial interval comprises exactly $m - 1$ intervals with Descartes test equal to 1 and a single interval I that presents a Descartes test greater than 1, then the algorithm terminates with an indication of success and reports these m intervals as isolating intervals for the m distinct real roots, with I distinguished as isolating interval for the unique multiple real root.

(F) If the Descartes test for every interval in the current partition is k or less, this certifies that there is more than one multiple complex root, and the algorithm terminates with an indication of failure.

If $F_{\mu\nu}(Y)$ has only one multiple *complex* root (as is the case in a generic coordinate system), the (m, k)-Descartes algorithm necessarily succeeds; if $F_{\mu\nu}(Y)$ has more than one multiple *real* root, the (m, k)-Descartes algorithm necessarily fails.[14] We refer to [EKW07, §5] for details, but not without emphasizing that a failure of the (m, k)-Descartes algorithm does not entail a failure or error of the curve analysis, it merely triggers a change of coordinates.

Of course, we do not propose to actually carry out the (m, k)-Descartes algorithm with exact coefficients. Instead, the techniques from §3.3 for approximate coefficients are applied: The *bitstream* (m, k)-*Descartes algorithm* consists in a breadth-first traversal of

[14] The indeterminism in the gap between these two conditions helps to avoid the costly computations to decide exactly whether there is a unique multiple complex root.

the subdivision tree constructed by the bitstream Descartes algorithm until the sets of ε-approximate sign variations approximating the Descartes test for each interval indicate with certainty that termination condition (S) or (F) is verified. In §3.3.4, we have seen two lemmas guaranteeing that we eventually gain certainty about the Descartes test being zero or one. Analyzing the failure case in the bitstream (m, k)-Descartes algorithm requires an analogous lemma for Descartes tests larger than 1; we return to this issue in §3.4.4. In the present section, we restrict ourselves to the analysis of the success case in the bitstream (m, k)-Descartes algorithm. More precisely, the rest of this section is devoted to the proof of the following result.

Theorem 3.71. *Let the polynomial $F(X, Y)$ as in (3.58) have integer coefficients with magnitudes $|a_{ij}| \leq 2^\tau$. Consider the executions of the bitstream (m, k)-Descartes algorithm on $F(\alpha, Y)$ for all real roots α of $R(X) = \operatorname{Res}(F, \mathrm{D}_Y F, Y)$, starting from initial intervals $I_0(\alpha)$ of the form $(-2^{r+1}, +2^{r+1})$ chosen with the approximate dyadic Fujiwara complex root bound (3.24). Suppose that all these executions have terminated via condition (S). Excluding the work done inside the procedures providing the coefficient bitstreams, all these executions together have performed no more than*

$$O(n^9 \log n \cdot (\tau + \log n)^2)$$

bit operations in expectancy, with respect to the randomized choices of subdivision points.

The proof of this theorem will be by reduction to the primary complexity result for the bitstream Descartes algorithm, Theorem 3.61 on page 99. This result depends on four explicit parameters, namely the degree n, the width $|I_0|$ of the initial interval, the minimum distance s between any two distinct complex roots, and a tsqd-bound P', cf. Definition 3.59; implicitly, it relies on the existence of a lower bound w as in Lemma 3.62 on the width of any subdivided interval. Therefore, our task for the remainder of this section is to derive bounds on these parameters for the polynomials $F(\alpha, Y)$, ignoring those with $m \leq 1$ distinct real roots. We will tackle the parameters in increasing order of difficulty. The first one is trivial and has already been dealt with: $\deg(F_{\mu\nu}) = n$, owing to the Y-regularity of F.

There is a common theme to the way we treat most of the remaining parameters: For attaining good estimates, it is helpful not to regard one polynomial $F_{\mu\nu}(Y) \in \mathbb{Z}[\xi_{\mu\nu}][Y]$ at a time, but all of them at once, because a single one may realize a worst case that is not possible for all of them simultaneously.

We begin with the **initial intervals** that are chosen for the polynomials $F(\alpha, Y)$ through the use of a complex root bound. The hypotheses of the following proposition cover, inter alia, the dyadic Fujiwara complex root bound functional from §2.4.1 and its approximate evaluation described in Equation (3.24) on page 78.

Proposition 3.72. *Consider a functional $\Phi\colon \mathbb{R}[Y] \setminus \mathbb{R} \to \mathbb{R}_{>0}$ that obeys the bound $\max\{0, \log \Phi(G)\} = O(\log \|G\|_\infty + \log \deg(G))$ for all polynomials G with $\|G\|_\infty \geq 1$. Let $F(X, Y)$ as in (3.58) have integer coefficients $|a_{ij}| \leq 2^\tau$. Consider the polynomial $R(X) = \operatorname{Res}(F, \mathrm{D}_Y F, Y)$ as in (3.59). Summing over all distinct real roots α of $R(X)$, it holds that $\sum_\alpha \max\{0, \log \Phi(F(\alpha, Y))\} = O(n^2(\tau + \log n))$.*

Proof. Lemma 3.70 implies $\deg(R) \leq n(n-1)$ and $\log \|R\|_\infty = O(n \cdot (\tau + \log n))$. Let us consider a real root α of $R(X)$. The leading coefficient of $F(\alpha, Y)$ is $a_{n0} \in \mathbb{Z} \setminus \{0\}$, so that $\|F(\alpha, Y)\|_\infty \geq 1$. Thus, $\max\{0, \log \Phi(F(\alpha, Y))\} = O(\log \|F(\alpha, Y)\|_\infty + \log n)$.

The coefficients of $F(\alpha, Y)$ satisfy $|f_i(\alpha)| \le \sum_j |a_{ij}|\,|\alpha|^j \le (n+1)\,2^\tau \max\{1,|\alpha|\}^n$, so $\log \|F(\alpha,Y)\|_\infty \le O(\tau + \log n) + n \log \max\{1,|\alpha|\}$.

Summing over all distinct real roots α of R, which are at most $n(n-1)$ in number, we obtain $\sum_\alpha \log \|F(\alpha,Y)\|_\infty = O(n^2(\tau+\log n)) + n \log \prod_\alpha \max\{1,|\alpha|\}$. The Mahler measure of R, see Definition 3.1 on page 52, provides an upper bound for $\prod_\alpha \max\{1,|\alpha|\}$ and is itself bounded in terms of a norm by Proposition 3.4. Taking logarithms, we obtain $\log \prod_\alpha \max\{1,|\alpha|\} \le \log \mathrm{Mea}(R) \le \log \|R\|_\infty + O(\log n) = O(n \cdot (\tau + \log n))$. Altogether, we arrive at $\sum_\alpha \max\{0, \log \Phi(F(\alpha,Y))\} = \sum_\alpha O(\log \|F(\alpha,Y)\|_\infty + \log n) = O(n^2(\tau+\log n))$, as was to be shown. □

Next on our agenda is the **minimum distance** between distinct complex roots.

Proposition 3.73. *Let the polynomial $F(X,Y)$ in (3.58) have integer coefficients $|a_{ij}| \le 2^\tau$, and consider its specializations $F_{\mu\nu}(Y)$ from (3.61). Let S be an arbitrary subset of the indices $\{(\mu,\nu) \mid 1 \le \mu \le M,\ 1 \le \nu \le h_\mu\}$ such that each $F_{\mu\nu}(Y)$ with $(\mu,\nu) \in S$ has more than one distinct complex root and minimum distance $s_{\mu\nu} > 0$ between any two of them. It holds that $\sum_S \log(1/s_{\mu\nu}) = O(n^3(\tau+\log n))$.*

We prove the proposition in this general form, but we will use it specifically on the set of those real roots $\alpha = \xi_{\mu\nu}$ of $R(X)$ for which $F(\alpha,Y)$ has $m \ge 2$ distinct real roots.

Proof. For each polynomial $F_{\mu\nu}$ with $(\mu,\nu) \in S$, we invoke our generalized Davenport-Mahler bound, Theorem 3.9 from page 54, on a single edge (α,β) that realizes the minimum distance $s_{\mu\nu} = |\alpha - \beta|$ and obtain the inequality

$$\log(1/s_{\mu\nu}) \le -\log \frac{|\mathrm{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2}}{\mathrm{Mea}(F_{\mu\nu})^{n-k_\mu-1}} + O(n \log n).$$

For each polynomial $F_{\mu\nu}$ with $(\mu,\nu) \notin S$, we invoke Theorem 3.9 on the empty edge set, resulting in the product 1, and obtain the inequality

$$0 \le -\log \frac{|\mathrm{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2}}{\mathrm{Mea}(F_{\mu\nu})^{n-k_\mu-1}} + O(n \log n). \tag{3.63}$$

Summing up over all complex roots $\xi_{\mu\nu}$, of which there are $O(n^2)$ many, we obtain

$$\sum_{(\mu,\nu)\in S} \log(1/s_{\mu\nu}) \le \sum_{\mu=1}^{M} \sum_{\nu=1}^{h_\mu} \left(-\log \frac{|\mathrm{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2}}{\mathrm{Mea}(F_{\mu\nu})^{n-k_\mu-1}} \right) + O(n^3 \log n).$$

The claim reduces directly to the next proposition. □

Proposition 3.74. *Let the polynomial $F(X,Y)$ as in (3.58) have integer coefficients with magnitudes $|a_{ij}| \le 2^\tau$. For $F_{\mu\nu}(Y)$ and k_μ as in (3.61) and (3.62), resp., it holds that*

$$\sum_{\mu=1}^{M} \sum_{\nu=1}^{h_\mu} \left(-\log \frac{|\mathrm{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2}}{\mathrm{Mea}(F_{\mu\nu})^{n-k_\mu-1}} \right) = O(n^3(\tau+\log n)). \tag{3.64}$$

Proof. Let $D_\mu(X) := \mathrm{sDisc}_{k_\mu}(F,Y) \in \mathbb{Z}[X]$, so that $D_\mu(\xi_{\mu\nu}) = \mathrm{sDisc}_{k_\mu}(F_{\mu\nu}) \in \mathbb{C} \setminus \{0\}$.

