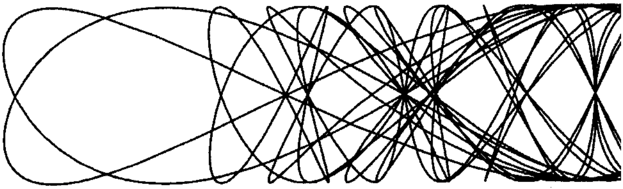
Foundations of Computational Mathematics

This page is intentionally left blank

Foundations of Computational Mathematics



Proceedings of the SMALEFEST 2000
Hong Kong, 13 – 17 2000

editors

Felipe Cucker
City University of Hong Kong

J. Maurice Rojas
Texas A&M University

World Scientific
New Jersey • London • Singapore • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.
P O Box 128, Farrer Road, Singapore 912805
USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
Proceedings of Smalefest 2000

ISBN 981-02-4845-8

Printed in Singapore by World Scientific Printers

FOREWORD

In August 1990 a conference celebrating the 60th birthday of Steve Smale was held at the University of California at Berkeley. The goal of that conference, in the words of its organizers, was "to gather in a single meeting mathematicians working in the many fields to which Smale has made lasting contributions." Thus, the contributed and invited lectures covered a broad scope of subjects including Differential Topology, Dynamical Systems, and Mathematical Economics, among many others. A volume containing most of those lectures was subsequently published by Springer-Verlag (*From Topology to Computation, Proceedings of the Smalefest*, M. W. Hirsch, J. E. Marsden, M. Shub (Eds.), Springer-Verlag, 1993).

Steve moved to City University of Hong Kong in 1995 and on July 15th 2000 he turned 70. It was a pleasure for his friends and colleagues to organize a conference to celebrate this event. On July 13–17, 2000, the second *Smalefest* was held in Hong Kong. Unlike the first one, however, the goal was to focus on the subject Steve had been working on since the early 80's: Theory of Computation. It was a simple matter to gather people who had been influenced by Steve's work on the Theory of Computation, and a glance at this volume shows that other subjects were quite well represented as well.

In the the first *Smalefest* volume, articles were grouped according to subjects and each group of articles was preceded by an article commenting on Steve's work on that subject. In this volume we have included one such article — "The Work of Steve Smale on the Theory of Computation: 1990–1999" — doing so for the period between the two conferences. We thank Singapore University Press and World Scientific which granted us permission to reprint this article. For the remaining articles, we thank the contributors for their valuable work.

Special thanks go to the referees, who helped us select and polish the papers in this volume; to the Liu Bie Ju Centre for Mathematical Sciences for its generous sponsorship; and to Ms. Robin Campbell for her lightning-fast LaTeX formatting.

Steve Smale has positively influenced not only our mathematics but — through his friendship, sincerity, and generosity — our lives. It is with great pleasure that we (the editors and the contributors) dedicate this volume to Steve Smale as a belated gift on his 70th birthday. Happy 70 Steve!

Felipe Cucker
J. Maurice Rojas
December 2001

This page is intentionally left blank

CONTENTS

EXTENDING TRIANGULATIONS AND SEMISTABLE REDUCTION

D. ABRAMOVICH*
Department of Mathematics, Boston University
111 Cummington, Boston, MA 02215, USA
abrmovic@math.bu.edu
http://math.bu.edu/INDIVIDUAL/abrmovic

J. M. ROJAS†
Department of Mathematics, Texas A&M University
College Station, TX 77843-3368, USA
rojas@math.tamu.edu
http://www.math.tamu.edu/~rojas

1 INTRODUCTION

In the past three decades, a strong relationship has been established between convex geometry, represented by convex polyhedra and polyhedral complexes, and algebraic geometry, represented by toric varieties and toroidal embeddings. In this note we exploit this relationship in the following manner. We address a basic problem in algebraic geometry: a certain version of **semistable reduction**.

Semistable reduction, for non-algebraic geometers, can be thought of as a far-reaching extension of Hironaka's famous **resolution of singularities** [8].[a] Technically, Hironaka's result is semistable reduction over a 0-dimensional base (see problem 1.3 below). Semistable reduction over a 1-dimensional base was proved in [13], and was later applied in the classification of algebraic threefolds [14] and the enumerative geometry of curves [4,5] to name but a few examples. Semistable reduction for families of surfaces and threefolds (i.e., part of the case of a 2-dimensional base), in characteristic 0, was proved in [11] but remains an open problem for a higher-dimensional base. This has motivated alternative constructions, e.g, **weak** semistable reduction (see theorem

*PARTIALLY SUPPORTED BY NSF GRANT DMS-9503276 AND AN ALFRED P. SLOAN RESEARCH FELLOWSHIP.

†PARTIALLY SUPPORTED BY AN NSF MATHEMATICAL SCIENCES POSTDOCTORAL FELLOWSHIP AND HONG KONG UGC GRANT #9040402-730.

[a] Roughly, his result is that any algebraic variety over an algebraically closed field of characteristic 0 is birationally equivalent to one without singularities.

1.6 below and the paragraph after the theorem), which could be proved in full generality in characteristic 0 [2], and has also yielded important applications [10,17].

Here, we will translate the local case of semistable reduction, over a base variety of dimension > 1, into a basic problem about polyhedral complexes: extending triangulations. Once we solve the second problem, the first follows. We have taken the opportunity with this note to try to extend some bridges between the terminologies of these two theories.

1.1 Semistable Reduction

We work over the field of complex numbers $\mathbb{C}$. Let $f : X \to B$ be a proper morphism of algebraic varieties, whose generic fiber is reduced and absolutely irreducible. Thus there exists a Zariski dense open set $U \subset B$ such that the fiber $f^{-1}(b)$ over any point in $b \in U$ is a compact complex algebraic variety.

Loosely speaking, semistable reduction for a morphism like f is a meta-problem of "desingularization of morphisms," where the goal is to "change f slightly" so that it becomes "as nice as possible". Of course, we need to specify more precisely what we mean by the clauses in quotation marks.

1.1.1 What do we mean by a morphism being "as nice as possible?"

First of all, X and B should be as nice as possible, namely nonsingular. Moreover, we want f to have a nice, explicit local description, so that the fibers of f have the simplest possible singularities.

Such a morphism will be called **semistable**. Here is the definition:

Definition 1.1 *Let $f : X \to B$ be a flat projective morphism, with connected fibers, of nonsingular varieties. We say that f is* **semistable** *if for each point $x \in X$ with $f(x) = b$ there is a choice of formal coordinates $B_b = \operatorname{Spec} \mathbb{C}[[t_1, \dots, t_m]]$ and $X_x = \operatorname{Spec} \mathbb{C}[[x_1, \dots, x_n]]$, such that f is given by:*

$$t_i = \prod_{j=l_{i-1}+1}^{l_i} x_j,$$

where $0 = l_0 < l_1 \cdots < l_m \leq n$, $n = \dim X$, and $m = \dim B$.

We must state right up front that in this note we will **not** end up with a semistable morphism, but we will get very close. In particular, our results form an additional step in work on semistable reduction [1], continuing the work of [2] for the case of a higher-dimensional base.

1.1.2 What do we mean by "changing f slightly?"

First we define two types of operations necessary for semistable reduction:

Definition 1.2 *An* **alteration** $B_1 \to B$ *is a proper, generically finite, surjective morphism. A* **modification** $Y \to X$ *is a birational proper morphism (equivalently, a birational alteration).*

Given a morphism $X \to B$ as before, and an alteration $B_1 \to B$, we call the component of $X \times_B B_1$ dominating B_1 the **main component** and denote it by $X \tilde{\times}_B B_1$.

We are now ready to state the semistable reduction problem in its ultimate form:

Problem 1.3 *Let $X \to B$ be a flat projective morphism, with connected fibers and B nonsingular. Find an alteration $B_1 \to B$, and a modification $Y \to X \tilde{\times}_B B_1$, such that $Y \to B_1$ is semistable.*

Note that thanks to resolution of singularities [8], we may assume in the characteristic 0 case that X is nonsingular.

1.1.3 Nearly Semistable Morphisms

We will need some terminology in order to state the weaker version of semistable reduction we actually address here. We will follow [13] for the basic definitions.[b]

Definition 1.4

1. *A* **toric variety** *is a normal*[c] *variety X with an open embedded copy T of $(\mathbb{C}^*)^n$, such that the natural $(\mathbb{C}^*)^n$-action on T extends to all of X. We sometimes call the pair (X, T) a* **torus embedding**.

2. *More generally, suppose Y is a normal variety with a smooth open subvariety U_Y satisfying the following condition: locally analytically at every point, (Y, U_Y) is isomorphic to a local analytic neighborhood of some torus embedding (X, T). We then call Y a* **toroidal variety** *and (Y, U_Y) a* **toroidal embedding**.[d]

[b] Also, mimicking standard notation from algebraic topology, $f : (X, A) \longrightarrow (Y, B)$ will be understood to mean that A and B are subvarieties of X and Y respectively; and that f is a morphism from X to Y satisfying $f(A) \subset B$.

[c] Although normality is not assumed in some contexts, all toric varieties will be normal in this paper.

[d] We will sometimes follow [13] and also refer to the inclusion $U_Y \subset Y$ as a toroidal embedding.

3. *A dominant morphism $f : (X,U_X) \to (B,U_B)$ of toroidal embeddings is called a* **toroidal morphism**, *if locally analytically near every point on X it is isomorphic to a torus equivariant morphism of toric varieties.*

Roughly speaking, a toric variety is "monomial:" an affine toric variety is always defined by binomial equations, and any toric variety can always be covered by affine charts in such a way that every overlap isomorphism is a monomial map. Similarly, a toroidal variety is "locally monomial" and a toroidal morphism is a "locally monomial morphism."

If $U_B \subset B$ is a toroidal embedding, then we may write $B \setminus U_B$ as a union of divisors $D_1 \cup \cdots \cup D_k$. More precisely, recall that $B \setminus U_B$ can be decomposed into strata of varying dimensions (see [13] or [7]). In particular, let us define $U_B^{(2)}$ to be the union of U_B and the codimension 0 strata of $B \setminus U_B$. This notation makes sense since we've actually only removed pieces of codimension ≥ 2 from B to construct $U_B^{(2)}$.

We now detail the type of morphisms we will treat:

Definition 1.5 *A proper toroidal morphism $f : (X,U_X) \to (B,U_B)$ is said to be* **nearly semistable** *if the following conditions hold:*

1. *There are no horizontal divisors in X, namely: $U_X = f^{-1}(U_B)$.*

2. *The base B is nonsingular.*

3. *The morphism f is equidimensional.*

4. *All the fibers of f are reduced.*

5. *The restriction of f to $U_B^{(2)}$ is semistable, i.e., "f is semistable in codimension ≤ 1."*

6. *The singularities of variety X are at worst finite quotient singularities.*

One may ask how far a nearly semistable morphism is from a semistable one. The answer is simple: every toroidal semistable morphism is nearly semistable; and a nearly semistable morphism $X \to B$ is semistable if and only if X is nonsingular (see [2]).

1.1.4 The Result

The problem addressed in this paper is a special (local) case of nearly semistable reduction:

Theorem 1.6 *Set $B = \mathbb{A}^n_{\mathbb{C}}$ and let U_B be the natural open subscheme of B whose underlying complex variety is $(\mathbb{C}^*)^n$. Note that the inclusion $U_B \subset B$ is a toroidal embedding, and let $f : X \to B$ be a proper morphism satisfying:*

1. *$U_X := f^{-1}(U_B) \subset X$ is a toroidal embedding, and $f : (X, U_X) \to (B, U_B)$ is a toroidal morphism;*
2. *f is equidimensional, with smooth and absolutely irreducible generic fiber;*
3. *every fiber of f is reduced.*

Then there exists a **finite** *toric morphism $(B_1, U_{B_1}) \to (B, U_B)$ and a toroidal modification $Y \to X \times_B B_1$, such that $Y \to B_1$ is nearly semistable.*

One may ask what right we have to make all these assumptions on the morphism f we start with. In [2] it is shown that given any morphism f, as in Problem 1.3, we can reduce it to a toroidal morphism f as in Theorem 1.6. Such morphisms are called **weakly semistable** in [2].

The methods of [2] are quite different from what we do here. In short, they involve:

1. Making $X \to B$ toroidal. This follows easily from the methods of [1].
2. Making a toroidal $X \to B$ satisfy the conditions in the theorem. Locally this can be done easily using toroidal modifications and finite base changes. To do it globally one uses a covering trick of Kawamata (see [12]).

Moreover, once the local results here are established, we can go back to [2] and, using Kawamata's covering trick, extend it to prove nearly semistable reduction in general.

1.2 Extending Triangulations

We now wear our polyhedral glasses.

For the concepts of a **compact polyhedral complex** Δ and a **conical polyhedral complex** Σ see [13] . An **integral structure** on a compact or conical polyhedral complex is defined in [13] . We will always assume that our complexes come equipped with an integral structure. From here on, we will simply say **polyhedral complex**, when we mean a compact polyhedral complex with integral structure.

Remark 1.7 *A useful example of a polyhedral complex to consider is a finite collection $\mathcal{P}$ of integral polyhedra in $\mathbb{R}^n$. (Recall that a polyhedron in $\mathbb{R}^n$ is* **integral** *iff all its vertices lie in $\mathbb{Z}^n$.) If $\mathcal{P}$ is closed under intersection and* *taking faces, then $\mathcal{P}$ is a polyhedral complex. Note, however, that* **not** *all polyhedral complexes arise this way. This accounts for some of the geometric richness of toroidal varieties.* ◇

Again, in [13], it is shown that for any compact polyhedral complex Δ, one can construct a conical polyhedral complex, which we denote $\Sigma(\Delta)$ — namely the cone over Δ. To reverse the process, define a **slicing function** $h : \Sigma \to \mathbb{R}$ to be a nonnegative continuous function, whose restriction to every cone $\sigma \in \Sigma$ is linear, which vanishes only at the origin $\mathbf{O} \in \Sigma$. Then the **slice** $h^{-1}(1)$ of Σ defines a compact polyhedral complex $\Delta(\Sigma, h)$.

We denote by $\mathrm{Sk}^k(\Delta)$ the k-skeleton of Δ. We will also use $\#S$ for the cardinality of a set S, and $\mathrm{Cone}(V)$ for the set of all nonnegative linear combinations of a set of vectors $V \subset \mathbb{R}^n$.

By a **subdivision** Δ' of Δ (resp. Σ' of Σ) we will mean a finite partial polyhedral decomposition of Δ (resp. Σ), as in [13], with the **completeness** property: $|\Delta'| = |\Delta|$. (Recall that the notation $|\Delta|$ simply means the topological space consisting of the union of all the cells of Δ.) A subdivision Δ' is called a **triangulation** or a **simplicial subdivision** if every cell of Δ' is a simplex.

A **lifting function** (or **order function**) $f : \Delta \to \mathbb{R}$ on a polyhedral complex is a continuous function, convex and piecewise linear on each cell of Δ, respecting the integral structure. (Briefly, the last appelation means that every maximal connected subdomain S on which f is in fact a **linear** function must satisfy the following conditions: (a) S is contained in some cell σ of Δ, (b) the underlying homeomorphism from σ to a polytope $\mathbb{R}^N$ with vertices in $\mathbb{Z}^n$ restricts to a homemorphism of S to a polytope $\tau \subset \sigma$ defined by linear inequalities with rational coefficients.) In the conical case ($f : \Sigma \to \mathbb{R}$) we add the requirement that f be homogeneous: $f(\lambda x) = \lambda f(x)$, for all $\lambda \geq 0$ and all $x \in |\Delta|$ [13].

Remark 1.8 *We follow the convention in* [13]*, where one requires a lifting function to be "convex down" on each cell, namely $f(\lambda x + \mu y) \geq \lambda f(x) + \mu f(y)$. Also, all our lifting functions take rational values on the lattices in the cells. This is in contrast with the polyhedral convention, as in* [18]*, where lifting functions are "convex up" and real values are allowed.* ◇

Given a lifting function $f : \Delta \to \mathbb{R}$, (resp. $f : \Sigma \to \mathbb{R}$) we define the subdivision Δ_f (resp. Σ_f) **induced by** f, to be the coarsest subdivision such that f is linear on each cell.

Remark 1.9 *The subdivision induced by f is clearly determined by the values of f on its vertices $\mathrm{Sk}^0(\Delta_f)$ (resp. its edges $\mathrm{Sk}^1(\Sigma_f)$). In fact, one can construct f from its values on $\mathrm{Sk}^0(\Delta_f)$ (resp. $\mathrm{Sk}^1(\Sigma_f)$) as the minimal func-*

tion which is convex on each cell, having the given values on $\mathrm{Sk}^0(\Delta_f)$ *(resp.* $\mathrm{Sk}^1(\Sigma_f)$*). However, note that* Δ_f *(resp.* Σ_f*) may have strictly more vertices (resp. edges) than* Δ *(resp.* Σ*)! Nevertheless, with some care, we can control this behavior.* ◇

We will prove the following result:

Theorem 1.10 *Let* Δ *be a polyhedral complex and* $\Delta_0 \subset \Delta$ *a subcomplex. Let* Δ_0' *be a triangulation of* Δ_0 *induced by a lifting function. Then there exists a triangulation* Δ' *of* Δ*, also induced by a lifting function, which extends* Δ_0' *and introduces no new vertices. That is,* $\mathrm{Sk}^0(\Delta') = \mathrm{Sk}^0(\Delta) \cup \mathrm{Sk}^0(\Delta_0')$.

Applying this to a slice of a conical polyhedral complex we obtain:

Corollary 1.11 *Let* Σ *be a conical polyhedral complex admitting a slicing function* $h : \Sigma \to \mathbb{R}$*, and let* $\Sigma_0 \subset \Sigma$ *be a subcomplex. Let* Σ_0' *be a triangulation of* Σ_0 *induced by a lifting function. Then there exists a triangulation* Σ' *of* Σ*, also induced by a lifting function, which extends* Σ_0' *and introduces no new edges. That is,* $\mathrm{Sk}^1(\Sigma') = \mathrm{Sk}^1(\Sigma) \cup \mathrm{Sk}^1(\Sigma_0')$.

One may ask, "Do we really need to assume that Δ_0' is induced by a lifting function?" The simplest example showing that this is indeed the case was communicated to us independently by R. Adin and B. Sturmfels:

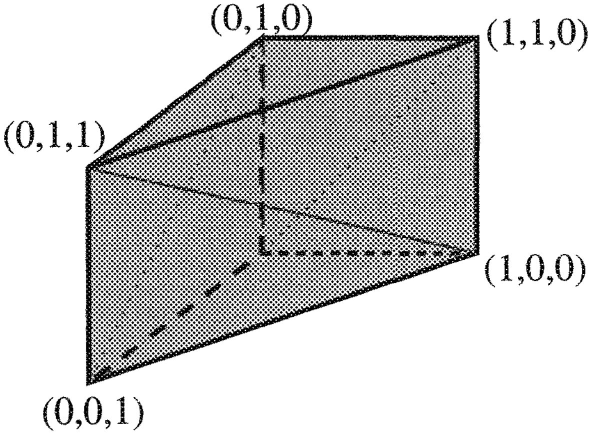


Figure 1. There is no subdivision of the solid prism which preserves the number of vertices and restricts to the subdivision depicted on the boundary.

Let $\Delta \subset \mathbb{R}^3$ be the triangular prism $\delta = \mathrm{Conv}\{v_{0,0}, \dots, v_{1,2}\}$, where:

$$\begin{array}{lll} v_{0,0} = (0,0,0); & v_{0,1} = (1,0,0); & v_{0,2} = (0,0,1) \\ v_{1,0} = (0,1,0); & v_{1,1} = (1,1,0); & v_{1,2} = (0,1,1) \end{array}$$

Let $\Delta_0 = \partial\Delta$ be the boundary of our prism.

Let Δ_0' be the subdivision of Δ_0 obtained by inserting the following new edges:

$$\overline{v_{0,0}v_{1,1}}, \overline{v_{0,1}v_{1,2}}, \overline{v_{0,2}v_{1,0}}$$

(So we've "cut" a new edge into each square 2-face of Δ_0.) It is an easy exercise to see that there is no extension of Δ_0' (to a triangulation of Δ) without new vertices: in particular, any 3-cell of such an extension must have an edge intersecting the midpoint of some edge of Δ_0' — a contradiction. It is also not hard to see that Δ_0' can not be induced by any lifting function [6] .

2 Reduction of Theorem 1.6 to 1.10

Let $f : X \to B$ be as in Theorem 1.6 and $f_\Sigma : \Sigma_X \to \Sigma_B$ the associated morphism of rational conical polyhedral complexes. (Recall that X is toroidal and f is a toroidal morphism, so these associated combinatorial structures indeed exist and are well-defined.) Note that Σ_B is a nonsingular cone (a simplicial cone of index 1): it is simply the nonnegative orthant in $\mathbb{R}^n$, generated by the standard basis vectors $\{\hat{e}_i\}$. Let τ_i be the edges of Σ_B, namely $\tau_i = \mathrm{Cone}(\hat{e}_i)$. We identify the lattice of τ_i with $\mathbb{Z}\hat{e}_i$.

Let $\Sigma_B^1 = \bigcup \tau_i$ be the 1-skeleton of Σ_B and $\Sigma_X^1 = f_\Sigma^{-1}(\Sigma_B^1)$. Also let $\Sigma_{X,i} = f_\Sigma^{-1}(\tau_i)$. For an integer k_i let $N_i(k_i)$ be the integral structure on $\Sigma_{X,i}$ obtained by intersecting the lattices in $\Sigma_{X,i}$ with $f_\Sigma^{-1}(\mathbb{Z}k_i \cdot \hat{e}_i)$.

By [13] , as interpreted in [13] , there exists an integer k_i and a simplicial subdivision $\Sigma_{X,i}'$ of $\Sigma_{X,i}$, **which is induced by a lifting function**, having index 1 with respect to the integral structure $N_i(k_i)$.

Let $B_1 \simeq \mathbb{A}_\mathbb{C}^n$ be complex affine space with coordinates $s_1, \dots, s_n$. The substitution $s_i^{k_i} = t_i$ gives a homomorphism $\mathbb{C}[t_1, \dots, t_n] \to \mathbb{C}[s_1, \dots, s_n]$, giving rise to a finite morphism $B_1 \to B$. Then Σ_{B_1} is the same as Σ_B but taken instead with the lattice $N_{B_1} = \prod \mathbb{Z}k_i\hat{e}_i$. Let $X_1 = X \times_B B_1$. Since the fibers of X are reduced, it follows that X_1 is normal and $X_1 \to B_1$ is again toroidal. Likewise, Σ_{X_1} is just Σ_X with integral structure given by intersecting the lattices in Σ_X with $f_\Sigma^{-1}(N_{B_1})$.

Putting the triangulations $\Sigma_{X,i}'$ of $\Sigma_{X,i}$ together, there exists a triangulation $\Sigma_X^{1'}$ of Σ_X^1 (induced by a lifting function) of index 1 with respect to the integral structure on Σ_{X_1}!

Let us verify that Σ_X admits a slicing function: let $h_B : \Sigma_B \to \mathbb{R}$ be the function defined by $h_b(\sum a_i \hat{e}_i) = \sum a_i$. Then the pullback $h_b \circ f_\Sigma$ is a slicing function on Σ_X.

By Corollary 1.11 of Theorem 1.10, there is an extension of $\Sigma_X^{1'}$ to a triangulation Σ'_X of Σ (induced by a lifting function) without added edges.

Let $Y \to X_1$ be the corresponding toroidal modification and let $f_1 : Y \to B_1$ the resulting morphism.

Note that since all the edges in the triangulation Σ'_X map to the edges τ_i of Σ_{B_1}, we have that f_1 is equidimensional [2]. Since the integral generator of every edge in Σ'_X maps to the generator of the image edge in Σ_{B_1}, and since B is nonsingular, all the fibers of f_1 are reduced [2]. By the construction of [13], f_1 is semistable in codimension 1. Since Δ'_X is simplicial, Y has at worst quotient singularities. Thus f_1 is nearly semistable. ■

Remark 2.1 *The variety Y may be singular, as the following example shows: let $\Sigma_Y \subset \mathbb{R}^4$ be the nonnegative orthant, generated by the standard basis vectors $\hat{e}_1, \ldots \hat{e}_4$. Let $w = (1/2, 1/2, 1/2, 1/2) \in \mathbb{R}^4$ and N_Y the lattice generated by $w, \hat{e}_1, \ldots \hat{e}_4$. Also let Y be the corresponding toric variety — the quotient of $\mathbb{A}^4_{\mathbb{C}}$ by the diagonal $\mathbb{Z}/2$ action given by $p \mapsto -p$ — which happens to be singular. Finally, let $\Sigma_B \subset \mathbb{R}^2$ be the first quadrant, generated by the standard basis vectors $\hat{e}_1, \hat{e}_2$, with the standard lattice $N_B = (\{0\} \cup \mathbb{N})^2$. We have a canonical morphism $\Sigma_Y \to \Sigma_B$ via*

$$(a, b, c, d) \mapsto (a + b, c + d)$$

which maps N_Y into N_B. The resulting morphism $Y \to \mathbb{A}^2_{\mathbb{C}}$ is nearly semistable, but not semistable. ◇

3 Proof of Theorem 1.10

It is a simple fact, made precise in Lemma 3.1 below, that any **generic** lifting function on a polyhedral complex induces a simplicial subdivision. This fact is used frequently in applications of subdivisions to the computation of mixed volumes, polyhedral homotopies, and **toric** (or **sparse**) resultants [16,9,3,15]. The last two constructions give effective recent techniques, sometimes more efficient than Gröbner bases, for solving systems of polynomial equations.

However, it should be emphasized that the lifting functions considered here and in [13] are more general than those in [16,9,3]: the lifting functions in the latter references are completely determined by the values assigned to the vertices of Δ. We will call these more restricted lifting functions **verticial**. The verticial lifting functions are a bit more "economical" in the sense that their corresponding subdivisions never introduce any new vertices.

There is a simple way to resolve this difference by passing to the verticial case from the start. In fact, we will reduce the proof of Theorem 1.10 to finding **any** triangulation (given by a verticial lifting function) in a new, specially constructed, polyhedral complex. The latter problem is then almost trivial to solve.

First recall (see [13], Corollary 1.12) that induced subdivisions are transitive: if Δ' is a subdivision of Δ induced by a lifting function f on Δ, and Δ'' is a subdivision of Δ' induced by a lifting function f' on Δ', then Δ'' is a subdivision of Δ as well. In fact, Δ'' is induced by $f + \varepsilon f'$ for sufficiently small $\varepsilon > 0$.

Thus let $f_0 : \Delta_0 \to \mathbb{R}$ be a lifting function which induces the given subdivision Δ_0' in our theorem. By adding a constant if necessary, we may assume f_0 is positive. Following Remark 1.9, we can take the values of f_0 on $\mathrm{Sk}^0(\Delta_0')$, extend them by zero to the other vertices $\mathrm{Sk}^0(\Delta) \setminus \mathrm{Sk}^0(\Delta_0')$, and take the minimal lifting function $f : \Delta \to \mathbb{R}$ which has these values on the vertices $\mathrm{Sk}^0(\Delta) \cup \mathrm{Sk}^0(\Delta_0')$. Clearly $f|_{\Delta_0} = f_0$. Let Δ_1 be the induced subdivision. Then clearly the restriction of Δ_1 to Δ_0 coincides with Δ_0'. If Δ' is any subdivision of Δ_1 without new vertices, then its restriction to Δ_0 must be Δ_0', since Δ_0' is already simplicial: any subdivision of a simplicial complex without new vertices is trivial. Thus all we need to do to prove Theorem 1.10 is find a **verticial** lifting function on Δ_1 giving a triangulation. In summary, by replacing Δ with Δ_1, we can assume that $\Delta_0 = \Delta_0'$ and then conclude by finding **any** triangulation of Δ_1 (given by a verticial lifting function) — a simpler problem than finding a triangulation of one complex extending some other triangulation.

To complete the proof of Theorem 1.10, recall the following lemma:

Lemma 3.1 *Supppose Δ is a polyhedral complex. Then*

1. *The set L_Δ of all verticial lifting functions on Δ is a finite-dimensional rational vector space.*

2. *The set of all lifting functions which do* **not** *induce simplicial subdivisions is a finite union of proper subspaces of L_Δ.*

Proof: Note that any verticial lifting function on Δ is uniquely determined by its values on $\mathrm{Sk}^0(\Delta)$, which are assumed to be rational, so part (1) follows immediately.

To prove (2), let $C := (c_v \mid v \in \mathrm{Sk}^0(\Delta))$ be a vector of rational constants. Let Δ_C denote the subdivision of Δ induced by the verticial lifting function sending $v \mapsto c_v$ for all $v \in \mathrm{Sk}^0(\Delta)$.

Now suppose that there is a nonsimplicial cell C, with vertex set $V(C)$, in Δ_C. Recall that the coordinates of $d+2$ points lying on a d-flat in $\mathbb{R}^n$ must make a certain $n \times n$ determinant vanish.[e] (In particular, this determinant is a nonconstant multilinear function in the coordinates of the points.) Then, by the definition of a cell in a subdivision induced by lifting, there must be a (nontrivial) linear relation satsified by $(c_v \mid v \in V(C))$. Furthermore, this linear relation depends only on Δ and the set of vertices $V(C)$. Since there are only finitely many possible nonsimplicial cells (since, by definition, our polyhedral complexes have only finitely many vertices), (2) follows immediately. ■

The following is an immediate corollary of our lemma.

Corollary 3.2 *Recall the notation of the proof of Lemma 3.1, and endow* $\mathbb{Q}^{\#\mathrm{Sk}^0(\Delta)}$ *with the standard Euclidean metric* $\|\cdot\|$. *Let* $C \in \mathbb{Q}^{\#\mathrm{Sk}^0(\Delta)}$. *Then for sufficiently small* $\varepsilon > 0$,

1. $\Delta_{C'}$ *is a simplicial subdivision for* **some** $C' \in \mathbb{Q}^{\#\mathrm{Sk}^0(\Delta)}$ *satisfying* $\|C' - C\| < \varepsilon$.

2. *If* Δ_C *is already a simplicial subdivision, then so is* $\Delta_{C'}$, **for all** $C' \in \mathbb{Q}^{\#\mathrm{Sk}^0(\Delta)}$ *satisfying* $\|C' - C\| < \varepsilon$. ■

Remark 3.3 *Put another way, simplicial subdivisions are a dense (via (1)) and open (via (2)) subset of the space of all subdivisions arising from verticial lifting functions. In fact, we really have the stronger statement that the set of all lifting values giving a* **particular** *simplicial subdivision forms an open cell within the space of all subdivisions.*

Note also two "nearby" subdivisions S_1 *and* S_2 *need not have the same extensions, even if* $S_1 = S_2$: *for example, consider the unit square* S *with vector of vertices (ordered clockwise)* (a, b, c, d), *and the subcomplex* E *consisting of the edges* $\{a, b\}$ *and* $\{c, d\}$. *Then* $C = (0, 0, 0, 0)$ *and* $C' = (-1, 1, -1, 1)$ *both generate the same (trivial) subdivision of* E. *However, these two liftings generate different subdivisions of* S, *the first being trivial.* ◇

Returning to the proof of Theorem 1.10, it follows by Corollary 3.2 that there exists a simplicial subdivision of Δ_1 without new vertices, which is what we needed to prove. ■

3.1 Acknowledgements

We would like to thank B. Sturmfels for inciting this collaboration, and R. Adin for a discussion of ideas relevant to this note.

[e] For example, (x_1, y_1), (x_2, y_2), and (x_3, y_3) lie on a line iff Det $\begin{bmatrix} x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \end{bmatrix} = 0$. Note also that this determinant is linear in the "last" coordinates $\{y_1, y_2, y_3\}$.

References

1. Abramovich, Dan and de Jong, Aise Johan, *"Smoothness, Semistability, and Toroidal Geometry,"*, J. Algebraic Geom. 6 (1997), no. 4, pp. 789–801. (Formerly available as Math ArXiV preprint alg-geom/9603018.)
2. Abramovich, Dan and Karu, Kalle, *"Weak Semistable Reduction in Characteristic 0,"* Invent. Math. 139 (2000), no. 2, pp. 241–273.
3. Canny, John F. and Emiris, Ioannis Z., *"A Subdivision-Based Algorithm for the Sparse Mixed Resultant,"* J. ACM **47** (2000), no. 3, pp. 417–451.
4. Caporaso, Lucia and Harris, Joe, *"Parameter Spaces for Curves on Surfaces and Enumeration of Rational Curves,"* Compositio Math. **113** (1998), no. 2, pp. 155–208.
5. ________________________, *"Enumerating Rational Curves: the Rational Fibration Method,"* Compositio Math. **113** (1998), no. 2, pp. 209–236.
6. Fulton, William, *Introduction to Toric Varieties*, Annals of Mathematics Studies, no. 131, Princeton University Press, Princeton, New Jersey, 1993.
7. Goresky, Mark and MacPherson, Robert, *Stratified Morse Theory,* Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag, 1988.
8. Hironaka, Heisuke, *"Resolution of Singularities of an Algebraic Variety over a Field of Characteristic Zero: I and II,"* Ann. of Math. (2) 79 (1964), pp. 109–203, 205–326.
9. Huber, Birkett and Sturmfels, Bernd, *"A Polyhedral Method for Solving Sparse Polynomial Systems,"* Mathematics of Computation, 64, pp. 1541–1555, 1995.
10. Karu, Kalle, *"Minimal Models and Boundedness of Stable Varieties,"* J. Algebraic Geom. 9 (2000), no. 1, pp. 93–109.
11. ________, *"Semistable Reduction in Characteristic Zero for Families of Surfaces and Threefolds,"* Discrete Comput. Geom. 23 (2000), no. 1, pp. 111–120.
12. Kawamata, Y., *"Characterization of Abelian Varieties,"* Comp. Math. 43, 1981 p 253-276.
13. Kempf, G.; Knudsen, F.; Mumford, D.; and Saint-Donat, B., *Toroidal Embeddings I*, Springer, LNM 339, 1973.
14. Kollár, János and Mori, Shigefumi, *"Birational Geometry of Algebraic Varieties,"* with the collaboration of C. H. Clemens and A. Corti, translated from the 1998 Japanese original, Cambridge Tracts in Mathematics, 134, Cambridge University Press, Cambridge, 1998.
15. Rojas, J. Maurice, *"Algebraic Geometry Over Four Rings and the Frontier to Tractability,"* Contemporary Mathematics, vol. 270, Proceedings of a Conference on Hilbert's Tenth Problem and Related Subjects (Uni-

versity of Gent, November 1999, edited by Jan Denef, Leonard Lipschitz, Thanases Pheidas, and Jan Van Geel), pgs. 270–321, AMS Press (2000).

16. Sturmfels, Bernd, *"Sparse Elimination Theory,"* In D. Eisenbud and L. Robbiano, editors, Proc. Computat. Algebraic Geom. and Commut. Algebra 1991, pp. 377–396, Cortona, Italy, 1993, Cambridge Univ. Press.
17. Viehweg, Eckart and Zuo, Kang, *"On the Brody Hyperbolicity of Moduli Spaces for Canonically Polarized Manifolds,"* math arXiv preprint math.AG/0101004 (`http://math.arXiv.org/abs/math.AG/0101004`).
18. Ziegler, Gunter M., *Lectures on Polytopes*, Graduate Texts in Mathematics, Springer Verlag, 1995.

THE WORK OF STEVE SMALE ON THE THEORY OF COMPUTATION: 1990–1999

FELIPE CUCKER
Dept. of Mathematics, City Univ. of Hong Kong, 83, Tat Chee Avenue, HONG KONG
E-mail: macucker@math.cityu.edu.hk

LENORE BLUM
Dept. of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA
E-mail: lblum@cs.cmu.edu

Steve Smale's work on the theory of computation during the 1980's has been carefully reviewed by Mike Shub in [21]. At the end of this period, two key aspects of Smale's work stand out. Firstly, there is the understanding he achieved of certain fundamental numerical algorithms. Quoting Shub's paper,

> He has firmly grounded himself in the mathematics of practical algorithms, Newton's method, and the simplex method of linear programming, inventing the tools and methodology for their analysis.

Secondly, there is the awareness of the need for a theory laying out the foundations for scientific computation. This need is a motivating goal behind [29] where a comparison between theoretical computer science and scientific computation constitutes the opening theme of the article. The following is taken from his paper.

> Algorithms in numerical analysis are primarily a means to solve practical problems, while in computer science, algorithms are studied systematically in their own right. [...] note that algorithms are the main object of study in scientific computation, yet there is not a formal definition of algorithm. I am reminded of how the development of the definition of differentiable manifold was so important in the history of differential topology.

Thus, Smale wants to construct foundations for scientific computation just as those for discrete computation are constructed in theoretical computer science. A good part of Smale's work during the 1980's points in this direction. For instance, in [28], complexity lower bounds are proved by what Smale called "tame machines", and notably in [6], a very general machine model is defined and a theory of computability and complexity is developed.

Two main results of [6] are the existence of universal machines and the NP-completeness of several feasibility problems. The latter reaffirmed the importance of equation solving as a computational problem and suggested two lines of research. On the one hand, there is the P $\neq$ NP conjecture which implies that, even deciding feasibility for systems of equations cannot be done efficiently. On the other hand, while accepting the difficulty of equation solving, one might try to find algorithms which behave well "in general" or with respect to some particular viewpoint. These two problems appear in a list of problems for the next century proposed in [32] in response to a request from V. I. Arnold (on behalf of the International Mathematical Union). Smale selected 18 problems from which the 3rd exactly asks whether P = NP and the 17th ("Solving polynomial equations") reads,

> Can a zero of n complex polynomial equations in n unknowns be found approximately, on the average, in polynomial time with a uniform algorithm?

In the next section we will review some of Smale's work in relation to the first problem. Then we will focus on the second.

1 The P vs. NP problem

1.1 By the end of the 1980's a complexity theory over the reals had been established [6]. This theory of complexity, although primarily devised to describe computations with real numbers, is actually much more general. It describes computations over arbitrary rings and can be "parameterized" by the following five features: 1) a base ring, 2) allowable operations, 3) allowable branching tests, 4) a cost measure, and 5) input size.

For instance, for the ring of integers $\mathbb{Z}$ with operations $\{+, -, \times\}$, branching on $\leq$, logarithmic (or bit) cost and bit input size, we recover, up to polynomial equivalence, the usual Turing machine model and the *classical* complexity theory of computer science. The same is true for the ring $\mathbb{Z}_2 = \{0, 1\}$ with operations $\{+, -, \times, /\}$, branching on $=$, unit cost and vector length as input size. If we replace $\mathbb{Z}_2$ by the complex numbers $\mathbb{C}$, we get a corresponding theory of computation and complexity over $\mathbb{C}$. For a theory over the reals $\mathbb{R}$, we allow branching on $\leq$.

For each combination of these parameters (let's call this a setting) one has a natural $\mathrm{P} \overset{?}{=} \mathrm{NP}$ problem. Here P is the class of (Yes-No decision) problems solvable in polynomial time in the setting; NP is the class of problems whose Yes instances are checkable in polynomial time in the setting.

For some settings, one can prove that P $\neq$ NP simply because NP contains undecidable problems. For a few, a proof that P $\neq$ NP can be obtained together with the standard inclusion NP $\subseteq$ EXP. But for the majority, the question remains wide open.

An issue raised by the variety of settings just mentioned is how the P$\overset{?}{=}$NP question relates between one setting and another. Thus in the foreword of [4] Richard Karp writes,

> It is interesting to speculate as to whether the questions of whether $\mathrm{P}_{\mathbb{R}} = \mathrm{NP}_{\mathbb{R}}$ and whether $\mathrm{P}_{\mathbb{C}} = \mathrm{NP}_{\mathbb{C}}$ are related to each other and to the classical P versus NP question [...].

A good part of the research in complexity theory over arbitrary rings during the last ten years has been dominated by this issue.

1.2 An example of a relation between different settings is found in [26]. Here an NP-complete problem over $\mathbb{C}$ is considered and its intractability is deduced from a hardness assumption concerning computations over $\mathbb{Z}$. This reduces a complexity question over $\mathbb{Z}$ to a complexity question over $\mathbb{C}$, thus considerably enhancing a classical tool in computer science (the idea of reduction between discrete problems) to apply now to a broader class of problems (reductions between problems in different settings).

The Hilbert Nullstellensatz as a decision problem can be stated as follows. Given polynomials $f_1, \ldots, f_s$ with complex coefficients, in n variables, decide whether or not there exists $z \in \mathbb{C}^n$ such that $f_i(z) = 0$ for $i = 1, \ldots, s$. Even restricted to the case when all the f_i have degree 2, the problem is $\mathrm{NP}_{\mathbb{C}}$-complete (cf. [6]).

Now consider the following computational problem over $\mathbb{Z}$. For an integer m denote by $\tau(m)$ the smallest integer ℓ such that there exists a sequence of integers $x_0, x_1, \ldots, x_\ell$ with $x_0 = 1$, $x_\ell = m$ and $x_k = x_i \circ x_j$ for $0 \leq i, j < k \leq \ell$. Here $\circ$ denotes addition, subtraction or multiplication. We say that a sequence of integers a_k is *easy to compute* if there exists $c \in \mathbb{N}$ such that $\tau(a_k) \leq (\log k)^c$ for all $k > 2$. We say that the sequence a_k is *ultimately easy to compute* if there exists a sequence m_k such that the sequence of products $m_k a_k$ is easy to compute. A sequence is *hard* (*ultimately hard*) to compute if it is not easy (resp. ultimately easy) to compute.

The main result in [26] is the following.

Theorem 1 *If the sequence $k!$ is ultimately hard to compute, then the Hilbert Nullstellensatz is intractable, and consequently* $\mathrm{P}_{\mathbb{C}} \neq \mathrm{NP}_{\mathbb{C}}$.

We remark here that the assumption of hardness of the sequence $k!$ in the sense above is related to the hardness of integer factorization classically (cf. [4]).

1.3 Although there is a general belief that P $\neq$ NP for every "reasonable" setting, it would appear wishful thinking to believe the resolution of the P$\overset{?}{=}$NP question over $\mathbb{C}$, say, would resolve the question over $\mathbb{Z}_2$, i.e. classically. Indeed, continuing the remarks quoted above, Karp adds,

> I am inclined to think that the three questions [over $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{Z}_2$] are very different and need to be attacked independently.

However, inroads are being made. For example, we now know that if P $=$ NP over $\mathbb{C}$ then BPP $\supseteq$ NP classically. Here BPP is the class of decision problems solvable in probabilistic polynomial time, the modern version of the concept of "feasible."

It is not difficult to prove that the answer to the P$\overset{?}{=}$NP question is the same for all finite fields. This is the case since, for any two finite fields K_1 and K_2, one can simulate computations over K_1 with a machine over K_2 with only a constant slowdown. In the same vein, one may ask whether the P$\overset{?}{=}$NP question has the same answer for all algebraically closed fields of characteristic zero. The main result of [2], which we now state, gives a positive answer to this question.

Theorem 2 *Let $\bar{\mathbb{Q}}$ be the algebraic closure of $\mathbb{Q}$ and K be any algebraically closed field of characteristic zero. Then* P $=$ NP *over K if and only if* P $=$ NP *over $\bar{\mathbb{Q}}$.*

The "if" part of Theorem 2 was first proved by Michaux [20] using model-theoretic arguments. The "only if" part required the use of properties of heights of algebraic numbers and establishing an Elimination of Constants theorem.

So far, there is no *transfer* result corresponding to Theorem 2 for real closed fields. Michaux [20] proves the "if" part, the "only if" part remains open. Partial results in this direction can be found in [7].

1.4 Another form of comparison arises naturally when considering $\{0,1\}$ (i.e. $\mathbb{Z}_2$) as a subset of a larger ring, most importantly as a subset of $\mathbb{R}$ or $\mathbb{C}$. One may wonder about the computational power of machines over $\mathbb{R}$, for instance, when input instances are assumed to be strings of zeros and ones.

Such questions are first considered by Koiran in [17]. Here he introduced a measure of cost for computations over $\mathbb{R}$ intended to be closer to the bit cost of the Turing model. More precisely, in Koiran's *weak cost* model, additions and comparisons are performed with unit cost but multiplications are penalized so that iterated multiplications becomes expensive (just as in the Turing model where one can compute 2^n with $\log n$ multiplications but with bit cost at least n).

Now, let $\mathcal{C}$ be a complexity class of subsets of $\mathbb{R}^\infty$. Denote by

$$\mathrm{BP}(\mathcal{C}) = \{S \cap \{0,1\}^\infty \mid S \in \mathcal{C}\},$$

i.e., $\mathrm{BP}(\mathcal{C})$ is the complexity class over $\mathbb{Z}_2$ consisting of all those sets (of bit strings) decidable by a machine[a] in class $\mathcal{C}$. The main result of [17] is the following.

Theorem 3 *Let* P_W *and* NP_W *be the classes of subsets of* $\mathbb{R}^\infty$ *decidable in weak deterministic and nondeterministic polynomial time respectively. Then*

$$\mathrm{BP}(\mathrm{P}_\mathrm{W}) = \mathrm{P}/\mathit{poly} \qquad \mathit{and} \qquad \mathrm{BP}(\mathrm{NP}_\mathrm{W}) \supseteq \mathrm{NP}/\mathit{poly}.$$

The "*/poly*" in the statement above introduces non-uniform complexity classes. These classes contain some undecidable sets but also they are known not to contain some decidable sets with high complexity. Also, P/poly is generally assumed to be strictly included in $\mathrm{NP}/\mathit{poly}$ since, by [15], if this is not the case then the polynomial hierarchy collapses at its second level[b]. Thus, Koiran's theorem yields a twofold insight. Firstly, it exactly describes the gain of using real constants and weak polynomial time for binary inputs (this gain is given by the "*/poly*"). Secondly, it gives evidence that $\mathrm{P}_\mathrm{W} \neq \mathrm{NP}_\mathrm{W}$ since the contrary would imply the collapse of the polynomial hierarchy.

In [10] it was shown that indeed $\mathrm{P}_\mathrm{W} \neq \mathrm{NP}_\mathrm{W}$. Actually, a more detailed analysis of the relationships between several complexity classes was made in that paper. Denote by PAR_W and PAR the classes of subsets of $\mathbb{R}^\infty$ decidable in parallel polynomial time for the weak and standard unit cost measures over $\mathbb{R}$ respectively. Also, let DNP_W be the class of subsets of $\mathbb{R}^\infty$ decidable in nondeterministic polynomial time but restricting the guesses to belong to $\{0,1\}$. Most of the results of [10] are sumarized in the following theorem.

Theorem 4 *The relations in the following diagram hold*

$$\begin{tikzcd}
 & & \mathrm{NP}_\mathrm{W} = \mathrm{NP}_\mathbb{R} \arrow[dr] \arrow[dd, leftrightarrow, "/" marking] & & \\
\mathrm{P}_\mathrm{W} \arrow[r] \arrow[dr, "\neq"] & \mathrm{DNP}_\mathrm{W} \arrow[ur, "\neq"] \arrow[dr] & & \mathrm{PAR} \arrow[r] & \mathrm{EXP}_\mathrm{W} \\
 & \mathrm{P}_\mathbb{R} \arrow[r, "/" marking] & \mathrm{PAR}_\mathrm{W} \arrow[ur, "\neq"] & &
\end{tikzcd}$$

[a] As common practice, we are identifying complexity classes with classes of machines.

[b] The polynomial hierarchy is an increasing sequence of complexity classes between P and EXP widely believed to be strict, i.e., no two classes in the hierarchy coincide.

where an arrow $\dashrightarrow$ *means inclusion, an arrow* $\overset{\neq}{\dashrightarrow}$ *means strict inclusion and a crossed arrow* $\not\dashrightarrow$ *means that the inclusion does not hold.*

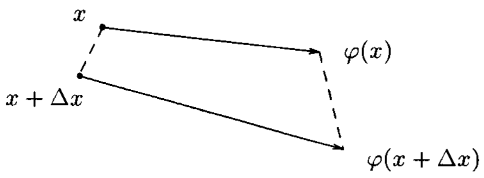
Other results in [10] include a proof of the equality $\mathrm{BP}(\mathrm{PAR}_\mathbb{W}) = \mathrm{PSPACE}/\mathit{poly}$. Here PSPACE denotes the class of subsets of $\{0,1\}^\infty$ decidable in polynomial space.

2 Solving equations

2.1 Algorithms for deciding the feasibility (and finding solutions, if appropriate) of complex systems of polynomial equations, or real systems of polynomial equations and inequalities, have a long history. For the most part, at least concerning feasibility, they rely on algebra or, more precisely, on elimination theory. These algorithms have several virtues (for instance, they show that NP is included in EXP, the class of problems decidable in exponential time). But they are slow and they do not appear to be stable when implemented with floating point numbers. A possible reason for these drawbacks is that these algorithms solve in exponential time *all* input systems. Therefore, they have to deal, on equal footing, with a collection of *ill-posed* systems (e.g. feasible overdetermined systems, systems with multiple roots, etc.).

The tradition in numerical analysis suggests a different strategy for the design and analysis of algorithms. A *condition number* is associated to an input. Several features of the algorithm and of the output corresponding to the input will depend on this number. In particular, ill-posed inputs, those having infinite condition number, may produce exceptional behavior of the algorithm.

Condition numbers were originally introduced to measure the sensitivity of a given input (for a specific computational problem) to perturbations. If φ is the function we are computing, the condition number of x measures how large $||\varphi(x + \Delta x) - \varphi(x)||$ may be compared to $||\Delta x||$ for small perturbations Δx.



Inputs with small condition number are *well-conditioned* and those with large condition number are *ill-conditioned.* This idea of conditioning is already present in a paper of Turing [33] from the early days of computers.

> We should describe the equations (8.2) as an *ill-conditioned* set, or, at any rate, as ill-conditioned when compared with (8.1). It is characteristic of ill-conditioned sets of equations that small percentage errors in the coefficients given may lead to large percentage errors in the solution.

In this paper Turing introduced the term *condition number* for linear equation solving. In this case, the condition number $\kappa(A)$ of a square matrix A is given by $\kappa(A) = \|A\|\|A^{-1}\|$ where the norm denotes the operator norm with respect to the Euclidean norm in both domain and target spaces.

Note that a matrix has infinite condition number if and only if it is not invertible. Thus, the set Σ of all ill-posed problems has measure zero in the space $\mathbb{R}^{n^2}$ of $n \times n$ matrices. The distance of a matrix A to Σ is closely related to $\kappa(A)$.

Theorem 5 (Condition Number Theorem) *For any $n \times n$ real matrix A one has*

$$d_F(A, \Sigma) = \frac{\|A\|}{\kappa(A)}.$$

Here d_F means distance in $\mathbb{R}^{n^2}$ with respect to the Frobenius norm, $\|A\|_F = \sqrt{\sum a_{ij}^2}$.

This theorem was first proved in [13] under the equivalent form $\|A^{-1}\| = d_F(A, \Sigma)^{-1}$.

For non-linear systems, the consideration of a condition number poses some difficulties. Suppose we want to compute some solution to a system of non-linear equations f. Since the system may have several solutions and we do not require any one in particular, $\varphi(f)$ is not well-defined.

A possible resolution is to consider the condition number $\mu(f, \xi)$ for a pair (f, ξ) with $\xi \in \mathbb{R}^m$ a solution of $f = 0$. Then, one may define the condition number $\mu(f)$ in terms of the worst conditioned solution ξ, i.e.,

$$\mu(f) = \max_{\xi | f(\xi)=0} \mu(x, \xi).$$

This is the tack taken by Smale and Mike Shub in the Bézout series of papers [Shub and Smale [22,23,24,27,25]]. Here an impressive development of homotopy methods for systems of complex polynomial equations provides a *non-uniform* solution to Problem 17 in Smale's list. We shall now attempt to summarize some of the main results in the series.

2.2 Let $d = (d_1, \ldots, d_n) \in \mathbb{N}^n$ and $\mathcal{H}_{(d)}$ denote the set of polynomial systems $f = (f_1, \ldots, f_n)$ where f_i is a complex homogeneous polynomial of degree d_i in $x_0, \ldots, x_n$. The problem at hand is: given $f \in \mathcal{H}_{(d)}$, find $\xi \in \mathbb{C}^{n+1}$, $\xi \neq 0$, such that $f(\xi) = 0$. Notice that replacing f by any nonzero multiple λf will not affect the problem. Also, since the f_i are homogeneous, if $f(\xi) = 0$ then $f(\lambda\xi) = 0$ for all $\lambda \in \mathbb{C}$, $\lambda \neq 0$. It is thus natural to consider a "scale invariant" version of the above problem. This is done by replacing the spaces $\mathcal{H}_{(d)}$ and $\mathbb{C}^{n+1}$ by their induced projective spaces. Let $\mathbb{P}(\mathcal{H}_{(d)})$ denote the complex projective space associated to $\mathcal{H}_{(d)}$. The problem now can be restated: given $f \in \mathbb{P}(\mathcal{H}_{(d)})$, find $\xi \in \mathbb{P}(\mathbb{C}^{n+1})$ such that $f(\xi) = 0$. For each sytem f and each root $\xi \in \mathbb{P}(\mathbb{C}^{n+1})$ Shub and Smale define a condition number $\mu(f, \xi)$ extending classical work in numerical analysis going back to Wilkinson [34] and Woźniakowski [35]. They then prove the following equality

$$\mu(f,\xi) = \|f\| \|Df(\xi)|_{T_\xi}^{-1} \Delta(\|\xi\|^{d_i - 1})\|.$$

Here, $\|f\|$ denotes the norm induced by the Weyl Hermitian product on $\mathcal{H}_{(d)}$[c]. Also, $\Delta(\|\xi\|^{d_i-1})$ denotes the diagonal matrix with diagonal $(\|\xi\|^{d_1-1}, \ldots, \|\xi\|^{d_n-1})$ and $T_\xi = \{v \in \mathbb{C}^{n+1} \mid \langle v, \xi\rangle = 0\}$.

From this characterization, a closed form of the set Σ' of ill-posed pairs (f, ξ) follows. More precisely,

$$\Sigma' = \{(f,\xi) \mid f(\xi) = 0 \text{ and } \ker(Df(\xi)|_{T_\xi}) \neq 0\},$$

i.e., Σ' is the set of pairs (f, ξ) such that ξ is a degenerate zero of f.

Shub and Smale then prove a result akin to a Condition Number Theorem. The standard Hermitian product in $\mathbb{C}^{n+1}$ naturally induces a Riemannian metric on $\mathbb{P}(\mathbb{C}^{n+1})$. In a similar way, the Weyl Hermitian product naturally induces a Riemannian metric on $\mathbb{P}(\mathcal{H}_{(d)})$. This allows us to consider distances in $\mathbb{P}(\mathcal{H}_{(d)}) \times \mathbb{P}(\mathbb{C}^{n+1})$. Let

$$V_\xi = \{f \in \mathbb{P}(\mathcal{H}_{(d)}) \mid f(\xi) = 0\}$$

and $d_\xi((f,\xi), \Sigma')$ denote the fiber distance, i.e. the distance in $V_\xi \times \{\xi\}$, between (f, ξ) and Σ'. Define the *normalized condition number* by

$$\mu_{\text{norm}}(f,\xi) = \|f\| \|Df(\xi)|_{T_\xi}^{-1} \Delta(\|\xi\|^{d_i-1}\sqrt{d_i})\|.$$

Then

$$\mu_{\text{norm}}(f,\xi) = \frac{1}{d_\xi((f,\xi),\Sigma')}.$$

[c] Although this is not relevant for what follows, we mention here that the Weyl Hermitian product is, essentially, the only such product in $\mathcal{H}_{(d)}$ invariant under unitary substitution of the variables. This means that if $\sigma : \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$ is unitary (i.e. $\|\sigma(z)\| = \|z\|$ for every $z \in \mathbb{C}^{n+1}$) then, for $f, g \in \mathcal{H}_{(d)}$, $\langle \sigma f, \sigma g\rangle = \langle f, g\rangle$. Here σf denotes the polynomial in $\mathcal{H}_{(d)}$ satisfying $(\sigma f)(z) = f(\sigma(z))$ for all $z \in \mathbb{C}^{n+1}$ and $\langle\ ,\ \rangle$ denotes the Weyl Hermitian product.

To obtain a condition number for f only, Shub and Smale define

$$\mu_{\text{norm}}(f) = \max_{\xi \mid f(\xi)=0} \mu_{\text{norm}}(f, \xi).$$

The condition number of a polynomial system is thus that of its worst conditioned zero. The set Σ of ill-posed systems is then the image of Σ' under the projection

$$\pi : \mathbb{P}(\mathcal{H}_{(d)}) \times \mathbb{P}(\mathbb{C}^{n+1}) \to \mathbb{P}(\mathcal{H}_{(d)}),$$

i.e., the set of all systems having a degenerate zero. Defining

$$\rho(f) = \min_{\xi \mid f(\xi)=0} d_\xi((f, \xi), \Sigma')$$

one gets $\mu_{\text{norm}}(f) = \rho(f)^{-1}$. This is not, strictly speaking, a Condition Number Theorem for $\mu_{\text{norm}}(f)$ since $\rho(f)$ is not the distance from f to Σ, but it is akin to one. A simplified account of this is presented in Chapter 12 of [4].

2.3 Let us consider again the problem: given $f \in \mathbb{P}(\mathcal{H}_{(d)})$, find $\zeta \in \mathbb{C}^{n+1}$, $\zeta \neq 0$, such that $f(\zeta) = 0$.

Consider an initial pair (g, ξ) with $g \in \mathbb{P}(\mathcal{H}_{(d)})$ and $\xi \in \mathbb{P}(\mathbb{C}^{n+1})$ satisfying $g(\xi) = 0$. Define, for $t \in [0, 1]$, the function $f_t = tf + (1-t)g$. In general, as t varies from 0 to 1, a curve C of pairs (f_t, ξ_t) with $f_t(\xi_t) = 0$ is generated in the product space $\mathbb{P}(\mathcal{H}_{(d)}) \times \mathbb{P}(\mathbb{C}^{n+1})$. Since $f_1 = f$, ξ_1 is the point ζ we are looking for. The homotopy algorithm in [22] produces a sequence

$$0 = t_0 < t_1 < \ldots < t_k = 1$$

and a sequence of pairs (f_{t_i}, ξ_i^*) which "closely follows" this curve provided k is large enough. By "closely follows" we mean that ξ_i^* is a good approximation of the root ξ_{t_i} of f_{t_i} (in the sense that Newton's method for f_{t_i} with initial point ξ_i^* converges quadratically to ξ_{t_i} from the first iteration). Notice that this property implies, in particular, that ξ_k^* is a good approximation of the desired root $\zeta = \xi_1$.

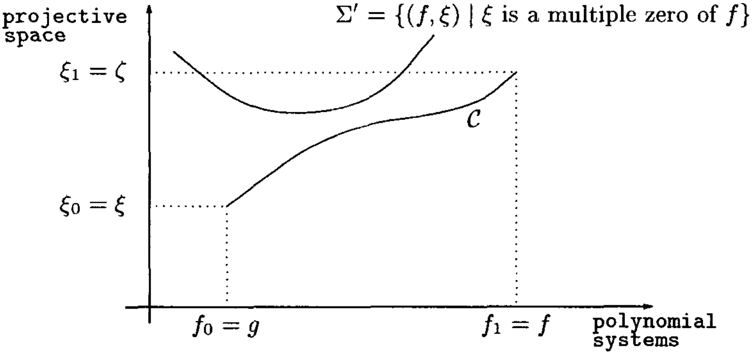
The expression "k is large enough" in the paragraph above has a precise meaning, namely,

$$k \geq cD^2 \mu(C)^2 L_f. \tag{1}$$

Here c is a universal constant and $D = \max\{d_1, \ldots, d_n\}$. In addition L_f is the length of the curve $\{f_t \mid 0 \leq t \leq 1\}$ in $\mathbb{P}(\mathcal{H}_{(d)})$ and $\mu(C)$ is a condition number for C (thus depending on f, g and ξ_0) defined by

$$\mu(C) = \max_{t \in [0,1]} \mu(f_t, \xi_t).$$

In some sense $\mu(C)$ measures (inversely) how close C is to the set Σ' (so again we have the scent of a Condition Number Theorem).



2.4 The algorithmic idea described above is not fully satisfactory. From an algorithmic viewpoint, the lack of a well specified initial pair (g, ξ_0) is certainly a drawback. And from a complexity viewpoint, one may want to eliminate $\mu(C)$ from the bound (1) on the number of iterations and replace it with some probabilistic argument. Note that $\mu(C)$ also depends on the choice of (g, ξ_0).

To date, these drawbacks have not been resolved. The contents of [25], which we now describe, provide the best results so far towards a solution. As we shall see, they trade-off between algorithmics and complexity.

A key ingredient of the homotopy algorithm in the Bézout series is the use of α-theory. This theory, developed by Smale in the 1980's, provides a test to check whether a point z is an approximate zero of a function f. The test is one-sided in the sense that if the answer is Yes, then z is an approximate zero of f, but if the answer is No then it may or may not be so. An approximate zero for which the answer is Yes will be called *certified*. The desired output ξ_k^* of the algorithm is, of course, a certified approximate zero.

Let's now be more precise about the way one eliminates $\mu(C)$ via a probabilistic argument.

The Hermitian structure on $\mathcal{H}_{(d)}$ also induces a probability measure on $\mathbb{P}(\mathcal{H}_{(d)})$. With this measure define a "probability of failure", for an initial pair (g, ξ) and a number of iterations k, by

$$\sigma = \operatorname*{Prob}_{f \in \mathbb{P}(\mathcal{H}_{(d)})} \{\text{the point } \xi_k^* \text{ is not a certified approximate zero of } f\}.$$

Intuitively, σ is related to the probability of having $\mu(C)$ large. This relation leads to the elimination of $\mu(C)$ in the next statement.

Theorem 6 *Fix $d = (d_1, \ldots, d_n)$ and $0 < \sigma < 1$. Then, there exists (g, ξ) with $g(\xi) = 0$ such that if the number k of iterates satisfies*

$$k \geq \frac{cN^3}{\sigma^{1-\varepsilon}} \qquad \textit{where} \qquad \varepsilon = \frac{1}{\log \mathcal{D}}$$

(or $k \geq cN^4/(\sigma^{1-\varepsilon})$ if some $d_i = 1$ or $n \leq 4$) then the probability of failure is at most σ.

Here N is the number of coefficients of f in the dense encoding. Thus, $N = N_1 + \cdots + N_n$ where each f_i has $N_i = \binom{n+d_i}{d_i}$ coefficients. The number N is a reasonable measure of the size of f. Also, $\mathcal{D} = \prod_{i=1}^{n} d_i$, the Bézout number of f.

Each iteration performs at most $\mathcal{O}(N)$ arithmetic operations. Hence (for $n > 4, d_i > 1$), Theorem 6 implies that, for each σ, there exists an algorithm which after $\frac{cN^4}{\sigma^{1-\varepsilon}}$ arithmetic operations either returns an approximate zero of its input f, or returns a failure message. The latter happens with probability at most σ on f. Notice, however, that Theorem 6 is a purely existential result since it gives no indication of the pair (g, ξ). The qualification of *non-uniform* for this algorithm refers to this dependance of (g, ξ) on the dimensions n, d of the problem and on σ.

Since σ is fixed, the bound on the number of arithmetic operations (i.e. the time complexity bound) is polynomial in N. Thus, the algorithm above is polynomial time in the worst-case setting. One can further eliminate the positive probability of failure by trading the worst-case setting for an average-case setting and adding additional non-uniformity.

For $\ell \geq 1$ consider a pair (g_ℓ, ξ_ℓ) as provided by Theorem 6 for $\sigma = 2^{-\ell}$. Now consider the following algorithm (we assume $n > 4, d_i > 1$).

```
      $\ell := 1$
(1)   $\sigma := 2^{-\ell}$
      $k := \frac{cN^3}{\sigma^{1-\varepsilon}}$
      perform the homotopy algorithm with $k$ iterations and
         initial pair $(g_\ell, \xi_\ell)$
      if $\xi_k^*$ is an approximate zero of $f$ then HALT and return $\xi_k^*$
      else $\ell := \ell + 1$ and go to (1)
```

Theorem 7 *The algorithm above performs, on the average (over $f \in \mathbb{P}(\mathcal{H}_{(d)})$), cN^4 arithmetic operations (or cN^5 if some $d_i = 1$ or $n \leq 4$). It yields an approximate zero of its input f in finite time provided $f \notin \Sigma$, i.e. for all its inputs except a set of measure zero.*

The algorithm in Theorem 7 has no failure return. One may consider the infinite running time for inputs $f \in \Sigma$ as a failure, but this event has measure zero in $\mathbb{P}(\mathcal{H}_{(d)})$. On the other hand, the polynomial time bound is only on the average and the non-uniform character has increased since we now need an infinite sequence of pairs (g_ℓ, ξ_ℓ) at hand.

The non-uniform character of these algorithms is certainly unsatisfactory. Shub and Smale conjecture, however, that making them uniform is easy. More concretely, let

$$\bar{g}_i = z_0^{d_i - 1} z_i \qquad \text{for } i = 1, \dots, n$$

and $\bar{\xi} = (1, 0, \dots, 0)$.

Conjecture *The pair $(\bar{g}, \bar{\xi})$, with $\bar{g} = (\bar{g}_1, \dots, \bar{g}_n)$, satisfies the hypothesis of Theorem 6. Consequently, one may take the constant sequence $(\bar{g}, \bar{\xi})$ for the algorithm in Theorem 7.*

2.5 The main focus of [29] was on complexity theory. Smale explained the path that led him (together with L. Blum and M. Shub) to the machine model in [6] and the complexity theory built upon this model. In an appendix, called "Round-off error, approximate solutions, and complexity theory," he proposed to integrate conditioning and round-off with complexity theory. (Other discussions along these lines appear in [5,1,30,3].)

In [31], by invitation of *Acta Numerica*, Smale wrote a paper with his views on complexity theory and numerical analysis. Here one can see advances towards the integration of complexity theory and conditioning, of which the Bézout series is a landmark. The consideration of round-off in a general complexity theory remained an open issue although Smale proposed some suggestions in the last part of the paper which he qualified as "tentative".

By the end of 1996, when the *Acta Numerica* paper was already in the printing process, Smale proposed to Cucker to study the feasibility problem for semi-algebraic systems from a round-off perspective. The goal was to provide an algorithm whose analysis would involve both complexity and round-off and in which conditioning would play a central role.

The turf for this goal was unclear. Traditional round-off analysis has dealt mainly with linear algebra problems. The field of semi-algebraic geometry had not been a major concern of numerical analysts. Thus, condition numbers for our problem had to be defined. But, since the feasibility problem is a decision problem, the definition of condition number using perturbations as in 2.1 is of no use. The condition number would be infinity for systems on the boundary between feasible and infeasible and zero otherwise.

The definition of condition number Cucker and Smale proposed, $\mu^*(\varphi)$, has two different expressions according to whether φ is feasible or not (cf. [11]). In the feasible case, if $x \in \mathbb{R}^n$ is a point satisfying φ, then the condition $\kappa(\varphi, x)$ of the pair (φ, x) is defined in a way which extends the condition number of the Bézout series. But unlike the latter, the condition number of φ is now taken to be the condition of its *best* conditioned solution. That is,

$$\mu^*(\varphi) = \inf_{x \in \mathrm{Sol}(\varphi)} \kappa(\varphi, x).$$

Here $\mathrm{Sol}(\varphi)$ denotes the set of solutions of φ. Thus, for φ to be ill-posed, all its solutions need be. One can see $\mu^*(\varphi)$ as a far-reaching generalization of the condition numbers used by Turing and Wilkinson.

Having defined $\mu^*(\varphi)$, there still remains the problem of what kind of result one may prove for a decision problem since the output of the problem does not allow for "perturbations", it is either Yes or No. The main result of [11] can be stated as follows.

Theorem 8 *Let φ denote a semi-algebraic system as follows*

$$\varphi = \begin{cases} f_i(x) = 0 \;\; i = 1, \dots, m \\ g_j(x) \geq 0 \;\; j = 1, \dots, r \\ h_k(x) > 0 \;\; k = 1, \dots, q \end{cases}$$

where f_i, g_j, h_k are polynomials in $x_1, \dots, x_n$ with real coefficients.

There is a machine M over $\mathbb{R}$ which decides, on input φ, whether there is a point $x \in \mathbb{R}^n$ satisfying φ. The halting time of the machine is bounded by

$$\mu^*(\varphi)^{2n} \mathrm{size}(\varphi)^{cn}$$

with c a universal constant (and thus, in particular, M may not halt on inputs φ such that $\mu^*(\varphi) = \infty$).

Moreover, on each arithmetic operation, a round-off error is permitted with precision polynomialy bounded in $\log \mu^*(\varphi)$ *and in* $\mathrm{size}(\varphi)$.

Here $\mathrm{size}(\varphi)$ *is the size of the dense encoding of φ and is independent of the coefficients of the f's, g's and h's.*

The round-off model considered in Theorem 8 is the absolute error model. An absolute *round-off unit* $\delta < 1$ is considered such that, the result of each arithmetic operation performed with round-off unit δ satisfies

$$x \tilde{\circ} y = (x \circ y) + \rho$$

with $|\rho| \leq \delta$. The *precision* of the computation is $|\log \delta|$. This roughly corresponds to the number of bits necessary to write down a number with round-off unit δ. It is also in agreement with the expression "infinite precision" (for $\delta = 0$) and with the idea that the higher the precision, the more accurate the final result. A similar result (polynomial precision) can also be obtained for the (more usual nowadays) model of relative error.

An additional feature, not present in the statement of Theorem 8, is that if no round-off is allowed and M halts on a feasible input then, in addition, M returns an approximate solution of φ. (By approximate solution we mean a point such that Newton's method, for a specific function associated to φ, will immediately converge quadratically to a solution of φ). The reason this additional bonus is not present in general here is made clear in Remark 24 of [11]. Rougly speaking, in the feasible case, $\mu^*(\varphi)$ is given by the best conditioned solution of φ. But there may be points which are not approximate solutions of φ but which can be erroneously tested as such if the machine precision is low. To avoid the return of such a point as an approximate solution, a more restrictive condition number is required to control the machine precision, one depending on *all* points in $\mathbb{R}^n$ and not just on the solutions of φ. In [8] another condition number, $\varrho^*(\varphi)$, is defined along these lines for which the following is true.

Theorem 9 *If the precision of the machine M in Theorem 8 satisfies a certain bound polynomial in* $\log \varrho^*(\varphi)$ *and* $\text{size}(\varphi)$ *then the following holds: for feasible inputs φ, if M halts it also returns an approximate solution of φ.*

2.6 Two more papers dealing with algorithms for equation solving are [12,9]. In the first one, lower bounds for the kind of algorithms used in the Bézout series are given. Firstly, the class of algorithms is formally defined. A *Newton Continuation Method sequence* (NCM sequence) is a sequence

$$(f_i, \zeta_i) \in \mathbb{P}(\mathcal{H}_{(d)}) \times \mathbb{P}(\mathbb{C}^{n+1}) \qquad 0 \le i \le k$$

satisfying $f_i(\zeta_i) = 0$ and ζ_i is a certified approximate zero of f_{i+1} for $0 \le i < k$.

The main result of [12] is the following.

Theorem 10 *For any NCM sequence (f_i, ζ_i), $0 \le i \le k$,*

(i) $k \ge c_1 \max\left(1, \frac{D-1}{2}\right) d_R(\zeta_0, \zeta_k)$, *and*

(ii) $k \ge c_2 \frac{d_R(\zeta_0, \zeta_k)}{k^{-1} \sum_{i=0}^{k} d_R(\zeta_i, \Sigma_{f_i})}$.

Here c_1 and c_2 are universal constants, d_R is the Riemannian distance in $\mathbb{P}(\mathbb{C}^{n+1})$ *and* $\Sigma_{f_i} = \{z \in \mathbb{P}(\mathbb{C}^{n+1}) \mid \text{rank} Df < n\}$.

Actually, the version of Theorem 10 appearing in [12] is more general in the sense that it holds also for underdetermined systems. That is, the functions f_i above may satisfy

$$f_i : \mathbb{C}^{n+1} \to \mathbb{C}^m$$

with $m \leq n$. To apply Newton's method in the underdetermined case, i.e., when $m < n$, one replaces the inverse $Df(z)|_{T_z}^{-1}$ by the Moore-Penrose inverse $Df(z)|_{T_z}^{\dagger}$.

In the second paper, [9], a totally diferent context is considered, that of diophantine equations. The general problem of deciding whether a polynomial equation has integer roots is known to be undecidable. For the special case of only one variable, algorithms exist which compute all the integer roots. If

$$f = \sum_{i=0}^{d} a_i x^i$$

with $a_i \in \mathbb{Z}$, $a_d \neq 0$, these algorithms return the integer roots of f in time polynomial in d and $L = \max\{\text{height}(a_i)\}$. Here, for an integer a, $\text{height}(a) = \log(1 + |a|)$. This is roughly the number of bits necessary to write down the binary expansion of a. We conclude that these algorithms are polynomial time in the dense encoding of f, i.e. in the encoding of f consisting of the list of all its coefficients,

$$\text{dense}(f) \equiv \{a_0, a_1, \ldots, a_d\}.$$

For polynomials with few non-zero coefficients this way of representing f can be artificially expensive. For such polynomials possibly a more sensible encoding is the sparse encoding in which only non-zero coefficients are specified, together with their indices,

$$\text{sparse}(f) \equiv \{(a_i, i) \mid 0 \leq i \leq d, a_i \neq 0\}.$$

This encoding uses at least L bits to write down the largest coefficient plus $\log d$ bits to write down the exponent d. But it may be exponentially more succint than the dense encoding since the latter specifies all $d+1$ coefficients. In particular, the algorithms mentioned above for computing the integer roots of f may take exponential time in the sparse encoding. The main result of [9] is the following.

Theorem 11 *There is a polynomial time algorithm which, given input $f \in \mathbb{Z}[t]$ in the sparse encoding, decides whether f has an integer root and, if this is the case, outputs the set of integer roots of f.*

3 Additional remarks

Many of the themes outlined in this article, and more, are developed in the book *Complexity and Real Computation* published toward the end of the 1990's [4]. An approach that initially was met with a certain degree of skepticism ("machines are finite so how can you have a theory of computation over the reals?" and "what use is a foundational theory for numerical analysis, anyway?") has led to fertile areas of research producing new insights, new algorithms, new methodologies for their analysis, and certainly a deeper understanding of computation. Connections are being made between the newer theories and the classical theory of computational complexity (e.g. tantalizing transfer results for the fundamental $P \overset{?}{=} NP$ problem), paving the way to employ techniques of mainstream (continuous) mathematics to grapple with hard (discrete) problems of computer science.

Steve Smale, indeed, has been the driving force behind the creation of an overarching community of researchers interested in the foundations of computational mathematics. One need only look at the titles of the workshops[d] offered at the Foundations of Computational Mathematics conference at Oxford University during the summer of 1999 to gleam an appreciation of the scope of this community. Smale's vision, drive, and personality —at once unassuming and compelling— has inspired many young (and some old) researchers to chart new territory. The wonderful work of Koiran [18,19], Kim [16], and Grädel and Meer [14], amongst others, testimony enough to Smale's influence, promises even more to come.

Acknowledgments

Partially supported by SRG grant 7001290. Thanks to the Computer Science Department, Carnegie Mellon University.

References

1. L. Blum, Lectures on a theory of computation and complexity over the reals (or an arbitrary ring), in *Lectures in the Sciences of Complexity II*, ed. E. Jen (Addison-Wesley, pp. 1–47, 1990).

[d] Approximation theory; Complexity theory, real machines and homotopy; Computational dynamics; Relations to computer science; Computational geometry and topology; Multiresolution, computer vision and PDEs; Optimization; Stochastic computation; Symbolic algebra and analysis; Computational number theory; Geometric integration and computation on manifolds; Information-based complexity; Numerical linear algebra.