We fix an arbitrary $1 \le \mu \le M$. By multiplicativity of the Mahler measure, we obtain

$$\begin{aligned}
&\prod_{\nu=1}^{h_\mu} \Big(|\mathrm{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2} \,\big/\, \mathrm{Mea}(F_{\mu\nu})^{n-k_\mu-1} \Big) \\
&\qquad = |\prod_{\nu=1}^{h_\mu} D_\mu(\xi_{\mu\nu})|^{1/2} \,\Big/\, \mathrm{Mea}(\prod_{\nu=1}^{h_\mu} F(\xi_{\mu\nu}, Y))^{n-k_\mu-1} \\
&\qquad = |\mathrm{Res}(D_\mu, R_\mu, X)/\ell_\mu^{(n-k_\mu)(n-k_\mu-1)}|^{1/2} \,\big/\, \mathrm{Mea}(\mathrm{Res}(F, R_\mu, X)/\ell_\mu^n)^{n-k_\mu-1}.
\end{aligned}$$

In the last equality, we have used Proposition 3.68 to express a product over all roots of $R_\mu(X)$ as resultant with R_μ. We have $\deg_X(D_\mu) \le (n-k_\mu)(n-k_\mu-1)$ by Lemma 3.70(ii) and $\deg_X(F) \le n$. To avoid case distinctions, we have taken the resultants with respect to these formal degrees. The total exponent of ℓ_μ is $n\cdot(n-k_\mu-1)-(1/2)(n-k_\mu)(n-k_\mu-1) = (1/2)(n+k_\mu)(n-k_\mu-1) \ge 0$, so we obtain

$$\begin{aligned}
\ldots &= |\ell_\mu|^{(1/2)(n+k_\mu)(n-k_\mu-1)} \cdot |\mathrm{Res}(D_\mu, R_\mu, X)|^{1/2} \,\big/\, \mathrm{Mea}(\mathrm{Res}(F, R_\mu, X))^{n-k_\mu-1} \\
&\ge 1 \big/ \mathrm{Mea}(\mathrm{Res}(F, R_\mu, X))^{n-2},
\end{aligned}$$

where we have estimated the exponents in the numerator from below by 0 and the exponent in the denominator from above by $n-2$. This is possible, because the integers $|\ell_\mu|$ and $|\mathrm{Res}(D_\mu, R_\mu, X)|$ and the measure of the polynomial $\mathrm{Res}(F, R_\mu, X) \in \mathbb{Z}[Y]$ are all larger than or equal to 1.

Let us now take the product over all factors R_μ. Combining the preceding estimate with the multiplicativity of the resultant (Corollary 3.69 on page 109) and the fact that $\mathrm{Mea}(A) \le \mathrm{Mea}(AB)$ for $A, B \in \mathbb{Z}[Y]$, we arrive at

$$\begin{aligned}
\prod_{\mu=1}^{M} \prod_{\nu=1}^{h_\mu} \frac{|\mathrm{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2}}{\mathrm{Mea}(F_{\mu\nu})^{n-k_\mu-1}} &\ge \prod_{\mu=1}^{M} \frac{1}{\mathrm{Mea}(\mathrm{Res}(F, R_\mu, X))^{n-2}} \\
&= 1 \big/ \mathrm{Mea}(\mathrm{Res}(F, \prod_{\mu=1}^{M} R_\mu, X))^{n-2} \\
&\ge 1 \big/ \mathrm{Mea}(\mathrm{Res}(F, R, X))^{n-2} \ge 1 \big/ \left\|\mathrm{Res}(F, R, X)\right\|_2^{n-2}
\end{aligned} \tag{3.65}$$

The last inequality holds by Proposition 3.4 from page 52.

Let us now study the logarithmic 2-norm of $S(Y) := \mathrm{Res}(F, R, X)$. We begin with a bound on the coefficient magnitudes of $R(X)$. By definition, R is the determinant of the Sylvester matrix of F and $\mathrm{D}_Y F$. The derivative $\mathrm{D}_Y F$ has degree $n-1$ in Y and coefficients of magnitudes $|i a_{ij}| \le n 2^\tau$. Thus, Lemma 3.70(iii) yields $\log \|R\|_\infty \le \log \|R\|_1 = O(n \cdot (\tau + \log n))$. We invoke Lemma 3.70(iii) again, now for F and R. We recall $\deg R \le n(n-1)$ and obtain $\log \|S\|_2 \le \log \|S\|_1 = O(n^2(\tau + \log n))$. After taking logarithms in (3.65), this estimate implies the proposition. □

For this proposition, our initial remark about the virtues of considering all $F_{\mu\nu}$ simultaneously is particularly fitting: In contrast to $\mathbb{Z}$, there is no immediate lower bound on the magnitude of a non-zero subdiscriminant in $\mathbb{Z}[\xi_{\mu\nu}]$, but by multiplying up all algebraic conjugates, necessarily including the imaginary ones, we got back to a situation with integer coefficients.

We observe at this point how the generalization of the Davenport-Mahler bound in §3.1.4 has helped to keep the preceding proof straightforward. Without it, we would have had to argue about the square-free part of $F_{\mu\nu}(Y)$. Its coefficients have closed-form expressions as Sylvester subdeterminants [BPR06, Cor. 10.15] in $f_i(X)$ evaluated at $X = \xi_{\mu\nu}$, but the resulting growth of the leading coefficient poses problems in bounding the product of Mahler measures in the denominator. The simple arguments for the integer case in the proof of Corollary 3.11 on page 57 are inapplicable, as they require a factorial ring.

Now we turn to the fourth parameter, namely the **tsqd-bound** P' for the bitstream (m, k)-Descartes algorithm, that is to say, a bound fulfilling conditions (i–iii) in Definition 3.59 for the bitstream (m, k)-Descartes algorithm and the respective subtree it traverses in any potential subdivision tree of the bitstream Descartes algorithm. We recall that the bitstream (m, k)-Descartes algorithm invariably performs breadth-first traversal.

Let $\mathcal{T}'$ be any of the subdivision trees possible for the bitstream Descartes method on a polynomial A_{in} with a unique multiple real root β and some initial interval I_0 that contains β and at least one other real root γ of A_{in}. Clearly, these root do not occur as subdivision points. Let $\mathcal{T}$ be the subdivision tree of the exact Descartes method executed with the same choice of subdivision points. Because of the unique multiple real root β, both of these trees have a unique path from the root downwards that has infinite length. This unique infinite path, seen as a sequence of intervals, is the same in $\mathcal{T}$ and $\mathcal{T}'$.

In the finite subdivision trees considered previously, any internal node was either terminal itself (i.e., a parent of two leaves, see Definition 3.18) or the ancestor of a terminal node. This has led to the distinguished role of rt-paths in the analysis of the Descartes method for polynomials with simple real roots. Now the existence of an infinite path has changed the situation. To see this, consider the extreme example of a unique infinite path that comprises all internal nodes, with the isolating intervals of simple real roots hanging off this path as leaves; in this case, not a single terminal node or rt-path exists. Thus, we first need to take a closer look at $\mathcal{T}$ before we can proceed to $\mathcal{T}'$.

Every node I on the unique infinite path in $\mathcal{T}$ has two children; one of them is again an element of the infinite path, we are interested in the other one and call it the *terminating child* of I. By the variation-diminishing property of subdivision (see Corollary 2.27 on page 27), there are only finitely many terminating children whose Descartes test is positive. Let I be the node of maximal depth on the infinite path with the property that its terminating child J has a positive Descartes test. Since A_{in} has more than one real root in I_0, such I exists; we call it the *final fork* of $\mathcal{T}$ and distinguish two cases. If the terminating child J of I is a leaf of $\mathcal{T}$ (meaning that its Descartes test is 1), then we call I a *final leaf fork*; otherwise a *final internal fork*.

In the case of a final internal fork, the terminating child J or one of its successors is a terminal node J', in fact a regular terminal node (because the unique multiple real root β is not contained in J'), and the rt-path from I_0 to J' includes the prefix of the unique infinite path down to the final fork, so that the length of this rt-path accounts for the number of nodes on the unique infinite path up to the fork, similar to the situation without a multiple real root.

In case of a final leaf fork, no such rt-path exists. (This is the crux in the "extreme example" that we regarded above.) To overcome this, we define an *rtfl-path* to be a path in $\mathcal{T}$ from the root down to a node I that is either a regular terminal node or the final leaf fork on the unique infinite path (if it exists).

Based on this notion of an rtfl-path, we call an internal node of $\mathcal{T}$ *accounted* if it sits on an rtfl-path; the other internal nodes of $\mathcal{T}$ are *unaccounted*. By construction, all unaccounted internal nodes sit on the unique infinite path.

We can now extend our primary bound on tree size, Theorem 3.19 from page 60, to the present situation. We remind ourselves that the constant $\rho = 4/3$ is a subdivision ratio bound for $\mathcal{T}$ in the sense of Definition 3.12. We denote by $\mathcal{G}_{\mathrm{mk}}(\mathcal{T})$ the directed graph on the distinct complex roots of A_{in} whose edge set consists of all pairs (α, β) that are responsible, in the sense of Definition 3.14, for subdivision of a regular terminal node or of the final leaf fork (if there is one). We emphasize that the final leaf fork I in $\mathcal{T}$, if existing, is a regular internal node, because it contains both the unique multiple real root β and the simple real root isolated by its terminating child J, and thus falls into case (iii) of Proposition 3.13 and Definition 3.14.