2. L. Blum, F. Cucker, M. Shub, and S. Smale, Algebraic settings for the problem "P $\neq$ NP". in *The Mathematics of Numerical Analysis*, eds. J. Renegar, M. Shub, and S. Smale (Volume 32 of *Lectures in Applied Mathematics*, American Mathematical Society, pp. 125–144, 1996a).
3. L. Blum, F. Cucker, M. Shub, and S. Smale, Complexity and real computation: a manifest, Int. J. of Bifurcation and Chaos **6**, 3–26 (1996b).
4. L. Blum, F. Cucker, M. Shub, and S. Smale, Complexity and Real Computation, (Springer-Verlag, 1998).
5. L. Blum and M. Shub, Evaluating rational functions: infinite precision is finite cost and tractable on average, SIAM Journal on Computing **15**, 384–398 (1986).
6. L. Blum, M. Shub, and S. Smale, On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines, Bulletin of the Amer. Math. Soc. **21**, 1–46 (1989).
7. O. Chapuis and P. Koiran, (1997), Saturation and stability in the theory of computation over the reals, to appear in *Annals of Pure and Applied Logic.*
8. F. Cucker, Approximate zeros and condition numbers, Journal of Complexity **15**, 214-226 (1999).
9. F. Cucker, P. Koiran, and S. Smale, A polynomial time algorithm for diophantine equations in one variable, Journal of Symbolic Computation **27**, 21–29 (1999).
10. F. Cucker, M. Shub, and S. Smale, Complexity separations in Koiran's weak model, Theoretical Computer Science **133**, 3–14 (1994).
11. F. Cucker and S. Smale, Complexity estimates depending on condition and round-off error, Journal of the ACM **46**, 113–184 (1999).
12. J.-P. Dedieu and S. Smale, Some lower bounds in the complexity of continuation methods, Journal of Complexity **14**, 454–465 (1998).
13. C. Eckart and G. Young, The approximation of one matrix by another of lower rank, Psychometrika **1**, 211–218 (1936).
14. E. Grädel and K. Meer, Descriptive complexity theory over the real numbers, in *The Mathematics of Numerical Analysis*, eds. J. Renegar, M. Shub, and S. Smale (Volume 32 of *Lectures in Applied Mathematics*, American Mathematical Society, pp. 381–404, 1996).
15. R. Karp and R. Lipton, Turing machines that take advice, L'Enseignement Mathématique **28**, 191–209 (1982).
16. M.-H. Kim, An average complexity estimate of a path-following method for a polynomial root, preprint, (1999).
17. P. Koiran, A weak version of the Blum, Shub & Smale model, J. Comput. System Sci. **54**, 177–189 (1997). A preliminary version appeared in *34th*

annual IEEE Symp. on Foundations of Computer Science, pp. 486–495, 1993.

18. P. Koiran, The real dimension problem is $\mathrm{NP}_{\mathbb{R}}$-complete, Journal of Complexity **15**, 227–238 (1999).
19. P. Koiran, Circuits versus trees in algebraic complexity, to appear in Proceedings of *STACS*, (2000).
20. C. Michaux, P $\neq$ NP over the nonstandard reals implies P $\neq$ NP over $\mathbb{R}$, Theoretical Computer Science **133**, 95–104 (1994).
21. M. Shub, On the work of Steve Smale on the theory of computation, in *From Topology to Computation: Proceedings of the Smalefest*, eds. M. Hirsch, J. Marsden, and M. Shub (Springer-Verlag, pp. 281–301, 1993).
22. M. Shub and S. Smale, Complexity of Bézout's theorem I: geometric aspects, Journal of the Amer. Math. Soc. **6**, 459–501 (1993a).
23. M. Shub and S. Smale, Complexity of Bézout's theorem II: volumes and probabilities, in *Computational Algebraic Geometry*, eds. F. Eyssette and A. Galligo, (Volume 109 of *Progress in Mathematics*, Birkhäuser, pp. 267–285, 1993b).
24. M. Shub and S. Smale, Complexity of Bézout's theorem III: condition number and packing, Journal of Complexity **9**, 4–14 (1993c).
25. M. Shub and S. Smale, Complexity of Bézout's theorem V: polynomial time, Theoretical Computer Science **133**, 141–164 (1994).
26. M. Shub and S. Smale, On the intractability of Hilbert's Nullstellensatz and an algebraic version of "P = NP", Duke Math. J. **81**, 47–54 (1995).
27. M. Shub and S. Smale, Complexity of Bézout's theorem IV: probability of success; extensions, SIAM J. of Numer. Anal. **33**, 128–148 (1996).
28. S. Smale, On the topology of algorithms I, Journal of Complexity **3**, 81–89 (1987).
29. S. Smale, Some remarks on the foundations of numerical analysis, SIAM Review **32**, 211–220 (1990).
30. S. Smale, Theory of computation, in *Symp. on the Current State and Prospects of Mathematics*, ed. M. Castellet, (Lect. Notes in Math., Springer-Verlag, pp. 59–69, 1991).
31. S. Smale, Complexity theory and numerical analysis, in *Acta Numerica*, ed. A. Iserles (Cambridge University Press, pp. 523–551, 1997).
32. S. Smale, Mathematical problems for the next century, Mathematical Intelligencer **20**, 7–15 (1998).
33. A. Turing, Rounding-off errors in matrix processes, Quart. J. Mech. Appl. Math. **1**, 287–308 (1948).
34. J. Wilkinson, *Rounding Errors in Algebraic Processes*, (Prentice Hall,

1963).

35. H. Woźniakowski, Numerical stability for solving non-linear equations, Numer. Math **27**, 373–390 (1977).

DATA COMPRESSION AND ADAPTIVE HISTOGRAMS

O. CATONI

Laboratoire de Probabilités et Modèles Aléatoires U.M.R. 7599 du C.N.R.S., Case 188, Université Paris 6, bureau 4 E 19 bât Chevaleret, 4, place Jussieu, F-75 252 Paris Cedex 05

catoni@ccr.jussieu.fr

We describe and study in this paper a two step estimation scheme for density estimation from i.i.d. observations. Each step is based on the Gibbs aggregation rule and computes an adaptive histogram for which a non asymptotic oracle inequality is satisfied. The estimator computed in the first step is used to code the data in the unit interval in a way that is inspired by arithmetic coding. The second estimator analyzes the coded sample and refines the first one. Numerical evidences are provided of the efficiency of the method.

1 Introduction

We will present an approach to adaptive inference that mixes ideas coming from data compression, statistical mechanics and model selection (or rather model aggregation, as we will see).

We will concentrate on the problem of density estimation by histograms. Numerical experiments will consist in estimating densities with respect to the Lebesgue measure on the unit interval. The observation will be an i.i.d. sample $(X_i)_{i=1}^N$ with joint distribution $P^{\otimes N}$, where $P \in \mathcal{M}_+^1([0,1], \mathcal{B})$ is a Borel probability measure on the unit interval (throughout this paper, $\mathcal{M}_+^1(\mathcal{X}, \mathcal{F})$ will be the set of probability measures on the measurable space $(\mathcal{X}, \mathcal{F})$, moreover we will forget to mention the sigma algebra $\mathcal{F}$ when its choice is obvious from the context).

This problem is interesting in itself, but also has some connection with the analysis of "symbolic sequences", such as DNA sequences or typed texts. Let us comment on this to start with, since it was the main motivation for writing this paper. More precisely, the observation $(X_i)_{i=1}^N$ which we were talking about can be constructed from an experiment on words: Let A be a finite alphabet, and assume that we observe an i.i.d. sample $(S_i)_{i=1}^N$ of infinite (or if you prefer very long when compared with the number of samples N) sequences of letters :

$$S_i = (S_i^k)_{k=1}^{\infty}, \qquad S_i^k \in A. \tag{1}$$

These sequences may for example be built by choosing at random a starting point in a DNA sequence or a digitized ASCII text.

We may be interested in coding a new-coming sequence S_{N+1} to achieve the best possible compression rate (i.e. the best possible average number of bits in the coded representation of S_{N+1}, supposed to be of variable length). If we knew the distribution of S_{N+1}, then we could code the prefix of length M of this sequence, namely $\left(S^1_{N+1}, \dots, S^M_{N+1}\right)$, or with shorter notations $S^{1,M}_{N+1}$, using arithmetic coding (sometimes called the Shannon-Fano-Elias algorithm, see [16]). Let us remind that an arithmetic code based on $\mathbb{P}(dS^{1,M}_{N+1})$ (our notation for the distribution of the random variable $S^{1,M}_{N+1} \in A^M$) is a mapping from A^M to the set of finite binary sequences $\{0,1\}^*$, $c : A^M \to \{0,1\}^*$, which is built from $\mathbb{P}(dS^{1,M}_{N+1})$ in such a way that the length of the code for word $s \in A^M$ is approximately equal to $-\log_2\left[\mathbb{P}(S^{1,M}_{N+1} = s)\right]$. The average code length is the expectation $\mathbb{E}\left[\ell\left[c\left(S^{1,M}_{N+1}\right)\right]\right]$ of the length ℓ of coded words. When arithmetic coding is used, the average code length is upper bounded by

$$H\left[\mathbb{P}\left(dS^{1,M}_{N+1}\right)\right] + 2,$$

where H is the Shannon entropy (expressed in bits : the Shannon entropy of a probability distribution P on a finite set E is equal to $-\sum_{s \in E} P(s) \log_2\left[P(s)\right]$, $\log_2$ being the logarithm with base two).

If we do not know this distribution, then we can replace it with an estimate $\widehat{P}$, computed from the observation of $(S_i)_{i=1}^N$. An arithmetic code based on $\widehat{P}$ will have an average length not greater than

$$H\left[\mathbb{P}\left(dS^{1,M}_{N+1}\right)\right] + \frac{1}{\log(2)} \mathcal{K}\left[\mathbb{P}\left(dS^{1,M}_{N+1}\right), \widehat{P}\right] + 2, \qquad (2)$$

where $\mathcal{K}$ is the Kullback Leibler divergence function : if μ and ν are two probability distributions, then

$$\mathcal{K}(\mu, \nu) = \begin{cases} \displaystyle\int \log\left(\frac{\mu}{\nu}\right) d\mu & \text{when } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Thus it is advisable in this situation to minimize what is usually called the *ideal redundancy*

$$\frac{1}{\log(2)} \mathcal{K}\left[\mathbb{P}\left(dS^{1,M}_{N+1}\right), \widehat{P}\right] \qquad (3)$$

of the "ideal code" corresponding to the coding distribution $\widehat{P}$, with respect to $\mathbb{P}\left(dS^{1,M}_{N+1}\right)$. Equation (3) defines a natural risk function to guide the choice of the estimator $\widehat{P}$.

We get back to the setting of density estimation on the unit interval by labeling the alphabet A with integers, taking $A = \{0, 1, \dots, |A| - 1\}$, and by defining X_i as

$$X_i = \sum_{k=1}^{\infty} S_i^k |A|^{-k}. \tag{4}$$

When this identification is made, the Lebesgue measure appears as the identity code in the binary case $A = \{0, 1\}$ and in the general case as a code of approximately constant length. If we replace the Lebesgue measure with an histogram, whose cells are defined from the first digits of X only (in the representation of X defined by (4)), then, apart from some rounding effects due to the discrete nature of coding, we essentially compress the representation of the first digits of X and let the remaining digits uncompressed. Thus, to an histogram model on $[0, 1]$, corresponds some kind of generalized n-gram model on the symbolic sequence S.

Another, maybe more familiar, interpretation of all this, is that we are looking for an estimator of the distribution of S_{N+1} which maximizes the mean log-likelihood of S_{N+1}. In this interpretation, the Lebesgue measure gives the same likelihood to all sequences of length M, and histograms based on the first digits give uniform weights to the following digits, not implied in the cell definition.

Note here that we neither assume that the distribution of the sequence S_{N+1} is time-homogeneous, nor that it satisfies the Markov property.

2 The model

Let A be some finite alphabet. Let $A^* = \bigcup_{i=1}^{\infty} A^i$ be the set of all finite sequences of letters (i.e. of finite words). Let $\varnothing$ be the empty word (of null length). For any word $s \in A^* \cup \{\varnothing\}$, let $\ell(s)$ be the length of s, i.e. its number of letters. (We put $\ell(\varnothing) = 0$.)

Let us say that $\mathcal{D} \subset A^*$ is a *complete prefix dictionary* if no word in $\mathcal{D}$ is the beginning (or as it is usual to say in these matters, the prefix) of another word in $\mathcal{D}$, and if the addition of any new word to $\mathcal{D}$ would break this rule. For any complete prefix dictionary, let $\pi(\mathcal{D})$ be the set $\{(s_1, \dots, s_k) : s \in \mathcal{D}, 1 \leq k \leq \ell(s)\} \cup \{\varnothing\}$ of all the (possibly empty) prefixes of all the words of $\mathcal{D}$. We will call $\pi(\mathcal{D})$ the *prefix set* of $\mathcal{D}$.

Consider some i.i.d. random variables $(X_i)_{i=1}^{N}$ with values in some measurable space $(\mathcal{X}, \mathfrak{B})$. Let $\mu \in \mathcal{M}_+^1(\mathcal{X}, \mathfrak{B})$ be some reference probability measure on $\mathcal{X}$.

Let $\mathcal{S}$ be a set of words that is either the set of all finite words $A^* \cup \{\varnothing\}$ or the prefix set $\pi(\mathcal{D}_\infty)$ of some large complete prefix dictionary $\mathcal{D}_\infty$. Assume that for each word $s \in \mathcal{S}$ a measurable cell $I_s \in \mathfrak{B}$ has been chosen, and that these cells satisfy the three following properties :

- $I_\varnothing = \mathcal{X}$,
- $\{I_{s,a} : a \in A\}$ is a partition of I_s, for any $s \in \pi(\mathcal{D}_\infty) \setminus \mathcal{D}_\infty$,
- $\mu(I_s) > 0$ for any $s \in \mathcal{S}$.

This *cell structure* $\{I_s : s \in \mathcal{S}\}$ induces a mapping of complete prefix dictionaries to measurable partitions of $\mathcal{X}$, as is stated in the following lemma :

Lemma 2.1. *For any complete prefix dictionary $\mathcal{D} \subset \mathcal{S}$, $\{I_s : s \in \mathcal{D}\}$ is a measurable partition of $\mathcal{X}$.*

The proof is left to the reader.

Let $\overline{\mathcal{D}} \subset \mathcal{S}$ be some complete prefix dictionary. Let us consider the family $\mathfrak{D}$ of all complete prefix dictionaries $\mathcal{D}$ such that $\mathcal{D} \subset \pi(\overline{\mathcal{D}})$. Let us consider the family of histogram probability distributions $q_{\mathcal{D},\theta}(x)\mu(dx) \in \mathcal{M}_+^1(\mathcal{X}, \mathfrak{B})$, where $\mathcal{D} \in \mathfrak{D}$ and where $\theta \in \mathcal{M}_+^1(\mathcal{D})$. We define the histogram density $q_{\mathcal{D},\theta}$ with respect to μ to be

$$q_{\mathcal{D},\theta}(x) = \sum_{s \in \mathcal{D}} \frac{\theta(s)}{\mu(I_s)} \mathbb{1}(x \in I_s). \tag{5}$$

Our aim will be to estimate the distribution $P^{\otimes N}$ of $(X_i)_{i=1}^N$ by an histogram $q_{\mathcal{D},\theta}(x)\mu(dx)$, trying to minimize the Kullback divergence

$$\mathcal{K}\big(P, q_{\mathcal{D},\theta}(x)\mu(dx)\big)$$

with respect to $\mathcal{D}$ and θ. Let us recall here that the Kullback divergence between two probability measures ν and ρ is defined to be

$$\mathcal{K}(\nu, \rho) = \begin{cases} \int \log\left(\frac{\nu}{\rho}(x)\right) \nu(dx), & \text{when } \nu \ll \rho, \\ +\infty & \text{otherwise.} \end{cases} \tag{6}$$

To make sure that we approximately minimize this *risk function*, we will produce an *oracle inequality*. This inequality provides an upper bound for the risk that implies no prior knowledge of the properties of the true marginal distribution P, and therefore shows that the estimator is in some sense *adaptive.*

3 Estimation scheme

Our strategy to compute an estimator will be the following :

- First aggregate the histogram models $q_{\mathcal{D},\theta}(x)\mu(dx)$ using the Gibbs estimator [15], to compute a first estimator $\widehat{Q}(dx) = \widehat{q}(x)\mu(dx)$. Base this first estimator on some part $(X_1, \ldots, X_K)$ of the observations only (where $0 \leq K \leq N$).
- Using $\widehat{Q}$ as a coding distribution, map $\mathcal{X}$ to the unit interval, in the spirit of arithmetic coding. More precisely, $A^* \cup \{\varnothing\}$ may be totally ordered by the lexicographic rule, which stipulates that $sas' < sbs''$ and $s \leq ss'$, whenever $s, s', s'' \in A^* \cup \{\varnothing\}$, and $a < b \in A$. We define a map F from $\mathcal{X}$ to $[0, 1]$ in two different ways, depending on whether $\mathcal{S}$ is finite or infinite.

1. When $\mathcal{S} = \pi(\mathcal{D}_\infty)$ is finite, we define a map $\sigma : \mathcal{X} \to \mathcal{D}_\infty$ by the non ambiguous rule $x \in I_{\sigma(x)}$. Then we define $F : \mathcal{X} \to [0, 1]$ by

$$F(x) = \widehat{Q}\big[\sigma(X) < \sigma(x)\big] + \frac{1}{2}\widehat{Q}\big[\sigma(X) = \sigma(x)\big]. \tag{7}$$

Note that the arithmetic code for $\sigma(X)$ based on the estimation $\widehat{Q}$ of P would be obtained by truncating $F(x)$ to the (approximately) shortest binary representation which allows to recover $\sigma(x)$ from $F(x)$ in a non ambiguous way (see [16]).

2. When $\mathcal{S} = A^* \cup \{\varnothing\}$ is infinite, we define $\sigma : \mathcal{X} \to A^{\mathbb{N}}$ by the rule $x \in I_s$ for all the prefixes s of $\sigma(x)$ (i.e. all finite words s of the form $(\sigma(x)_1, \ldots, \sigma(x)_k)$, with k ranging in $\mathbb{N}$). Then we define again $F : \mathcal{X} \to [0, 1]$ by

$$F(x) = \widehat{Q}\big[\sigma(X) < \sigma(x)\big] + \frac{1}{2}\widehat{Q}\big[\sigma(X) = \sigma(x)\big]. \tag{8}$$

Lemma 3.1. *In both cases for any $x \in \mathcal{X}$*

$$\frac{1}{2}\Big\{\widehat{Q}\big[F(X) \leq F(x)\big] + \widehat{Q}\big[F(X) < F(x)\big]\Big\} = F(x). \tag{9}$$

Consequently, when $\mathcal{S} = A^* \cup \{\varnothing\}$, the following statements are equivalent :

1. The image $\widehat{Q} \circ F^{-1}$ of $\widehat{Q}$ by F has no atom.

2. The image measure $\widehat{Q} \circ F^{-1}$ is the Lebesgue measure on $[0, 1]$.

3. For any $x \in \mathcal{X}$

$$\widehat{Q}\big[\sigma(X) = \sigma(x)\big] = 0.$$

Remark 3.2. The proof can be found in the appendix. To summarize things, what we wanted to express by this lemma is that $F(X)$ is almost uniformly distributed on $[0,1]$ under P, up to the estimation error between P and $\widehat{Q}$, and discretization phenomena.

• Code the remaining part of the observations $(X_{K+1}, X_{K+2}, \dots, X_N)$, to create

$$Y_i = F(X_i), \qquad K < i \leq N.$$

The distribution of $(Y_i)_{i=K+1}^N$ is i.i.d., and its marginal distribution should be close to the Lebesgue measure. Therefore, it is natural to consider the Lebesgue measure λ as the reference measure on $[0,1]$. A second estimator $\tilde{Q}$ is built for the distribution of Y_i. It is computed along the same principles as the first estimator. The unit interval is equipped with the cell structure of the dyadic intervals : for any $s \in \{0,1\}^* \cup \{\varnothing\}$ we put

$$J_s = \left(\sum_{k=1}^{\ell(s)} s_k 2^{-k} \right) + [0, 2^{-\ell(s)}[. \tag{10}$$

The second estimator $\tilde{Q}$ is built by aggregating histogram distributions based on the cell structure $\big\{ J_s : s \in \{0,1\}^* \cup \{\varnothing\} \big\}$, using the Lebesgue measure as our reference measure on the unit interval.

• Eventually we compute an estimator $\check{Q}$ in the following way. We put for every I_s, $s \in \mathcal{S}$

$$\check{Q}(I_s) = \tilde{Q}\Big(\widehat{Q}\Big(\bigcup_{\substack{s' < s, \\ s' \in \mathcal{S}}} I_{s'} \Big) + \Big[0, \widehat{Q}(I_s) \Big[\Big). \tag{11}$$

This characterizes a probability measure on the sigma algebra $\mathfrak{J}$ generated by $\{I_s : s \in \mathcal{S}\}$. Indeed, in the case when $\mathcal{S} = \pi(\mathcal{D}_\infty)$ is finite, $\mathfrak{J}$ is finite, and it is elementary to check that $\check{Q}$ is additive on $\mathfrak{J}$. Indeed, if we put

$$\tilde{J}_s = \widehat{Q}\Big(\bigcup_{\substack{s' < s, \\ s' \in \mathcal{S}}} I_{s'} \Big) + \Big[0, \widehat{Q}(I_s) \Big[, \qquad s \in \mathcal{S}, \tag{12}$$

we see immediately that $\tilde{J}_s = \bigcup_{a \in A} \tilde{J}_{sa}$, and that this is a disjoint union. In the infinite case, when $\mathcal{S} = A^{\mathbb{N}}$, $\mathfrak{J}$ is isomorphic to the sigma algebra of $A^{\mathbb{N}}$ generated by the coordinate maps. The same reasoning as in the finite case shows that $\check{Q}$ is additive on any sigma algebra generated by a finite number of coordinates. Therefore, by a well known extension theorem (see e.g. [30]), it can be uniquely extended to a probability measure on $\mathfrak{I}$. Then we complete the definition of $\check{Q}$ by taking

$$\check{Q}(\cdot \,|\, \mathfrak{I}) = \mu(\cdot \,|\, \mathfrak{I}). \tag{13}$$

The reason for using this two step estimation scheme is that we can save some computing and memory resources by using in the first and second estimates two maximal dictionaries (i.e. two values for $\overline{\mathcal{D}}$) which are both smaller than what would have been required in a one step estimation algorithm to achieve the same accuracy. This allows to use a static implementation of the tree structure $\pi(\overline{\mathcal{D}})$, resulting in faster computations. We are planning to use this approach in application fields where it is crucial to optimize memory requirements, due to the necessity to use complex models. Performing successive zooming on high probability regions by successive recoding of the data is a way to "factorize" the choice of a model adapted to the data.

4 Details of estimator definition

We build the first estimator in the following way. This variant of what we proposed in [15] was influenced by Alain Trouvé, who suggested to estimate the conditional distributions $\theta(s_k \,|\, s_1^{k-1})$ instead of estimating directly $\theta(s)$. We are glad to thank him for this contribution. We cut $(X_1, \ldots, X_K)$ into $(X_1, \ldots, X_M)$ and $(X_{M+1}, \ldots, X_K)$. We use $(X_1, \ldots, X_M)$ first to estimate the parameter $\theta \in \mathcal{M}_+^1(\mathcal{D})$. We write

$$\theta(s) = \prod_{k=1}^{\ell(s)} \theta\bigl(s_k \,|\, s_1^{k-1}\bigr). \tag{14}$$

Each conditional distribution $\theta\bigl(s_k \,|\, s_1^{k-1}\bigr)$ takes its range in the set Θ of all possible distributions on the alphabet A, which is nothing but the $|A| - 1$ dimensional simplex. We estimate these conditional distributions using the Laplace estimator. An oracle inequality for this estimator applied to any exchangeable sample is derived in [14]. The estimator itself was introduced by Laplace a long time ago (as mentioned in Rissanen's lecture notes [32]), and is for instance used by J. Rissanen in his papers about the *context* algorithm [33,36]. Let us introduce the counters

$$b(s) = \sum_{i=1}^{M} \mathbb{1}\bigl(X_i \in I_s\bigr), \qquad s \in \pi(\mathcal{D}). \tag{15}$$

Let $S_i \in \mathcal{S}$ be the random variable $S_i = \sigma(X_i)$ (this coincides with the definition given in the introduction in the case of the unit interval). The Laplace estimator is

$$\widehat{\theta}_M\bigl(S^k_{M+1} = s_k \mid S_{M+1} = s_1^{k-1}\bigr) \overset{\text{def}}{=} \frac{b(s_1^k)+1}{b(s_1^{k-1})+|A|}. \tag{16}$$

Theorem 4.1. *With the previous notations and hypotheses,*

$$\begin{aligned}\mathbb{E}_{P^{\otimes(M+1)}(dX_1^{M+1})}\Bigl\{-\log\Bigl[q_{\mathcal{D},\widehat{\theta}_M}\bigl(X_{M+1}\bigr)\Bigr]\Bigr\} \\ \leq \inf_{\theta\in\mathcal{M}^1_+(\mathcal{D})} \mathbb{E}_{P(dX_{M+1})}\Bigl\{-\log\Bigl[q_{\mathcal{D},\theta}\bigl(X_{M+1}\bigr)\Bigr]\Bigr\} + \frac{|\mathcal{D}|-1}{M+1}.\end{aligned} \tag{17}$$

Proof. For any $x \in \mathcal{X}$, let $\mathcal{D}(x) \in \mathcal{D}$ be defined by the relation $x \in I_{\mathcal{D}(x)}$. Let us notice that $q_{\mathcal{D},\theta}(x) = \dfrac{\theta\bigl[\mathcal{D}(x)\bigr]}{\mu\bigl[I_{\mathcal{D}(x)}\bigr]}$, and therefore that it is enough to prove the theorem with $q_{\mathcal{D},\theta}(X_{M+1})$ replaced with $\theta\bigl[\mathcal{D}(X_{M+1})\bigr]$.

Let us change within this proof the definition of the counters $b(s)$ and put

$$b(s) = \sum_{i=1}^{M+1} \mathbb{1}\bigl(X_i \in I_s\bigr). \tag{18}$$

With this modified definition

$$\widehat{\theta}_M\bigl(S^k_{M+1} \mid S^{1,k-1}_{M+1}\bigr) = \frac{b(s_1^k)}{b(s_1^{k-1})+|A|-1}. \tag{19}$$

Let us also introduce the notation $\widehat{\theta}_{M,i}$ to indicate the estimator based on the modified sample $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_M, X_{M+1})$. We can use the fact that $P^{\otimes(M+1)}$ is exchangeable to write

$$-\mathbb{E}\Bigl\{\log\Bigl[\widehat{\theta}_M\bigl[\mathcal{D}(X_{M+1})\bigr]\Bigr]\Bigr\} = -\mathbb{E}\left\{\frac{1}{M+1}\sum_{i=1}^{M+1}\log\Bigl[\widehat{\theta}_{M,i}\bigl[\mathcal{D}(X_i)\bigr]\Bigr]\right\} \tag{20}$$

$$= -\mathbb{E}\left\{\sum_{s\in\mathcal{D}}\frac{b(s)}{M+1}\log\left[\prod_{k=1}^{\ell(s)}\frac{b(s_1^k)}{b(s_1^{k-1})+|A|-1}\right]\right\} \tag{21}$$

... To get this second expression, we have grouped the indices according to the values taken by $\mathcal{D}(X_i)$ *and used the definition of* $\widehat{\theta}_M$. *Moreover we have used the convention* $b(s_1^0) = M + 1$ ...

$$= \mathbb{E}\Bigg\{ - \sum_{s \in \mathcal{D}} \frac{b(s)}{M+1} \log\left(\frac{b(s)}{M+1}\right) \tag{22}$$

$$+ \sum_{s \in \mathcal{D}} \frac{b(s)}{M+1} \sum_{k=1}^{\ell(s)} \log\left(1 + \frac{|A|-1}{b(s_1^{k-1})}\right) \Bigg\} \tag{23}$$

... We have used the identity $\frac{b(s_1^k)}{b(s_1^{k-1})+|A|-1} = \frac{b(s_1^k)}{b(s_1^{k-1})}\left(1 + \frac{|A|-1}{b(s_1^{k-1})}\right)^{-1}$ ...

$$= \mathbb{E}\Bigg\{ \sum_{s \in \mathcal{D}} - \frac{b(s)}{M+1} \log\left(\frac{b(s)}{M+1}\right) \tag{24}$$

$$+ \sum_{s \in \pi(\mathcal{D}) \setminus \mathcal{D}} \frac{b(s)}{M+1} \log\left(1 + \frac{|A|-1}{b(s)}\right) \Bigg\} \tag{25}$$

... We have exchanged the order of summations in the second term ...

$$\leq \mathbb{E}\left[\sum_{s \in \mathcal{D}} - \frac{b(s)}{M+1} \log\left(\frac{b(s)}{M+1}\right) \right] \tag{26}$$

$$+ \frac{|\pi(\mathcal{D}) \setminus \mathcal{D}|(|A|-1)}{M+1} \tag{27}$$

... and used the inequality $\log(1+r) \leq r$ ...

$$= \mathbb{E}\left\{ \inf_{\theta \in \mathcal{M}_+^1(\mathcal{D})} \left[- \sum_{s \in \mathcal{D}} \frac{b(s)}{M+1} \log[\theta(s)] \right] \right\} + \frac{|\mathcal{D}|-1}{M+1} \tag{28}$$

... *We have rewritten both terms. The transformation of the first one is justified by the fact that the Kullback divergence function is non-negative and zero on the diagonal, the second term has been rewritten from the identity* $|\pi(\mathcal{D}) \setminus \mathcal{D}| = \frac{|\mathcal{D}|-1}{|A|-1}$ ...

$$= \mathbb{E}\left\{ \inf_{\theta \in \mathcal{M}^1_+(\mathcal{D})} \left[-\frac{1}{M+1} \sum_{i=1}^{M+1} \log\left\{ \theta\left[\mathcal{D}(X_i)\right] \right\} \right] \right\} \tag{29}$$

$$+ \frac{|\mathcal{D}|-1}{M+1} \tag{30}$$

... *where we have just switched back to a summation with respect to time indices* ...

$$\leq \inf_{\theta \in \mathcal{M}^1_+(\mathcal{X})} \mathbb{E}\left\{ -\log\left[\theta\left[\mathcal{D}(X_{M+1})\right] \right] \right\} + \frac{|\mathcal{D}|-1}{M+1} \tag{31}$$

... *where we have pulled back* inf *from the expectation and used the fact that the sample is exchangeable once more.* □

Once we have built the estimators $q_{\mathcal{D},\widehat{\theta}_M}$, we can aggregate them, using the Gibbs aggregation rule based on the independent subsample $(X_{M+1}, \ldots, X_K)$. Let $\rho \in \mathcal{M}^1_+(\mathfrak{D})$ be the probability distribution

$$\rho(\mathcal{D}) = \alpha^{(|\mathcal{D}|-1)/(|A|-1)}(1-\alpha)^{|\mathcal{D} \setminus \overline{\mathcal{D}}|}. \tag{32}$$

(We remind the reader that $\mathfrak{D}$ is the set of all prefix dictionaries $\mathcal{D} \subset \pi(\overline{\mathcal{D}})$.) It is obtained by considering a branching process starting at the empty word, where generation d is made of a set of words of length d. The transition between generation d and $d+1$ is the following : each word $s \in \pi(\overline{\mathcal{D}}) - \overline{\mathcal{D}}$ of the population independently gives birth to $\{sa : a \in A\}$ with probability α, and dies without heirs with probability $(1-\alpha)$, whereas the words of $\overline{\mathcal{D}}$ die with probability 1.

We define the Gibbs estimator

$$\widehat{q}(x) = \frac{\sum_{\mathcal{D} \in \mathfrak{D}} \rho(\mathcal{D}) \left(\prod_{i=M+1}^{K} q_{\mathcal{D},\widehat{\theta}_M}(X_i) \right)^{\beta} q_{\mathcal{D},\widehat{\theta}_M}(x)}{\sum_{\mathcal{D} \in \mathfrak{D}} \rho(\mathcal{D}) \left(\prod_{i=M+1}^{K} q_{\mathcal{D},\widehat{\theta}_M}(X_i) \right)^{\beta}}, \tag{33}$$

where β is a parameter in the range $0 < \beta < 1/2$.

The following theorem holds :

Theorem 4.2. *Let*

$$\chi = \max_{\mathcal{D},\mathcal{D}' \in \mathfrak{D}} \max_{x \in \mathcal{X}} \log \left(\frac{q_{\mathcal{D},\widehat{\theta}_M}(x)}{q_{\mathcal{D}',\widehat{\theta}_M}(x)} \right), \tag{34}$$

and let

$$\beta = \sup_{\alpha \in]0,1]} \frac{1}{e^{\alpha\chi} - 1} \left(\sqrt{1 + \alpha(2-\alpha)(e^{\alpha\chi} - 1)} - 1 \right). \tag{35}$$

With the previous notations and hypotheses

$$\mathbb{E}\Big\{ - \log\big[\widehat{q}(X_{K+1})\big] \Big\} \leq \inf_{\mathcal{D} \in \mathfrak{D}} \inf_{\theta \in \mathcal{M}^1_+(\mathcal{D})} \mathbb{E}\Big\{ - \log\big[q_{\mathcal{D},\theta}(X_{K+1})\big] \Big\} + \frac{|\mathcal{D}| - 1}{M+1} - \frac{\log\big(\rho(\mathcal{D})\big)}{(K-M+1)} \mathbb{E}(\beta^{-1}). \tag{36}$$

Remark 4.3. Note that $\chi \leq \log(M+1)$, and that consequently in any case, taking $\alpha = \frac{\log(\chi)}{\chi}$ we obtain that

$$\mathbb{E}(\beta^{-1}) \leq \frac{\log(M+1) - 1}{\sqrt{1 + \frac{\log(\log(M+1))}{\log(M+1)} \left(2 - \frac{\log(\log(M+1))}{\log(M+1)}\right) \Big(\log(M+1) - 1\Big)} - 1} \underset{M \to \infty}{\sim} \frac{\log(M+1)}{\sqrt{2 \log(\log(M+1))}}. \tag{37}$$

Remark 4.4. Note also that the theorem implies that

$$\mathbb{E}\big[\mathcal{K}(P, \widehat{q}\mu)\big] \leq \inf_{\mathcal{D} \in \mathfrak{D}} \inf_{\theta \in \mathcal{M}^1_+(\mathcal{D})} \mathcal{K}(P, q_{\mathcal{D},\theta}\mu) + \frac{|\mathcal{D}| - 1}{M+1} - \frac{\log\big[\rho(\mathcal{D})\big]}{K - M + 1} \mathbb{E}\left(\beta^{-1}\right). \tag{38}$$

Remark 4.5. For a large dictionary $\mathcal{D}$, the maximum of $\rho(\mathcal{D})$ is reached when the branching process is critical, that is when $\alpha = \frac{1}{|A|}$. Therefore it is advisable to set α to this particular value.

For a proof of theorem 4.2, we refer the reader to [15].

Hopefully, there are fast algorithms to compute $\widehat{q}(x)$. We propose one which is inspired by the data compression algorithm described in [37], although we have slightly modified the induction step, because it is more efficient here to work with conditional probabilities. Define the counters

$$c(s) = \sum_{i=M+1}^{K} \mathbb{1}\big(X_i \in I_s\big), \qquad s \in \pi(\overline{\mathcal{D}}). \tag{39}$$

Attach to each node of the tree $\pi(\overline{\mathcal{D}})$ a weight $\Upsilon_s(x)$ defined in a recursive way by

$$\Upsilon_s(x) = \begin{cases} (1-\alpha) + \alpha \prod_{a \in A} \left\{ \Upsilon_{sa}(x) \left[\dfrac{\widehat{\theta}_M(a \mid s)}{\mu(I_{sa} \mid I_s)} \right]^{\beta c(sa) + \mathbb{1}(x \in I_{sa})} \right\} \\ \qquad \text{when } s \in \pi(\overline{\mathcal{D}}) \setminus \overline{\mathcal{D}}, \\ 1 \quad \text{when } s \in \overline{\mathcal{D}}. \end{cases} \tag{40}$$

The Gibbs estimator $\widehat{q}$ is given by the formula

$$\widehat{q}(x) = \frac{\Upsilon_\varnothing(x)}{\int_{\mathcal{X}} \Upsilon_\varnothing(y)\mu(dy)}. \tag{41}$$

It is constant on each cell I_s, $s \in \overline{\mathcal{D}}$ of the partition defined by the maximal dictionary $\overline{\mathcal{D}}$. Therefore there are $|\overline{\mathcal{D}}|$ numbers to compute. The computation of $\widehat{q}(x)$ for $x \in I_s$ and $s \in \overline{\mathcal{D}}$ requires an update of $\Upsilon(s_1^k)(x)$, for $k = 1, \dots, \ell(s)$, starting from a tree of weights $\Upsilon_s(\varnothing)$, where "$\mathbb{1}(x \in I_s)$ has been set to 0 everywhere", namely $\Upsilon_s(\varnothing)$ is defined by

$$\Upsilon_s(\varnothing) = \begin{cases} (1-\alpha) + \alpha \prod_{a \in A} \left\{ \Upsilon_{sa}(\varnothing) \left[\dfrac{\widehat{\theta}_M(a \mid s)}{\mu(I_{sa} \mid I_s)} \right]^{\beta c(sa)} \right\} \\ \qquad \text{when } s \in \pi(\overline{\mathcal{D}}) \setminus \overline{\mathcal{D}}, \\ 1 \quad \text{when } s \in \overline{\mathcal{D}}. \end{cases} \tag{42}$$

Therefore the number of operations involved to compute $\widehat{q}$ is of order at most $|\overline{\mathcal{D}}|\ell(\overline{\mathcal{D}})$, where $\ell(\overline{\mathcal{D}}) = \max_{s \in \overline{\mathcal{D}}} \ell(s)$. Note also that

$$\int_{\mathcal{X}} \Upsilon_\varnothing(y)\mu(dy) = \Upsilon_\varnothing(\varnothing). \tag{43}$$

The second estimator $\tilde{Q} \in \mathcal{M}_+^1([0,1])$ is built from $(Y_{K+1}, \dots, Y_N)$ and the cell structure $\{J_s : s \in \{0,1\}^* \cup \{\varnothing\}\}$ using the Lebesgue measure as the reference measure, in the same way as $\widehat{Q}$ is built from $(X_1, \dots, X_K)$, $\{I_s : s \in \mathcal{S}\}$ and μ. The last appendix contains the source code of the function `WeightMix` computing equation (40), with some comments about numerical stability issues.

5 Simulations

All the simulations presented here are made in the case when $\mathcal{X} = [0,1]$ and $\mu = \lambda$ (the Lebesgue measure). To perform them we wrote some piece of software which can be downloaded and tested from the web address

`http://www.proba.jussieu.fr/users/catoni/homepage/homepage-en.html`. We work with simulated data with a known distribution. This allows to compute the Kullback distances $\mathcal{K}(P,\widehat{Q})$ and $\mathcal{K}(P,\check{Q})$. Also, we always take $M = K/2$ in the computation of the Gibbs estimators.

We will test first the one step estimation scheme (obtained by taking $M = N$).

Let us start with an example where $P = q_{\mathfrak{D},\theta}\lambda$.

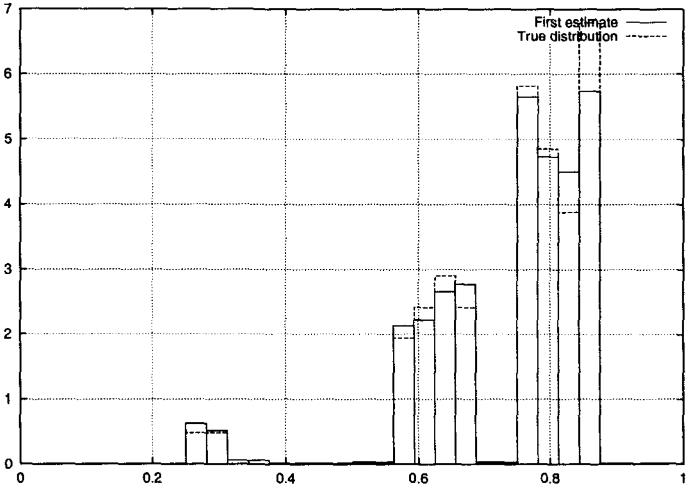


Figure 1. $N = 1000$, $\overline{\mathfrak{D}} = \{0,1\}^5$, $\mathcal{K}(P,\widehat{Q}) = 0.021$, $\beta = 0.159$

Here the true distribution P is an histogram based on $\overline{\mathfrak{D}}$. The example shown in figure 2 shows that we can take $\overline{\mathfrak{D}}$ to be much larger (namely $\overline{\mathfrak{D}} = \{0,1\}^{10}$) without falling into over-fitting problems.

To get more accurate results, we can run a batch session with 1000 trials with both sets of parameters. We obtained the following outputs :

N	$\overline{\mathfrak{D}}$	mean of $\mathcal{K}(P,\widehat{Q})$	std dev
1000	$\{0,1\}^5$	0.027	0.005
1000	$\{0,1\}^{10}$	0.029	0.005

We can then increase the sample size to check that the risk is proportional to its inverse :

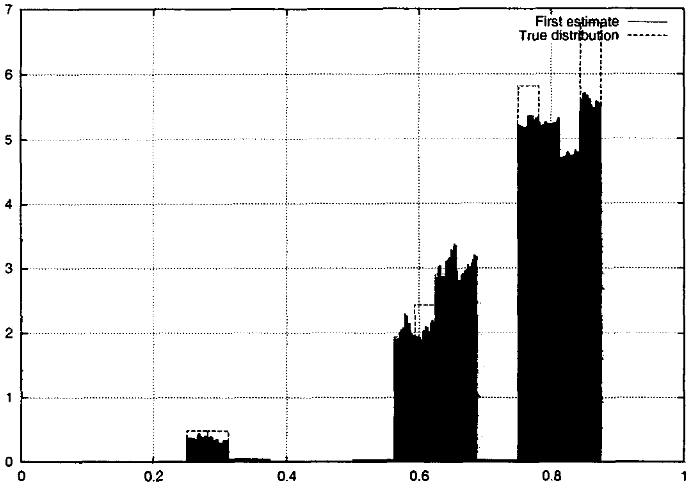


Figure 2. $N = 1000$, $\mathcal{D} = \{0,1\}^{10}$, $\mathcal{K}(P,\widehat{Q}) = 0.026$, $\beta = 0.155$

N	mean of $\mathcal{K}(P,\widehat{Q})$	std dev
100	0.24	0.02
1000	0.029	0.005
10 000	0.002 4	0.000 46
100 000	0.000 23	5×10^{-5}

Let us notice that the proportionality is remarkably well maintained through a large scale of sample sizes. Considering that the probability to be estimated depends on ten parameters (counting for one parameter the definition of the support, which is clearly an underestimation), we see that the risk observed in the simulations is less than $3d/N$, where d is the dimension of the problem. This is better than what is proved by the theory, which gives an upper bound larger than $2d\bigl[1 + 2\log(2)\mathbb{E}(\beta^{-1})\bigr]/N$ with $\beta \leq 1/2$ (note that $2\bigl[1 + 4\log(2)\bigr] \geq 7.5$).

Let us present now an example where the true distribution is not contained in the models used by the first estimator. In this second example, the true distribution is a mixture of four Gaussian distributions. Figures 3, 4, 5 and 6 show the evolution of the quality of estimation with the sample size. As expected, the decrease is slower than what is observed when the true distribution belongs to one of the models used by the estimator, because of the influence of the bias term. Of course, the speed of approximation of such a smooth density function by histograms is not optimal. It would certainly be more appropriate to work with smooth density models in this case : anyhow, the design of a fast algorithm to aggregate more general density models is beyond the scope of this paper, and is, to our knowledge, an open question deserving further investigations.

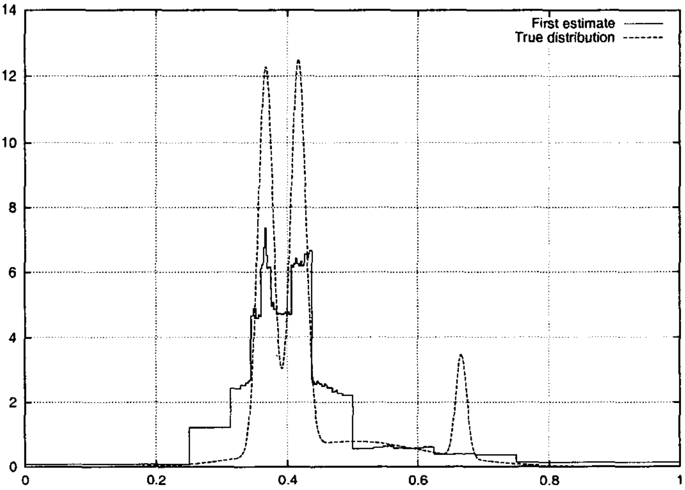


Figure 3. $N = 100$, $\overline{\mathcal{D}} = \{0,1\}^9$, $\mathcal{K}(P,\widehat{Q}) = 0.321$, $\beta = 0.179$

To show the benefit of the two step estimation scheme, we end with experiments on binary sequences extracted from an ASCII text (namely a short story by Oscar Wilde). The statistical experiment is built in the following way. We cut the text into chunks of four bytes (32 bits), which we represent as real numbers in the unit interval. Let $(Z_1, \dots, Z_L)$ be these numbers. Then we draw $(n_1, \dots, n_N)$, an i.i.d. sample from the uniform distribution on the integers $\{1, 2, \dots, L\}$ and we let $X_i = Z_{n_i}$. To estimate adaptive histograms, we choose a maximum dictionary of the form $\overline{\mathcal{D}} = \{0,1\}^d$, and we let the true distribution be the empirical distribution of $\left(\mathcal{D}_\infty(Z_k)\right)_{k=1}^{L}$, where $\mathcal{D}_\infty = \{0,1\}^D$, with $D \gg d$.

We see, as expected, on figure 7 that the estimated densities are very irregular. We did not plot the "true distribution", because it would have been meaningless: indeed we can expect the true distribution to be virtually singular with respect to the Lebesgue measure. Anyhow it is still meaningful to approximate it by histograms, using our method, since theorem 4.2 does

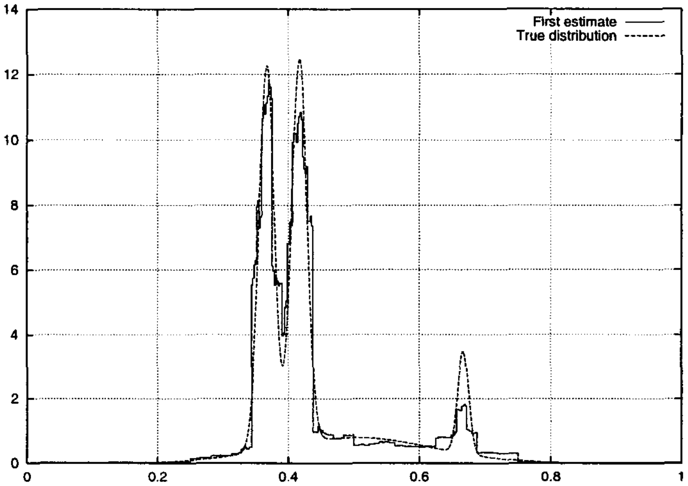


Figure 4. $N = 1000$, $\overline{\mathcal{D}} = \{0, 1\}^9$, $\mathcal{K}(P, \widehat{Q}) = 0.0533$, $\beta = 0.148$

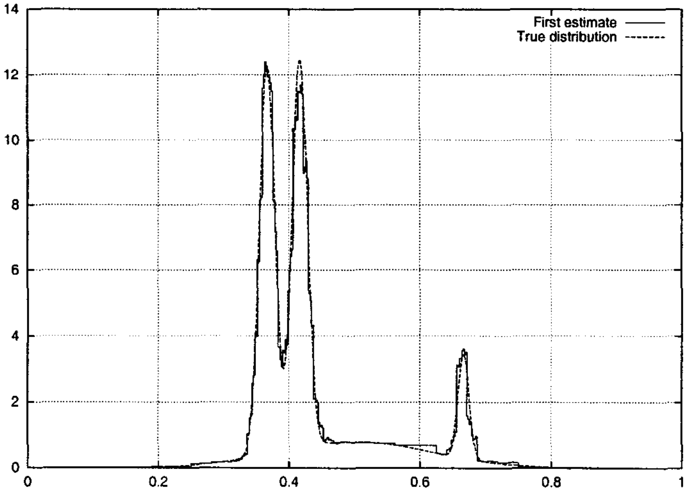


Figure 5. $N = 10000$, $\overline{\mathcal{D}} = \{0, 1\}^9$, $\mathcal{K}(P, \widehat{Q}) = 0.0111$, $\beta = 0.136$

not require that $P \ll \mu$.

L	N	K	$\overline{\mathcal{D}}$	$\mathcal{D}_\infty$
7072	1000	100	$\{0,1\}^8$	$\{0,1\}^{18}$

mean $\mathcal{K}(P,\widehat{Q})\pm$ std dev	mean $\mathcal{K}(P,\check{Q})\pm$ std dev
5.12 ± 0.08	3.70 ± 0.07

Table 1.

Table 1 gathers the results of a batch session illustrating the use of a two step estimation scheme. It contains 50 trials: As can be seen, the parameters are the same as in figure 7. Profile data reproduced in appendix indicate the time spent in the different functions of the program (when it is run on a Pentium II processor at 266 MHz with 144 Mb of RAM). They show that (discarding ancillary functions added to compute the risk) most computing time is spent in the critical function `WeightMix`, where densities are aggregated according to equation (40), and that the time spent in this function jumps from 0.38 to 1.69 seconds when one goes from a two step scheme to a one step scheme of comparable accuracy.

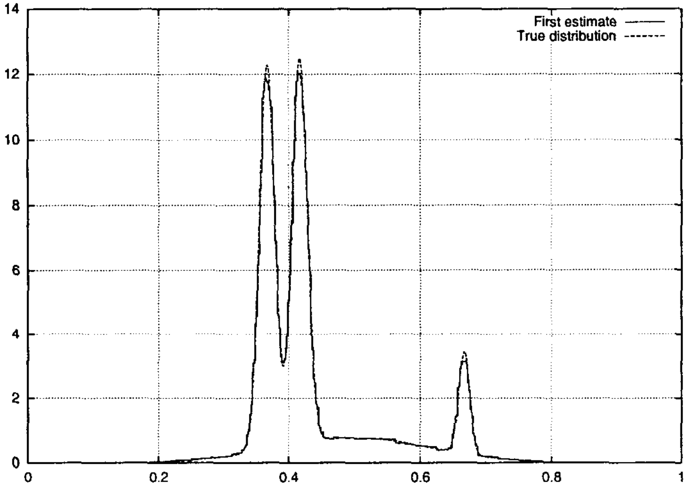


Figure 6. $N = 100000$, $\overline{\mathcal{D}} = \{0,1\}^9$, $\mathcal{K}(P,\widehat{Q}) = 0.00232$, $\beta = 0.121$

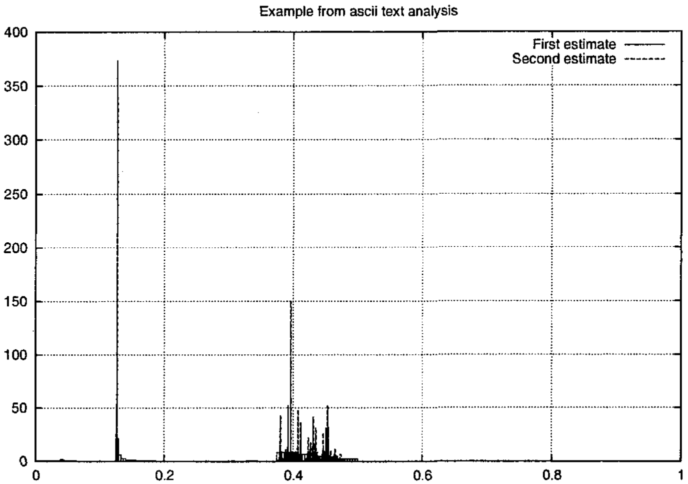


Figure 7. $N = 1000$, $K = 100$, $\overline{\mathcal{D}} = \{0,1\}^8$, $\mathcal{D}_\infty = \{0,1\}^{18}$, $\mathcal{K}(P, \widehat{Q}) = 5.0088$, $\mathcal{K}(P, \check{Q}) = 3.6160$, $\beta = 0.162$ (second estimate)

Conclusion

We have shown in this paper that the Gibbs estimator for adaptive histogram aggregation is not only a theoretical object with a nice oracle inequality, but also an efficient practical way of performing density estimation. Nice features are that no fine hand tuning of parameters is needed to make the algorithm work (considering that it is reasonable to take $M = K/2$, $\alpha = 1/|A|$ and $K = 0.1N$), and that overfitting is avoided even when the finer cell size is very small with respect to the sample size. The results shown here are obtained with the algorithm corresponding to theorem 4.2. The implementation is strictly faithful to the theory. To make a better use of the sample, we could however think about using some cross validation scheme, instead of cutting the sample into two independent chunks, $(X_1, \dots, X_M)$ and $(X_{M+1}, \dots, X_K)$.

The second point discussed in the paper is a two step estimation scheme, where a first coarse estimate is used to code the data into the unit interval, in a close to uniform way. We showed that it is a way to cut down computation and memory requirements when the distribution to be estimated is close to be singular with respect to the reference measure used to define the histograms. This idea could be generalized to multiple step schemes.

Appendix A: Proof of lemma 3.1

In both cases

$$\begin{aligned}
\widehat{Q}[F(X) \leq F(x)] &= \widehat{Q}\big[\sigma(X) \leq \sigma(x)\big] + \widehat{Q}\big[\sigma(X) > \sigma(x), F(X) = F(x)\big] \\
&= \widehat{Q}\big[\sigma(X) \leq \sigma(x)\big] + \widehat{Q}\Big\{\sigma(X) > \sigma(x), \frac{1}{2}\widehat{Q}\big[\sigma^{-1}(]\sigma(x), \sigma(X)])\big] \\
&\qquad + \frac{1}{2}\widehat{Q}\big[\sigma^{-1}([\sigma(x), \sigma(X)[)\big] = 0\Big\} \\
&= \widehat{Q}\Big[\sigma(X) \leq \sigma(x)\Big] \\
&\qquad + \widehat{Q}\Big\{\sigma(X) > \sigma(x), \widehat{Q}\big[\sigma^{-1}([\sigma(x), \sigma(X)])\big] = 0\Big\} \\
&= \widehat{Q}\big[\sigma(X) \leq \sigma(x)\big].
\end{aligned}$$

In the same way

$$\begin{aligned}
\widehat{Q}\Big[F(X) < F(x)\Big] &= \widehat{Q}\Big[\sigma(X) < \sigma(x)\Big] - \widehat{Q}\Big[\sigma(X) < \sigma(x), F(X) = F(x)\Big] \\
&= \widehat{Q}\Big[\sigma(X) < \sigma(x)\Big] \\
&\qquad - \widehat{Q}\Big\{\sigma(X) < \sigma(x), \widehat{Q}\big[\sigma^{-1}([\sigma(X), \sigma(x)])\big] = 0\Big\} \\
&= \widehat{Q}\Big[\sigma(X) < \sigma(x)\Big].
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{1}{2}\Big\{\widehat{Q}\big[F(X) \leq F(x)\big] + \widehat{Q}\big[F(X) < F(x)\big]\Big\} \\
&\qquad = \frac{1}{2}\Big\{\widehat{Q}\big[\sigma(X) \leq \sigma(x)\big] + \widehat{Q}\big[\sigma(X) < \sigma(x)\big]\Big\} = F(x).
\end{aligned}$$

Let us assume now that we are in the case when $\mathcal{S} = A^* \cup \{\varnothing\}$ is an infinite set.

Let us prove that 1. is equivalent to 3.. For any $r \in [0,1]$, $\widehat{Q} \circ F^{-1}(\{r\}) = \widehat{Q} \circ \sigma^{-1}\big(\{s \in A^{\mathbb{N}} : F \circ \sigma^{-1}(s) = r\}\big)$, because $F(x)$ is a function of $\sigma(x)$ only. Moreover, as F is non decreasing, $\{s \in A^{\mathbb{N}} : F \circ \sigma^{-1}(s) = r\}$ is an interval. If this interval contains two distinct points $s < s'$, then, as $F(\sigma^{-1}(s)) = F(\sigma^{-1}(s'))$, $\widehat{Q} \circ \sigma^{-1}([s, s']) = 0$, from the definition of F. It follows in this case that $\widehat{Q} \circ F^{-1}(r) = 0$, because any interval of $A^{\mathbb{N}}$ which is not reduced to one point is a countable union of closed intervals of the form $[s, s']$, with $s < s'$. On the other hand, if $\{s \in A^{\mathbb{N}} : F \circ \sigma^{-1} = r\}$ is a one point set, it is of the form $\{\sigma(x)\}$, because σ is surjective, thus in this case

$\widehat{Q} \circ F^{-1}(r) = \widehat{Q}(\sigma(X) = \sigma(x))$. Thus 3. implies 1.. In the other direction, $\widehat{Q}\big(\sigma(X) = \sigma(x)\big) \leq \widehat{Q} \circ F^{-1}\big(F(x)\big)$, and therefore 1. is equivalent to 3..

The Lebesgue measure having no atom, 2. implies 1. and 3. Now assume 1. (and consequently 3.). Remark that if $r < r' \in [0,1]$ are such that $\widehat{Q} \circ F^{-1}([0,r]) < \widehat{Q} \circ F^{-1}([0,r'])$, then there is $x \in \mathcal{X}$ such that $F(x) \in]r, r']$. Using 1. and the first part of the lemma, we see that $\widehat{Q} \circ F^{-1}([0, F(x)]) = F(x)$, and therefore that

$$\widehat{Q} \circ F^{-1}([0,r]) \leq F(x) \leq r',$$
$$\widehat{Q} \circ F^{-1}([0,r']) \geq F(x) \geq r.$$

Now assume that for some $r \in [0,1]$, $\widehat{Q} \circ F^{-1}([0,r]) < r$. Let

$$r' = \sup\{p : \widehat{Q} \circ F^{-1}([0,p]) = \widehat{Q} \circ F^{-1}([0,r])\}.$$

Then for any $r'' > r'$, we should have that

$$\widehat{Q} \circ F^{-1}([0,r'']) \geq r' \geq r > \widehat{Q} \circ F^{-1}([0,r]) = \widehat{Q} \circ F^{-1}([0,r']).$$

This would imply that $p \mapsto \widehat{Q} \circ F^{-1}([0,p])$ is not right continuous at r', which is a contradiction.

In the same way, if for some $r \in [0,1]$, $\widehat{Q} \circ F^{-1}([0,r]) > r$, we can consider

$$r' = \inf\{p : \widehat{Q} \circ F^{-1}([0,p]) = \widehat{Q} \circ F^{-1}([0,r])\}.$$

Then for any $r'' < r'$, we would have that

$$\widehat{Q} \circ F^{-1}([0,r'']) \leq r' \leq r < \widehat{Q} \circ F^{-1}([0,r]) = \widehat{Q} \circ F^{-1}([0,r']),$$

and $\widehat{Q} \circ F^{-1}$ would have an atom at r'. This shows that necessarily $\widehat{Q} \circ F^{-1}\big([0,r]\big) = r$ for any $r \in [0,1]$, and therefore that $\widehat{Q} \circ F^{-1}$ is the Lebesgue measure as stated in statement number 2.

Appendix B: Listings from experiments on text

This is a listing of profile data showing how much time is spent in the different functions of the program, while executing 50 trials of the two step text experiment described in table 1. Mon Oct 15 10:58:02 2001

```
Flat profile:

Each sample counts as 0.01 seconds.
  %   cumulative   self              self     total
 time   seconds   seconds    calls  us/call  us/call  name
 87.62      4.67     4.67      100 46700.00 46700.00  Kullback
```

```
  7.13       5.05     0.38       100  3800.00  3800.00  WeightMix
  1.88       5.15     0.10       201   497.51   497.51  CountFill
  1.50       5.23     0.08         1 80000.00 80000.00  Count2Dens
  0.75       5.27     0.04       200   200.00   200.00  BootStrap
  0.38       5.29     0.02       100   200.00   200.00  SampleCode
  0.19       5.30     0.01     12800     0.78     0.78  QuantileApply
  0.19       5.31     0.01       100   100.00   100.00  BetaOpt
  0.19       5.32     0.01       100   100.00   100.00  WeightFill
  0.19       5.33     0.01        50   200.00   200.00  Dist2Dens
```

We cut the list when, up to the accuracy of the profile measurements, the cumulative time reaches its final value. The most interesting function is WeightMix, which computes the weights $\Upsilon_s(x)$. It is where most time is spent. We reproduce its code here, to give a flavor of the way we implemented our method. Notice that we move through the tree $\pi(\overline{\mathfrak{D}})$ by shifting array indices.

```
#define LOGP(x,y) \
(((x)>(y))?(x)+log1p(exp((y)-(x))):(y)+log1p(exp((x-y))))
#define LASTB 1
#define OTHERB (~1)
#define BROTHER(i) (((i)&OTHERB)|((~(i))&LASTB))

Weight *WeightMix(Weight *w,Count *c,double alpha, double beta) {
    int depth, dd;
    int i,j,M,brother;
    Weight *mixW;
    double buff, *wp, *mixWp;
    int *cp;
    double sup;
    double ac, al, acl;
    double right;
    depth = w->depth;
    if (c->depth < depth) {
        depth = c->depth;
    }
    mixW = WeightNew(depth);
    M = 1 << depth;
    ac = 1 - alpha;
    al = log(alpha);
    acl = log(1-alpha);
    cp=c->first;
```

```
wp=w->first;
mixWp=mixW->first;
for (i=(M>>1);i<M;i++) {
    right = al+(beta*cp[i<<1])*wp[i<<1]
        + (beta*cp[(i<<1)+1])*wp[(i<<1)+1];
    mixWp[i] = LOGP(acl,right);
}
for(i=(M>>1)-1;i;i--) {
    right = al+mixWp[i<<1]+mixWp[(i<<1)+1]
        + (beta*cp[i<<1])*wp[i<<1]
        + (beta*cp[(i<<1)+1])*wp[(i<<1)+1];
    mixWp[i] = LOGP(acl,right);
}
M = 1 << (depth+1);
sup = 0;
for(i=(1<<depth);i<M;i++) {
    brother = BROTHER(i);
    right = al+((beta*cp[i])+1)*wp[i]
        + (beta*cp[brother])*wp[brother];
    buff = LOGP(acl,right);
    for(j=(i>>1),dd=depth-1;j>1;j>>=1,dd--) {
        brother = BROTHER(j);
        right = al+buff+mixWp[brother]
            + ((beta*cp[j])+1)*wp[j]
            + (beta*cp[brother])*wp[brother];
        buff = LOGP(acl,right);
    }
    mixWp[i] = buff;
    if (buff > sup) {
        sup = buff;
}
/* normalizing the weights in two steps to
    make things numerically more stable */
for (i=(1<<depth);i<M;i++) {
    mixWp[i] -= sup;
}
for (i=(1<<depth)-1;i;i--) {
    mixWp[i] = LOGP(mixW->first[i<<1],mixW->first[(i<<1)+1]);
}
for (i=M-1;i;i--) { /* back from the log
```

```
                                        representation of weights */
        mixWp[i] = exp(mixWp[i]-mixWp[1]);
    }
    return mixW;
}
```

Note that we preferred *not* to use equation (43) to compute the normalizing factor of the estimated distribution, in order to increase numerical stability and be sure that round off errors do not prevent the estimated distribution from summing up to one within a good accuracy. Numerical stability was checked for samples up to size $N = 10^7$ (which is the maximum size we tried).

The following reports of batch sessions show that it is necessary to increase the depth d of the maximal dictionary by two units, going from 8 to 10, to reach the same accuracy with a one step estimation scheme :

L	N	K	$\overline{\mathcal{D}}$	$\mathcal{D}_\infty$	$\mathbb{E}[\mathcal{K}(P,\widehat{Q})]$	$\mathbb{E}[\mathcal{K}(P,\check{Q})]$
7072	1000	100	$\{0,1\}^8$	$\{0,1\}^{18}$	5.12 ± 0.08	3.70 ± 0.07
7072	1000	1000	$\{0,1\}^8$	$\{0,1\}^{18}$	4.56 ± 0.01	4.561 ± 0.01
7072	1000	1000	$\{0,1\}^9$	$\{0,1\}^{18}$	4.017 ± 0.009	4.017 ± 0.009
7072	1000	1000	$\{0,1\}^{10}$	$\{0,1\}^{18}$	3.788 ± 0.014	3.788 ± 0.014

Table 2.

These are the profile data for the last batch session :

```
Flat profile:

Each sample counts as 0.01 seconds.
  %   cumulative   self              self     total
 time   seconds   seconds    calls  us/call  us/call  name
 69.72      4.72     4.72      100 47200.00 47200.00  Kullback
 24.96      6.41     1.69      100 16900.00 16900.00  WeightMix
  1.62      6.52     0.11      201   547.26   547.26  CountFill
  1.18      6.60     0.08        1 80000.00 80000.00  Count2Dens
  1.03      6.67     0.07      100   700.00   700.00  WeightLog
  0.59      6.71     0.04    51200     0.78     0.78  QuantileApply
  0.30      6.73     0.02      200   100.00   100.00  BootStrap
  0.30      6.75     0.02      100   200.00   200.00  Weight2Dist
  0.15      6.76     0.01      100   100.00   100.00  BetaSet
  0.15      6.77     0.01       50   200.00  1000.00  Est2Dens
```

The `Kullback` function is used only to compute the Kullback divergence between estimated and true distributions, therefore it is not part of the estimation algorithm itself, but is just a tool used to test its performance.

The important thing is the time spent in `WeightMix`, which jumps to 1.69 seconds, whereas it was only 0.38 seconds for the two step scheme.

N.B.: We included in the following bibliography some references about adaptive regression estimation, because of its numerous links with density estimation.

References

1. Y. Baraud (1998) *Sélection de modèles et estimation adaptative dans différents cadres de régression,* PHD, Université Paris 11.
2. Y. Baraud, F. Comte and G. Viennet (1998) Adaptive estimation in an autoregressive and a geometrical beta-mixing framework, *preprint 98-07* CREST-INSEE.
3. Y. Baraud, F. Comte and G. Viennet (1999) Model selection for (auto-)regression with dependent data, *preprint.*
4. A. Barron (1987) Are Bayes Rules Consistent in Information? *Open Problems in Communication and Computation, T. M. Cover and B. Gopinath Ed.*, Springer Verlag 1987.
5. A. Barron and Y. Yang (1999) Information Theoretic Determination of Minimax Rates of Convergence, *Ann. Statist.* **27**, no. 5, 1564–1599.
6. A. Barron, L. Birgé and P. Massart, (1999) Risk bounds for model selection via penalisation, *Probab. Theory Related Fields* **113**, no 3, 301-413.
7. L. Birgé and P. Massart (1997) From model selection to adaptive estimation, *Festschrift for Lucien Le Cam*, 55–87, Springer, New York.
8. L. Birgé and P. Massart (1998) Minimum contrast estimators on sieves : exponential bounds and rates of convergence, *Bernoulli*, **4** (1998), no. 3, 329–375.
9. L. Birgé and P. Massart (1996) An adaptive compression algorithm in Besov spaces, *Prépublication d'Orsay numéro 39,* `http://www.math.u-psud.fr/~biblio/pub/1996/ppo_1996.html`
10. L. Birgé and Y. Rozenholc (1999) How many bins should be put in a regular histogram?, *preprint.*
11. G. Castellan (1999) Modified Akaike's criterion for histogram density estimation, *preprint* `http://www.math.u-psud.fr/~biblio/pub/1999/ppo_1999.html`
12. Castellan, G. (2000) Sélection d'histogrammes à l'aide d'un critère de type Akaike. [Histogram selection with an Akaike-type criterion] *C. R. Acad. Sci.* Paris Sr. I Math. **330**, no. 8, 729–732.
13. O. Catoni (1997) The mixture approach to universal model selection,

preprint LMENS-97-22
`http://www.dmi.ens.fr/EDITION/preprints/1997/resu9722.html`

14. Olivier Catoni. Universal aggregation rules with sharp oracle inequalities. *Annals of Statistics, to appear*, pages 1–37, 1999. Revised and augmented from *A mixture approach to universal model selection.*
15. Olivier Catoni. Gibbs estimators. *Probab. Theory Relat. Fields, to appear*, pages 1–23, 2000. see `http://www.dmi.ens.fr/preprints`.
16. T. M. Cover and J. A. Thomas (1991) *Elements of information theory,* John Wiley, New York, 542 pages.
17. Csiszár I. and Körner J. (1981) *Information theory : coding theorems for discrete memoryless systems,* Academic Press, New York.
18. Härdle, Wolfgang; Kerkyacharian, Gerard; Picard, Dominique; Tsybakov, Alexander (1998) Wavelets, approximation, and statistical applications. Lecture Notes in Statistics, 129. Springer-Verlag, New York. xviii+265 pp.
19. Donoho, David L.; Johnstone, Iain M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* 90, no. 432, 1200–1224.
20. Donoho, David L.; Johnstone, Iain M. (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, no. 3, 425–455.
21. Donoho, David L.; Johnstone, Iain M. (1999) Asymptotic minimaxity of wavelet estimators with sampled data. *Statist. Sinica* 9, no. 1, 1–32.
22. Donoho, David L.; Johnstone, Iain M.; Kerkyacharian, Gérard; Picard, Dominique (1996) Density estimation by wavelet thresholding. Ann. Statist. 24, no. 2, 508–539.
23. Donoho, D. L.; Johnstone, I. M.; Kerkyacharian, G.; Picard, D. (1997) Universal near minimaxity of wavelet shrinkage. Festschrift for Lucien Le Cam, 183–218, Springer, New York.
24. M. Feder and N. Merhav, (1996) Hierarchical Universal Coding, *IEEE Trans. Inform. Theory,* vol 42, no 5, Sept.
25. Goldenshluger, A.; Nemirovski, A. (1997) On spatially adaptive estimation of nonparametric regression. Math. Methods Statist. 6, no. 2, 135–170.
26. I. Johnstone (1998) Function estimation and wavelets, *Lecture Notes, ENS Paris.*
27. A. Juditsky, A. Nemirovski (1996) Functional aggregation for nonparametric regression in 4th World Congress of the Bernouilli Society, Vienna, Austria, August 26-31.
28. Lugosi, Gábor; Nobel, Andrew (1996) Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.*

24, no. 2, 687–706.

29. A. Nemirovski (1998) Topics in Non-Parametric Statistics *Saint-Flour Summer School on Probability Theory.*
30. Neveu, Jacques, (1965) *Mathematical foundations of the calculus of probability*, Holden-Day, San-Francisco.
31. Nobel, Andrew (1996) Histogram regression estimation using data-dependent partitions. *Ann. Statist.* 24, no. 3, 1084–1105.
32. J. Rissanen (2001) Lectures on statistical modeling theory, *Lecture notes*, avaible at the author's site `http://www.cs.tut.fi/~rissanen/`
33. J. Rissanen (1983) A universal data compression system, *IEEE Trans. Inform. Theory*, Vol. IT-29, no 5, 656-664.
34. B. Y. Ryabko, (1984) Twice-universal coding, *Probl. Inform. Transm.*, vol 20, no 3, pp. 24-28, July-Sept.
35. Vapnik V. (1998) *Statistical Learning Theory*, Wiley.
36. M. J. Weinberger, J. Rissanen and M. Feder (1995) A universal finite memory source, *IEEE Trans. Inform. Theory*, Vol. IT-41, no 3, pp 643-652.
37. F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens (1995) The Context-Tree Weighting Method: Basic Properties, *IEEE Trans. Inform. Theory*, vol 41, no 3, May, 1995.
38. F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens (1996) Context Weighting for General Finite-Context Sources, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.
39. Yang, Y. (2000) Mixing strategies for density estimation. *Ann. Statist.* **28**, no. 1, 75–87.
40. Yang Yuhong (1998) Combining Different Procedures for Adaptive Regression, *preprint* `http://www.public.iastate.edu/~stat/update/reports.html`

BIFURCATIONS OF LIMIT CYCLES IN Z_Q-EQUIVARIANT PLANAR VECTOR FIELDS OF DEGREE 5

H.S.Y.CHAN AND K.W.CHUNG

Department of Mathematics, City University of Hong Kong, Hong Kong, China

JIBIN LI

School of Science, Kunming University of Science and Technology, Kunming, China

Consider the weakened Hilbert's 16th problem for symmetric planar perturbed polynomial Hamiltonian systems. In particular, for perturbed Hamiltonian polynomial vector fields of degree 5, numerical examples of Z_q-equivariant planar perturbed Hamiltonian systems are constructed, having maximal number of centers. They give rise to different configurations of limit cycles forming compound eyes. These are studied by the bifurcation theory of planar dynamical systems and the method of detection functions. With the help of numerical analysis, it is shown that there exist parameter groups such that a polynomial vector field of degree 5 has at least 20 limit cycles with Z_5 symmetry and at least 24 limit cycles with Z_6 symmetry. There is reason to conjecture that the Hilbert number $H(2k+1) \geq (2k+1)^2-1$ for the perturbed Hamiltonian systems.

1 Introduction

One of the problem posed by Smale[26] in his *"Mathematical Problems for the Next Century"* is **Hilbert's 16th problem.** It is well known that Hilbert's 16th problem consists of two parts. The first part studies the mutual disposition of maximal number(in the sense of Harnack) of separate branches of an algebraic curve, and also the "corresponding investigation" for non-singular real algebraic varieties; and the second part poses the questions of the maximal number and relative dispositions of limit cycles of the planar polynomial vector field:

$$\frac{dx}{dt} = P_n(x,y), \qquad \frac{dy}{dt} = Q_n(x,y), \tag{E_n}$$

where P_n and Q_n are polynomials of degree n(See Hilbert[11], Farkas[8], Ye[29], Zhang et al [31]). As professor Smale said: "Except for the Riemann hypothesis, it seems to be the most elusive of Hilbert's problems." In fact, for the first part, the specialists of the real algebraic geometry usually study the topology of non-singular real planar projective algebraic curves of degree m. Up to now, we know the schemes of mutual arrangement of ovals realized by M-curves only for $m \leq 7$(see Gudkov[10], Viro[27] and Wilson[28] etc).

For the second part, the answer still seems to be far away. Let $H(n)$ be the maximal number of limit cycles of (E_n). Up to now, we only know that the system (E_n) always has a finite number of limit cycles(see Ilyashenko[12]) and that $H(2) \geq 4, H(3) \geq 11$ (see recent discussions in Chan et al[4], Li[14,15], ,Lloyd[20], Luo[21], Perko[23], Ye[29] and etc). Also by considering a small neighbourhood of a singular point, $H(n) \geq (n^2 + 5n - 20 - 6(-1)^n)/2$ for $n \geq 6$(see Otrokov[22]). Recently, Christopher and Lloyd[6] showed that $H(2^k - 1) \geq 4^{k-1}(2k - \frac{35}{6}) + 3.2^k - \frac{5}{3}$ (for example $H(7) \geq 25$) by perturbing some families of closed orbits of a Hamiltonian system. However, in global phase plane, when $n > 3$, how many cycles can be created by (E_n), and what global limit cycle configurations can appear? These are still interesting open problems. In order to obtain more limit cycles and various configuration patterns of their relative dispositions, one of us indicated in [13]-[17] that an efficient method is to perturb the symmetric Hamiltonian systems having maximal number of centers, i.e., to study the weakened Hilbert's 16th problem posed by V.I.Arnold[2] in 1977 for the symmetric planar polynomial Hamiltonian systems, since bifurcation and symmetry are closely connected and symmetric systems play pivotal roles as a bifurcation point in all planar Hamiltonian system class. To investigate perturbed Hamiltonian systems, we should first know the global behaviour of unperturbed polynomial systems, namely, determine the global property for the families of real planar algebraic curves defined by the Hamiltonian functions. Then by using proper perturbation techniques, we shall obtain the global information of bifurcations for the perturbed non-integrable systems. In this sense, we say that our study method will utilize both parts of Hilbert's 16th problem.

On the basis of the method of detection functions posed by Li[13], we give in this paper a method of control parameters in order to obtain more limit cycles for some Z_q-equivariant perturbed polynomial Hamiltonian system of degree $n = 5$ $(q = 2-6)$. With the help of numerical analysis (using Maple[1] or Mathematica) we show that there exist parameter groups such that there exists at least 15 to 24 limit cycles having the configurations of compound eyes in these systems. The cases for $q = 2, 3$ and 4 are being reported separately in [5] and [19] where at least 15 and 23 limit cycles are respectively obtained. The case for $q = 6$ is discussed in [18]. Here we will give an overall view of the process and fill in more details using $q = 5$ and 6 as examples.

The paper is divided into five sections. Section 2 gives a general Z_q-equivariant fifth degree planar polynomial vector fields and their Hamiltonian forms. As examples, for $q = 5, 6$ we discuss the behaviour of sextic algebraic curves defined by the Hamiltonian vector fields. In Section 3 we introduce the detection function and its properties for the perturbed planar Hamiltonian systems. In Section 4 we add Z_5-invariant perturbations to the Hamiltonian system and consider five detection functions which correspond to different period annuluses. We determine the perturbed parameter values from given bifurcation conditions and by using the method of control parameters. These bifurcation parameters ensure that the perturbed systems have at least 15 and 20 limit cycles. In Section 5, for $q = 6$, we introduce some new results of bifurcations of phase portraits and give a conjucture: $H(2k+1) \geq (2k+1)^2 - 1$. In other words, we believe that $H(n)$ is increasing at least as order n^2 for the perturbed symmetric Hamiltonian systems.

2 Z_q-equivariant planar vector fields

Definition 2.1 Let G be a compact Lie group of transformations acting on R^n. A mapping $\Phi : R^n \to R^n$ is called G-equivariant if for all $g \in G$ and $x \in R^n, \Phi(gx) = g\Phi(x)$. A function $H : R^n \to R$, is called G-invariant function if for all $x \in R^n, H(gx) = H(x)$. If Φ is a G-equivariant mapping, the vector field $dx/dt = \Phi(x)$ is called a G-equivariant vector field.