Theorem 3.75. *Let A_{in} be a real polynomial of degree $n \geq 2$ with a unique multiple real root β and another real root γ. Let I_0 be an open interval containing β and γ. Consider the infinite subdivision tree $\mathcal{T}$ generated by an execution of the exact Descartes method on A_{in} and I_0 in which β is not chosen as subdivision point. Let ρ be a subdivision ratio bound for $\mathcal{T}$.*

(i) The graph $\mathcal{G}_{\mathrm{mk}}(\mathcal{T})$ satisfies conditions (i–iii) of Theorem 3.9 and has at most $n/2$ edges.

(ii) The sum P of the lengths of all rtfl-paths of $\mathcal{T}$ satisfies

$$P \leq \log_\rho \Big(\prod_{(\alpha,\beta)} \frac{|I_0|}{|\alpha - \beta|} \Big) + \frac{n}{2} \log_\rho \Big(\frac{2}{\sqrt{3}} \Big), \tag{3.66}$$

with (α, β) ranging over the edges of $\mathcal{G}_{\mathrm{mk}}(\mathcal{T})$.

(iii) The number of all internal nodes of $\mathcal{T}$ that lie on an rtfl-path is at most $P + 1$.

Proof. The following arguments closely resemble the proofs of the corresponding items (i), (iii) and (iv) in Theorem 3.19.

Ad (i). No terminal node is a descendant of another, and no terminal node is a descendant or ancestor of the final leaf fork in $\mathcal{T}$, if existing. Therefore, the intervals from the definition of $\mathcal{G}_{\mathrm{mk}}(\mathcal{T})$ are pairwise disjoint, and the first claim reduces to Proposition 3.15. Each of these disjoint intervals has a Descartes test of at least 2, but according to Corollary 2.27, their sum is at most $\mathit{DescartesTest}(A_{\mathrm{in}}, I_0) \leq n$, which shows the second claim.

Ad (ii). Consider any rtfl-path $(I_0, \ldots, I_k)$. Its length k is the depth of I_k, which is bounded by Lemma 3.17(ii) in terms of the pair (α, β) responsible for subdivision of I_k as $k < \log_\rho(|I_0| \,/\, |\alpha - \beta|) + \log_\rho(2/\sqrt{3})$. Summing over all rtfl-paths, of which there are at most $n/2$, we attain the claimed bound.

Ad (iii). To each non-root node I on an rtfl-path, we associate the edge to the parent of I; this map is injective, and each edge in its image is counted at least once in P. Adding 1 to account for the root node, we obtain the claimed bound. □

Proposition 3.76. *In the situation of Theorem 3.71, consider all real roots α of $R(X)$ for which $F(\alpha, Y)$ has $m \geq 2$ distinct real roots. There exist numbers $P(\alpha) \geq 0$ such that $\sum_\alpha P(\alpha) = O(n^3(\tau + \log n))$ and $P(\alpha)$ is an upper bound on the sum of the lengths of the rtfl-paths in any possible subdivision tree $\mathcal{T}$ constructed by the exact Descartes method on $F(\alpha, Y)$ and $I_0(\alpha)$.*

Proof. Consider the statement of Theorem 3.75 for $A_{\text{in}}(Y) = F(\alpha, Y)$. We invoke Theorem 3.9 for the first term in (3.66). Using the estimate $\#E \le n/2$, we arrive at an upper bound $P(\alpha)$ for the sum of all rtfl-path lengths that is valid for all possible subdivision trees and satisfies

$$P(\alpha) \le -\log_\rho \frac{|\text{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2}}{\text{Mea}(F_{\mu\nu})^{n-k_\mu-1}} + O(n \log n) + \frac{n}{2} \max\{0, \log |I_0(\alpha)|\}.$$

It remains to demonstrate the bound on $\sum_\alpha P(\alpha)$. For the first and second term together, we use the same trick as in the proof of Proposition 3.73 and add the trivial inequalities (3.63) for all missing complex roots $\xi_{\mu\nu}$ of $R(X)$ to arrive at the upper bound

$$\sum_{\mu=1}^{M} \sum_{\nu=1}^{h_\mu} \left(-\log \frac{|\text{sDisc}_{k_\mu}(F_{\mu\nu})|^{1/2}}{\text{Mea}(F_{\mu\nu})^{n-k_\mu-1}} \right) + O(n^3 \log n),$$

which reduces to $O(n^3(\tau + \log n))$ by Proposition 3.74. According to Proposition 3.72, this bound also covers the remaining term $n/2 \cdot \sum_\alpha \max\{0, \log |I_0(\alpha)|\}$. □

We now make the transition from $P(\alpha)$ to a tsqd-bound, in analogy to Proposition 3.60.

Lemma 3.77. *Let A_{in} be a real polynomial of degree $n \ge 2$ that possesses a unique multiple real root β and another real root γ. Let $I_0 \supseteq \{\beta, \gamma\}$ be an initial interval for the bitstream Descartes algorithm on A_{in}. If $P \ge n/4$ is an upper bound on the sum of the lengths of all rtfl-paths in any subdivision tree constructed by the exact Descartes method on A_{in} and I_0 with subdivision points different from β, then $P' := 5P + 4n + 2 = O(P)$ is a tsqd-bound for the bitstream (m,k)-Descartes algorithm on A_{in} and I_0.*

The condition $P \ge n/4$ is satisfied automatically by bounds P obtained from Theorem 3.75(ii) when using the subdivision ratio bound $\rho = 4/3$.

Proof. Let $\mathcal{T}'$ be any of the subdivision trees possible for the bitstream Descartes method on A_{in} and I_0; clearly, the root β does not occur as a subdivision point. Let $\mathcal{T}$ be the subdivision tree of the exact Descartes method when making the same choice of subdivision points. The bitstream (m,k)-Descartes algorithm traverses some subtree $\mathcal{T}'_{\text{mk}}$ of $\mathcal{T}'$ in breadth-first order. There are three kinds of internal nodes in $\mathcal{T}'_{\text{mk}}$: those that occur as internal nodes in $\mathcal{T}$ and are *accounted*, those that occur as internal nodes in $\mathcal{T}$ and are *unaccounted*, and those that do not occur in $\mathcal{T}$, we call them *extraneous*. All unaccounted internal nodes lie on the unique infinite path.

We prove the lemma through a sequence of observations on $\mathcal{T}'_{\text{mk}}$.

1. The parent of an extraneous internal node is accounted.

An extraneous internal node I is an interval that has been subdivided because it had type (M1), even though its true Descartes test is 1. I is not the root node, so it has a parent J. By Lemma 3.47, J is an internal node of $\mathcal{T}$. Due to our consideration of the final fork on the infinite path, J is accounted, even if it sits on the infinite path.

2. An unaccounted node has depth at most $P + 2$.

Suppose the subdivision of an interval I at depth $d-1$ in $\mathcal{T}'$ produces an unaccounted interval I' at depth d. At the time when I is chosen from Q for subdivision, Q also records another interval J; otherwise termination condition (S) would hold. The fact that subdivision of I gets priority over subdivision of J in breadth-first traversal means that

the depth of J is no less than $d-1$. Being disjoint from I, the interval J does not contain β and thus is either accounted or extraneous. If J is accounted, then its depth is at most P, so the depth of I' is at most $P+1$. If J is extraneous, then its parent is accounted, the depth of J is at most $P+2$, and so the depth of I' is at most $P+2$.

3. The total number of accounted, unaccounted and extraneous node is at most $3P+5$. The number of unaccounted nodes is at most $P+2$, because there is at most one unaccounted node on levels $1, \ldots, P+2$. The number of accounted nodes is at most $P+1$ by Theorem 3.75(iii). They form a subtree of $\mathcal{T}$ with size at most $P+1$, so they can carry up to $P+2$ children. Each extraneous node is one such child, so there are at most $P+2$ extraneous nodes. Adding up, we arrive at $3P+5 \leq 5P+4n+2+1$ nodes in total and have thus established property (i) for the claimed tsqd-bound.

4. At any stage of the algorithm, the sum of the depths of the intervals recorded in Q is at most $5P+4n+2$.

Q may contain at most one unaccounted interval I of depth at most $P+2$. For Q without I, the corresponding argument from the proof of Proposition 3.60(ii) from page 98 carries over and yields a bound of $4P+4n$ on the sum of depths. Adding $P+2$, property (ii) for the claimed tsqd-bound is established and the lemma is proved. □

Finally, we show that a **lower bound on subdivided intervals** as in Lemma 3.62 still holds.

Lemma 3.78. *In the situation of Theorem 3.71, consider one of the polynomials $F(\alpha, Y)$ with $m \geq 2$ distinct real roots and the corresponding initial interval $I_0 = I_0(\alpha)$. If $s > 0$ is a lower bound on the distance between any two distinct complex roots of $F(\alpha, Y)$, then there exists a number $0 < w \leq s$ with $\log(|I_0|/w) = O(\log(|I_0|/s))$ such that w is a lower bound on the length of any interval subdivided in any possible execution of the bitstream (m,k)-Descartes algorithm on $F(\alpha, y)$ and I_0.*

Proof. We recall the general observation that an interval I at depth d in any possible subdivision tree satisfies $(1/4)^d |I_0| \leq |I| \leq (3/4)^d |I_0|$.

Let us first consider the case of an arbitrary interval I that does not lie on the unique infinite path and is subdivided. For this case, the arguments from the proof of Lemma 3.62 apply and yield $(\sqrt{3}/8) \cdot s < |I| \leq (3/4)^d |I_0|$ when combined with our initial observation. Thus, the depth of I is $d < d_1 := (\log(|I_0|/s) + \log(8/\sqrt{3}))/\log(4/3)$.

As a consequence of breadth-first traversal, any subdivided interval I, even if on the infinite path, has depth less than $d_1 + 1$, so that the other part of our initial observation implies $|I| > w := (1/4)^{d_1+1} |I_0|$, where $\log(|I_0|/w) = 2(d_1+1) = O(\log(|I_0|/s))$. □

We are now ready to prove the **main result** announced on page 112.