Let q be an integer. A group Z_q is called a cyclic group if it is generated by a planar counterclockwise rotation through $2\pi/q$ about the origin. Making the transformation $z = x + iy, \bar{z} = x - iy$, the system (E_n) becomes

$$\frac{dz}{dt} = F(z, \bar{z}), \quad \frac{d\bar{z}}{dt} = \bar{F}(z, \bar{z}), \tag{2.1}$$

where $F(z, \bar{z}) = P(u, v) + iQ(u, v), u = (z + \bar{z})/2, v = (z - \bar{z})/2i$.

Theorem 2.1 (Li and Zhao[16]) A vector field defined by (2.1) is Z_q-equivariant, if and only if the function $F(z, \bar{z})$ has the following form:

$$F(z, \bar{z}) = \sum_{l=1} g_l(|z|^2)\bar{z}^{lq-1} + \sum_{l=0} h_l(|z|^2)z^{lq+1}, \tag{2.2}$$

where $g_l(w)$ and $h_l(w)$ are polynomials with complex coefficients in w. In addition, (2.1) is a Hamiltonian system having Z_q-equivariance, if and only if (2.2) holds and

$$\frac{\partial F}{\partial z} + \frac{\partial \bar{F}}{\partial \bar{z}} \equiv 0. \tag{2.3}$$

Theorem 2.2 (Z_q-invariant function) A Z_q-invariant function $I(z, \bar{z})$ has the following form:

$$I(z, \bar{z}) = \sum_{l=0} g_l(|z|^2)z^{lq} + \sum_{l=1} h_l(|z|^2)\bar{z}^{lq}. \tag{2.4}$$

Corollary 2.3 For the planar polynomial systems of degree 5, all nontrivial Z_q-equivariant vector fields have the following forms:

(1) $q = 6, F(z,\bar{z}) = (A_0 + A_1|z|^2 + A_2|z|^4)z + A_3\bar{z}^5$;

(2) $q = 5, F(z,\bar{z}) = (A_0 + A_1|z|^2 + A_2|z|^4)z + A_3\bar{z}^4$;

(3) $q = 4, F(z,\bar{z}) = (A_0 + A_1|z|^2 + A_2|z|^4)z + (A_3 + A_4|z|^2)\bar{z}^3 + A_5\bar{z}^5$;

(4) $q = 3, F(z,\bar{z}) = (A_0+A_1|z|^2+A_2|z|^4)z+(A_3+A_4|z|^2)\bar{z}^2+A_5z^4+A_6\bar{z}^5$;

(5) $q = 2, F(z,\bar{z}) = (A_0 + A_1|z|^2 + A_2|z|^4)z + (A_3 + A_4|z|^2 + A_5|z|^4)\bar{z}$ $+(A_6+A_7|z|^2)z^3+(A_8+A_9|z|^2)\bar{z}^3+A_{10}z^5+A_{11}\bar{z}^5$, where $A_j = a_j+ib_j (j = 0-11)$ are complex. The above $F(z,\bar{z})$ define Z_q-equivariant Hamiltonian vector fields if and only if $a_0 = a_1 = a_2 = 0$ and for $q = 4, A_4 = -5\bar{A}_5$; for $q = 3, A_4 = -4\bar{A}_5$; for $q = 2, A_4 = -3\bar{A}_6, A_5 = -2\bar{A}_7, A_9 = -5\bar{A}_{10}$.

The orbits of these Hamiltonian polynomial systems define different families of sextic ($m = 6$) algebraic curves having Z_q-equivariance. One of the main questions in real algebraic geometry is to describe what schemes of the mutual arrangement (schemes or configurations) of ovals can be realized by curves of given degree. By using some Z_q-equivariant Hamiltonian systems, we can realize a lot of configurations of ovals for planar algebraic curves of degree m. Rokhlin[25] listed all real schemes of ovals with $m \leq 5$. Leaving aside the simple case $m = 1,2,3$, we know that for $m = 4$, there are six real schemes: a nest of depth 2 and five unnested schemes with $l = 0,1,2,3,4$; for $m = 5$, there are eight real schemes: a scheme with a nest of depth 2, and seven unnested schemes with $l = 0,1,2,3,4,5,6$.

Theorem 2.4 For $m = 4$ and 5, all real schemes of projective algebraic curves in RP^2 can be realized by the orbits of Z_q-equivariant Hamiltonian vector fields.

For $m = 6$, the system (E_5) is Z_6-equivariant Hamiltonian system if and only if it can be reduced to the following 3-parameter family:

$$\frac{dr}{dt} = \beta r^5 \sin 6\theta,$$

$$\frac{d\theta}{dt} = 1 - 2\delta r^2 + (\alpha + \beta\cos 6\theta)r^4, \tag{2.5}$$

which has the Hamiltonian

$$H(r,\theta) = -\frac{1}{2}r^2 + \frac{1}{2}\delta r^4 - \frac{1}{6}(\alpha + \beta\cos 6\theta)r^6 \tag{2.6}$$

Suppose that $\alpha > \beta > 0, \delta = (\alpha + \beta + 1)/2$. Then, there exist 25 finite singular points of (2.5) at $(0,0),(z_1,0),(z_2,0),(z_3,\pi/6),(z_4,\pi/6)$ and their equivariant symmetric points, where

$$z_1 = \frac{1}{\sqrt{\alpha+\beta}}, z_2 = 1, z_3, z_4 = \sqrt{\frac{\delta \mp \sqrt{\delta^2 - \alpha + \beta}}{(\alpha - \beta)}}.$$

Proposition 2.5(see [18]) If the system (2.5) has 25 non-degenerate singular points and $\alpha > \beta > 0$, then there exist 3 topologically different phase portraits . In the affine real plane, the orbits of (2.5) can realize the following oval schemes:$1, 6, 7, \frac{6}{1}, \frac{1}{1}6, (1, 1, 1)$.

We next consider Z_5-equivariant Hamiltonian systems for $m = 6$. A system (E_5) is Z_5-equivariant Hamiltonian system if and only if it can be reduced to the following form in the polar coordinates:

$$\frac{dr}{dt} = \beta r^4 \sin 5\theta,$$

$$\frac{d\theta}{dt} = \alpha - \delta r^2 + (\beta \cos 5\theta) r^3 - r^4 = \Theta(r, \theta), \tag{2.7}$$

which has the Hamiltonian

$$H(r, \theta) = -\frac{1}{2}\alpha r^2 + \frac{1}{4}\delta r^4 - \frac{1}{5}(\beta \cos 5\theta) r^5 + \frac{1}{6} r^6. \tag{2.8}$$

It is easy to show that System (2.7) has at most 21 non-degenerate singular points in the phase plane. Without loss of generality, suppose that $(r, \theta) = (1, 0)$ is a singular point of (2.7). Then from $\Theta(r, 0) = 0$ and $\Theta(r, \pi/5) = 0$ we have $\delta = \alpha + \beta - 1$ and

$$f(r) = \alpha - \delta r^2 + \beta r^3 - r^4 = (1-r)(r^3 - (\beta - 1) r^2 + \alpha r + \alpha) = (1-r) f_3(r) = 0, \quad f(-r) = 0.$$

We may assume that (2.7) has 21 singular points. This implies that the function $f(r)$ has four real zeros, i.e., the following parameter condition holds:

$$\Delta = 4\alpha^3 + \alpha^2(8 + 20\beta - \beta^2) + \frac{4}{3}\alpha(\beta - 1)^4 - \frac{4}{27}(\beta - 1)^6 < 0.$$

Notice that $f_3'(r) = 3r^2 - 2(\beta - 1) r + \alpha$ has two real zeros at

$$\bar{r_1} = \frac{1}{3}((\beta - 1) - \sqrt{(\beta - 1)^2 - 3\alpha}), \qquad \bar{r_2} = \frac{1}{3}((\beta - 1) + \sqrt{(\beta - 1)^2 - 3\alpha}),$$

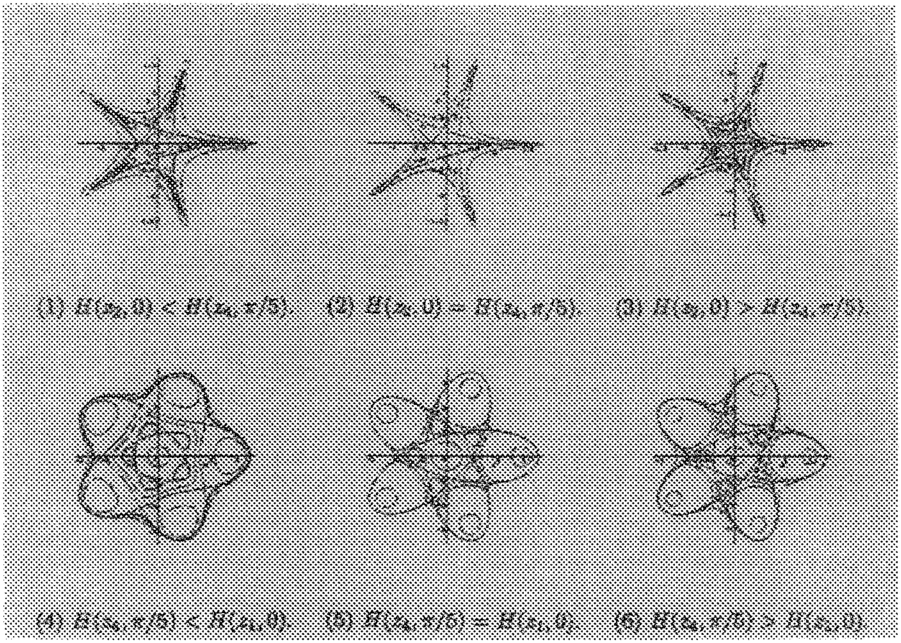
if $\Delta_1 = (\beta - 1)^2 - 3\alpha > 0$. Thus, we have the following conclusions.

(i) When $\alpha > 0, \Delta_1 > 0$, if $\beta > 1$, then there exist three singular points of (2.7) at $(r, \theta) = (z_1, 0), (z_2, 0)$ and $(z_3, 0)$ with $z_i > 0$. And there exists one singular point of (2.7) at $(r, \theta) = (z_4, \pi/5)$ with $z_4 > 0$. If $\beta < 1$, then there exists one singular point of (2.7) at the axis $\theta = 0$, and there exist three singular points of (2.7) at the line $\theta = \pi/5$.

(ii) When $\alpha < 0$, there exist two singular points $(z_i, 0)$ and $(z_j, \pi/5), i = 1, 2, j = i + 2$ of (2.7) at the axes $\theta = 0$ and $\theta = \pi/5$, respectively.

To determine the type (center or saddle point) of singular poinits $(z_i, 0)$ and $(z_j, \pi/5)$ of (2.7). We only need to use the signs of the Jacobians of the linearized system of (2.7):

$$J(z_i, 0) = -5\beta z_i^6 f'(z_i), \qquad J(z_j, \pi/5) = 5\beta z_j^6 f'(-z_j).$$



(1) $H(z_2, 0) < H(z_4, \pi/5)$. (2) $H(z_2, 0) = H(z_4, \pi/5)$. (3) $H(z_2, 0) > H(z_4, \pi/5)$.

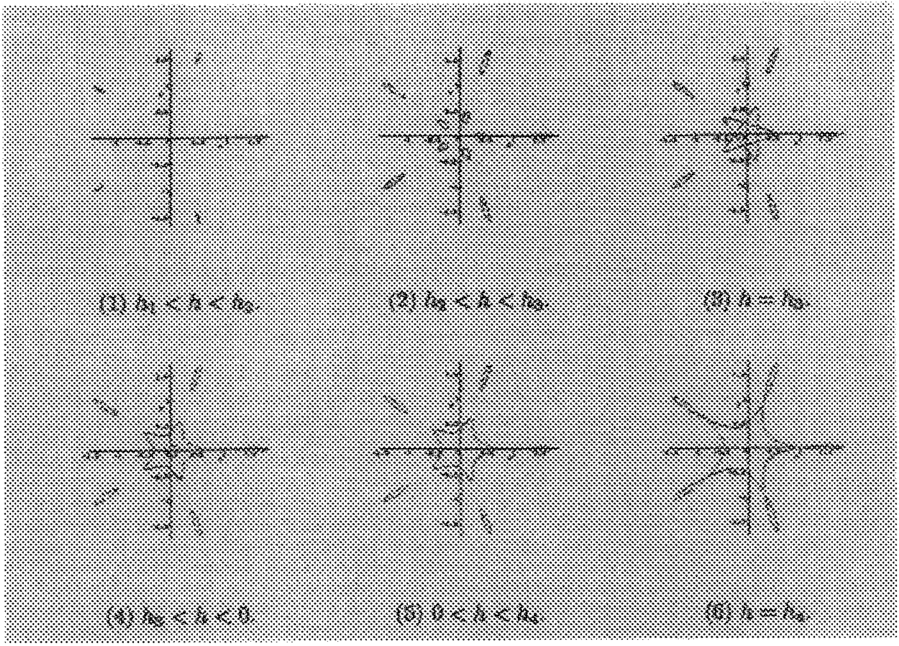
(4) $H(z_4, \pi/5) < H(z_1, 0)$. (5) $H(z_4, \pi/5) = H(z_1, 0)$. (6) $H(z_4, \pi/5) > H(z_1, 0)$.

Figure 1: Z_5-equivariant polynomial Hamiltonian vector fields of degree 5

Obviously, for $\alpha > 0$, the singular points $(z_2, 0), (z_4, \pi/5)$ are saddle points and $(z_1, 0), (z_3, 0)$ are centers; for $\alpha < 0$, the singular points $(z_1, 0), (z_4, \pi/5)$ are saddle points and $(z_2, 0), (z_3, \pi/5)$ are centers. Since the Hamiltonian defined by (2.8) is of definite sign near the origin $(0, 0)$ and at infinity, hence the origin ia a center and there exists a global periodic family of (2.7) surrounding all 21 singular points. By the above qualitative analysis and Z_5-equivariance of (2.7), we have the following result.

Proposition 2.6 For the Z_5-equivariant polynomial Hamiltonian system (2.7) having 21 non-degenerate singular points, there exist 6 topologically different phase portraits as shown in Figure 1.

As h varies, the level curves $H(r, \theta) = h$ of the Hamiltonian defined by (2.8) give rise to a family of sextic algebraic curves in the affine real plane. As an example, we



(1) $h_1 < h < h_2$. (2) $h_2 < h < h_3$. (3) $h = h_3$.

(4) $h_3 < h < 0$. (5) $0 < h < h_4$. (6) $h = h_4$.

Figure 2:Different schemes of ovals defined by (2.3) under the parameter condition G_1.

consider Figure 1 (3). Let $G_1 = (\alpha, \beta) = (0.1551515151, 2.7575757575)$ and $\delta = 1.91272727$. We have $z_1 = 0.4,\ z_2 = 1,\ z_3 = 1.6,\ z_4 = 0.2424242424$ and

$$h_1 = H(z_3, 0) = -0.05163442391,\ h_2 = H(z_1, 0) = -0.005135515151,$$

$$h_3 = H(z_4, \pi/5) = -.002411905557,\ h_4 = H(z_2, 0) = 0.0157575757.$$

As h increases from h_1 to ∞, the schemes of ovals of the sextic algebraic curves will be varied. This change process is shown in Figure 2. Similarly, we can discuss other portraits in Figure 1. Therefore, by using Gudkov's oval notation (see [10]), we have the conclusion as follows.

Proposition 2.7 In the affine real plane, the integral curves of the Hamiltonian system (2.7) can realize the following oval schemes:$1, 5, 6, 10, \frac{5}{1}, \frac{1}{1}5, (1, 1, 1)$.

3 The method of detection functions

Consider the following perturbed planar Hamiltonian system

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H}{\partial y} - \epsilon x(p(x,y) - \lambda), \\ \frac{dy}{dt} &= -\frac{\partial H}{\partial x} - \epsilon y(p(x,y) - \lambda), \end{aligned} \qquad (3.1)$$

where $H(x,y)$ is the Hamiltonian, $p(0,0) = 0, 0 < \epsilon \ll 1, \lambda \in R$.

Suppose that the origin in the phase plane is a singular point of (3.1) and the following conditions hold:

(A1) The unperturbed system $(3.1)_{\epsilon=0}$ is a Z_q-equivariant Hamiltonian vector field. For $h \in (h_1, h_2)$ one branch family of the curves $\{\Gamma^h\}$ defined by the Hamiltonian function $H(x,y) = h$ lies in a period annulus enclosing at least one singular point. As h increases, Γ^h expands outwards. When $h \to h_1, \Gamma^h$approaches a singular point or an inner boundary of the period annulus consisting of a heteroclinic (or homoclinic) loop.

(A2) Surrounding the period annulus, there exists a heteroclinic (or homoclinic) loop Γ^{h_2} at $h = h_2$ connecting some hyperbolic saddle points $(\alpha_i, \beta_i), 1 \le i \le q$.

(A3) The divergence $2(\lambda - F(x,y)) \equiv 2(\lambda - \frac{x}{2}\frac{\partial p}{\partial x} - \frac{y}{2}\frac{\partial p}{\partial y} - p(x,y))$ of the perturbed vector field is a Z_q -invariant function.

We define the function

$$\lambda = \lambda(h) = (\int\int_{D^h} F(x,y)dxdy)/(\int\int_{D^h} dxdy) = \psi(h)/\phi(h), \qquad (3.2)$$

which is called a detection function corresponding to the periodic family $\{\Gamma^h\}$. The graph of $\lambda = \lambda(h)$ in the plane (h, λ) is called a detection curve, where D^h is the area inside Γ^h.

Clearly, if $H(x,y) = h$ is a polynomial, then $\lambda(h)$ is a ratio between two Abelian integrals (see Carr, Chow and Hale[3]). In this case, $\lambda(h)$ is a differentiable function with respect to h. Of course, when the degree of $H(x,y)$ is more than 4, classical mathematical analysis cannot provide the calculating method for $\lambda(h)$ in general. We must use a numerical technique to compute these Abelian integrals.

On the basis of the Poincaré-Pontrjagin-Andronov theorem on the global center bifurcation and Melnikov method (see Perko[23], Guckenheimer and Holmes [9]), we have the following two conclusions (as in Li Jibin etc.[13][15]):

Theorem 3.1 (Bifurcation of limit cycles) Suppose that the conditions (A1) and (A3) hold. For a given $\lambda = \lambda_0$, considering the set S of the intersection points of the straight line $\lambda = \lambda_0$ and the curve $\lambda = \lambda(h)$ in the (h, λ)-plane, we have

(i) if S consists of exactly one point (h_0, λ_0) and $\lambda'(h) > 0(< 0)$, then there exists a stable(unstable) limit cycle of (3.1) near Γ^{h_0};

(ii) if S consists of two points (h_0, λ_0) and $(\tilde{h}_0, \lambda_0)$ having $\tilde{h}_0 > h_0$ and $\lambda'(\tilde{h}_0) > 0, \lambda'(h_0) < 0$, then there exist two limit cycles near $\Gamma^{\tilde{h}_0}$ and Γ^{h_0} respectively, the former is stable and the latter is unstable;

(iii) if S contains a point (h_0, λ_0) and $\lambda'(h_0) = \lambda''(h_0) = \cdots = \lambda^{(k-1)}(h_0) = 0$, but $\lambda^{(k)}(h_0) \neq 0$, then (3.1) has at most k limit cycles near Γ^{h_0};

(iv) if S is empty, then (3.1) has no limit cycle.

Theorem 3.2 (Bifurcation parameter created by a heteroclinic or homoclinic loop) Suppose that the conditions (A1),(A2) and (A3) hold. Then for $0 < \epsilon \ll 1$, when $\lambda = \lambda(h_2) + O(\epsilon)$, System (3.1) has a heteroclinic (or homoclinic) loop having Z_q-equivariance.

The following two propositions describe the properties of the detection function at the boundary values of h.

Proposition 3.3 (The parameter value of Hopf bifurcation) Suppose that as $h \to h_1$, the periodic orbit Γ^h of (3.1) approaches a singular point (ξ, η), then at this point the Hopf bifurcation parameter value is given by

$$b_H = \lambda(h_1) + O(\epsilon) = \lim_{h \to h_1} \lambda(h) + O(\epsilon) = F(\xi, \eta) + O(\epsilon). \tag{3.3}$$

Proposition 3.4 (Bifurcation direction of heteroclinic or homoclinic loop) Suppose that as $h \to h_2$, the periodic orbit Γ^h of (3.1) approaches a heteroclinic (or homoclinic) loop connecting a hyperbolic saddle point (α, β), where the saddle point value satisfies

$$SQ(\alpha, \beta) = 2\epsilon\sigma(\alpha, \beta) \equiv 2\epsilon(\lambda(h_2) - F(\alpha, \beta)) > 0(< 0),$$

then we have

$$\lambda'(h_2) = \lim_{h \to h_2} \lambda'(h) = -\infty(+\infty). \tag{3.4}$$

Remark 3.5

(1) If Γ^h contracts inwards as h increases, then the stability of limit cycles mentioned in Theorem 3.1 and the sign of $\lambda'(h_2)$ in (3.4) have the opposite conclusion.

(2) If the curve Γ^h defined by $H(x, y) = h$ $(h \in (h_1, h_2))$ consists of m components of oval families having Z_q-equivariance, then Theorem 3.1 gives rise to simultaneous global bifurcations of limit cycles from all these m oval families.

(3) If (3.1) has several different period annuluses filled by periodic orbit families $\{\Gamma_i^h\}$, then by calculating detection functions for every oval families, the global information of bifurcations of System (3.1) can be obtained.

4 Bifurcations of limit cycles of Z_5-equivariant perturbed Hamiltonian systems

In this section we consider the perturbed Z_5-equivariant vector field:

$$\frac{dr}{dt} = \beta r^4 \sin 5\theta - \epsilon r(pr^4 + qr^2 - \lambda),$$

$$\frac{d\theta}{dt} = \alpha - \delta r^2 + (\beta \cos 5\theta) r^3 - r^4 = \Theta(r,\theta). \tag{4.1}$$

Corresponding to (4.1), the function $F(x,y)$ in the divergence of vector field of hypothesis (A3) in Section 3 has the form:

$$F(r,\theta) = 3pr^4 + 2qr^2. \tag{4.2}$$

We first consider the case $G_2 = (\alpha,\beta) = (-7, 0.95)$ i.e., the unperturbed system (2.3) has the phase portrait of Figure 1 (6). Under this parameter condition group G_2, we have

$$z_1 = 1,\ z_2 = 3.026773093,\ z_3 = 1.306138497,\ z_4 = 1.770634596$$

and

$$H(z_2,0) < 0 < H(z_1,0) < H(z_4,\pi/5) < H(z_3,\pi/5), \tag{4.3}$$

where

$$h_1 = H(z_2,0) = -35.97718144,\ h_2 = H(z_1,0) = 1.714166667,$$

$$h_3 = H(z_4,\pi/5) = 2.09184063,\ h_4 = H(z_3,\pi/5) = 2.391163167.$$

From (4.3), we see that as h increases from h_1 to ∞, the schemes of ovals of the sextic algebraic curves defined by $H(r,\theta) = h$ will be varied as follows:

(1) $h \in (h_1, h_2)$: there exist five period annuluses $\{\Gamma_{1i}^h\}, i = 1-5$, enclosing the center $(z_2, 0)$ and its Z_5-equivariant symmetry points.

(2) $h \in (0, h_2)$: there is a period annulus $\{\Gamma_0^h\}$ enclosing the origin $(0,0)$. Together with $\{\Gamma_{1i}^h\}$, there exist 6 period annuluses.

(3) $h = h_2$: there exist 5 homoclinic orbits $\{\Gamma_{1i}^{h_2}\}$ and a heteroclinic 5-cycle enclosing the origin.

(4) $h \in (h_2, h_3)$: there is a period annulus $\{\Gamma_2^h\}$ enclosing 11 singular points.

(5) $h = h_3$: there exist 5 homoclinic orbits $\{\Gamma_{3i}^{h_3}\}$ enclosing the singular point $(z_3, \pi/5)$ and its Z_5-equivariant symmetry points, and there is a heteroclinic 5-cycles enclosing 11 singular points.

(6) $h \in (h_3, h_4)$: there exist 5 period annuluses $\{\Gamma_{3i}^h\}, i = 1-5$, enclosing the singular point $(z_3, \pi/5)$ and its Z_5-equivariant symmetry points, and there is a global period annulus $\{\Gamma_4^h\}$ enclosing all 21 singular points.

(7) $h \in (h_4, \infty)$: there is a global period annulus $\{\Gamma_4^h\}$ enclosing all 21 singular points.

Notice that as h increases, the periodic orbits Γ_{3i}^h constract inwards while all other periodic orbits expand outwards.

We compute five detection functions defined by (3.2) which correspond to the above five types of period annuluses $\{\Gamma_0^h\} - \{\Gamma_4^h\}$.

$$\lambda_i(h) = \frac{\int\int_{D_i^h} F(r,\theta) r dr d\theta}{\int\int_{D_i^h} r dr d\theta} = \frac{\psi_i(h)}{\phi_i(h)} = \frac{1}{\phi_i(h)}[3pI_{i1}(h) + 2qI_{i2}(h)]$$

$$= 3pJ_{i1}(h) + 2qJ_{i2}(h), \quad i = 0, \cdots, 4, \tag{4.4}$$

where $J_{ij}(h) = I_{ij}(h)/\phi_i(h), \quad j = 1, 2$, and

$$\phi_i(h) = \int\int_{D_i^h} r dr d\theta, \quad D_i^h \textit{ is the area inside } \Gamma_i^h,$$

$$I_{i1}(h) = \int\int_{D_i^h} r^5 dr d\theta, \quad I_{i2}(h) = \int\int_{D_i^h} r^3 dr d\theta.$$

For the given parameter group G_2 the functions $J_{ij}(h)$ can be numerically calculated to a given degree of accuracy. We will give these results in the Appendix.

By using the theory given in Section 3, we immediately obtain the following values of bifurcation parameters and bifurcation direction detections.

(i) **Hopf bifurcation parameters:**

(1) Bifurcation from the origin $(0,0)$:

$$b_0^H = \lambda_0(0) + O(\epsilon) = 0 + O(\epsilon);$$

(2) Simultaneous bifurcations from the center $(z_2, 0)$ and its Z_5- equivariant symmetry points:

$$b_1^H = \lambda_1(h_1) + O(\epsilon) = F(z_2, 0) + O(\epsilon) = 3pz_2^4 + 2qz_2^2 + O(\epsilon)$$

$$= 251.7912959p + 18.32271071q + O(\epsilon);$$

(3) Simultaneous bifurcations from the center $(z_3, \pi/5)$ and its Z_5- equivariant symmetry points:

$$b_3^H = \lambda_3(h_4) + O(\epsilon) = F(z_3, \pi/5) + O(\epsilon) = 3pz_3^4 + 2qz_3^2 + O(\epsilon)$$

$$= 8.731285209p + 3.411995546q + O(\epsilon);$$

(ii) **Bifurcations from heteroclinic or homoclinic loops:**

(1) The heteroclinic bifurcation value from $\Gamma_0^{h_2}$:

$$\lambda_0(h_2) = 3pJ_{01}(h_2) + 2qJ_{02}(h_2) = 0.6447758181p + 0.7977862274q;$$

(2) The homoclinic bifurcation value from $\Gamma_{1i}^{h_2}$:

$$\lambda_1(h_2) = 3pJ_{11}(h_2) + 2qJ_{12}(h_2) = 132.5289769p + 11.90914478q;$$

(3) The homoclinic and heteroclinic loop bifurcation value from $\Gamma_2^{h_2}$:

$$\lambda_2(h_2) = 3pJ_{21}(h_2) + 2qJ_{22}(h_2) = 117.8557157p + 10.67290976q;$$

(4) The heteroclinic 5-cycle bifurcation value from $\Gamma_2^{h_3}$:

$$\lambda_2(h_3) = 3pJ_{21}(h_3) + 2qJ_{22}(h_3) = 107.7755921p + 9.989393194q;$$

(5) The homoclinic bifurcation value from $\Gamma_{3i}^{h_3}$:

$$\lambda_3(h_3) = \lambda_3(h_3) = 3pJ_{31}(h_3) + 2qJ_{32}(h_3) = 9.777634251p + 3.48267226q;$$

(6) The heteroclinic 5-cycle bifurcation value from $\Gamma_4^{h_3}$:

$$\lambda_4(h_3) = 3pJ_{41}(h_3) + 2qJ_{42}(h_3) = 102.7553682p + 9.656067918q.$$

(iii) **The values of bifurcation direction detections of heteroclinic and homoclinic loops:**

(1) $\sigma_0(z_1, 0) = \lambda_0(h_2) - F(z_1, 0) = -2.355224182p - 1.202213773q;$
(2) $\sigma_1(z_1, 0) = \lambda_1(h_2) - F(z_1, 0) = 129.5289769p + 9.90914478q;$
(3) $\sigma_2(z_1, 0) = \lambda_2(h_2) - F(z_1, 0) = 114.8557157p + 8.67290976q;$
(4) $\sigma_2(z_4, \pi/5) = \lambda_2(h_3) - F(z_4, \pi/5) = 78.28815436p + 3.719099448q;$
(5) $\sigma_3(z_4, \pi/5) = \lambda_3(h_3) - F(z_4, \pi/5) = -19.70980349p - 2.787621486q;$
(6) $\sigma_4(z_4, \pi/5) = \lambda_4(h_3) - F(z_4, \pi/5) = 73.26793046p + 3.385774172q;$

To control the perturbed parameters (p, q) such that System (4.1) has more limit cycles, we suppose that the following two conditions hold:

(C1) $\lambda_1(h_2) = \lambda_3(h_3) - 0.001, i.e., 122.7513426p + 8.42647252q + 0.001 = 0;$
(C2) $\sigma_1(z_1, 0) > 0, i.e., 129.5289769p + 9.90914478q > 0;$

These conditions imply that

$$q = -14.56734622p - 0.0001186736203, \quad p < -0.00007934395757.$$

As an example, taking $PG_2 = (p,q) = (-1, 14.56722755)$, then we have

$$\lambda_0(0) = 0, \quad \lambda_0(h_2) = 10.97675769;$$

$$\lambda_1(h_1) = 15.1198005, \quad \lambda_1(h_2) = 40.95424505, \quad max(\lambda_1(h)) \approx 41.0007;$$

$$\lambda_2(h_2) = 37.6189894, \quad \lambda_2(h_3) = 37.7421716;$$

$$\lambda_3(h_3) = 40.95524505, \quad \lambda_3(h_4) = 40.97203031, \quad min(\lambda_3(h)) \approx 40.94019;$$

$$\lambda_4(h_3) = 37.9067704, \quad max(\lambda_4(h) \approx 38.449682, \quad \lim_{h\to+\infty} \lambda_4(h) = -\infty;$$

and

$$\sigma_0 = -15.15769741 < 0,\ \sigma_1 = 14.8197899 > 0,\ \sigma_2(z_1, 0) = 11.4845343 > 0,$$

$$\sigma_2(z_4, \pi/5) = -24.11118646 < 0,\ \sigma_3 = -20.898113 < 0,\ \sigma_4 = -23.94658766 < 0.$$

It follows that under the parameter conditions of G_2 and PG_2, the system (4.1) has the graphs of detection curves as shown in Figure 3. We see from Figure 3 that when

$$\tilde{\lambda} \in (\lambda_1(h_2), \lambda_3(h_3)) = (40.95424505, 40.95524505), \tag{4.5}$$

in the (h,λ)-plane the straight line $\lambda = \tilde{\lambda}$ intersects the curves $\lambda = \lambda_1(h)$, $\lambda = \lambda_3(h)$ at two points respectively. By using the Z_5-equivariance of (4.1), we obtain from Theorem 3.1 the following result.

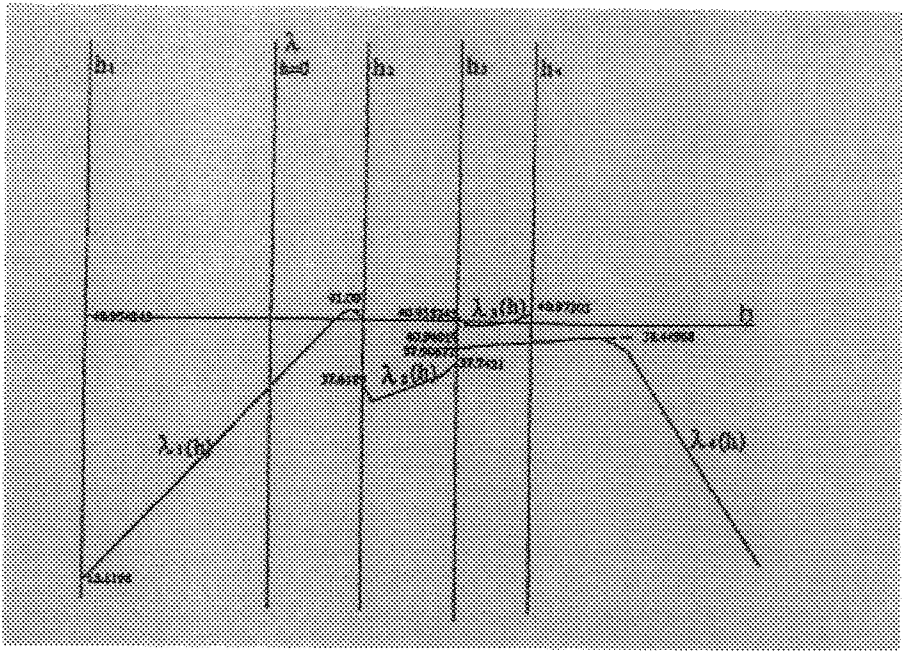


Figure 3: Graphs of detection curves of (4.1) with parameters G_2 and GP_2.

Theorem 4.1 For the parameter group G_2 and PG_2 and small $\epsilon > 0$, the system (4.1) has at least 20 limit cycles having the configuration shown in Figure 4 (1) for $\lambda = \tilde{\lambda}$ satisfies (4.5).

We now consider the system (4.1) having unperturbed parameter group $G_3 = (\alpha, \beta) = (0.2992021277, 2.680851063)$, i.e., the case of unperturbed system (2.3) has the phase portait shown as Figure 1(1). Under the parameter group G_3, we have

$$z_1 = 0.75,\ z_2 = 1,\ z_3 = 1.25,\ z_4 = 0.3191489362$$

and

$$H(z_3, 0) < H(z_1, 0) < H(z_2, 0) < H(z_4, \pi/5) < 0, \tag{4.6}$$

where

$$h_1 = H(z_3, 0) = -0.02570186111,\ h_2 = H(z_1, 0) = -0.02509791599,$$

$$h_3 = H(z_2, 0) = -0.0240913121,\ h_4 = H(z_4, \pi/5) = -0.008150769223.$$

From (4.6), we see that as h increases from h_1 to $+\infty$, the schemes of ovals of the sextic algebraic curves defined by $H(r, \theta) = h$ will be varied as follows:

(1) $h \in (h_1, h_3)$: there exist five period annuluses $\{\Gamma^h_{2i}\}, i = 1-5$, enclosing the center $(z_3, 0)$ and its Z_5-equivariant symmetry points.

(2) $h \in (h_2, h_3)$: there is a period annulus $\{\Gamma^h_{1i}\}, i = 1-5$ enclosing the center $(z_1, 0)$ and its Z_5-equivariant symmetry points. Together with $\{\Gamma^h_{1i}\}$, there exist 10 period annuluses.

(3) $h = h_3$: there exist 5 homoclinic loops $\{\Gamma^{h_3}_3\}$ in 'figure of eight' fashion connecting the saddle $(z_2, 0)$ and its Z_5-equivariant symmetry points.

(4) $h \in (h_3, h_4)$: there exist 5 period annuluses $\{\Gamma^h_{3i}\}, i = 1-5$ enclosing 3 singular points $(z_1, 0), (z_2, 0), (z_3, 0)$ and their Z_5-equivariant symmetry points.

(5) $h = h_4$:there exist 10 heteroclinic orbits $\{\Gamma^{h_4}_{3i}\}$ and $\{\Gamma^{h_4}_{0i}\}$ connecting the singular point $(z_3, \pi/5)$ and its Z_5-equivariant symmetry points.

(6) $h \in (h_4, 0)$: there exist two period annuluses $\{\Gamma^h_0\}$ enclosing the origin and $\{\Gamma^h_4\}$ enclosing all 21 finite singular points of (2.7).

(7) $h \in (0, \infty)$: there is a global period annulus $\{\Gamma^h_4\}$ enclosing all 21 singular points.

Notice that as h increases, the periodic orbits Γ^h_0 constract inwards, all other periodic orbits expand outwards. Corresponding to the above 5 different classes $\{\Gamma^h_0\} - \{\Gamma^h_4\}$ of the period annuluses, we also calculate 5 detection functions defined by (4.4) and obtain the bifurcation parameter values. To control the perturbed parameters (p, q) such that the system (4.1) has also more limit cycles, we suppose that

(C3) $\lambda_1(h_3) - \lambda_2(h_3) = 0$, (C4) $\sigma_1(z_1, 0) > 0$. This condition group implies that $p > 0$, $q = -3.092791861p$. As an example, taking $PG_3 = (p, q) = (1, -3.092791861)$, then we have

$$\lambda_0(0) = 0, \quad \lambda_0(h_4) = -0.2325782105;$$

$$\lambda_1(h_2) = -2.530172094, \quad \lambda_1(h_3) = -2.578978734;$$

$$\lambda_2(h_1) = -2.340755816, \quad \lambda_2(h_3) = -2.578978734;$$

$$\lambda_3(h_3) = -2.578978734, \quad \lambda_3(h_4) = -1.725044127;$$

$$\lambda_4(h_4) = -1.427510837, \quad \lim_{h \to +\infty} \lambda_4(h) = +\infty;$$

and

$$\sigma_0 = 0.3663369133 > 0, \ \sigma_1 = 0.606604988 > 0, \ \sigma_2 = 0.606604989 > 0,$$

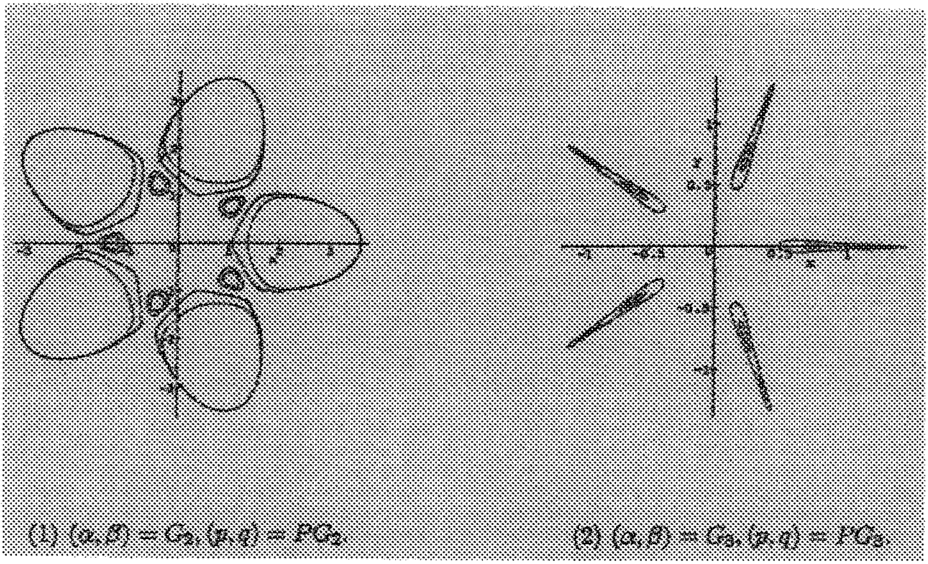
$$\sigma_3(z_4, \pi/5) = -1.126129003 < 0, \ \sigma_4 = -0.8285957134 < 0.$$

It follows that under the parameter conditions of G_3 and PG_3 with

$$\tilde{\lambda} \in (\lambda_1(h_3), \lambda_1(h_2)) = (-2.578978734, -2.530172094), \qquad (4.7)$$

in the (h, λ)-plane the straight line $\lambda = \tilde{\lambda}$ intersects the curves $\lambda = \lambda_1(h)$, $\lambda = \lambda_2(h)$ and $\lambda = \lambda_3(h)$ at one point respectively. By using the Z_5-equivariance of (4.1), we obtain from Theorem 3.1 the following result.

Theorem 4.2 For the parameter group G_3 and PG_3 and small $\epsilon > 0$, the system (4.1) has at least 15 limit cycles having the configuration shown in Figure 4(2) when $\lambda = \tilde{\lambda}$ satisfies (4.7).



(1) $(\alpha, \beta) = G_2$, $(p, q) = PG_2$

(2) $(\alpha, \beta) = G_3$, $(p, q) = PG_3$

Figure 4: The configurations of 20 and 15 limit cycles of System (4.1)

5 Bifurcations in Z_6-equivariant vector fields and conjucture

Consider the perturbed Z_6-equivariant vector field:

$$\frac{dr}{dt} = \beta r^5 \sin 6\theta - \epsilon r(pr^4 + qr^2 - \lambda),$$

$$\frac{d\theta}{dt} = 1 - 2\delta r^2 + (\alpha + \beta \cos 6\theta) r^4. \qquad (5.1)$$

Corresponding to (5.1), the function $F(x,y)$ in the divergence of vector field of the hypothesis (A3) in Section 3 has the same form as (4.2). We discuss the case $G_4 = (\alpha,\beta) = (1.4, 0.25)$. Under this parameter condition group, we have

$$H(z_1,0) < H(z_2,0) < H(z_3,\pi/6) < 0 < H(z_4,\pi/6),\ i.e,\ h_1 < h_2 < h_3 < 0 < h_4, \tag{5.2}$$

where $H(x,y)$ is defined by (2.6). From (5.2) we see that as h increases from $-\infty$ to h_4 the schemes of ovals of the sextic algebraic curves defined by $H(r,\theta) = h$ will be varied and there exist five types of period annuluses $\{\Gamma_0^h\} - \{\Gamma_4^h\}$ of (2.5). We define 5 detection functions $\lambda_i(h), i = 0-4$ similar to (4.4). By taking $PG_4 = (p,q) = (1, -3.376113233)$, we have the following values of bifurcation parameters and bifurcation direction detections(in detail, see[18]):

$$\lambda_0(0) = 0, \quad \lambda_0(h_3) = -1.207972277;$$

$$\lambda_1(h_1) = -2.990330089, \quad \lambda_1(h_2) = -2.96704993, \quad max(\lambda_1(h)) \approx -2.97;$$

$$\lambda_2(h_2) = -2.51962967, \quad \lambda_2(h_3) = -2.4842287;$$

$$\lambda_3(h_3) = -2.96604993, \quad \lambda_3(h_4) = -2.31457964, \quad min(\lambda_3(h)) \approx -2.967;$$

$$\lambda_4(h_2) = -2.564531384, \quad max(\lambda_4(h) \approx -2.5779, \quad \lim_{h\to-\infty} \lambda_4(h) = +\infty;$$

And

$$\sigma_0 > 0,\ \sigma_1 > 0,\ \sigma_2(z_2,0) > 0,$$

$$\sigma_2(z_3,\pi/6) > 0,\ \sigma_3 < 0,\ \sigma_4 > 0.$$

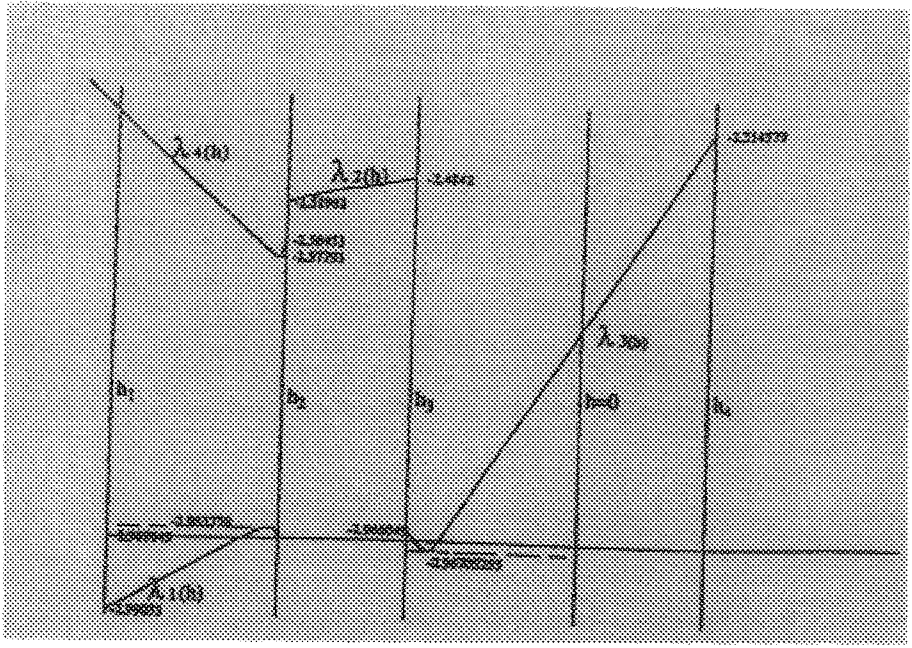


Figure 5: Graphs of detection curves of (5.1) with parameters G_4 and GP_4.

The above information follows that under the parameter conditions of G_4 and PG_4, the system (5.1) has the graphs of detection curves shown as Figure 5. We see from Figure 5 that when

$$\tilde{\lambda} \in (\lambda_1(h_2), \lambda_3(h_3)) = (-2.96704993, -2.96604993), \qquad (5.3)$$

in the (h, λ)-plane the straight line $\lambda = \tilde{\lambda}$ intersects the curves $\lambda = \lambda_1(h)$, $\lambda = \lambda_3(h)$ at two points respectively. By using the Z_6-equivariance of (5.1), we obtain from Theorem 3.1 the following result.

Theorem 5.1 For the parameter group G_4 and PG_4 and small $\epsilon > 0$, when $\lambda = \tilde{\lambda}$ satisfies (5.3), the system (5.1) has at least 24 limit cycles having the configuration shown in Figure 6. Thus, we have $H(5) \geq 24 = 5^2 - 1$.

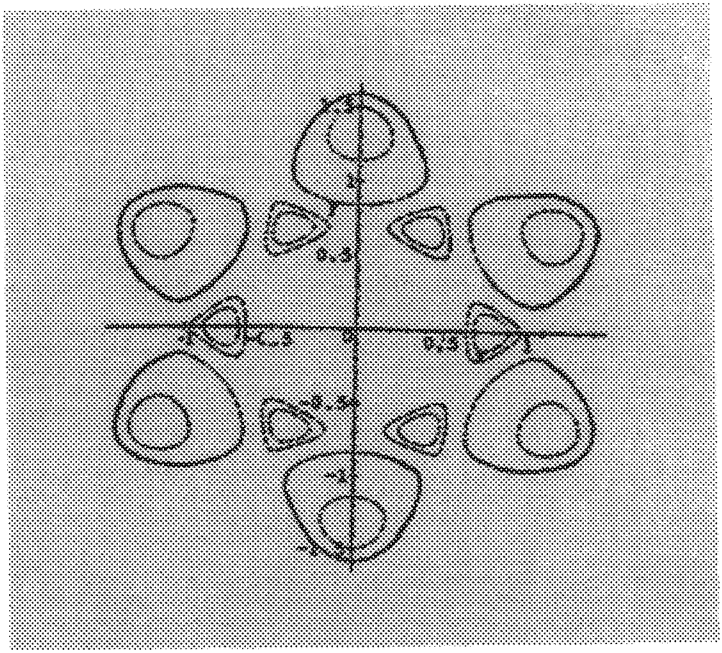


Figure 6: The configurations of 24 limit cycles of System (5.1).

In general, let k be an integer. A Z_{2k+2}-equivariant Hamiltonian system of degree $2k+1$ has the form:

$$\frac{dr}{dt} = r^{2k+1}(a_{k+1}\cos(2k+2)\theta + b_{k+1}\sin(2k+2)\theta),$$

$$\frac{d\theta}{dt} = b_0 + b_1 r^2 + b_2 r^4 + \cdots + b_k r^{2k} + (a_{k+1}\cos(2k+2)\theta - a_{k+1}\sin(2k+2)\theta)r^{2k}. \tag{5.4}$$

By changing the polar axis and time scale, we can reduce (5.4) by two parameters to the following $(k+1)$-parameter system:

$$\frac{dr}{dt} = \beta r^{2k+1}\sin(2k+2)\theta,$$

$$\frac{d\theta}{dt} = 1 + b_1 r^2 + b_2 r^4 + \cdots + b_{k-1} r^{2k-2} + (\alpha + \beta\cos(2k+2)\theta)r^{2k} = \Theta(r,\theta), \tag{5.5}$$

which has the Hamiltonian

$$H(r,\theta) = -\frac{1}{2}r^2 - \frac{1}{4}b_1 r^4 - \cdots - \frac{1}{2k}b_{k-1}r^{2k} - \frac{1}{2k+2}(\alpha + \beta\cos(2k+2)\theta)r^{2k+2}. \tag{5.6}$$

Suppose that $\alpha > \beta > 0$ and the algebraic equation

$$(\alpha+\beta)r^{2k}+b_{k-1}r^{2k-2}+\cdots+b_2r^4+b_1r^2+1=0 \qquad (5.7)$$

and

$$(\alpha-\beta)r^{2k}+b_{k-1}r^{2k-2}+\cdots+b_2r^4+b_1r^2+1=0 \qquad (5.8)$$

have respectively k different positive roots, i.e., in the polar axis $\theta=0$ and $\theta=\pi/(2k+2)$, the system (5.5) has repectively k singular points at

$$(z_1,0),(z_2,0),\cdots,(z_k,0),(z_{k+1},\pi/(2k+2)),(z_{k+2},\pi/(2k+2)),\cdots,(z_{2k},\pi/(2k+2)).$$

By the Z_{2k+2}-equivariance, there exist $(2k+1)^2$ singular points of (5.5). Thus, we can choose a parameter group, such that the system (5.5) has k different period annuluses with Z_{2k+2}-equivariant symmetry. In other words, there are $k(2k+2)+1$ centers of (5.5). We next consider the perturbed system

$$\frac{dr}{dt}=\beta r^{2k+1}\sin(2k+2)\theta-r(p_1r^{2k}+p_2r^{2k-2}+\cdots+p_kr^2-\lambda),$$

$$\frac{d\theta}{dt}=1+b_1r^2+b_2r^4+\cdots+b_{k-1}r^{2k-2}+(\alpha+\beta\cos(2k+2)\theta)r^{2k}. \qquad (5.9)$$

It is similar to the discussion for the system (5.1). By controlling the perturbed parameters $p_i, i=1-k$, such that enclosing every centers (except the origin), there exist two limit cycles created by homoclinic or heteroclinic bifurcations. Therefore, we may obtain $2k(2k+2)=(2k+1)^2-1$ limit cycles of the system (5.9). It means that the conjecture: $H(2k+1)\geq(2k+1)^2-1$ holds.

Acknowledgement The research work is supported by the strategic research grant No.7000934 from the City University of Hong Kong.

Appendix: Ratio Between Double Integrals and Areas: the values of $J_{ij}(h_s)$.

For the fixed unperturbed parameter group $G_2=(\alpha,\beta)=(-7,-0.95)$, we obtain the values of $J_{ij}(h_s)$ by numerical integration to four digits accuracy after the decimal point as follows:

$$J_{01}(h_2)=0.2149252727,\quad J_{02}(h_2)=0.3988931137;$$

$$J_{11}(h_2)=44.17632564,\quad J_{12}(h_2)=5.954572388;$$

$$J_{21}(h_2)=39.28523855,\quad J_{22}(h_2)=5.33645488;$$

$$J_{21}(h_3)=35.92519738,\quad J_{22}(h_3)=4.994696597;$$

$$J_{31}(h_3)=3.259211417,\quad J_{32}(h_3)=1.74133613;$$

$$J_{41}(h_3) = 34.2517894, \quad J_{42}(h_3) = 4.828033959.$$

For the fixed unperturbed parameter group $G_3 = (\alpha, \beta) = (0.2992021277, 2.680851063)$, we have the values of $J_{ij}(h_s)$ by numerical integration to four digits accuracy after the decimal point as follows:

$$J_{01}(h_4) = 0.002025353778, \quad J_{02}(h_4) = 0.03858233638;$$

$$J_{11}(h_3) = 0.3790647646, \quad J_{12}(h_3) = 0.6007796831;$$

$$J_{21}(h_3) = 2.109516141, \quad J_{22}(h_3) = 1.440046333;$$

$$J_{31}(h_4) = 0.5477732675, \quad J_{32}(h_4) = 0.5445506974;$$

$$J_{41}(h_4) = 0.4389746838, \quad J_{42}(h_4) = 0.4436824416.$$

References

1. M.L. Abell and J.P. Braselton, *Maple V: by Example*, AP Professional, Boston, 1994.
2. V.I. Arnold, *Geometric Methods in Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.
3. J. Carr, S.N. Chow and J.K. Hale, Abelian integrals and bifurcation theory, Journal of Differential Equations **59**, 413-417 (1985).
4. H.S.Y. Chan, K.W. Chung and Dongwen Qi, Some bifurcation diagrams for limit cycles of quadratic differential systems, Int. J. Bifurcation and Chaos **11**, 197-206 (2001).
5. H.S.Y. Chan, K.W. Chung and Jibin Li, Bifurcations of limit cycles in a Z_3-equivariant planar vector field of degree 5, Int. J. Bifurcation and Chaos **11**, 2287-2298 (2001).
6. C.J.Christopher and N.G. Lloyd, Polynomial systems: lower bound for the Hilbert numbers, Proc. Royal Soc. London Ser. A **450**, 219-224 (1995).
7. G.B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J., 1963.
8. M. Farkas, *Periodic Motion*, Springer-Verlag, New York, 1994.
9. J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
10. D.A. Gudkov, The topology of real projective algebraic varieties, Russian Math. Surveys **29**, 4:1-79 (1974).

11. D. Hilbert, Mathematical problems, In Proceeding of Symposia in Pure Mathematics **28**, 1-34 (1976).
12. Yu. Ilyashenko, *Finiteness Theorem for Limit Cycles*, American Mathematical Society, Providence, RI, 1991.
13. Li Jibin and Li Cunfu, Planar cubic Hamiltonian systems and distribution of limit cycles of (E_3), Acta. Math.Sinica **28**, 4:509-521 (1985).
14. Li Jibin and Huang Qimin, Bifurcations of limit cycles forming compound eyes in the cubic system, Chinese Ann. of Math. **8B**, 391-403 (1987).
15. J. Li and Liu Zhengrong, Bifurcation set and compound eyes in a perturbed cubic Hamiltonian system, in *Ordinary and Delay Differential Equations*,π (Pitman Research Notes in Math. Series 272, Longman, England, 116-128, 1992).
16. J. Li and Zhao Xiaohua, Rotation symmetry groups of planar Hamiltonian systems, Ann. of Diff. Eqs. **5**, 25-33 (1989).
17. J. Li and Liu Zhenrong, Bifurcation set and limit cycles forming compound eyes in a perturbed Hamiltonian system, Publications Mathmatiques **35**, 487-506 (1991, Spain).
18. J. Li, H.S.Y. Chan and K.W. Chung, Bifurcations of limit cycles in a Z_6-equivariant planar vector field of degree 5, to appear.
19. J. Li, H.S.Y.Chan and K.W.Chung, Investigations of bifurcations of limit cycles in a Z_2-equivariant planar vector field of degree 5, *Int. J. Bifurcation and Chaos*, to appear.
20. N.G. Lloyd, Limit cycles of polynomial systems, in *New Directions in Dynamical Systems*, T. Bedford and J. Swift, eds. 40, (London Mathematical Society Lecture Notes, 192-234, 1988).
21. Luo Dingjun, Wang Xian, Zhu Deming and Han Mouan, *Bifurcation Theory and Methods of Dynamical Systems*, (World Scientific, Singapore, 1997).
22. N.T. Otrokov, On the number of limit cycles of a differential equation in a neighbourhood of a singular point (in Russian), Mat. Sb. **34**, 127-144 (1955).
23. L.M. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1991.
24. L.M. Perko, Limit cycles of quadratic systems in the plane, Rocky Mountain Journal of Math. **14**, 619-645 (1984).
25. V.A. Rokhlin, Complex topological characteristics of real algebraic curves, Russian Math. Surveys **33**, 85-98 (1978).
26. S. Smale, Mathematical problems for the next century, The Mathematical Intelligencer **20**, 2:7-15 (1998).
27. O.Ya. Viro, Progress in topology of real algebraic varieties over the last

six years, Usp. Mat. Nauk. **41**, 3:45-67 (1986).

28. G. Wilson, Hilbert's sixteenth problem, Topology **17**, 53-73 (1978).
29. Ye Yanqian, *Theory of Limit Cycles*, Transl. Math. Monographs 66 (Amer. Math.Soc., Providence, RI, 1986).
30. Ye Yanqian, *Qualitative Theory of Polynomial Differential Systems*, (Modern Mathematics Series, Shanghai Scientific and Technical Publishers, Shanghai, 1995) (in Chinese).
31. Zhang Zhifeng, Ding Tongren, Huang Wenzao and Dong Zhenxi, *Qualitative Theory of Differential Equations*, Transl. Math. Monographs 101. (Amer. Math.Soc., Providence, RI, 1992).

SYSTEMS OF INEQUALITIES AND THE STABILITY OF DECISION MACHINES

JEAN-PIERRE DEDIEU
MIP, Département de Mathématique, Université Paul Sabatier, 31077 Toulouse Cedex 4. France
E-mail: dedieu@@mip.ups-tlse.fr

1 Introduction

In their 1989 paper, Lenore Blum, Mike Shub and Steve Smale introduced a model of computation over the real numbers in order to "[give] a foundation to the theory of modern scientific computation, where most of the algorithms ... are real number algorithms." Real number computations are described by machines over the real numbers. A finite dimensional machine over the real numbers consists of a finite directed connected graph with four types of nodes: an input node, the input is a vector $x \in \mathbf{R}^n$, computation nodes where real polynomials or rational fractions $f(x)$ are computed, branching nodes defined by polynomial inequalities $f(x) \geq 0$, and an output node. This model of computation is called the BSS model. It is described in [5] or in [4].

In real life computations we do not use real numbers but floating point numbers and a finite precision arithmetic. For this reason we do not compute exactly the quantities appearing in our theoretical machine but only approximations. In other words we have two machines: a theoretical machine designed over the real numbers and an approximate machine, designed similarly, which uses floating point numbers and a finite precision arithmetic.

On a certain input $x \in \mathbf{R}^n$, due to the presence of branching nodes, these two machines may have a different behaviour: to the question "is $f(x) \geq 0$" the answer may be "yes" in the exact machine and "no" in the approximate machine. In such a case the computation paths and the outputs may be completely different. This is a typical example of instability introduced by floating point numbers and a finite precision arithmetic.

In this paper we consider *decision machines* i.e. machines where the output is *yes* or *no*. To analyse the stability of such machines we adopt a backward analysis viewpoint. We show that, under a certain hypothesis, the decision taken on $x \in \mathbf{R}^n$ by the approximate machine is identical to the decision taken by the exact machine on a nearby input $y \in \mathbf{R}^n$.

Since the computation path on input x is described by the answers at the branching nodes, our starting point is a system of inequalities

$$f_i(x) \geq 0, \; 1 \leq i \leq m,$$

where $f_i : \mathbf{R}^n \to \mathbf{R}$ is a polynomial or, more generally, an analytic function.

To a vector $a \in \mathbf{R}^m$ we associate its positive part a^+ and its negative part a^-. The i-th coordinate of a^+ (resp. a^-) is a_i (resp. 0) when a_i is nonnegative and 0 (resp. $-a_i$) otherwise so that, as for real numbers, $a = a^+ - a^-$. In our context we use $\|f(x)^-\|$ to measure the deviation of the vector $f(x)$ from positivity.

To begin we show that when $\|f(x)^-\|$ is small enough, there exists a certain $y \in \mathbf{R}^n$ close to x such that the system of inequalities is satisfied exactly at y:

$$f_i(y) \geq 0, \; 1 \leq i \leq m$$

and we give an estimate for the distance of y from x in terms of $\|f(x)^-\|$. Let us denote by σ the sum of the following series:

$$\sigma = \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^{2^k - 1} = 1.63284\ldots$$

To an analytic map $f : \mathbf{R}^n \to \mathbf{R}^m$ and to $x \in \mathbf{R}^n$ we associate the two following numbers:

$$\Gamma(f, x) = \sup_{k \geq 2} \left\| \frac{D^k f(x)}{k!} \right\|^{\frac{1}{k-1}}$$

and

$$\delta(f, x) = \left\| \left(Df(x) Df(x)^* + 4 \operatorname{Diag}\left(f(x)^+\right)\right)^{-1} \right\|^{\frac{1}{2}}.$$

We also let $\delta(f, x) = \infty$ when the matrix $Df(x)Df(x)^* + 4\operatorname{Diag}(f(x)^+)$ is singular. Here A^* denotes the adjoint of the matrix A, $\| \; \|$ is the operator norm associated with the usual Euclidean norms in R^n and R^m and $\operatorname{Diag}(d)$ is the diagonal matrix with diagonal entries $d_1, d_2, \ldots$ Notice that $\Gamma(f, x)$ is always finite because f is analytic.

Theorem 1. *Let $x \in \mathbf{R}^n$ be such that*

$$\|f(x)^-\| \leq \frac{1}{8(1 + \Gamma(f, x))\delta(f, x) \max(1, \delta(f, x))}.$$

Then the set $f^{\geq 0} = \{x \in \mathbf{R}^n \; : \; f(x) \geq 0\}$ is nonempty and

$$\operatorname{Dist}(x, f^{\geq 0}) \leq \sigma \delta(f, x) \|f(x)^-\|.$$

When f is a polynomial system with Degree $f_i \leq 2$ *this result holds when*

$$\|f(x)^-\| \leq \frac{1}{4(2+\|D^2f(x)\|)\delta(f,x)^2}.$$

In the following we consider the case of inequalities and strict inequalities. Like in Theorem 1 we assume $\|f(x)^-\|$ small enough so that $f(y) \geq 0$ for a certain y close to x. Let us assume that $f_i(x) > 0$ for each i in a certain set $J \subset \{1,\dots,m\}$. If the quantities $f_i(x)$, $i \in J$, are far enough from 0 we show that these strict inequalities are also valid at y.

Theorem 2. *Let $x \in \mathbf{R}^n$ be such that*

$$\|f(x)^-\| \leq \frac{1}{8(1+\Gamma(f,x))\delta(f,x)\max(1,\delta(f,x))}.$$

Let us also suppose that

$$f_i(x) > (\sigma\delta(f,x)\|f(x)^-\|)^2$$

for each i in a certain set $J \subset \{1,\dots,m\}$. Then there exists $y \in f^{\geq 0}$ such that

$$f_i(y) > 0 \textit{ for each } i \in J$$

and

$$\|x-y\| \leq \sigma\delta(f,x)\|f(x)^-\|.$$

The following corollary is another formulation of these theorems. We introduce here a computation error ϵ. In an exact machine we branch on the inequality $f_i(x) \geq 0$ while in the approximate machine we have to take into account computation errors: we branch on $f_i(x)+\epsilon_i \geq 0$.

Corollary 3. *Let $x \in \mathbf{R}^n$ and $\epsilon \in \mathbf{R}^m$ be given. Let us denote $\epsilon^{<0}$ the vector in $\mathbf{R}^m$ with $i-th$ component equal to ϵ_i when $f_i(x) < 0$ and* 0 *otherwise. When the three following conditions are satisfied*

- $f_i(x)+\epsilon_i \geq 0$ *for each* $i = 1\dots p$,
- $f_i(x) > (\sigma\delta(f,x)\|\epsilon^{<0}\|)^2$ *for each* $i = p+1\dots m$,
- $\|\epsilon^{<0}\| \leq 1/(8(1+\Gamma(f,x))\delta(f,x)\max(1,\delta(f,x)))$,

then, there exists $y \in \mathbf{R}^n$ such that

- $f_i(y) \geq 0$ *for each* $i = 1\dots p$,
- $f_i(y) > 0$ *for each* $i = p+1\dots m$,
- $\|x-y\| \leq \sigma\delta(f,x)\|\epsilon^{<0}\|$.

2 A word about the proofs

To prove these theorems we use a powerful argument based on Smale's alpha-theory. We associate to the system of inequalities $f(y) \geq 0$ the underdetermined system

$$F_i(y,t) = f_i(y) - t_i^2, \ 1 \leq i \leq m,$$

which is a system of m equations in $m+n$ unknowns: $(y,t) \in \mathbf{R}^n \times \mathbf{R}^m$. We see easily that $F(y,t) = 0$ implies $f(y) \geq 0$. To prove the existence of a zero (y,t) for F with y close to x we show that Newton's sequence $(x_{k+1}, t_{k+1}) = N_F(x_k, t_k)$ starting at $(x_0, t_0) = (x, \sqrt{f(x)^+})$ is converging. Its limit is a zero for F. This process is interesting because it provides a very efficient way to compute y.

Such a method has already been used by F. Cucker and S. Smale in [9] where the authors study the complexity of the feasibility of a system of polynomial equalities and inequalities in n variables.

Newton's method for underdetermined systems of equations was introduced for the first time in 1966 by Ben-Israel [3]. This iteration is defined by

$$N_F(z) = z - DF(z)^\dagger F(z), \quad z_{k+1} = N_F(z_k),$$

where z_0 is given. We denote here by $DF(z)^\dagger$ the Moore-Penrose inverse of the derivative $DF(z)$. When $DF(z)$ is onto and more generally for a surjective linear operator L between two Euclidean spaces, its Moore-Penrose inverse ie given by $L^\dagger = L^*(LL^*)^{-1}$ with L^* the adjoint of L.

To prove the convergence of the sequence $(x_{k+1}, t_{k+1}) = N_F(x_k, t_k)$ we use a theorem due to M. Shub and S. Smale 1996 [19] .

Let $F : \mathbf{E} \to \mathbf{F}$ be an analytic function between two Euclidean spaces. We suppose here that dim $\mathbf{E} \geq$ dim $\mathbf{F}$. To F and a given $z \in \mathbf{E}$ we associate the three following numbers:

$$\alpha(F,z) = \beta(F,z)\gamma(F,z),$$

$$\beta(F,z) = \|DF(z)^\dagger F(z)\|,$$

$$\gamma(F,z) = \sup_{k \geq 2} \left\| DF(z)^\dagger \frac{D^k F(z)}{k!} \right\|^{\frac{1}{k-1}}.$$

Theorem 4. *There is a universal constant α_0, approximately $1/7$, with the following property. For any $z_0 \in \mathbf{E}$ with $\alpha(F, z_0) < \alpha_0$, all the Newton iterates*

$$z_{k+1} = z_k - DF(z_k)^\dagger F(z_k), \ k \geq 0,$$

are defined, converge to $\zeta \in \mathbf{E}$ with $F(\zeta) = 0$ and for all $k \geq 0$

$$||z_{k+1} - z_k|| \leq \left(\frac{1}{2}\right)^{2^k - 1} ||z_1 - z_0||.$$

In particular

$$||\zeta - z_0|| \leq \sigma\beta(F, z_0).$$

3 Comments and similar results

The material contained in this paper is taken from Cucker-Dedieu 1998 [8] and from Dedieu 2000 [10]. In the first paper the authors study the behaviour of round-off decision machines. This talk is inspirated by these ideas. Decision problems and round-off computations are also considered by Cucker and Smale 1999 [9] where the authors design robust algorithms to solve decision problems.

The main ingredient to study such problems is given by the relation between the number $||f(x)^-||$ and the distance to the feasible set $f^{\geq 0}$. Classical results relate these two quantities.

First of all, Hoffman's Theorem, published in 1952 [13], and reconsidered by Güler, Hoffman and Rothblum in 1995 [11]. Hoffman considers linear inequalities:

Theorem 5. *(Hoffman) Let $A \in \mathbf{R}^{m\times n}$. Then there exists a scalar $K(A)$, such that for each $b \in \mathbf{R}^m$ for which the set $A^{\leq b} = \{x' \in \mathbf{R}^n \ : \ Ax' \leq b\}$ is not empty and for each $x \in \mathbf{R}^n$*

$$\mathrm{Dist}\,(x, A^{\leq b}) \leq K(A)||(Ax - b)^+||.$$

There have been a number of generalizations of Hoffman's Theorem to nonlinear cases. A first class of results uses a convexity assumption and is proved via convex analysis: Robinson 1975 [18], Mangasarian 1985 [17], Auslender and Crouzeix 1988 [1] . A recent paper in these directions is "Error Bounds for Convex Inequality Systems" by Lewis and Pang 1998 [15].

Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ be an extended-valued closed proper convex function and $\mathcal{S}$ the closed convex set defined by $f(x) \leq 0$. We denote by $f'(x, d)$ the directional derivative of f at x along a direction d and by $\mathcal{N}(x, \mathcal{S})$ the normal cone of $\mathcal{S}$ at a vector $x \in \mathcal{S}$. With these notations Lewis and Pang prove the following:

Theorem 6. *(Lewis-Pang) The following statements are equivalent:*

- $\mathrm{Dist}\,(x, \mathcal{S}) \leq \gamma f(x)^+$ *for any* $x \in \mathbf{R}^n$*,*

- *For any $x \in f^{-1}(0)$ and $d \in \mathcal{N}(x,S)$*

$$f'(x,d) \geq \gamma^{-1}\|d\|.$$

Another generalization of Hoffman's Theorem to a nonlinear and nonconvex case may be obtained via Lojasiewicz's Inequality. This result was proved for the first time by Lojasiewicz 1964 [16] for semianalytic or semialgebraic sets and functions and then extended to the subanalytic case by Hironaka 1973 [12]. A good exposé of such questions is contained in Bierstone-Milman 1988 [6].

Theorem 7. *(Lojasiewiecz) Let K be a compact and subanalytic set contained in $\mathbf{R}^n$. Let $f : K \to \mathbf{R}^m$ be a continuous subanalytic function. There exist $\alpha > 0$ and an integer $N > 0$ such that for all $x \in K$,*

$$\alpha \ \mathrm{Dist}\,(x, f^{\geq 0})^N \leq \|f(x)^-\|.$$

Hoffman's Theorem has also been extended by Ioffe (1979) [14] to locally Lipschitz functions using Clarke's subgradient, and more recently by Azé, Corvellec and Lucchetti (1999) [2] who consider the case of lower semicontinuous functions defined over Banach spaces. Let us recall the definition of Clarke's subgradient [7]. From Rademacher's Theorem, a function which is Lipschitz on an open subset of $\mathbf{R}^n$ is differentiable almost everywhere on that subset. Based on this result, the generalized gradient at x, denoted $\partial g(x)$ has the following characterization: for any set S of measure zero

$$\partial g(x) = co\{\lim_{i\to\infty} \nabla g(x_i) \mid g \text{ is differentiable at } x_i,\ x_i \notin S,\ x_i \to x\}$$

where co denotes the convex hull.

Theorem 8. *(Ioffe) Let given $f : \mathbf{R}^n \to \mathbf{R}^m$ locally Lipschitz, $x \in \mathbf{R}^n$ with $f(x) \geq 0$ and $\epsilon > 0$. Let us define $c = \min\|y^*\|$ where the minimum is taken for $\|y - x\| \leq \epsilon$ with $y \notin f^{\geq 0}$ and $y^* \in \partial\|f^-\|(y)$. Then*

$$c\,\mathrm{Dist}\,(y, f^{\geq 0}) \leq \|f(y)^-\|$$

for any y with $\mathrm{Dist}\,(y, f^{\geq 0}) \leq \epsilon/2$.

References

1. AUSLENDER A. AND J.-P. CROUZEIX, *Global Regularity Theorems*, Math. Oper. Res., 13 (1988) 243-253.
2. AZÉ D., J.-N. CORVELLEC AND R. F. LUCCHETTI, *Variational Pairs and Applications to Stability in Nonsmooth Analysis*, (1999) Preprint.

3. BEN-ISRAEL A., *A Newton-Raphson Method for the Solution of Systems of Equations*, J. Math. Anal. Appl. 15 (1966) 243-252.
4. BLUM, L., F. CUCKER, M. SHUB, S. SMALE , *Complexity and Real Computation*, (1998) Springer Verlag.
5. BLUM, L., M. SHUB, S. SMALE , *On a Theory of Computation and Complexity over the Real Numbers; NP Completeness, Recursive Functions and Universal Machines*, Bull. Amer. Math. Soc. (New Series) Vol. 21 (1989) 1-46.
6. BIERSTONE E. AND P. MILMAN, *Semianalytic and Subanalytic sets*, IHES Pub. Math. 67 (1988) 5-42.
7. CLARKE F. *Optimization and nonsmooth analysis*, J. Wiley and Sons. New York. 1983.
8. CUCKER, F. AND J.-P. DEDIEU, *Decision Problems and Round-Off Machines*, to appear in: Theory of Computing Systems, 2001.
9. CUCKER, F. AND S. SMALE, *Complexity Estimates Depending on Condition and Round-off Error*, Journal of the ACM, 46 (1999) 113-184.
10. DEDIEU, J.-P., *Approximate Solutions of Analytic Inequality Systems*, SIAM Opt., to appear.
11. GÜLER, O., A. HOFFMAN, U. ROTHBLUM, *On Approximations to Solutions to Systems of Linear Inequalities*, SIAM J. Matrix Anal. Appl., 16 (1995) 688-696.
12. HIRONAKA H., *Introduction to Real-Analytic Sets and Real-Analytic Maps*, Preprint, Pisa (1973).
13. HOFFMAN, A., *On Approximate Solutions of Systems of Linear Inequalities*, J. Res. Nat. Bur. Stand., 49 (1952) 263-265.
14. IOFFE A., *Regular points of Lipschitz functions*, Trans. of the Amer. Math. Soc. 251 (1979) 61-69.
15. LEWIS, A. AND J.-S. PANG, *Error Bounds for Convex Inequality Systems*, in: J.-P. Crouzeix, J.-E. Martinez-Legaz and M. Volle (eds), Generalized Convexity, Generalized Monotonicity, Kluwer (1998) 75-110.
16. LOJASIEWICZ S., *Ensembles Semi-Analytiques.*, IHES Mimeographed Notes (1964).
17. MANGASARIAN O., *A Condition Number for Differentiable Convex Inequalities*, Math. Oper. Res., 10 (1985) 175-179.
18. ROBINSON S., *An Application of Error Bounds for Convex Programming in Linear Spaces*, SIAM Journal of Control and Opt. 13 (1975) 271-273.
19. SHUB, M. AND S. SMALE, *Complexity of Bézout's Theorem IV: Probability of Success, Extensions*, SIAM J. Numer. Anal., 33 (1996) 128-148.

RECONCILIATION OF VARIOUS COMPLEXITY AND CONDITION MEASURES FOR LINEAR PROGRAMMING PROBLEMS AND A GENERALIZATION OF TARDOS' THEOREM

JACKIE C. K. HO

Ph.D. student, Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

LEVENT TUNÇEL

Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

First, we review and clarify the relationships amongst various complexity and condition measures for linear programming problems. Then, we generalize Tardos' Theorem for linear programming problems with integer data to linear programming problems with real number data. Our generalization, in contrast to the only previous such generalization due to Vavasis and Ye, shows that many conventional, polynomial-time (in the sense of the Turing Machine Model, with integer data) primal-dual interior-point algorithms can be adapted in a Tardos' like scheme, to solve linear programming problems with real number data in time polynomial in the dimensions of the coefficient matrix and the logarithms of certain measures of the coefficient matrix (independent of the objective function and the right-hand-side vectors).

Keywords: linear programming, computational complexity, complexity measures, interior-point methods
AMS Subject Classification: 90C05, 90C60, 49K40, 90C31, 90C51

1 Introduction

Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^n$. In this paper, one of our main concerns is the computational complexity of solving linear programming (LP) problems with data (A, b, c) in a way that the number of arithmetic operations is bounded by polynomial functions determined only by A.

For $t \in \mathbb{R}_+$, poly(t) denotes a polynomial function of t. For $\alpha \in \mathbb{Z}$, we define

$$\text{size}(\alpha) := \lceil \log(|\alpha| + 1) \rceil + 1;$$

for $A \in \mathbb{Z}^{m\times n}$,

$$\text{size}(A) := \sum_{i,j} \text{size}(a_{ij}).$$

When $A \in \mathbb{Z}^{m\times n}$, $b \in \mathbb{Z}^m$, $c \in \mathbb{Z}^n$, Tardos [22] proved that the existence of an algorithm for LP which performs only polynomially many elementary arithmetic operations in $\text{size}(A, b, c)$ implies the existence of an algorithm for LP which performs only $\text{poly}(\text{size}(A))$ elementary arithmetic operations. (Her results also apply in the more general case $A \in \mathbb{Q}^{m\times n}$, $b \in \mathbb{Q}^m$, $c \in \mathbb{Q}^n$, also see [23] for network flow problems.)

Tardos' proof is constructive in the sense that it shows how to use *any* polynomial time algorithm for LP as a subroutine to achieve the goal of solving LP problems in $\text{poly}(\text{size}(A))$ time complexity. However, the proof requires calling the subroutine (the LP solver with $\text{poly}(\text{size}(A, b, c))$ time complexity), polynomially many times using modified data so that the sizes of the modified LP instances can be bounded by $\text{poly}(\text{size}(A))$.