Proof of Theorem 3.71. Let $\alpha = \xi_{\mu\nu}$ be an arbitrary real root of $R(X)$, and let m denote the number of distinct real roots of the corresponding polynomial $F_{\mu\nu}(Y) = F(\alpha, Y)$.

If $m \leq 1$, no work is done; we ignore these roots completely in the rest of this proof.

If $m \geq 2$, we consider the execution of the bitstream (m,k)-Descartes algorithm on the polynomial $F(\alpha, Y)$ and the corresponding initial interval $I_0(\alpha)$. Lemma 3.77 provides a tsqd-bound $P'(\alpha) = O(P(\alpha))$, where $P(\alpha)$ is the bound from Proposition 3.76. We write $s(\alpha)$ for the minimum distance between any two distinct complex roots of $F_{\mu\nu}$.

Lemma 3.78 shows that the statement of Lemma 3.62 also holds in the present situation (albeit with a different constant hidden in the O-notation), so that Theorem 3.61 extends to the present situation. It shows that the bitstream (m,k)-Descartes algorithm needs

$O(n^3 \log n \cdot P'(\alpha) \cdot (\log(|I_0(\alpha)|/s(\alpha)) + \log n))$ bit operations in expectancy. Summing over all α that make $m \geq 2$, we obtain

$$\sum_{\alpha} O\big(n^3 \log n \cdot P'(\alpha) \cdot (\log(|I_0(\alpha)|/s(\alpha)) + \log n)\big) \\ \leq O(n^3 \log n) \cdot \big(\sum_{\alpha} P'(\alpha)\big) \cdot \big(\sum_{\alpha} \log(|I_0(\alpha)|/s(\alpha)) + \log n\big),$$

where the inequality follows from $\sum_i a_i b_i \leq (\sum_i a_i)(\sum_i b_i)$ for $a_i, b_i \in \mathbb{R}_{\geq 0}$. The second factor is $O(\sum_\alpha P(\alpha)) = O(n^3(\tau + \log n))$ by Proposition 3.76. The third factor is $O(n^3(\tau + \log n))$ by Propositions 3.72 and 3.73. Altogether, this entails the claimed bound $O(n^9 \log n \cdot (\tau + \log n)^2)$. □

3.4.4 Outlook: subdivision tree size with several multiple roots

In this section, we study the following theoretical variant of the exact Descartes method for a real polynomial A_{in} of degree $n \geq 2$ that may possess several multiple real roots. An open initial interval I_0 is subdivided recursively until each subinterval I satisfies one of the following conditions:

(i) I does not contain a root of A_{in} and it holds that $\mathit{DescartesTest}(A_{\text{in}}, I) = 0$;

(ii) I contains a unique root of A_{in}, whose multiplicity we denote by k, and it holds that $\mathit{DescartesTest}(A_{\text{in}}, I) = k$.

Throughout this section, we let $\mathcal{T}$ denote the recursion tree constructed by this method for A_{in} and I_0 with some arbitrary choice of subdivision parameters $1/4 \leq \alpha \leq 3/4$, and it is our goal to give bounds on $\mathcal{T}$ in the style of our previous treatment of polynomials with simple real roots (§3.1.5). The task of analyzing $\mathcal{T}$ is a natural generalization of analyzing the recursion tree in the failure case of the (m, k)-Descartes algorithm.

It is instructive to compare this task to the topic of the previous section: analysis of the success case in the bitstream (m, k)-Descartes algorithm. For that case, we did not need to know much about the effects of multiple real roots, except that a unique multiple real root produces a unique infinite path in the subdivision tree. Beyond that, it was sufficient to analyze how quickly the $m - 1$ intervals for the simple real roots have been found and all intervals free of roots have disappeared, similar to the case of simple real roots. We did not even need a partial converse to Descartes' rule for the case of more than one sign variation. Moreover, to cope with approximate coefficients, it was good enough that the lemmas from §3.3.4 guaranteed eventual certainty about zero or one sign variations; a sufficient condition for $\max \mathrm{var}_\varepsilon(\ldots) \leq k$ with $k > 1$ was not needed. By contrast, such a lemma is required for an analysis of the failure case in the bitstream (m, k)-Descartes algorithm. In the present thesis, this problem it not addressed; hence our restriction to a setting with exact arithmetic for this section. Instead, we treat the other problem pertinent to this analysis: How can we bound the size of $\mathcal{T}$ in the style of §3.1.5 by distances between pairs of roots such that Theorem 3.9, the generalized Davenport-Mahler bound from §3.1.4, remains applicable?

In our previous treatment of simple real roots, an interval I containing a single real root was subdivided only if $\mathit{DescartesTest}(A_{\text{in}}, I)$ was driven beyond the value 1 by a pair of complex-conjugate roots in the Obreshkoff range $\mathrm{OL}_{\leq}(1, 1, n; I)$, see Theorem 2.32 on page 31. Using the particular geometry of $\mathrm{OL}_{\leq}(1, 1, n; I)$, we showed that no two non-adjacent subintervals of I_0 are affected by the same complex-conjugate pair in this way

(Lemma 3.16); this allowed us to justify our choice of pairs for use in the Davenport-Mahler bound (Proposition 3.15). However, in the present setting, this distinguished role of $\mathrm{OL}_{\leq}(1,1,n;I)$ is lost; we would have to consider the larger ranges $\mathrm{OL}_{\leq}(k,k,n;I)$ in its place. We avoid this; instead, we use this occasion to show what one can do with the Davenport-Mahler bound when considering the nearest neighbour of each real root irrespective of specific regions of influence.

To each real root ϑ of A_{in}, we associate the nearest other root $N(\vartheta) \neq \vartheta$ of A_{in} (arbitrarily in case that several other roots realize the minimum distance). If an imaginary root ζ is to be chosen as $N(\vartheta)$, we make the additional requirement that $\operatorname{Im} N(\vartheta) > 0$ if $\vartheta \leq \operatorname{Re} N(\vartheta)$ and $\operatorname{Im} N(\vartheta) < 0$ otherwise. This is no restriction, since imaginary roots of A_{in} occur in complex-conjugate pairs with equal distance to any real root ϑ.

Proposition 3.79. *If I is an internal node of $\mathcal{T}$, then exactly one of the following two conditions holds:*

(i) I contains a real root ϑ of A_{in} such that $|N(\vartheta) - \vartheta| < |I| \cdot O(n^2)$.

(ii) I does not contain a real root of A_{in}, and there is a complex-conjugate pair of imaginary roots $\xi \pm \mathrm{i}\eta$ of A_{in} such that $\xi \in I$ and $0 < |(\xi + \mathrm{i}\eta) - (\xi - \mathrm{i}\eta)| < |I|$.

Proof. Ad (i). This is obtained immediately as the contrapositive of either Proposition 2.36 from page 33 (using the estimate from Proposition 2.35(iv) and $\Theta(n) = O(n^2)$), or Proposition 2.46(ii) from page 38.

Ad (ii). This is obtained immediately as the contrapositive of the "one-circle theorem", Proposition 2.33 on page 32. □

With reference to this case distinction, we declare one pair of roots to be *responsible for subdivision* of an internal node I: either $(\vartheta, N(\vartheta))$ as in (i) or $(\xi - \mathrm{i}\eta, \xi + \mathrm{i}\eta)$ as in (ii). If two nodes of $\mathcal{T}$ are disjoint intervals, then they are assigned different pairs; but it may happen that this difference lies only in the order of the pairs' components, namely if two real roots are their mutual nearest neighbours.

We can now formulate an analogue to Lemma 3.17.

Lemma 3.80. *Consider an internal node I of $\mathcal{T}$ at depth $d \geq 0$. If (α, β) is responsible for subdivision of I, then $d < \log_\rho(|I_0| \,/\, |\alpha - \beta|) + O(\log n)$, where $\rho \geq 4/3$ is a subdivision ratio bound for $\mathcal{T}$.*

Proof. By choice of (α, β), it holds that $|\alpha - \beta| < |I| \cdot O(n^2) \leq |I_0| \,/ \rho^d \cdot O(n^2)$. The claim is proved by solving for d. □

As before, we call an internal node of $\mathcal{T}$ *terminal* if both of its children are leaves, and a path from the root down to a terminal node we call a *terminal path*. The terminal nodes of $\mathcal{T}$ are pairwise disjoint intervals, and each of them has a Descartes test of at least 2. Thus, by the variation-diminishing property of subdivision (see Corollary 2.27 on page 27), their number is at most $n/2$. Summing over all terminal nodes and the pairs (α, β) responsible for their subdivision, we obtain that the sum of the lengths of all terminal paths is less than $\log_\rho \prod(|I_0| \,/\, |\alpha - \beta|) + O(n \log n)$, similar to our previous approach.

However, we form a product over a set of edges which does, in general, not satisfy the conditions of Theorem 3.9. We need to partition this edge set suitably; one possibility is given by the following classification of pairs of roots responsible for subdivision:

- a pair $(\xi - \mathrm{i}\eta, \xi + \mathrm{i}\eta)$ is called *imaginary*,
- a pair $(\vartheta, N(\vartheta))$ with $N(\vartheta) \notin \mathbb{R}$ is called *complex*,
- a pair $(\vartheta, N(\vartheta))$ with $N(\vartheta) \in \mathbb{R}$ is called *outward* if $|\vartheta| < |N(\vartheta)|$ and *inward* if $|N(\vartheta)| < |\vartheta|$; in the special case $\vartheta = -N(\vartheta)$, we call the pair inward if $\vartheta > 0$ and outward if $\vartheta < 0$.