Later Vavasis and Ye [29], in another seminal paper (with many new insights), proposed a new kind of interior-point algorithm and proved that their algorithm can solve LP problems with data $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, in $O\left(n^{3.5}\left(\log \bar{\chi}(A) + \log(n)\right)\log\log\bar{\chi}(A)\right)$ interior-point iterations. Also, see Adler and Beling's [1] paper which is more specialized than the Vavasis-Ye paper since it is concerned with the polynomial-time LP algorithms over the algebraic numbers. When specialized to integer (or rational) data, Vavasis-Ye result gives another proof of Tardos' theorem (using $\bar{\chi}(A) = 2^{O(\text{size}(A))}$—see Section 2). So, in this sense, Vavasis-Ye result generalizes Tardos' theorem to LP problems with data $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$. Vavasis-Ye proof is even "more constructive" in the sense that their algorithm is a specialized algorithm designed for such a purpose, and need not be called many times (except to guess an upper bound for $\bar{\chi}(A)$—accounted for in the above quoted iteration bound by the $\log\log\bar{\chi}(A)$ term; also see [15]).

One advantage of Vavasis-Ye algorithm is that it has the potential of becoming a practical algorithm. However, theoretically speaking, Vavasis and Ye left open the question of whether conventional polynomial time interior-point algorithms (or perhaps some others) can be adapted in a scheme more directly related to Tardos' to solve the LP problems with data $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ in polynomially many elementary arithmetic operations where the polynomial bound depends only on the (properly defined) "size" of $A \in \mathbb{R}^{m\times n}$. In fact, Vavasis and Ye [29] state that

"Tardos uses the assumption of integer data in a fairly central way: an *important tool in* [22] *is the operation of rounding down to the nearest integer.* It is not clear how to generalize the rounding operation to noninteger data."

For example, let $A \in \mathbb{Z}^{m\times n}$, $c \in \mathbb{Z}^n$. Then if d is an extreme ray of $\{x \in \mathbb{R}^n : Ax = 0, x \geq 0\}$ such that $c^T d < 0$, then we know that there exists an integral extreme ray d in the above cone such that $c^T d \leq (-1)$. Of course, such arguments do not directly apply in general when the entries of A and c are real numbers. When A and c have only rational entries, the data can be multiplied by a large enough (but not too large) integer such that the new scaled data contain only integers. This again ensures a notion of a "unit" to round to, even after a normalization of the integral d such that $\sum_{j=1}^n d_j = 1$, so that the arguments similar to the above still work (e.g., after such a normalization, $c^T d \leq -1/\Delta(A)$, where $\Delta(A)$ denotes the largest absolute value of a subdeterminant of A). In addition to this, a few other obstacles arise in an attempt to obtain such a generalization of Tardos' theorem and proof to the real number model.

In this paper, we overcome these obstacles, and generalize Tardos' theorem *and* a significant part of her proof to the case when $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$. Our results also generalize Vavasis and Ye's result in the sense that in our scheme almost any polynomial time LP algorithm can be adapted, whereas their result uses a new, specialized algorithm.

Before we describe the generalization of Tardos' theorem, we review and clarify (with many new results) relationships amongst various complexity and condition measures such as $\chi(A)$, $\bar{\chi}(A)$, the condition number of (AA^T) denoted by $\kappa(AA^T)$, Hoffman's bound (or the Lipschitz bound) for systems of linear inequalities, Ye's complexity measure for LP (also known as the smallest large variable bound), $\Delta(A)$ and the smallest nonzero absolute value of a subdeterminant of A, denoted $\delta(A)$. Special emphasis is put on establishing various fundamental properties of $\bar{\chi}$, which becomes one of the central tools in the last section when we deal with generalization of Tardos' result. While our proof of the generalization of her theorem is very similar to hers, a key part of the proof which makes it work in the real number case, is the generalization of the rounding operation to noninteger data (in the sense of choosing an appropriate "unit" for the data at hand). For this, we rely heavily on those fundamental properties of $\bar{\chi}$ mentioned above. We first perform our analysis on deciding the feasibility of a system of inequalities, and then use the resulting algorithm as a subroutine to solve the whole primal-dual LP problem. In both cases, we solve the original problem by solving a sequence of polynomially many "nicer" or smaller LP problems, each of which has integral right hand side vector (and cost vector, in the latter case) whose size is bounded by a polynomial function of our complexity measures. This is one of the fundamental tools for eliminating the dependence on b and c in the overall complexity bound of the algorithms. Solving these "nicer" LP problems gives us important information about the structure of the optimal solutions of the main LP problem in terms of the linear algebraic structures of the input data. For example, "there exists an optimal solution at which the jth inequality is tight" or "at all optimal solutions, the jth inequality is strictly satisfied." Such information helps us reduce the dimensions of the problem at hand; but, it also requires us to analyze the complexity measures for the subproblems.

The sizes of all the integers making up the right hand side and objective vectors of these "nicer" LP problems are bounded above by a polynomial function of n and the logarithm of $\left(\frac{\Delta(A)}{\delta(A)}\right)$. Many are also bounded by a polynomial function of n and $\log \bar{\chi}(A)$.

As mentioned, we need to use an LP solver as a subroutine in our proof of Tardos' theorem. While any polynomial time LP solver can be used, we describe a very useful formulation – the homogeneous self-dual form – in Section 5. The complexity of running an interior-point algorithm (with a certain termination rule) on such a form can be expressed in terms of Ye's complexity measure, which becomes convenient in our complexity analysis.

This paper is organized as follows. In Section 2, we review definitions and characterizations of some complexity measures which are relevant to our stated interest in this paper. We also present some new results in this section. Section 3 includes the Cauchy-Binet formula and an application of it to obtain a bound on the condition number of (AA^T). In Section 4, we discuss Hoffman's Theorem and relate the Hoffman constant to $\chi(A)$. In Section 5, we discuss Ye's complexity measure for LP problems and relate it to the Hoffman constant. Also in Section 5, we show that the number of iterations of many primal-dual interior-point algorithms to solve LP problems with data (A, b, c), with arbitrary A and special b and c, can be bounded by a polynomial function of n and logarithms of certain complexity measures. We review a sensitivity bound result of Cook, Gerards, Schrijver, Tardos [3] in Section 6 and establish various variants of it based on the complexity measures $\chi(A)$ and $\bar{\chi}(A)$. Section 7 contains our main result – a generalization of Tardos' Theorem – based on the results obtained in the preceding sections. We conclude with a very brief discussion of the special cases when A is integral and totally unimodular.

2 Complexity and Condition Measures: χ and $\bar{\chi}$

We denote by $\mathcal{N}(A)$, the null-space of A; $\mathcal{R}(A)$ denotes the range (or column-space) of A. We assume $A \neq 0, n > m \geq 3$. Recall the definitions:

$$\|A\|_p := \max_{\|x\|_p=1} \|Ax\|_p, \text{ for } 1 \leq p \leq \infty,$$

$$\|A\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|^2}.$$

It is not hard to show that

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |A_{ij}|, \tag{1}$$

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |A_{ij}|. \tag{2}$$

We have the following well-known matrix norm inequalities:

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2, \tag{3}$$

$$\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty, \tag{4}$$

$$\frac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1, \tag{5}$$

We also have the submultiplicative property for p-norms, $1 \leq p \leq \infty$. For all $A \in \mathbb{R}^{m\times n}, C \in \mathbb{R}^{n\times q}$, we have

$$\|AC\|_p \leq \|A\|_p\|C\|_p. \tag{6}$$

For the rest of the paper, the 2-norm is assumed when norms are mentioned, unless stated otherwise.

We assume throughout this section that A has full row rank. Define

$$\bar{\chi}(A) := \sup\{\|A^T(ADA^T)^{-1}AD\| : D \in \mathcal{D}\},$$

where $\mathcal{D}$ is the set of all positive definite $n \times n$ diagonal matrices. Note that $\bar{\chi}(RA) = \bar{\chi}(A)$ for all nonsingular $R \in \mathbb{R}^{m\times m}$. In fact, $\bar{\chi}(A)$ depends only on the pair of orthogonal subspaces, $\mathcal{N}(A)$ and $\mathcal{R}(A^T)$. So, it can be defined on subspaces instead. Note that for all $D \in \mathcal{D}$,

$$\|A^T\| = \|A^T(ADA^T)^{-1}ADA^T\| \leq \|A^T(ADA^T)^{-1}AD\| \cdot \|A^T\|.$$

Hence $\|A^T(ADA^T)^{-1}AD\| \geq 1$, and thus $\bar{\chi}(A) \geq 1$.

Similarly we define

$$\chi(A) := \sup\{\|(ADA^T)^{-1}AD\| : D \in \mathcal{D}\}.$$

Note that both $\bar{\chi}(A)$ and $\chi(A)$ are finite. Also,

$$\|A^T(ADA^T)^{-1}AD\| \leq \|A^T\| \cdot \|(ADA^T)^{-1}AD\|,$$
$$\|(ADA^T)^{-1}AD\| \leq \|(AA^T)^{-1}A\| \cdot \|A^T(ADA^T)^{-1}AD\|.$$

Therefore, we have

$$\frac{1}{\|A\|}\bar{\chi}(A) \leq \chi(A) \leq \|(AA^T)^{-1}A\|\bar{\chi}(A) = \frac{\sqrt{\kappa(AA^T)}\bar{\chi}(A)}{\|A\|}, \tag{7}$$

where $\kappa(R) := \|R\| \cdot \|R^{-1}\|$, the condition number of R, for any nonsingular matrix R. Note that if $m = n$, then $\kappa(AA^T) = \|A\|^2 \cdot \|A^{-1}\|^2 = (\|A\|\chi(A))^2$.

An equivalent way to define these parameters is in terms of weighted least squares:

$$\bar{\chi}(A) = \sup\left\{\frac{\| A^T y \|}{\| c \|} : y \text{ minimizes } \| D^{1/2}(A^T y - c) \| \right.$$
$$\left. \text{for some } c \in \mathbb{R}^n, D \in \mathcal{D}\right\},$$
$$\chi(A) = \sup\left\{\frac{\| y \|}{\| c \|} : y \text{ minimizes } \| D^{1/2}(A^T y - c) \| \right.$$
$$\left. \text{for some } c \in \mathbb{R}^n, D \in \mathcal{D}\right\}.$$

Let us define, for $1 \leq \alpha, \beta \leq \infty$,

$$\rho_{\alpha,\beta}(A) := \inf\{\|x - y\|_\beta : x \in X, y \in Y_\alpha\},$$

where $X := \{D\xi : \xi \in \mathcal{N}(A), D \in \text{cl}(\mathcal{D})\}, Y_\alpha := \{\gamma : \gamma \in \mathcal{R}(A^T), \|\gamma\|_\alpha = 1\}$, and $\text{cl}(\mathcal{D})$ denotes the closure of the set $\mathcal{D}$, that is, the set of nonnegative diagonal matrices. Note that $\rho_{\alpha,\beta}(\cdot) > 0$. If we have $\| \cdot \|_\beta \leq c\| \cdot \|_\alpha$, then $\rho_{\alpha,\beta}(\cdot) \leq c$, as $0 \in X$. In particular, $\rho_{\alpha,\alpha}(\cdot) \leq 1$. Also note that the definition of $\rho_{\alpha,\beta}(A)$ depends only on $\mathcal{R}(A^T)$ and its orthogonal complement $\mathcal{N}(A)$. Gonzaga and Lara [8] prove that when $\alpha = \beta = 2$, the subspaces $\mathcal{N}(A)$ and $\mathcal{R}(A^T)$ can be interchanged in the definition of $\rho_{\alpha,\beta}(A)$. In the following, we denote $\rho_{\alpha,\alpha}$ simply by ρ_α. We are mostly interested in ρ_2, which we denote simply by ρ.

All vector p-norms are equivalent, that is, given α, β such that $1 \leq \alpha, \beta \leq \infty$, there exist positive c_1, c_2 such that $c_1\| \cdot \|_\alpha \leq \| \cdot \|_\beta \leq c_2\| \cdot \|_\alpha$. This property also applies to $\rho_{\alpha,\beta}$:

Proposition 2.1 *Suppose $1 \leq \alpha, \beta, \gamma, \delta \leq \infty$, and $c_1, c_2, d_1, d_2 > 0$ such that $d_1\|\cdot\|_\alpha \leq \|\cdot\|_\delta \leq d_2\|\cdot\|_\alpha$ and $c_1\|\cdot\|_\gamma \leq \|\cdot\|_\beta \leq c_2\|\cdot\|_\gamma$. Then*

$$c_1 d_1 \rho_{\delta,\gamma}(\cdot) \leq \rho_{\alpha,\beta}(\cdot) \leq c_2 d_2 \rho_{\delta,\gamma}(\cdot).$$

Proof
Note that $\rho_{\delta,\beta}(A) = \inf\{\|x-y\|_\beta : x \in X, y \in Y_\delta\}$ is attained, by say, $\bar{x}$ and $\bar{y}$. Let $y^* := \bar{y}/\|\bar{y}\|_\alpha$. Then $y^* \in Y_\alpha$, and we have

$$\rho_{\delta,\beta}(A) = \|\bar{x}-\bar{y}\|_\beta = \|\bar{y}\|_\alpha \left\| \frac{\bar{x}}{\|\bar{y}\|_\alpha} - y^* \right\|_\beta \geq \frac{\|\bar{y}\|_\delta}{d_2}\rho_{\alpha,\beta}(A) = \frac{1}{d_2}\rho_{\alpha,\beta}(A).$$

By considering the infimum in $\rho_{\alpha,\beta}(A)$, similarly we have $\rho_{\alpha,\beta}(A) \geq d_1\rho_{\delta,\beta}(A)$. Combining the above, we have $d_1\rho_{\delta,\beta}(A) \leq \rho_{\alpha,\beta}(A) \leq d_2\rho_{\delta,\beta}(A)$. By using $c_1\|\cdot\|_\gamma \leq \|\cdot\|_\beta \leq c_2\|\cdot\|_\gamma$, we have $c_1 d_1 \rho_{\delta,\gamma}(\cdot) \leq \rho_{\alpha,\beta}(\cdot) \leq c_2 d_2 \rho_{\delta,\gamma}(\cdot)$. □

In particular, we have

$$\frac{1}{\sqrt{n}}\rho(\cdot) \leq \rho_\infty(\cdot) \leq \sqrt{n}\rho(\cdot). \tag{8}$$

The following is a well-known fact.

Proposition 2.2 *(Stewart [21])*

$$\bar{\chi}(A) = 1/\rho(A).$$

A *basis* of A is a set of indices $B \subseteq \{1, \ldots, n\}$ such that $|B| = m$ and the columns of A_B are linearly independent. We denote the set of all bases of A by $\mathcal{B}(A)$.

Proposition 2.3 *(Vavasis and Ye [29], Todd, Tunçel and Ye [24])*

$$\bar{\chi}(A) = \max\{\|A_B^{-1}A\| : B \in \mathcal{B}(A)\}.$$

Here, "$\geq$" is proven in [29] and "$\leq$" is proven in [24]. It is known and not hard to show that an analogous characterization for $\chi(A)$ also exists:

$$\chi(A) = \max\{\|A_B^{-1}\| : B \in \mathcal{B}(A)\}. \tag{9}$$

Using the above proposition, we prove that $\bar{\chi}$ cannot increase if any column is removed.

Proposition 2.4 *Suppose $\tilde{A}$ is obtained by removing a column $a \in \mathbb{R}^m$ from $A \in \mathbb{R}^{m\times n}$. We have the following:*

- *If* $\operatorname{rank}(\tilde{A}) = m$*, then* $\bar{\chi}(\tilde{A}) \leq \bar{\chi}(A)$.

- *If* $\operatorname{rank}(\tilde{A}) \leq m-1$, *then let* $\bar{A}$ *be obtained by removing any dependent row from* $\tilde{A}$. *We have* $\operatorname{rank}(\bar{A}) = m-1$ *and* $\bar{\chi}(\bar{A}) = \bar{\chi}(A)$.

Proof
If $\operatorname{rank}(\tilde{A}) = m$, we have

$$\begin{aligned}\bar{\chi}(\tilde{A}) &= \|\tilde{A}_B^{-1}\tilde{A}\|, \text{ for some basis } B \text{ of } \tilde{A}\\ &\leq \|[\tilde{A}_B^{-1}\tilde{A}|\tilde{A}_B^{-1}a]\| = \|\tilde{A}_B^{-1}A\| \leq \bar{\chi}(A).\end{aligned}$$

We used the fact that B is also a basis of A. Now, consider the case where $\operatorname{rank}(\tilde{A}) \leq m-1$. Without loss of generality, assume a is the last column of A. Then by row reduction, there exists a nonsingular $G \in \mathbb{R}^{m\times m}$ such that

$$GA = G[\tilde{A}|a] = \begin{pmatrix} A' & 0 \\ 0^T & 1 \end{pmatrix},$$

for some $A' \in \mathbb{R}^{(m-1)\times(n-1)}$ having full row rank (hence, $\operatorname{rank}(\bar{A}) = m-1$). Then

$$\mathcal{R}(\bar{A}^T) = \mathcal{R}(\tilde{A}^T) = \mathcal{R}((G\tilde{A})^T) = \mathcal{R}(A'^T),$$

and hence $\mathcal{N}(\bar{A}) = \mathcal{N}(A')$. So, $\bar{\chi}(\bar{A}) = \bar{\chi}(A')$. Now, since every basis of GA must include the last column,

$$\begin{aligned}\bar{\chi}(GA) &= \left\| \begin{pmatrix} (A'_B)^{-1} & 0 \\ 0^T & 1 \end{pmatrix} \begin{pmatrix} A' & 0 \\ 0^T & 1 \end{pmatrix} \right\|, \text{ for some basis } B \text{ of } A'\\ &= \left\| \begin{pmatrix} (A'_B)^{-1}A' & 0 \\ 0^T & 1 \end{pmatrix} \right\| = \max\{\| (A'_B)^{-1} A'\|, 1\} \leq \bar{\chi}(A').\end{aligned}$$

The proof of $\bar{\chi}(A') \leq \bar{\chi}(GA)$ is similar. Therefore, we have

$$\bar{\chi}(\bar{A}) = \bar{\chi}(A') = \bar{\chi}(GA) = \bar{\chi}(A).$$

□

Consider $A \in \mathbb{Q}^{m\times n}$. Let L denote the total number of bits required to store A. We have the following.

Proposition 2.5 *(Vavasis and Ye [29])*
If $A \in \mathbb{Q}^{m\times n}$, $\bar{\chi}(A)$ *and* $\chi(A)$ *are both bounded by* $2^{O(L)}$.

Khachiyan [14] proved that approximating $\chi(A)$ within a factor of $2^{\operatorname{poly}(n)}$ is NP-hard. Similarly, approximating $\bar{\chi}(A)$ within a factor of $2^{\operatorname{poly}(n)}$ is also NP-hard [25].

The following observation is due to O'Leary [17]. Naturally, for $\alpha \in \mathbb{R}$, $\operatorname{sign}(\alpha)$ is either $+, 0$, or $-$ depending on the sign of α.

Proposition 2.6 *(O'Leary [17])*
Considering J, γ, ξ as the variables, we have

$$\begin{aligned} \rho_{\alpha,\beta}(A) = \min \quad & \|\gamma_J\|_\beta \\ \text{subject to } & \operatorname{sign}(\gamma_j) = \operatorname{sign}(\xi_j), j \notin J \\ & \|\gamma\|_\alpha = 1, \\ & \gamma \in \mathcal{R}(A^T), \\ & \xi \in \mathcal{N}(A), \\ & J \subseteq \{1, 2, \ldots, n\}, J \neq \emptyset. \end{aligned}$$

Consider the matrix:

$$A_C := \begin{pmatrix} A & 0 \\ C & C \end{pmatrix},$$

where C is an $n \times n$ invertible matrix. Obviously A_C also has full row rank. We have the following result.

Proposition 2.7 *(Ho [11])*

$$\bar{\chi}(A_C) = \sqrt{2}\bar{\chi}(A).$$

Proof
It is easy to see that

$$\mathcal{N}(A_C) = \left\{ \begin{pmatrix} \xi \\ -\xi \end{pmatrix} : \xi \in \mathcal{N}(A) \right\} \text{ and}$$
$$\mathcal{R}(A_C^T) = \left\{ \begin{pmatrix} \gamma + y \\ y \end{pmatrix} : \gamma \in \mathcal{R}(A^T), y \in \Re^n \right\}.$$

We will prove that $\rho(A) = \sqrt{2}\rho(A_C)$ using the characterization of ρ in Proposition 2.6 with $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ interchanged (which we can do since we are working with the 2-norms). Let us denote this minimization problem as $Q(A)$.

1. $\rho(A) \geq \sqrt{2}\rho(A_C)$
   Let (ξ^*, γ^*, J^*) be an optimal solution of $Q(A)$. We now define a y^* that satisfies certain sign conditions. If $j \in J^*$, let y_j^* be such that $\operatorname{sign}(-\xi_j^*) = \operatorname{sign}(y_j^*)$. Therefore $\operatorname{sign}(\xi_j^*) \neq \operatorname{sign}(\gamma_j^* + y_j^*)$. Now if $j \notin J^*$, we can let y_j^* be such that $\operatorname{sign}(-\xi_j^*) = \operatorname{sign}(y_j^*)$, and $\operatorname{sign}(\xi_j^*) = \operatorname{sign}(\gamma_j^* + y_j^*)$, by ensuring $|y_j^*|$ is small enough. Thus, the 3-tuple

   $$\left(\frac{1}{\sqrt{2}} \begin{pmatrix} \xi^* \\ -\xi^* \end{pmatrix}, \begin{pmatrix} \gamma^* + y^* \\ y^* \end{pmatrix}, J^* \right)$$

   is feasible for $Q(A_C)$. Therefore, $\sqrt{2}\rho(A_C) \leq \|\xi_{J^*}^*\| = \rho(A)$.

2. $\rho(A) \leq \sqrt{2}\rho(A_C)$
   Let

$$\left(\begin{pmatrix} \xi^* \\ -\xi^* \end{pmatrix}, \begin{pmatrix} \gamma^* + y^* \\ y^* \end{pmatrix}, J^*\right)$$

be an optimal solution of $Q(A_C)$. Let

$$\hat{J} := \{j \in \{1, \ldots, n\} : \operatorname{sign}(\xi_j^*) \neq \operatorname{sign}(\gamma_j^*)\}.$$

Since ξ^* is orthogonal to γ^*, and $\xi^* \neq 0$, there must exist j such that $\operatorname{sign}(\xi_j^*) \neq \operatorname{sign}(\gamma_j^*)$. Hence $\hat{J} \neq \emptyset$. The 3-tuple $(\sqrt{2}\xi^*, \gamma^*, \hat{J})$ is feasible for $Q(A)$, and therefore $\rho(A) \leq \sqrt{2}\|\xi_{\hat{J}}^*\|$. Now take any $j \in \hat{J}$. Since $\operatorname{sign}(\xi_j^*) \neq \operatorname{sign}(\gamma_j^*)$, there does not exist a y_j which satisfies both $\operatorname{sign}(\xi_j^*) = \operatorname{sign}(\gamma_j^* + y_j)$ and $\operatorname{sign}(-\xi_j^*) = \operatorname{sign}(y_j)$ at the same time. Hence, at least one of j or $n + j$ is in J^*. Therefore we have

$$\frac{1}{\sqrt{2}}\rho(A) \leq \|\xi_{\hat{J}}^*\| \leq \left\| \begin{pmatrix} \xi^* \\ -\xi^* \end{pmatrix}_{J^*} \right\| = \rho(A_C).$$

□

Using a proof similar to the above or using Proposition 2.3, we easily prove the following fact.

Proposition 2.8 *(Ho [11])*

$$\chi\left([A| - A]\right) = \sqrt{2}\chi(A).$$

Recall that the *singular values* of A are the square roots of the eigenvalues of the matrix A^TA. The largest singular value of A is simply $\|A\|_2$. Let $\sigma_{\min}(A)$ denote the smallest nonzero singular value of A. We have the following connection to $\rho(A)$.

Proposition 2.9 *(Stewart [21] and O'Leary [17])*
Let the columns of $U \in \mathbb{R}^{n\times m}$ form an orthonormal basis for $\mathcal{R}(A^T)$. Then

$$\rho(A) = \min_{\emptyset \neq I \subseteq \{1,\ldots,n\}} \sigma_{\min}(U_I),$$

where U_I denotes the submatrix formed from a set I of rows of U.

First, Stewart proved "$\leq$", next O'Leary proved "$\geq$". A nonzero $x \in \mathcal{N}(A)$ (with nonzero entries in positions $\{i_1, \ldots, i_p\} \subseteq \{1, \ldots, n\}$) is said to define a *minimal linear dependence* amongst the columns of A if for every subset I of size at most $(p-1)$ of $\{i_1, \ldots, i_p\}$, the columns of A indexed by I are linearly independent. We have the following proposition due to Vavasis.

Proposition 2.10 *(Vavasis* [28]*)*
Let $x \in \mathcal{N}(A)$ be a nonzero vector defining a minimal linear dependence amongst the columns of A. We have

$$\frac{\min\{|x_j| : x_j \neq 0\}}{\max\{|x_j| : x_j \neq 0\}} \geq \rho(A).$$

We now give a new proof that is different from Vavasis'.

Proof
Let k and l be such that $\min\{|x_j| : x_j \neq 0\} = |x_k| = l$. Let us denote the jth column of A as A_j, for all $j \in \{1, \ldots, n\}$. Then there exists $J \subseteq \{1, \ldots, n\} \backslash \{k\}$ such that $A_J x_J = \pm l A_k$, where x_J contains precisely the nonzero entries of x other than x_k. Since x defines a minimal linear dependence, the columns of A_J must be linearly independent. So we can extend J to a basis B of A to get $A_B x_B = \pm l A_k$. Now,

$$||x||_\infty = ||x_B||_\infty \leq ||x_B|| = l||A_B^{-1}A_k|| \leq l||A_B^{-1}A|| \leq l\bar{\chi}(A).$$

In other words,

$$\bar{\chi}(A) \geq \frac{||x||_\infty}{l},$$

or equivalently,

$$\rho(A) \leq \frac{l}{||x||_\infty} = \frac{\min\{|x_j| : x_j \neq 0\}}{\max\{|x_j| : x_j \neq 0\}}.$$

□

Using these arguments, it is not hard to show that the same result holds for any extreme ray x of the cone $\{x : Ax = 0, x \geq 0\}$.

Corollary 2.11 *Suppose $\{d \in \mathbb{R}^n : Ad = 0, e^T d = 1, d \geq 0\}$ is not empty. Then, it is compact and every extreme point $\bar{d}$ of it has the property*

$$\min\{\bar{d}_j : \bar{d}_j \neq 0\} \geq \frac{\rho(A)}{n}.$$

Proof
Compactness of the set is clear. Every extreme point corresponds to an extreme ray (and hence a basic feasible direction) of $\{x : Ax = 0, x \geq 0\}$. For every basic feasible direction $\bar{x}$, we identify the smallest nonzero component $\bar{x}_k$ first, and then $B \in \mathcal{B}(A)$ such that all other nonzero components of $\bar{x}$ are determined by the system of equations

$$A_B x_B = -\bar{x}_k A_k.$$

Then, as in the proof of Proposition 2.10, we get $||\bar{x}||_\infty \leq \bar{x}_k\bar{\chi}(A)$. Letting $\bar{d} := \bar{x}/(e^T\bar{x})$, we see that

$$\min\{\bar{d}_j : \bar{d}_j \neq 0\} \geq \frac{\bar{x}_k}{\bar{x}_k\,[m\bar{\chi}(A) + 1]} \geq \frac{1}{n\bar{\chi}(A)} = \frac{\rho(A)}{n}.$$

We used the facts that $n \geq (m+1)$ and $\bar{\chi}(A) \geq 1$. □

Recall that $\Delta(A)$ and $\delta(A)$ denote the maximum and minimum (respectively) of the absolute values of the determinants of all the nonsingular square submatrices of A. We have the following relationship among $\rho_\infty, \Delta(A)$ and $\delta(A)$, proven via exploitation of the sign pattern characterization and Cramer's Rule.

Proposition 2.12 *(Tunçel*[27]*)*

$$\rho_\infty(A) \geq \frac{\delta(A)}{m\Delta(A)}.$$

Proof
Recall the definition

$$\rho_\infty(A) := \inf\{||Dx - y||_\infty : D \in \mathrm{cl}(\mathcal{D}), x \in \mathcal{N}(A), y \in \mathcal{R}(A^T), ||y||_\infty = 1\}.$$

Clearly here we can restrict x to be in $\{x \in \mathcal{N}(A) : ||x|| \leq 1\}$. Let $\{(D^k, x^k, y^k)\}$ be a sequence of feasible solutions such that $||D^k x^k - y^k||_\infty$ converges to $\rho_\infty(A)$. Since $\{x^k\}$ and $\{y^k\}$ are in compact feasible sets, we may assume $\{(x^k, y^k)\}$ converges to, say, (x^*, y^*). Let J^* be the set of indices such that the signs of x^* and y^* disagree. Note that $J^* \neq \emptyset$ because otherwise we can choose $D \in \mathrm{cl}(\mathcal{D})$ such that $Dx^* - y^* = 0$, contradicting the fact that $\rho_\infty(A) > 0$. Note that for the pair (x^*, y^*), a best D^* is such that

$$D^*_{ii} = \begin{cases} 0, & i \in J^*, \\ 1, & i \notin J^*, x^*_i = 0, \\ \frac{y^*_i}{x^*_i}, & i \notin J^*, x^*_i \neq 0. \end{cases}$$

So $\rho_\infty(A) = ||y^*_{J^*}||_\infty$, and therefore

$$\rho_\infty(A) = \min\{||y_{J^*}||_\infty : y \in \mathcal{R}(A^T), ||y||_\infty = 1, \mathrm{sign}(y) = \mathrm{sign}(y^*)\}.$$

Then it is easy to see that

$$\begin{aligned} \frac{1}{\rho_\infty(A)} &= \max\{||y||_\infty : y \in \mathcal{R}(A^T), \mathrm{sign}(y) = \mathrm{sign}(y^*), ||y_{J^*}||_\infty \leq 1\} \\ &= \max\{||A^T w||_\infty : \mathrm{sign}(A^T w) = \mathrm{sign}(y^*), ||(A^T w)_{J^*}||_\infty \leq 1\}. \end{aligned}$$

Let w^* be a maximizer of this expression,

$$\epsilon := \min\{|(A^T w^*)_j| : (A^T w^*)_j \neq 0\},$$

$$F(\text{sign}(y^*), J^*) := \left\{ w : \begin{array}{ll} (A^T w)_j \geq \epsilon, & \text{if sign}(y_j^*) = 1, \\ (A^T w)_j = 0, & \text{if sign}(y_j^*) = 0, \\ (A^T w)_j \leq -\epsilon, & \text{if sign}(y_j^*) = -1, \\ (A^T w)_j \leq 1, & \text{if } j \in J^* \end{array} \right\}.$$

Then

$$\begin{aligned} \frac{1}{\rho_\infty(A)} &= \max\{\|A^T w\|_\infty : w \in F(\text{sign}(y^*), J^*)\} \\ &= \max\{a^T w : w \in F(\text{sign}(y^*), J^*)\}, \end{aligned}$$

where a is a column of A (or its negation) such that $a^T w^* = \|A^T w\|_\infty$. Equivalently this is the optimal value of the LP:

$$\begin{array}{rl} (P) \max \eta & \text{subject to} \\ & w \in F(\text{sign}(y^*), J^*), \\ & a^T w - \eta \geq 0. \end{array}$$

Suppose the feasible region of (P) contains a line. So there exist (w, η) and $(d, t) \neq 0$ such that $w + kd \in F(\text{sign}(y^*), J^*)$ and $a^T(w + kd) \geq \eta + kt$, for all $k \in \mathbb{R}$. So $A^T d = 0$. If $d \neq 0$, then it contradicts that fact that A has full row rank. So $d = 0$ and $t \neq 0$. But then $a^T w = a^T(w + kd) \geq \eta + kt$ for all $k \in \mathbb{R}$ also gives a contradiction. So the feasible region of (P) is pointed, and hence contains an optimal basic feasible solution. Let $f(\epsilon)$ be the vector representing the right-hand-side values in the definition of $F(\text{sign}(y^*), J^*)$ (entries of $f(\epsilon)$ are $0,1,\epsilon,-\epsilon$). Then using Cramer's Rule, we have

$$\frac{1}{\rho_\infty(A)} = \left| \frac{\text{subdet}\begin{bmatrix} \begin{pmatrix} A^T \\ A^T_{J^*} \end{pmatrix} & f(\epsilon) \\ a^T & 0 \end{bmatrix}}{\text{subdet}\begin{bmatrix} \begin{pmatrix} A^T \\ A^T_{J^*} \end{pmatrix} & 0 \\ a^T & -1 \end{bmatrix}} \right| \leq \frac{m\Delta(A)}{\delta(A)}.$$

Here, we used that fact that $\epsilon \leq 1$; because $0 \neq \|(A^T w^*)_{J^*}\|_\infty \leq 1$ (as otherwise, it would contradict $\rho_\infty(A) > 0$). □

In fact, the above was originally stated for $A \in \mathbb{Z}^{m \times n}$ in [27], in which case we have $\rho_\infty(A) \geq 1/(m\Delta(A))$.

Proposition 2.5 is a consequence of Proposition 2.12 and (8). Indeed,

$$\bar{\chi}(A) = \frac{1}{\rho(A)} \leq \frac{\sqrt{n}}{\rho_\infty(A)} \leq \sqrt{n}m\frac{\Delta(A)}{\delta(A)}. \tag{10}$$

Therefore, for $A \in \mathbb{Q}^{m\times n}$,

$$\log(\bar{\chi}(A)) \leq \log\left(\frac{\Delta(A)}{\delta(A)}\right) + \log(m) + \frac{1}{2}\log(n) = O(L).$$

Directly utilizing equation (9) and Proposition 2.3, we also bound χ and $\bar{\chi}$ in terms of Δ/δ in the following two propositions.

Proposition 2.13

$$\chi(A) \leq m\frac{\Delta(A)}{\delta(A)}.$$

Proof

Suppose $B \in \mathcal{B}(A)$ maximizes (9). Let y be such that $||y|| = 1$ and $||A_B^{-1}|| = ||A_B^{-1}y||$. Let $x \in \mathbb{R}^m$ such that $A_Bx = y$. Then by Cramer's rule, for each $i \in \{1,\dots,m\}$,

$$|x_i| = \frac{|\text{subdet}([A_B|y])|}{|\det(A_B)|} \leq ||y||_1\frac{\Delta(A)}{\delta(A)} \leq \sqrt{m}\frac{\Delta(A)}{\delta(A)}.$$

So,

$$||A_B^{-1}||^2 = ||x||^2 \leq m^2\frac{\Delta(A)^2}{\delta(A)^2}.$$

Therefore,

$$\chi(A) = ||A_B^{-1}|| \leq m\frac{\Delta(A)}{\delta(A)}.$$

□

Proposition 2.14

$$\bar{\chi}(A) \leq \sqrt{m(n-m)+1}\;\frac{\Delta(A)}{\delta(A)}.$$

Proof

Suppose $B \in \mathcal{B}(A)$ maximizes the expression in Proposition 2.3. Let $\{y^1,\dots,y^{n-m}\}$ be the columns of A that are not in A_B. Let $x^l \in \mathbb{R}^m$ such that $A_Bx^l = y^l$, for all $l \in \{1,\dots,n-m\}$. Then by Cramer's rule, for each $i \in \{1,\dots,m\}$,

$$|x_i^l| = \frac{|\det(C)|}{|\det(A_B)|} \leq \frac{\Delta(A)}{\delta(A)},$$

for some $m \times m$ submatrix C of A. So, denoting the maximum eigenvalue of a matrix by $\lambda_{\max}(\cdot)$, we have

$$\begin{aligned}\|A_B^{-1}A\|^2 &= \|[I|x^1|\cdots|x^{n-m}]\|^2 = \lambda_{\max}[I + x^1(x^1)^T + \cdots + x^{n-m}(x^{n-m})^T] \\ &\leq 1 + (x^1)^T x^1 + \cdots + (x^{n-m})^T x^{n-m} \\ &\leq 1 + m(n-m)\frac{\Delta(A)^2}{\delta(A)^2} \leq [m(n-m)+1]\frac{\Delta(A)^2}{\delta(A)^2}.\end{aligned}$$

Therefore,

$$\bar{\chi}(A) = \|A_B^{-1}A\| \leq \sqrt{m(n-m)+1}\frac{\Delta(A)}{\delta(A)}.$$

□

Facts similar to those given in last three propositions can also be obtained by employing the Cauchy-Binet Formula. This goes back at least to Dikin [4]. (For a historical account and related results, see Forsgren [7] and the references therein.)

3 Cauchy-Binet Formula and the Condition Number of AA^T

Recall that $\mathcal{B}(A)$ denotes the set of all bases of A. We represent each basis B of A as a m-subset of the set of numbers from the natural numbering of the columns of A.

Proposition 3.1 *(Cauchy-Binet Formula)*
Let $A, \tilde{A} \in \mathbb{R}^{m\times n}$ with full row rank. Then

$$\det(A\tilde{A}^T) = \sum_{B\in\mathcal{B}(A)\cap\mathcal{B}(\tilde{A})} \det(A_B)\det(\tilde{A}_B).$$

Using this, we can prove the following relationship among κ, Δ and δ.

Proposition 3.2 *Suppose $A \in \mathbb{R}^{m\times n}$ has full row rank. Then*

$$\kappa(AA^T) \leq m^{3/2}n^{m+1}\frac{\Delta(A)^4}{\delta(A)^2}.$$

Proof
We have $|A_{ij}| \leq \Delta(A)$ for all i, j, and hence by (3),

$$\|AA^T\| = \|A\|^2 \leq \|A\|_F^2 \leq mn\Delta(A)^2.$$

On the other hand,

$$\|(AA^T)^{-1}\| \leq \sqrt{m}\|(AA^T)^{-1}\|_\infty \leq \frac{m^{3/2}\Delta(AA^T)}{\det(AA^T)}.$$

Now, by Proposition 3.1,

$$\begin{aligned}\Delta(AA^T) &= \det(A_{I,*}A^T_{J,*}) \text{ (for some sets } I, J \subseteq \{1,\dots,m\}, |I| = |J|) \\ &= \sum_{B\in\mathcal{B}(A_{I,*})\cap\mathcal{B}(A_{J,*})} \det(A_{I,*B})\det(A_{J,*B}) \\ &\leq \binom{n}{m}\Delta(A)^2 \leq \frac{n^m\Delta(A)^2}{m},\end{aligned}$$

and $\det(AA^T) = \sum_{B\in\mathcal{B}(A)} \det(A_B)^2 \geq \delta(A)^2$. Therefore,

$$\kappa(AA^T) = \|AA^T\| \cdot \|(AA^T)^{-1}\| \leq \frac{m^{3/2}n^{m+1}\Delta(A)^4}{\delta(A)^2}.$$

□

4 Hoffman's Bound and χ

For a vector $u \in \mathbb{R}^n$, let pos$(u) \in \mathbb{R}^n$ be such that $(\text{pos}(u))_j := \max\{u_j, 0\}$ for each $j \in \{1,\dots,n\}$. The following result gives an upper bound on the distance of a point to a polyhedron, in terms of its violation of the constraints defining the polyhedron.

Theorem 4.1 *(Hoffman [12])*
Let $A \in \mathbb{R}^{m\times n}$ (not necessarily full row rank) and let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on $\mathbb{R}^m$ and on $\mathbb{R}^n$, respectively. Then there exists a scalar $K_{\alpha,\beta}(A)$, such that for every $c \in \mathbb{R}^n$ for which the set $\{y \in \mathbb{R}^m : A^Ty \leq c\} \neq \emptyset$, and for every $y' \in \mathbb{R}^m$,

$$\min_{y:A^Ty\leq c} \| y - y' \|_\alpha \leq K_{\alpha,\beta}(A) \| pos(A^Ty' - c) \|_\beta .$$

The coefficient $K_{\alpha,\beta}(A)$ is sometimes called a *Lipschitz bound* of A. For a norm $\|\cdot\|$ on $\mathbb{R}^n$, let $\|\cdot\|^*$ be the *dual norm* defined by

$$\| v \|^* := \max\{v^Tx : x \in \mathbb{R}^n, \| x \| \leq 1\},$$

for each $v \in \mathbb{R}^n$. Note that for p-norms $(1 \leq p \leq \infty)$, we have $\|\cdot\|_p^* = \|\cdot\|_q$, where q is such that $p^{-1} + q^{-1} = 1$. In particular, $\|\cdot\|_2^* = \|\cdot\|_2$. Let ext$(S)$ denote the set of extreme points of a set S. We have the following geometric representation of the Lipschitz bound.

Proposition 4.2 *(Güler, Hoffman and Rothblum [9])*
Theorem 4.1 holds with $K_{\alpha,\beta}(A) := \max\{\| v \|_\beta^* : v \in \text{ext}(V_\alpha(A))\}$, where $V_\alpha(A) := \{v \in \mathbb{R}^n : v \geq 0, \| Av \|_\alpha^* \leq 1\}$.

We write $K_2(A) := K_{2,2}(A)$ for all A. There is also a representation of the Lipschitz bound via singular values. For any $E \in R^{n\times m}$. Let $U(E)$ be the set of subsets of $\{1,\dots,n\}$ for which the corresponding rows of E are linearly independent. Let $U^*(E)$ be the maximal elements in $U(E)$.

Proposition 4.3 *(Güler, Hoffman and Rothblum [9])*

$$K_2(A) \leq \max_{J\in U^*(A^T)} \frac{1}{\sigma_{\min}(A_J^T)}.$$

Note that $\min_{J\in U^*(A^T)} \sigma_{\min}(A_J^T) = \min_{\emptyset\neq J\subseteq\{1,\dots,n\}} \sigma_{\min}(A_J^T)$. To prove this, first note that "$\geq$" is clear. Take $A \in \mathbb{R}^{m\times n}$ with rank, say, r. Take any nonempty $J \subseteq \{1,\dots,n\}$. Let $\sigma_i(E)$ denote the ith largest singular value of any matrix E, and $k := \text{rank}(A_J^T)$. Then $\sigma_{\min}(A_J^T) = \sigma_k(A_J^T)$. Let $I \subseteq J$ be such that $\text{rank}(A_I^T) = k = |I|$. Then by the interlacing property of singular values, $\sigma_k(A_I^T) \leq \sigma_k(A_J^T)$. Let $M \in U^*(A^T)$ be such that $I \subseteq M$. Then

$$\sigma_{\min}(A_M^T) = \sigma_r(A_M^T) \leq \sigma_k(A_I^T) \leq \sigma_{\min}(A_J^T),$$

where we used the interlacing property again in the first inequality above. Therefore,

$$\max_{J\in U^*(A^T)} \frac{1}{\sigma_{\min}(A_J^T)} = \max_{\emptyset\neq J\subseteq\{1,\dots,n\}} \frac{1}{\sigma_{\min}(A_J^T)}.$$

The next proposition gives a connection between K_2 and $\bar{\chi}$ via singular values.

Proposition 4.4 *Suppose $A \in \mathbb{R}^{m\times n}$ has full row rank. Then*

$$\|A\|K_2(A) \leq \bar{\chi}(A).$$

Proof

Consider the singular value decomposition of A. Let $A = UDV^T$, where $U \in \mathbb{R}^{m\times m}$ is orthogonal, $D \in \mathbb{R}^{m\times n}$ is diagonal (with singular values $\sigma_1,\dots,\sigma_m$ of A on the diagonal, in that order), and $V \in \mathbb{R}^{n\times n}$ is orthogonal as well. Suppose $V = [v_1|\cdots|v_n]$, i.e., $\{v_1,\dots,v_n\}$ are the columns of V. Let $\bar{V} := [v_1|\cdots|v_m]$ and $\Sigma := \text{Diag}(\sigma_1,\dots,\sigma_m)$. Then $A = U\Sigma\bar{V}^T$. Since A has full row rank, $\sigma_1,\dots,\sigma_m > 0$, and hence Σ is invertible. We have $A^T = \bar{V}\Sigma U^T$, and $\bar{V} = A^T U\Sigma^{-1}$. So $\mathcal{R}(A^T) = \mathcal{R}(\bar{V})$, and $\bar{V}$ has orthonormal columns. By Propositions 2.9 and 4.3,

$$K_2(\bar{V}^T) \leq \max_{I\in U^*(\bar{V})} \frac{1}{\sigma_{\min}(\bar{V}_I)} = \max_{\emptyset\neq I\subseteq\{1,\dots,n\}} \frac{1}{\sigma_{\min}(\bar{V}_I)} = \bar{\chi}(A).$$

Now it remains to show $\|A\|K_2(A) \leq K_2(\bar{V}^T)$. Note that

$$\|A^T y\|^2 = y^T U\Sigma\bar{V}^T\bar{V}\Sigma U^T y = \|\Sigma U^T y\|^2.$$

Therefore,

$$\|A\| = \|A^T\| = \max_{\|y\|=1} \|A^T y\| = \|\Sigma U^T\| = \|U\Sigma\|.$$

Now we consider the relationship between $K_2(A)$ and $K_2(\bar{V}^T)$. Suppose $K_2(A) = \|\hat{v}\|$, where $\hat{v}$ is an extreme point of $V_2(A)$. Let $\bar{v} := \|A\|\hat{v}$. We will prove that $\bar{v}$ is an extreme point of $V_2(\bar{V}^T)$. Suppose $\bar{v} = \lambda w + (1-\lambda)z$, where $\lambda \in (0,1)$, and $w, z \in V_2(\bar{V}^T)$. Then

$$\hat{v} = \lambda \frac{w}{\|A\|} + (1-\lambda)\frac{z}{\|A\|}.$$

Since $w \in V_2(\bar{V}^T)$, $w \geq 0$ and therefore $w/\|A\| \geq 0$. Also,

$$\left\| A\left(\frac{w}{\|A\|}\right) \right\| = \frac{1}{\|A\|}\|U\Sigma\Sigma^{-1}U^T A w\| \leq \frac{1}{\|A\|}\|U\Sigma\|\|\bar{V}^T w\| \leq 1.$$

So $w/\|A\| \in V_2(A)$, and similarly so does $z/\|A\|$. Therefore, $w = z$, implying that $\bar{v}$ is an extreme point of $V_2(\bar{V}^T)$. Now,

$$\|A\| K_2(A) = \|A\|\|\hat{v}\| = \|\bar{v}\| \leq K_2(\bar{V}^T) \leq \bar{\chi}(A).$$

□

As a corollary, since $\bar{\chi}(A) \leq \|A\|\chi(A)$, we have $K_2(A) \leq \chi(A)$. During the review of our paper, we became aware of [33]. Note that the relation $K_2(A) \leq \chi(A)$ implies Theorem 3.6 from [33] which states that Theorem 4.1 holds with $K_{\alpha,\beta}(A)$ replaced by $\chi(A)$, when $\alpha = \beta = 2$ and A has full row rank. Also Lemmas 3.3, 3.4 and 3.5 of [33] follow from equation (9) and the fact that whenever $\{x : Ax = b, x \geq 0\}$ is nonempty, it contains a basic feasible solution.

We also note that, by Proposition 2.12, we have

$$K_2(A) \leq \frac{m\Delta(A)}{\|A\|\delta(A)}.$$

Let $\mathcal{G}$ be the set of diagonal matrices in $\mathbb{R}^{n\times n}$ with diagonal entries from $\{1,-1\}$. Take $G \in \mathcal{G}$. Then $\|AG\| = \|A\|$. Also for any diagonal matrix $D \in \mathbb{R}^{n\times n}$, $\|(AG)^T(AGD(AG)^T)^{-1}AGD\| = \|A^T(ADA^T)^{-1}AD\|$, and hence $\bar{\chi}(AG) = \bar{\chi}(A)$. (Similarly, $\chi(AG) = \chi(A)$.) Therefore, we have

$$\max_{G\in\mathcal{G}} K_2(AG) \leq \frac{\bar{\chi}(A)}{\|A\|} \leq \chi(A). \qquad (11)$$

Also, $\Delta(AG) = \Delta(A)$ and $\delta(AG) = \delta(A)$. So we also have

$$\max_{G\in\mathcal{G}} K_2(AG) \leq \frac{m\Delta(A)}{\|A\|\delta(A)}.$$

We now characterize the extreme points of $V_1(A)$. Recall that

$$V_1(A) = \left\{ v \in \mathbb{R}^n : \begin{pmatrix} A \\ -A \\ -I \end{pmatrix} v \leq \begin{pmatrix} e \\ e \\ 0 \end{pmatrix} \right\},$$

which is a polyhedron, and the constraint matrix in the above description has full column rank. Let $J \subseteq \{1, \ldots, n\}$ such that $|J| \leq m$. Then we pick $I_1, I_2 \subseteq \{1, \ldots, m\}$ such that $I_1 \cap I_2 = \emptyset$ and $|I_1| + |I_2| = |J|$. Assume that the matrix

$$\begin{pmatrix} A_{I_1,J} \\ -A_{I_2,J} \end{pmatrix}$$

is nonsingular. Here $A_{I_1,J}$ denotes the submatrix of A with rows indexed by I_1 and columns indexed by J. Let $x \in \mathbb{R}^n$ be such that $x_j = 0$ if $j \notin J$ and

$$\begin{pmatrix} A_{I_1,J} \\ -A_{I_2,J} \end{pmatrix} x_J = \begin{pmatrix} e \\ e \end{pmatrix}.$$

If $x \in V_1(A)$, then x is an extreme point of $V_1(A)$. Vice versa, any given $x \in \text{ext}(V_1(A))$ must satisfy the above for some J, I_1 and I_2. So using Cramer's rule, for each $j \in J$,

$$x_j = \frac{\left|\text{subdet}\begin{pmatrix} A_{I_1,J} \; e \\ -A_{I_2,J} \; e \end{pmatrix}\right|}{\left|\det\begin{pmatrix} A_{I_1,J} \\ -A_{I_2,J} \end{pmatrix}\right|} \leq \frac{|J|\Delta(A)}{\delta(A)} \leq \frac{m\Delta(A)}{\delta(A)}.$$

So,

$$K_1(A) = \|x\|_\infty \leq \frac{m\Delta(A)}{\delta(A)}, \text{ and } K_{1,\infty}(A) = \|x\|_1 \leq \frac{m^2\Delta(A)}{\delta(A)}.$$

Therefore, we have

$$\max_{G \in \mathcal{G}} K_1(AG) \leq \frac{m\Delta(A)}{\delta(A)}, \text{ and } \max_{G \in \mathcal{G}} K_{1,\infty}(AG) \leq \frac{m^2\Delta(A)}{\delta(A)}.$$

In fact, the extreme points of $V_1(AG)$ can be characterized in a similar way. The only difference is that we require x to satisfy the sign pattern given by G, instead of $x \geq 0$. Now, we give another proof of the implication $K_2(AG) \leq \chi(A)$, $\forall G \in \mathcal{G}$ of (11). We use the following characterization of $K_2(A)$ for this purpose.

Lemma 4.5

$$K_2(A) = \max\left\{\|A_B^{-1}A\gamma\| : \gamma \in \mathcal{R}(A^T), B \in \mathcal{B}(A), \|A\gamma\| = 1, \gamma_B \leq -A_B^{-1}A_N\gamma_N\right\}.$$

Proof
Note that

$$K_2(A) = \max\left\{||v|| : \; v \in \operatorname{ext}(V_2(A)) \right.$$
$$\left. \cap\{\xi - \gamma : \; \xi \geq \gamma, \xi \in \mathcal{N}(A), \gamma \in \mathcal{R}(A^T), ||A\gamma|| = 1\}\right\}.$$

Let $\gamma \in \mathcal{R}(A^T)$ such that $||A\gamma|| = 1$; also let $B \in \mathcal{B}(A)$ such that $\gamma_B \leq -A_B^{-1}A_N\gamma_N$ and $||A_B^{-1}A\gamma||$ is equal to the maximum value in the statement of the lemma. Define $\xi \in \mathbb{R}^n$ as follows. $\xi_N := \gamma_N$, $\xi_B := -A_B^{-1}A_N\gamma_N$. Thus, we have $\xi \in \mathcal{N}(A)$, $\xi \geq \gamma$. Next, we claim $v := (\xi - \gamma) \in \operatorname{ext}(V_2(A))$. Suppose not. Then, there exist $v^{(1)}, v^{(2)} \in V_2(A)$ such that $\frac{1}{2}\left(v^{(1)} + v^{(2)}\right) = \xi - \gamma$, $v^{(1)} \neq v^{(2)}$. We immediately have $v_N^{(1)} = v_N^{(2)} = 0$. Thus,

$$1 = ||A_B v_B|| \leq \frac{1}{2}||A_B v_B^{(1)}|| + \frac{1}{2}||A_B v_B^{(2)}|| \leq 1$$

which implies

$$||A_B v_B|| = ||A_B v_B^{(1)}|| = ||A_B v_B^{(2)}|| = 1.$$

Therefore (since $v_B = \frac{1}{2}v_B^{(1)} + \frac{1}{2}v_B^{(2)}$), by the characterization of the equality case in the Cauchy-Schwarz inequality, we must have $A_B v_B = A_B v_B^{(1)} = A_B v_B^{(2)}$. Since $v_B^{(1)} \neq v_B^{(2)}$, A_B must be singular, we arrived at a contradiction. In addition to $(\xi - \gamma)$ being an extreme point of $V_2(A)$, we have

$$||\xi - \gamma|| = ||A_B^{-1}A_N\gamma_N + \gamma_B|| = ||A_B^{-1}A\gamma||.$$

Therefore,

$$K_2(A) \geq \max\left\{||A_B^{-1}A\gamma|| : \; \gamma \in \mathcal{R}(A^T), B \in \mathcal{B}(A), ||A\gamma|| = 1, \right.$$
$$\left. \gamma_B \leq -A_B^{-1}A_N\gamma_N\right\},$$

as desired.
To prove the reversed inequality, we let $\xi \in \mathcal{N}(A)$, $\gamma \in \mathcal{R}(A^T)$ such that $||A\gamma|| = 1$, $\xi \geq \gamma$, $(\xi - \gamma) \in \operatorname{ext}(V_2(A))$ and $||\xi - \gamma|| = K_2(A)$. Let $J \subset \{1, 2, \ldots, n\}$ be such that $\xi_J = \gamma_J$ and $\xi_{\bar{J}} > \gamma_{\bar{J}}$. Then since $(\xi - \gamma)$ is in $\operatorname{ext}(V_2(A))$, we must have $\operatorname{rank}(A_{\bar{J}}) = |\bar{J}| \leq m$ (otherwise, we can find $\bar{\xi} \in \mathcal{N}(A_{\bar{J}})\backslash\{0\}$ such that

$$\tilde{\xi}_j := \begin{cases} 0 & \text{if } j \in J \\ \bar{\xi}_j & \text{if } j \notin J; \end{cases}$$

now $\tilde{\xi} \in \mathcal{N}(A)$ and for small enough $\epsilon > 0$, $(\xi + \epsilon\tilde{\xi} - \gamma)$ and $(\xi - \epsilon\tilde{\xi} - \gamma)$ $\in V_2(A)$, a contradiction). Complete $\bar{J}$ to a basis B of A. Then $\xi_N = \gamma_N$ and

$A_B\xi_B = -A_N\gamma_N$. The latter implies $\xi_B = -A_B^{-1}A_N\gamma_N$. Thus,

$$K_2(A) = \|\xi - \gamma\| = \|\xi_B - \gamma_B\| = \|A_B^{-1}A_N\gamma_N + \gamma_B\| = \|A_B^{-1}A\gamma\|.$$

Hence yielding the desired inequality

$$K_2(A) \leq \max\{\|A_B^{-1}A\gamma\|: \ \gamma \in \mathcal{R}(A^T), B \in \mathcal{B}(A), \|A\gamma\| = 1, \\ \gamma_B \leq -A_B^{-1}A_N\gamma_N\}.$$

□

Theorem 4.6

$$\max_{G\in\mathcal{G}} K_2(AG) \leq \chi(A).$$

Proof

By Lemma 4.5,

$K_2(A) = \max\{\|A_B^{-1}A\gamma\| : \gamma \in \mathcal{R}(A^T), B \in \mathcal{B}(A), \|A\gamma\| = 1, \gamma_B \leq -A_B^{-1}A_N\gamma_N\}$.

So,

$$\begin{aligned}\max_{G\in\mathcal{G}} K(AG) &\leq \max\{\|A_B^{-1}AG\gamma\| : \gamma \in \mathcal{R}(GA^T), B \in \mathcal{B}(AG), \|AG\gamma\| = 1, G \in \mathcal{G}\}\\ &= \max\{\|A_B^{-1}AG\gamma\| : G\gamma \in \mathcal{R}(A^T), B \in \mathcal{B}(A), \|AG\gamma\| = 1, G \in \mathcal{G}\}\\ &= \max\{\|A_B^{-1}A\gamma\| : \gamma \in \mathcal{R}(A^T), B \in \mathcal{B}(A), \|A\gamma\| = 1\}\\ &= \max\{\|A_B^{-1}Ax\| : x \in \mathbb{R}^n, B \in \mathcal{B}(A), \|Ax\| = 1\}\\ &= \max\{\|A_B^{-1}y\| : B \in \mathcal{B}(A), \|y\| = 1\}\\ &= \max\{\|A_B^{-1}\| : B \in \mathcal{B}(A)\} = \chi(A).\end{aligned}$$

□

We note that the inequality above may be strict. Otherwise, using (11) we would have had $\bar{\chi}(A) = \|A\|\chi(A)$ which is clearly false in general —take for instance $A := \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$.

5 Ye's Complexity Measure for LP and Hoffman's Bound

We are going to look at two more complexity measures, η and symm. These complexity measures relate closely to the symmetry of certain geometric objects of the LP problem. Let us consider the LP problem in the following primal form:

$$\begin{array}{lll}(P) & \min & c^Tx \\ & \text{subject to} & Ax = b \\ & & x \in \mathbb{R}^n_+,\end{array}$$

and the corresponding dual form:

$$\begin{array}{lrl} (D)\ \max & b^T y & \\ \text{subject to} & A^T y + s & = c \\ & y & \in \mathbb{R}^m \\ & s & \in \mathbb{R}^n_+ \end{array}$$

where $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$.

Under the assumption that both (P) and (D) have feasible solutions, Ye [31] first defines a complexity measure for each of the problems (P) and (D):

$$\eta_P := \min_{j \in B} \max_{x \in \mathrm{opt}(P)} x_j,$$

$$\eta_D := \min_{j \in N} \max_{s \in \mathrm{opt}(D)} s_j.$$

Then, Ye [31] defines the complexity measure of the primal dual pair as the minimum of the two:

$$\eta(P, D) := \min\{\eta_P, \eta_D\},$$

where $\mathrm{opt}(P)$ and $\mathrm{opt}(D)$ denote the sets of optimal solutions of (P) and (D) respectively, and (B, N) denotes the strict complementarity partition.

Let us study these measures for feasibility problems over polyhedra expressed in Karmarkar's ([13]) standard form:

$$\mathcal{P} := \{x : Ax = 0, e^T x = 1, x \geq 0\}.$$

(This form is relevant in Subsection 5.1 as well.) We assume A to have full row rank and no zero columns because, without loss of generality, we can always eliminate the variables that correspond to zero columns in A. Let $S := \mathcal{N}(A)$ and (hence) $S^\perp = \mathcal{R}(A^T)$. (P) and its dual can now be written as a primal-dual pair of feasibility problems. See Vavasis and Ye [30] and [26].

$$\begin{array}{lrlr} (FP) & x \in S, & (FD) & s \in S^\perp, \\ & \|x\|_1 = 1, & & \|s\|_1 = 1, \\ & x \geq 0. & & s \geq 0. \end{array}$$

(FD) is the dual of (FP) in the sense that every feasible solution to the dual problem of maximizing 0 over the constraints defined by (FP), corresponds to a feasible solution of (FD), except for $s = 0$ which does not correspond to a feasible solution in (FD). In this setting, even though (FP) is always bounded, (FD) can still be infeasible (for example, $A := [1, -1]$).

When (FP) is feasible, there exists a pair (x, s) such that $x \in S, x_N = 0, x_B > 0, s \in S^\perp, s_N > 0, s_B = 0$, where $[B, N]$ is the corresponding strict complementarity partition with B nonempty. Furthermore, all feasible solutions of (FP) and (FD) must satisfy $x_N = 0$ and $s_B = 0$. We allow B or N to be empty. The condition $B \neq \emptyset$ is equivalent to (FP) being feasible. Similarly, $N \neq \emptyset$ is equivalent to (FD) being feasible.

Since the problems (FP) and (FD) are written in terms of the subspaces S and $S^\perp$, let us redefine Ye's measures accordingly. For any subspace C, define $C(1) := \{x \in C : \|x\|_1 = 1\}$. Let

$$\begin{aligned}
\eta(S) &:= \min_{j \in B} \max_{x \in S(1), x \geq 0} x_j, \\
\eta(S^\perp) &:= \min_{j \in N} \max_{s \in S^\perp(1), s \geq 0} s_j, \\
\eta(A) &:= \min\{\eta(S), \eta(S^\perp)\}.
\end{aligned}$$

We define $\eta(S)$ to be 1, when (FP) is infeasible (similarly, $\eta(S^\perp)$ is 1 if (FD) is infeasible). Notice that all of $\eta(S), \eta(S^\perp)$ and $\eta(A)$ are positive for all A. $\eta(S)$ measures some kind of symmetry of the columns vectors of A_B about the origin. The set $\{A_B x_B : \|x_B\|_1 = 1, x_B \geq 0\}$ is the set of all convex combinations of the columns of A_B. Therefore $\{x \in S(1) : x \geq 0\}$ corresponds to the coefficients when 0 is written as convex combinations of the columns of A_B, and hence η measures their sizes. If the columns of A_B are perfectly symmetric about the origin, $\eta(S)$ would be 1/2. And if the columns are highly asymmetric, $\eta(S)$ would be much smaller than 1/2.

The following results give dual descriptions for $\eta(S)$ and $\eta(S^\perp)$. For $v \in \mathbb{R}^n, J \in \{1, 2, \ldots, n\}$, let

$$v_J^+ := \begin{cases} -\infty & \text{if } v_J \leq 0, \\ \max_{j \in J} v_j & \text{otherwise,} \end{cases}$$

$$v_J^- := \begin{cases} +\infty & \text{if } v_J \geq 0, \\ \min_{j \in J} v_j & \text{otherwise.} \end{cases}$$

Proposition 5.1 *(Tunçel*[26]*)*
Suppose $\{e_j : j \in \{1, 2, \ldots, n\}\} \cap S = \emptyset$ and $B \neq \emptyset$. Then

$$\eta(S) = \min\{\gamma_B^+ : \gamma \in S^\perp, 0 < \gamma_B^+ < 1, \gamma_B^+ - \gamma_B^- = 1\}.$$

Proposition 5.2 *(Tunçel*[26]*)*
Suppose $\{e_j : j \in \{1, 2, \ldots, n\}\} \cap S^\perp = \emptyset$ and $N \neq \emptyset$. Then

$$\eta(S^\perp) = \min\{\xi_N^+ : \xi \in S, 0 < \xi_N^+ < 1, \xi_N^+ - \xi_N^- = 1\}.$$

Note that under our assumptions, we always have $\{e_j : j \in \{1,2,\ldots,n\}\} \cap S = \emptyset$ because A has no zero columns. Also, the condition $B \neq \emptyset$ is equivalent to (FP) being feasible. Similarly, $N \neq \emptyset$ is equivalent to (FD) being feasible.

Recently, Epelman [5], Epelman and Freund [6] presented another complexity measure based on A. Let $\mathcal{H}(A_B) := \{A_B x_B : x_B \geq 0, \|x_B\|_1 = 1\}$. That is, $\mathcal{H}(A_B)$ is the convex hull of the column vectors of A_B. Let

$$\operatorname{symm}(A) := \max\{t : -tv \in \mathcal{H}(A_B) \text{ for all } v \in \mathcal{H}(A_B)\}.$$

Note that a generalized version of this measure has been used before by Renegar [18] to estimate complexity for convex optimization problems.

It is clear that $\operatorname{symm}(A)$ measures precisely the degree of symmetry of $\mathcal{H}(A_B)$ about the origin in $\mathbb{R}^m$ $(A \in \mathbb{R}^{m \times n})$. When $\mathcal{H}(A_B)$ is centrally symmetric (about the origin), $\operatorname{symm}(A) = 1$.

Proposition 5.3 *(Epelman [5], Epelman and Freund [6])*

$$\frac{\operatorname{symm}(A)}{1 + \operatorname{symm}(A)} = \eta(S).$$

The above proposition gives an explicit relation between the two complexity measures, $\eta(S)$ and $\operatorname{symm}(A)$. Since the function $x/(1+x)$ is strictly increasing on $(0,1]$, $\eta(S)$ also measures the degree of symmetry of $\mathcal{H}(A_B)$ about the origin. In fact, by combining Proposition 5.1 and Proposition 5.3, we get the following.

Corollary 5.4 *(Ho [11])*

$$\operatorname{symm}(A) = \min_{\gamma \in S^\perp, \|\gamma_B\|=1} -\frac{\gamma_B^+}{\gamma_B^-}.$$

We can state similar results for $\eta(S^\perp)$. Let us define $H \in \mathbb{R}^{(n-m)\times n}$ to be a full row rank matrix obtained by deleting linearly dependent rows from $P_A := I - A^T(AA^T)^{-1}A$.

Corollary 5.5 *(Ho [11]) Suppose (FD) is feasible and $\{e_j : j \in \{1,2,\ldots,n\}\} \cap S^\perp = \emptyset$. Then*

$$\frac{\operatorname{symm}(H)}{1 + \operatorname{symm}(H)} = \eta(S^\perp).$$

Similarly, we can combine Proposition 5.2 and Corollary 5.5.

Corollary 5.6 *(Ho [11]) Suppose (FD) is feasible and $\{e_j : j \in \{1,2,\ldots,n\}\} \cap S^\perp = \emptyset$. Then*

$$\operatorname{symm}(H) = \min_{\xi \in S, \|\xi_N\|=1} -\frac{\xi_N^+}{\xi_N^-}.$$

We now look at a relationship between the complexity measures $\eta(A)$ and $\rho(A)$. We call AG a *signing* of A, where $G \in \mathcal{G}$ and $\mathcal{G}$ is the set of diagonal matrices in $\mathbb{R}^{n\times n}$ with diagonal entries from $\{1,-1\}$. Note that $\bar{\chi}(AG) = \bar{\chi}(A)$.

Define $\underline{\eta}(A) := \min_{G\in\mathcal{G}} \eta(AG)$. We have the following fact.

Proposition 5.7 *(Todd, Tunçel and Ye* [24]*)*

$$\frac{1}{\sqrt{n}}\underline{\eta}(A) \le \rho(A) \le \underline{\eta}(A).$$

The second inequality above can be obtained easily from the results of Vavasis and Ye [30] and Gonzaga and Lara [8]. The first inequality can be proved using Propositions 2.6 and 5.1. The second author [26] showed that in general, η may carry no information about ρ. Indeed, suppose the columns of A define an almost centrally symmetric polytope. Then there is a signing of A such that the new polytope is highly asymmetric and therefore has a very small η value, which in turn implies a very small ρ value. This suggests that $\bar{\chi}$ may not be a good complexity measure as it tends to grossly overestimate the complexity of interior-point algorithms. Even though $\bar{\chi}(A)$ grossly overestimates the amount of computational work to solve LP problems with data (A, b, c), it has been useful in estimating the work for LP problems having A as the coefficient matrix, with arbitrary b and c and arbitrary orientation of inequalities. Also, $\Delta(A)/\delta(A)$ has a similar role.

Proposition 5.7 shows that $\frac{1}{\bar{\chi}(A)}$ behaves like $\underline{\eta}(A)$ or like $\eta(AG)$, where G is "the worst signing of A" in this context. Notice that Theorem 4.6 relates Hoffman's bound to $\chi(A)$ in a similar way. It shows that $\chi(A)$ is at least $K_2(AG)$, where G is "the worst signing of A" in this latter context. Since $\eta(S)$ is essentially $\mathrm{symm}(A)$ and we have noticed the above parallel, we give below a brief geometric interpretation of $K_{1,\infty}$, in a special but illustrative case. Note that the essential difference between η and K is that of formulation. They both measure similar quantities; considering the problem (D), K works in the y-space and η in the s-space. See the next section for similar situations between χ and $\bar{\chi}$.

Let us now look at $K_{\alpha,\beta}(A)$ more closely. For this brief discussion, we assume that $V_\alpha(A)$ is bounded. This is true if and only if $\{v : Av = 0, v \ge 0, v \ne 0\} = \emptyset$, if and only if there exists $y \in \mathbb{R}^m$ such that $A^T y > 0$, by LP duality theory. Under this assumption,

$$\begin{aligned} K_{\alpha,\beta}(A) &= \max\{\|v\|_\beta^* : v \ge 0, \|Av\|_\alpha^* \le 1\} \\ &= \max\{\|v\|_\beta^* : v \ge 0, \|Av\|_\alpha^* = 1\} \end{aligned}$$

$$= \max\left\{\frac{\|v\|_\beta^*}{\|Av\|_\alpha^*} : v \geq 0, Av \neq 0\right\}.$$

Also $v = 0$ if and only if $Av = 0$. Hence,

$$\frac{1}{K_{\alpha,\beta}(A)} = \min\left\{\frac{\|Av\|_\alpha^*}{\|v\|_\beta^*} : v \geq 0, v \neq 0\right\}$$
$$= \min\{\|Av\|_\alpha^* : v \geq 0, \|v\|_\beta^* = 1\}.$$

For the case $\alpha = 1, \beta = \infty$, we have

$$\frac{1}{K_{1,\infty}(A)} = \min\{\|Av\|_\infty : v \geq 0, e^T v = 1\}.$$

This is precisely the ∞-norm distance of the origin of $\mathbb{R}^m$ to the convex hull of the column vectors of A. Since we assume that V_1 is bounded, 0 is not in this convex hull. On the other hand,

$$\frac{1}{K_{1,\infty}(A)} = \min\{t : \|Av\|_\infty \leq t, v \geq 0, e^T v = 1\}$$
$$= \min\left\{t : \begin{pmatrix} A \\ -A \end{pmatrix} v + te \geq 0, e^T v = 1, v \geq 0\right\}.$$

This is an LP problem. So by LP duality theory,

$$\frac{1}{K_{1,\infty}(A)} = \max\{\eta : [A^T| - A^T]y + \eta e \leq 0, e^T y = 1, y \geq 0\}$$
$$= \max\{\text{smallest entry of } [-A^T|A^T]y : e^T y = 1, y \geq 0\}.$$

In other words, it is the maximum of the smallest entry of any vector in the convex hull of the rows of A and their negations.

5.1 Linear Programming Solver Subroutine

In Section 7, we generalize Tardos' scheme. To do so, we need to solve LP problems with the following data. Define

$$\bar{q} := \max\left\{2\left\lceil \log\left(\frac{\Delta(A)}{\delta(A)}\right)\right\rceil, n\right\}, \tilde{q} := \bar{q}^2,$$

$$\bar{p} := 2^{\lceil \log\left(2(2m+n)^{3/2}(2mn+1)\right)\rceil} 2^{\bar{q}} \text{ and } \tilde{p} := 2^{\lceil \log\left(2(2m+n)^{3/2}(2mn+1)\right)\rceil} 2^{\tilde{q}}. \quad (12)$$

Let p be a positive integer power of two and $p \leq \tilde{p}$. We will not have any restriction on the entries of A, except that we want A to have full row rank (easily ensured). The rest of the data, b and c, for the LP solver subroutine will be restricted to the following two cases.

(i) We set $l := \left[(p+1),(p+1)^2,\ldots,(p+1)^n\right]^T$, and $b := Al$. We have $c \in \mathbb{Z}^n$ such that $||c||_\infty \leq \tilde{p}$.

(ii) We have $b \in \mathbb{Z}^m$, $c \in \mathbb{Z}^n$ such that $||b||_\infty \leq \tilde{p}$ and $||c||_\infty \leq \tilde{p}$.

In this subsection, we assume that b and c satisfy at least one of (i) and (ii). We also need the following function of A in our estimations.

Definition 5.8 *Let $\bar{A} := [A|I]$. For every $B \in \mathcal{B}(\bar{A})$ (N is the complement of B) consider the smallest absolute value of nonzero entries of*

$$\bar{A}_B^{-1}u,\ \forall\, u \in \mathbb{Z}^m \text{ such that } ||u||_\infty \leq \tilde{p},$$

$$\bar{A}_B^{-1}\bar{A}w,\ \forall\, w \in \mathbb{Z}^n,\ \text{with entries from } (p+1),(p+1)^2,\ldots,(p+1)^n,$$

where p is a positive integer power of two and $p \leq \tilde{p}$,

$$\left[-\bar{A}_N^T\bar{A}_B^{-T}|I\right]v,\ \forall\, v \in \mathbb{Z}^{n+m},\ \text{such that } ||v||_\infty \leq \tilde{p}.$$

Also consider the entries of the vectors for the same construction in which $\bar{A}$ is replaced by

$$\bar{A} := \left[A^T| - A^T| - I\right].$$

These generate a finite collection of positive real numbers depending only on A. We call the minimum of all these numbers $\delta_\delta(A)$.

Note that $0 < \delta_\delta(A) \leq 1$ for all $A \in \mathbb{R}^{m\times n}$. If $A \in \mathbb{Z}^{m\times n}$ then $\delta_\delta(A) \geq 1/\Delta(A)$.

The LP problems with b and c described as above (in (i) and (ii)) depend only on A. As we show in this subsection, many algorithms can be adapted to solve such LP problems in poly $\left(n, |\log(\delta_\delta(A))|, \log\left(\frac{\Delta(A)}{\delta(A)}\right)\right)$ elementary arithmetic operations. In particular, we show that such polynomial bounds can be satisfied by employing almost any primal-dual interior-point algorithm with (mild centrality properties and) polynomial-time complexity in the Turing Machine Model. Consider the homogeneous self-dual linear programming problem $(HSDLP)$:

$$\begin{array}{ll}
\min & (n+1)\theta \\
\text{subject to} & \begin{pmatrix} 0 & A & -b & b-Ae \\ -A^T & 0 & c & e-c \\ b^T & -c^T & 0 & e^Tc+1 \\ (Ae-b)^T & (c-e)^T & -(e^Tc+1) & 0 \end{pmatrix}\begin{pmatrix} y \\ x \\ \tau \\ \theta \end{pmatrix} \begin{array}{c} = \\ \geq \\ \geq \\ = \end{array} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -(n+1) \end{pmatrix}, \\
& y \text{ free}, \\
& x \geq 0, \\
& \tau \geq 0, \\
& \theta \text{ free}.
\end{array}$$

Note that $(HSDLP)$ is self-dual, and that $\theta = 0$ at every optimal solution of $(HSDLP)$. Let us define the surplus variables for the inequalities above:

$$s := -A^T y + \tau c + \theta(e - c),$$
$$\psi := b^T y - c^T x + \theta(e^T c + 1).$$

Then $\bar{y} := 0, \bar{x} := e, \bar{s} := e, \bar{\tau} := 1, \bar{\psi} := 1, \bar{\theta} := 1$ is feasible in $(HSDLP)$. For various facts on such formulations, see the book by Roos, Terlaky and Vial [19].

Theorem 5.9 *(Ye, Todd and Mizuno[32])*
Let $(y^*, x^*, \tau^*, \theta^* = 0, s^*, \psi^*)$ be a strictly self-complementary solution for $(HSDLP)$. Then,

1. *(P) has a solution if and only $\tau^* > 0$. In this case, x^*/τ^* is an optimal solution for (P) and $(y^*/\tau^*, s^*/\tau^*)$ is an optimal solution for (D),*

2. *if $\tau^* = 0$, then $\psi^* > 0$, which implies that $c^T x^* - b^T y^* < 0$, that is, at least one of $c^T x^*$ and $-b^T y^*$ is strictly less than zero. If $c^T x^* < 0$, then (D) is infeasible; if $-b^T y^* < 0$, then (P) is infeasible; if both $c^T x^* < 0$ and $-b^T y^* < 0$, then both (P) and (D) are infeasible.*

Consider the setting at the very beginning of Section 5. Assume both (P) and (D) have feasible solutions. Let $\left\{\left(x^{(k)}, s^{(k)}\right)\right\}$, $k \in \mathbb{Z}_+$ denote the iterates of a primal-dual interior-point algorithm (with feasible iterates). Güler and Ye [10] proved that the mild, wide neighborhood condition (or centrality condition)

$$\frac{\min_j \left\{x_j^{(k)} s_j^{(k)}\right\}}{\left(x^{(k)}\right)^T s^{(k)}} \geq \Omega\left(\frac{1}{n}\right) \tag{13}$$

guarantees that every limit point of $\left\{\left(x^{(k)}, s^{(k)}\right)\right\}$ is a strictly complementary pair. Mehrotra and Ye [16] and Ye [31] showed how to make such polynomial-time primal-dual interior-point algorithms terminate in $O\left(\sqrt{n}\,|\log(\eta(P, D)|\right)$ iterations.

Results of Ye-Todd-Mizuno [32] and Ye [31] also show how to terminate primal-dual interior-point algorithms (those converging to a strictly complementary pair) after $O(\sqrt{n}|\log(\eta(HSDLP))|)$ iterations. We denoted by $\eta(HSDLP)$, Ye's complexity measure applied to the problem $(HSDLP)$. Since the problem is self-dual, the notation is consistent.

Next, we will estimate $\eta(HSDLP)$. The optimal value of $(HSDLP)$ is 0. Therefore, we can represent the set of optimal solutions of $(HSDLP)$ as $(FHSDLP)$:

$$Ax = \tau b,$$

$$\begin{aligned} A^T y + s &= \tau c, \\ b^T y - c^T x &= \psi, \\ e^T x + e^T s + \tau + \psi &= n + 1, \\ x, s, \tau, \psi &\geq 0. \end{aligned}$$

By the last equation and the nonnegativity constraints, we have

$$0 < \eta(HSDLP) \leq n + 1.$$

It remains to bound $\eta(HSDLP)$ from below and away from zero. We want to maximize each restricted variable (say x_j for some j) subject to $(FHSDLP)$. We will split the analysis into three exhaustive cases:

1. (P) and (D) both have feasible solutions,
2.(a) (D) is infeasible,
2.(b) (P) is infeasible.

As mentioned before, we will assume that b and c satisfy (i) or (ii), and we will differentiate the analysis of these two cases, whenever necessary.

Case 1.: (P) and (D) both have feasible solutions

Every solution of $(FHSDLP)$ satisfies $\psi = b^T y - c^T x = 0$, by LP weak duality and the constraint $\psi \geq 0$. Also, there exists a solution of $(FHSDLP)$ with $\tau > 0$. Let $(\bar{x}, \bar{y}, \bar{s})$ be a basic primal-dual pair of optimal solutions for (P) and (D). So for some $B \in \mathcal{B}(A)$, we have

$$\bar{x}_B = A_B^{-1} b, \bar{s}_N = c_N - A_N^T A_B^{-T} c_B,$$

where $N := \{1, \ldots, n\} \setminus B$. For case (i), we have

$$\begin{aligned} e^T \bar{x} = \|\bar{x}_B\|_1 \leq \sqrt{m} \|\bar{x}_B\| = \sqrt{m} \|A_B^{-1} A l\| &\leq \sqrt{m} \|A_B^{-1} A\| \cdot \|l\| \\ &\leq \sqrt{nm} \bar{\chi}(A)(\tilde{p} + 1)^n. \end{aligned}$$

For case (ii), we have

$$\begin{aligned} e^T \bar{x} &\leq \sqrt{m} \|A_B^{-1} b\| \leq \sqrt{m} \|A_B^{-1}\| \cdot \|b\| \leq m \chi(A) \tilde{p} \\ &\leq m^2 \tilde{p} \frac{\Delta(A)}{\delta(A)} \leq \tilde{p}^3 \leq \sqrt{nm} \bar{\chi}(A)(\tilde{p} + 1)^n, \end{aligned}$$

where the fourth inequality uses Proposition 2.13. Similarly,

$$e^T \bar{s} \leq \|c_N\|_1 + \|A_N^T A_B^{-T} c_B\|_1 \leq \sqrt{n} \bar{\chi}(A) \|c\|_1 \leq n^{3/2} \bar{\chi}(A) \tilde{p}.$$

Let

$$\bar{\tau} := \frac{n + 1}{e^T \bar{x} + e^T \bar{s} + 1}.$$

Then $(\bar\tau\bar y, \bar\tau\bar x, \bar\tau\bar s, \bar\tau, \bar\psi := 0)$ is a solution of $(FHSDLP)$. Hence,

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} \tau \ge \bar\tau \ge \frac{n+1}{\sqrt{nm}\bar\chi(A)(\tilde p+1)^n + n^{3/2}\bar\chi(A)\tilde p + 1}$$
$$\ge \frac{n+1}{3n^{3/2}\bar\chi(A)(\tilde p+1)^n}.$$

Let $[B', N']$ be the (unique) strict complementarity partition (restricted to just the indices x, or s) for $(HSDLP)$. Let $j \in B'$. Then there exists a basic primal-dual pair of optimal solutions for (P) and (D), $(\bar x, \bar y, \bar s)$, corresponding to some new basis B, such that $\bar x_j > 0$. Then all the above arguments apply with this new B. Since $\bar x_j = (A_B^{-1}b)_j \ge \delta_\delta(A)$, we have

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} x_j \ge \bar\tau\bar x_j \ge \frac{(n+1)\delta_\delta(A)}{3n^{3/2}\bar\chi(A)(\tilde p+1)^n}.$$

Similarly, for each $j \in N'$, there exists $\bar s$ corresponding to some basis B such that $\bar s_j > 0$. Then $j \in N$ and

$$\bar s_j = \left([-A_N^T A_B^{-T} | I] \begin{bmatrix} c_B \\ c_N \end{bmatrix}\right)_j \ge \delta_\delta(A).$$

Hence, we have

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} s_j \ge \bar\tau\bar s_j \ge \frac{(n+1)\delta_\delta(A)}{3n^{3/2}\bar\chi(A)(\tilde p+1)^n}.$$

Therefore, since $\delta_\delta(A) \le 1$,

$$\eta(HSDLP) \ge \frac{(n+1)\delta_\delta(A)}{3n^{3/2}\bar\chi(A)(\tilde p+1)^n}$$

in this case.

Case 2.(a): (D) is infeasible

Every solution of $(FHSDLP)$ satisfies $\tau = 0$, because if (y, x, s, τ, ψ) is a solution such that $\tau > 0$, then $(y/\tau, s/\tau)$ is a feasible solution of (D). On the other hand, by Farkas' lemma,

$$\min\{c^T x : Ax = 0, e^T x = 1, x \ge 0\} < 0.$$

Let $\bar x$ be a basic optimal solution of this problem. So for some $B \in \mathcal{B}(A)$ and $k \in \{1, \dots, n\} \setminus B$ such that $\bar x_k \ne 0$, we have $A_B\bar x_B = -A_k\bar x_k$ and $\bar x_j = 0$ for all $j \notin B \cup \{k\}$. It is easy to see that

$$\left(0, \frac{(n+1)\bar x}{1 - c^T\bar x}, 0, 0, \frac{-(n+1)c^T\bar x}{1 - c^T\bar x}\right) \in (FHSDLP).$$

Note that $1 = e^T\bar{x} = \bar{x}_k - \bar{x}_k e^T A_B^{-1} A_k$, which implies

$$\bar{x}_k = \frac{1}{1 - e^T A_B^{-1} A_k} > 0.$$

Now, $1 - e^T A_B^{-1} A_k \leq 1 + \|A_B^{-1} A_k\|_1 \leq 1 + \sqrt{m}\|A_B^{-1} A_k\| \leq 1 + \sqrt{m}\bar{\chi}(A)$. So,

$$\begin{aligned} 0 < -c^T\bar{x} &= |c^T\bar{x}| = |c_k\bar{x}_k + c_B^T(-A_B^{-1}A_k\bar{x}_k)| \\ &= \frac{|c_k - c_B^T A_B^{-1} A_k|}{1 - e^T A_B^{-1} A_k} \geq \frac{\delta_\delta(A)}{1 + \sqrt{m}\bar{\chi}(A)}. \end{aligned} \tag{14}$$

Also, $-c^T\bar{x} \leq \tilde{p}$ (since $\bar{x} \geq 0$ and $e^T\bar{x} = 1$). So,

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} \psi \geq \frac{-(n+1)c^T\bar{x}}{1 - c^T\bar{x}} \geq \frac{(n+1)\delta_\delta(A)}{(1 + \sqrt{m}\bar{\chi}(A))(\tilde{p} + 1)}.$$

Let $j \in B'$ where $[B', N']$ is, as before, the (unique) strict complementarity partition (restricted to just the subvectors x and s) for $(HSDLP)$. Let $\tilde{x}$ be a maximizer of

$$\max\{x_j : Ax = 0, e^T x = 1, x \geq 0\}.$$

Note that $\tilde{x}_j \geq \eta(\mathcal{N}(A)) \geq \eta(A)$. Also $|c^T\tilde{x}| \leq \tilde{p}$. Let

$$\hat{x} := (1 + \sqrt{m}\bar{\chi}(A))\bar{x} + \frac{\delta_\delta(A)}{\tilde{p}}\tilde{x}.$$

Now, by (14), $c^T\hat{x} \leq 0$. So,

$$\left(0, \frac{(n+1)\hat{x}}{e^T\hat{x} - c^T\hat{x}}, 0, 0, \frac{-(n+1)c^T\hat{x}}{e^T\hat{x} - c^T\hat{x}}\right) \in (FHSDLP).$$

Note that $-c^T\hat{x} \leq (1 + \sqrt{m}\bar{\chi}(A))\tilde{p} + \delta_\delta(A) \leq (1 + \sqrt{m}\bar{\chi}(A))\tilde{p} + 1$ and $e^T\hat{x} \leq 2 + \sqrt{m}\bar{\chi}(A)$. Therefore,

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} x_j \geq \frac{(n+1)\hat{x}_j}{e^T\hat{x} - c^T\hat{x}} \geq \frac{(n+1)\delta_\delta(A)\eta(A)}{\tilde{p}\left[(1 + \sqrt{m}\bar{\chi}(A))(\tilde{p} + 1) + 2\right]}.$$

Now let $j \in N'$. Consider the problem

$$\max\{s_j : s \in \mathcal{R}(A^T), e^T s = 1, s \geq 0\}.$$

First note that if this problem is infeasible, then every solution of $(FHSDLP)$ satisfies $s = 0$ and hence N' is empty; and we are done. So we assume the problem has a feasible solution and because the feasible set is compact, the maximum is attained by some basic solution, say $\tilde{s}$, corresponding to some basis $B \in \mathcal{B}(A)$. Note that $\tilde{s}_j \geq \eta(\mathcal{R}(A^T)) \geq \eta(A)$. We (again) let

$N := \{1, \ldots, n\} \setminus B$. Let $\tilde{y}$ be the unique vector in $\mathbb{R}^m$ such that $A^T\tilde{y} = -\tilde{s}$. For case (i), we let $l \in \mathbb{R}^n$ be as in the assumption given before. For case (ii), we let $l \in \mathbb{R}^n$ be such that $l_B := A_B^{-1}b$ and $l_N := 0$. In both cases, we have $Al = b$ and

$$b^T\tilde{y} = l^T A^T \tilde{y} = -l^T\tilde{s}.$$

For case (i), it is clear that $|l^T\tilde{s}| \leq (\tilde{p}+1)^n$. This is also true for case (ii) because

$$\begin{aligned} |l^T\tilde{s}| &= |l_B^T\tilde{s}_B| = |(A_B^{-1}b)^T\tilde{s}_B| \leq \|A_B^{-1}b\| \cdot \|\tilde{s}_B\| \leq \chi(A)\sqrt{m} \cdot \|b\|_\infty \|\tilde{s}_B\|_1 \\ &\leq \sqrt{m}\tilde{p}\chi(A) \leq m^{3/2}\tilde{p}\left(\frac{\Delta(A)}{\delta(A)}\right) \leq \tilde{p}^3 \leq (\tilde{p}+1)^n, \end{aligned}$$

where we use Proposition 2.13 and the fact that $n \geq 3$. If $l^T\tilde{s} \leq 0$, then

$$\left(\frac{(n+1)\tilde{y}}{1-l^T\tilde{s}}, 0, \frac{(n+1)\tilde{s}}{1-l^T\tilde{s}}, 0, \frac{-(n+1)l^T\tilde{s}}{1-l^T\tilde{s}}\right) \in (FHSDLP).$$

We then have

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} s_j \geq \frac{(n+1)\tilde{s}_j}{1-l^T\tilde{s}} \geq \frac{(n+1)\eta(A)}{(\tilde{p}+1)^n+1}.$$

If $l^T\tilde{s} > 0$, then we can easily show that

$$\left(\frac{-(n+1)(c^T\bar{x})\tilde{y}}{l^T\tilde{s}-c^T\bar{x}}, \frac{(n+1)(l^T\tilde{s})\bar{x}}{l^T\tilde{s}-c^T\bar{x}}, \frac{-(n+1)(c^T\bar{x})\tilde{s}}{l^T\tilde{s}-c^T\bar{x}}, 0, 0\right) \in (FHSDLP).$$

Now, using the fact that $-c^T\bar{x} \leq \tilde{p}$, we have

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} s_j \geq \frac{-(n+1)(c^T\bar{x})\tilde{s}_j}{l^T\tilde{s}-c^T\bar{x}} \geq \frac{(n+1)\delta_\delta(A)\eta(A)}{(1+\sqrt{m}\chi(A))\left[(\tilde{p}+1)^n+\tilde{p}\right]}.$$

Case 2.(b): (P) is infeasible

Note that this case does not apply to case (i), since by construction, $Al = b, l \geq 0$, and therefore (P) must have a feasible solution. So we only need to consider case (ii).

Every solution of $(FHSDLP)$ satisfies $\tau = 0$, because if (y, x, s, τ, ψ) is a solution such that $\tau > 0$, then x/τ is a feasible solution of (P). On the other hand, by Farkas' lemma,

$$\max\{b^Ty : A^Ty \leq 0, e^TA^Ty = 1\} > 0.$$

Let $s = -A^Ty$. Then as before, we have $b^Ty = -l^Ts$. So the above problem can be rewritten as

$$\max\{-l^Ts : s \in \mathcal{R}(A^T), e^Ts = 1, s \geq 0\}.$$

Now let $D \in \mathbb{R}^{(n-m)\times n}$ be such that the rows are precisely a basis of $\mathcal{N}(A)$. We know $\mathcal{R}(D^T) = \mathcal{N}(A)$ and $\mathcal{N}(D) = \mathcal{R}(A^T)$. In particular, $\bar{\chi}(D) = \bar{\chi}(A)$, which we will use later on. Therefore, the above problem can be further rewritten as

$$\max\{-l^Ts : Ds = 0, e^Ts = 1, s \geq 0\}.$$

Let $\bar{s}$ be a basic optimal solution of this problem. So for some $N \in \mathcal{B}(D)$ and $k \in \{1, \ldots, n\} \setminus N$ such that $\bar{s}_k \neq 0$, we have $D_N\bar{s}_N = -D_k\bar{s}_k$ and $\bar{s}_j = 0$ for all $j \notin N \cup \{k\}$. Let $\bar{y}$ be the unique vector in $\mathbb{R}^m$ such that $A^T\bar{y} = -\bar{s}$. It is easy to see that

$$\left(\frac{(n+1)\bar{y}}{1+b^T\bar{y}}, 0, \frac{(n+1)\bar{s}}{1+b^T\bar{y}}, 0, \frac{(n+1)b^T\bar{y}}{1+b^T\bar{y}},\right) \in (FHSDLP).$$

Note that $1 = e^T\bar{s} = \bar{s}_k - \bar{s}_k e^T D_N^{-1} D_k$, which implies

$$\bar{s}_k = \frac{1}{1 - e^T D_N^{-1} D_k} > 0.$$

Now,

$$\begin{aligned} 1 - e^T D_N^{-1} D_k &\leq 1 + \|D_N^{-1} D_k\|_1 \leq 1 + \sqrt{n-m}\|D_N^{-1} D_k\| \\ &\leq 1 + \sqrt{n-m}\bar{\chi}(D) = 1 + \sqrt{n-m}\bar{\chi}(A). \end{aligned}$$

Since the choice of B in the definition of l (for case (ii) in case 2(a)) does not affect the previous arguments, we can redefine l using $B := \{1, \ldots, n\} \setminus N$. It is not hard to see that $B \in \mathcal{B}(A)$. So we have

$$0 < b^T\bar{y} = |l^T\bar{s}| = |l_B^T\bar{s}_B| = |(A_B^{-1}b)^T\bar{s}_B| = |(A_B^{-1}b)_k|\bar{s}_k \geq \frac{\delta_\delta(A)}{1+\sqrt{n-m}\bar{\chi}(A)}. \tag{15}$$

Also, $b^T\bar{y} = -l^T\bar{s} \leq (\tilde{p}+1)^n$, as we have shown before. So,

$$\max_{(y,x,s,\tau,\psi)\in(FHSDLP)} \psi \geq \frac{(n+1)b^T\bar{y}}{1+b^T\bar{y}} \geq \frac{(n+1)\delta_\delta(A)}{(1+\sqrt{n-m}\bar{\chi}(A))\left[(\tilde{p}+1)^n+1\right]}.$$

Recall that $[B', N']$ denotes, as in the previous cases, the (unique) strict complementarity partition (restricted to the subvectors x and s) for $(HSDLP)$. Now let $j \in N'$. Let $\tilde{s}$ be a maximizer of

$$\max\{s_j : s \in \mathcal{R}(A^T), e^Ts = 1, s \geq 0\}.$$

Note that $\tilde{s}_j \geq \eta(A)$, and $|l^T\tilde{s}| \leq (\tilde{p}+1)^n$. Let

$$\hat{s} := (\tilde{p}+1)^n(1+\sqrt{n-m}\bar{\chi}(A))\tilde{s} + \delta_\delta(A)\bar{s}.$$

Let $\hat{y}$ be the unique vector in $\mathbb{R}^m$ such that $A^T\hat{y} = -\hat{s}$. Now, by (15), $l^T\hat{s} \leq 0$. So we have $b^T\hat{y} = -l^T\hat{s} \geq 0$. Therefore,

$$\left(\frac{(n+1)\hat{y}}{e^T\hat{s}+b^T\hat{y}}, 0, \frac{(n+1)\hat{s}}{e^T\hat{s}+b^T\hat{y}}, 0, \frac{(n+1)b^T\hat{y}}{e^T\hat{s}+b^T\hat{y}}\right) \in (FHSDLP).$$

Now,

$$-l^T\hat{s} \leq (\tilde{p}+1)^{2n}(1+\sqrt{n-m}\bar{\chi}(A)) + (\tilde{p}+1)^n.$$

Therefore,

$$\begin{aligned} \max_{(y,x,s,\tau,\psi)\in(FHSDLP)} s_j &\geq \frac{(n+1)\hat{s}_j}{e^T\hat{s}+b^T\hat{y}} \\ &\geq \frac{(n+1)\delta_\delta(A)\eta(A)}{[(\tilde{p}+1)^{2n}+(\tilde{p}+1)^n](1+\sqrt{n-m}\bar{\chi}(A))+(\tilde{p}+1)^n+1}. \end{aligned}$$

If B' is empty, then we are done. Otherwise, let $j \in B'$. Consider the problem

$$\max\{x_j : Ax = 0, e^Tx = 1, x \geq 0\}.$$

Let $\tilde{x}$ be a basic optimal solution of this problem such that $A_B\tilde{x}_B = -\tilde{x}_jA_j$, where we called the corresponding basis B. First, we have $\tilde{x}_j \geq \eta(S) \geq \eta(A)$ by definitions. Also, $|c^T\tilde{x}| \leq \tilde{p}$. If $c^T\tilde{x} \leq 0$, then

$$\left(0, \frac{(n+1)\tilde{x}}{1-c^T\tilde{x}}, 0, 0, -\frac{(n+1)c^T\tilde{x}}{1-c^T\tilde{x}}\right) \in (FHSDLP).$$

If $c^T\tilde{x} > 0$, then

$$\left(\frac{(n+1)(c^T\tilde{x})\bar{y}}{b^T\bar{y}+c^T\tilde{x}}, \frac{(n+1)(b^T\bar{y})\tilde{x}}{b^T\bar{y}+c^T\tilde{x}}, \frac{(n+1)(c^T\tilde{x})\bar{s}}{b^T\bar{y}+c^T\tilde{x}}, 0, 0\right) \in (FHSDLP).$$

Using $\frac{\delta_\delta(A)}{1+\sqrt{n-m}\bar{\chi}(A)} \leq b^T\bar{y} \leq (\tilde{p}+1)^n$ and $c^T\tilde{x} \leq \tilde{p}$, we conclude

$$\begin{aligned} \max_{(y,x,s,\tau,\psi)\in(FHSDLP)} x_j &\geq \frac{(n+1)(b^T\bar{y})\eta(A)}{b^T\bar{y}+c^T\tilde{x}} \\ &\geq \frac{(n+1)\delta_\delta(A)\eta(A)}{[(\tilde{p}+1)^n+\tilde{p}](1+\sqrt{n-m}\bar{\chi}(A))}. \end{aligned}$$

The above lower bound on x_j also applies in the case that $c^T\tilde{x} \leq 0$. We proved the following fact.

Theorem 5.10 *Consider feasible-start primal-dual interior-point algorithms satisfying condition (13) above and have been proven to run in polynomial time, with $O(\sqrt{n}\,|\log(\eta(P,D))|)$ iteration complexity. Every such algorithm when applied to $(HSDLP)$ with the staring point $\bar{y} := 0, \bar{x} := e, \bar{s} := e, \bar{\tau} := 1, \bar{\psi} := 1, \bar{\theta} := 1$, terminates correctly in*

$$O\left(\sqrt{n}\left(|\log(\delta_\delta(A))| + n\log\left(\frac{\Delta(A)}{\delta(A)}\right) + n\log(n)\right)\right)$$

iterations.

Here we used Propositions 5.7 and 2.14 to see that

$$\eta(A) \geq \rho(A) = \frac{1}{\bar{\chi}(A)} \geq \frac{1}{\sqrt{m(n-m)+1}} \cdot \frac{\delta(A)}{\Delta(A)},$$

and so conclude that

$$|\log(\eta(A))| \leq O\left(\log(n) + \log\left(\frac{\Delta(A)}{\delta(A)}\right)\right).$$

The last inequality above can also be obtained directly from the definition of $\eta(A)$ by utilizing the techniques in Section 2. Note that the above theorem stays valid if we replace A by any submatrix of it. This is one of the reasons why in Definition 5.8, we chose $\bar{A}$ as $[A|I]$, rather than just A. Each iteration can be performed in $O(n^3)$ elementary arithmetic operations.

6 Sensitivity Analysis, Hoffman's Bound, $\chi, \bar{\chi}, \Delta$, and δ

Given an LP $\max\{b^T y : A^T y \leq c\}$, we are interested in the change in the set of optimal solutions as the vector c is varied. Let $\bar{\Delta}(A)$ denote the maximum of the absolute values of the entries of C^{-1} over all nonsingular submatrices C of A.

Proposition 6.1 *(Cook, Gerards, Schrijver, Tardos [3], [20])*
Suppose $A \in \mathbb{R}^{m\times n}$ (not necessarily full row rank), $c, c' \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$, such that both LP problems $\max\{b^T y : A^T y \leq c\}$ and $\max\{b^T y : A^T y \leq c'\}$ have optimal solutions. Then for every optimal solution $\bar{y}$ of $\max\{b^T y : A^T y \leq c\}$, there exists an optimal solution $\bar{y}'$ of $\max\{b^T y : A^T y \leq c'\}$ with

$$\|\bar{y} - \bar{y}'\|_\infty \leq m\bar{\Delta}(A)\|c - c'\|_\infty.$$

Note that $\bar{\Delta}(A) \leq \Delta(A)/\delta(A)$ for all A, by Cramer's Rule. In particular, if $A \in \mathbb{Z}^{m\times n}$, then $\bar{\Delta}(A) \leq \Delta(A)$. In fact, Cook et al. state the above proposition in [3] for integral A, and $\bar{\Delta}(A)$ above is replaced by $\Delta(A)$.

We define, for A with full row rank,

$$\chi_1(A) := \max\{\|A_B^{-1}\|_1 : B \in \mathcal{B}(A)\},$$

and

$$\bar{\chi}_1(A) := \max\{\|A_B^{-1}A\|_1 : B \in \mathcal{B}(A)\}.$$

Using almost exactly the same arguments as in the above proof, together with Proposition 2.3, we can give an alternative sensitivity bound in terms of $\chi(A)$.

Corollary 6.2 *If the A in Proposition 6.1 has full row rank, then*

$$\|\bar{y} - \bar{y}'\|_\infty \leq \chi_1(A)\|c - c'\|_\infty.$$

Following the proof of Cook et al. we also have the following useful theorem in terms of $\bar{\chi}(A)$.

Theorem 6.3 *Let $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank}(A) = m$, $c, c' \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$, such that both LP problems $\max\{b^T y : A^T y + s = c, s \geq 0\}$ and $\max\{b^T y : A^T y + s = c', s \geq 0\}$ have optimal solutions. Then for every optimal solution $(\bar{y}, \bar{s})$ of the former problem, there exists an optimal solution $(\bar{y}', \bar{s}')$ of the latter problem with*

$$\|\bar{s} - \bar{s}'\|_\infty \leq (\bar{\chi}_1(A) + 1)\|c - c'\|_\infty.$$

Proof
We first show the inequality for the special case $b = 0$. Then we use the special case to establish the theorem. Assume for now that $b = 0$. Suppose for a contradiction that there exists $(\bar{y}, \bar{s})$ feasible for the first problem such that no feasible solution $(\bar{y}', \bar{s}')$ of the latter problem satisfies

$$\|\bar{s} - \bar{s}'\|_\infty \leq (\bar{\chi}_1(A) + 1)\|c - c'\|_\infty.$$

Then the system

$$A^T y + s = c', s \leq \bar{s} + pe, -s \leq -\bar{s} + pe, s \geq 0,$$

where $p := (\bar{\chi}_1(A) + 1)\|c - c'\|_\infty$, has no solution. By Farkas' lemma, there exist $x \in \mathbb{R}^n, u, v \in \mathbb{R}^n_+$ such that

$$Ax = 0, x + u - v \geq 0, (c')^T x + \bar{s}^T(u - v) + p(e^T u + e^T v) < 0.$$

Note that if $u = v = 0$, then the above x proves that the system $\{A^T y + s = c', s \geq 0\}$ is infeasible, a contradiction. Therefore, $u + v \neq 0$. Let

$$\bar{u} := \frac{u}{\|u + v\|_1}, \bar{v} := \frac{v}{\|u + v\|_1},$$

so that $||\bar{u}+\bar{v}||_1 = 1$. Let $\bar{x}$ be a basic optimal solution of

$$\min\{(c')^T x : Ax = 0, x \geq -(\bar{u}-\bar{v})\}.$$

Note that this problem has a feasible solution (for example, $x/||u+v||_1$). Also it is bounded, because otherwise there exists $d \in \mathbb{R}^n_+$ such that $d \neq 0, Ad = 0, (c')^T d < 0$ which implies that $\{A^T y + s = c', s \geq 0\}$ is infeasible, a contradiction. Note that $\bar{x} = \tilde{x} - (\bar{u}-\bar{v})$, where, for some $B \in \mathcal{B}(A)$,

$$\tilde{x}_B = A_B^{-1} A(\bar{u}-\bar{v}) \geq 0, \tilde{x}_N = 0.$$

Thus,

$$||\bar{x}||_1 \leq ||\tilde{x}||_1 + ||\bar{u}-\bar{v}||_1 \leq ||A_B^{-1}A(\bar{u}-\bar{v})||_1 + ||\bar{u}+\bar{v}||_1 \leq ||A_B^{-1}A||_1 + 1 \leq \bar{\chi}_1(A)+1.$$

This gives a contradiction since

$$\begin{aligned}
0 &> (c')^T \left(\frac{x}{||u+v||_1}\right) + \bar{s}^T(\bar{u}-\bar{v}) + p \\
&\geq (c')^T \bar{x} + \bar{s}^T(\bar{u}-\bar{v}) + p \\
&\geq (c')^T \bar{x} - (c - A^T\bar{y})^T \bar{x} + p \\
&= (c'-c)^T \bar{x} + p \\
&\geq -||c-c'||_\infty ||\bar{x}||_1 + p \\
&\geq -(\bar{\chi}_1(A)+1)||c-c'||_\infty + p = 0.
\end{aligned}$$

So there exists $(\bar{y}', \bar{s}')$ feasible in the second system of the theorem such that

$$||\bar{s}-\bar{s}'||_\infty \leq (\bar{\chi}_1(A)+1)||c-c'||_\infty.$$

This completes the proof for the special case $b = 0$.
Now, consider the general case. Let $(\bar{y}, \bar{s}, \bar{x})$ be an optimal solution of

$$\max\{b^T y : A^T y + s = c, s \geq 0\}$$

and its dual. Let $J := \{j : \bar{s}_j = 0\}$. Let (y^*, s^*) be an optimal solution of

$$\max\{b^T y : A^T y + s = c', s \geq 0\}.$$

We have, by complementary slackness, $\bar{x}_j = 0$ for all $j \notin J$, and so

$$A_J \bar{x}_J = b, \bar{x}_J \geq 0.$$

Also,

$$A_J^T \bar{y} = c_J \geq c'_J - ||c_J - c'_J||_\infty e \geq A_J^T y^* - ||c-c'||_\infty e.$$

We proved that

$$A^T \bar{y} \leq c, -A_J^T \bar{y} \leq ||c-c'||_\infty e - A_J^T y^*.$$

Also the system

$$A^T y \le c', -A_J^T y \le -A_J^T y^*$$

has a feasible solution (for example, y^*). Therefore, by applying the first part of the proof (with $b = 0$) to these two systems of inequalities, we conclude that there exists $(\bar{y}', \bar{s}')$ such that

$$A^T \bar{y}' + \bar{s}' = c', -A_J^T \bar{y}' \le -A_J^T y^*, \bar{s}' \ge 0,$$

and

$$\|\bar{s} - \bar{s}'\|_\infty \le (\bar{\chi}_1([A| - A_J]) + 1)\|c - c'\|_\infty.$$

Note that

$$b^T \bar{y}' = \bar{x}_J^T A_J^T \bar{y}' \ge \bar{x}_J^T A_J^T y^* = b^T y^*.$$

Therefore, $(\bar{y}', \bar{s}')$ is an optimal solution of $\max\{b^T y : A^T y + s = c', s \ge 0\}$. We have (trivially, from (1))

$$\bar{\chi}_1([A| - A_J]) = \bar{\chi}_1(A).$$

We conclude

$$\|\bar{s} - \bar{s}'\|_\infty \le (\bar{\chi}_1(A) + 1)\|c - c'\|_\infty$$

and this completes the proof. □

Using (5), we easily have the following facts.

Corollary 6.4 *Under the same assumptions as in Theorem 6.3, we have*

$$\|\bar{y} - \bar{y}'\|_\infty \le \sqrt{m}\chi(A)\|c - c'\|_\infty$$

and

$$\|\bar{s} - \bar{s}'\|_\infty \le (\sqrt{m}\chi(A) + 1)\|c - c'\|_\infty.$$

Note that converting norms inside the proof of Theorem 6.3 would also give the same constant for the bound in terms of χ; however, for $\bar{\chi}$, we would have to resort to Proposition 2.8, leading to an unnecessary factor of $\sqrt{2}$ in the upper bound.

For the LP problems in the primal form, we define

$$\bar{\chi}_\infty(A) := \max\{\|A_B^{-1}A\|_\infty : B \in \mathcal{B}(A)\}$$

and prove by the above techniques the following fact.

Theorem 6.5 *Suppose $A \in \mathbb{R}^{m\times n}$ has full row rank, $c \in \mathbb{R}^n$ and $l, l' \in \mathbb{R}^n$ such that both LP problems $\min\{c^Tx : Ax = 0, x \geq -l\}$ and $\min\{c^Tx : Ax = 0, x \geq -l'\}$ have optimal solution(s). Then for every optimal solution $\bar{x}$ of the former problem, there exists an optimal solution $\bar{x}'$ of the latter problem with*

$$\|\bar{x} - \bar{x}'\|_\infty \leq (\bar{\chi}_\infty(A) + 2) \cdot \|l - l'\|_\infty.$$

7 Tardos' Theorem

Tardos [22] shows that any LP problem $\max\{b^Ty : A^Ty \leq c\}$ (with integer or rational data) can be solved in at most poly(size(A)) elementary arithmetic operations on numbers of size polynomially bounded by size(A, b, c). Here we extend her ideas to the case of real number data. The following proofs are very similar to Tardos', and Schrijver's presentation in [20].

7.1 Assumptions

Tardos [22] works with integer (can also easily handle rational numbers) data and the Turing Machine Model. So, not only the number of arithmetic operations but also the sizes of the numbers in intermediate steps are to be bounded by polynomial functions of the input size. In this section, we work with real numbers and utilize Blum-Shub-Smale (BSS) Model (see the book by Blum, Cucker, Shub and Smale [2]). Our final complexity bounds involve complexity measures of the input other than the dimension n. Therefore, to unify the approaches of Vavasis-Ye and Tardos, we introduce below some integers to the complexity model. The sizes of the integers are polynomially bounded in terms of the sizes of the integers closest to our complexity measures. We allow comparison of real numbers to such integers in $O(1)$ time. As a result, determining the "ceiling" of a real number arising from the input data in polynomially many steps of BSS model becomes a polynomial operation for our purposes in this paper. For simplicity, we assume that we can compute the ceiling of such real numbers in $O(1)$ time and consider this operation an *elementary operation.*

Here are some other assumptions that we will make:

1. $A \in \mathbb{R}^{m\times n}$ has full row rank.

2. We can solve the LP problems of the form $(D) : \max\{b^Ty : A^Ty \leq c\}$, where $c \in \{-1, 0, 1\}^n$, $b \in \{-1, 0, 1\}^m$, in at most poly($n, \log(\bar{\chi}(A))$) elementary arithmetic operations.

As we noted before in various settings, Assumption 1 can be made without loss of generality, and is assumed throughout Section 7. Also note that Assumption 2 holds for the Vavasis-Ye algorithm. It is possible that there exists simpler algorithms than Vavasis-Ye's (and with better complexity bounds) for LP problems with the above-mentioned special data.

In this section, we first do our analysis under Assumption 2. This will lay down most of the main ideas and main technical tools needed. Using these, we then show that removing Assumption 2 is possible by utilizing the results of Subsection 5.1.

Proposition 7.1 *Suppose Assumption 2 holds. Then we can solve (D), where $c \in \mathbb{R}^n \setminus \{0\}, b \in \mathbb{R}^m \setminus \{0\}$, in at most*

$$\text{poly}\left(n, \log(\bar{\chi}(A)), \log\left(\frac{\|c\|_\infty}{\min_{c_j \neq 0} |c_j|}\right)\right)$$

elementary arithmetic operations.

Proof

The feasible set $\{A^T y \leq c\}$ can be rewritten as $\{CA^T y \leq Cc\}$, where $C \in \mathbb{R}^{n \times n}$, diagonal, such that for all $j \in \{1, \ldots, n\}$,

$$C_{jj} := \begin{cases} 1/|c_j|, & \text{if } c_j \neq 0, \\ 1/\|c\|_\infty, & \text{if } c_j = 0. \end{cases}$$

Now the problem $\max\{b^T y : CA^T y \leq Cc\}$ is equivalent to $\max\{(Bb)^T w : CA^T Bw \leq Cc\}$, where $w := B^{-1} y$ and $B \in \mathbb{R}^{m \times m}$, diagonal, such that for all $i \in \{1, \ldots, m\}$,

$$B_{ii} := \begin{cases} 1/|b_i|, & \text{if } b_i \neq 0, \\ 1/\|b\|_\infty, & \text{if } b_i = 0. \end{cases}$$

Now $Cc \in \{-1, 0, 1\}^n$ and $Bb \in \{-1, 0, 1\}^m$. So by Assumption 2, we can solve $\max\{b^T y : A^T y \leq c\}$ in at most poly$(n, \log(\bar{\chi}(BAC)))$ elementary arithmetic operations. Now $\bar{\chi}(BAC) = \bar{\chi}(AC)$ since B is nonsingular. Also,

$$\begin{aligned} \|(AC)^T (ACD(AC)^T)^{-1} ACD\| &= \|CA^T (A(CDC)A^T)^{-1} A(CDC)C^{-1}\| \\ &\leq \|C\| \cdot \|C^{-1}\| \cdot \|A^T (A(CDC)A^T)^{-1} A(CDC)\|, \end{aligned}$$

for all positive definite diagonal $n \times n$ matrices D. Therefore,

$$\bar{\chi}(AC) \leq \|C\| \cdot \|C^{-1}\| \cdot \bar{\chi}(A) = \frac{\max_j C_{jj}}{\min_j C_{jj}} \bar{\chi}(A) = \frac{\|c\|_\infty}{\min_{c_j \neq 0} |c_j|} \bar{\chi}(A).$$

So we get the bound

$$\text{poly}\left(n, \log(\bar{\chi}(A)), \log\left(\frac{\|c\|_\infty}{\min_{c_j \neq 0} |c_j|}\right)\right). \qquad \square$$

7.2 Deciding the Feasibility of $A^Ty \leq c$

In this subsection, we describe an iterative algorithm to determine whether $A^Ty \leq c$ has a solution and if not, find a certificate of its infeasibility.

We first use Gaussian elimination to remove any redundant rows of A, to get $\bar{A}$. (Clearly, the given data A has no redundant rows since it has full row rank; but, this procedure is necessary beyond the first iteration as our A changes.) As before, we can replace A by $\bar{A}$ without changing our problem. Now A has full row rank.

Let $c' := (I - A^T(AA^T)^{-1}A)c$. Then for all $d \in \mathcal{N}(A)$,

$$c'^Td = c^Td - d^TA^T(AA^T)^{-1}Ac = c^Td.$$

Since c' is the orthogonal projection of c onto $\mathcal{N}(A)$,

$$\{y : A^Ty \leq c\} = \emptyset \Leftrightarrow \{y : A^Ty \leq c'\} = \emptyset.$$

Therefore, we can replace c by c' without changing our problem. Now we have $c \in \mathcal{N}(A)$.

If $c = 0$, then $y = 0$ is a feasible solution, and we are done. So, we replace c by $c/||c||_\infty$. This does not change our problem since the feasibility of the system is invariant under positive scalar multiplication of c (or independently A). Now we have $||c||_\infty = 1$.

Suppose we are given an integer p such that $p \geq 2n^{3/2}(\bar{\chi}(A))^2$. We first solve $A^Ty \leq \lceil pc \rceil$. If it has no solution, then we have a $d \geq 0$ such that $Ad = 0$ and $\lceil pc \rceil^Td < 0$. This d is also a certificate of the infeasibility of $A^Ty \leq c$, since $(pc)^Td \leq \lceil pc \rceil^Td < 0$, which implies $c^Td < 0$. So we stop.

Therefore, we assume we get $(\bar{y}, \bar{s})$ such that

$$A^T\bar{y} + \bar{s} = \lceil pc \rceil, \bar{s} \geq 0. \tag{16}$$

Lemma 7.2 *Let $c \in \mathcal{N}(A)$, $c \neq 0$. Suppose (y, s) is given such that $A^Ty+s = c$. Then $||s|| \geq ||c||/\bar{\chi}(A)$.*

Proof

We use Proposition 2.6. Note that since the 2-norms are used here, we can interchange $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ in Proposition 2.6, as we noted earlier. Let $\gamma := c/||c||, \xi := A^Ty$, and

$$J := \{j \in \{1, \ldots, n\} : \text{sign}(\gamma_j) \neq \text{sign}(\xi_j)\}.$$

Note that $J \neq \emptyset$ because otherwise $\text{sign}(c) = \text{sign}(A^Ty)$ together with $c \in \mathcal{N}(A)$ would imply $c = 0$, a contradiction. So (γ, ξ, J) is a feasible solution to the minimization problem in Proposition 2.6, and hence $||c_J|| \geq ||c||/\bar{\chi}(A)$. Now, for each $j \in J$,

$$|s_j| = |c_j + (-(A^T y)_j)| = |c_j| + |(A^T y)_j| \geq |c_j|,$$

where the second equality above uses the fact that c_j and $-(A^T y)_j$ either have the same sign or at least one of them is 0. So, $||s|| \geq ||s_J|| \geq ||c_J|| \geq ||c||/\bar{\chi}(A)$.
□

From (16), we have

$$A^T \bar{y} + \bar{s} + pc - \lceil pc \rceil = pc,$$

and hence by Lemma 7.2,

$$||\bar{s} + pc - \lceil pc \rceil|| \geq \frac{||pc||}{\bar{\chi}(A)} \geq \frac{p||c||_\infty}{\bar{\chi}(A)} = \frac{p}{\bar{\chi}(A)}.$$

So,

$$||\bar{s}|| \geq \frac{p}{\bar{\chi}(A)} - ||pc - \lceil pc \rceil|| > \frac{p}{\bar{\chi}(A)} - \sqrt{n},$$

and hence,

$$||\bar{s}||_\infty \geq \frac{||\bar{s}||}{\sqrt{n}} > \frac{p}{\sqrt{n}\bar{\chi}(A)} - 1 \geq 2n\bar{\chi}(A) - 1 \geq n\bar{\chi}(A). \tag{17}$$

Let $J := \{j \in \{1, \ldots, n\} : \bar{s}_j < ||\bar{s}||_\infty\}$. (We could have defined $J := \{j \in \{1, \ldots, n\} : \bar{s}_j < n\bar{\chi}(A)\}$ and the following arguments would work as well. But the difficulty is we cannot compute $n\bar{\chi}(A)$ efficiently.)

Lemma 7.3 *The system $A^T y \leq c$ has a feasible solution if and only if $A_J^T y \leq c_J$ has a feasible solution.*

Proof

Clearly, if $A^T y \leq c$ has a feasible solution, so does $A_J^T y \leq c_J$ since the latter has possibly fewer constraints. If $A^T y \leq c$ has no solution, then by Farkas' lemma, there exists $d \geq 0$ such that $Ad = 0, c^T d < 0$, and (without loss of generality) $e^T d = 1$. We can assume that d is an extreme point of the compact set

$$\{d : Ad = 0, e^T d = 1, d \geq 0\}.$$

So, by Corollary 2.11, we have

$$\min\{|d_j| : d_j \neq 0\} \geq \frac{1}{n\bar{\chi}(A)}.$$

Now,

$$d^T \bar{s} = d^T(\lceil pc \rceil - pc) + pc^T d - d^T A^T \bar{y} < 1.$$

For each $j \notin J, \bar{s}_j \geq n\bar{\chi}(A)$, and so if $d_j > 0$, then $d_j \bar{s}_j \geq 1$, which contradicts $d^T \bar{s} < 1$. Therefore, $d_j = 0$ for all $j \notin J$. So d_J satisfies $d_J \geq 0, A_J d_J = 0$ and $c_J^T d_J < 0$. Hence by Farkas' lemma, $A_J^T y \leq c_J$ has no solution. □

If $A_J^T y \leq c_J$ has no solution, then we have a $d_J \geq 0$ such that $A_J d_J = 0$ and $c_J^T d_J < 0$. By inserting zero(es) to d_J, we have a $d \geq 0$ such that $Ad = 0$ and $c^T d < 0$. This is a certificate of the infeasibility of $A^T y \leq c$.

Therefore, we can repeat this algorithm with the data (A_J, c_J). Since we remove at least one column from A to get A_J in each iteration, the algorithm will terminate in at most n iterations.

We now look at the complexity of running the above algorithm. In each iteration, we solve $A^T y \leq \lceil pc \rceil$. Note that

$$||\lceil pc \rceil||_\infty = \lceil ||pc||_\infty \rceil = p,$$

and

$$\min_{\lceil pc_j \rceil \neq 0} |\lceil pc_j \rceil| \geq 1.$$

Therefore, by using the proof of Proposition 7.1 for the case $b = 0$, we have proven that if Assumption 2 holds, we can solve $A^T y \leq \lceil pc \rceil$ in at most poly$(n, \log(\bar{\chi}(A)), \log(p))$ elementary arithmetic operations. Here we use Proposition 2.4 repeatedly to conclude that $\bar{\chi}(A_J) \leq \bar{\chi}(A)$ in every iteration.

Proposition 7.4 *Suppose Assumption 2 holds and that we are given an integer $p \geq 2n^{3/2}(\bar{\chi}(A))^2$. Then in at most* poly$(n, \log(\bar{\chi}(A)), \log(p))$ *elementary arithmetic operations, we can determine whether $A^T y \leq c$ has a solution, and if not, find a certificate of its infeasibility.*

Similarly we have the following result, in which we use the algorithm and the analysis in Subsection 5.1 and the relation (10).

Proposition 7.5 *Suppose we are given p, an integer power of 2, that is at least as large as $2n^{3/2}(\bar{\chi}(A))^2$. Then in at most* poly$(n, |\log(\delta_\delta(A))|, \log(\Delta(A)/\delta(A)), \log(p))$ *elementary arithmetic operations, we can determine whether $A^T y \leq c$ has a solution, and if not, find a certificate of its infeasibility.*

7.3 Main Results

From now on, we assume that $c \in \mathbb{R}^n \setminus \{0\}$, and $b \in \mathbb{R}^m \setminus \{0\}$.

Proposition 7.6 *Suppose Assumption 2 holds, (D) is feasible and we are given an integer $p \geq 2n^{3/2}(\bar{\chi}(A))^2$. Then in at most* $\operatorname{poly}(n, \log(\bar{\chi}(A)), \log(p))$ *elementary arithmetic operations, we can either:*

(i) find z such that $A^T z = c$, or
(ii) detect that (D) is unbounded, or
(iii) find an inequality $a^T y \leq \gamma$ in $A^T y \leq c$ such that $a^T y^* < \gamma$ for some optimal solution y^* of (D).

Proof
Let z be the (unique) minimizer of $\|A^T z - c\|$. z can be computed by solving $AA^T z = Ac$ using a good implementation of Gaussian elimination, in $\operatorname{poly}(n)$ elementary arithmetic operations. Let $c' := c - A^T z$. If $c' = 0$, then we have found z that satisfies condition (i) above. So we assume $c' \neq 0$. Let

$$c'' := \frac{p}{\|c'\|_\infty} c'.$$

Note that $A^T y \leq c''$ arises from $A^T y \leq c$ by a translation and a scaling. Hence maximizing $b^T y$ over $A^T y \leq c$ is equivalent to maximizing $b^T y$ over $A^T y \leq c''$ in the sense that y^* is an optimal solution of $\max\{b^T y : A^T y \leq c\}$ if and only if $(p/\|c'\|_\infty)(y^* - z)$ is an optimal solution of $\max\{b^T y : A^T y \leq c''\}$. Also note that $c'' \in \mathcal{N}(A)$, since c' is.
Now we solve the problem $(D') : \max\{b^T y : A^T y \leq \lceil c'' \rceil\}$. Note that (D') is feasible since (D) is and $\{y : A^T y \leq c''\} \subseteq \{y : A^T y \leq \lceil c'' \rceil\}$. Also, (D') is unbounded if and only if (D) is unbounded because each of these is true if and only if there exists $d \neq 0$ such that $A^T d \leq 0$ and $b^T d > 0$. Hence condition (ii) is satisfied. We can now assume both (D) and (D') are bounded. Let $(\bar{y}, \bar{s})$ be an optimal solution of (D'). We have by (17) that $\|\bar{s}\|_\infty \geq n\bar{\chi}(A)$. Corollary 6.4 implies that there exists an optimal solution $(\bar{y}', \bar{s}')$ of $\max\{b^T y : A^T y \leq c''\}$ such that

$$\|\bar{s} - \bar{s}'\|_\infty \leq \left[\sqrt{m}\bar{\chi}(A) + 1\right] \|c'' - \lceil c'' \rceil\|_\infty < \sqrt{m}\bar{\chi}(A) + 1.$$

Therefore, we pick the inequality with the largest $\bar{s}_j$ among the inequalities $A^T y \leq \lceil c'' \rceil$, and condition (iii) is satisfied.
We now look at the complexity of solving (D') using Proposition 7.1. We have

$$\|\lceil c'' \rceil\|_\infty = \lceil \|c''\|_\infty \rceil = p.$$

Also, $\lceil c'' \rceil \neq 0$ and $\min_{\lceil c''_j \rceil \neq 0} |\lceil c''_j \rceil| \geq 1$. So,

$$\log\left(\frac{\|\lceil c'' \rceil\|_\infty}{\min_{\lceil c''_j \rceil \neq 0} |\lceil c''_j \rceil|}\right) \leq \log(p+1);$$

therefore, the required time bound is satisfied. □

Proposition 7.7 *Suppose Assumption 2 holds and that we are given an integer $p \geq 2n^{5/2}m\left(\frac{\Delta(A)}{\delta(A)}\right)^2$. Then we can find a solution of the system $A^Ty \leq c$ or a certificate of its infeasibility in at most*

$$\text{poly}\left(n, \log\left(\frac{\Delta(A)}{\delta(A)}\right), \log(p)\right)$$

elementary arithmetic operations.

Proof
Let $\hat{b} := A\left(p+1, (p+1)^2, \cdots, (p+1)^n\right)^T$. We apply Proposition 7.4 to test whether $(\hat{D})$: $\max\{\hat{b}^Ty : A^Ty \leq c\}$ has a feasible solution, and if not, we obtain a certificate of its infeasibility. Therefore, we assume that $(\hat{D})$ is feasible. Since $(\hat{D})$ is not unbounded (by construction of $\hat{b}$), $(\hat{D})$ has optimal solution(s).

Suppose $\hat{b}$ is a linear combination of fewer than m columns of A. Then there exists an $m \times (m-1)$ submatrix C of A of rank $m-1$, so that the matrix $[C|\hat{b}]$ is singular. Hence,

$$\begin{aligned} 0 &= \det[C|\hat{b}] \\ &= (p+1)\det[C|A_1] + (p+1)^2\det[C|A_2] + \cdots + (p+1)^n\det[C|A_n], \end{aligned}$$

where A_j denotes the jth column of A. Suppose $\det[C|A_j] \neq 0$ for some j. Let k be the largest j such that $\det[C|A_j] \neq 0$. Then

$$\begin{aligned} 0 &= \sum_{j=1}^{k-1}\left[(p+1)^j(\pm\det[C|A_j])\right] + (p+1)^k|\det[C|A_k]| \\ &\geq -\Delta(A)\sum_{j=1}^{k-1}(p+1)^j + (p+1)^k\delta(A) \\ &= -\Delta(A)(p+1)\frac{(p+1)^{k-1}-1}{(p+1)-1} + (p+1)^k\delta(A) \\ &= (p+1)^k\left(\delta(A) - \frac{\Delta(A)}{p}\right) + \frac{\Delta(A)(p+1)}{p} > 0, \end{aligned}$$

since

$$p \geq \frac{\Delta(A)}{\delta(A)}.$$

This gives a contradiction. So $\det[C|A_j] = 0$ for all $j \in \{1, \ldots, n\}$, contradicting the fact that A has rank m. So $\hat{b}$ is not a linear combination of fewer than m columns of A. Therefore, $(\hat{D})$ is attained at a unique minimal face.
We now apply Proposition 7.6 to $(\hat{D})$. If it returns a z such that $A^T z = c$, we stop. Otherwise, we have an inequality $a^T y \leq \gamma$ in $A^T y \leq c$ such that $a^T y^* < \gamma$ for some optimal solution y^* of $(\hat{D})$. Let $\tilde{A} \in \mathbb{R}^{m \times (n-1)}$ be A with the column a removed, and $\tilde{c} \in \mathbb{R}^{(n-1)}$ be c with the corresponding entry γ removed. We then solve the more relaxed problem $\max\{\hat{b}^T y : \tilde{A}^T y \leq \tilde{c}\}$ and repeat the above. Note that $\tilde{A}$ must have full row rank in order to apply Proposition 7.6 to the new relaxed problem. So we perform the following procedures to reformulate this problem. We do Gaussian elimination to eliminate any redundant row of $[\tilde{A}|\hat{b}]$ to get $[\bar{A}|\bar{b}]$. Now,

$$\begin{aligned} \max\{\hat{b}^T y : \tilde{A}^T y \leq \tilde{c}\} &= \min\{\tilde{c}^T \tilde{x} : \tilde{A}\tilde{x} = \hat{b}, \tilde{x} \geq 0\} \\ &= \min\{\tilde{c}^T \tilde{x} : \bar{A}\tilde{x} = \bar{b}, \tilde{x} \geq 0\} \\ &= \max\{\bar{b}^T \bar{y} : \bar{A}^T \bar{y} \leq \tilde{c}\}. \end{aligned}$$

It is not hard to see that the first problem (and hence all of them) has an optimal solution (so the equations above are justified). Since the system $\tilde{A}\tilde{x} = \hat{b}$ is consistent, $\bar{A}$ must have full row rank. So we apply Proposition 7.6 to the last problem above. If it returns a $\bar{z}$ such that $\bar{A}^T \bar{z} = \tilde{c}$, then $\tilde{A}^T z = \tilde{c}$, where z is obtained from $\bar{z}$ by adding a zero entry in the place that corresponds to the redundant row of $\tilde{A}$ being eliminated earlier. Otherwise, it returns an inequality $\bar{a}^T \bar{y} \leq \tilde{\gamma}$ in $\bar{A}^T \bar{y} \leq \tilde{c}$ such that $\bar{a}^T \bar{y}^* < \tilde{\gamma}$ for some optimal solution $\bar{y}^*$. Let y^* be obtained from $\bar{y}^*$ by adding a zero entry as before. Then y^* is an optimal solution of $\max\{\hat{b}^T y : \tilde{A}^T y \leq \tilde{c}\}$ because $\tilde{A}^T y^* = \bar{A}^T \bar{y}^* \leq \tilde{c}$ and $\hat{b}^T y^* = \bar{b}^T \bar{y}^*$. Also, $\tilde{a}^T y^* = \bar{a}^T \bar{y}^* < \tilde{\gamma}$.
Note that for each submatrix C of A, we have (using Proposition 2.14),

$$2n^{3/2} \left(\bar{\chi}(C)\right)^2 \leq 2n^{5/2} m \left(\frac{\Delta(C)}{\delta(C)}\right)^2 \leq p.$$

Hence p satisfies the supposition of Proposition 7.6 *every* time it is being called.
By repeatedly applying Proposition 7.6, we obtain an ordering of the inequalities in $A^T y \leq c$, say, $\alpha_1^T y \leq \gamma_1$, $\alpha_2^T y \leq \gamma_2, \ldots, \alpha_n^T y \leq \gamma_n$, such that for some $r, 1 \leq r \leq n-1$, and some $z \in \mathbb{R}^m$:

- $\alpha_j^T z = \gamma_j$, for all $r + 1 \leq j \leq n$,
- for each $1 \leq j \leq r$, $\alpha_j^T y^j < \gamma_j$ for some optimal solution y^j of $\max\{\hat{b}^T y : \alpha_k^T y \leq \gamma_k, \forall k \geq j\}$.

That is, we run Proposition 7.6 r times, by removing one inequality each time from $A^Ty \leq c$ until we find a z that satisfies the remaining inequalities as equalities. Since the maximum is attained at a unique minimal face, the optimal solution set can be written as

$$\{y : A_=^T y = c_=\} = \{y : A_=^T y = c_=, A_<^T y < c_<\},$$

where $([A_<^T|c_<], [A_=^T|c_=])$ is a row-partition of $[A^T|c]$. It is easy to see that the rows of $A_<^T$ are precisely $\{a_j^T : 1 \leq j \leq r\}$, whereas the rows of $A_=^T$ are precisely $\{a_j^T : r+1 \leq j \leq n\}$. So $A_=^T z = c_=$, which implies $A_<^T z < c_<$. Therefore, z is a feasible solution of $(\hat{D})$.

We now look at the complexity of the above algorithm. We apply Proposition 7.4 once to $(\hat{D})$, which takes time

$$\text{poly}(n, \log(\bar{\chi}(A)), \log(p)) \leq \text{poly}\left(n, \log\left(\frac{\Delta(A)}{\delta(A)}\right), \log(p)\right),$$

by (10).

Afterwards, we apply Proposition 7.6 at most n times. At the kth time $(1 \leq k \leq r+1)$, Proposition 7.6 takes at most $\text{poly}(n, \log(\bar{\chi}(A^{(k)})), \log(p))$ elementary arithmetic operations, where $A^{(1)} := A$ and for $k \geq 2$, $A^{(k)}$ is obtained by first removing some column of $A^{(k-1)}$, and then removing any redundant row. By Proposition 2.4, we have $\bar{\chi}(A^{(k)}) \leq \bar{\chi}(A^{(k-1)}) \leq \bar{\chi}(A)$, for all $k \geq 2$, and we can again use (10). □

Theorem 7.8 *If Assumption 2 holds, then we can solve the primal-dual LP problems*

$$(P) : \min\{c^T x : Ax = b, x \geq 0\} \textit{ and } (D) : \max\{b^T y : A^T y \leq c\}$$

in at most $\text{poly}\left(n, \log\left(\frac{\Delta(A)}{\delta(A)}\right)\right)$ *elementary arithmetic operations.*

Proof

Suppose we are given an integer $p \geq \bar{p}$, where $\bar{p}$ is defined in (12). We first describe an algorithm for solving the given LPs, and later explain how to obtain such a p. We apply Proposition 7.7 to test if $\{A^T y \leq c\}$ and $\{Ax = b, x \geq 0\}$ are feasible, where the latter is the same as

$$\left\{\begin{pmatrix} A \\ -A \\ -I \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \\ 0 \end{pmatrix}\right\}.$$

(To use Proposition 7.7 for the above displayed data, we apply Propositions 2.14 and 2.4 to the matrix $[A^T| - A^T| - I]$ and note that $\bar{p}$ is large enough for the application of Proposition 7.7—and the results it uses—to this matrix too.) If one of them is infeasible, then we stop (having determined the status of each problem). Therefore, we may assume that both (P) and (D) are feasible.

By repeated application of Proposition 7.6 (as in the proof of Proposition 7.7, and we again have $2n^{3/2}\left(\bar{\chi}(C)\right)^2 \leq p$, for all submatrices C of A), we can split $\{A^T y \leq c\}$ into $\{A^T_{(1)}y \leq c_{(1)}, A^T_{(2)}y \leq c_{(2)}\}$ and find a vector z, such that $A^T_{(2)}z = c_{(2)}$ and $A^T_{(1)}y^* < c_{(1)}$ for some optimal solution y^* of $\max\{b^T y : A^T y \leq c\}$. Let $(x^T_{(1)}, x^T_{(2)})^T$ be a partition of any primal solution x such that $x^T_{(1)}$ corresponds to $A^T_{(1)}$ and $x^T_{(2)}$ corresponds to $A^T_{(2)}$. Hence every primal optimal solution x satisfies $x_{(1)} = 0$. So,

$$\begin{aligned}\min\{c^T x : Ax = b, x \geq 0\} &= \min\{c^T_{(2)}x_{(2)} : A_{(2)}x_{(2)} = b, x_{(2)} \geq 0\}\\ &= \max\{b^T y : A^T_{(2)}y \leq c_{(2)}\}.\end{aligned}$$

Using Proposition 7.7, we can find a feasible solution $x^*_{(2)}$ of

$$\left\{\begin{pmatrix} A_{(2)} \\ -A_{(2)} \\ -I \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \\ 0 \end{pmatrix}\right\}.$$

Then $c^T_{(2)}x^*_{(2)} = z^T A_{(2)}x^*_{(2)} = b^T z$, and by LP duality, $x^*_{(2)}$ is an optimal solution of $\min\{c^T_{(2)}x_{(2)} : A_{(2)}x_{(2)} = b, x_{(2)} \geq 0\}$. Let $x^*_{(1)} := 0$. Then x^* is an optimal solution of $\min\{c^T x : Ax = b, x \geq 0\}$.

Let $A^T_{(3)}y \leq c_{(3)}$ be the subsystem of $A^T_{(2)}y \leq c_{(2)}$ corresponding to the positive components of $x^*_{(2)}$. By complementary slackness, it follows that $\{y : A^T y \leq c, A^T_{(3)}y = c_{(3)}\}$ is the set of optimal solutions of $\max\{b^T y : A^T y \leq c\}$. We can use Proposition 7.7 to find such a solution.

As in the proof of Proposition 7.7, identifying the partition $[A_{(1)}|A_{(2)}]$ of A takes at most poly $\left(n, \log\left(\frac{\Delta(A)}{\delta(A)}\right), \log(p)\right)$ elementary arithmetic operations. Also, note that Δ/δ values for $[A^T|-A^T|-I]$, $[A^T_{(2)}|-A^T_{(2)}|-I]$, $[A|A_{(3)}|-A_{(3)}]$ are all bounded by $\frac{\Delta(A)}{\delta(A)}$. Therefore, the algorithm terminates in at most poly $\left(n, \log\left(\frac{\Delta(A)}{\delta(A)}\right), \log(p)\right)$ elementary arithmetic operations.

The correctness of the above algorithm is guaranteed by the assumption that $p \geq \bar{p}$. Without a prior knowledge of $\left(\frac{\Delta(A)}{\delta(A)}\right)$, we will use the following *"log-squaring trick"*. (Similar tricks have been used before for similar purposes; see [29].) Initially, we can guess n for the value of $\log\left(\frac{\Delta(A)}{\delta(A)}\right)$ and run the above algorithm so that our initial p is roughly $2(2m+n)^{3/2}(2mn+1)2^{2n}$. If the algorithm fails, we replace the current guess by its square, update p, and repeat the algorithm. We also check the output of the above algorithm. If it concludes that (P) (or (D)) is infeasible, we use the corresponding infeasibility certificate to ensure that (P) (or (D)) is indeed infeasible. Similarly, if the algorithm returns a primal-dual "optimal" solution pair, we use complementary slackness conditions to ensure it is indeed optimal. All of these can be done efficiently. If any of the output is false, we again square the most recent guess for $\log\left(\frac{\Delta(A)}{\delta(A)}\right)$, update p, and repeat the algorithm. It is easy to show that after

$$O\left(\log\left(\frac{\log\log\left(\frac{\Delta(A)}{\delta(A)}\right)}{\log(n)}\right)\right)$$

guesses, we have the current guess for p between $\bar{p}$ and $\tilde{p}$. (Here we assume that $\log\log\left(\frac{\Delta(A)}{\delta(A)}\right) \geq 2\log(n)$; otherwise, our first or second guess works and no additional iterations are necessary.) Also, clearly all the guesses for p is at most $\tilde{p}$; moreover, $\log(\tilde{p}) = O\left(\text{poly}\left(n, \log\left(\frac{\Delta(A)}{\delta(A)}\right)\right)\right)$. Therefore, the claimed overall complexity bound is established. □

Note that in the proof of the above theorem, one cannot increase the size of the guess significantly faster than we did, since the sizes of all the integers used by our algorithm must be bounded by a polynomial function of the sizes of the complexity measures we are using.

Theorem 7.9 *We can solve the primal-dual LP problems*

$$(P) : \min\{c^T x : Ax = b, x \geq 0\} \textit{ and } (D) : \max\{b^T y : A^T y \leq c\}$$

by utilizing the LP solver subroutine of Subsection 5.1 $O(n^2)$ times and therefore in at most $\text{poly}\left(n, |\log(\delta_\delta(A))|, \log\left(\frac{\Delta(A)}{\delta(A)}\right)\right)$ *elementary arithmetic operations.*

Proof

We assume that we are given an integer $p \geq \bar{p}$. (We can remove this assumption as in the proof of Theorem 7.8, by applying a *log-squaring trick.*) First we check the feasibility of (P) and (D) using Proposition 7.5 and the underlying algorithm. If any of (P), (D) is infeasible, we have the certificates of such fact and we are done. So, we assume that both (P) and (D) have feasible solutions. Then we apply the proof of Proposition 7.6 to (D) and have the problem

$$(D') : \max\{b^T y : A^T y \leq \lceil c'' \rceil\}.$$

Our theorem in Subsection 5.1 cannot deal with this LP problem (since the objective function of (D') is arbitrary). We form the dual (call it (P')) of (D') and apply the proof of Proposition 7.6 to (P'). Now, the LP problems arising from the applications of Proposition 7.6 to (P') all satisfy the conditions needed in Subsection 5.1 (namely, condition (ii) of the subsection for b and c). So, calling this subroutine $O(n)$ times, as in the proof of Theorem 7.8, we can compute optimal solutions of (P') and (D'). (At some point, during this process, inside the proof of Theorem 7.8, the method in the proof of Proposition 7.7 is used. This requires the LP solver subroutine to be called with data satisfying condition (i)—potentially not satisfying condition (ii)—of Subsection 5.1.) Now, we have an optimal solution of (D') and we can keep applying this technique in using the proof of Theorem 7.8 to solve (P) and (D). This clearly requires no more than $O(n)$ problems of the type (D') to be solved. Since each such problem can be solved with $O(n)$ calls to the LP solver subroutine, the $O(n^2)$ bound follows. □

7.4 Overall Complexity Bounds

Suppose we have an interior-point algorithm satisfying Assumption 2, with an $O\left(n^{\alpha}(\log(\bar{\chi}(A)))^{\beta}\right)$ iteration bound, for some $\alpha \geq 0$, $\beta \geq 0$. Then Theorem 7.8 implies an iteration bound of

$$O\left(n^{1+\alpha}\left[\log\left(\frac{\Delta(A)}{\delta(A)}\right)+\log(n)\right]^{\beta}\left(\log\left(\frac{\log\log\left(\frac{\Delta(A)}{\delta(A)}\right)}{\log(n)}\right)\right)\right).$$

On the other hand, using the methods of Subsection 5.1 and Theorem 7.9, we obtain the iteration bound

$$O\left(n^{2.5}\left(|\log(\delta_\delta(A))|+n\log\left(\frac{\Delta(A)}{\delta(A)}\right)+n\log(n)\right)\right.$$
$$\left.\left(\log\left(\frac{\log\log\left(\frac{\Delta(A)}{\delta(A)}\right)}{\log(n)}\right)\right)\right).$$

The above bound is not better than Vavasis-Ye's and can be much worse in general. However, in the case that A is totally unimodular, it becomes the same. In this very special case, we can omit the factor of $(\log\log(\bar{\chi}(A)))$

(caused by a *log-squaring* type trick) in the iteration bound of Vavasis-Ye algorithm. See, for instance, Proposition 7.10 and the discussion following it. In the case that A is integral, the bounds can be considered close. See below.

7.5 Integer Data and Network Flow Problems

- Integer Data:

  When the data is integer, $\delta(A) = 1$, $\delta_\delta(A) \geq \frac{1}{\Delta(A)}$ and $\log(\Delta(A)) \leq n\log(n) + \text{size}(A)$. Therefore, we have Tardos' theorem as a special case. Also, in this case it is very easy to get upper bounds (whose sizes are bounded by polynomial functions of the input size) for $\tilde{p}$ so that the multiplicative factor $\left(\log\left(\frac{\log\log\left(\frac{\Delta(A)}{\delta(A)}\right)}{\log(n)}\right)\right)$ in the complexity bound can be removed.

- Totally Unimodular Matrix A:

  Recall that a matrix is *totally unimodular* if all of its square submatrices have determinants -1, 0 or 1. That is, $\delta_\delta(A) = \delta(A) = \Delta(A) = 1$. The following is special case of Proposition 2.14.

Proposition 7.10 *(Ho* [11]*) Let $A \in \Re^{m\times n}$ be a full row rank totally unimodular matrix. Then $\bar{\chi}(A) \leq \sqrt{mn}$.*

Proof
Take any basis B of A. It is elementary to show that $A_B^{-1}A$ is also totally unimodular. Then for all x such that $||x||_2 = 1$,

$$||A_B^{-1}Ax||_2 = \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}(A_B^{-1}A)_{ij}x_j\right)^2} \leq \sqrt{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|x_j|\right)^2}$$

$$= \sqrt{\sum_{i=1}^{m}(||x||_1)^2} \leq \sqrt{mn},$$

because $\max\limits_{||x||_2=1} ||x||_1 = \sqrt{n}$ when $x = \frac{1}{\sqrt{n}}e$. Therefore $\bar{\chi}(A) \leq \sqrt{mn}$ by Proposition 2.3. □

In fact we can exhibit a totally unimodular matrix A with $\bar{\chi}(A) = \Theta(\sqrt{mn})$. Consider the complete graph on vertices $\{1, \ldots, m+1\}$, with arcs ij if $i < j$. Let A be its node-arc incidence matrix, with any one row deleted. Then A is a totally unimodular $m \times n$ full row rank matrix, where $n = m(m+1)/2$. It can be easily shown that if we choose $x = e$ and B such that the columns of A_B correspond to a spanning tree that is also a path, i.e., a Hamiltonian path with the correspoding incidence matrix:

$$A_B = \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & -1 & \\ & & & & & 1 \end{pmatrix},$$

then $(A_B^{-1}Ax)_j = j(m-j+1)$. Therefore

$$\bar{\chi}(A) \geq \sqrt{\frac{\sum_{j=1}^{m} j^2(m-j+1)^2}{\frac{m(m+1)}{2}}} = \Theta(m^{1.5}) = \Theta(\sqrt{mn}).$$

Therefore, the upper bound proven in Proposition 7.10 is tight up to the order.

Note that we used above, the fact that A_B^{-1} is the all ones upper-triangular matrix. As it is well-known, for every $B \in \mathcal{B}(A)$, there exist permutations of the rows and the columns of A_B such that the resulting matrix is upper-triangular. Since A_B^{-1} is also totally unimodular, it can only have $-1, 0, 1$ entries. Therefore, in this special setting, $B \in \mathcal{B}(A)$, corresponding to Hamiltonian paths, maximize $\|A_B^{-1}\|$.

- Minimum Cost Flow Problems:

  Consider the minimum cost flow problem with the constraints $Ax = b$ and $0 \leq x \leq u$, where A is the node-arc incidence matrix of a given directed graph with any one row deleted (so that it has full row rank). By introducing the slack variables v, we convert the constraints into standard equality form:

$$\hat{A}\begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} b \\ u \end{pmatrix}, \begin{pmatrix} x \\ v \end{pmatrix} \geq 0,$$

  where

$$\hat{A} := \begin{pmatrix} A & 0 \\ I & I \end{pmatrix}.$$

This structure arises whenever we convert an upper bounded LP problem to the standard equality form. Vavasis and Ye [29] prove that $\bar{\chi}(\hat{A}) = O(mn)$. Using Propositions 2.7 and 7.10 (and the arguments following that), we have $\bar{\chi}(\hat{A}) = \Theta(\sqrt{mn})$ when A is totally unimodular.

Acknowledgment

Part of this work was done while both authors were visiting The Fields Institute, Toronto, Ontario, Canada (Fall, 1999). Research of the first author was supported in part by an OGSST, an NSERC PGSB, and by an NSERC research grant and a PREA of the second author. Research of the second author was supported in part by a Faculty Research Grant from University of Waterloo, research grants from NSERC, and a PREA from Ontario, Canada.

Part of this work was done while the second author was visiting the Department of Mathematical and Computing Sciences of Tokyo Institute of Technology on a sabbatical leave from the University of Waterloo (October 1997–March 1998). Support and hospitality of Masakazu Kojima and T.I.T. are gratefully acknowledged.

References

1. I. Adler and P.A. Beling, Polynomial algorithms for linear programming over the algebraic numbers, Algorithmica **12**, 436–457 (1994).
2. L. Blum, F. Cucker, M. Shub and S. Smale, *Complexity and Real Computation*, Springer-Verlag, New York, NY, U.S.A., 1998.
3. W. Cook, A.M.H. Gerards, A. Schrijver and É. Tardos, Sensitivity theorems in integer programming, Mathematical Programming A **34**, 251-264 (1986).
4. I.I. Dikin, On the speed of an iterative process, Upravlaemye Sistemy **12**, 54–60 (1974).
5. M. Epelman, *Complexity, Conditions Numbers, and Conic Linear Systems*, Ph. D. Thesis, Massachusetts Institute of Technology, 1999.
6. M. Epelman and R.M. Freund, Pre-conditioners and relations between different measures of conditioning for conic linear systems, *Working Paper OR-344-001*, Operations Research Center, M.I.T., 2000.
7. A. Forsgren, On linear least-squares problems with diagonally dominant weight matrices, SIAM J. Matrix Anal. Appl. **17**, 763–788 (1996).
8. C.C. Gonzaga and H. Lara, A note on properties of condition numbers, Linear Algebra And Its Applications **261**, 269–273 (1997).
9. O. Güler, A.J. Hoffman and U.G. Rothblum, Approximations to solutions

to systems of linear inequalities, SIAM J. Matrix Anal. Appl. **16**, 688–696 (1995).

10. O. Güler ad Y. Ye, Convergence behavior of interior-point algorithms, Mathematical Programming A **60**, 215–228 (1993).
11. J.C.K. Ho, *Structure of the Central Path and Related Complexity Issues for lLnear Programming*, M. Math. Thesis, University of Waterloo, December 1998.
12. A.J. Hoffman, On approximate solutions of systems of linear inequalities, J. Res. Nat. Bur. Stand. **49**, 263–265 (1952).
13. N. Karmarkar, A new polynomial time algorithm for linear programming, Combinatorica **4**, 373–395 (1984).
14. L. Khachiyan, On the complexity of approximating extremal determinants in matrices, Journal of Complexity **11**, 138–153 (1995).
15. N. Megiddo, S. Mizuno and T. Tsuchiya, A modified layered-step interior point algorithm for linear programming, Mathematical Programming A **82**, 339-355 (1998).
16. S. Mehrotra and Y. Ye, Finding an interior point in the optimal face of linear programs, Mathematical Programming A **62**, 497–515 (1993).
17. D.P. O'Leary, On bounds for scaled projections and pseudoinverses, Linear Algebra And Its Applications **132**, 115–117 (1990).
18. J. Renegar, Linear programming, complexity theory and elementary functional analysis, Mathematical Programming A **70**, 279–351 (1995).
19. C. Roos, T. Terlaky and J.-Ph. Vial, *Theory and algorithms for linear optimization. An interior point approach*, Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley and Sons, Ltd., Chichester, 1997.
20. A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley and Sons, 1986.
21. G.W. Stewart, On scaled projections and pseudoinverses, Linear Algebra And Its Applications **112**, 189–193 (1989).
22. É. Tardos, A strongly polynomial algorithm to solve combinatorial linear programs, Operations Research **34**, 250–256 (1986).
23. É. Tardos, A strongly polynomial mimimum cost circulation algorithm, Combinatorica **5**, 247–255 (1985).
24. M.J. Todd, L. Tunçel and Y. Ye, Probabilistic analysis of two complexity measures for linear programming problems, Mathematical Programming A **90**, 59–69 (2001).
25. L. Tunçel, Approximating the complexity measure of Vavasis-Ye algorithm is NP-hard, Mathematical Programming A **86**, 219–223 (1999).
26. L. Tunçel, On the condition numbers for polyhedra in Karmarkar's form,

Operations Research Letters **24**, 149–155 (1999).

27. L. Tunçel, A pseudo-polynomial complexity analysis for interior-point algorithms, *Dept. of Comb. and Opt. Research Report* 93-16 (June 1993, Revised November 1993).
28. S.A. Vavasis, Stable numerical algorithms for equilibrium systems, SIAM J. Matrix Anal. Appl. **15**, 1108–1131 (1994).
29. S.A. Vavasis and Y. Ye, A primal-dual interior point method whose running time depends only on the constraint matrix, Mathematical Programming A **74**, 79–120 (1996).
30. S.A. Vavasis and Y. Ye, Condition numbers for polyhedra with real number data, Operations Research Letters **17**, 209–214 (1995).
31. Y. Ye, Toward probabilistic analysis of interior-point algorithms for linear programming, Mathematics of Operations Research **19**, 38–52 (1994).
32. Y. Ye, M.J. Todd and S. Mizuno, An $O(\sqrt{n}L)$-iteration homogeneous and self-dual linear programming algorithm, Mathematics of Operations Research **19**, 53-67 (1994).
33. S. Zhang, Global error bounds for convex conic problems, SIAM J. Optimization **10**, 836–851 (2000).

ON THE EXPECTED NUMBER OF REAL ROOTS OF A SYSTEM OF RANDOM POLYNOMIAL EQUATIONS

ERIC KOSTLAN

eric@developmentserver.com

We unify and generalize several known results about systems of random polynomials. We first classify all orthogonally invariant normal measures for spaces of polynomial mappings. For each such measure we calculate the expected number of real zeros. The results for invariant measures extend to underdetermined systems, giving the expected volume for orthogonally invariant random real projective varieties. We then consider noninvariant measures, and show how the real zeros of random polynomials behave under direct sum, tensor product and composition.

Part I – Introduction

1 Overview

To motivate our investigation, we begin this chapter with some known results about systems of random polynomials. In particular, we discuss results of Mark Kac [12], Edelman and Kostlan [7], Shub and Smale [23], and Rojas [21]. We conclude Part I with a detailed discussion of the level of generality adopted in this chapter. The titles for Part II and Part III were chosen to indicate something about the level of generality of the results in each respective section.

Part II is devoted to random polynomials whose coefficient vectors form orthogonally invariant normal measures. We classify all such measures. We then consider a systems of such polynomials. The system may be mixed – that is, the polynomials making up the system do not have to be identical – and the system may be underdetermined. In all cases we calculate the expected volume of the corresponding random real projective variety. For a completely determined system, this is simply the expected number of real zeros.

In Part III of this chapter we consider the direct sum, tensor product and composition of random polynomials. This is a familiar theme in mathematics. We have systems of random polynomials, and we wish to use them to construct new systems of random polynomials. Our focus will be on how the expected number of real zeros behaves under sum, product and composition.

This chapter is not self-contained. It is written with [7], Sections 1-4 and Section 7, as a necessary prerequisite. Although this chapter does discuss volumes of random varieties, the focus is on the expected number of real zeros of systems. A more detailed discussion of random real varieties is in preparation [15]. An open-ended web-based project, [15] completes a series of five papers: [13], [14], [7], this paper, and [15]. All these papers are available at http://www.developmentserver.com/randompolynomials. One result appearing in [13] and [14] that might be of particular interest to readers of this chapter is the calculation (over the complexes) of the joint distribution of zeros for pairs of plane conics. For most readers interested in random polynomials, [7] is sufficient.