This classification, in particular the last case, makes sure that no two distinct terminal nodes give rise to the same pair of roots within one class, not even in reversed order.

For each of the four classes separately, we define a directed graph on the distinct complex roots of A_{in} whose edges are those pairs responsible for subdivision of a terminal node that have the respective class; we denote these graphs by $\mathcal{G}_{\mathrm{im}}$, $\mathcal{G}_{\mathrm{c}}$, $\mathcal{G}_{\mathrm{ro}}$, and $\mathcal{G}_{\mathrm{ri}}$. Each of these four graphs taken on its own satisfies the conditions of Theorem 3.9, perhaps after some edge reversals. For the first three, this is obvious.

Lemma 3.81. *The graph $\mathcal{G}_{\mathrm{im}}$, the graph $\mathcal{G}_{\mathrm{ro}}$, as well as the graph $\mathcal{G}_{\mathrm{ri}}$ with all edges reversed, satisfy conditions (i–iii) in Theorem 3.9.*

For the last graph $\mathcal{G}_{\mathrm{co}}$, this is not entirely trivial.

Lemma 3.82. *The graph $\mathcal{G}_{\mathrm{co}}$ with edges (α, β) reoriented such that $|\alpha| \le |\beta|$ satisfies conditions (i–iii) in Theorem 3.9.*

Proof. Any edge in $\mathcal{G}_{\mathrm{co}}$ connects a real root ϑ and an imaginary root ζ; in particular, there are no cycles of length 1. By construction, every real root has at most one incident edge. To prove the claim, it suffices to show that every imaginary root has at most one incident edge as well. Suppose the real roots $\vartheta < \vartheta'$ are both adjacent to the imaginary root ζ. We may assume w.l.o.g. that $\operatorname{Im}\zeta > 0$. Then, by construction, $\vartheta < \vartheta' \le \operatorname{Re}\zeta$ and so $|\vartheta - \vartheta'| < |\vartheta - \zeta|$, a contradiction to $N(\vartheta) = \zeta$. □

Our preceding deliberations have proved the following analogue to Theorem 3.19.

Theorem 3.83. *Consider the subdivision tree $\mathcal{T}$ defined at the beginning of the section.*

(i) $\mathcal{T}$ is finite and all its nodes lie on a terminal path.

(ii) There are four sets $E_1, \dots, E_4$ of pairs of roots of A_{in} with total cardinality not exceeding $n/2$ such that the graph on the distinct complex roots of A_{in} with edge set E_i satisfies conditions (i–iii) of Theorem 3.9 for any $1 \le i \le 4$.

(iii) The sum P of the lengths of all terminal paths of $\mathcal{T}$ satisfies

$$P \le \sum_{i=1}^{4} \log_\rho \Big(\prod_{E_i} \frac{|I_0|}{|\alpha - \beta|} \Big) + O(n \log n), \tag{3.67}$$

with each product ranging over the edges $(\alpha, \beta) \in E_i$.

(iv) The number of all internal nodes that lie on a terminal path is at most $P + 1$.

The usual Descartes method for isolating simple real roots is a special case of its theoretical variant considered here. In particular, Theorem 3.83 can be used as a substitute for Theorem 3.19 in the analysis of the Descartes method. Our new theorem is worse only by a constant factor: When we estimate (3.67) with the generalized Davenport-Mahler bound, its dominant factor, that is $-\log(\mathrm{sDisc}_{n-r}(A_{\mathrm{in}})^{1/2} / \mathrm{Mea}(A_{\mathrm{in}})^{n-r-1})$, now comes in multiplied by four. But of course, this is irrelevant for the asymptotic bounds in a complexity analysis.

We point out that the factor $O(n^2)$ from Proposition 3.79(i) turned into $O(\log n)$ in Lemma 3.80, and ended up as the $O(n \log n)$ term in (3.67), which is codominant with the $O(n \log n)$ term that comes in anyway when estimating (3.67) with the Davenport-Mahler bound. Any power of n in Proposition 3.79(i) would lead to the same bound on P in O-notation. In particular, the partial converse from §2.3.3, which gave rise to this factor $O(n^2)$, works just as well as Obreshkoff's partial converse from §2.3.2 that seemed superior during the comparison in §2.3.5.

Appendix A

Additions

A.1 Subdivision of $(0,\infty)$ and the Budan-Fourier Theorem

We recall our discussion of polar forms from §2.2.2 and take a homogeneous polar form F of degree n. Evaluating it as $F\big[\binom{0}{1}^{n-i}\binom{1}{0}^{i}\big]$, see (2.5), gives the coefficient of $\binom{n}{i}X^i$ in the dehomogenized polynomial $F(X,1)$. This resembles Proposition 2.20(i), and in this regard, the basis $(\binom{n}{i}X^i)_{i=0}^n$ parallels a "Bernstein basis" for the interval $(0,\infty)$. The purpose of this section is to explore the analogue of de Casteljau's algorithm arising from this parallelism. In particular, its variation-diminishing property will turn out to be the classical Budan-Fourier Theorem.

```
 1: procedure HomogDeCasteljau((f_0, ..., f_n), m)
 2:     (f_{0,0}, f_{0,1}, ..., f_{0,n}) ← (f_0, ..., f_n);    // input goes to top side
 3:     for j from 1 to n do
 4:         for i from 0 to n − j do
 5:             f_{j,i} ← f_{j−1,i} + m f_{j−1,i+1};
 6:         od;
 7:     od;
 8:     (f'_0, f'_1, ..., f'_n) ← (f_{0,0}, f_{1,0}, ..., f_{n,0});     // left side
 9:     (f''_0, f''_1, ..., f''_n) ← (f_{n,0}, f_{n−1,1}, ..., f_{0,n}); // right side
10:     return ((f'_j)_{j=0}^n, (f''_i)_{i=0}^n);
11: end procedure;
```

Proposition A.1. *Let $m \neq 0$. Let F be a homogeneous polar form of degree n with dehomogenized diagonal $F(X) = \sum_{i=0}^n f_i\binom{n}{i}X^i$. Consider the execution of the procedure above invoked as $((f'_j)_j, (f''_i)_i) \leftarrow$ HomogDeCasteljau$((f_i)_{i=0}^n, m)$.*

(i) We have $f_{j,i} = F\big[\binom{0}{1}^{n-(i+j)}\binom{m}{1}^{j}\binom{1}{0}^{i}\big]$ for $0 \le j \le n$ and $0 \le i \le n-j$.

(ii) $(X+1)^n F(m/(X+1)) = \sum_{i=0}^n f'_{n-i}\binom{n}{i}X^i$ and $F(X+m) = \sum_{i=0}^n f''_i\binom{n}{i}X^i$.

Proof. Claim (i) follows by induction, using the multiaffinity of F. Both equations in (ii) follow from Lemma 2.13, using

$$f'_{n-i} = F\big[\binom{0}{1}^{i}\binom{m}{1}^{n-i}\big] = F\big[(M'\binom{1}{0})^i(M'\binom{0}{1})^{n-i}\big] \quad \text{with} \quad M' = \left(\begin{smallmatrix}0 & m\\ 1 & 1\end{smallmatrix}\right),$$
$$f''_i = F\big[\binom{1}{0}^{i}\binom{m}{1}^{n-i}\big] = F\big[(M''\binom{1}{0})^i(M''\binom{0}{1})^{n-i}\big] \quad \text{with} \quad M'' = \left(\begin{smallmatrix}1 & m\\ 0 & 1\end{smallmatrix}\right). \qquad \square$$

For $m = 1$, we can thus compute $(X+1)^n F(1/(X+1))$ and $F(X+1)$ simultaneously with $(n+1)n/2$ additions and no multiplications, provided that F is represented in the basis $(\binom{n}{i}X^i)_{i=0}^n$. This may be useful for implementing the subdivision step in the Continued Fractions method (see §3.1.2).

The recurrence in *HomogDeCasteljau*$(\,\cdot\,,1)$, namely $f_{j,i} \leftarrow f_{j-1,i} + f_{j-1,i+1}$, resembles that of *DeCasteljau*$(\,\cdot\,,1/2)$, which is $b_{j,i} \leftarrow (b_{j-1,i}+b_{j-1,i+1})/2$. Thus, if $f_{0,i} = b_{0,n-i}$, then $f_{j,i} = 2^j b_{j,n-i}$. This is reflected by the different statements of Proposition A.1(ii) above and Proposition 2.28(ii) on page 28: *HomogDeCasteljau*$(\,\cdot\,,1)$ performs just a translation $X \leftarrow X + 1$; *DeCasteljau*$(\,\cdot\,,1/2)$ additionally scales the coefficients and thus performs a subsequent homothetic transformation $X \leftarrow 2X$.

Proposition A.2. *Let $m > 0$. Let F be a homogeneous polar form of degree n. Let $k \geq 0$ denote the multiplicity of $[m:1]$ as a root of the homogeneous polynomial $F(X,Y)$. Then*

$$\operatorname{var}((F\big[\tbinom{0}{1}^{n-i}\tbinom{1}{0}^{i}\big])_{i=0}^{n}) \geq \operatorname{var}((F\big[\tbinom{0}{1}^{n-i}\tbinom{m}{1}^{i}\big])_{i=0}^{n}) + k + \operatorname{var}((F\big[\tbinom{m}{1}^{n-i}\tbinom{1}{0}^{i}\big])_{i=0}^{n}).$$

The difference between both sides is an even number.