2 Summary of previous results

2.1 Polynomials with independent standard normal coefficients

Consider a system of m polynomials in m variables with independent standard normal coefficients. Assume that each variable has degree at most d. Therefore, the Newton polytope, that is, the convex hull of the support of each polynomial, is an m-dimensional hypercube. If $E_d^{(m)}$ represents the expected number of zeros for the system,

$$E_d^{(m)} = \pi^{-\frac{m+1}{2}} \Gamma\left(\frac{m+1}{2}\right) (\pi E_d^{(1)})^m .$$

As $d \to \infty$,

$$E_d^{(m)} \sim \pi^{-\frac{m+1}{2}} \Gamma\left(\frac{m+1}{2}\right) (2 \log d)^m .$$

The univariate case was established by Kac [12], and generalized to systems of equations in [7]. In Section 10.1 we show that this result may be viewed as a special case of the tensor product of random polynomials.

We conjecture that the same asymptotic formula holds for a large class of monomial term structures, as long as all the coefficients are i.i.d. central normal random variables. This is because for any fixed C, as $d \to \infty$,

$$E_d^{(m)} \sim E_{Cd}^{(m)} .$$

We may choose $C << 1$ and $C >> 1$, to inscribe and circumscribe any given polytope with cubes.

Conjecture *Let K be any bounded set in $\{x \in \mathbf{R} | x \geq 0\}^m$ with nonzero interior. Assume that either K or $\{x \in \mathbf{R} | x \geq 0\}^m - K$ is the finite union of convex sets. Consider a sequence $\{P_d\}$ of completely determined systems of polynomials with independent standard normal coefficients. Assume that the support of P_d is the set of integer lattice points contained in dK. If E_d*

represents the expected number of zeros for the system, then as $d \to \infty$,

$$E_d \sim \pi^{-\frac{m+1}{2}} \Gamma\left(\frac{m+1}{2}\right) (2 \log d)^m \ .$$

The assumption about K is to avoid pathologies and make a precise conjecture. I doubt the correct hypothesis would have anything to do with convexity. I suspect the same asymptotic formula holds for mixed systems as well.

The study of random polynomials began with the assumption that the coefficients were identically distributed. These early works include the 1932 paper by Block and Polya [2], and the seminal work of Mark Kac [12], published in 1943. The extensive literature that has grown out of this work (and this assumption) has been documented by Bharucha-Reid and Sambandham [1].

I feel that the subject of random polynomials is blessed by the fact that these *obvious* random polynomials (polynomials with independent standard normal coefficients) are not the *natural* random polynomials. The attempt to replace these measures with more natural measures was, in part, the motivation for [7], [13], [14], [21], and Part II of this chapter. Furthermore, when we study random polynomials, we can use this case as a convenient non-invariant random polynomial. For example, comparing Section 7.1 of [7] and [21], we see that orthogonal invariance is *not* required to make the tensor product theorem work.

2.2 The most natural random polynomial

Consider a random polynomials

$$\sum_{i_1,\dots,i_m} a_{i_1\dots i_m} \Pi_{k=1}^m x_k^{i_k} \ ,$$

where $\sum_{k=1}^m i_k \leq d$ and where the $a_{i_1\dots i_m}$ are independent normals with mean zero and variances equal to multinomial coefficients:

$$\binom{d}{i_1, \dots, i_m} = \frac{d!}{(d - \sum_{k=1}^m i_k)! \prod_{k=1}^m i_k!} \ .$$

Consider m independent equations of degrees $d_1, \dots, d_m$, each defined this way. Then the expected number of real zeros of the system is

$$\sqrt{\prod_{k=1}^m d_k} .$$

The general case was established by Shub and Smale [23]. For the case where all the degrees are equal, the result was established in [14]. Furthermore, [14] stated the result for underdetermined systems as well, thus giving the expected volume of a real projective variety of dimension k and degree d:

$$\frac{d^{(m-k)/2}\pi^{(k+1)/2}}{\Gamma[(k+1)/2]}.$$

In Part II of this chapter we will unify these results. Following [14], we will consider arbitrary codimension, but like [23], we will avoid the assumption that the polynomials are of equal degree. In this chapter we will refer to any of these results as the **square root result**.

We can characterize these random polynomials using a combination of invariance and independence. See [7], [13] and [14] for detailed discussions. We define the action of the orthogonal group on random polynomials in Section 3.1. The following result, that appears in [14], is true for both real and complex random polynomials.

A central probability measure on a vector space of real polynomial mappings has the following two properties, (1) orthogonal invariance of the measure and (2) statistical independence of the coefficients, iff it is, up to a scalar multiple, the polynomial discussed above. That is, the coefficients must be independent, and the variances of the coefficients are some constant multiplied by the multinomial coefficients.

Notice we are *not* assuming normality of the coefficients. This is one of many theorems in statistics of the form

$$\text{independence} \;+\; \text{invariance} \;\rightarrow\; \text{normallydistributed}$$

In [7] we referred to these polynomials as "a random polynomial with a nice answer" and discussed the geometry of these polynomials in detail. A homogeneous real quadratic system may be written as a real symmetric matrix. For this case, the above result reduces to the characterization of *Gaussian orthogonal ensemble* given by [19].

Over the complex numbers, we may replace the assumption of independence with the assumption of normality of coefficients. This is equivalent to saying that there is (up to a constant) a unique unitarily invariant multivariate normal. In Section 5.1 of this chapter, we give a simple proof of this classical result. The significance of these complex random polynomials has also been noted by physicists [3]. Therefore, it should come as no surprise that the real versions of these random polynomials have nice properties as well.

However, there are other orthogonally invariant normal random polynomials. One goal of this chapter is to give a unified treatment of them. We will produce a complete list of all random polynomials with orthogonally invariant normal coefficients – we say exactly what we mean by orthogonally invariant in Section 3.1. For each we will calculate the expected volume of the real hypersurface determined by the polynomials. We will also establish a product formula that will give us the expected volume of a variety of any codimension generated by a set of (possibly distinct) random polynomials of this type.

2.3 Random harmonic polynomials

Consider the vector space of homogeneous polynomials of degree d in $m+1$ variables that are *harmonic*, that is, the Laplacians of the polynomials are equal to zero. There is, up to a constant, a unique normal measure on harmonic polynomials that is invariant under the orthogonal action defined in Section 3.1. The expected number of real zeros for a system of m such random harmonic polynomial is

$$\left(\frac{d(d+m-1)}{m}\right)^{m/2} .$$

This result appears in [7]. Here are random polynomials with *natural* measures. It seems that the random polynomials described in Section 2.2 have rivals for the status of *most natural random polynomial.* We shall address this issue by considering, in Part II of this chapter, *all* orthogonally invariant normal random polynomials.

2.4 Rojas polynomials

Assume that the support for a random polynomial is a product of simplices

$$\{\prod_{k=1}^{p} z_k^{I_k}; 0 \leq |I_k| \leq \delta_k, 1 \leq k \leq p\} ,$$

where I_k are multi-indices, $|I_k|$ is the sum of the indices of I_K, and $z_k \in \mathbf{R}^{m_k}$. Assume that the coefficients of the polynomials are independent central normals with variances

$$\{\prod_{k=1}^{p} \binom{\delta_k}{I_k}\} .$$

The expected number of real zeros of a system of such polynomials is equal to

$$\frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{\frac{m+1}{2}}} \Pi_{k=1}^{p} \left[\frac{\pi^{\frac{m_k+1}{2}}}{\Gamma\left(\frac{m_k+1}{2}\right)} \sqrt{\delta_k^{m_k}}\right] ,$$

where m is the sum of the m_k. These polynomials were introduced by Rojas in 1996 [21]. In Section 10.1 we see that the product formula used by Rojas can be applied to the tensor product of any central multivariate normal random polynomials.

3 Level of generality

The emphasis of [7] is on generality. For the most part, however, we proved results for polynomial systems with any central multivariate normal coefficients. However, we did call attention to the existence of preferred orthogonally invariant random polynomials with particularly nice properties. This paper expands on both of these subjects. Part II is dedicated to orthogonally invariant random polynomials, with coefficients that are multivariate normal random variables, not necessarily central. Both underdetermined and mixed systems are considered. In Part III we drop invariance and add centrality. In fact, like [7] many of of the results of Part III may be generalized to finite-dimensional space of rectifiable functions. However, in Part III we add the assumption that the system is central. Furthermore, we will not allow mixed or underdetermined systems in Part III.

Throughout this chapter we will allow measures to be restricted to proper subsets of our function space. The domain of our functions may be any measurable subset U of $\mathbf{R}^{m+1}$. Therefore, is some sense, all of our results are local in nature. For simplicity, we will not specify the domain U in each theorem and corollary, as we did in [7]. However, when explicit formulas are given for the expected number of real zeros, we are assuming that the domain is all of $\mathbf{R}^{m+1}$.

Let P be a finite dimensional function space, and consider a (random) $p \in P$, $p : \mathbf{R}^{m+1} \to \mathbf{R}$. The evaluation mapping to the dual of P:

$$ev : R^{m+1} \to P^*,$$

is defined by $ev(x)(f) = f(x)$. The function p thus corresponds to a hyperplane in P^*. The intersection of this (random) hyperplane with ev pulls back to the zero set for p. In P^* integral geometry provides a simple intersection theory to handle (unmixed) systems of equations. This simple theory pulls back to a simple intersection theory in $\mathbf{R}^{m+1}$.

Ideally, we would like to prove every result for systems that are possibly mixed and possibly underdetermined. If the system in underdetermined, we have the problem that the random variety is distorted by the pullback. Thus our intersection theory gives us answers about expected projective volumes in $(ev)(\mathbf{R}^{\mathbf{m}+1})$ instead of in $\mathbf{R}^{m+1}$. For varieties of dimension zero this distortion is not an issue. For mixed systems the picture is complicated by the presence of several different evaluation mapping, and therefore no intersections can be pulled back. We must work with intersections of different random hypersurfaces in $\mathbf{R^{m+1}}$. For mixed *invariant* systems, these random hypersurfaces are orthogonally invariant and therefore classical integral geometry provides simple answers.

3.1 Homogeneous and inhomogeneous systems

Given an inhomogeneous polynomial with m variables $z_1, \ldots, z_m$, we can always consider, instead, the corresponding homogeneous polynomial in $m+1$ homogeneous variables $z_0, \ldots, z_m$. For completely determined systems, this should cause no confusion, because all of our results about expected number of real zeros apply to both homogeneous and inhomogeneous systems. When considering underdetermined systems, we must make sure that the domain is projective space.

When we speak of unitary or orthogonal invariance, we will always be considering the homogeneous version of the random polynomial. We assume the unitary and orthogonal groups act on the right,

$$U(p)(z) \equiv p(U(z)) \ .$$

Actions of products of orthogonal groups have also been considered [17,?,?], but we will not consider them here. We will be studying both inner products and multivariate normal measures on vector spaces of polynomials. Our focus will be on the real case, so we will usually just consider orthogonal invariance. For the complex case we replace orthogonal invariance with unitary invariance.

By convention, we will make no effort to distinguish between an inhomogeneous system and its homogeneous counterpart. For example, by a *univariate polynomial* we will mean a random polynomial that may be written as

$$p(t) = \sum_{i=0}^{d} a_i t^i \quad \textit{or} \quad p(x,y) = \sum_{i=0}^{d} a_i x^i y^{d-i} \ .$$

Elements of the orthogonal group reflect or rotate p to give

$$r^d \sum_{i=0}^{d} a_i \cos^i(\theta+\delta)^i \sin^{d-i}(\theta+\delta) \ ,$$

where $x = r\cos(\theta)$, $y = r\sin(\theta)$. The random polynomial will be said to be *orthogonally invariant* if the probability measure is invariant under this action. Using this convention, the random polynomial

$$p(t) = a(t^2+1) \ ,$$

where a is a standard normal random variable, is considered to be orthogonally invariant.

3.2 Underdetermined and overdetermined systems

For invariant random polynomials we will prove results for underdetermined systems thus giving results about the expected volume of random projective varieties. Several interesting problems suggest themselves when considering random real varieties, and little is known. In this chapter, underdetermined and overdetermined systems, and the corresponding random varieties, are strictly tools to study the expected number of zeros of (possible mixed) systems. We consider only the expected volumes of such varieties, ignoring their more interesting invariants. The key motivation for [15] is to focus on the geometric properties of random real varieties that are ignored in this chapter. For this chapter, we will only need one result about overdetermined systems.

Lemma 3.1 *Let U be a measurable subset of $\mathbf{R}^m$, and consider a real-valued random function*

$$a_0 f_0(t) + a_1 f_1(t) + \ldots + a_n f_n(t), \qquad t \in U,$$

where the a_i are independent standard normals, $n \geq m$. Generate a random variety of dimension m in $\mathbf{R}^{k+1}$, $k \geq m$, by choosing an independent sample of $k+1$ such functions. The expected volume of the projection of this variety onto the unit sphere in $\mathbf{R}^{k+1}$ is equal to the expected number of zeros in U of a system of m independent random functions of this form, multiplied the volume of m-dimensional real projective space

$$\frac{\pi^{\frac{m+1}{2}}}{\Gamma(\frac{m+1}{2})}.$$

Proof Let N be the random variety generated in the lemma, and let M to be the subspace of $\mathbf{R}^{k+1}$, defined by $x_0 = \ldots = x_{m-1} = 0$. The proof is then a straightforward generalization of the proof of Lemma 6.1 in [7]. □

3.3 Mixed systems

I recall having two contrary emotions when Steve Smale told me that he and Mike Shub [23] had generalized the square root result of [14] to mixed completely determined systems. The first was a pleasant lack of surprise. The answer derived by Shub and Smale was the only reasonable result. Furthermore, my research had received an unusually strong endorsement. On the other hand, I was discouraged. I felt that [14] had presented *the correct picture* of random systems. But in [14], a system of equation was a plane in the space of real-valued polynomial mapping, or a point in the Grassmann manifold. Unfortunately, this viewpoint precludes the consideration of mixed systems. Nonetheless, the generalization to mixed systems is clearly desirable. In Part II of this chapter, we will consider invariant mixed systems. We will use the viewpoint of [14] to get results about random hypersurfaces, and then develop an intersection theory similar to [23] to deal with mixed systems. The Grassmann manifold will never be used. We will also allow mixed systems in Section 11, where we discuss the composition of random polynomials.

The generalization of the square root result to mixed systems by Shub and Smale has motivated at least three resent papers. The first of these papers was by Rojas [21]. Rojas attempts to define, for any given support structure, distinguished measures for the corresponding random polynomials. I do not believe this issue has been completely resolved. Rojas also considers mixed systems, and conjectures a relationship between the square root of the mixed volume and the expected number of real zeros. McLennan [18] considers (the multihomogeneous version of) the Rojas polynomials, and shows that the square root of the (normalized) mixed volume provides a lower bound for the expected number of real zeros. In the words of McLennan, "The mean exceeds the square root of the maximum." The third paper, by Malajovich and Rojas [17], produce an explicit, albeit coarse, upper bound for the expected number of real zeros in terms of the square root of the mixed volume for arbitrary central normal random polynomial systems. Reading these three papers, one sees a wealth of new ideas emerging.

Part II – Orthogonally Invariant Normal Coefficients

4 Classification of invariant inner products

We will now classify invariant inner products on vector spaces of homogeneous polynomials. These are the classical results we will need for the rest of the chapter. We include proofs because they are surprisingly easy. Before we try to understand the real case, we discuss the complex case, which is considerably easier.

4.1 The complex analogue

Over the complexes, there is, up to a constant, a unique unitarily invariant Hermitian inner product. This is a well known classical result. The generalization to multihomogenous polynomials may be found in [18].

Theorem 4.1 *Assume a Hermitian inner product is defined on a vector space of complex homogeneous polynomials of degree d, and assume it is invariant under the right action of the unitary group. Then monomials are perpendicular to each other, and the lengths of the monomials $\{x^I\}$ are, up to a constant, square roots of multinomial coefficients*

$$\sqrt{\binom{d}{I}}\,,$$

where I is a multi-index.

Proof This is the corollary following Theorem 3.2 of [13], and is Theorem 4.1 in [14]. We now outline a simplified version of the proof given in [13]. Fix y and think of $v(x)^T Cv(\bar{y})$ as a function of x on $y_\perp$. Assume that for any unitary matrix U, $v(Ux)^T Cv(Uy) = v(x)^T Cv(y)$. Then for any U that fixes y, we must have $v(Ux)^T Cv(y) = v(x)^T Cv(y)$. Thus $v(Ux)^T Cv(y)$ must be constant for x on the unit sphere in $y_\perp$, and therefore it's gradient has rank one. But for an analytic function this implies that $v(x)^T Cv(\bar{y})$ is constant on $y_\perp$, and therefore, by symmetry, zero. Thus

$$x \perp y \ \rightarrow\ v(x)^T Cv(\bar{y}) = 0\ .$$

Therefore $x^T \bar{y}$ divides $v(x)^T Cv(\bar{y})$. We then apply the same argument to the $v(x)^T Cv(\bar{y})\ /\ x^T \bar{y}$, and so on. We ultimately deduce that $v(x)^T Cv(\bar{y}) = a(x^T \bar{y})^d$, which completes the proof. □

We have just established that the space of homogeneous complex valued polynomials of a fixed degree in $m+1$ complex variables is an *irreducible representation* of the unitary group $U(m+1)$. As stated in [14], this may also be shown using classical invariant theory.

4.2 Classification of indefinite inner products

Assume that for any orthogonal matrix Q, $v(Qx)^T Cv(Qy) = v(x)^T Cv(y)$. This implies that $v(x)^T Cv(y)$ must be a polynomial in $x \cdot x$, $x \cdot y$, and $y \cdot y$. This is classical invariant theory. For proofs and discussion of such results, see [24] . We thus deduce that there must exist β_i such that

$$v(x)^T Cv(y) \;=\; \sum_{k=0}^{[d/2]} \beta_k (x \cdot x)^k (y \cdot y)^k (x \cdot y)^{d-2k}. \qquad (1)$$

Thus we have a $[d/2]$ parameter family of orthogonally invariant inner products. For example, if we set $\beta_0 = 1$ and $\beta_k = 0$ for $k > 1$, we produce a inner product for which the monomials are orthogonal.

This completely classified all orthogonally invariant inner products on spaces of homogeneous polynomials of degree d. But this classification is unsatisfactory. We wish to use these inner products as covariance matrices, so we must identify the *positive definite* inner products – actually positive semidefinite, since we allow normal measures restricted to proper subspaces of polynomials. Unfortunately, if we use the parameters β_k it is difficult to see which inner products are positive definite. We therefore will construct a different parameterization of these inner products. First we review some basic facts about Gegenbauer polynomials.

4.3 Gegenbauer polynomials

The reference for all the results in this section is [10] and [25]. This section includes formulas we use in this chapter, along with other well-known and useful formulas. The Gegenbauer polynomials may be defined in terms of their generating function

$$(1-2tz+z^2)^{-\nu} = \sum_{n=0}^{\infty} C_n^{\nu}(t) z^n \ .$$

In [25] these functions are called "ultraspherical functions" and are denoted as $P_n^{(\nu)}(t)$. The Gegenbauer polynomials are Gaussian hypergeometric functions,

$$C_n^{\nu}(t) = \frac{\Gamma(2\nu+n)}{\Gamma(n+1)\Gamma(2\nu)} F\left(2\nu+n, -n; \nu+\frac{1}{2}; \frac{1-t}{2}\right) \ .$$

If n is a non-negative integer,

$$C_n^{\nu}(t) = \frac{2^n\Gamma(\nu+n)}{n!\Gamma(\nu)} t^n F\left(-\frac{n}{2}, \frac{1-n}{2}; 1-\nu-n; \frac{1}{t^2}\right) \ .$$

The polynomials

$$2^{\nu}\Gamma(\nu)\left[\frac{(n+\nu)n!}{2\pi\Gamma(2\nu+n)}\right]^{1/2} C_n^{\nu}(t) \ ,$$

$n = 0, \ldots, d$, form an orthonormal bases for polynomials of degree d, with domain $t \in [-1,1]$ and weight $(1-t^2)^{\nu-\frac{1}{2}}$. The first few Gegenbauer polynomials are

$$C_0^{\nu}(t) = 1 \ , \quad C_1^{\nu}(t) = 2\nu t \ , \quad C_2^{\nu}(t) = 2\nu(\nu+1)t^2 - \nu \ ,$$
$$C_3^{\nu}(t) = \frac{4}{3}\nu(\nu+1)(\nu+2)t^3 - 2\nu(\nu+1)t \ .$$

Subsequent Gegenbauer polynomials may be calculated using either of the following formulas.

$$(n+2)C_{n+2}^{\nu}(t) = 2(\nu+n+1)tC_{n+1}^{\nu}(t)-(2\nu+n)C_{n}^{\nu}(t)\ ;\ \frac{d}{dt}C_{n}^{\nu}(t) = 2\nu C_{n-1}^{\nu+1}(t)\,.$$

We also have

$$C_n^{\nu}(1) = \frac{\Gamma(2\nu+n)}{\Gamma(n+1)\Gamma(2\nu)}\,.$$

Associated Legendre functions may be written in terms of Gegenbauer polynomials:

$$P_l^m(t) = \frac{(-1)^m(2m)!(1-t^2)^{m/2}}{m!2^m}C_{l-m}^{m+\frac{1}{2}}(t)\,.$$

If m and l are non-negative integers,

$$P_l^m(t) = \frac{(-1)^{l+m}}{2^l l!}(1-t^2)^{m/2}\frac{d^{l+m}}{dt^{l+m}}(1-t^2)^l\,.$$

The spherical harmonics that arise in the solution of the Schrödinger equation for the hydrogen atom are

$$Y_l^m(\theta,\phi) = \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}P_l^m(\cos\,\theta)e^{im\phi}\,,$$

where l is the azimuthal quantum number, and $m = -l,\ldots,l$ is the magnetic quantum number. Here $l = 0,1,2,3,4,5,\ldots$ for the $s,p,d,f,g,h,\ldots$ orbitals. Unfortunately, the definitions of associated Legendre functions and spherical harmonics are not entirely uniform through the literature. Compare [8] , [10], [20] and [25].

The Gegenbauer polynomials are Jacobi polynomials, but with different normalization constants,

$$C_n^{\nu}(t) = \frac{\Gamma(n+2\nu)\Gamma(\nu+\frac{1}{2})}{\Gamma(2\nu)\Gamma(n+\nu+\frac{1}{2})}P_n^{(\nu-\frac{1}{2},\nu-\frac{1}{2})}\,.$$

Special cases of the Gegenbauer polynomials include the Legendre polynomials,

$$P_n(t) = C_n^{1/2}(t)\,,$$

and the Chebyshev polynomials of the second kind,

$$U_n(t) = C_n^1(t) = \frac{\sin[n\arccos(t)]}{\sin[\arccos(t)]}\,.$$

We will need to define Gegenbauer polynomials for which $\nu = 0$. Unfortunately, $C_n^0(t) \equiv 0$, unless $n = 0$. However, we can renormalize, because

$$\lim_{\nu \to 0} \frac{C_n^\nu(t)}{\nu} = \frac{2\cos[n \arccos(t)]}{n} = \frac{2}{n} T_n(t) ,$$

where T_n are the Chebyshev polynomials of the first kind. We choose a normalization that is suited to our needs, and that is continuous as $\nu \to 0$. We use that fact that

$$\lim_{\nu \to 0} \frac{C_n^\nu(1)}{\nu} = \frac{2}{n} .$$

Notation. *We define*

$$\tilde{C}_n^\nu(t) \equiv \frac{C_n^\nu(t)}{C_n^\nu(1)} ,$$

when $\nu \neq 0$, and we define

$$\tilde{C}_n^0(t) \equiv T_n(t) .$$

The first few renormalized Gegenbauer polynomials are

$$\tilde{C}_0^\nu(t) = 1 , \quad \tilde{C}_1^\nu(t) = t , \quad \tilde{C}_2^\nu(t) = \frac{2(\nu+1)}{2\nu+1} t^2 - \frac{1}{2\nu+1} ,$$
$$\tilde{C}_3^\nu(t) = \frac{2(\nu+2)}{2\nu+1} t^3 - \frac{3}{2\nu+1} t .$$

If we rewrite the results in this chapter in terms of Jacobi polynomials, the problem at $\nu = 0$ does not arise. Nonetheless, we will use Gegenbauer polynomials, because the Jacobi polynomials are too general for the problems we will consider.

We will make use of the following explicit formula:

$$\tilde{C}_n^\nu(t) = \frac{2^{n-1} n! \Gamma(2\nu+1)}{\Gamma(2\nu+n)\Gamma(\nu+1)} \sum_{j=0}^{[n/2]} \frac{(-1)^j \Gamma(\nu+n-j)}{2^{2j} j!(n-2j)!} t^{n-2j} , \qquad (2)$$

where n is a non-negative integer. Here we define

$$\frac{\Gamma(\nu+n-j)}{\Gamma(2\nu+n)} \equiv 2 , \quad \textbf{not } 1 ,$$

when $n = j = \nu = 0$. This is because we set $n = j = 0$ *before* we set $\nu = 0$. This formula may be inverted, to give

$$t^n = \frac{n!\Gamma(\nu+1)}{2^{n-1}\Gamma(2\nu+1)} \sum_{j=0}^{[n/2]} \frac{(\nu+n-2j)\Gamma(2\nu+n-2j)}{j!(n-2j)!\Gamma(\nu+n-j+1)} \tilde{C}^{\nu}_{n-2j}(t) , \quad (3)$$

where we define

$$(\nu+n-2j)\Gamma(2\nu+n-2j) \equiv \frac{1}{2} , \quad \textbf{not } 1 ,$$

when $n = 2j$ and $\nu = 0$. Using functional properties of the Gamma function, these formulas may be rewritten in many ways.

4.4 The eigenspaces of $r^2\nabla^2$

A straightforward calculation using multivariate calculus yields the following **Lemma 4.1** *If f is a homogeneous polynomial of degree $d - 2i$ in $m + 1$ variables, then*

$$\nabla^2 f = 0 \rightarrow r^2\nabla^2(r^{2i}f) = 2i(m+2d-2i-1)(r^{2i}f) .$$

Let P_d be the space of homogeneous real-valued polynomials of degree d in $m+1$ real variables. The operator $r^2\nabla^2$ maps this space to itself. Let H_d be the subspace of this space for which ∇^2 is equal to zero. By the lemma, $\{r^{2i}H_{d-2i}\}$, $i = 0, \ldots, [d/2]$, are eigenspaces of $r^2\nabla^2$ with distinct eigenvalues $\{2i(m+2d-2i-1)\}$. Furthermore, since H_d is the kernel of $\nabla^2 : P_d \rightarrow P_{d-2}$,

$$\dim H_d \geq \dim P_d - \dim P_{d-2} .$$

Therefore the sum of the dimensions of these eigenspaces is at least the dimension of P_d. We conclude that

$$P_d = \sum_{i=0}^{[d/2]} r^{2i}H_{d-2i} \quad \textit{(direct sum)}.$$

As a consequence of this argument, we see that this sum is exactly the decomposition of P_d into the eigenspaces of $r^2\nabla^2$. Furthermore,

$$\dim H_d = \dim P_d - \dim P_{d-2} ,$$

and $\nabla^2 : P_d \rightarrow P_{d-2}$ is onto. This result may be stated as

$$P_d = H_d \oplus r^2 P_{d-2} .$$

We will now show that for each eigenspace, there is a unique orthogonally invariant inner product, up to a constant. Thus each of these subspaces form irreducible representations of the orthogonal group $O(m+1)$. We fix y, and consider $v(x)^T Cv(y)$ to be a function of $x \equiv r$ alone. We then consider the following equation:

$$r^2 \nabla^2 v(x)^T Cv(y) = \lambda v(x)^T Cv(y) \ .$$

By substituting (1) into this equation, and applying elementary calculus, we obtain a first order difference equation, $[2k(m+2d-2k-1)-\lambda]\beta_k + (d-2k+2)(d-2k+1)\beta_{k-1} = 0$. But $\lambda = 2i(m+2d-2i-1)$, where $0 \leq i \leq [d/2]$. Therefore, $\beta_k = 0$, for $0 \leq k < i$. We therefore replace k with $k = i$, and conclude that for $0 \leq i \leq [d/2]$ and $1 \leq k \leq [d-2i]$,

$$2k[m+2(d-2i)-2k-1]\beta_{k+i}+[(d-2i)-2k+2][(d-2i)-2k+1]\beta_{k+i-1} = 0 \ . \quad (4)$$

We see that $v(x)^T Cv(y)$ is uniquely determined, up to a constant. We may now use (2) to see that the coefficients of certain Gegenbauer polynomials satisfy (4). Therefore, we can write

$$v(x)^T Cv(y) \ = \ \beta(x \cdot x)^{d/2}(y \cdot y)^{d/2} \ \tilde{C}_{d-2i}^{\frac{m-1}{2}}(\cos\ \theta),$$

where θ is the angle between the vectors x and y in the Euclidean norm. Notice that these inner products are positive definite iff $\beta > 0$. We set $\beta = 1$ to make the definition of inner product unambiguous.

4.5 Classification of positive definite inner products

Any invariant inner product on P_d can be written as a weighted sum of the invariant inner products on the eigenspaces of $r^2\nabla^2$ derived above. The inner product on P_d is positive definite iff all these weights are positive. We therefore use Gegenbauer polynomials to write any inner product as

$$v(x)^T Cv(y) \ = \ (x \cdot x)^{d/2}(y \cdot y)^{d/2} \sum_{i=0}^{[d/2]} \alpha_i \tilde{C}_{d-2i}^{\frac{m-1}{2}}(\cos\ \theta) \ . \quad (5)$$

If all the α_i are greater than zero, then the inner product is positive definite. To indicate this, we write

$$v(x)^T Cv(y) \ = \ (x \cdot x)^{d/2}(y \cdot y)^{d/2} \sum_{i=0}^{[d/2]} r_i^2 \tilde{C}_{d-2i}^{\frac{m-1}{2}}(\cos\ \theta) \ . \quad (6)$$

The $\{r_i\}$ have a clear geometric interpretation. They are the lengths of the projections of $v(x)$ onto the subspaces $\{r^{2i}H_{d-2i}\}$, where x is any unit vector in $\mathbf{R}^{m+1}$. We have normalized the Gegenbauer polynomials in such a way that

$$\sum_{k=0}^{[d/2]} \beta_k = \sum_{i=0}^{[d/2]} r_i^2$$

Setting $n = d - 2i$, $j = k - i$, and $\nu = (m-1)/2$ in (2), gives

$$\beta_k = \frac{2^{d-2k-1}(m-1)!}{\Gamma\left(\frac{m+1}{2}\right)(d-2k)!} \sum_{i=0}^{k} \frac{(-1)^{k-i}(d-2i)!\Gamma\left(\frac{m-1}{2} + d - i - k\right)}{\Gamma(m + d - 2i - 1)(k-i)!} r_i^2 \, , \tag{7}$$

where we define

$$\frac{\Gamma\left(\frac{m-1}{2} + d - i - k\right)}{\Gamma\left(\Gamma(m + d - 2i - 1)\right)} \equiv 2$$

when $m = 1$, $d = 2k$ and $k = i$. Equivalently, we could set $n = d - 2k$, $j = i - k$, and $\nu = (m-1)/2$ in (3) to produce the inverse

$$r_i^2 = \frac{\Gamma\left(\frac{m+1}{2}\right)\left(\frac{m-1}{2} + d - 2i\right)\Gamma(m - 1 + d - 2i)}{(m-1)!(d-2i)!}$$

$$\sum_{k=0}^{i} \frac{(d-2k)!}{2^{d-2k-1}(i-k)!\Gamma\left(\frac{m+1}{2} + d - k - i\right)} \beta_k \, , \tag{8}$$

where we define

$$\left(\frac{m-1}{2} + d - 2i\right)\Gamma(m - 1 + d - 2i) \equiv \frac{1}{2} \, ,$$

when $d = 2i$ and $m = 1$.

If we think of (2) and (3) as linear systems, (2) and (3) are the duals (or transposes) of (7) and (8), respectively.

5 Invariant normal random polynomials

We will classify all orthogonally invariant normal random homogeneous polynomials. We first observe that a positive definite inner product on a vector space corresponds to a central normal measure,

$$\rho(p) \; = \; K \; \exp(-\frac{1}{2} < p, p >) \; .$$

If we use the monomials $\{x^I\}$ as an basis, the matrix for $< .,. >$ is the inverse of the covariance matrix of the coefficients. We will discuss these covariance matrices in detail in Section 8. Therefore, at least for central random polynomials, we reduce the problem of invariant normal measures to the problem that we considered in the previous section.

We first resolve the easier problem of unitarily invariant random polynomials. We then turn our attention to real invariant random polynomials. We produce a product formula that reduces our problem to the study of random hypersurfaces, and allows use to treat mixed invariant systems. We then reduce the problem to the study of a single *central* invariant polynomial. For each such random polynomial, we calculate the expected volume of the corresponding random projective hypersurface. This completes the calculation of the expected volume for all invariant normal random polynomials, and concludes this section. We conclude Part II with detailed discussions of two special cases: random symmetric matrices, and univariate polynomials.

5.1 The complex analogue

Over the complexes, there is, up to a constant, a unique unitarily invariant normal measure.

Theorem 5.1 *Assume a measure on a vector space of complex polynomial mappings has the following two properties: (1) the measure is unitarily invariant and (2) the coefficients of the polynomial are a multivariate normal random variable. Then the coefficients must be independent, the real and imaginary part of each coefficient must be independent, and the variances of the coefficients are some constant times the binomial coefficients.*

Proof This follows from the classification of unitarily invariant inner products. See Section 7.1. □

This theorem resembles the characterization of these random polynomials given in Section 2.2, but it is quite different. In Section 2.2, independence was part of the assumption, and normality was part of the conclusion. But in 5.1, normality is assumed, and independence is deduced. For unitarily invariant complex random polynomials, independence of the coefficients is equivalent to normality of the coefficients.

This characterization has an obvious, but interesting, consequence observed in [13]. It is a central limit theorem for unitarily invariant random polynomials.

Theorem 5.2 [13] *Assume we are given any unitarily invariant probability measure on the space of complex-valued polynomials of a fixed degree. Let $\{p_i\}$ be an independent sample from this space. Then the measures of the* *random polynomial*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} p_i$$

converge weakly to a measure that is, up to a constant, the independent invariant measure we are considering in this section.

In [13] this result is stated for systems, using the two sided action of the unitary group, but the proof is similar.

This is the central limit theorem for unitarily invariant random polynomials. It is precisely the central limit theorem that is the explanation of why normal distributions are observed in practice. Therefore, this theorem would suggest that the polynomials we discuss in this section would be observed, in nature, if some complicated unitarily invariant process was generating complex valued polynomials. Indeed, some physicists have began to show interest in these random polynomials [3].

5.2 The expected volume of a real projective variety

Definition 5.1 *Consider any orthogonally invariant normal measure on a vector space of real valued, real polynomials of degree d, in the variables $t_0, \ldots, t_m$. Let μ and σ be the mean and standard deviation of the t_0^d coefficient, and let σ' be the standard deviation of the $t_0^{d-1}t_1$ coefficient. Then we define*

$$D \equiv \frac{\sigma'}{\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

as the **expected geometric degree** *for that random polynomial.*

For the random polynomials in Section 2.2, , $\sigma = 1$ and $\sigma' = \sqrt{d}$. For the random harmonic polynomials of Section 2.3,

$$\frac{\sigma'}{\sigma} = \sqrt{\frac{d(d+m-1)}{m}}\ .$$

Theorem 5.3 *For any orthogonally invariant normal homogeneous random polynomial of degree d in $m+1$ homogeneous variable,*

$$\frac{1-(-1)^d}{2} \leq \frac{\sigma'}{\sigma} \leq \sqrt{\frac{d(d+m-1)}{m}}\ .$$

The right hand inequality becomes equality for harmonic polynomials. The left hand inequality is attained by concentrating the measure on $\{ar^d\}$ if d is even, and on $\{$linear polynomials $\times\, r^{(d-1)/2}\}$ when d is odd.

By using Section 9, we can make D equal to any value in this range as a weighted average of the extreme cases. What is less obvious is that the harmonic polynomials are extremal. For us, this will be an immediate consequence of the calculation of D for all invariant measures.

Theorem 5.4 *Assume for a system of k (possibly different) independent orthogonally invariant random polynomials in m variables (or $m+1$ homogeneous variables), we have defined $D_1, ..., D_k$ as above. Then, the expected volume of the real projective variety (of codimension k) corresponding to this system will be product of these D_i multiplied by the volume of the real projective space of codimension k:*

$$\pi^{\frac{m-k+1}{2}}\Gamma\left(\frac{m-k+1}{2}\right)^{-1}\prod_{i=1}^{k}D_i \ .$$

Proof [16] First consider a completely determined (possibly mixed) system. By invariance we need only understand what is happening in some ϵ neighborhood B_ϵ of any zero of the random system, say $(0,\ldots,0)$ – where here we are using inhomogeneous coordinates. The system may be written

$$\sigma_i(b_i+\mu_i)+\sum_{j=1}^{m}\sigma_i{}'a_{ij}t_j+O(t^2)\ ,\ i=1,\ldots,m\ ,$$

and where the b_j and a_{ij} are independent standard normal random variables. For this linearized system, we must determine which values of the random vector (b_i) and the random matrix (a_{ij}) yield systems with zeros in B_ϵ, as $\epsilon \to 0$. Fix the matrix (a_{ij}) and consider and inverse (α_{ij}). We solve the linearized system to give

$$t_i \ = \ \sum_{j=1}^{m}\alpha_{ij}\frac{\sigma_j}{\sigma_j{}'}(-b_j-\mu_j)\ .$$

We must compute

$$Prob[(t_i)\in B_\epsilon] \ = \ Prob[(b_i)\in(\mu_i)-\left(\frac{\sigma_i}{\sigma_i{}'}(a_{ij})B_\epsilon\right)]\ .$$

But clearly the measure of this (infinitesimal) ellipsoid is proportional to

$$\prod_{i=1}^{m}\frac{\sigma_i{}'}{\sigma_i}exp(-\frac{\mu_i^2}{2\sigma_i^2}) \ = \ \prod_{i=1}^{m}D_i\ .$$

Having established the theorem for completely determined systems, orthogonal invariance allows us to extended the result to underdetermined systems using classical integral geometry [22] . □

5.3 Non-central invariant normal random polynomials

Fortunately, we will dispense with the non-central case quickly, and then focus on the central case.

Theorem 5.5 *The mean of any invariant random polynomial must be of the form $\mu(\Sigma x_i^2)^{d/2}$ where μ is the mean of the t_0^d coefficient (as defined in the previous section). If d is odd, any invariant measure is central.*

Proof This is classical invariant theory. For proofs and discussion of such results, see [24] . □

Thus any non central invariant random polynomial can be studied in terms of the central case. We need only adjust D by the factor

$$\exp(-\frac{\mu^2}{2\sigma^2})$$

in the calculations of expected volume and expected number of real zeros.

5.4 Central invariant normal random polynomials

We now calculate D for every central invariant normal random polynomial system.

Theorem 5.6 *For any invariant normal random polynomial, we may write*

$$v(x)^T Cv(y) \;=\; \sum_{k=0}^{[d/2]} \beta_k (x\cdot x)^k (y\cdot y)^k (x\cdot y)^{d-2k}\;. \qquad (9)$$

Then $\sigma^2 = \sum_{k=0}^{[d/2]} \beta_k$ and $\sigma'^2 = \sum_{k=0}^{[d/2]}(d-2k)\beta_k$, and therefore

$$D = \sqrt{\frac{\sum_{k=0}^{[d/2]}(d-2k)\beta_k}{\sum_{k=0}^{[d/2]}\beta_k}}.$$

Proof By orthogonal invariance of the random polynomial, we know that any orthogonal matrix Q, $v(Qx)^T Cv(Qy) = v(x)^T Cv(y)$. Then $v(x)^T Cv(y)$ must be a polynomial in $x\cdot x$, $x\cdot y$, and $y\cdot y$. This is classical invariant theory. For proofs and discussion of such results, see [24] . But $v(x)^T Cv(y)$ is homogeneous in x and y separately, so it must be of the form (1). Since the matrix C is the covariance matrix of the coefficients of the polynomial, the value of σ and σ' may be deduced from (1). The value of D then follows from Definition 5.1 and Theorem 5.3. □

Thus we have a $[d/2]$ parameter family of central invariant normal measures. For example, if we set $\beta_0 = 1$ and $\beta_k = 0$ for $k > 1$, we recover the random polynomials in Section 2.2. Another example was given in Section 7.3 of [7].

5.5 Classification of invariant normal measures

It would seem that we are done. We have a list of all central invariant normal measures and have calculated D for each of them. But we are really missing something. We do not know which values of β_i are admissible. If we let them be arbitrary, we can generate non positive definite metrics, which do not correspond to any measures at all. In some sense, the β_i are the wrong parameters to use. That is why we will express $v^T(x)Cv(y)$ in terms of Gegenbauer polynomials. In Section 7.3 of [7], we saw that for invariant normal random harmonic polynomials

$$D = \sqrt{\frac{d(d+m-1)}{m}} .$$

Theorem 5.7 *For any central invariant measure, we may write*

$$v(x)^T Cv(y) = (x \cdot x)^{d/2} (y \cdot y)^{d/2} \sum_{i=0}^{[d/2]} r_i^2 \tilde{C}_{d-2i}^{\frac{m-1}{2}}(\cos\ \theta) .$$

Then

$$D = \sqrt{\frac{\sum_{i=0}^{[d/2]}(d-2i)(d-2i+m-1)r_i^2}{m\sum_{i=0}^{[d/2]} r_i^2}} .$$

Proof For H_{d-2i} we have, by Section 7.3 of [7],

$$D = \sqrt{\frac{(d-2i)(d-2i+m-1)}{m}} .$$

We may now calculate the β_k using (7), and then apply Theorem 5.6 to complete the proof. □

We will give an alternate proof of this theorem in Section 9.

From this theorem we can calculate the expected number of real zeros of any set of m independent orthogonally invariant normal random polynomials.

6 Quadratic forms

It will be instructive to reproduce all of the work in this section for the special case $d = 2$. Our problem reduces to a discussion of random real symmetric $(m+1) \times (m+1)$ matrices. The orthogonal group acts by conjugation

$$Q(M) \equiv Q^{-1}MQ .$$

There is a three parameter family of orthogonally invariant random symmetric $(m+1) \times (m+1)$ matrices. The joint probability density for the elements of the matrix may be written in the form by

$$\rho(M) = Cexp\left\{ -\frac{1}{2}\left[b_0 tr((M-\mu I)^2) + b_1 (tr(M-\mu I))^2\right]\right\} ,$$

where $b_0 > 0$ and $b_1 > -b_0/(m+1)$. Harmonic quadratic forms correspond to traceless matrices ($b_0/b_1 \to 0$). The other extremal case is when the measure is concentrated on scalar multiples of the identity matrix ($b_0/b_1 \to -(m+1)$). The case $b_1 = 0$ has been studied extensively (see [4] [5] and [19]). This is the only case for which the elements of the random matrix are independent.

Formula (1) reduces to

$$v(x)^T C v(y) \;=\; \beta_0 (x \cdot y)^2 \;+\; \beta_1 (x \cdot x)(y \cdot y) \; . \tag{10}$$

Since C is the $(m+1)(m+2)/2 \times (m+1)(m+2)/2$ covariance matrix of the coefficients of the quadratic form, it must be inverted, to express $\{b_0, b_1\}$ in terms of $\{\beta_0, \beta_1\}$. If we define $e = (1, \ldots, 1) \in \mathbf{R}^{m+1}$,

$$(\beta_0 I + \beta_1 e^T e)^{-1} \;=\; \frac{1}{\beta_0} I + \frac{-\beta_1}{\beta_0[(m+1)\beta_1 + \beta_0]} e^T e \; .$$

We deduce that

$$b_0 = \frac{1}{\beta_0} \; ; \quad b_1 = \frac{-\beta_1}{\beta_0[(m+1)\beta_1 + \beta_0]} \; .$$

For $d = 2$ and $\nu = (m-1)/2$, the renormalized Gegenbauer polynomials are For $d = 2$ and $\nu = (m-1)/2$,

$$\tilde{C}_0^{\frac{m-1}{2}}(t) = 1 \; , \quad \tilde{C}_2^{\frac{m-1}{2}} \; = \; \frac{m+1}{m} t^2 - \frac{1}{m} \; ,$$

and the triangular systems (7) and (8) reduce to

$$\begin{pmatrix} \frac{m+1}{m} & 0 \\ -\frac{1}{m} & 1 \end{pmatrix} \begin{pmatrix} r_0^2 \\ r_1^2 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \; ; \quad \begin{pmatrix} \frac{m}{m+1} & 0 \\ \frac{1}{m+1} & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} r_0^2 \\ r_1^2 \end{pmatrix} \; .$$

Therefore, we conclude that

$$b_0 = \frac{m}{(m+1) r_0^2} \; ; \quad b_1 = \frac{r_0^2 - m r_1^2}{(m+1)^2 r_0^2 r_1^2} \; ,$$

where $r_1^2 \to 0$ for traceless random matrices, where $r_0^2 \to 0$ for measures concentrated on scalar multiples of the identity. A straightforward calculation shows that

$$Var(M_{ii}) = \beta_0 + \beta_1 = r_0^2 + r_1^2 \; ,$$

and

$$Var(M_{ij}) = \frac{\beta_0}{2} = \frac{m+1}{2m} r_0^2 \ ,$$

when $i \neq j$.

An easy way to generate these random matrices is to start with a random nonsymmetric $(m+1) \times (m+1)$ matrix with independent standard normal coefficients, and then compute

$$2r_0\sqrt{\frac{m}{m+1}}\left[M + M^T - \frac{2tr(M)}{m+1} I\right] \ + \ (r_1 z + \mu)I \ , \qquad (11)$$

where z is a standard normal random variable independent of M.

6.1 The expected volume of random conics

For random conics

$$\sigma^2 \equiv Var(M_{ii}) = \beta_0 + \beta_1 = r_0^2 + r_1^2 \ ,$$

and

$$\sigma'^2 \equiv Var(2M_{ij}) = 2\beta_0 = \frac{2(m+1)}{m} r_0^2 \ ,$$

when $i \neq j$. Furthermore, we know that the mean of $M_{ii} = \mu$. We combine Theorem 3.6 with Theorem 3.7 to get

$$\begin{aligned} D &= \sqrt{\frac{2\beta_0}{\beta_0 + \beta_1}} \exp\left(-\frac{\mu^2}{2(\beta_0 + \beta_1)}\right) \\ &= \sqrt{\frac{2(m+1)r_0^2}{m(r_0^2 + r_1^2)}} \exp\left(-\frac{\mu^2}{2(r_0^2 + r_1^2)}\right) \leq \sqrt{\frac{2(m+1)}{m}} \ . \end{aligned}$$

If the coefficients are independent this reduces to $D = \sqrt{2}$ and for the traceless case this reduces to $D = \sqrt{2(m+1)/m}$. We then can apply Theorem 2.2 to calculate expected volumes of projective varieties.

Example 6.1 *If the coefficients are independent $D = \sqrt{2}$. For a random harmonic conic $D = \sqrt{2(m+1)/m}$. We then can apply Theorem 2.2 to calculate expected volumes of projective varieties. We thus recover the results of Section 2.2 and Section 2.3, for the case $d = 2$.*

Example 6.2 *Consider an invariant normal random plane projective conic (m=2). The expected length of such a curve is πD which is always less than* *or equal to $\sqrt{3}\pi$. The random projective curve of expected length $\sqrt{3}\pi$ is given by*

$$(x \;\; y \;\; z)M\begin{pmatrix} x \\ y \\ z \end{pmatrix} = 0 \ ,$$

where M is a 3×3 invariant normal traceless matrix.

6.2 Wigner's semicircular law

Although this is a digression from the subject of this chapter, we can hardly mention random matrices without mentioning their eigenvalues. Invariant random matrices with independent coefficients have been studied intensely. In particular, it is known that the marginal density of the eigenvalues, when divided by the square root of the size of the matrix, converges to a semicircle as the size of the matrix tends to infinity. This is known as Wigner's semicircular law. We now compute the limit density for any central invariant random matrix. If $\mu \neq 0$, the following picture is simply shifted the appropriate amount. Note that μ, r_0 or r_1 may vary with m.

Theorem 6.1 *Assume we have a central invariant normal random matrix. Assume $\lim_{m\to\infty} r_1\sqrt{2m}/r_0 = C$. If $C = 0$ the density of the eigenvalues divided by $\sqrt{2m}$ converges to the unit semi-circle $\{(x,y)|x \in [-1,1],\ y = \sqrt{1-x^2}\}$. If $C = +\infty$ the density of the eigenvalues divided by r_1 converges to a standard normal distribution. Otherwise, the density of the eigenvalues divided by $\sqrt{2m}$ converges to a Fourier convolution of a unit semicircle and a normal density with variance C^2:*

$$\rho(x) \;=\; \frac{1}{\pi\sqrt{2\pi}C}\int_{-1}^{1}\sqrt{1-t^2}\exp(-\frac{(x-t)^2}{2C^2})dt \ .$$

Proof We first observe that Wigner's law holds for traceless invariant matrices. We generate an invariant random matrix M with independent coefficients, and consider it's projection $\pi(M)$ onto the traceless matrices. Now consider the equation

$$\frac{1}{\sqrt{2(m+1)}}M \;=\; \frac{1}{\sqrt{2(m+1)}}\pi(M) \;+\; \frac{1}{(m+1)\sqrt{2(m+1)}}tr(M)I \ ,$$

and let $m \to \infty$. By the classical version of Wigner's law [19] we know that the eigenvalue density of the left hand side converge to unit semicircle. On the other hand, $tr(M)$ is normally distributed with mean zero and variance $m+1$. Therefore the eigenvalue distribution for the last term of the sum converges to a measure concentrated at zero. Therefore the first term of the sum must converge to a semicircle.

We then decompose any invariant random matrix using (11). The eigenvalues of the first term of (11), when divided by $r_0\sqrt{2(m+1)}$, converge to a unit semicircle. The eigenvalues of the second term of (11) when divided by r_1, are (dependent) standard normals. The eigenvalues of the sum are convolutions of the eigenvalues of each summand. Therefore, as $m \to \infty$, we are convolving a semicircle with a normal density. We need only pay attention to the ratio of $r_0\sqrt{2(m+1)}$ to r_1.

Finally we observe that to make the statement of the proof shorter, we have replaced $m+1$ with m. For large m this difference can be ignored. □

Although there is no difference in the asymptotic density for the independent and traceless cases, for finite size the distributions are qualitatively different. See [15] for examples. All of these eigenvalue densities can be obtained by convolving a normal density with the eigenvalue density for the traceless matrices. This convolution blurres the eigenvalue distribution. Therefore the traceless matrices yield the sharpest pictures.

7 Univariate polynomials

As a final special case, we reproduce the work in this section for univariate polynomials, that is we will assume $m = 1$.

Using the substitution $t = \tan(\theta)$, we see that these random polynomials are actually special case of the random trigonometric sums studied in Section 3.2.4., and Section 5.3 Case I Example 2, of [7]. However, for the random trigonometric sums we are concerned with, that is, we can say more.

The renormalized Gegenbauer polynomials have reduced to the Chebyshev polynomials of the first kind,

$$\tilde{C}^0_{d-2i}(t) \equiv T_{d-2i}(t) = \cos[(d-2i)\arccos(t)] \ ,$$

and the eigenspaces of $r^2\nabla^2$ are given by

$$r^{2i}H_{d-2i} \equiv \{r^d \sin(d-2i)\theta, r^d \cos(d-2i)\theta\} \ ,$$

$i = 0, \ldots, [d/2]$. These are all two dimensional eigenspaces, unless d is even and $i = d/2$. Note that the number of real zeros is exactly $d-2i$ for any polynomial in H_{d-2i}.

Formula (7) reduces to

$$\beta_k = 2^{d-2k-1} \sum_{i=0}^{k} \frac{(-1)^{k-i}(d-2i)}{d-i-k} \binom{d-i-k}{k-i} r_i^2 \ ,$$

where we define

$$\frac{d-2i}{d-i-k} \equiv 2 ,$$

when $d = 2i$ and $k = i$. Similarly, (8) reduces to

$$r_i^2 = 2^{-\delta} \sum_{k=0}^{i} 2^{2k} \binom{d-2k}{i-k} \beta_k ,$$

where we define $\delta \equiv d$ if $d = 2i$, and $\delta \equiv d-1$ otherwise.

We see that Theorem 5.6 remains unchanged,

$$D = \sqrt{\frac{\sum_{k=0}^{[d/2]}(d-2k)\beta_k}{\sum_{k=0}^{[d/2]}\beta_k}} \exp\left(-\frac{\mu}{2\sum_{k=0}^{[d/2]}\beta_k}\right) ,$$

and Theorem 5.7 reduces to

$$D = \sqrt{\frac{\sum_{i=0}^{[d/2]}(d-2i)^2 r_i^2}{\sum_{i=0}^{[d/2]} r_i^2}} \exp\left(-\frac{\mu}{2\sum_{i=0}^{[d/2]} r_i^2}\right) .$$

Note that $\sum_{k=0}^{[d/2]}\beta_k = \sum_{i=0}^{[d/2]} r_i^2$.

The simplest non-trivial example is $d = 2$. This is just the case $m = 1$ of the previous subsection, but we will instead approach it as a generalization of Section 5.3 Case I Example 2 of [7].

Example 7.1 *Let a_0, a_1, and a_2 be independent standard normal random variables. Any orthogonally invariant normal random quadratic polynomial may be written*

$$a_0 r_0 (t^2-1) + a_1 2 r_0 t + (a_2 r_1 + \mu)(t^2+1) .$$

The conversion formulas (7) and (8) reduce to

$$\beta_0 = 2r_0^2 \ ; \quad \beta_1 = -r_0^2 + r_1^2$$

and

$$r_0^2 = \frac{1}{2}\beta_0 \ ; \quad r_1^2 = \frac{1}{2}\beta_0 + \beta_1 .$$

We see that the expected number of real zeros is

$$\sqrt{\frac{2\beta_0}{\beta_0+\beta_1}} \exp\left(-\frac{\mu^2}{2(\beta_0+\beta_1)}\right) = \frac{2r_0}{\sqrt{r_0^2+r_1^2}} \exp\left(-\frac{\mu^2}{2(r_0^2+r_1^2)}\right) .$$

Example 7.2 *Any orthogonally invariant normal random cubic polynomial may be written*

$$a_0 r_0(t^3 - 3t) + a_1 r_0(3t^2 - 1) + a_2 r_1 t(t^2 + 1) + a_3 r_1(t^2 + 1) \ .$$

The conversion formulas are

$$\beta_0 = 4r_0^2 \ ; \quad \beta_1 = -3r_0^2 + r_1^2$$

and

$$r_0^2 = \frac{1}{4}\beta_0 \ ; \quad r_1^2 = \frac{3}{4}\beta_0 + \beta_1 \ ,$$

and the expected number of real zeros is

$$\sqrt{\frac{3\beta_0 + \beta_1}{\beta_0 + \beta_1}} = \sqrt{\frac{9r_0^2 + r_1^2}{r_0^2 + r_1^2}} \ .$$

Part III – Central normal coefficients

8 The Metric Potential

Here we introduce definitions and notations that we will use throughout Part III. A more detailed discussion of these ideas may be found in [7]. Consider a finite dimensional real vector space of differentiable functions, $f : \mathbf{R}^m \to \mathbf{R}$, and give this vector space a central multivariate normal measure μ. We have an inner product on the dual space defined by

$$f \cdot g \equiv \int_\mu f(t)g(t) \ .$$

This induces an inner product on the primal space. If we use monomials as a basis, the matrix of this inner product is the covariance matrix of the coefficients of the random polynomial. Let v to be the dual of the evaluation mapping: $v(t) \cdot f = f(t)$.

Definition 8.1 *We define the* **metric potential** *of the random function to be*

$$\Phi(x,y) \ = \ v(x) \cdot v(y) \ = v(x)^T C v(y)$$

where C is the covariance matrix of the coefficients of the random polynomial.

Example 8.1 *Consider the random quadratic polynomials of Example 7.1.*

$$\Phi(x,y) \ = \ (x^2 \ \ x \ \ 1) \begin{pmatrix} r_0^2 + r_1^2 & 0 & -r_0^2 + r_1^2 \\ 0 & 4r_0^2 & 0 \\ r_0^2 + r_1^2 & 0 & -r_0^2 + r_1^2 \end{pmatrix} \begin{pmatrix} y^2 \\ y \\ 1 \end{pmatrix} \ .$$

This 3×3 matrix is the covariance matrix for the coefficients of the random quadratic. From this we see immediately that $\sigma^2 = r_0^2 + r_1^2$ and $\sigma'^2 = 4r_0^2$, and therefore the expected number of real zeros is

$$D = \frac{2r_0}{\sqrt{r_0^2 + r_1^2}} \; .$$

Example 8.2 *Consider the random cubic polynomials of Example 7.2.*

$$\Phi(x,y) \;=\; (x^3 \;\; x^2 \;\; x \;\; 1) \begin{pmatrix} r_0^2 + r_1^2 & 0 & -3r_0^2 + r_1^2 & 0 \\ 0 & 9r_0^2 + r_1^2 & 0 & -3r_0^2 + r_1^2 \\ -3r_0^2 + r_1^2 & 0 & 9r_0^2 + r_1^2 & 0 \\ 0 & -3r_0^2 + r_1^2 & 0 & r_0^2 + r_1^2 \end{pmatrix} \begin{pmatrix} y^3 \\ y^2 \\ y \\ 1 \end{pmatrix} .$$

This 4×4 matrix is the covariance matrix for the coefficients of the random cubic. From this we see immediately that $\sigma^2 = r_0^2 + r_1^2$ and $\sigma'^2 = 9r_0^2 + r_1^2$, and therefore the expected number of real zeros is

$$D = \sqrt{\frac{9r_0^2 + r_1^2}{r_0^2 + r_1^2}} \; .$$

Example 8.3 . *Consider a random polynomial with normal independent coefficients. Assume the variance of the $I - th$ coefficient is σ_i^2. Then*

$$\Phi(x,y) = \sum_I \sigma_I^2 x^I y^I \; .$$

Definition 8.2 *We define the* **metric** *of the random polynomial $G(P)$ to be the pullback of the projective metric to the space of zeros, using v:*

$$G(P) = \left[\frac{\partial^2}{\partial x_i \partial y_j} \log \Phi(x,y) |_{y=x=t} \right]_{ij} \; .$$

Definition 8.3 *For any random polynomial P, we define $\rho(P)$ to be the density of real zeros for a completely determined system of such random polynomials, and we define $E(P)$ to be the expected number of real zeros of such a system.*

Therefore

$$\rho(P) = \pi^{-\frac{m+1}{2}} \Gamma\left(\frac{m+1}{2}\right) \sqrt{\det(G(P))} \; .$$

Definition 8.4 *For any system of random homogeneous polynomials P, we define $Vol(P)$ to be the expected volume of the real projective variety determined by P.*

For a completely determined system, $Vol(P) \equiv E(P)$.

9 Direct Sums

Let P be the direct sum of central normal random polynomials $P_i : \mathbf{R}^m \to \mathbf{R}$. Let r_i be the lengths of the evaluation mappings v_i, and r be the length of the evaluation mapping v:

$$r_i(t) \equiv \sqrt{v_i(t)^T C_i v_i(t)}\ , \quad r(t) \equiv \sqrt{v(t)^T C v(t)}\ ,$$

for all $t \in \mathbf{R}^m$. Our results regarding direct sums follow from simple geometric considerations. In particular, we compare the projections of the tangent spaces of v_i onto the unit sphere in P_i and the unit sphere in P. For brevity we will omit proofs.

Theorem 9.1 *For any such random polynomials,*

$$\rho(P) \geq \sqrt{\sum_i \frac{r_i^{2m}}{r^{2m}} \rho_i^2(P_i)}.$$

Theorem 9.2 *Let $P_i(t)$ be m independent central normal random polynomials in m variables with proportional metrics $\alpha_i G(t)$. We may therefore write the zero densities as $\alpha_i^{m/2}\rho(t)$, and expected number of real roots as $\alpha_i^{m/2}E$. Then the direct sum has metric*

$$\sum_i \frac{\alpha_i r_i^2}{r^2} G(t)$$

zero density

$$\left(\sum_i \frac{\alpha_i r_i^2}{r^2}\right)^{m/2} \rho(t)\ ,$$

and expected number of real roots

$$\left(\sum_i \frac{\alpha_i r_i^2}{r^2}\right)^{m/2} E\ .$$

Example 9.1 *Consider the eigenspaces $r^{2i}H_{d-2i}$ of $r^2\nabla^2$ introduced in Section 4.4. As pointed out in Section 4.5, we defined r_i so that $||v(t)|| = r_i||t||$. Also, by Section 7.3 of*[7]*, the expected number of real zeros for a system of m such random polynomials is*

$$\left(\frac{(d-2i)(d-2i+m-1)}{m}\right)^{m/2}\ .$$

Therefore we may apply Theorem 9.2 with

$$\alpha_i = \frac{(d-2i)(d-2i+m-1)}{m},$$

and $E = 1$, to see that Theorem 5.7 may be considered as a special case of direct sums of random polynomials.

If we restrict our attention to univariate polynomials, we can get sharper estimates.

Theorem 9.3 *Let P be the direct sum of univariate central normal random polynomials P_i. Then*

$$\rho(P) = \sqrt{\sum_i \frac{r_i^2}{r^2}\rho_i^2(P_i) + \frac{1}{r^2}\sum_i (r_i')^2 - \frac{1}{r^2}(r')^2}\ .$$

Since $\sum_i r_i^2 = r^2$, $\sum_i (r_i')^2 \geq (r')^2$. If $\forall i$ the r_i are constant, then

$$\rho(P) = \sqrt{\sum_i \frac{r_i^2}{r^2}\rho_i^2(P_i)} \quad and \quad E(P) \leq \sum_i \frac{r_i}{r}E_i(P_i).$$

10 Tensor products

We now consider the tensor product of random polynomials. Much of this section is motivate by the proof of Main Theorem 2 in [21]. Be aware that we must distinguish beween the tensor product of a space with itself, and the tensor product of copies of the same space. For example, the tensor product of a random univariate polynomial of degree d with itself is a univariate polynomial of degree $2d$. But the tensor product of two univariate polynomials defined on different domains (this is, with distinct variables) is a bivariate polynomials. We will consider both constructions in this section. Our definition of tensor product of random polynomials is essentially the standard definition of the tensor product of Hilbert spaces.

Definition 10.1 *Let us consider central normal random polynomials P_i defined on (possibly distinct) sets of m_i variables. We define the tensor product P of the P_i to be a central normal random polynomial defined on the tensor product of the domains for the P_i, We define the covariance of the coefficients of the tensor product to be equal to the tensor product of the corresponding covariance matrices. If there are overlaps in the variables of any of the P_i, these tensor products must by symmetrized over the common variables.*

Actually, there is no need to symmetrize. We can as easily work with noncommutative polynomials, defining zeros in the usual way. This is, however, a bit unorthodox, so we have chosen to symmetrize, so that the tensor product of spaces of (commutative) polynomials are (commutative) polynomials. If the variables of P_i and P_j are distinct for all $i \neq j$, this issue does not arise, and the support of P is simply the Cartesian product of the support of P_i.

Theorem 10.1 *For each i let ρ_i be the density of the real zeros of a system of m_i independent central normal random polynomials P_i, with corresponding metrics G_i. Consider the tensor product P of the random polynomials P_i, with corresponding metric potential $\Phi(x,y)$ and metric G. Then*

$$\Phi(x,y) = \prod_i \Phi_i(x_i,y_i) \quad \textit{and} \quad G = \sum_i G_i \; .$$

Proof Consider the matrix defined in Theorem 7.1 of [7]:

$$G(t) = \left[\frac{\partial^2}{\partial x_i \partial y_j} (\log \Phi(x,y))\big|_{y=x=t} \right]_{ij} , \tag{12}$$

where $\Phi(x,y) \equiv v(x)^T C v(y)$. Here $G(t)$ is the pullback of the metric from the space of polynomials to the space of zeros, and $\Phi(x,y)$ could be called the *metric potential.* Let P_i be random polynomials, and for each we have potentials $\Phi_i(x_i,y_i)$ and corresponding metrics G_i. Then the potential of the tensor product is given by

$$\Phi(x,y) = \prod_i \Phi_i(x_i,y_i) \; ,$$

and therefore

$$G(t) = \sum_i G_i(t_i) \; . \; \square$$

We would like to replace this theorem with one about the density of real zeros. Unfortunately, the determinant of a sum of matrices is not, in general, well behaved. In the following two subsections, we consider two special cases. First we assume the factors are defined on different domains, that is, we assume each random polynomial has distinct variables. Later we assume the factors are defined on the same domain, that is, we assume the variables are identical.

10.1 Distinct variables

Theorem 10.2 *For each i let E_i be the expected number of real zeros of a system of m_i independent identically distributed central normal random polynomials P_i. Assume that that sets of variables for P_i and P_j are disjoint for*

$i \neq j$. Then the density of the real zeros of the tensor product is

$$\pi^{-\frac{m+1}{2}}\Gamma\left(\frac{m+1}{2}\right)\prod_i \pi^{\frac{m_i+1}{2}}\Gamma\left(\frac{m_i+1}{2}\right)^{-1} \rho_i \ ,$$

and the expected number of the real zeros is given by

$$\pi^{-\frac{m+1}{2}}\Gamma\left(\frac{m+1}{2}\right)\prod_i \pi^{\frac{m_i+1}{2}}\Gamma\left(\frac{m_i+1}{2}\right)^{-1} E_i \ ,$$

where m is the sum of the m_i

Proof Because the variables are distinct, the sum $G = \sum_i G_i$ is direct, that is, the G_i form a block decomposition of G. So $\det(G) = \prod_i \det(G_i)$, and therefore $\rho(t) = \prod_i \rho_i(t)$. Since each P_i has distinct variable, the integral of $\rho(t)$, decomposes into a product of separate integrals. Each integral represents the expected number of zeros of P_i, at least up to some constant. The correct normalization constant is calculated by comparing the normalization constants found in Theorem 7.1 of [7] when the number of variables is m_i and m. □

Example 10.1 *Let each P_i be univariate random polynomials with independent standard normal coefficients. The above theorem then gives the results stated in Section 2.1.*

Example 10.2 *Let P_i be the random polynomials defined in Section 2.2. Then the above theorem yields the Rojas polynomials discussed in Section 2.4.*

10.2 Identical variables

Theorem 10.3 *For each i let ρ_i and E_i be the density and expected number of real zeros of a system of m independent identically distributed central normal random polynomials from P_i, with metrics G_i, densities ρ_i, and expected number of real zeros E_i. Assume that for all i the sets of m variables of P_i are the same. Then the density ρ of a system of m independent polynomials from the tensor product of the P_i satisfies*

$$\rho(t) \geq \sqrt{\sum_i \rho_i^2(t)} \ .$$

Proof The proof follows from the fact that, for any symmetric positive definite matrices G_i,

$$\det(\sum_i G_i) \geq \sum_i \det(G_i) \ . \ \square$$

Notice that we *cannot* deduce from this theorem that $E \geq \sqrt{\sum_i E_i^2}$, even if we assume that the $\rho_i(t)$ are all proportional. However, if we make a much stronger assumption, *that the G_i are proportional*, everything becomes trivial.

Theorem 10.4 *Let $P_i(t)$ be m independent central normal random polynomials in m variables with metrics $\alpha_i(t)G(t)$ and densities $\alpha_i^{m/2}(t)\rho(t)$. Then the density of the real zeros of a system of m independent polynomials from the tensor product of the P_i is*

$$\left(\sum_i \alpha_i(t)\right)^{m/2} \rho(t) \ .$$

If furthermore the α_i do not depend on t, then if we write expected number of real zeros for P_i as $\alpha_i^{m/2}E$, and the expected number of real zeros is equal to

$$\left(\sum_i \alpha_i\right)^{m/2} E \ .$$

Corollary 10.1 *Consider the (symmetric) tensor product of n independent identically distributed central normal random polynomials P_i, each in $m+1$ homogeneous variables. Assume that the density and expected number of real zeros for a system of m such P_i is given by ρ and E respectively. Then the density and expected number of zeros for a system of m independent elements of the tensor product is equal to $n^{m/2}\rho(t)$ and $n^{m/2}E$ respectively.*

Example 10.3 *Consider the symmetric product of n linear polynomials in $m+1$ homogeneous variables. Assume the coefficients of the linear polynomial are independent standard normal coefficients. According to the theorem, the expected number of zeros for a system of n such polynomials is $n^{m/2}$. But these tensor products are exactly the random polynomials considered in Section 2.2. We therefore recover the square root result of* [14]*.*

For random polynomials with independent coefficients, we have the following immediate consequence of Theorem 10.1, and Example 8.3.

Theorem 10.5 *Let P_i be random polynomial with normal independent coefficients, and assume the variance of the $I-th$ coefficient of P_i is σ_{iI}^2. Let P be the symmetric tensor product of the P_i, and assume the variance of the $I-th$ coefficient of P is σ_I^2. Then*

$$\sum_I \sigma_I^2 z^I = \prod_i \sum_I \sigma_{iI}^2 z^I \ .$$

Example 10.4 *Consider a univariate random polynomials of even degree d with independent coefficients. Assume that the variance of the i-th coefficient* *is either i or $d-i$, whichever is less. Let E_d be the expected number of real zeros. This random polynomial is the two-fold tensor product of the random polynomial of degree $d/2$ with independent standard normal coefficients. By the asymptotic result in Section 2.1, we see that as $d \to \infty$,*

$$E_d \sim \frac{2\sqrt{2}}{\pi} \log d \ .$$

Example 10.5 *Consider univariate random polynomials of even degrees $d = 0, 2, 4, 6, 8, \ldots$, with independent coefficients. Assume that the variances of the coefficients of these random polynomials form the* **trinomial triangle** [25]*:*

$$\begin{array}{c}
1 \\
1 \ \ 1 \ \ 1 \\
1 \ \ 2 \ \ 3 \ \ 2 \ \ 1 \\
1 \ \ 3 \ \ 6 \ \ 7 \ \ 6 \ \ 3 \ \ 1 \\
1 \ 4 \ 10 \ 16 \ 19 \ 16 \ 10 \ 4 \ 1 \\
\cdot \quad \cdot \quad \cdot
\end{array}$$

Each number in this triangle is the sum of the three closest numbers in the previous row. Let E_d be the expected number of real zeros. Than

$$E_d = \sqrt{\frac{d}{2}} E_2 \ .$$

Note that

$$E_2 = \frac{4}{\pi} \int_0^1 \frac{\sqrt{t^4 + 4t^2 + 1}}{t^4 + t^2 + 1} dt \ ,$$

or approximately 1.297023574.

Corollary 10.2 *Consider the (symmetric) tensor product of m independent orthogonally central invariant normal random polynomials P_i, each in $m+1$ homogeneous variables. Assume that the density and expected number of real zeros for a system of m such P_i is given by ρ and E_i respectively. Then the density and expected number of zeros for a system of m independent elements of the tensor product is equal to*

$$\left(\sum_i \alpha_i \right)^{m/2} \rho(t)$$

and

$$\left(\sum_i \alpha_i \right)^{m/2} E$$

respectively.

Notice that the symmetric tensor product of orthogonally central invariant random polynomials is orthogonally invariant. This is not true if the variables are distinct. For example, the Rojas polynomials discussed in Section 2.4, and in the previous subsection, are *not* orthogonally invariant. However, Rojas polynomials are indeed invariant with respect to a similarly defined action of a *product* of orthogonal groups.

Example 10.6 *Consider a simple nontrivial case of Corollary 10.2, the product of a linear and a (central) quadratic. Following Example 7.1, we may write this tensor product as*

$$a_0 r_0 t(t^2-1) + a_1 2r_0 t^2 + a_2 r_1 t(t^2+1) + a_3 r_0 (t^2-1) + a_4 2r_0 t + a_5 r_1 (t^2+1)\ ,$$

where the a_i are independent standard normal random variables. This may be rewritten as

$$(a_0r_0+a_2r_1)t^3+(a_1 2r_0+a_3r_0+a_5r_1)t^2+(-a_0r_0+a_2r_1+a_4 2r_0)t+(-a_3r_0+a_5r_1)\ .$$

Let

$$F \equiv \frac{\partial(a_0r_0+a_2r_1, a_1 2r_0+a_3r_0+a_5r_1, -a_0r_0+a_2r_1+a_4 2r_0, -a_3r_0+a_5r_1)}{\partial(a_0,a_1,a_2,a_3,a_4,a_5)}\ .$$

Since

$$F = \begin{pmatrix} r_0 & 0 & r_1 & 0 & 0 & 0 \\ 0 & 2r_0 & 0 & r_0 & 0 & r_1 \\ -r_0 & 0 & r_1 & 0 & 2r_0 & 0 \\ 0 & 0 & 0 & -r_0 & 0 & r_1 \end{pmatrix}\ ,$$

we see that

$$FF^T = \begin{pmatrix} r_0^2+r_1^2 & 0 & -r_0^2+r_1^2 & 0 \\ 0 & 5r_0^2+r_1^2 & 0 & -r_0^2+r_1^2 \\ -r_0^2+r_1^2 & 0 & 5r_0^2+r_1^2 & 0 \\ 0 & -r_0^2+r_1^2 & 0 & r_0^2+r_1^2 \end{pmatrix}\ ,$$

and therefore

$$\Phi(x,y) = (x^3\ \ x^2\ \ x\ \ 1)\begin{pmatrix} r_0^2+r_1^2 & 0 & -r_0^2+r_1^2 & 0 \\ 0 & 5r_0^2+r_1^2 & 0 & -r_0^2+r_1^2 \\ -r_0^2+r_1^2 & 0 & 5r_0^2+r_1^2 & 0 \\ 0 & -r_0^2+r_1^2 & 0 & r_0^2+r_1^2 \end{pmatrix}\begin{pmatrix} y^3 \\ y^2 \\ y \\ 1 \end{pmatrix}\ .$$

From this we see immediately that $\sigma^2 = r_0^2 + r_1^2$ and $\sigma'^2 = 5r_0^2 + r_1^2$, and therefore the expected number of real zeros is

$$D = \sqrt{\frac{5r_0^2+r_1^2}{r_0^2+r_1^2}}\ .$$

Combining this with Example 7.1, we confirm Corollary 10.2 for this particular case:

$$\sqrt{\frac{5r_0^2+r_1^2}{r_0^2+r_1^2}} = \sqrt{1^2+\left(\frac{2r_0}{\sqrt{r_0^2+r_1^2}}\right)^2} .$$

If we replace (r_0^2, r_1^2) *with* $(2r_0^2, -r_0^2+r_1^2)$*, we recover Example 8.2, with the added restriction that* $r_1^2 \geq r_0^2$*.*

Theorem 10.6 *Let* ρ_i *and* E_i *be densities of zeros for central univariate normal random polynomials, and let* ρ *be the density of zeros of their tensor product. Then*

$$\rho(t) = \sqrt{\sum_i \rho_i^2(t)} .$$

If the ρ_i *are proportional, we may integrate to give*

$$E = \sqrt{\sum_i E_i^2} .$$

Proof This is just the univariate version of Theorem 10.4, along with the observation that all 1 × 1 matrices are proportional. □

Example 10.7 *Let* P_d *be the random polynomial* $a+bx^d$*, where* a *and* b *are independent standard normal random variables. It may be seen in a number of ways that the density of real zeros is*

$$\rho_d(t) \equiv \frac{dt^{d-1}}{\pi(1+t^{2d})} .$$

Note that the expected number of real zeros is one for all d*. Since*

$$\prod_{i=0}^{k-1}(1+x^{2^i}) = \sum_{i=0}^{2^k-1} x^i ,$$

we see that the univariate random polynomial discussed in Section 2.1 is, when the degree $n = 2^k-1$*, a tensor product of* $k-1$ *random polynomials with different zero densities. Applying Theorem 10.6, we recover the density of zeros originally derived by Kac* [12]*:*

$$\rho^2(t) = \frac{(t^{2n+2}-1)^2-(n+1)^2t^{2n}(t^2-1)^2}{\pi^2(t^2-1)(t^{2n+2}-1)}$$
$$= \sum_{i=0}^{k-1} \frac{2^{2i}t^{2^{i+1}-2}}{\pi^2(1+t^{2^{i+1}})^2} = \sum_{i=0}^{k-1} \rho_{2^i}^2(t) .$$

11 Compositions

Given random univariate polynomials $f(t)$ and $g(t)$, we could try to determine the expected number of real zeros of $f \circ g(t)$. Unfortunately, we cannot express this in terms of the expected number of real zeros of $f(t)$ and $g(t)$. Consider, for example, $g(t) = a(t^2 + c^2)$, where a is a standard normal random variable, and c is a constant. Instead, we will consider polynomial compositions that are homogeneous in nature. In Example 11.1 we replace $f(t)$ with a multivariate homogeneous function, and in Theorem 11.2 we replace $g(t)$ with a rational function.

Theorem 11.1 *Let $P_{ij} : \mathbf{R}^m \to \mathbf{R}$, $i = 0, \ldots, k$, $j = 1, \ldots, m$, be $(k+1) \times m$ i.i.d. central normal random polynomials, and let $Q : \mathbf{R}^{k+1} \to \mathbf{R}^m$ be any homogeneous random system, $k \geq m$. Let $E(P)$ be the expected number of real roots of any one of the $k+1$ systems $P_{i1} = \ldots = P_{im} = 0$. Let $E(Q \circ P)$ be the expected number of real roots of the system $Q(P_{01}, \ldots, P_{k1}) = 0$. Let $Vol(Q)$ be the expected volume of $\{t : \ Q(t) = 0\} \cap \{t : \ ||t|| = 1\}$. Then*

$$E(Q \circ P) \ = \ \frac{1}{2}\pi^{-\frac{k-m+1}{2}}\Gamma\left(\frac{k-m+1}{2}\right) \ E(P)Vol(Q).$$

Proof Fix j and consider the random variety $\wp : \mathbf{R}^m \to \mathbf{R}^{k+1}$, defined to be the image of any one of the random vectors (P_{ij}), $i = 0, \ldots, k$. Note that this random variety is invariant under the left action of the orthogonal group on $\mathbf{R}^{k+1}$. By Lemma 3.1, the expected volume of the projection of this variety onto the unit sphere in $\mathbf{R}^{k+1}$ is

$$\pi^{-\frac{m+1}{2}}\Gamma\left(\frac{m+1}{2}\right) E(P) \ .$$

We then apply formula (11) from Section 4.5 of [7], with $M \equiv \{t : \ Q(t) = 0\} \cap \{t : \ ||t|| = 1\}$, and where N is the projection of $\wp$ onto the unit sphere in $\mathbf{R}^{k+1}$. □

Note that Q need not be normal. In fact Q can be a measure concentrated on a single polynomial. In otherwords, we may assume, as a special case, that Q is a fixed (non-random) polynomial.

Example 11.1 *(Theorem 6.1 of [7]) Let A be a $p \times p$ random matrix polynomial. Assume the p^2 elements of A are i.i.d. central normal random polynomials. Let α_p denote the expected number of real solutions, in some interval $[a, b]$ of $\mathbf{R}$, of the equation* $\det(A) = 0$. *Then*

$$\alpha_p/\alpha_1 \ = \ \sqrt{\pi}\frac{\Gamma((p+1)/2)}{\Gamma(p/2)} \ .$$

To prove this let $Q : \mathbf{R}^{p^2} \to \mathbf{R}$ be the determinant. The volume of $\{t : Q(t) = 0\} \cap \{t : ||t|| = 1\}$ is known [6] *to be*

$$\frac{2\pi^{p^2/2}\Gamma((p+1)/2)}{\Gamma(p/2)\Gamma((p^2-1)/2)} \,.$$

We then apply Theorem 11.1 with $k = p^2 - 1$, and observe that $\alpha_1 = E(P)$ and $\alpha_p = E(Q \circ P)$. □

We now apply Theorem 11.1 to calculate the expected number of real zeros of a composition of univariate random rational functions.

Theorem 11.2 *Let $R(t) = P_1(t)/P_2(t)$ be a random rational function, where $P_1(t)$ and $P_2(t)$ are i.i.d. central normal random polynomials of the same degree. Let $S(t)$ be any random rational function, independent of $R(t)$. Let E_R, E_S and $E_{S\circ R}$ be the expected number of real zeros for $R(t)$, $S(t)$ and $S(R(t))$, respectively. Then*

$$E_{S\circ R} \;=\; E_S E_R \;.$$

Proof Define $Q(x,y) \equiv S(x/y)$. Clearly $Q(P_1(t), P_2(t)) = Q(-P_1(t), -P_2(t)) = S(R(t))$. We then apply Theorem 11.1 with $m = k = 1$, and $P \equiv P_1$. Note that $Vol(Q) = 2E(S)$. □
Note that since $Q(t)$ may be any random rational function. For example, we could take $Q(t)$ to be fixed (concentrated at a point), or we could assume $Q(t)$ is a univariate random polynomial.

Example 11.2 *Let $P_i(t)$, $i = 0, \ldots 2k+1$ be independent univariate central normal random polynomials from Section 2.2, and assume that the degree of P_{2i} and P_{2i+1} is d_i. Define random rational functions $R_i(t) = P_{2i}(t)/P_{2i+1}(t)$, $i = 0, \ldots, k$. Then the expected number of real zeros of $R_0 \circ \ldots \circ R_k$ is equal to $\sqrt{\prod_{i=0}^{k} d_i}$. In Section 3.1.2 of* [7] *we used the equation $P_{2i}(t) - tP_{2i+1}(t) = 0$ to show that the expected number of fixed points of the rational mapping $R_i(t) : \mathbf{R} \cap \infty \to \mathbf{R} \cap \infty$ is exactly $\sqrt{d_i + 1}$.*

Acknowledgments

I thank Maurice Rojas and Gregorio Malajovich for their input while producing the paper, and the City University of Hong Kong for its support during Smalefest 2000. Of particular value were conversations with Maurice Rojas and Andrei Zelevinsky while at the Foundations of Computational Mathematics Conference, IMPA, Rio de Janeiro, January 5–12, 1997. I thank Alan Edelman for allowing me to present part of this paper during his random matrix seminar at the University of California at Berkeley, on April 1, 1999.

I thank the referee for providing many useful suggestions. This chapter is dedicated to Steve Smale on the occasion of his seventieth birthday.

References

1. A.T. Bharucha-Reid and M. Sambandham, *Random Polynomials*, Academic Press, New York, 1986.
2. A. Bloch and G. Pólya, On the roots of a certain algebraic equations, *Proc. London Math. Soc.* **33** (1932), 102–114.
3. E. Bogomolny, O. Bohias, and P. Lebœuf, Distribution of roots of random polynomials, *Phys. Rev. Lett.* **68** (1992), 2726–2729.
4. A. Edelman, Eigenvalues and condition numbers of random matrices, *SIAM J. Matrix Anal. Appl.* **9** (1988), 543–560.
5. A. Edelman, *Bibliography of random eigenvalue literature*, available by anonymous ftp from math.berkeley.edu in the directory /pub/edelman. PhD thesis, Department of Mathematics, MIT, 1989.
6. A. Edelman, E. Kostlan, and M. Shub, How many eigenvalues of a random matrix are real?, *J. Amer. Math. Soc.* **7** (1994), 247–267.
7. A. Edelman and E. Kostlan, How many roots of a random polynomial are real?, *Bull. Amer. Math. Soc.* **32** (1995), 1–37.
8. R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Addison-Wesley, Reading, 1965.
9. J. Ginibre, Statistical ensembles of complex, quaternion, and real matrices, *Ann. Probab.* **14** (1986), 1318–1328.
10. I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.
11. J. Hammersley, The zeros of a random polynomial, *Proc. Third Berkeley Symp. Math. Statist. Prob.* **2** (1956), 89–111.
12. M. Kac, On the average number of roots of a random algebraic equation, *Bull. Amer. Math. Soc.* **49** (1943), 314–320.
13. E. Kostlan, Random polynomials and the statistical fundamental theorem of algebra, (1987). Available at http://developmentserver.com/randompolynomials
14. E. Kostlan, On the distribution of roots of random polynomials, Chapter 38 (pp. 419-431) of *From Topology to Computation: Proceedings of the Smalefest* edited by M.W. Hirsch, J.E. Marsden and M. Shub, Springer-Verlag, New York, 1993.
15. E. Kostlan, On the expected volume of a real algebraic variety. An open-ended web-based project. http://developmentserver.com/randompolynomials
16. E. Kostlan and A. Zelevinsky, Conversation, *Foundations of Computational Mathematics Conference*, IMPA, Rio de Janeiro, January 5–12, 1997.

17. G. Malajovich and J.M. Rojas, Random sparse polynomial systems, preprint, July 2000. This paper is included as a chapter of these conference proceedings.
18. A. McLennan, The expected number of real roots of a multihomogeneous system of polynomial equations, preprint, February 2000. Available at http://www.econ.umn.edu/~mclennan/Papers/papers.html
19. M.L. Mehta, *Random Matrices*, Academic Press, New York, 1991.
20. A. Messiah, *Quantum Mechanics*, translated from French by G.M. Temmer, Wiley, New York, 1958.
21. J.M. Rojas, On the average number of real roots of certain random sparse polynomial system, *Lectures in Applied Mathematics, American Mathematical Society* **32** (1996), 689–699.
22. L.A. Santaló, *Integral Geometry and Geometric Probability*, Volume 1 of *Encyclopedia of Mathematics and Its Applications*, Addison-Wesley, Reading, 1976.
23. M. Shub and S. Smale, Complexity of Bezout's Theorem II: Volumes and Probabilities, in *Computational Algebraic Geometry*, F. Eyssette and A. Galligo, eds, Progress in Mathematics, v 109, Birkhauser, 1993, 267–285.
24. M. Spivak, *A Comprehensive Introduction to Differential Geometry*(2nd Ed.) Publish or Perish, Berkeley, 1979.
25. E. Weisstein, *CRC Concise Encyclopedia of Mathematics* Chapman & Hall/CRC, Boca Raton, 1999.

ALMOST PERIODICITY AND DISTRIBUTIONAL CHAOS

GONGFU LIAO

Department of Mathematics, Jilin University, Changchun, Jilin, People's Republic of China

E-mail: liaogf@public.cc.jl.cn

LIDONG WANG

Department of Calculator, Siping Normal College, Siping, Jilin, People's Republic of China

In this paper, we discuss for some compact systems the existence of distributionally scrambled set. It is proved that (1) There exists a distributionally chaotic subshift of the one-sided full two shift which is strictly ergodic; (2) If a continuous self-map of a compact metric space has a regular shift invariant set, then it has an uncountable distributionally scrambled set in which each point is almost periodic; (3) If a continuous self-map of an interval has positive topological entropy, then it has an uncountable distributionally scrambled set in which each point is almost periodic.

1 Introduction

Throughout this paper, X will denote a compact metric space with metric d; I is the closed interval [0, 1].

Let $f : X \to X$ be a continuous map. For any integer $n \geq 0$, we use f^n to denote the nth iterate.

The notion of distributional chaos first occurred in ref. [9] (where, however, "distributional chaos" is called "strong chaos"), which is characterized by distribution function of distances between trajectories of two points, The concrete version is as follows.

For any $x, y \in X$, any real t and any positive integer n, let

$$\xi_n(f, x, y, t) = \#\{i; \quad d(f^i(x), f^i(y)) < t, 0 \leq i < n\},$$

where $\#\{\cdot\}$ denotes the cardinality. Let

$$F(f, x, y, t) = \liminf_{n\to\infty} \frac{1}{n}\xi_n(f, x, y, t)$$

and let

$$F^*(f, x, y, t) = \limsup_{n\to\infty} \frac{1}{n}\xi_n(f, x, y, t).$$

Call $D \subset X$ a **distributionally scrambled set** of f or, simply, a **DS scrambled set**, if for any distinct points $x, y \in D$,
(i) $F(f, x, y, t) = 0$ for some $t > 0$, and
(ii) $F^*(f, x, y, t) = 1$ for all $t > 0$.

f is said to be **distributionally chaotic** or, simply, **DS chaotic**, if it has a DS scrambled set which is uncountable.

For a continuous map $f : I \to I$, Schweizer and Smítal[9] have proven:
(C1) If f has zero topological entropy, then any pair of points can not form a DS scrambled set and therefore f is not DS chaotic;
(C2) If f has positive topological entropy, then there exists an uncountable DS scrambled set contained in an ω-limit set of some point and therefore f is DS chaotic.

One may pose the following questions:
(Q1) Is (C1) still true for a continuous map of any compact metric space X ?
(Q2) Is there an uncountable DS scrambled set in which each member is an almost periodic point of f under the hypothesis of (C2) ?

A negative answer for (Q1) has been given in [5], where a minimal DS chaotic subshift having zero topological entropy was formed.

In the present paper, we first derive in Theorem A a stronger formulation of a result in [5]. And then we discuss in Theorem B the existence of DS scrambled set and prove in Theorem C that f is DS chaotic iff so is f^n. Finally, a positive answer to (Q2) is given in Theorem D.
The main results are stated as follows.

Theorem A. There exists a DS chaotic subshift of the one-sided full two shift which is strictly ergodic and has zero topological entropy.

Theorem B. Let $f : X \to X$ be continuous. If f has a regular shift invariant set (see §4 for the definition), then it has an uncountable DS scrambled set in which each point is almost periodic and therefore f is DS chaotic.

Theorem C. Let $f : X \to X$ be a continuous map and $n > 0$ an integer. Then f is DS chaotic iff so is f^n.

Theorem D. Let $f : I \to I$ be continuous. If f has positive topological entropy, then it has an uncountable DS scrambled set in which each point is almost periodic.

The proof of Theorem A will be given in §3, the proof of Theorem B in §4, and the proofs of Theorem C and D in §5. Here we use an immediate consequence of Theorem D and some results in [9] to end this section.

Corollary E. Let $f : I \to I$ be continuous. Then the following are equivalent.

(1) f has positive topological entropy.

(2) f has an uncountable DS scrambled set in which each point is almost periodic.

(3) f has an uncountable DS scrambled set in which each member is an ω-limit point of f.

(4) f has a DS scrambled set containing two points.

Remark. In Corollary E, (1)⇒(3) and (4)⇒(1) are the results in [9]; (2)⇒(3) holds because each almost periodic point is an ω-limit one (see (2.1)); (3)⇒(4) is obvious; However, (1)⇒(2) is new. Many authors are interested in the scrambled set in the sense of Li and Yorke (see [7] or [2] for the definition). Since any DS scrambled set must be scrambled, some results in [2], [6], [8] and [11] may be also deduced directly from our Theorem D.

2 Basic definitions and preparations

Let $f : X \to X$ be a continuous map.

Let $x \in X$. $y \in X$ is said to be an ω-limit point of x, if the sequence $f(x), f^2(x), \cdots$, has a subsequence converging to y. The set of ω-limit points of x is denoted by $\omega(x, f)$. Each point in the set $\cup_{x \in X}\omega(x, f)$ is called an ω-limit point of f.

$x \in X$ is called **almost periodic** for f, if for any $\varepsilon > 0$, one can find $K > 0$ such that for any integer $q \geq 0$, there is an integer r with $q \leq r < q+K$ satisfying $d(f^r(x), x) < \varepsilon$. Denote by $A(f)$ the set of all almost periodic points of f. Obviously,

$$A(f) \subset \bigcup_{x \in X} \omega(x, f). \tag{2.1}$$

$Y \subset X$ is said to be a **minimal set** of f, if for any $x \in Y, \omega(x, f) = Y$.

Lemma 2.1. For any $x \in X$ and any $N > 0$, the following are equivalent.

(1) $x \in A(f)$.

(2) $x \in A(f^N)$.
(3) $x \in \omega(x, f)$ and $\omega(x, f)$ is a minimal set of f.

For a proof see [3] and [4].

Let $\mathcal{B}$ denote the σ-algebra of Borel sets of X. Call a probability measure μ on $(X, \mathcal{B})$ **invariant** under f, if $\mu(f^{-1}(B)) = \mu(B)$ for any $B \in \mathcal{B}$. The set of all the invariant measure of f will be denoted by $M(X, f)$. $\mu \in M(X, f)$ is said to be **ergodic** if for $B \in \mathcal{B}$, $f^{-1}(B) = B$ implies $\mu(B) = 0$ or 1. If μ is the only member of $M(X, f)$, then it must be ergodic([10]). In this case, we call f **uniquely ergodic.** A minimal and uniquely ergodic map is simply said to be **strictly ergodic**.

Lemma 2.2. Let $X, \mathcal{B}$, $M(X, f)$ be defined as above. The following are equivalent.

(1) There exists $\mu \in M(X, f)$ such that for all $x \in X, \frac{1}{n}\sum\limits_{i=0}^{n-1} \delta_{f^i(x)} \to \mu$, where $\delta_y(B)$ is 1 if $y \in B$ and 0 otherwise for any $B \in \mathcal{B}$.

(2) There exists $\mu \in M(X, f)$ such that for all complex-valued continuous function g on X and all $x \in X$,

$$\frac{1}{n}\sum_{i=0}^{n-1} g(f^i(x)) \to \int f d\mu.$$

(3) f is uniquely ergodic.

For a proof see [10].

Let $S = \{0, 1\}, \Sigma = \{x = x_1x_2\cdots; \quad x_i \in S, i = 1, 2, \cdots\}$. Define ρ : $\Sigma \times \Sigma \to R$ as follows: for any $x, y \in \Sigma$, if $x = x_1x_2\cdots, y = y_1y_2\cdots$, then

$$\rho(x, y) = \begin{cases} 0, & \text{if } x = y, \\ 1/2^k, & \text{if } x \neq y \text{ and k} = \min\{n \geq 1; x_n \neq y_n\} - 1. \end{cases}$$

It is not difficult to check that ρ is a metric on $\Sigma.(\Sigma, \rho)$ is compact and called the **one-sided symbolic space**(with two symbols).

Define $\sigma : \Sigma \to \Sigma$ by

$$\sigma(x) = x_2x_3\cdots \text{for any } x = x_1x_2\cdots \in \Sigma.$$

σ is continuous and called the **one-sided full two shift** or, simply, the **shift** on Σ. If $Y \subset \Sigma$ is closed and $\sigma(Y) \subset Y$, then $\sigma|_Y : Y \to Y$ is called a **subshift** of σ.

Call A a **tuple**, if it is a finite arrangement of the elements in S. If $A = a_1 \cdots a_m$, where $a_i \in S, 1 \leq i \leq m$, then the length of A is said to be m, denoted by $|A| = m$. Let $B = b_1 \cdots b_n$ be another tuple. Denote

$$AB = a_1 \cdots a_m b_1 \cdots b_n.$$

Then AB is also a tuple. We say B occurs in A, denoted by $B \prec A$, if there is an $i \geq 0$ such that

$$b_j = a_{i+j} \text{ for each } j = 1, 2, \cdots, n. \tag{2.2}$$

The number of i satisfying (2.2) is called the **occurrence number** of B in A, denoted by $L_B(A)$.

For any tuple $B = b_1 \cdots b_n$, denote

$$[B] = \{x = x_1 x_2 \cdots \in \Sigma; \quad x_i = b_i, 1 \leq i \leq n\}, \tag{2.3}$$

which will be called a **cylinder** generated by B.

The following Lemmas 2.3 and 2.4 may be simply deduced from the definition.

Lemma 2.3. For any two tuples A and $B, L_B(A) \leq |A|$.

Lemma 2.4. If $B, P_1 \cdots P_n$ are all tuples, then

$$\sum_{i=1}^{n} L_B(P_i) \leq L_B(P_1 \ldots P_n) \leq \sum_{i=1}^{n} L_B(P_i) + (n-1)|B|.$$

3 Proof of theorem A

In this section we shall use the subshift formed in [5] to prove Theorem A. For this we first restate, for completeness, the construction of the subshift as follows:

For any tuple $A = a_1 \cdots a_n$, we denote $\overline{A} = \overline{a}_1 \cdots \overline{a}_n$, and call it the inverse of A, where

$$\overline{a}_i = \begin{cases} 0, & a_i = 1, \\ 1, & a_i = 0, \end{cases} \text{ for } i = 1, 2, \ldots, n.$$

Take a tuple, denoted by A_1. Let A_2 be an arrangement of A_1 and $\overline{A}_1$, say $A_2 = A_1 \overline{A}_1$ (or $\overline{A}_1 A_1$). Define inductively the tuples $A_2, A_3, \cdots$, such that for any $n \geq 2$, A_n is exactly a finite arrangement of all the tuples of the set

$$\mathcal{P}_{n-1} = \{J_1 J_2 \cdots J_{n-1}; \quad J_i \in \{A_i, \overline{A}_i\}, 1 \leq i \leq n-1\}. \tag{3.1}$$

Denote $a = A_1A_2\cdots$ and let $Y = \omega(a, \sigma)$. Then $\sigma|_Y : Y \to Y$ is the subshift.

Theorem A will be completed by proving the following three propositions:

Proposition 3.1. $\sigma|_Y$ is minimal and DS chaotic.

Proposition 3.2. $\sigma|_Y$ has zero topological entropy.

Proposition 3.3. $\sigma|_Y$ is uniquely ergodic.

A proof of the first proposition will be given in the Appendix. For a proof of the second proposition see [5]. And here we only prove the third proposition. To do this we first give several lemmas.

Lemma 3.1. For any $n \geq 2, |A_n| = |\overline{A}_n| = 2^{n-1}|A_1A_2\cdots A_{n-1}|$.

Lemma 3.2. For any $n \geq 1, a = A_1A_2\cdots$ is an infinite arrangement of tuples in $\mathcal{P}_n$, where $\mathcal{P}_n$ is defined as in (3.1).

These two lemmas may be simply deduced from the definitions, here the proofs are omitted.

Lemma 3.3. If B is any given tuple, then when $n \to \infty$, the sequence $L_B(J_n)/|J_n|$ converges to a real number uniformly for $J_n \in \{A_n, \overline{A}_n\}$.

Proof: For a given tuple B, we put

$$q_n = \sum_{P\in\mathcal{P}_{n-1}} L_B(P), \qquad r_n = \frac{q_n}{|A_n|}.$$

It is easy to check from the definition that $0 \leq r_n \leq 1$ for any $n \geq 1$. Note that $Q \in \mathcal{P}_n$ iff there exists $P \in \mathcal{P}_{n-1}$ such that $Q = PA_n$ or $P\overline{A}_n$. By using Lemma 2.4 repeatedly, we get

$$\begin{aligned} q_{n+1} &= \sum_{Q\in\mathcal{P}_n} L_B(Q) \\ &= \sum_{P\in\mathcal{P}_{n-1}} L_B(PA_n) + \sum_{P\in\mathcal{P}_{n-1}} L_B(P\overline{A}_n) \\ &\geq 2 \sum_{P\in\mathcal{P}_{n-1}} (L_B(P) + q_n) \\ &= 2(2^{n-1}q_n + q_n) \\ &= (2^n + 2)q_n. \end{aligned}$$

On the other hand, by Lemma 3.1,

$$\begin{aligned}|A_{n+1}| &= 2^n|A_1A_2\dots A_n|\\ &= 2^n|A_1\cdots A_{n-1}| + 2^n|A_n|\\ &= 2\cdot 2^{n-1}|A_1\cdots A_{n-1}| + 2^n|A_n|\\ &= 2|A_n| + 2^n|A_n|\\ &= (2^n+2)|A_n|.\end{aligned}$$

Thus for each $n \geq 1$,

$$r_{n+1} = \frac{q_{n+1}}{|A_{n+1}|} \geq \frac{(2^n+2)q_n}{(2^n+2)|A_n|} = \frac{q_n}{|A_n|} = r_n.$$

So when $n \to \infty$, the sequence $\{r_n\}$ admits a limit, denoted by d_B. We will prove

$$\lim_{n\to\infty} \frac{L_B(J_n)}{|J_n|} = d_B$$

uniformly for $J_n \in \{A_n, \overline{A}_n\}$. For given $\varepsilon > 0$, there is an $N > 0$ such that for any $n \geq N$,

$$|r_n - d_B| < \frac{\varepsilon}{2}, \quad \frac{(2^{n-1}-1)|B|}{|J_n|} < \frac{\varepsilon}{2}.$$

The later can hold because $|J_n| = 2^{n-1}|A_1\cdots A_{n-1}|$ and $|B|/|A_1\cdots A_{n-1}| \to 0$ as $n \to \infty$. By Lemma 2.4,

$$\begin{aligned}q_n &= \sum_{P\in\mathcal{P}_{n-1}} L_B(P)\\ &\leq L_B(J_n)\\ &\leq \sum_{P\in\mathcal{P}_{n-1}} L_B(P) + (2^{n-1}-1)|B|\\ &= q_n + (2^{n-1}-1)|B|.\end{aligned}$$

So

$$0 \leq L_B(J_n) - q_n \leq (2^{n-1}-1)|B|.$$

Moreover,

$$\left|\frac{L_B(J_n)}{|J_n|} - r_n\right| \leq \frac{(2^{n-1}-1)|B|}{|J_n|}.$$

It follows that for $n \geq N$,

$$\begin{aligned}\left|\frac{L_B(J_n)}{|J_n|} - d_B\right| &\leq \left|\frac{L_B(J_n)}{|J_n|} - r_n\right| + |r_n - d_B| \\ &< \frac{(2^{n-1} - 1)|B|}{|J_n|} + \frac{\varepsilon}{2} \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon.\end{aligned}$$

The lemma is shown.

Lemma 3.4. When $n \to \infty, L_B(J_1 J_2 \cdots J_n)/m_n \to d_B$ uniformly for $J_i \in \{A_i, \overline{A}_i\}, i = 1, 2, \cdots$, where $d_B = \lim_{n\to\infty} L_B(J_n)/|J_n|$ as shown in Lemma 3.3, and $m_n = |A_1 A_2 \ldots A_n|$.

Proof: Since, by Lemma 2.4,

$$\begin{aligned}&L_B(J_1 \cdots J_{n-1}) + L_B(J_n) \\ \leq \quad &L_B(J_1 \cdots J_n) \\ \leq\ &L_B(J_1 \cdots J_{n-1}) + L_B(J_n) + |B|,\end{aligned}$$

we have

$$\frac{L_B(J_1 \cdots J_{n-1}) + L_B(J_n)}{m_n} \leq \frac{L_B(J_1 \cdots J_n)}{m_n}$$

$$\leq \frac{L_B(J_1 \cdots J_{n-1}) + L_B(J_n) + |B|}{m_n}. \tag{3.2}$$

By Lemma 2.3,

$$0 \leq \frac{L_B(J_1 \cdots J_{n-1})}{m_{n-1}} \leq \frac{m_{n-1}}{m_{n-1}} = 1$$

and for $n \to \infty, m_n \to \infty$ and $m_{n-1}/m_n \to 0$, we have

$$\frac{L_B(J_1 \cdots J_{n-1}) + L_B(J_n)}{m_n}$$

$$= \frac{L_B(J_1 \cdots J_{n-1})}{m_n} + \frac{L_B(J_n)}{m_n} \tag{3.3}$$

$$= \frac{L_B(J_1 \cdots J_{n-1})}{m_{n-1}} \frac{m_{n-1}}{m_n} + \frac{L_B(J_n)}{m_n} \to d_B$$

uniformly for $J_i \in \{A_i, \overline{A}_i\}, i = 1, 2, \cdots$. Moreover,

$$\frac{L_B(J_1 \cdots J_{n-1}) + L_B(J_n) + |B|}{m_n}$$

$$= \frac{L_B(J_1 \cdots J_{n-1}) + L_B(J_n)}{m_n} + \frac{|B|}{m_n} \to d_B \tag{3.4}$$

uniformly for $J_i \in \{A_i, \overline{A}_i\}, i = 1, 2, \cdots$.

From (3.2)-(3.4), we see immediately that the conclusion follows.

Proof of proposition 3.3. For any tuple $B = b_1 \cdots b_n$, denote $[\tilde{B}] = [B] \cap \mathrm{Y}$, where [B] is as in (2.3). Then all such $[\tilde{B}]$ form a subalgebra of subsets of Y, which generates the σ-algebra $\mathcal{B}(\mathrm{Y})$. Write $a = A_1 A_2 \cdots$ as $a = a_1 a_2 \cdots, a_i \in \{0, 1\}$. Define $\mu : \mathcal{B}(\mathrm{Y}) \to R$ by

$$\mu([\tilde{B}]) = \lim_{n \to \infty} \frac{1}{m_n} \#\{i \le m_n; \quad a_i \cdots a_{i+|B|-1} = B\},$$

where $m_n = |A_1 A_2 \cdots A_n|$ is as in Lemma 3.4. By Lemma 3.4, $\mu([\tilde{B}]) = d_B$. In other words,

$$\lim_{n \to \infty} \frac{1}{m_n} \sum_{i=0}^{m_n - 1} \delta_{\sigma^i(a)} = \mu \in M(Y, \sigma|_Y)$$

(see [10]). To show the proposition, by Lemma 2.2, we have to prove that for any tuple B,

$$\lim_{N \to \infty} \frac{L_B(a_i \cdots a_{i+N})}{N+1} = \mu([\tilde{B}]) \tag{3.5}$$

uniformly in i.

Let $N \gg n$. By Lemma 3.2, we may decompose

$$a_i \cdots a_{i+N} = K P_{n_1} P_{n_2} \cdots P_{n_l} Q, \tag{3.6}$$

where K, Q are tuples with length $\le m_n$ and $P_{n_j} \in \mathcal{P}_n$ for each $j = 1, 2, \cdots, l$. It follows that $N + 1 = |K| + |Q| + l \cdot m_n$. By Lemma 2.4, we get

$$\sum_{j=1}^{l} L_B(P_{n_j}) \le L_B(a_i \cdots a_{i+N}) \tag{3.7}$$

$$\leq L_B(K) + L_B(Q) + \sum_{j=1}^{l} L_B(P_{n_j}) + (l+1)|B|.$$

For any given $\varepsilon > 0$, we first take an n large enough, such that

$$\frac{|B|}{m_n} < \frac{\varepsilon}{6}$$

and for all $P_n \in \mathcal{P}_n$,

$$\left|\frac{L_B(P_n)}{m_n} - \mu([\tilde{B}])\right| < \frac{\varepsilon}{3}.$$

We then take $N_0 \gg n$, such that for $N > N_0$,

$$\frac{2(\mu([\tilde{B}])+1)}{l(N)} < \frac{\varepsilon}{3}, \quad \frac{l(N)+1}{l(N)} < 2,$$

where $l(N) = l$ is as in (3.6) (Note that for given $n, l(N) \to \infty$ as $N \to \infty$).

Since, by Lemma 2.3, $L_B(K) \leq |K| \leq m_n, L_B(Q) \leq |Q| \leq m_n$, it follows from (3.7) that for any $N > N_0$ and all $i > 1$,

$$\left|(1/(N+1))L_B(a_i \cdots a_{i+N}) - \mu([\tilde{B}])\right|$$

$$\leq \left|(1/(N+1)) \sum_{j=1}^{l} L_B(P_{n_j}) - \mu([\tilde{B}])\right| + (1/(N+1))(2m_n + (l+1)|B|)$$

$$= (1/(N+1)) \left|\sum_{j=1}^{l} L_B(P_{n_j}) - (N+1)\mu([\tilde{B}])\right|$$

$$+(1/(N+1))(2m_n + (l+1)|B|)$$

$$= (1/(N+1)) \left|\sum_{j=1}^{l} L_B(P_{n_j}) - (l\, m_n + |P| + |Q|)\mu([\tilde{B}])\right|$$

$$+(1/(N+1))(2m_n + (l+1)|B|)$$

$$\leq (1/(N+1)) \left|\sum_{j=1}^{l} L_B(P_{n_j}) - l\, m_n \mu([\tilde{B}])\right|$$

$$+(1/(N+1))(|K| + |Q|)\mu([\tilde{B}]) + (1/(N+1))(2m_n + (l+1)|B|)$$

$$\leq (1/l\,m_n)\left|\sum_{j=1}^{l} L_B(P_{n_j}) - l\,m_n\mu([\tilde{B}])\right|$$

$$+(2m_n/l\,m_n)(\mu([\tilde{B}])+1)+(1/l\,m_n)(l+1)|B|$$

$$\leq (1/l)\sum_{j=1}^{l}\left|(1/m_n)L_B(P_{n_j}) - \mu([\tilde{B}])\right|$$

$$+(2/l)(\mu([\tilde{B}])+1)+((l+1)/l)(|B|/m_n)$$

$$< (1/l)\cdot l\cdot(\varepsilon/3)+\varepsilon/3+2\cdot(\varepsilon/6)=\varepsilon.$$

Thus (3.5) is true and so the result follows.

Proof of theorem A. This follows clearly from Propositions 3.1, 3.2 and 3.3.

4. Proof of theorem B

For any tuple $B = b_1\cdots b_n$, let $[B] = [b_1\cdots b_n]$ be a cylinder generated by B as defined in (2.3). For any $n \geq 1$, let

$$\mathcal{B}_n = \{[b_1\cdots b_n];\ b_i = 0 \text{ or } 1, 1 \leq i \leq n\}.$$

Then the collection $\cup_{n=1}^{\infty}\mathcal{B}_n$ is a subalgebra which generates the σ-algebra of Borel sets of Σ.

Definition 4.1. Let $f : X \to X$ be continuous. A compact set $\Lambda \subset X$ is said to be **regular shift invariant**, if:

(1) $f(\Lambda) \subset \Lambda$;

(2) there exists a continuous surjection $h : \Lambda \to \Sigma$ satisfying

(a) $h \circ f|_\Lambda = \sigma \circ h$,

(b) there exists $M > 0$ such that for any $n \geq 1$,

$$\sum_{[B]\in\mathcal{B}_n} \mathit{diam}\, h^{-1}([B]) \leq M,$$

where *diam* denotes the diameter.

Example 4.2. Let $f : X \to X$ be continuous. If there exists an isometric homeomorphism $h : X \to \Sigma$ such that $h \circ f = \sigma \circ h$, then X is regular shift invariant under f. This is because, in this case,

$$\sum_{[B]\in\mathcal{B}_n} diam\, h^{-1}([B]) = \sum_{[B]\in\mathcal{B}_n} diam[B] \leq 1$$

and therefore (2) – (b) in Definition 4.1 is satisfied. As a special case, Σ is regular shift invariant under σ, since the identity mapping of Σ into itself is an isometric homeomorphism satisfying the requirement.

Example 4.3. Let $f : I \to I$ be continuous. Call f strictly turbulent, if there exist disjoint compact subintervals $J, K \subset I$ such that $J \cup K \subseteq f(J) \cap f(K)$. One can see from the proof of Prop. 15 of Chap. II in [1] that if f is strictly turbulent, then it has a regular shift invariant set.

Lemma 4.4. Let X be infinite and let $f : X \to X$ be minimal. If μ is the only invariant probability measure of f, then it has no atoms (i.e., each point of X has zero μ-measure).

Proof: Let $x \in X$. We first claim that $\{x\}, f^{-1}(x), f^{-2}(x), \cdots$ are pairwise disjoint. Assume the claim to be false, then $f^{-m}(x) \cap f^{-n}(x) \neq \emptyset$ for some m, n with $m > n \geq 0$. Take any $y \in f^{-m}(x) \cap f^{-n}(x)$, we have $f^m(y) = f^n(y) = x$. Furthermore,

$$f^{m-n}(x) = f^{m-n}(f^n(y)) = f^m(y) = x,$$

which contradicts the minimality of f, and so the claim follows. Thus by the property of μ, we get $\mu(\{x\}) = 0$.

Lemma 4.5. Let $Y \subset \Sigma$ be an infinite minimal set of σ and let μ be the only invariant probability measure of $\sigma|_Y$. Then when $n \to \infty$, the sequence of real numbers $\mu([b_1 b_2 \cdots b_n])$ converges to zero uniformly for $b_i \in \{0,1\}, 1 \leq i \leq n$.

Proof: Fix $\varepsilon > 0$. For any $x \in Y$, by Lemma 4.4, there exists an open neighborhood V_x of x such that $\mu(V_x) < \varepsilon$. Since Y is compact, the open cover $\{V_x\ ; x \in Y\}$ of Y has a Lebesgue number, say $\delta > 0$. There is an $N > 0$ such that for all $n \geq N$, $diam[b_1 \cdots b_n] < \delta$ uniformly for $b_i \in \{0,1\}, 1 \leq i \leq n$. Thus if $n \geq N$, then any member of the form $[b_1 \cdots b_n] \cap Y$ is contained in some V_x and so $\mu([b_1 \cdots b_n]) = \mu([b_1 \cdots b_n] \cap Y) < \varepsilon$, which proves the lemma.

Lemma 4.6. Let $f : X \to X, g : Y \to Y$ be continuous, where X, Y are compact metric space. If there exists a continuous surjection $h : X \to Y$ such that $g \circ h = h \circ f$, then $h(A(f)) = A(g)$.

Proof: By the definition of almost periodic points, we have obviously

$$h(A(f)) \subset A(g).$$

To prove the lemma, it suffices to show $h(A(f)) \supset A(g)$. For any $y \in A(g), h^{-1}(\omega(y,g))$ is an invariant subset, so it contains a minimal set M of f. Clearly, $h(M) \subset \omega(y,g)$ is invariant under g. By minimality of $\omega(y,g), h(M) = \omega(y,g)$. Thus there exists an almost periodic point $x \in M$ such that $h(x) = y$, which proves

$$h(A(f)) \supset A(g).$$

Proof of theorem B. By the hypothesis, f has a regular shift invariant set Λ, thus there is a continuous surjection $h : \Lambda \to \Sigma$ such that for any $x \in \Lambda$,

$$h \circ f(x) = \sigma \circ h(x).$$

By Theorem A, there is a minimal set $Y' \subset \Sigma$ such that $\sigma|_{Y'}$ has a uniquely ergodic measure μ and Y' contains an uncountable DS scrambled set D' of σ. Again by Lemma 2.1, each point of Y' is almost periodic. Denote, for simplicity, $g = f|_\Lambda$. By Lemma 4.6, for each $y \in D'$, we can take an $x \in A(g)$ such that $h(x) = y$. All of these points form an uncountable set of Λ, which we will denote by D. To complete the theorem, it suffices to show that D is an DS scrambled set of g.

For any distinct $x_1, x_2 \in D$, there exist distinct $y_1, y_2 \in D'$ such that $h(x_i) = y_i, i = 1, 2$. Since y_1, y_2 are in an DS scrambled set of σ, there exist $s > 0$ and a sequence $\{n_k\}$ of positive integers such that for $n_k \to \infty$,

$$\frac{1}{n_k}\xi_{n_k}(\sigma, y_1, y_2, s) \to 0. \tag{4.1}$$

Choose an $N > 0$ such that $diam[B] < s$ for all $[B] \in \mathcal{B}_N$. Denote, for simplicity,

$$I_{[B]} = h^{-1}([B])$$

for any $[B] \in \mathcal{B}_N$, and let

$$t = \min\{d(I_{[B]}, I_{[C]}); \quad [B], [C] \in \mathcal{B}_N \text{ with } [B] \neq [C]\},$$

where $d(I_{[B]}, I_{[C]}) = \inf\{d(p,q); p \in I_{[B]}, q \in I_{[C]}\}$. Since all members in $\mathcal{B}_N$ are pairwise disjoint and closed, it follows that if $[B], [C] \in \mathcal{B}_N$ with $[B] \neq [C]$, then $I_{[B]}$ and $I_{[C]}$ are disjoint compact subsets in Λ and so $d(I_{[B]}, I_{[C]}) > 0$. Therefore, by the definition, $t > 0$. It is easily seen that for any $i \geq 0$,

$\rho(\sigma^i(y_1), \sigma^i(y_2)) \geq s$
$\Rightarrow \sigma^i(y_1) \in [B], \sigma^i(y_2) \in [C]$ for some distinct $[B], [C] \in \mathcal{B}_N$
$\Rightarrow g^i(x_1) \in I_{[B]}, g^i(x_2) \in I_{[C]}$ and $d(I_{[B]}, I_{[C]}) \geq t$
$\Rightarrow d(g^i(x_1), g^i(x_2)) \geq t$, therefore, we have for each k

$$\xi_{n_k}(g, x_1, x_2, t) \leq \xi_{n_k}(\sigma, y_1, y_2, s).$$

It follows from (4.1) that for $n_k \to \infty$,

$$\frac{1}{n_k}\xi_{n_k}(g, x_1, x_2, t) \to 0,$$

and hence

$$F(g, x_1, x_2, t) = 0. \tag{4.2}$$

We now prove $F^*(g, x_1, x_2, t) = 1$ for all $t > 0$.

Choose $M > 0$ such that for any fixed $n > 0$

$$\sum_{[B] \in \mathcal{B}_n} diam I_{[B]} \leq M.$$

Such an N exists by the hypothesis of the theorem. Fix $t > 0, \varepsilon > 0$. Choose an integer $k > 0$, such that $tk > M$. And by Lemma 4.5, we may also choose an N_1 large enough, such that for any $[B] \in \mathcal{B}_{N_1}, \mu([B]) < \frac{\varepsilon}{2k}$, i.e., for any $y \in Y$,

$$\lim_{n\to\infty} \frac{1}{n} \#\{i;\ \sigma^i(y) \in [B], 0 \leq i < n\} < \frac{\varepsilon}{2k}. \tag{4.3}$$

Put

$$s = \frac{1}{2^{N_1}}.$$

Since $F^*(\sigma, y_1, y_2, s) = 1$, there exists a sequence $\{n_j\}$ of positive integers such that for $n_j \to \infty$,

$$\frac{1}{n_j}\xi_{n_j}(\sigma, y_1, y_2, s) \to 1. \tag{4.4}$$

Denote, for simplicity,

$$\theta_{n_j} = \sum_{[B] \in \mathcal{B}_{N_1}} \frac{1}{n_j} \#\{i;\ g^i(x_1), g^i(x_2) \in I_{[B]}, 0 \leq i < n_j\}.$$

Noting that

$$\begin{aligned}
&\rho(\sigma^i(y_1), \sigma^i(y_2)) < s \\
&\Longleftrightarrow \sigma^i(y_1), \sigma^i(y_2) \in [B] \text{ for some } [B] \in \mathcal{B}_{N_1} \\
&\Longleftrightarrow g^i(x_1), g^i(x_2) \in I_{[B]} \text{ for some } [B] \in \mathcal{B}_{N_1},
\end{aligned}$$

by (4.4), we have for $n_j \to \infty$

$$\theta_{n_j} \to 1. \tag{4.5}$$

Thus we can from (4.3) and (4.5) choose N large enough, such that for any $n_j > N$ and any $[B] \in \mathcal{B}_{N_1}$,

$$\frac{1}{n_j}\#\{i;\ g^i(x_1), g^i(x_2) \in I_{[B]}, 0 \le i < n_j\} < \frac{\varepsilon}{2k} \tag{4.6}$$

and

$$1 - \theta_{n_j} < \frac{\varepsilon}{2}. \tag{4.7}$$

On one hand, by the definition of θ_{n_j},

$$\begin{aligned}
&\theta_{n_j} - \sum_{[B]\in\mathcal{B}_{N_1}, diamI_{[B]}\ge t} \frac{1}{n_j}\#\{i;\ g^i(x_1), g^i(x_2) \in I_{[B]}, 0 \le i < n_j\} \\
&= \sum_{[B]\in\mathcal{B}_{N_1}, diamI_{[B]}<t} \frac{1}{n_j}\#\{i;\ g^i(x_1), g^i(x_2) \in I_{[B]}, 0 \le i < n_j\} \qquad (4.8) \\
&\le \tfrac{1}{n_j}\xi_{n_j}(g, x_1, x_2, t).
\end{aligned}$$

On the other hand, because of the choice of k, there exist in $\mathcal{B}_{N_1}$ at most k different $[B]'s$ with $diamI_{[B]} \ge t$, it follows from (4.6) and (4.8) that

$$\theta_{n_j} - \frac{\varepsilon}{2} = \theta_{n_j} - k \cdot \frac{\varepsilon}{2k} \le \frac{1}{n_j}\xi_{n_j}(g, x_1, x_2, t).$$

Combining this with (4.7), we see that for $n_j > N$,

$$0 \le 1 - \frac{1}{n_j}\xi_{n_j}(g, x_1, x_2, t) < \varepsilon$$

which gives

$$F^*(g, x_1, x_2, t) = 1. \tag{4.9}$$

By (4.2), (4.9) and the arbitrariness of x_1 and x_2, we know that D is a DS scrambled set of g.

5. Proofs of theorems C and D

Lemma 5.1. Let $f : X \to X$ be continuous, $x, y \in X, N > 0$ and $t > 0$. Then the following follow.

(i) If $F(f, x, y, t) = 0$, then $F(f^N, x, y, t) = 0$;

(ii) If $F^*(f, x, y, t) = 1$, then $F^*(f^N, x, y, t) = 1$.

Proof: (i) If $F(f, x, y, t) = 0$, then there is an increasing sequence $\{n_k\}$ of positive integers such that for $k \to \infty$,

$$\frac{1}{n_k}\xi_{n_k}(f, x, y, t) \to 0. \tag{5.1}$$

Put

$$m_k = [\frac{n_k}{N}],$$

where $[\frac{n_k}{N}]$ denotes the integral part of $\frac{n_k}{N}$. Then for each k,

$$\xi_{m_k}(f^N, x, y, t) \leq \xi_{n_k}(f, x, y, t).$$

It follows from (5.1) that for $k \to \infty$,

$$\frac{1}{n_k}\xi_{m_k}(f^N, x, y, t) \to 0$$

and further,

$$\frac{N}{n_k}\xi_{m_k}(f^N, x, y, t) \to 0.$$

This gives for $k \to \infty$,

$$\frac{1}{m_k}\xi_{m_k}(f^N, x, y, t) \to 0.$$

Hence $F(f^N, x, y, t) = 0$.

(ii) If $F^*(f, x, y, t) = 1$, then there exists an increasing sequence $\{n_k\}$ of positive integers such that for $k \to \infty$,

$$\frac{1}{n_k}\xi_{n_k}(f, x, y, t) \to 1. \tag{5.2}$$

Set

$$\delta_{n_k}(f, x, y, t) = \#\{i;\ d(f^i(x), f^i(y)) \geq t, 0 \leq i < n_k\}. \tag{5.3}$$

Then by (5.2), for $k \to \infty$,

$$\frac{1}{n_k}\delta_{n_k}(f, x, y, t) \to 0,$$

which is because for each n_k,

$$\frac{1}{n_k}\xi_{n_k}(f, x, y, t) + \frac{1}{n_k}\delta_{n_k}(f, x, y, t) = 1. \tag{5.4}$$

Put

$$m_k = [\frac{n_k}{N}].$$

By a similar argument given above, we get that for $k \to \infty$,

$$\frac{1}{m_k}\delta_{m_k}(f^N, x, y, t) \to 0$$

and further

$$\frac{1}{m_k}\xi_{m_k}(f^N, x, y, t) = 1 - \frac{1}{m_k}\delta_{m_k}(f^N, x, y, t) \to 1.$$

This proves

$$F^*(f^N, x, y, t) = 1.$$

Lemma 5.2 Let $f : X \to X$ be continuous, $x, y \in X$ and $N > 0$. Then the following follow.

(i) If for $s > 0, F(f^N, x, y, s) = 0$, then there exists $t > 0$ such that $F(f, x, y, t) = 0$.

(ii) If $F^*(f^N, x, y, s) = 1$ for all $s > 0$, then $F^*(f, x, y, t) = 1$ for all $t > 0$.

Proof: (i) If $F(f^N, x, y, s) = 0$ for $s > 0$, then there exists an increasing sequence $\{n_k\}$ of positive integers such that for $k \to \infty$

$$\frac{1}{n_k}\xi_{n_k}(f^N, x, y, s) \to 0. \tag{5.5}$$

Since X is compact, f^i is uniformly continuous for each $i = 1, 2, \ldots, N$. Consequently, for fixed $s > 0$, there exists $t > 0$ such that for all $p, q \in X$ and each $i = 1, 2, \cdots, N, d(f^i(p), f^i(q)) \geq t$ provided $d(f^N(p), f^N(q)) \geq s$. So we have

$$N(\delta_{n_k}(f^N, x, y, s) - 1) \leq \delta_{n_k N}(f, x, y, t), \tag{5.6}$$

where $\delta_{n_k}(\cdot)$ and $\delta_{n_k N}(\cdot)$ are as in (5.3). Put

$$m_k = n_k N.$$

By a simple calculation, we may derive from (5.4) and (5.6) that

$$\frac{1}{m_k}\xi_{m_k}(f,x,y,t) \leq \frac{1}{n_k}\xi_{n_k}(f^N,x,y,s) + \frac{1}{n_k}. \tag{5.7}$$

Noting that $\frac{1}{n_k} \to 0$ for $k \to \infty$, by (5.7) and (5.5) we have for $k \to \infty$,

$$\frac{1}{m_k}\xi_{m_k}(f,x,y,t) \to 0.$$

This shows

$$F(f,x,y,t) = 0.$$

(ii) Suppose $F^*(f^N,x,y,s) = 1$ for all $s > 0$. Fix $t > 0$. Since for each $i = 0,1,\cdots,N-1, f^i$ is uniformly continuous as indicated above, there exists $s > 0$ such that for all $p,q \in X$ and each $i = 0,1,\cdots,N-1, d(f^i(p),f^i(q)) < t$ provided $d(p,q) < s$. For such an $s, F^*(f^N,x,y,s) = 1$ by the hypothesis. So there exists an increasing sequence $\{n_k\}$ of positive integers, such that for $k \to \infty$

$$\frac{1}{n_k}\xi_{n_k}(f^N,x,y,s) \to 1. \tag{5.8}$$

Put

$$m_k = n_k N.$$

We easily see that

$$N\xi_{n_k}(f^N,x,y,s) \leq \xi_{m_k}(f,x,y,t).$$

Dividing both sides by m_k gives for each k,

$$\frac{1}{n_k}\xi_{n_k}(f^N,x,y,s) \leq \frac{1}{m_k}\xi_{m_k}(f^N,x,y,t).$$

Therefore by (5.8),for $k \to \infty$

$$\frac{1}{m_k}\xi_{m_k}(f^N,x,y,t) \to 1,$$

and further

$$\frac{1}{m_k}\xi_{m_k}(f,x,y,t) \to 1,$$

which shows

$$F^*(f,x,y,t) = 1.$$

Proof of theorem C. The necessity holds by Lemma 5.1 and the sufficiency by Lemma 5.2.

Proof of theorem D. If f has positive topological entropy, then by [1] f^N is strictly turbulent for some $N > 0$ (cf. Prop. 34 of Chapt.VIII in [1]). Hence f^N has a regular shift invariant set as indicated in Example 4.3. It follows from Theorem B that f^N has an uncountable DS scrambled set, say D, in which each point is almost periodic under f^N. By Lemma 5.2, D is also an DS scrambled set for f. And by Lemma 2.1, $D \subset A(f)$. Hence the result follows.

Appendix

The aim of this appendix is to prove Proposition 3.1. The argument is patterned on that given in [5].

We will continue to use the notations $\mathcal{P}_n$ and m_n, whose definitions are as in (3.1) and Lemma 3.4 respectively. And we will use the following lemmas in which Lemmas A.1 and A.2 may be simply deduced from the definitions.

Lemma A.1. For any $n > 1, m_n - m_{n-1} = 2^{n-1}m_{n-1}$.

Lemma A.2. Let $x = x_1x_2\cdots \in \Sigma$. If for any $n \geq 1$ there is a $K > 0$ such that for each $i \geq 1$,

$$x_1\cdots x_n \prec x_ix_{i+1}\cdots x_{i+K},$$

then $x \in A(\sigma)$.

Lemma A.3. $a = A_1A_2\cdots \in A(\sigma)$.

Proof of lemma A.3. Write a as $a = a_1a_2\cdots, a_i \in \{0,1\}, i = 1,2,\cdots$. Fixing an $n \geq 1$, we have obviously

$$a_1a_2\cdots a_n \prec A_1A_2\cdots A_n. \tag{A.1}$$

By the definition of A_{n+1},

$$A_1A_2\cdots A_n \prec A_{n+1}, \tag{A.2}$$

$$\overline{A}_1\overline{A}_2\cdots\overline{A}_n \prec A_{n+1}. \tag{A.3}$$

Taking inverses on both sides of (A.3), we also have

$$A_1A_2\cdots A_n \prec \overline{A}_{n+1}. \tag{A.4}$$

Combining (A.4) with (A.2), we know that

$$A_1A_2\cdots A_n \prec J_{n+1} \prec J_1J_2\cdots J_{n+1}, \tag{A.5}$$

provided $J_1J_2\cdots J_{n+1} \in \mathcal{P}_{n+1}$. Thus for given n, we may take $K = 3m_{n+1}$. For any $i \geq 1$, by Lemma 3.2, there is a tuple $J_1J_2\cdots J_{n+1} \in \mathcal{P}_{n+1}$ occurring in $a_ia_{i+1}\cdots a_{i+K}$, i.e.

$$J_1J_2\cdots J_{n+1} \prec a_ia_{i+1}\cdots a_{i+K}. \tag{A.6}$$

This is because the length of any $J_1J_2\cdots J_{n+1} \in \mathcal{P}_{n+1}$ is m_{n+1}. Summing up (A.1), (A.5) and (A.6) gives $a_1a_2\cdots a_n \prec a_ia_{i+1}\cdots a_{i+K}$. Thus by Lemma A.2, $a \in A(\sigma)$.

Lemma A.4. There is an uncountable subset E in Σ such that for any different points $x = x_1x_2\cdots, y = y_1y_2\cdots \in \Sigma, x_n = y_n$ for infinitely many n and $x_m \neq y_m$ for infinitely many m.

Proof: For any $x = x_1x_2\cdots, y = y_1y_2\cdots \in \Sigma$, denote $x \sim y$, if $x_n = y_n$ holds only for finitely many n or $x_m \neq y_m$ holds only for finitely many m. We easily check that $\sim$ is an equivalence relation on Σ. Let $x \in \Sigma$. It is easy to see that the set $\{y \in \Sigma;\, y \sim x\}$ is countable and so the quotient set $\Sigma/\sim$ is uncountable. Taking a representative in each equivalent class of $\Sigma/\sim$, we get an uncountable set E which satisfies the requirement.

Proof of proposition 3.1. By Lemmas 2.1 and A.3, we easily see that $\sigma|_Y$ is minimal. So to complete the proof of the proposition, it suffices to show that $\sigma|_Y$ is DS chaotic.

Take an uncountable subset E in Σ such that for any different points $x = x_1x_2\cdots, y = y_1y_2\cdots \in E, x_n = y_n$ holds for infinitely many n and $x_m \neq y_m$ holds for infinitely many m. By Lemma A.4, such a subset is existent. Define $\varphi : E \to \Sigma$ by $\varphi(x) = J_1J_2\cdots$, where

$$J_i = \begin{cases} A_i, & \text{if } x_i = 1, \\ \overline{A}_i, & \text{if } x_i = 0, \end{cases} \text{ for } i = 1, 2, \cdots.$$

Set $D = \varphi(E)$. Since for fixed $i, J_1\cdots J_i \prec A_{i+1} \prec a$, no matter what $J_j (1 \leq j \leq i)$ is taken, there is a $k \geq 0$ such that the first m_i symbols of $\sigma^k(a)$ is exactly the tuple $J_1\cdots J_i$(note that $|J_1\cdots J_i| = m_i$). This shows $\varphi(x) \in \omega(a, \sigma) = Y$ for all $x \in E$. And therefore $D \subset Y$. Since E is uncountable and φ is injective, D is uncountable.

Let $b = B_1B_2\cdots, c = C_1C_2\cdots$ be different points in D, where $B_i, C_i \in \{A_i, \overline{A}_i\}, i = 1, 2, \cdots$. By the definition we know that there exist sequences of positive integers $p_i \to \infty$ and $q_i \to \infty$ such that $B_{p_i} = C_{p_i}$ and $B_{q_i} = \overline{C}_{q_i}$ for all i. Put, for simplicity,

$$\delta_{bc}(j) = \rho(\sigma^j(b), \sigma^j(c)),\ j = 1, 2, \cdots.$$

First, it is easily seen that for given $p_i > 1$, if $m_{p_i-1} \leq j < m_{p_i} - m_{p_i-1}$, then the first m_{p_i-1} symbols of $\sigma^j(b)$ and $\sigma^j(c)$ coincide correspondingly. So for such j $\delta_{bc}(j) \leq 1/2^{m_{p_i-1}}$. Thus for given $t > 0, \delta_{bc}(j) < t$ provided p_i is large enough. Furthermore,

$$\frac{1}{m_{p_i} - m_{p_i-1}} \xi_{m_{p_i} - m_{p_i-1}}(\sigma, b, c, t)$$

$$= \frac{1}{m_{p_i} - m_{p_i-1}} \#\{j;\ \delta_{bc}(j) < t, 0 \leq j < m_{p_i} - m_{p_i-1}\}$$

$$\geq \frac{1}{m_{p_i} - m_{p_i-1}} \#\{i;\ \delta_{bc}(j) < t,\ m_{p_i-1} \leq j < m_{p_i} - m_{p_i-1}\}$$

$$= \frac{m_{p_i} - m_{p_i-1} - m_{p_i-1}}{m_{p_i} - m_{p_i-1}} = 1 - \frac{m_{p_i-1}}{m_{p_i} - m_{p_i-1}}$$

$$= 1 - \frac{m_{p_i-1}}{2^{p_i-1} m_{p_i-1}} \to 1\ (p_i \to \infty),$$

where the last equality is by Lemma A.1. This proves

$$F^*(\sigma, b, c, t) = 1. \tag{A.7}$$

Secondly, it is easy to see that for given $q_i > 1$, if $m_{q_i-1} \leq j < m_{q_i} - m_{q_i-1}$, then the first m_{q_i-1} symbols of $\sigma^j(b)$ and $\sigma^j(c)$ are all distinct correspondingly. So for such $j, \delta_{bc}(j) = 1$. Then for any $t \in (0, 1]$, we have

$$\frac{1}{m_{q_i} - m_{q_i-1}} \xi_{m_{q_i} - m_{q_i-1}}(\sigma, b, c, t)$$

$$= \frac{1}{m_{q_i} - m_{q_i-1}} \#\{j;\ \delta_{bc}(j) < t,\ 0 \leq j < m_{q_i} - m_{q_i-1}\}$$

$$= \frac{1}{m_{q_i} - m_{q_i-1}} \#\{j;\ \delta_{bc}(j) < t,\ 0 \leq j < m_{q_i-1}\}$$

$$\leq \frac{m_{q_i-1}}{m_{q_i} - m_{q_i-1}} = \frac{m_{q_i-1}}{2^{q_i-1} m_{q_i-1}} \to 0\ (q_i \to \infty).$$

This shows

$$F(\sigma, b, c, t) = 0. \tag{A.8}$$

(A.7) and (A.8) prove that b and c are a pair of DS scrambled points. By the arbitrariness of b and c, $\sigma|_Y$ is DS chaotic.

Acknowledgements

Supported by the National Natural Science Foundation of China

References

1. L. S. Block and W. A. Coppel, Dynamics in one dimension, Lecture Notes in Math., 1513, Springer-Verlag, New York, Berlin Heidelberg, 1992.
2. B. S. Du, Every chaotic interval map has a scrambled set in the recurrent set, Bull. Aust. Math. Soc. **39**(1989), 259 - 264.
3. P. Erdös and A. H. Stone, Some remarks on almost transformation, Bull. Amer. Math. Soc. **51**(1945), 126 - 130.
4. W. H. Gottschalk, Orbit-closure decompositions and almost periodic properties, Bull. Amer. Math. Soc. **50**(1944), 915 - 919.
5. G. F. Liao and Q. J. Fan, Minimal subshifts which display Schweizer-Smítal chaos and have zero topological entropy, Science in China, Series **A 41**(1988), 33 - 38.
6. G. F. Liao and L. Y. Wang, Almost periodicity, chain recurrence and chaos, Israel J. Math. **93**(1996), 145 - 156.
7. T. Y. Li and J. A. Yorke, Period three implies chaos, Amer. Math. Monthly **82**(1975), 985 - 992.
8. R. S. Yang, Pseudo shift invariant sets and chaos (Chinese), Chinese Ann. Math. Series **A 13**(1992), 22 - 25.
9. B. Schweizer and J. Smítal, Measures of chaos and a spetral decomposition of dynamical systems on the interval, Tran. Amer. Math. Soc. **344**(1994), 737 - 754.
10. P. Walters, An introduction to ergodic theory, Springer-Verlag, New York, Berlin Heidelberg, 1982.
11. Z. L. Zhou, Chaos and topological entropy (Chinese), Acta Math. Sintica **31**(1988), 83 - 87.

POLYNOMIALS OF BOUNDED TREE-WIDTH

JANOS A. MAKOWSKY
Department of Computer Science, Technion - Israel Institute of Technology, Haifa, Israel
E-mail: janos@cs.technion.ac.is

KLAUS MEER
SDU Odense, Department of Mathematics and Computer Science, Campusvej 55, DK-5230 Odense M, Denmark
E-mail: meer@imada.sdu.dk

We introduce a new sparsity conditions, the *tree-width*, on multivariate polynomials in n variables (over some ring R) and show that under this condition many otherwise intractable computational problems involving these polynomials become solvable in polynomial (in some cases even linear) time in n in the Blum-Shub-Smale-model over R. To define our sparsity condition we associate with these polynomials a hypergraph and study classes of polynomials where this hypergraph has tree-width at most k for some fixed $k \in \mathbb{N}$.
We are interested in three cases:
(1) The evaluation of multivariate polynomials where the number of monomials is $O(2^n)$. Examples are the permanent or the hamiltonian polynomials.
(2) For finite fields $\mathbb{F}$ the question whether a system of n polynomials $p_i(\bar{x}) \in \mathbb{F}[\bar{x}]$ of fixed degree d in n variables has a root in $\mathbb{F}^n$.
(3) For infinite ordered rings (or fields) R_{ord}, a polynomial of fixed degree d in n variables $p(\bar{x}) \in R_{ord}[\bar{x}]$ and a finite subset $A \subset R_{ord}$ we want to know whether $p(\bar{a}) > 0$ for all $\bar{a} \in R^n_{ord}$.
Our method uses graph theoretic and model theoretic tools developed in the last 15 years and applies them to the algebraic setting. This work is an extension of work by B. Courcelle, J.A. Makowsky, and U. Rotics and by Arneborg, Lagergren, and Seese.

Key words: Blum-Shub-Smale model, meta-finite structures, bounded tree-width, monadic second order logic

1 Introduction and results

It is well known that deciding the solvability of polynomial systems and approximating solutions, if they exist, are in general difficult computational tasks. In a complexity theoretic framework for real or algebraically closed fields Blum, Shub, and Smale [10] substantiated this experience by introducing a computational model over $\mathbb{R}, \mathbb{C}$ and more general ring structures and showing **NP**-completeness results for the above mentioned decision problem. The same is true over finite fields. For an introduction into the Blum-Shub-Smale (shortly BSS) model of computation see [5].

Another interesting problem is the computation of families of polynomials having exponentially many monomials. Prominent examples are the permanent and the hamiltonian polynomials. Their computational complexity has been studied by Valiant [42].

A lot of work has been done in analyzing subclasses of such hard problems hoping for better algorithms when problem instances are restricted to these classes. For example, in relation with polynomial systems sparsity conditions were previously used in order to analyze location of zeros of multivariate polynomials. One of the most spectacular stems from the *Newton polytope* associated with a system of multivariate polynomials over the complex numbers. Bernstein's theorem then relates the number of isolated zeros of this system to the *mixed volume of the Newton polytope*; for more on this subject see [12,26,30].

In our approach we are more interested in deciding the existence of zeros than in counting. For this purpose we introduce a new sparsity condition, the tree-width, on systems of multivariate polynomials in n variables (over some ring R) and show that under this condition many otherwise intractable computational problems involving these polynomials become solvable in polynomial or even linear time in n (in the BSS-model over R).

We associate with these polynomials a hypergraph and study classes of polynomials where this hypergraph has tree-width at most k for some fixed $k \in \mathbb{N}$. Tree-width of graphs is a useful concept with a long history and a plethora of results, cf. [20]. A definition for hypergraphs is given in section 2 below.

We are interested in three cases:

(i) The evaluation of multivariate polynomials where the number of monomials is $O(2^n)$, such as the permanent of a matrix, the permanent of the Hadamard powers of a matrix, the hamiltonian or many other *generating functions of graph properties*, see [11]. In general most of these polynomials are not known to allow evaluation in polynomial time. Here, the sparsity condition we impose is the bound k on the tree-width of the underlying graph. We show that all these generating functions can be evaluated in time $O(n)$ where the constant depends (super-exponentially) on k.

(ii) The feasibility problem over a finite field $\mathbb{F}$. Here, the question is whether a system of n polynomials $p_i(\bar{x}) \in \mathbb{F}[\bar{x}]$ of fixed degree d in n variables has a root in $\mathbb{F}^n$.

This problem is $NP_{\mathbb{F}}$ hard for large enough degree d. The sparsity condition we impose is the bound k on the tree-width of the d-hypergraph of non-vanishing monomials. We show that for finite fields the problem is solvable in time $O(n)$ where the constant depends super-exponentially on k and the size of the field $\mathbb{F}$. The same is true for finite rings.

(iii) We analyze an extension of (ii) to infinite fields or rings. It turns out that we have to impose further conditions on the decision problem being considered. For infinite ordered rings R_{ord}, a polynomial of fixed degree d in n variables $p(\bar{x}) \in R_{ord}[\bar{x}]$ and a finite subset $A \subset R_{ord}$ we want to know both whether there exists an $\bar{a} \in A^n$ such that $p(\bar{a}) = 0$ or whether $p(\bar{a}) > 0$ for all $\bar{a} \in A^n$. Though not known to be NP_R complete any more, these problems are important members of the subclass DNP_R of NP_R where the search space for verification is restricted to be finite. Feasibility and positivity turn out to be decidable in polynomial time in n for polynomials of tree-width at most k if we impose some further restriction on the coefficients of p and A.

For positivity we finally show how this additional coefficient condition can be avoided.

Our approach applies to a general setting where properties being expressible in a specific logical manner are considered. It uses methods from graph theory and model theoretic tools developed in the last 15 years and applies them to the algebraic setting. This work is an extension of work by B. Courcelle, J.A. Makowsky and U. Rotics [15], which extends [13] and [1]. The main new aspect with respect to those works is the ability to deal with a much larger class of algebraic properties captured by the logical framework we are going to define. This allows treatment of problems like the existence of zeros for polynomials, linear programming and many more.

The paper is organized as follows. In section 2 we introduce tree-width of matrices, polynomials and systems of polynomials. In section 3 we state a result from [15] to illustrate the definition. Section 4 collects problems in relation with polynomial systems our approach applies to. The main results are then stated. The mathematical development begins in section 5. We define the logical framework in which we express our problems. This mainly refers to developing so-called existential monadic second-order logic for meta-finite structures. The latter is crucial in order to combine algebraic issues with the concept of bounded tree-width. Proofs of the theorems are given in section 6 and further discussions follow in section 7. Since the arguments in sections 5 and 6 use a lot of previous work which we could not find presented in a compact fashion in literature, a detailed appendix is added. With the appendix the paper is self-contained. It explains the construction of parse-trees starting from tree-decompositions of hypergraphs; full proofs of the crucial theorems by Fraisse-Hintikka and Feferman-Vaught are given together with its algorithmic use.

2 Tree-width of polynomials and matrices

Let $V = \{0, 1, 2, \dots, n\}$ be the index set of the variables of

$$p(x) = p(x_0, x_1, \dots, x_n) = \sum_{(i_1, \dots i_d) \in E} c_{i_1, \dots, i_d} x_{i_1} \cdot x_{i_2} \cdot \dots \cdot x_{i_d}$$

with $E \subseteq V^d$ and $x_0 = 1$. E is the set of d-tuples of indices $(i_1, \dots i_d)$ such that the coefficient $c_{i_1, \dots, i_d} \neq 0$.

Definition 1. With $p(x)$ we associate the d-hypergraph $G = \langle V, E \rangle$ and define the tree-width of $p(x)$ as the tree-width of G.

For systems of polynomials p_j of degree d we look at the hypergraph of the non-vanishing coefficients $c_{j,\alpha}$, i.e. the induced $d+1$-hypergraph.

Definition 2 (Tree-width of a d-hypergraph). A k-tree decomposition of G is defined as follows:

(i) $\mathcal{T} = \langle T, <_T \rangle$ is a tree with $t <_T s$ expressing that t is a child of s.

(ii) For each $t \in T$ we have a subset $V_t \subseteq V$ of size at most $k + 1$.

(iii) For each hyperedge $(i_1, \dots, i_d) \in E$ there is a $t \in T$ such that $\{i_1, \dots i_d\} \subseteq V_t$.

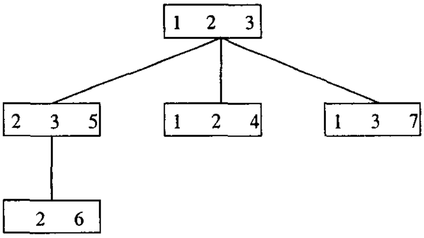
(iv) For each $i \in V$ the set $V(i) = \{t \in T \mid i \in V_t\}$ forms a (connected) subtree of $\mathcal{T}$.

G has tree-width at most k if there exists a k-tree decomposition of G. The *tree width* of G is the smallest such k.

Examples 3. (i) The polynomial $p_1(x) = \sum_{i=1}^{n} x_i^4$ has tree-width 0.

Note that the Newton polytope of p_1 is maximal with respect to polynomials of degree 4 in n variables.

(ii) Consider the polynomial $p(x_1, \dots, x_7) := x_1 \cdot x_2 \cdot x_3 + x_2^3 \cdot x_3^5 + x_1^3 + x_2^2 \cdot x_6 + x_2 \cdot x_3 \cdot x_5 + x_1 \cdot x_2^5 \cdot x_4 + x_1 \cdot x_3 \cdot x_7 + x_7^2 + x_3 \cdot x_7 - x_1 \cdot x_3 - x_3^5$. It has tree-width 2 according to the following tree decomposition:



Note that there is no tree decomposition of width 1 because there are monomials involving 3 factors.

(iii) The polynomials

$$p_2(x) = \sum_{i=1}^{n} c_{i,i+_3 1,i+_3 2} x_i x_{i+_3 1} x_{i+_3 2}$$

and

$$p_3(x) = \sum_{i=1}^{n} d_{i,i+_3 1,i+_3 2} x_i^3 x_{i+_3 1}^5 x_{i+_3 2}$$

(where $+_3$ is addition $\pmod 3$)) have tree-width 2.

(iv) For $p_1(x)$ from above, the polynomial $p_1^2(x)$ has tree-width $n-1$. This is so because all monomials appear and hence the 4-hypergraph associated with the polynomial is a hyperclique.

Boundedness of the tree-width is a sparsity condition. If $p(x)$ in n variables of degree d has tree-width k, the number of monomials is $O(n)$ with a constant depending on k, d only.

Remark 1. The tree-width of a polynomial depends on its particular representation. We can easily see that p_1 and p_1^2 have different tree-width but they represent the same variety. The tree-width of a system of polynomials which consists of a single polynomial does differ at most by 1 from the tree-width of the polynomial as such.

Similarly, we can define the tree-width of an $(n \times n)$ matrix $M = (m_{i,j})$.

Definition 4. The tree-width of an $(n \times n)$ matrix $M = (m_{i,j})$ is the tree-width of the graph $G_M = \langle V_M, E_M \rangle$ with $V_M = \{1, 2, \ldots, n\}$ and $(i,j) \in E_M$ iff $m_{i,j} \neq 0$.

Examples 5.

(1) The $(n \times n)$ matrix $M_1 = (m_{i,j})$ with $m_{i,j} = 1$ for all i, j has tree-width $n - 1$. Note that M_1 has linear rank 1.

(2) The $(n \times n)$ matrix $\mathbf{1} = (m_{i,j})$ with $m_{i,i} = 1$ and $m_{i,j} = 0$ for $i \neq j$ has tree-width 0. Note that $\mathbf{1}$ has linear rank n.

Theorem 6 (Bodlaender). *a) There is a linear time algorithm (with bad constants) which decides, given a hypergraph G whether it has a k-tree decomposition, and if yes, constructs one.*

b) If a hypergraph G over n vertices has a k-tree decomposition, then one can construct in linear time a balanced $O(k)$-tree decomposition of depth $O(logn)$.

Though originally formulated for graphs, the extension of the above results to hypergraphs is straightforward.

A survey of such results may be found in [8,7].

3 Generating functions of graph properties

The first use of tree-width of a matrix was presented in [15]. There the computational complexity of computing the *permanent* and *hamiltonian* of a matrix was studied.

Let $M = \{m_{i,j}\}$ be an $(n \times n)$ matrix over a field K. The *permanent* $per(M)$ of M is defined as

$$\sum_{\pi \in \mathcal{S}_n} \prod_i m_{i,\pi(i)}$$

The *hamiltonian* $ham(M)$ of M is defined as

$$\sum_{\pi \in \mathcal{H}_n} \prod_i m_{i,\pi(i)}$$

where $\mathcal{H}_n$ is the set of hamiltonian permutations of $\{1, \ldots, n\}$. Recall that a permutation $\pi \in \mathcal{S}_n$ is *hamiltonian* if the relation $\{(i, \pi(i)) : i \leq n\}$ is connected and forms a cycle.

In general, both the *permanent* and the *hamiltonian* are hard to compute and the best algorithms known so far are exponential in n, [4,11]. This applies also for the computational model due to Blum, Shub and Smale (BSS model), cf. [5]. Barvinok in [2] has shown that if the (linear) rank of the matrix is bounded by r both the *permanent* and the *hamiltonian* can be computed in polynomial (not linear) time. Hence these problems are parametrically tractable in the sense of [19]. Linear rank and tree-width are independent notions: The $(n \times n)$ matrix consisting of 1's only has rank 1 but tree-width $n-1$ (it is a clique). The corresponding unit matrix has rank n but tree-width 1, as the graph consists of isolated points. Tree-width of a matrix also makes the *permanent* and the *hamiltonian* parametrically tractable.

Theorem 7 (Courcelle, Makowsky, Rotics, 1998). *Let M be a real $(n \times n)$ matrix of tree-width k. Then $per(M)$ and $ham(M)$ can be computed in time $O(n)$.*

The same technique can also be applied to other families of multivariate polynomials such as *cycle format polynomials*, and, more generally, *generating functions of graph properties*, cf. [29,11].

Let $G = \langle V, E, w \rangle$ be an edge weighted graph with weights in a field K and $\mathcal{E}$ be a class of (unweighted) graphs closed under isomorphisms. We extend w to subsets of E by defining $w(E') = \prod_{e \in E'} w(e)$. The *generating function corresponding to G and $\mathcal{E}$* is defined by

$$GF(G, \mathcal{E}) =_{def} \sum \{w(E') : \langle V, E' \rangle \in \mathcal{E} \text{ and } E' \subseteq E\}$$

Strictly speaking $GF(G, \mathcal{E})$ is a function with argument w and value in K. Furthermore, w is a function

$$w : \{1, \ldots, n\}^2 \to K$$

which can be interpreted as an $(n \times n)$ matrix over K. If we view $w(i,j) = u_{i,j}$ as indeterminates, $GF(G, \mathcal{E})$ is a multivariate polynomial in $K[u_{i,j} : i, j \leq n]$.

The *permanent* is the generating function for $G = K_n$, the clique on n vertices, and $\mathcal{E}_{per}$ the perfect matchings. The *hamiltonian*, similarly, is the generating function for $\mathcal{E}_{ham}$, the class of n-cycles.

Similarly, $K_{m,n}$ is the complete bipartite graph on m and n vertices, R_n is the two-dimensional $(n \times n)$ grid and C_n is the corresponding three-dimensional grid.

In [11] the complexity of many generating functions is discussed. Among his examples we have also:

Cliques: Let $\mathcal{E}_{Clique}$ be the class of cliques, which is an MS_1 property. By [11], $GF(K_n, \mathcal{E}_{Clique})$ is $\sharp$**P** hard (or **VNP**$_K$-complete).

Maximal Clique: Let $\mathcal{E}_{MaxClique}$ be the class of maximal cliques, which is an MS_1 property. By [42], $GF(K_n, \mathcal{E}_{MaxClique})$ is $\sharp$**P** hard (or **VNP**$_K$-complete).

Perfect Matchings: Let $\mathcal{E}_{PerfM}$ be the class of perfect matchings, which is an MS_2 property. By [42], $GF(C_n, \mathcal{E}_{PerfM})$ is $\sharp$**P** hard (or **VNP**$_K$-complete).

Partial permanent: Let $\mathcal{E}_{PartM}$ be the class of partial matchings, which is an MS_2 property. By [29], both $GF(K_n, \mathcal{E}_{pm})$ and $GF(R_n, \mathcal{E}_{pm})$ are $\sharp$**P** hard (or **VNP**$_K$-complete).

The proof of theorem 7 relies on the observation that in these two (and many more) cases $\mathcal{E}$ is definable in Monadic Second Order Logic. This allows us to apply techniques first used in a more restrictive framework in [1] and extended in [15]. For more details, cf. section 5.

In [25] $(n \times n)$-matrices M over $\{0,1\}$ are considered where one knows in advance that the permanent is bounded by a polynomial, i.e. $per(M) \leq k \cdot n^q$ for some constants $k, q \in \mathbb{N}$. Grigoriev and Karpinski prove that under this assumption $per(M)$ can be computed in $\mathbf{NC^3}$, and hence in $\mathbf{P}$. To the best of our knowledge no similar result is known for $ham(M)$.

More about the complexity of generating functions of graph properties in the BSS framework can be found in [35,34].

4 Feasibility and positivity of polynomial systems

We now want to explore how far these techniques can be pushed further. We will state our main results; proofs then follow in the next sections.

Is the tree-width of a system of polynomials an appropriate tool to decide the existence of zeros?