Proof by reference. The proof is completely analogous to that of Proposition 2.26 (page 26), with the following changes: Instead of de Casteljau's algorithm and the array $b_{j,i}$, we take its homogeneous counterpart and the array $f_{j,i}$ from above, whose elements satisfy $f_{j,i} = 1 \cdot f_{j-1,i} + m \cdot f_{j-1,i+1}$ with positive factors 1 and m. The invocation of Proposition 2.15 uses $S = \{\binom{0}{1}, \binom{1}{0}\}$. □

Before Jacobi [Jac35, IV.] published the generalization of Descartes' Rule to an arbitrary affine open interval (see §2.2.3) that we used throughout Chapters 2 and 3, already Budan and Fourier had extended Descartes' Rule in that direction. Their bound is different, we discuss it now. (For history and citations of the original sources see [RS02, §10.7].) Originally, Budan and Fourier considered an open interval (c,m), but we give their result in the refined form due to Hurwitz [Hur12][1] that is precise about the asymmetric roles of the interval endpoints.

Theorem A.3 (Budan-Fourier). *Let $F(X)$ be a polynomial of degree n with real coefficients that has exactly p roots in the open interval (c,m), counted with multiplicities, and a k-fold root at m. Let $w = \operatorname{var}(F(c), F'(c), \ldots, F^{(n)}(c)) - \operatorname{var}(F(m), F'(m), \ldots, F^{(n)}(m))$. Then $w \geq p + k$ and $w \equiv p + k \pmod 2$. If all roots of F are real, then $w = p + k$.*

Descartes' Rule of Signs is often presented as a corollary to this theorem for $c = 0$ and $m \to \infty$; see, e.g., [RS02, §10.1] [BPR06, §2.2.1]. Following Schoenberg[2] [Sch34], we take the opposite point of view and use the variation-diminishing property of subdivision to reduce the Budan-Fourier Theorem to Descartes' Rule.

Proof of Theorem A.3. Observe that $(c,m] = (c,\infty) \setminus (m,\infty)$. In the special case that all roots of F are real, this theorem is an immediate consequence of Descartes' Rule, which gives the exact numbers of roots in (c,∞) and (m,∞), so we can just subtract them. In the presence of imaginary roots, however, Descartes' Rule only gives upper bounds, and an argument is needed why the difference of two upper bounds gives an upper bound for the difference.

By a suitable translation of the indeterminate X, we can reduce to the special case $c = 0$. We can express w using Lemma 2.14 as follows:

$$w = \operatorname{var}((F\big[\tbinom{0}{1}^{n-i}\tbinom{1}{0}^{i}\big])_{i=0}^{n}) - \operatorname{var}((F\big[\tbinom{m}{1}^{n-i}\tbinom{1}{0}^{i}\big])_{i=0}^{n}).$$

[1] The main achievement of [Hur12] is to generalize the Budan-Fourier theorem to holomorphic functions.

[2] Isaac Jacob Schoenberg (1903–1990), Romanian-American mathematician widely celebrated as "father of splines". Interest in Descartes' Rule led to his pioneering work [Sch30] on variation-diminishing transformations. The *Journal of Approximation Theory* reports on his life and achievements in volumes **8** (1973), issue 1, pp. vi–ix, and **63** (1990), issue 1, pp. 1–2.

Thus, Proposition A.2 implies that

$$w \geq \operatorname{var}((F\big[\tbinom{0}{1}^{n-i}\tbinom{m}{1}^{i}\big])_{i=0}^{n}) + k$$

with an even difference. On the other hand, Corollary 2.18 yields that

$$v := \operatorname{var}((F\big[\tbinom{0}{1}^{n-i}\tbinom{m}{1}^{i}\big])_{i=0}^{n}) \geq p.$$

with an even difference. Combining these two statements, the claim follows. □

Corollary A.4. *Let v, w and k be as above. Then $w - k \geq v$ and $w - k \equiv v \pmod 2$.*

For general degree n, this proof of the Budan-Fourier Theorem and its pivotal inequality $w \geq v$ were first given by Schoenberg [Sch34]. He also uses the variation-diminishing property of subdivision, but lacking de Casteljau's algorithm, he has to prove it first in a rather technical fashion, drawing on his famous result about variation-diminishing linear transformations [Sch30]. Also, Schoenberg's argument, just like the formulation of the variation-diminishing property common today (see, e.g., [BPR06, Prop. 10.41]), does not account precisely for the contribution of the multiplicity k. We have achieved this through the respective improvement in our twin Propositions 2.26 and A.2.

We conclude with a concrete example for which the Budan-Fourier theorem counts too much: $F(X) = X^3 + X$. The sequence $(F(x), \ldots, F'''(x))$ exhibits the sign pattern $(-,+,-,+)$ for $x < 0$ and $(+,+,+,+)$ for $x > 0$. Hence $w = 3$ for any interval $(c, m]$ containing the simple root 0 in its interior. By contrast, $v = 1 = p$ if the interval (c, m) around zero is small enough, as we saw in §2.3.

Appendix B

Bibliography

[ACM84a] Dennis S. Arnon, George E. Collins, and Scott McCallum. Cylindrical algebraic decomposition I: the basic algorithm. *SIAM J. Computing* **13** (1984), 865–877. Reprinted in Caviness and Johnson [CJ98], pp. 136–151.

[ACM84b] Dennis S. Arnon, George E. Collins, and Scott McCallum. Cylindrical algebraic decomposition II: an adjacency algorithm for the plane. *SIAM J. Computing* **13** (1984), 878–889. Reprinted in Caviness and Johnson [CJ98], pp. 152–165.

[AG98] Alberto Alesina and Massimo Galuzzi. A new proof of Vincent's theorem. *L'Enseignement Mathématique* **44** (1998), 219–256.

[Akr86] Alkiviadis G. Akritas. There is no "Uspensky's method". *Proc. 1986 ACM Symposium on Symbolic and Algebraic Computation (SYMSAC 1986)*, pp. 88–90. ACM, 1986.

[And99] James W. Anderson. *Hyperbolic Geometry*. Springer, 1999.

[AS00] Pankaj K. Agarwal and Micha Sharir. Arrangements and their applications. *Handbook of Computational Geometry*, edited by J.-R. Sack and J. Urrutia, pp. 49–119. Elsevier, 2000.

[AS05] A. G. Akritas and A. W. Strzeboński. A comparative study of two real root isolation methods. *Nonlinear Analysis: Modelling and Control* **10** (2005), 297–304.

[ASU75] A. V. Aho, K. Steiglitz, and J. D. Ullman. Evaluating polynomials at fixed sets of points. *SIAM J. Computing* **4** (1975), 533–539.

[ASV07] Alkiviadis G. Akritas, Adam W. Strzeboński, and Panagiotis S. Vigklas. Advances on the continued fractions method using better estimations of positive root bounds. *Computer Algebra in Scientific Computing, 10th Internat. Workshop, CASC 2007*, vol. 4770 of *LNCS*, pp. 24–30. Springer, 2007.

[Bat99] Prashant Batra. *Abschätzungen und Iterationsverfahren für Polynom-Nullstellen*. Ph.D. thesis, Technical University Hamburg-Harburg, Germany, 1999.

[Bat04] Prashant Batra. A property of the nearly optimal root-bound. *J. Computational and Applied Mathematics* **167** (2004), 489–491.

[BEH+05] E. Berberich, A. Eigenwillig, M. Hemmer, S. Hert, L. Kettner, K. Mehlhorn, J. Reichel, S. Schmitt, E. Schömer, and N. Wolpert. EXACUS: Efficient and exact algorithms for curves and surfaces. *Algorithms, ESA 2005*, vol. 3669 of *LNCS*, pp. 155–166. Springer, 2005.

[Ber12] S. Bernstein. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Soobshcheniia Khar'kovskago Matematicheskago Obshchestva (= Communications de la Société Mathématiques de Kharkow)* **13** (1912), 1–2. In French. Available from `http://www.math.technion.ac.il/hat/`.

[BF00] Dario Andrea Bini and Giuseppe Fiorentino. Design, analysis, and implementation of a multiprecision polynomial rootfinder. *Numerical Algorithms* **23** (2000), 127–173.

[BK07] Eric Berberich and Lutz Kettner. Linear-time reordering in a sweep-line algorithm for algebraic curves intersecting in a common point. Research Report MPI-I-2007-1-001, Max-Planck-Institut für Informatik, 66123 Saarbrücken, Germany, 2007. `http://domino.mpi-inf.mpg.de/internet/reports.nsf/NumberView/2007-1-001`.

[BK08] Eric Berberich and Michael Kerber. Exact arrangements on tori and Dupin cyclides. *Proc. 2008 ACM Symposium on Solid and Physical Modeling (SPM 2008)*. ACM, 2008. To appear.

[BKS08] Eric Berberich, Michael Kerber, and Michael Sagraloff. Exact geometric-topological analysis of algebraic surfaces. *Proc. 24th Annual Symposium on Computational Geometry (SCG 2008)*. ACM, 2008. To appear.

[BM99] Wolfgang Boehm and Andreas Müller. On de Casteljau's algorithm. *Computer Aided Geometric Design* **16** (1999), 587–605.

[BM04] Yann Bugeaud and Maurice Mignotte. On the distance between roots of integer polynomials. *Proc. Edinburgh Mathematical Society (2nd Ser.)* **47** (2004), 553–556.

[BPR06] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*. Springer, 2nd ed., 2006. Electronic versions at `http://www.math.gatech.edu/~saugata/bpr-posted1.html`.

[CA76] George E. Collins and Alkiviadis G. Akritas. Polynomial real root isolation using Descarte's [*sic!*] rule of signs. *Proc. 1976 ACM Symposium on Symbolic and Algebraic Computation (SYMSAC 1976)*, pp. 272–275. ACM, 1976.

[Cau21] Augustin-Louis Cauchy. *Cours d'Analyse de l'École Royale Polytechnique*, vol. 1: Analyse algébrique. Debure Frères, Paris, 1821. Reprinted as: *Œuvres complètes d'Augustin Cauchy, II^{e} Série*, vol. 3. Gauthier-Villars, Paris, 1897.

[Cau29] Augustin-Louis Cauchy. *Exercices de Mathématiques*, vol. 4, chap. 7: Sur la résolution des équations numériques et sur la théorie de l'élimination. De Bure Frères, Paris, 1829. Reprinted in: *Œuvres complètes d'Augustin Cauchy, II^{e} Série*, vol. 9, pp. 87–161. Gauthier-Villars, Paris, 1891.

[CJ98] B. F. Caviness and J. R. Johnson (eds.). *Quantifier Elimination and Cylindrical Algebraic Decomposition*. Springer, 1998.

[CJK02] George E. Collins, Jeremy R. Johnson, and Werner Krandick. Interval arithmetic in cylindrical algebraic decomposition. *J. Symbolic Computation* **34** (2002), 145–157.

[Coh93] Henri Cohen. *A Course in Computational Algebraic Number Theory*, vol. 138 of *GTM*. Springer, 1993.

[Col75] George E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. *Automata Theory and Formal Languages, 2nd GI Conference*, vol. 33 of *LNCS*, pp. 134–183. Springer, 1975. Reprinted with corrections in Caviness and Johnson [CJ98], pp. 85–121.

[Dav85] James H. Davenport. Computer algebra for cylindrical algebraic decomposition. Tech. Rep., Royal Inst. of Technology, Dept. of Numer. Analysis and Computing Science, Stockholm, Sweden, 1985. Reprinted as Tech. Rep. 88-10, U. of Bath, School of Math. Sciences, Bath, England. `http://www.bath.ac.uk/~masjhd/TRITA.pdf`.

[Dim98] Dimitar K. Dimitrov. A refinement of the Gauss-Lucas theorem. *Proc. American Mathematical Society* **126** (1998), 2065–2070.

[DSY07] Zilin Du, Vikram Sharma, and Chee K. Yap. Amortized bound for root isolation via Sturm sequences. *Symbolic-Numeric Computation*, edited by Dongming Wang and Lihong Zhi, Trends in Mathematics, pp. 113–129. Birkhäuser, 2007.

[Eig07] Arno Eigenwillig. On multiple roots in Descartes' rule and their distance to roots of higher derivatives. *J. Computational and Applied Mathematics* **200** (2007), 226–230.

[EK08] Arno Eigenwillig and Michael Kerber. Exact and efficient 2D-arrangements of arbitrary algebraic curves. *Proc. 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2008)*, pp. 122–131. ACM/SIAM, 2008.

[EKK+05] Arno Eigenwillig, Lutz Kettner, Werner Krandick, Kurt Mehlhorn, Susanne Schmitt, and Nicola Wolpert. A Descartes algorithm for polynomials with bitstream coefficients. *Computer Algebra in Scientific Computing, 8th Internat. Workshop, CASC 2005*, vol. 3718 of *LNCS*, pp. 138–149. Springer, 2005.

[EKSW06] Arno Eigenwillig, Lutz Kettner, Elmar Schömer, and Nicola Wolpert. Exact, efficient, and complete arrangement computation for cubic curves. *Computational Geometry* **35** (2006), 36–73.

[EKW07] Arno Eigenwillig, Michael Kerber, and Nicola Wolpert. Fast and exact geometric analysis of real algebraic plane curves. *Proc. 2007 Internat. Symposium on Symbolic and Algebraic Computation (ISSAC 2007)*, pp. 151–158. ACM, 2007.

[EMT06] I. Z. Emiris, B. Mourrain, and E. Tsigaridas. Real algebraic numbers: Complexity analysis and experimentations. Research Report 5897, INRIA, 2006. `http://www.inria.fr/rrrt/rr-5897.html`.

[ESY06] Arno Eigenwillig, Vikram Sharma, and Chee K. Yap. Almost tight recursion tree bounds for the Descartes method. *Proc. 2006 Internat. Symposium on Symbolic and Algebraic Computation (ISSAC 2006)*, pp. 71–78. ACM, 2006.

[Far97] Gerald Farin. *Curves and Surfaces for Computer-Aided Geometric Design*. Academic Press, 4th ed., 1997.

[Fuj16] Matsusaburô Fujiwara. Über die obere Schranke des absoluten Betrages der Wurzeln einer algebraischen Gleichung. *Tôhoku Mathematical Journal (1st Ser.)* **10** (1916), 167–171.

[Ger04] Jürgen Gerhard. *Modular Algorithms in Symbolic Summation and Symbolic Integration*, vol. 3218 of *LNCS*. Springer, 2004.

[Gib98] C. G. Gibson. *Elementary Geometry of Algebraic Curves: an Undergraduate Introduction*. Cambridge University Press, 1998.

[GO04] Jacob E. Goodman and Joseph O'Rourke (eds.). *Handbook of Discrete and Computational Geometry*. Chapman&Hall/CRC, 2nd ed., 2004.

[Hal04] Dan Halperin. Arrangements. Goodman and O'Rourke [GO04], pp. 529–562.

[Hen74] Peter Henrici. *Applied and Computational Complex Analysis*, vol. 1. Wiley, 1974.

[Hoa62] C. A. R. Hoare. Quicksort. *The Computer Journal* **5** (1962), 10–15.

[Hur12] A. Hurwitz. Über den Satz von Budan-Fourier. *Mathematische Annalen* **71** (1912), 584–591.

[Jac35] C. G. J. Jacobi. Observatiunculae ad theoriam aequationum pertinentes. *J. für die reine und angewandte Mathematik* **13** (1835), 340–352.

[Jac09] C. Jaccottet. Une démonstration du théorème de Descartes. *L'Enseignement Mathématique* **11** (1909), 118–120.

[JK97] J. R. Johnson and Werner Krandick. Polynomial real root isolation using approximate arithmetic. *Proc. 1997 Internat. Symposium on Symbolic and Algebraic Computation (ISSAC 1997)*, pp. 225–232. ACM, 1997.

[JKL+06] Jeremy R. Johnson, Werner Krandick, Kevin Lynch, David G. Richardson, and Anatole D. Ruslanov. High-performance implementations of the Descartes method. *Proc. 2006 Internat. Symposium on Symbolic and Algebraic Computation (ISSAC 2006)*, pp. 154–161. ACM, 2006.

[JKR05] Jeremy R. Johnson, Werner Krandick, and Anatole D. Ruslanov. Architecture-aware classical Taylor shift by 1. *Proc. 2005 Internat. Symposium on Symbolic and Algebraic Computation (ISSAC 2005)*, pp. 200–207. ACM, 2005.

[Joh91] Jeremy R. Johnson. *Algorithms for Polynomial Real Root Isolation*. Ph.D. thesis, Ohio State University, Columbus, OH, USA, 1991.

[Joh98] J. R. Johnson. Algorithms for polynomial real root isolation. Caviness and Johnson [CJ98], pp. 269–299.

[Ker0X] Michael Kerber. Ph.D. thesis, Saarland University, Saarbrücken, Germany, 200X. In preparation.

[Kio86] John B. Kioustelidis. Bounds for positive roots of polynomials. *J. Computational and Applied Mathematics* **16** (1986), 241–244.

[KM06] Werner Krandick and Kurt Mehlhorn. New bounds for the Descartes method. *J. Symbolic Computation* **41** (2006), 49–66.

[Knu69] Donald E. Knuth. *The Art of Computer Programming*, vol. 2: Seminumerical Algorithms. Addison-Wesley, 1969.

[Knu76] Donald E. Knuth. Big omicron and big omega and big theta. *SIGACT News* **8** (1976), 18–24.

[Knu97] Donald E. Knuth. *The Art of Computer Programming*, vol. 1: Fundamental Algorithms. Addison-Wesley, 3rd ed., 1997.

[Kra95] Werner Krandick. Isolierung reeller Nullstellen von Polynomen. *Wissenschaftliches Rechnen*, edited by Jürgen Herzberger, pp. 105–154. Akademie-Verlag, 1995.

[Lag69] J. L. Lagrange. Sur la résolution des équations numériques. *Mémoires de l'Académie royale des Sciences et Belles-Lettres de Berlin* **23** (1769). Reprinted in: J.-A. Serret (ed.), *Œuvres de Lagrange*, vol. 2, pp. 539–578. Gauthier-Villars, Paris, 1868.

[Lag08] J. L. Lagrange. *Traité de la résolution des équations numériques de tous les degrés*. Courcier, Paris, new ed., 1808. Reprinted as vol. 8 of: J.-A. Serret (ed.), *Œuvres de Lagrange*, Gauthier-Villars, Paris, 1879.

[Lip41a] István Lipka. Néhány új jelváltási tétel (Über einige Sätze von Zeichenwechsel). *Matematikai és természettudományi értesítő (= Mathematischer und naturwissenschaftlicher Anzeiger der Ungarischen Akademie der Wissenschaften)* **60** (1941), 70–82. Im Hungarian, with German summary.

[Lip41b] Stephan Lipka. Über die Abzählung der reellen Wurzeln von algebraischen Gleichungen. *Mathematische Zeitschrift* **47** (1941), 343–351.

[Lip42] Stephan Lipka. Über die Vorzeichenregeln von Budan-Fourier und Descartes. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **52** (1942), 204–217.

[Loo83] R. Loos. Computing in algebraic extensions. *Computer Algebra: Symbolic and Algebraic Computation*, edited by B. Buchberger, G. E. Collins, and R. Loos, pp. 173–187. Springer, 2nd ed., 1983.

[LR81] Jeffrey M. Lane and R. F. Riesenfeld. Bounds on a polynomial. *BIT* **21** (1981), 112–117.

[Mah64] K. Mahler. An inequality for the discriminant of a polynomial. *Michigan Mathematical Journal* **11** (1964), 257–262.

[Mar66] Morris Marden. *Geometry of Polynomials*. AMS, 2nd ed., 1966.

[Mig81] Maurice Mignotte. Some inequalities about univariate polynomials. *Proc. 1981 ACM Symposium on Symbolic and Algebraic Computation (SYMSAC 1981)*, pp. 195–199. ACM, 1981.

[Mig95] Maurice Mignotte. On the distance between the roots of a polynomial. *Applicable Algebra in Engineering, Communication and Computing* **6** (1995), 327–332.

[MRR05] Bernard Mourrain, Fabrice Rouillier, and Marie-Françoise Roy. The Bernstein basis and real root isolation. *Combinatorial and Computational Geometry*, edited by Jacob E. Goodman, János Pach, and Emo Welzl, no. 52 in MSRI Publications, pp. 459–478. Cambridge University Press, 2005.

[MVY02] B. Mourrain, M. N. Vrahatis, and J. C. Yakoubsohn. On the complexity of isolating real roots and computing with certainty the topological degree. *J. Complexity* **18** (2002), 612–640.

[Neu03] Arnold Neumaier. Enclosing clusters of zeros of polynomials. *J. Computational and Applied Mathematics* **156** (2003), 389–401.

[Obr25] N. Obreschkoff. Über die Wurzeln von algebraischen Gleichungen. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **33** (1925), 52–64.

[Obr52a] Nikola Obreshkov. Generalization of Descartes' theorem for imaginary roots. *Doklady Akademii Nauk SSSR (N. S.)* **85** (1952), 489–492. In Russian.

[Obr52b] Nikola Obreshkov. Sur les racines des equations algébriques. *Godishnik na Sofiĭskiĭa Universitet, Fiziko-Matematicheski Fakultet (= Annuaire de l'Université de Sofia, Faculté des Sciences Physiques et Mathématiques), livre 1, partie 2* **47** (1952), 67–83. In Bulgarian, with French summary. Reprinted in: Nikola Obrechkoff, *Opera*, vol. 1, Birkhäuser, 1978, pp. 314–326.

[Obr63] Nikola Obreschkoff. *Verteilung und Berechnung der Nullstellen reeller Polynome*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1963.

[Obr03] Nikola Obrechkoff. *Zeros of Polynomials*. Marin Drinov, Sofia, 2003. Translation from the original edition in Bulgarian, 1963.

[Ost50] A. M. Ostrowski. Note on Vincent's theorem. *Annals of Mathematics (2nd Ser.)* **52** (1950), 702–707. Reprinted in: Alexander Ostrowski, *Collected Mathematical Papers*, vol. 1, pp. 728–733, Birkhäuser, 1983.

[Pan02] Victor Y. Pan. Univariate polynomials: Nearly optimal algorithms for numerical factorization and root-finding. *J. Symbolic Computation* **33** (2002), 701–733.

[PBP02] Hartmut Prautzsch, Wolfgang Boehm, and Marco Paluszny. *Bézier and B-Spline Techniques*. Springer, 2002.

[PS58] G. Pólya and I. J. Schoenberg. Remarks on de la Vallée Poussin means and convex conformal maps of the circle. *Pacific J. Mathematics* **8** (1958), 295–334.

[Ram87] Lyle Ramshaw. Blossoming: A connect-the-dots approach to splines. Research Report 19, DEC Systems Research Center, Palo Alto, CA, 1987.

[Ram89] Lyle Ramshaw. Blossoms are polar forms. Research Report 34, DEC Systems Research Center, Palo Alto, CA, 1989.

[Ric97] Daniel Richardson. How to recognize zero. *J. Symbolic Computation* **24** (1997), 627–645.

[RS02] Q. I. Rahman and G. Schmeisser. *Analytic Theory of Polynomials*. Oxford University Press, 2002.

[RZ04] Fabrice Rouillier and Paul Zimmermann. Efficient isolation of [a] polynomial's real roots. *J. Computational and Applied Mathematics* **162** (2004), 33–50.

[Sch30] Isac Schoenberg. Über variationsvermindernde lineare Transformationen. *Mathematische Zeitschrift* **32** (1930), 321–328.

[Sch34] I. J. Schoenberg. Zur Abzählung der reellen Wurzeln algebraischer Gleichungen. *Mathematische Zeitschrift* **38** (1934), 546–564.

[Sch59] I. J. Schoenberg. On variation diminishing approximation methods. *On Numerical Approximation*, edited by R. E. Langer, pp. 249–274. University of Wisconsin Press, Madison, 1959.

[Sch82] Arnold Schönhage. The fundamental theorem of algebra in terms of computational complexity. Preliminary report, Mathematisches Institut der Universität Tübingen, 1982. Electronic version (2004) at `http://www.cs.uni-bonn.de/~schoe/fdthmrep.ps.gz`.

[Sch03] Arnold Schönhage. Adaptive raising strategies optimizing relative efficiency. *Automata, Languages and Programming, 30th Internat. Colloquium, ICALP 2003*, vol. 2719 of *LNCS*, pp. 611–623. Springer, 2003.

[Sch06] Arnold Schönhage. Polynomial root separation examples. *J. Symbolic Computation* **41** (2006), 1080–1090.

[Sha07a] Vikram Sharma. *Complexity Analysis of Algorithms in Algebraic Computation*. Ph.D. thesis, New York University, NY, USA, 2007.

[Sha07b] Vikram Sharma. Complexity of real root isolation using continued fractions. *Proc. 2007 Internat. Symposium on Symbolic and Algebraic Computation (ISSAC 2007)*, pp. 339–346. ACM, 2007.

[Syl39] James Joseph Sylvester. On rational derivation from equations of coexistence, that is to say, a new and extended theory of elimination, part I. *Philosophical Magazine* **15** (1839), 428–435. Reprinted in: *Collected Mathematical Papers of James Joseph Sylvester*, vol. 1, pp. 40–46, Cambridge University Press, 1904.

[Sze22] G. Szegö. Bemerkungen zu einem Satz von J. H. Grace über die Wurzeln algebraischer Gleichungen. *Mathematische Zeitschrift* **13** (1922), 28–55.

[vdS70] A. van der Sluis. Upperbounds for roots of polynomials. *Numerische Mathematik* **15** (1970), 250–262.

[vdW93] B. L. van der Waerden. *Algebra I*. Springer, 9th ed., 1993. In German; English translations of earlier editions under the same title.

[Vin36] M[onsieur] Vincent. Sur la résolution des équations numériques. *J. de Mathématiques Pures et Appliquées* **1** (1836), 341–372. With an addendum in **3** (1838), 235–243.

[vzG90] Joachim von zur Gathen. Functional decomposition of polynomials: the tame case. *J. Symbolic Computation* **9** (1990), 281–299.

[vzGG97] Joachim von zur Gathen and Jürgen Gerhard. Fast algorithms for Taylor shifts and certain difference equations. *Proc. 1997 Internat. Symposium on Symbolic and Algebraic Computation (ISSAC 1997)*, pp. 40–47. ACM, 1997.

[Wal50] Robert Walker. *Algebraic Curves*. Princeton University Press, 1950.

[Wan04] Xiaoshen Wang. A simple proof of Descartes' rule of signs. *American Mathematical Monthly* **111** (2004), 525–526.

[Wes31] E. C. Westerfield. New bounds for the roots of an algebraic equation. *American Mathematical Monthly* **38** (1931), 30–35.

[Wil63] J. H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice-Hall, 1963. Republished by Dover, 1994.

[Yap00] Chee Keng Yap. *Fundamental Problems of Algorithmic Algebra*. Oxford University Press, 2000.

[Yap04a] Chee K. Yap. On guaranteed accuracy computation. *Geometric Computation*, edited by Falai Chen and Dongming Wang, chap. 12, pp. 322–373. World Scientific, 2004.

[Yap04b] Chee K. Yap. Robust geometric computation. Goodman and O'Rourke [GO04], pp. 927–952.

Note on transliteration

The romanization of Russian and Bulgarian journal names attempts to follow the ALA-LC (1997) tables available at `http://www.loc.gov/catdir/cpso/roman.html`

Note on electronic availability

The following retrodigitization archives on the World Wide Web may be helpful for retrieving some of the older research literature. Some require an institutional subscription.

- Göttinger Digitalisierungs-Zentrum (GDZ) `http://gdz.sub.uni-goettingen.de`
  esp. for *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *J. für die reine und angewandte Mathematik*, *Mathematische Annalen*, *Mathematische Zeitschrift*.
- Gallica `http://gallica.bnf.fr`
  esp. for the *Œuvres* of French mathematicians.
- JSTOR `http://www.jstor.org`
  esp. for *American Mathematical Monthly*, *Annals of Mathematics*, *Proc. American Mathematical Society*.
- The Internet Archive `http://www.archive.org`
  esp. for Collected Works of mathematicians in English language.
- Project Euclid `http://projecteuclid.org`
  esp. for *Michigan Mathematical Journal*, *Pacific J. Mathematics*.
- Swiss Electronic Academic Library Services (SEALS) `http://retro.seals.ch`
  Commentarii Mathematici Helvetici, *L'Enseignement Mathématique*.
- `http://www-mathdoc.ujf-grenoble.fr/JMPA/`
  J. de Mathématiques Pures et Appliquées.
- `ftp://ftp.digital.com/pub/DEC/SRC/research-reports/`
  Research Reports of the DEC Systems Research Center, Palo Alto.

Further sources

The following general reference works have been used during the preparation of this thesis:

- The online literature indices *MathSciNet*, *ZMATH*, *DBLP*, and *ACM Guide*.
- MathWorld `http://mathworld.wolfram.com`
- Wikipedia `http://de.wikipedia.org`, `http://en.wikipedia.org`
- The free English dictionary at `http://dict.leo.org` and the online editions of the *Oxford English Dictionary* and *Merriam-Webster Unabridged*.

All figures have been created by the author, using Maple and `gnuplot` for Figure 2.6 on page 39 and `xfig` for all others.

Index