We look at the following problems:

Definition 8. Let $\mathbb{F}$ be a field (finite or infinite). Let $A \subseteq \mathbb{F}$ be finite of cardinality a. Let $p(x)$ be a polynomial in $\mathbb{F}[x]$ in the variables $x = (x_1, \ldots, x_n)$ of degree d and tree-width k, and Σ be a system of such polynomials, whose $d+1$-hypergraph is of tree-width at most k.

$(d,k) - FEAS_{\mathbb{F}}$: Does p have a zero in $\mathbb{F}^n$?

$(d,k) - FEAS(A)_{\mathbb{F}}$: Does p have a zero in A^n?

$(d,k) - HN_{\mathbb{F}}$: Does Σ have a common zero in $\mathbb{F}^n$?

$(d,k) - HN(A)_{\mathbb{F}}$: Does Σ have a common zero in A^n?

For $\mathbb{F}$ an ordered field we can also ask

$(d,k) - POS_{\mathbb{F}}$: Is $p(r) > 0$ for all $r \in \mathbb{F}^n$?

$(d,k) - POS(A)_{\mathbb{F}}$: Is $p(r) > 0$ for all $r \in A^n$?

If the tree-width is not bounded we write $(d, \infty) - FEAS(A)_{\mathbb{F}}$, etc. If R is an (finite, infinite, ordered) ring rather than a field, we use the analogous notation $(d,k) - FEAS(A)_R$, etc.

Remark 2. The finite set $A \subseteq \mathbb{F}$ can be viewed as condition of zero-dimensionality. A can be encoded by an additional polynomial which has exactly the elements of A as its zeros. Such problems are considered in [39].

Remark 3. The way we have set up the definitions there is a minor discrepancy between $(d,k) - FEAS_{\mathbb{F}}$ and $(d,k) - HN_{\mathbb{F}}$ for systems consisting of one polynomial. We have $(d,k) - FEAS_{\mathbb{F}} \subseteq (d,k+1) - HN_{\mathbb{F}}$, but not necessarily $(d,k) - FEAS_{\mathbb{F}} \subseteq (d,k) - HN_{\mathbb{F}}$. This does not affect the results.

$(d,\infty) - FEAS_{\mathbb{F}}$, $(d,\infty) - HN_{\mathbb{F}}$ and, for ordered fields $(d,\infty) - POS_{\mathbb{F}}$ are discussed in [5,37]. In general they are $\mathbf{NP}_{\mathbb{F}}$ resp. co-$\mathbf{NP}_{\mathbb{F}}$ hard, and no subexponential algorithms are known for their solution. If we relativize the problems to $(d,\infty) - FEAS(A)_{\mathbb{F}}$ and $(d,\infty) - POS(A)_{\mathbb{F}}$ they are in $\mathbf{DNP}_{\mathbb{F}}$ resp. co-$\mathbf{DNP}_{\mathbb{F}}$, the classes *digital* **NP** *resp. digital co-***NP** *over* $\mathbb{F}$.

It is known that $\mathbf{P}_{\mathbb{F}} \subseteq \mathbf{DNP}_{\mathbb{F}} \subseteq \mathbf{NP}_{\mathbb{F}}$ for any field $\mathbb{F}$ but it is not known whether the inclusions are proper, cf. [38,37].

Problem 1. Are the problems $(d,\infty) - FEAS(A)_{\mathbb{R}}$ and $(d,\infty) - POS(A)_{\mathbb{R}}$ $\mathbf{DNP}_{\mathbb{R}}$ resp. co-$\mathbf{DNP}_{\mathbb{F}}$ complete over the reals? Is $(d,\infty) - HN_{\mathbb{C}}(A)$ is $\mathbf{DNP}_{\mathbb{C}}$ complete over the complex numbers?

The problem with proving one of these questions is the following: If we mimic the original proof of $\mathbf{NP}_{\mathbb{R}}$ completeness of $(4,\infty) - FEAS_{\mathbb{R}}$ in [10] it does not apply to $(4,\infty) - FEAS_{\mathbb{R}}(\{0,1\})$ if we restrict to problems in $\mathbf{DNP}_{\mathbb{R}}$. The reason is that intermediate results of a computation are also used as guesses. They might be reals even though the initial guesses are zeros and ones. However, some seemingly hard problems can be reduced to $(4,\infty) - FEAS_{\mathbb{R}}(A)$ for $|A| \geq 2$. This includes for example the real Knapsack problem and, of course, also the classical SAT problem. It is therefore reasonable considering $(4,\infty) - FEAS_{\mathbb{R}}(A)$ to be a difficult problem in the BSS setting. Complete problems for $DNP_{\mathbb{R}}$ do exist ([18]).

Our first main theorem is

Theorem 9. *For finite fields* $\mathbb{F}$ *of size* f *the problems* $(d,k) - FEAS_{\mathbb{F}}$ *and* $(d,k) - HN_{\mathbb{F}}$ *can be solved in time* $O(n)$ *where the constant depends on* k, d, f. *The same holds for finite rings.*

The proof exploits the finiteness of the field (ring) by making all elements of it part of the underlying logic. In this way the problem becomes a problem of the weighted hypergraph of the non-vanishing monomials (with the coefficients as weights). The proof has no particular algebraic content, but shows that the notion of tree-width of systems of polynomials has surprising algebraic applications. Vardi pointed out that, using the methods of [22], the result can be sharpened a little by making it independent of the size of the finite field. Unfortunately, this does not help in the case of infinite structures, cf. the conclusion section.

The same technique as for the proof of Theorem 9 can also be used for infinite fields like $\mathbb{R}$ when dealing with so called $\exists$-MSO$_{\mathbb{R}}$ decision and evaluation problems (to be defined later on). This includes the evaluation of polynomials with exponentially many monomials like the permanent of a matrix of bounded tree-width and many other. We obtain

Theorem 10. *For $\mathbb{R}$-structures of tree-width at most k, $\exists$-MSO$_{\mathbb{R}}$-decision and evaluation problems can be solved in linear time (in the size of the structure).*

We next generalize the discussion of the feasibility and positivity problems to infinite fields. Here it turns out that the tree-width condition alone seems not to be strong enough to allow similar results. We can afford infinite fields (or rings), but restrict the set of tuples for which we want to evaluate the polynomial. This results in decision problems belonging to the class DNP$_{\mathbb{R}}$. Without bounded tree-width no polynomial algorithm is know for these problems.

We then prove

Theorem 11. *a) For arbitrary fields $\mathbb{F}$ the problem $(d,k)-FEAS(A)_{\mathbb{F}}$ can be solved in polynomial time in n where the constants of the polynomial bound depend on k, d and a condition concerning the number of different values the monomials take over A.*

b) For ordered fields $\mathbb{F}$ the problem $(d,k) - POS(A)_{\mathbb{F}}$ can be solved in polynomial time in n where the constants of the polynomial bound depend on k, d and a condition concerning the number of different values the monomials take over A.

The same holds for any ordered ring R.

The condition concerning the coefficients and A basically requests that the set of partial sums of the monomials evaluated in A be bounded by a polynomial function in n.

A similar theorem can be proved also for $(d,k) - PIS(A)_R$, systems of polynomial inequalities of polynomials of bounded degree $d \geq 2$ over an ordered ring R.

The most general result on infinite fields in this paper is on $\exists$-MSO$_{\mathbb{R}}$ extended decision problems (to be defined):

Theorem 12. *For structures of tree-width at most k $\exists$-MSO$_{\mathbb{R}}$-extended decision problems can be solved in time $O(n \cdot t(n))$ provided the number of possible values of the subterms appearing in the formalization of the problem is bounded by $O(t(n))$ where n is the size of the structure.*

Finally, we discuss how the coefficient condition mentioned in Theorems

11 and 12 in some cases can be avoided. A kind of linearization of the positivity problem in fact allows us to strengthen Theorem 11, b). Here, by linearization we mean that the weight terms are linear in the variables. We obtain

Theorem 13. *For ordered fields $\mathbb{F}$ the problem $(d,k)-POS(A)_{\mathbb{F}}$ can be solved in linear time in n where the constants of the linear bound depend on k, d. The same holds for any ordered ring R.*

5 Meta-finite Monadic Second Order Logic

In this and the next section the main mathematical ideas behind the results stated above are developed. Complete proofs are given.

Before going into details let us first outline the overall foregoing: the major point of our approach is to extend the (previously known) handling of MS_2 properties over finite structures to algebraic issues, i.e. decision and evaluation problems as they naturally appear in the Blum-Shub-Smale framework. To this aim we consider problem instances as specific finite, relational structures together with real-valued weight functions. The latter are called $\mathbb{R}$-structures in [24,27]. Problems are then given as conjunction of two formulas; one is expressed in monadic second order logic MS_2 over the underlying finite structure and the other is given in existential monadic second order logic over the corresponding $\mathbb{R}$-structure (a logic to be defined). This generalizes the framework of *Extended Monadic Second order Logic EMSOL* proposed in [1] and unifies it with the framework of *Meta-finite Model Theory* of [24].

On the underlying finite structure we evaluate MS_2 properties using the Feferman-Vaught theorem. The latter is rigorously proved in the Appendix together with its use for algorithmic purposes. Those readers being not familiar with it are strongly encouraged to study the Appendix before continuing.

The results of this paper rely on a careful definition of the meta-finite monadic second order logic for expressing algebraic issues. It has to be strong enough to capture interesting problems; on the other hand, it must be defined in such a way that we can perform a decomposition on the meta-finite structure in parallel to the one given by the bounded tree-width decomposition and the Feferman-Vaught theorem on the underlying finite structure. This will be done now.

5.1 Hypergraphs and parse trees

Let us formalize the kind of meta-finite structures and decompositions of them we are interested in. Some missing proofs are in the Appendix.

We consider problem instances as logical structures representing particular hypergraphs. To capture the combinatorial aspects of a problem we start with a finite relational structure $(V, E, R_1, \ldots, R_\ell)$ of signature τ. Here, (V, E) is a hypergraph, i.e. $V := \{1, \ldots, n\}$ for some $n \in \mathbb{N}$ and any element of E is a (non-void) subset of V (of arbitrary but fixed arity). Every relation R_i is a subset of V^{n_i}, where $n_i \in \mathbb{N}$ denotes its corresponding arity.

We consider $(V, E, R_1, \ldots, R_\ell)$ as a two-sorted structure with universe $V \uplus E$; by convention, there is a relation $R_{inc} \subseteq V \times E$ among the symbols in τ which gives the incidence relation between vertices and edges. If more comfortable, we can also choose $V \uplus V^2$ as the two-sorted universe and include a relation $R_E \subseteq V^2$ representing the edge set, see below.

The monadic second order logic $MS_2(\tau)$ over hypergraphs is defined as sublogic of second order logic, where we allow quantified and free second order variables of arity 1 only. In addition, set variables range over subsets of V or E.

Example 1. (cf. [15]). In order to give an idea about the expressiveness of the logics we are using two examples related to graph properties are studied. As we will see both can be captured by MS_2 logic.

We will be very precise here concerning the logical representation of graphs as structures. Later on, we will just keep in mind how to represent (weighted) hypergraphs as logical structures without going through all the details every time. Firstly, a graph $G = (V, E)$ is represented as a two-sorted finite structure. The universe U of this structure consists of the union of two sorts, namely the vertex set V and the set V^2 of possible edges (sometimes, we also directly take E as second sort). Several relations are included in the structure whose interpretation will give the graph G. More precisely, a unary relation $R_V \subseteq U$ is interpreted as $R_V(x) \Leftrightarrow x \in V$, a unary relation $R_E \subseteq U$ is interpreted as $R_E(e) \Leftrightarrow e \in E$, and a binary incidence relation R_{inc} is interpreted as $R_{inc}(v, e) \Leftrightarrow R_V(v) \wedge R_E(e) \wedge v$ is incident with e. Thus, the graph G is represented as the finite structure $\langle V \cup V^2, R_V, R_E, R_{inc}\rangle$. We can as well add unary predicates either on V or on V^2 (the two sorts); for example, such unary relations might indicate labels of either the vertices or the edges.

Note that for a subset $X \subseteq U$ we can easily express in MS_2 logic that $X \subseteq V$ (resp. $X \subseteq E$) by

$$\forall x \in U \; X(x) \;\Rightarrow\; R_V(x) \text{ resp. } \forall x \in U \; X(x) \;\Rightarrow\; R_E(x)\ .$$

This will be used implicitly throughout the paper.

a) <u>3-Colorability:</u> given a graph G as $\langle V \cup V^2, R_V, R_E, R_{inc}\rangle$ we define a MS_2 formula Φ as follows:

$$\begin{aligned}\Phi \equiv \exists\ & X_1 \subseteq V, X_2 \subseteq V, X_3 \subseteq V \text{ such that}\\ & PART(X_1, X_2, X_3) \wedge NO\text{-}EDGE(X_1) \wedge NO\text{-}EDGE(X_2)\\ & \wedge NO\text{-}EDGE(X_3).\end{aligned}$$

Here, $PART(X_1, X_2, X_3)$ is true if and only if the sets X_1, X_2, X_3 build a partition of V, i.e.

$$\begin{aligned}PART(X_1, X_2, X_3) \equiv \forall v \in V\ (X_1(v) \vee X_2(v) \vee X_3(v))\ \wedge\\ \neg\exists\ u \in V\ (\{X_1(u) \wedge X_2(u)\} \vee \{X_1(u) \wedge X_3(u)\} \vee\\ \{X_2(u) \wedge X_3(u)\})\end{aligned}$$

and $NO\text{-}EDGE(X)$ is true if and only if no two vertices in X are incident with the same edge of G :

$$NO\text{-}EDGE(X) \equiv \forall u, v \in V\ (X(u) \wedge X(v) \Rightarrow \neg E(u, v))$$

Clearly, G is 3-colorable iff $G \models \Phi$. In the above example we do not really need the full power of $MS_2(\tau)$ logic; the quantification is just extending to subsets of the first sort V of the universe. Therefore we could represent G as well as a one-sorted structure only and still express colorability. This will be different with the next example.

b) Another MS_2-definable property is that of being a perfect matching. This time we need set quantifiers over both sorts of the universe. Again, let a graph G being given as $\langle V \cup V^2, R_E, R_V, R_{inc}\rangle$. For a subset $X \subseteq V^2$ we want to find a MS_2 formula $\Phi(X)$ such that $(G, X) \models \Phi(X)$ iff X is a perfect matching of G. Using R_{inc} we can either deal with an edge $e \in E$ or with the two incident vertices; just consider the MS_2 formula

$$\varphi(u, v, e) \equiv R_{inc}(u, e) \wedge R_{inc}(v, e) \wedge u \neq v$$

which gives the incident vertices of an edge. Having this in mind we define the formula

$$\begin{aligned}\Phi(X) \equiv\ & \forall u, v, \in V\ \{(u, v) \in X \Longrightarrow E(u, v)\}\\ \wedge\ & \forall v \in V\ \exists u \in V\ \{(u, v) \in X\ \vee\ (v, u) \in V\}\\ \wedge\ & \forall u, v, w \in V\ \{((u, v) \in X\ \wedge\ (u, w) \in X) \Longrightarrow v = w\}\\ \wedge\ & \forall u, v, w \in V\ \{((v, u) \in X\ \wedge\ (w, u) \in X) \Longrightarrow v = w\}\end{aligned}$$

which satisfies the properties we are looking for.

Besides the combinatorial part of our structures the use of weights in some algebraic structure (ring, field, ordered ring, etc.) has to be incorporated. To simplify our notation we assume here that weights are in the ordered field of real numbers $\mathbb{R}$. Structures of this kind are particular $\mathbb{R}$-structures in the sense of [24,27].

Towards this aim a weight function $C : E \mapsto \mathbb{R}$ is added to the structure, thereby turning it into a meta-finite one. The ordered structure

$(\mathbb{R}, +, *, \leq\ , 1, 0, -1, r_1, \ldots, r_s)$ is included as well. Here, the r_i are fixed real constants.

We could also think about more than one weight function or weights on the vertices, but the above is sufficient for our purposes.

The properties which will be checked on such $\mathbb{R}$-structures are twofold. One is combinatorial and expressed by a $MS_2(\tau)$ formula. The other involves the weight functions and the real number part. It is given as well as a specific monadic second order property, this time defined for the real number part of the structure.

For structures of bounded tree-width one major ingredient of our algorithm is a decomposition. This is first done on the underlying finite structure (the hypergraph without weights).

Definition 14 (cf. [19]). A hypergraph G is k-boundaried if exactly k of its vertices are labelled by $\{1, \ldots, k\}$ (i.e. every label appears with one vertex). The labelled vertices are called the boundary $\delta(G)$ of G.

We next define the gluing operations we are interested in:

a) **create** : this operation creates a $k+1$ boundaried hypergraph with no edges (i.e. $k+1$ vertices all of which are labelled);

b) **join** : the join $\oplus$ operates on two $k+1$ boundaried hypergraphs $G_1 = (V_1, E_1)$ with boundary $\delta(G_1)$ and $G_2 = (V_2, E_2)$ with boundary $\delta(G_2)$. The graph $G_1 \oplus G_2$ is obtained by joining V_1 and V_2 in such a way that those vertices in $\delta(G_1)$ and $\delta(G_2)$ having the same label are identified. The vertices in $\widehat{V_1} := V_1 \setminus \delta(G_1)$ and those in $\widehat{V_2} := V_2 \setminus \delta(G_2)$ obtain an own copy. After this identification the hyperedges E_1 and E_2 are unified;

c) $\mathbf{change}_{i,j}(G)$: in the $k+1$ boundaried hypergraph G labels i and j are interchanged;

d) $\mathbf{add}_{i_1, i_2, \ldots, i_s}(G)$: adds an hyperedge between vertices with labels $i_1, i_2, \ldots, i_s$;

e) $\mathbf{new}_i(G)$: adds a new vertex, labels it i and removes label i from the previously labelled vertex.

Remark 4. Later on we decompose a given structure into substructures using the above operations in an inverse manner. Note that in that situation a decomposition is not always unique. Most important for our purposes is the case where a structure G is divided into substructures A and B such that $G = A \oplus B$. We fix by convention that hyperedges being made of vertices in the common boundary $\delta(A) = \delta(B)$ then are only transferred to substructure A, not to B.

Definition 15. Let G be a $k+1$ boundaried hypergraph. A parse-tree for G is a tree whose vertices are labelled by one of the above operators such that the following holds:

- the leafs are labelled by *create*;
- the branch nodes are labelled by $\oplus$;
- the other nodes are labelled by one of the other operations;
- G is the hypergraph obtained at the root after performing all the operations along the tree bottom-up.

The main relation between parse-trees and hypergraphs of bounded tree-width is the following:

Theorem 16 (cf. [19]). *Let G be a hypergraph of tree-width at most k. Starting from a tree-decomposition of G we can compute in linear time w.r.t. $|G|$ a parse-tree of G.*

A tree-decomposition itself can also be obtained in linear time according to theorem 6.

The Feferman-Vaught theorem gives information about the way $MS_2(\tau)$ properties on a finite structure can be evaluated on substructures if the relation between these structures are given by one of the above operations. We state it here for the join operation $\oplus$. The proof can be found in the Appendix.

Let $\mathfrak{A}$ and $\mathfrak{B}$ be $k+1$ boundaried hypergraphs over τ with universes A and B resp., and boundaries $\delta(A) \subset A, \delta(B) \subset B, \widehat{A} := A \setminus \delta(A), \widehat{B} := B \setminus \delta(B)$. Let $\mathfrak{C} := \mathfrak{A} \oplus \mathfrak{B}$ be the join of $\mathfrak{A}$ and $\mathfrak{B}$.

Let z be an assignment of the variables in a $MS_2(\tau)$ formula $\Phi(\underline{x}, \underline{y}, \underline{w}, \underline{X})$ into the universe C of $\mathfrak{C}$ such that the following is true:

- z maps the variables x_i of block $\underline{x}$ to the set $\widehat{A}$
- z maps the variables y_i of block $\underline{y}$ to the set $\widehat{B}$
- z maps the variables w_i of block $\underline{w}$ to the boundary $\delta(A) = \delta(B)$ (recall that the two boundaries were identified). In particular, $z(w_i)$ is member of both the universes A and B.

Denote by z_A, z_B the assignments with $z_A(X) = X \cap A, z_A(s) = z(s)$ for $s \in A$ and $z_B(X) = X \cap B, z_B(s) = z(s)$ for $s \in B$.

Theorem 17 (Feferman-Vaught). *For every $MS_2(\tau)$ formula $\psi(\underline{x}, \underline{y}, \underline{w}, \underline{X})$ there are finitely many so called Hintikka formulas $h_{1,\alpha}(\underline{x}, \underline{w}, \underline{X})$ and $h_{2,\alpha}(\underline{y}, \underline{w}, \underline{X})$ such that for every τ-structure $\mathfrak{C}$ given as join $\mathfrak{A} \oplus \mathfrak{B}$ and every assignment z as above we have*

$$(\mathfrak{C}, z) \models \psi(\underline{x}, \underline{y}, \underline{w}, \underline{X}) \Leftrightarrow \exists\, \alpha\, (\mathfrak{A}, z_A) \models h_{1,\alpha}(\underline{x}, \underline{w}, \underline{X}) \wedge (\mathfrak{B}, z_B) \models h_{2,\alpha}(\underline{y}, \underline{w}, \underline{X}).$$

If ψ itself is a Hintikka formula then there exist uniquely determined Hintikka formulas h_1 and h_2 such that

$$(\mathfrak{C}, z) \models \psi(\underline{x}, \underline{y}, \underline{w}, \underline{X}) \Leftrightarrow (\mathfrak{A}, z_A) \models h_1(\underline{x}, \underline{w}, \underline{X}) \wedge (\mathfrak{B}, z_B) \models h_2(\underline{y}, \underline{w}, \underline{X}).$$

5.2 MSO logic over $\mathbb{R}$-structures

In order to extend the Feferman-Vaught approach to meta-finite structures we have to look for an appropriate definition of monadic second order logic $\mathrm{MSO}_{\mathbb{R}}$ for these structures. The major task is to define it in such a way that the decomposition operations of section 5.1 extend in a natural manner also to weighted structures and the $\mathrm{MSO}_{\mathbb{R}}$ formulas.

Let us inductively define terms and formulas for obtaining this monadic second order logic over weighted hypergraphs (V, E, C). In each step we first define the class of terms or formulas which will be used. Then, it is shown how the corresponding terms and formulas can be decomposed to the substructures involved. We restrict the presentation again to the $\oplus$ operation. For the other parse operations similar statements hold true.

Simple terms are of the following form:

- for any edge $e \in E$ the expression $C(e)$ is a simple term;
- the constants $r_1, \ldots, r_s \in \mathbb{R}$ are simple terms;
- if t_1 and t_2 are simple terms so are $t_1 + t_2$ and $t_1 \cdot t_2$.

Thus, simple terms are of the form $pol(C(e_1), \ldots, C(e_m))$ for a polynomial pol and edges $e_1, \ldots, e_m$.

Decomposition: For a decomposition $\mathfrak{C} = \mathfrak{A} \oplus \mathfrak{B}$ a term $C(e)$ is taken over to that substructure e belongs to. Note that for $e \in \mathfrak{P}(\delta(A))$ by convention e appears in the second sort of the substructure $\mathfrak{A}$ only. The constants can be used in both substructures. A polynomial $pol(C(e_1), \ldots, C(e_m))$ is evaluated on $\mathfrak{C}$ by first evaluating $C(e_i), 1 \le i \le$

m (and further parts as far as possible) on the corresponding substructure and then putting together the results on $\mathfrak{C}$. Note that the complexity of the latter step only depends on the size of the polynomial *pol*, not on $\mathfrak{C}$.

Summation and product terms are expressions of the following form:

- for a subset $U \subseteq E$ a summation term has the form $T(U) := \sum_{e \in U} C(e)$ and a product term has the form $T(U) := \prod_{e \in U} C(e)$; for the empty set we can define $T(\emptyset)$ to be any fixed real value.
- if $T_1(U_1)$ and $T_2(U_2)$ are summation and product terms so are $T_1(U_1) + T_2(U_2)$ and $T_1(U_1) \cdot T_2(U_2)$.

Decomposition: For $\mathfrak{C} = \mathfrak{A} \oplus \mathfrak{B}$ and $T(U) = \sum_{e \in U} C(e)$ denote $U^A := U \cap E^A$ and $U^B := U \cap E^B$. Then $T_1(U^A) := \sum_{e \in U^A} C(e)$ and $T_2(U^B) := \sum_{e \in U^B} C(e)$ can be evaluated on $\mathfrak{A}$ and $\mathfrak{B}$ resp. Their sum gives the result of T on $\mathfrak{C}$. The same holds true for $T(U) := \prod_{e \in U} C(e)$. Polynomials in summation or product terms are evaluated as it was explained for simple terms.

Remark 5. Note that according to our convention $E^A \cap E^B = \emptyset$. Thus, there will be no hyperedge taken into account twice. We could have modeled the decomposition by including all hyperedges relating vertices from the boundary in both substructures and giving them the weight 0 in one of it. But this would raise some technical problems; for example, in a product term we would obtain the value 0 above.

Min/max terms are defined as summation and product terms but using

$$T(U) := \max_{e \in U} C(e) \text{ and } T(U) := \min_{e \in U} C(e)$$

instead of $\sum$ and $\prod$.

Monadic second order $\mathrm{MSO}_{\mathbb{R}}$ **terms** are all terms above together with terms of the following form:

- for a summation or product term of the form $T(U)$ (not a polynomial in such terms!) and a MS_2 formula ϱ on the underlying finite structure the term $\sum_{U, \varrho(U)} T(U)$ is a $\mathrm{MSO}_{\mathbb{R}}$ term;

- for a summation or product term of the form $T(U)$ and a univariate polynomial *pol* the term $\sum\limits_{U,\varrho(U)} pol(T(U))$ is a $\mathrm{MSO}_{\mathbb{R}}$ term;
- for $1 \leq i \leq m$ let $T_i(U_i) := \prod\limits_{e \in U_i} C(e)$ be product terms and $\varrho(X_1, \dots, X_m)$ be a MS_2 formula on the underlying finite structure. Then $\sum\limits_{(U_1,\dots,U_m),\varrho(\underline{U})} T_1(U_1) \cdot \ldots \cdot T_m(U_m)$ is a $\mathrm{MSO}_{\mathbb{R}}$ term;
- sums and products of $\mathrm{MSO}_{\mathbb{R}}$ terms are $\mathrm{MSO}_{\mathbb{R}}$ terms.

Decomposition: Let $\mathfrak{C} = \mathfrak{A} \oplus \mathfrak{B}$ and consider as example a $\mathrm{MSO}_{\mathbb{R}}$ term of the form $\sum\limits_{(U_1,\dots,U_m),\varrho(\underline{U})} T_1(U_1) \cdot \ldots \cdot T_m(U_m)$, where all T_i are product terms. The evaluation on $\mathfrak{C}$ uses the Feferman-Vaught theorem for ϱ as well as evaluations on $\mathfrak{A}$ and $\mathfrak{B}$ as follows. W.l.o.g. suppose ϱ to be a Hintikka formula and h_1 as well as h_2 to be the corresponding unique formulas for $\mathfrak{A}$ and $\mathfrak{B}$ given by theorem 17. Let $\underline{U}_j := (U_{1j}, \dots, U_{mj}), j \in J$ be those assignments with $(\mathfrak{C}, \underline{U}_j) \models \varrho(\underline{U}_j)$. Furthermore, let $\underline{U}_j^A$ and $\underline{U}_j^B$ be its corresponding decompositions for the substructures $\mathfrak{A}$ and $\mathfrak{B}$, see the definition of $\oplus$ and remark 4.

Now

$$\begin{aligned}
&\sum_{\underline{U},\varrho(\underline{U})} T_1(U_1) \cdot \ldots \cdot T_m(U_m) = \sum_{j \in J} T_1(U_{1j}) \cdot \ldots \cdot T_m(U_{mj}) \\
&= \sum_{j \in J} T_1(U_{1j}^A) \cdot T_1(U_{1j}^B) \cdot \ldots \cdot T_m(U_{mj}^A) \cdot T_m(U_{mj}^B) \\
&= \sum_{j \in J} T_1(U_{1j}^A) \cdot \ldots \cdot T_m(U_{mj}^A) \cdot T_1(U_{1j}^B) \cdot \ldots \cdot T_m(U_{mj}^B) \\
&= \left(\sum_{\underline{U}^A, h_1(\underline{U}^A)} T_1(U_1^A) \cdot \ldots \cdot T_m(U_m^A) \right) \cdot \\
&\quad \left(\sum_{\underline{U}^B, h_2(\underline{U}^B)} T_1(U_1^B) \cdot \ldots \cdot T_m(U_m^B) \right)
\end{aligned}$$

where the latter equality holds because of the Feferman-Vaught theorem. For the other types of $\mathrm{MSO}_{\mathbb{R}}$ terms the decomposition can be done in the same manner.

Remark 6. The above definition of $\mathrm{MSO}_{\mathbb{R}}$ terms probably is not the most general one. However, the reader might check the difficulties arising when performing a decomposition for a term like $\prod\limits_{\underline{U},\varrho(\underline{U})} \sum\limits_{e \in U} C(e)$ or one like $\sum\limits_{(U_1,U_2),\varrho(U_1,U_2)} T_1(U_1)+T_2(U_2)$ where T_1 and T_2 are product terms. The above setting nevertheless is strong enough to capture important problems.

For the other operations $op \in \{add_{i_1,\ldots,i_s}, change_{ij}, new_i\}$ the evaluation of a $\mathrm{MSO}_{\mathbb{R}}$ term on a structure $op(\mathfrak{C})$ can be done similarly by first evaluating a corresponding term on $\mathfrak{C}$ and then extending the result to $op(\mathfrak{C})$.

After having introduced monadic second order terms we turn to monadic second order formulas built from these terms.

Basic $\mathrm{MSO}_{\mathbb{R}}$ formulas are expressions of the form $T(v_1 \ldots, v_k, U_1, \ldots, U_t)\Delta 0$ with $\Delta \in \{=, >, \geq\}$; here, T is a monadic second order term, the v_i are elements of the universe and the U_i are subsets of E.

$\mathrm{MSO}_{\mathbb{R}}$ formulas are all basic $\mathrm{MSO}_{\mathbb{R}}$ formulas together with

- if $\psi(v_1, \ldots, v_k, W_1, \ldots, W_t)$ is a $\mathrm{MSO}_{\mathbb{R}}$ formula and $U_1, \ldots, U_k$ are subsets of E, then $\varrho(W_1, \ldots, W_t) \equiv \bigwedge\limits_{v_i \in U_i} \psi(v_1, \ldots, v_k, W_1, \ldots, W_t)$ is a $\mathrm{MSO}_{\mathbb{R}}$ formula;
- if $\psi(v_1, \ldots, v_k, U_1, \ldots, U_s, W_1, \ldots, W_t)$ is a $\mathrm{MSO}_{\mathbb{R}}$ formula and $\varrho(v_1, \ldots, v_k, U_1, \ldots, U_s)$ is a $MSOL(\tau)$ formula then $\phi(W_1, \ldots, W_t) \equiv \bigwedge\limits_{\underline{v},\underline{U},\varrho(\underline{v},\underline{U})} \psi(\underline{v}, \underline{U}, W_1, \ldots, W_t)$ is a $\mathrm{MSO}_{\mathbb{R}}$ formula (for $W_i \subseteq E$);
- the closure of the above construction scheme under logical conjunction, disjunction and negation gives the set of $\mathrm{MSO}_{\mathbb{R}}$ formulas.

$\exists$-$\mathrm{MSO}_{\mathbb{R}}$ (Existential monadic second order logic) over $\mathbb{R}$-structures is obtained from $\mathrm{MSO}_{\mathbb{R}}$ logic by defining

- all formulas in $\mathrm{MSO}_{\mathbb{R}}$ to belong to $\exists$-$\mathrm{MSO}_{\mathbb{R}}$
- if $\psi(W_1, \ldots, W_t) \in \exists$-$\mathrm{MSO}_{\mathbb{R}}$ then $\exists W_1, \ldots, W_t\ \psi(W_1, \ldots, W_t) \in \exists$-$\mathrm{MSO}_{\mathbb{R}}$.

The problems considered in this paper are of the following type:

Decision problems: For a fixed MS_2 formula ψ, decide whether $\mathcal{G} \models \psi$.

Evaluation problem: For a fixed $\mathrm{MSO}_{\mathbb{R}}$ term T, compute its value over $\mathcal{G}$.

Extended decision problem: Given a $MS_2(\tau)$ formula ψ as well as a $\exists$-MSO$_\mathbb{R}$ formula Φ we want to decide whether $\mathcal{G} \models \psi \wedge \Phi$.

Optimization problems: These are like the extended decision problems, *but with Φ quantifier free, but possibly involving the functions max and min.*

For other fields $\mathbb{F}$ the definition of $\exists$-MSO$_\mathbb{F}$ is done on a similar way.

6 Proofs

After clarifying the above way to define $\exists$-MSO$_\mathbb{R}$ logic by dealing with some examples we will turn to rigorous proofs of the theorems stated in section 4.

6.1 Guiding examples

Let us consider some examples and the way they fit into the formal setting of the previous section.
Example 2. The generating functions of section 3

$$GF(G, \mathcal{E}) =_{def} \sum \{w(E') : \langle V, E' \rangle \in \mathcal{E} \text{ and } E' \subseteq E\}$$

with $\mathcal{E}$ $MSOL$-definable by $\psi(E')$ can be written as a Monadic Second Order Term

$$\sum_{\psi(E') \wedge E' \subseteq E} \prod_{e \in E'} w(e)$$

Hence they are $\exists$-MSO$_\mathbb{R}$-evaluation problems.
Example 3 (#4-$FEAS(A)$). In order to formalize the #4-$FEAS(A)$ problem as an $\exists$-MSO$_\mathbb{R}$ extended decision problem we use the representation of degree 4 polynomials as given in [27]: let $V = \{0, 1, \ldots, n\}, E := V^4$ and $C : E \to \mathbb{R}$ be a weight function giving the coefficients of f in the following sense: for $(i, j, k, l) \in E$ the value $C(i, j, k, l)$ is the coefficient of the monomial $x_i \cdot x_j \cdot x_k \cdot x_l$ in f. Note that here we assume f to be homogeneous of degree 4. Later on f is dehomogenized by adding the condition $x_0 := 1$.

Let $A := \{s_1, \ldots, s_m\}$; for every $1 \leq i \leq m$ define a function $w_i : V \to A$ such that $\forall x \in V \; w_i(x) := s_i$. We are looking for disjunct subsets $U_1, \ldots, U_m$ of V such that the following holds: $\bigcup_{i=1}^{m} U_i = V$ and if we assign to every $x \in U_i$ the value $w_i(x) = s_i$ then this assignment of variables gives a zero of f.

For any quadruple $\lambda = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ where $\varepsilon_i \in \{1, \ldots, m\}$ define the set $E_\lambda := E \cap (U_{\varepsilon_1} \times U_{\varepsilon_2} \times U_{\varepsilon_3} \times U_{\varepsilon_4})$ (i.e. a point $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in V^4$ belongs to E_λ iff every component α_i lies in U_{ε_i}).

For $\alpha \in E_\lambda$ the corresponding monomial $C(\alpha) \cdot x^\alpha$ gives the value $C(\alpha) \cdot s_{\varepsilon_1} \cdot s_{\varepsilon_2} \cdot s_{\varepsilon_3} \cdot s_{\varepsilon_4}$ under the above assignment. The $MS_2(\tau)$ formula for #4-$FEAS(A)$ simply is

$$\psi \equiv \exists U_1, \ldots, U_m \subseteq V \text{ such that } \forall i, j \in \{1, \ldots, m\} U_i \cap U_j = \emptyset \wedge U_1 \cup \ldots \cup U_m = V$$

(note that m is independent of the structure).

The real number part of the logical description is $\Phi \equiv \sum\limits_{\lambda \in \{1, \ldots, m\}^4} T(E_\lambda) = 0$, where $T(E_\lambda) := \sum\limits_{\alpha \in E_\lambda} C(\alpha) \cdot s_{\varepsilon_1} \cdot s_{\varepsilon_2} \cdot s_{\varepsilon_3} \cdot s_{\varepsilon_4}$ is a summation/product term. Again note that m is independent of the structure; hence, the same is true for the size of the first sum above. For dehomogenization one can introduce a further subset $U_0 \subseteq V$ which only consists of the element 0 and put $s_0 := 1$. It follows that f has a zero in A if and only if $(V, E, C) \models \psi \wedge \Phi$.

Note that over a *finite* field $\mathbb{F}$ this logical description works as well taking $A := \mathbb{F}$, but could of course be simplified. We used the above description in order to cover already the infinite field case we are interested in later on.

In a completely similar fashion the feasibility of polynomial inequality systems $2\text{-}PIS(A)_{\mathbb{R}}$ fits into the framework of $\exists$-MSO$_{\mathbb{R}}$ extended decision problems and $(d, \infty) - POS(A)_{\mathbb{F}}$ can be coded as an optimization problem. If the ring (field) is finite the coding yields an $\exists$-MSO$_{\mathbb{R}}$ decision problem.

Example 4. Another example fitting into the framework is the computation of the determinant of a square matrix $A \in \mathbb{R}^{n \times n}$ (whose computation is not that interesting in its own but in relation with more complicated problems like, for example, the linear programming problem). It can be expressed as a MSO$_{\mathbb{R}}$ term as follows.

Since A is not necessarily symmetric we represent it as a directed weighted graph (G, c) with vertices $V := \{1, \ldots, n\}$. The weight function $c : V^2 \to \mathbb{R}$ for $(i, j) \in V^2$ gives the value $c(i, j) := a_{ij}$. This implies that in the model theoretic representation of A we use a vocabulary τ with two incidence relations R_1, R_2 instead of R_{inc} (and the obvious meaning: $R_1(i, e) \Leftrightarrow i$ is the start node for edge e). Moreover, we include a linear order $<$ on V (with the same interpretation as the natural order $1 < 2 < \ldots < n$). The interested reader might try to define such an order in MS_2 logic.

In the first step we now define a formula $PERM(\Pi)$ such that $(G, \Pi) \models PERM(\Pi)$ iff $\Pi \subseteq V^2$ is a permutation of V (where $(i, j) \in \Pi$ stands for $\Pi(i) = j$). This formula can easily be written down in MS_2 logic.

Secondly, for a permutation Π we define the set $INV(\Pi) \subseteq V$ of its inversions by

$$i \in INV(\Pi) \Leftrightarrow \exists j \in V : (i,j) \in \Pi \wedge i > j \,.$$

Here, we use the linear order on V. Thirdly, we need two product terms T_1 and T_2; the first is defined on subsets U of V :

$$T_1(U) := \prod_{i \in U}(-1) \qquad (\text{and } T_1(\emptyset) := 1);$$

the second is defined on subsets Π of V^2 (the second sort of the universe)

$$T_2(\Pi) := \prod_{(i,j)\in\Pi} c(i,j) \,.$$

If now $\varrho(U,\Pi)$ is the MS_2 formula expressing

$$\varrho(U,\Pi) \equiv U \subseteq V \ \wedge \ \Pi \subseteq V^2 \ \wedge \ U = INV(\Pi) \ \wedge \ PERM(\Pi)$$

we obtain the determinant of A as

$$DET(A) \ = \sum_{U,\Pi,\varrho(U,\Pi)} T_1(U) \cdot T_2(\Pi)$$

which is a $\mathrm{MSO}_{\mathbb{R}}$ term according to the definition of section 5.2.

Although the determinant of a matrix can be computed quickly, the example can be modified, such as to include the permament and other matrix functions, which are #P hard to compute, cf. [15].

We will now turn to the proof of Theorem 9.

Proof of Theorem 9:

We restrict ourselves to the case $(4,k) - FEAS_{\mathbb{F}}$. The other cases can be treated in the same way. Let the input polynomial be represented as in example 3 by a structure $\mathfrak{D} = (V,E,C)$ of tree-width at most k and size n; let ψ and Φ be the corresponding MS_2 and $\exists$-$\mathrm{MSO}_{\mathbb{F}}$ formulas.

Perform the Feferman-Vaught decomposition of $\mathfrak{D}$ and ψ according to theorem 17. Suppose ψ to be a disjunction $\bigvee_{\gamma} h_\gamma$ of Hintikka formulas h_γ. We check for each of it whether $\mathfrak{D} \models h_\gamma \wedge \Phi$ holds. With respect to the MS_2 formulas h_γ we do this as shown in the Appendix: We precompute all triples of Hintikka formulas of a given quantifier rank and number of variables which are linked by the Feferman-Vaught theorem according to the gluing operation $\oplus$. The corresponding tables are precomputed as well for the other operations. Climbing up the parse tree for $\mathfrak{D}$ by looking into these precomputed tables we check in linear time whether $\mathfrak{D} \models h_\gamma$.

For the $\exists$-MSO$_{\mathbb{F}}$ formula Φ we proceed as follows: first, we decompose the MS_2 formulas involved in Φ in the same way using Theorem 17. Next, we begin the evaluation at the leave structures of the parse tree for $\mathfrak{D}$. Since all of them have a universe of size at most k we can use a brute algorithm which evaluates the corresponding MSO$_{\mathbb{R}}$ terms for the leave structures by considering all possible choices for second order variables. This takes exponential time only in the fixed parameter k. We store the computed results in order to use them again on the further structures appearing when we glue substructures and climb up the parse tree.

Note that there are at most $|\mathbb{F}|$ many different subresults, i.e. a constant number.

We continue along the parse tree until the root, representing $\mathfrak{D}$, is reached. We check the possible results of the evaluation (which are at most $|\mathbb{F}|$ many) and decide whether $\mathfrak{D} \models h_\gamma \wedge \Phi$ for all γ. Climbing up the tree according to the decomposition has running time $O(n)$.

A technical problem appears during the evaluation of MSO$_{\mathbb{F}}$ terms. Here, joining two substructures $\mathfrak{A}$ and $\mathfrak{B}$ and identifying their boundaries makes it necessary to consider only assignments treating the boundary elements the same in $\mathfrak{A}$ and $\mathfrak{B}$. If we decompose a structure top down that's of course no problem. But in the above algorithm we compute bottom up. Suppose we want to evaluate a term $\sum\limits_{\underline{U},\varrho(\underline{U})} T(\underline{U})$ over $\mathfrak{A} \oplus \mathfrak{B}$ using the Feferman-Vaught theorem. Suppose furthermore w.l.o.g. ϱ to be a Hintikka formula and

$$(\mathfrak{A} \oplus \mathfrak{B}, \underline{U}) \models \varrho(\underline{X}) \;\Leftrightarrow\; (\mathfrak{A}, \underline{U}^A) \models h_1(\underline{X}) \;\wedge\; (\mathfrak{B}, \underline{U}^B) \models h_2(\underline{X}) \;.$$

If we evaluate bottom up then only assignments $\underline{U}^A, \underline{U}^B$ can be combined which give the same pattern for elements in the boundary of the two structures. That is, only if $\forall x \in \delta(A) = \delta(B)$ and $\forall i \in \{1, \ldots, m\}$ $x \in U_i^A \Leftrightarrow x \in U_i^B$. This control can be organized either by enlarging the vocabulary or by keeping track of the possible patterns arising when checking which boundary elements belong to which of the U_i. Because at every step there are only $k+1$ boundary values this bookkeeping does not effect the overall linear running time. □

For decision and evaluation problems even over infinite fields the above proof can be adapted as well:

Proof of Theorem 10:

Let $\mathfrak{D} = (V, E, C)$ be a $\mathbb{R}$-structure of bounded tree-width. If we are dealing with decision problems we just have to check the validity of $\mathfrak{D} \models \psi$ for a MS_2-formula ψ. We do that in the same manner as explained in the proof of Theorem 9. If a MSO$_{\mathbb{R}}$ term T has to be computed over $\mathfrak{D}$, on every substructure of the tree decomposition we evaluate that term corresponding to the decomposition of T; climbing up the tree we combine the two values computed at the two substructures to obtain the value on the joined structure. This needs constant time for every substructure in the decomposition. □

6.2 Extended decision problems over infinite fields

If we consider extended decision problems like the feasibility problem over infinite fields there arise some problems with the tree decomposition approach.

The first observation is that in $\exists$-MSO$_{\mathbb{R}}$ logic we cannot quantify over the real numbers. Thus, at first sight it is not clear whether the existence of a real zero of a polynomial can be expressed in this logic. We therefore restrict ourselves to ask for zeros the components of which belong to a fixed finite subset $A \subset \mathbb{R}$ (cf. Remark 2). Nevertheless, there are still many interesting and potentially hard problems captured in this situation, recall the discussion in section 4. Furthermore it turns out that for some problems we have to require an additional condition on the values of the weight function. The necessity of such a condition will become evident in the proof of Theorem 11. A precise definition is then provided after it.

Proof of Theorem 11:
We again restrict to the case $(4,k)-FEAS(A)$ and the representation of an input polynomial by $\mathfrak{D} = (V, E, C)$ as in Example 3. Let ψ and Φ be the corresponding formulas in MS_2 and $\exists$-MSO$_{\mathbb{R}}$. The decomposition of $\mathfrak{D}, \psi$ and Φ is done as in the proof of Theorem 10. The same is true for checking $\mathfrak{D} \models h_\gamma$ for at least one of the Hintikka formulas ψ decomposes into. The validity of an additional condition has to be required when climbing up the parse tree in order to evaluate MSO$_{\mathbb{R}}$ terms in Φ. The number of different intermediate values which are taken by the MSO$_{\mathbb{F}}$ terms during the evaluation process has to be bounded by a polynomial in the size of the input-structure.

Note that in the current situation this condition is satisfied if the set of partial sums of the monomials evaluated on the finite set A is bounded by a function in $O(p(n))$ for p a polynomial. This is true because the decomposition of Φ according to the definition of MSO$_{\mathbb{R}}$ logic in the previous section implies that the MSO$_{\mathbb{R}}$ terms evaluated on substructures of the input polynomial precisely are such partial sums.

At the root of the parse tree we check all possible results of the evaluation, which are now polynomially many. The rest of the proof again works as the one of Theorem 9.

The proof of part b) works precisely the same. Instead of checking whether at least one result is zero we only have to check the positivity of all results (which are at most polynomially many).

□

The above proof substantiates the interest in the following condition on a meta-finite structure $\mathfrak{D}$ together with a MS_2 formula ψ and a $\exists$-MSO$_{\mathbb{R}}$ formula Φ :

Definition 18. Let p be a polynomial and $\mathfrak{D}, \psi, \Phi$ as above. The triple $(\mathfrak{D}, \psi, \Phi)$ satisfies the "coefficient condition" w.r.t. polynomial p if the number of intermediate results of MSO$_{\mathbb{R}}$ terms which have to be evaluated when applying the above algorithm is bounded by $p(size(\mathfrak{D}))$.

In some situations we will restrict ourselves to structures $\mathfrak{D}$ of bounded tree-width which in addition satisfy the coefficient condition with respect to a given polynomial and the decision problem $\mathfrak{D} \models \psi \wedge \Phi$ we are interested in.

Let us comment on the coefficient condition. Since it depends on ψ and Φ as well as on the used algorithmic implementation of the Feferman-Vaught theorem it might be possible to strengthen the results by carefully considering the formulation of a problem and the evaluation process. However, this might result in quite complicated conditions. For example, consider the proof of theorem 11. Instead of dealing with all partial sums of monomials it would actually be sufficient to take into account only those appearing along the decomposition. Nevertheless, it seems to be more practical to look for sufficient conditions like the one mentioned in the proof.

It is not always obvious whether the additional coefficient condition on the partial sums of a polynomial implies a real restriction. For an example like the real Knapsack problem (given n real numbers $x_1, \ldots, x_n$, is there a subset $S \subseteq \{1, \ldots, n\}$ s.t. $\sum\limits_{i \in S} x_i = 1$) it implies a serious restriction because under the additional hypothesis that all sums $\sum\limits_{i \in \tilde{S}} x_i$ only take polynomially many different values the problem lies in P$_{\mathbb{R}}$. On the other hand, if we reduce a problem like 3-SAT to an instance of $4 - FEAS(\{0,1\})$ the coefficient condition is automatically fulfilled because of the lemma below. Thus, in related situations our approach gives linear algorithms on structures of bounded tree-width without any additional assumption on the weights.

Lemma 1. *Let q be a polynomial, $A \subset \mathbb{R}^n$ finite and $d \in \mathbb{N}$. Then for all polynomial functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ of degree d which only have integer coefficients in $\{-q(n), -q(n)+1, \ldots, q(n)\}$ the number of different values for the partial sums of the monomials of f if evaluated on A is bounded by $O(N \cdot q(n))$. Here, N is the number of non-zero monomials in f.*

Proof. For fixed degree d the normalized monomials $x_{i_1} \cdot \ldots \cdot x_{i_d}, i_1, \ldots, i_d \in$

$\{1,\ldots,n\}$ of f only take polynomially in $|A|$ many different values if evaluated in elements of A. Let S be the set of these results. Each value in S is multiplied by the sum of the coefficients of those monomials actually giving the value (or by 0). Since the number of monomials is bounded by a N (which itself is bounded by $O(n^d)$) every such sum gives an integer in $\{-q(n)\cdot N,\ldots,q(n)\cdot N\}$. Thus, there are only polynomially many different possibilities to multiply one of the finitely many elements in S by a particular sums of the coefficients. The assertion follows. □

Recall that in particular for polynomials of bounded tree-width the number of non-zero monomials is linear in the variable number.

From the previous proofs it should be clear that Theorem 12 summarizes the general way for obtaining polynomial time algorithms by exploiting the above methods. We therefore omit its proof.

6.3 Getting around the coefficient condition

For some of the problems studied above it is possible to avoid the coefficient condition over infinite fields. The ideas are already present in [1] (in an automata theoretic framework) and [14] (in a logical framework). Therefore we just outline how some of these problems can be putted into their framework.

The problems we can handle that way have to be optimization problems where the objective function has a linear structure. In the framework of weighted hypergraphs $\mathfrak{D} = (V, E, C)$ linear structure means that we want to compute

$$\min_{\phi(\underline{X})} \sum_{t=1}^{s} a_t \cdot |X_t| \ , \ \text{where } \underline{X} = (X_1, X_2, \ldots, X_s).$$

Here, ϕ is a MS_2 formula and $|X_t|$ is an abbreviation for $\sum\limits_{e \in X_i} C(e)$. In such a situation analyzing the Feferman-Vaught theorem for all of the parse operations shows that the special linear structure of the evaluation term (which is a $\text{MSO}_{\mathbb{R}}$ term according to our definition) allows to compute the minimum (or maximum) of a $\mathbb{R}$-structure obtained after applying a parse operation in constant time from the corresponding extremal values on the substructure (see [1,14]). In particular, we do not have to store too many intermediate results. This ideas can be applied to the $POS(A)$ problem:

Theorem 10. *For ordered fields $\mathbb{F}$ the problem $(d,k){-}POS(A)_{\mathbb{F}}$ can be solved in linear time in n where the constants of the linear bound depend on k, d.*

Proof. We restrict ourselves to proving how $(d,k)-POS(A)$ can be expressed as a linear optimization problem in the above sense. The linear algorithm can then be obtained by the algorithm given in the proofs of Theorems 9 and 11 and the above remarks.

Let $A := \{s_1, \ldots, s_m\} \subset \mathbb{R}$ and $\mathfrak{D} = (V, E, C)$ be a $\mathbb{R}$-structure representing a polynomial f in n variables of degree d. Obviously, $f(x) > 0 \; \forall x \in A^n$ iff $\min\{f(x) | x \in A^n\} > 0$. We enlarge $\mathfrak{D}$ by d incidence relations $R_1, \ldots, R_d \subset V \times E$ having the interpretation $R_i(j,e) \;\; \Leftrightarrow\;$ the i-th factor in monomial $e \in E$ is x_j. Note that for this enlarged structure the tree-width concept and the Feferman-Vaught approach work similarly as before.

In order to make the problem linear we consider all (at most m^d many) values $a_1, \ldots a_{m^d}$ which can be obtained from A by multiplying d elements. The a_t thus give possible values of a monomial (with coefficient 1) if an assignment from A^n is chosen.

For any a_t there is a set $\mathfrak{P}_t$ of finitely many patterns in $\{1, \ldots, m\}^d$ such that for every $(k_1, \ldots, k_d) \in \mathfrak{P}_t$ we have $s_{k_1} \cdot \ldots \cdot s_{k_d} = a_t$. Now consider m^d subsets $X_t \subset E$. The intended meaning of the MS_2 formula $\phi(X_1, \ldots, X_{m^d})$ we want to be fulfilled on $\mathfrak{D}$ is a decomposition of all monomials in E into sets X_t such that there is a minimizing assignment which for every monomial $e \in X_t$ gives the value $a_t, 1 \leq t \leq m^d$. Thus

$$\begin{aligned} \phi(X_1, \ldots, X_{m^d}) \equiv \; & \exists U_1, \ldots, U_m \subset V, \bigcup_i U_i = V, \text{ all disjoint} \\ & \bigwedge \; \forall 1 \leq t \leq m^d \; \forall \, i_1, \ldots, i_d \in V : \\ & \text{if } e \in X_t \text{ and } R_1(i_1, e) \wedge \ldots \wedge R_d(i_d, e) \\ & \text{then } (i_1, \ldots, i_d) \in \mathfrak{P}_t \end{aligned}$$

If $(\mathfrak{D}, \underline{X}) \models \phi(\underline{X})$ then there is an assignment $(U_1, \ldots, U_m)$ for $x_1, \ldots, x_n$ from A^n which is compatible with $\underline{X}$ in the sense that for every monomial $e \in X_t$ the assignment given by the U_i yields the corresponding value a_t. Finally, the linear objective function to be minimized is $\sum\limits_{t=1}^{m^d} a_t \cdot \sum\limits_{e \in X_t} C(e)$. □

7 Conclusions and further research

We have shown how the concept of tree-width of multivariate polynomial systems can be used in finding polynomial time algorithms for some otherwise difficult computational problems provided the tree-width is bounded by a constant. The method has further extensions to linear and quadratic programming, previously analyzed in [36]. This will be worked out in a future paper.

Our proofs use a detour through logic. An alternative route would be through automata theory, as in [1]. But automata theory and Monadic Second Order Logic are just two faces of the same definability phenomenon, cf. [17,16,41].

Using the methods developed in [21,22,31] theorem 9 can be improved to

Theorem 19. *For finite rings R of size $r, (d,k) - FEAS_R$ and $(d,k) - HN_R$ can be solved in time $O(n \cdot r)$ where the constant depends on k, d.*

The point is to view feasibility of systems of polynomials over finite structures as *constraint satisfaction problems.* However, this method does not give improvements for infinite structures.

Another direction of further research is the restriction of our feasibility problems to finite subsets A for the components of possible zeros.

Using very effective versions of quantifier elimination over the reals $\mathbb{R}$ one might ask:

Problem 2. Can $\mathbb{R}$ (d,k) – $POS_{\mathbb{R}}$ and (d,k) – $FEAS_{\mathbb{R}}$ be solved in time $O(n)$ over the reals, where the constant in $O(n)$ depends on k, d only?

For the best known algorithms to solve $(d,\infty) - FEAS_{\mathbb{R}}$ and for quantifier elimination the reader should consult the surveys [9,40,3].

It remains a challenging problem to find direct algebraic proofs and to overcome the limitations imposed by our coding technique.

8 Appendix

In this Appendix we want to give a rigorous presentation of the results used in the previous sections. This will make the paper self-contained; though well-known to logicians people in complexity theory of algebraic problems might not be that familiar with the Feferman-Vaught theorem. However, a lot of different results and concepts have to be put together which to our knowledge are hard to find at a single place in literature. So even for readers familiar with the background we believe it might be useful having all this available in an Appendix.

Moreover, we adapted the proofs to the framework in which we need the corresponding statements.

8.1 MSO logic over finite structures

We suppose the reader to be familiar with the notions of a finite structure and second order logic SOL over a given signature (see [23]). the finite structures we are interested in are hypergraphs. These are two-sorted structures the universe of which consists V and E, the vertices and hyperedges respectively. In addition, there is one binary relation symbol R_{inc} for the incidence relation between edges and vertices. (This relation is not always needed, for example when dealing with the zero-existence problem for a polynomial). In case E is the set of edges of a directed graph we might also include 'another binary relation in order to distinguish the direction of an edge. Furthermore, we allow an arbitrary but finite number of constant symbols and unary predicate symbols. In MSO logic the set variables range over subsets of V or E only. We denote the vocabulary by τ and the MSO logic over τ by $MS_2(\tau)$ (the index "2" indicates the two sorts), see also [15].

8.2 Hintikka formulas; the Fraisse-Hintikka theorem

Fix a relational, finite vocabulary τ as above and consider the monadic second order logic $MS_2(\tau)$ over τ. In $MS_2(\tau)$ we are interested in formulas with first-order variables among $x_1, \ldots, x_{n_1}$ and second-order variables among $X_1, \ldots, X_{n_2}$. Later on it will be convenient to split the block $x_1, \ldots, x_{n_1}$ into further blocks.

Proviso: In order to avoid confusion by mixing too many things we reduce our description of theorem 20 and 17 to $MS_2(\tau)$ formulas with free first order variables for vertices only and free second order variables for subsets of vertices.

However, it should be clear from the presentation that a generalization including the edges as second sort of our structures is straightforward from that.

Let $n := n_1 + n_2$. We denote by $F^{\tau}_{n,r}$ the set of $MS_2(\tau)$ formula with n variables and quantifier rank at r (the quantifier rank is the maximal number of nested quantifiers in a formula).

Theorem 20. *Let τ be a signature. For each $r, n \in \mathbb{N}$ we can effectively find a finite set $H_{n,r}$ of unnested formulas in $F^{\tau}_{n,r}$ such that the following is true:*

a) for every τ-structure $\mathfrak{A}$ and for every tuple $(\underline{a}, \underline{M}) := (a_1, \ldots, a_{n_1}, M_1, \ldots, M_{n_2})$, where the a_i are elements of the finite universe A of $\mathfrak{A}$ and the M_j are subsets of A there exists exactly one formula $\theta \in H_{n,r}$ such that $(\mathfrak{A}, \underline{a}, \underline{M}) \models \theta(\underline{a}, \underline{M})$.

b) For every $r \in \mathbb{N}$ and every un-nested formula $\phi \in F^{\tau}_{n,r}$ there is a disjunction $h_1 \vee \ldots \vee h_m$ of formulas $h_i \in H_{n,r}$ which is equivalent to ϕ for any τ-structure. The h_i can be computed effectively.

Proof. We follow closely [28], where this theorem is established for first-order logic.

Ad a) We define the sets $H_{n,r}$ recursively with respect to the parameter r. First, consider all unnested atomic formula in n variables. There are only finitely many such atomic formulas, say $\psi_1, \ldots, \psi_s$. Define $H_{n,0}$ to be the set of all formulas

$$\psi_1^{\alpha_1} \wedge \ldots \wedge \psi_s^{\alpha_s} ,$$

where $\alpha_i \in \{0,1\}$ and $\psi^0 := \neg\psi, \psi^1 := \psi$. Obviously, $\mathfrak{A}$ together with a tuple $(\underline{a}, \underline{M})$ satisfies either ψ_i or $\neg\psi_i$ (of course, we regard only those formulas with the same shape of variables as $(\underline{a}, \underline{M})$). Thus, $(\mathfrak{A}, \underline{a}, \underline{M})$ satisfies exactly one of the formulas in $H_{n,0}$.

Now let the set $H_{l,r}$ be defined for all $l \in \mathbb{N}$ according to the assertion of part a). In particular, suppose the set $H_{n+1,r}$ to consist of formulas $\theta_1, \ldots, \theta_m$. Define the set $H_{n,r+1}$ as follows:

First, divide $H_{n+1,r}$ into a part $H^1_{n+1,r}$ where the $n+1$-st free variable is first-order and a part $H^2_{n+1,r}$ where the $n+1$-st free variable is second-order. That way one obtains two disjunct index sets I_1, I_2 such that $I_1 \cup I_2 = \{1, \ldots, m\}$. We consider two ways to define formulas in $MS_2(\tau)$ with n variables and quantifier rank $r+1$ from those in $H^1_{n+1,r}$ and in $H^2_{n+1,r}$:

for any nonempty subset $X \subseteq I_1$ build the formula

$$\bigwedge_{i \in X} \exists a_{n+1} \in A \; \theta_i(\underline{a}, \underline{M}, a_{n+1}) \;\wedge\; \bigvee a_{n+1} \in A \; \exists i \in X \; \theta_i(\underline{a}, \underline{M}, a_{n+1}) \quad . \tag{1}$$

Accordingly, for any nonempty subset $\tilde{X} \subseteq I_2$ build the formula

$$\bigwedge_{i \in \tilde{X}} \exists M_{n+1} \subseteq A \; \theta_i(\underline{a}, \underline{M}, M_{n+1}) \;\wedge\; \bigvee M_{n+1} \subseteq A \; \exists i \in \tilde{X} \; \theta_i(\underline{a}, \underline{M}, M_{n+1}) \quad . \tag{2}$$

Now, combine any formula of type (1) with a formula of type (2) by conjunction. This gives for any pair $(X, \tilde{X})$ the formula

$$\left\{ \begin{array}{l} \bigwedge_{i \in X} \exists a_{n+1} \in A \; \theta_i(\underline{a}, \underline{M}, a_{n+1}) \;\wedge\; \bigvee a_{n+1} \in A \; \exists i \in X \; \theta_i(\underline{a}, \underline{M}, a_{n+1}) \; \bigwedge \\ \\ \bigwedge_{i \in \tilde{X}} \exists M_{n+1} \subseteq A \; \theta_i(\underline{a}, \underline{M}, M_{n+1}) \;\wedge\; \bigvee M_{n+1} \subseteq A \; \exists i \in \tilde{X} \; \theta_i(\underline{a}, \underline{M}, M_{n+1}) \end{array} \right. \tag{3}$$

where $X \subseteq I_1$ and $\tilde{X} \subseteq I_2$ and not both of them are empty.

Consider a structure $\mathfrak{A}$ over τ and assignments $a_1, \dots, a_{n_1} \in A, M_1, \dots, M_{n_2} \subseteq A$.

For every $a_{n+1} \in A$ the induction hypothesis yields the existence of a formula $\theta_{i_0} \in H^1_{n+1,r}$ such that $(\mathfrak{A}, \underline{a}, \underline{M}, a_{n+1}) \models \theta_{i_0}(\underline{a}, \underline{M}, a_{n+1})$. Define $X \subseteq I_1$ as the set of all indices $i \in I_1$ such that there is a $a_{n+1} \in A$ with $(\mathfrak{A}, \underline{a}, \underline{M}, a_{n+1}) \models \theta_i(\underline{a}, \underline{M}, a_{n+1})$. Similarly, for every $M_{n+1} \subseteq A$ the induction hypothesis yields the existence of a formula $\theta_{j_0} \in H^2_{n+1,r}$ such that $(\mathfrak{A}, \underline{a}, \underline{M}, M_{n+1}) \models \theta_{j_0}(\underline{a}, \underline{M}, M_{n+1})$. Define $\tilde{X} \subseteq I_2$ as the set of all indices $i \in I_2$ such that there is a $M_{n+1} \subseteq A$ with $(\mathfrak{A}, \underline{a}, \underline{M}, M_{n+1}) \models \theta_i(\underline{a}, \underline{M}, M_{n+1})$. Now it is easy to see that $(\mathfrak{A}, \underline{a}, \underline{M})$ satisfies exactly that formula built according to rule (3) in which the sets X and $\tilde{X}$ are chosen as explained above. This gives claim a).

Ad b) Again via induction over r : For $r = 0$ let ψ be a quantifier free formula in $MS_2(\tau)$ with n variables; ψ is a disjunction of conjunctions of atomic formulas. Let $\varrho_1 \wedge \dots \wedge \varrho_t$ be such a conjunction; consider all elements $\theta \in H_{n,0}$ containing the ϱ_i in exactly the same form. Now, $\varrho_1 \wedge \dots \wedge \varrho_t$ is equivalent to the disjunction of all the θ's chosen above. The formula ψ itself then is equivalent to all the corresponding disjunctions.

For the induction step from r to $r+1$ we assume the set $H^2_{n+1,r}$ to be $\{\theta_1, \dots, \theta_m\}$. Without loss of generality let $\phi \equiv \theta_1 \vee \dots \vee \theta_s$ (i.e. we assume the $n+1$-st variable to be a set-variable). Consider the formula

$$\psi(\underline{x}, \underline{Y}) := \exists\, Y_{n+1}\ \phi(\underline{x}, \underline{Y}, Y_{n+1}) \ .$$

In order to express ψ as a disjunction of formulas in $H_{n,r+1}$, among the formulas in (3) take all those where X is the empty set and $\tilde{X} \subseteq I_2$ contains at least one index in $\{1, \dots, s\}$. This gives a new set of formulas

$$\tilde{\theta}_1, \dots, \tilde{\theta}_{\tilde{s}} \in H_{n,r+1}. \tag{4}$$

<u>Claim:</u> $\psi \equiv \tilde{\theta}_1 \vee \dots \vee \tilde{\theta}_{\tilde{s}}$.

To prove the claim suppose $(\mathfrak{A}, \underline{x}, \underline{Y}) \models \tilde{\theta}_1(\underline{x}, \underline{Y}) \vee \dots \vee \tilde{\theta}_{\tilde{s}}(\underline{x}, \underline{Y})$ holds. Then there exists a formula, say $\tilde{\theta}_1$, such that $(\mathfrak{A}, \underline{x}, \underline{Y}) \models \tilde{\theta}_1(\underline{x}, \underline{Y})$. According to rule (3) we find a set $\tilde{X}$ containing an element $i_0 \in \{1, \dots, s\}$ and

$$(\mathfrak{A}, \underline{x}, \underline{Y}) \models \bigwedge_{i \in \tilde{X}} \exists\, Y_{n+1}\ \theta_i(\underline{x}, \underline{Y}, Y_{n+1}).$$

In particular, we obtain $(\mathfrak{A}, \underline{x}, \underline{Y}) \models \exists\, Y_{n+1}\ \tilde{\theta}_{i_0}(\underline{x}, \underline{Y}, Y_{n+1})$ which implies $(\mathfrak{A}, \underline{x}, \underline{Y}) \models \psi(\underline{x}, \underline{Y})$.

To show the opposite, assume $(\mathfrak{A}, \underline{x}, \underline{Y}) \models \psi(\underline{x}, \underline{Y})$, say

$$(\mathfrak{A}, \underline{x}, \underline{Y}) \models \exists\, M_{n+1}\ \tilde{\theta}_{i_1}(\underline{x}, \underline{Y}, M_{n+1}).$$

We define $\tilde{X}$ as set of all those indices $i \in \{1, \ldots, m\}$ for which $\theta_i \in H^2_{n+1,r}$ as well as $(\mathfrak{A}, \underline{x}, \underline{Y}) \models \exists\, M_{n+1}\ \tilde{\theta}_i(\underline{x}, \underline{Y}, M_{n+1})$. According to the definition we get $i_1 \in \tilde{X}$; thus, the formula obtained by choosing our particular $\tilde{X}$ together with $X = \emptyset$ in (3) belongs to the formulas under (4), which immediately gives the reverse direction.

If ψ has the shape $\forall M_{n+1}\ \phi(\underline{x}, \underline{Y}, M_{n+1})$, for building the corresponding set of formulas in (4) consider all choices $\tilde{X} \subseteq \{1, \ldots, s\}$.

If ψ is obtained by quantification of a first-order variable x_{n+1} the proof works similarly after replacing $\tilde{X}$ by X and the formulas in $H^2_{n+1,r}$ by those in $H^1_{n+1,r}$. □

Remark 7. In the above proof there are several ways to obtain a conjunction of atomic formulas which is false over any structure with an equality relation, for example by including $x \neq x$ for one or several variables. By convention, we include only one such formula into each of the sets $H_{n,0}$. This will be needed later on in order to avoid ambiguities.

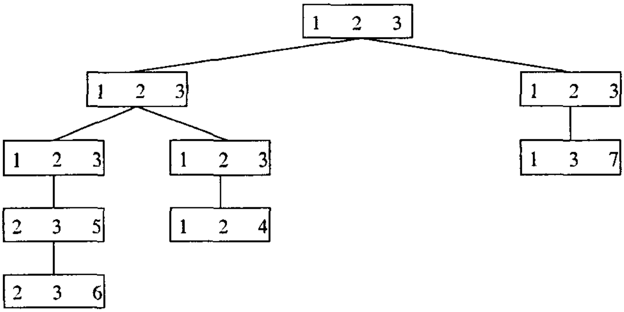
8.3 Parse-trees for hypergraphs of bounded tree-width

Let G be a hypergraph of bounded tree-width k. Our goal is to create G from some elementary hypergraphs by applying finitely many operations to the latter. The initial hypergraphs will be structures of size at most k. The operations have to be chosen carefully in order to guarantee validity of the theorem by Feferman-Vaught.

The hypergraphs in-between the building procedure of G all are $(k+1)$-boundaried hypergraphs in the in the sense defined in section 5. The following example shows how the decomposition of a hypergraph of tree-width k into $(k = 1)$ boundaried hypergraphs works.

Example 5. We consider once again the polynomial p from Example 3, ii). Starting with a tree decomposition of p the latter is first turned into a binary tree. To this aim branch nodes are simply duplicated sufficiently many times. By adding further nodes to those subsets V_t with cardinality less than $k+1$ (without destroying the tree decomposition requirements) we can assume all sets V_t to be of cardinality $k+1$. Furthermore, by duplication once again branch nodes every branch has identical children. Finally, including additional nodes we can guarantee that two neighbored sets V_t differ by at most one element.

For the polynomial p from Example 3,ii) this results in the following normalized tree decomposition:



The parse tree is now constructed bottom up. At the leaves we apply the *create* operations for the $k + 1$ vertices included in the corresponding sets V_t. To include new hyperedges the *add* operation is applied. To change an already treated vertex with a new one we use the operation *new* ; *change* is used to fit labels. Finally, due to the normalization of the tree decomposition *join* is only applied to substructures with the same labels on the same vertices (the boundary).

Full details can be found in [19].

A tree-decomposition itself can also be obtained in linear time according to theorem 6.

8.4 Feferman-Vaught theorem for finite structures of bounded tree-width

We combine the previous subsections in order to obtain the main result for the structures we are interested in.

The overall idea is to decide a $MS_2(\tau)$ formula for a given structure by reducing it to substructures and deciding corresponding formulas there. The substructures are obtained by analyzing the parse-tree of the given structure. The formulas to be decided on the substructures are determined through the given one according to theorem 20.

We will explicitly proof the according statement for a decomposition related to a join operation $\oplus$. The according statements for the other parse operations can be established similarly.

For the following recall our general proviso!

Let $\mathfrak{A}$ and $\mathfrak{B}$ be $k+1$ boundaried hypergraphs over τ with universes A and B resp., and boundaries $\delta(A) \subset A, \delta(B) \subset B, \widehat{A} : +A \setminus \delta(A), \widehat{B} := B \setminus \delta(B)$. Let $\mathfrak{C} = \mathfrak{A} \oplus \mathfrak{B}$ be the join of $\mathfrak{A}$ and $\mathfrak{B}$ as explained in subsection 8.4.

Let z be an assignment of the variables in a $MS_2(\tau)$ formula $\Phi(\underline{x}, \underline{y}, \underline{w}, \underline{X})$ into the universe C of $\mathfrak{C}$ such that the following holds:

- z maps the variables x_i of block $\underline{x}$ to the set $\widehat{A}$
- z maps the variables y_i of block $\underline{y}$ to the set $\widehat{B}$
- z maps the variables w_i of block $\underline{w}$ to the boundary $\delta(A) = \delta(B)$ (recall that the two boundaries were identified). In particular, $z(w_i)$ is member of both the universes A and B.

Denote by z_A, z_B the assignments with $z_A(X) = X \cap A, z_A(s) = z(s)$ for $s \in A$ and $z_B(X) = X \cap B, z_B(s) = z(s)$ for $s \in B$.

Then the following is true

Theorem 21. *Let $\mathfrak{C}$ be the join of $\mathfrak{A}$ and $\mathfrak{B}$: $\mathfrak{C} = \mathfrak{A} \oplus \mathfrak{B}$ and let $h(\underline{x}, \underline{y}, \underline{w}, \underline{X})$ be a $MS_2(\tau)$ formula which is a Hintikka formula in some $H_{n,r}(\underline{x}, \underline{y}, \underline{w}, \underline{X})$. Then there are unique Hintikka formulas $h_1(\underline{x}, \underline{w}, \underline{X}) \in H_{n,r}(\underline{x}, \underline{w}, \underline{X})$ and $h_2(\underline{y}, \underline{w}, \underline{X}) \in H_{n,r}(\underline{y}, \underline{w}, \underline{X})$ such that for every assignment z as above we have $(\mathfrak{C}, z) \models h(\underline{x}, \underline{y}, \underline{w}, \underline{X}) \;\Leftrightarrow\; (\mathfrak{A}, z_A) \models h_1(\underline{x}, \underline{w}, \underline{X}) \;\wedge\; (\mathfrak{B}, z_B) \models h_2(\underline{y}, \underline{w}, \underline{X})$.*

Proof. Induction on the quantifier rank r of h:

$\underline{r = 0:}$

According to the proof of theorem 20 the quantifier free Hintikka formula h is a conjunction of all atomic formulas (where the latter appear either negated or not). We indicate with some examples how h_1 and h_2 are being built given these parts in h:

i) suppose an atomic formula $x_i = y_j$ appears unnegated in h. Then $(\mathfrak{C}, z) \not\models h$ because z assigns different values to x_i and y_j. This holds independently of $\mathfrak{C}$. We are therefore done by including in h_1 (or in h_2) an atomic formula $x \neq x$ which is always false as well (according to the uniqueness condition mentioned in remark 7).

ii) Suppose an atomic formula $w_i \neq w_j$ appears in h. Then we include it both in h_1 and in h_2. A formula like $x_i \neq x_j$ is only included in h_1, similarly for variables from the block $\underline{y}$ and h_2.

iii) For a unary relation R in τ a formula $R(x_i)$ is maintained in h_1, whereas $R(y_j)$ is transferred to h_2. An atomic formula $R(w_i)$ will again be both present in h_1 and in h_2.

Thus, given as unary relations which only hold true on exactly one element of the boundary, the labels are transformed to both substructures.

Note that by convention in the general case where we deal with two-sorted structures we decompose the set E in such a way that multi-edges from $\mathfrak{P}(\delta(A))$ are only maintained in $\mathfrak{A}$, not in $\mathfrak{B}$.

$\underline{r-1 \to r:}$ Let

$\psi(\underline{x}, \underline{y}, \underline{w}, \underline{M}) :=$

$$\bigwedge_{i \in X} \exists a_{n+1} \in A\ \theta_i(\underline{x}, \underline{y}, \underline{w}, \underline{M}, a_{n+1}) \ \wedge\ \forall a_{n+1} \in A\ \exists i \in X\ \theta_i(\underline{x}, \underline{y}, \underline{w}, \underline{M}, a_{n+1}) \ \bigwedge$$

$$\bigwedge_{i \in \tilde{X}} \exists M_{n+1} \subseteq A\ \theta_i(\underline{x}, \underline{y}, \underline{w}, \underline{M}, M_{n+1}) \ \wedge\ \forall\ M_{n+1} \subseteq A\ \exists i \in \tilde{X}\ \theta_i(\underline{x}, \underline{y}, \underline{w}, \underline{M}, M_{n+1})$$

where $\theta_i \in H^1_{n+1,r-1}(\underline{x}, \underline{y}, \underline{w}, \underline{M}, a_{n+1})$ for $i \in X$ and $\theta_i \in H^2_{n+1,r-1}(\underline{x}, \underline{y}, \underline{w}, \underline{M}, M_{n+1})$ for $i \in \tilde{X}$.

Suppose $(\mathfrak{A} \oplus \mathfrak{B}, z) \models \psi$. Then

- for every $i \in X$ there exists a $c_i^{\mathfrak{A}\oplus\mathfrak{B}} \in A \oplus B$ such that $(\mathfrak{A} \oplus \mathfrak{B}, z) \models \theta_i(\underline{x}, \underline{y}, \underline{w}, \underline{M}, c_i^{\mathfrak{A}\oplus\mathfrak{B}})$

  and

- for every $i \in \tilde{X}$ there exists a $M_i^{\mathfrak{A}\oplus\mathfrak{B}} \subseteq A \oplus B$ such that $(\mathfrak{A} \oplus \mathfrak{B}, z) \models \theta_i(\underline{x}, \underline{y}, \underline{w}, \underline{M}, M_i^{\mathfrak{A}\oplus\mathfrak{B}})$.

The induction hypothesis yields the existence of four Hintikka formulas $\theta^{\mathfrak{A}}_{i,1}, \theta^{\mathfrak{A}}_{i,2} \in H_{n_1,r-1}$ and $\theta^{\mathfrak{B}}_{i,1}, \theta^{\mathfrak{B}}_{i,2} \in H_{n_2,r-1}$ such that

$$(*) \quad \begin{cases} (\mathfrak{A}, z_A) \models \theta^{\mathfrak{A}}_{i,1}(\underline{x}, \underline{w}, \underline{M}, c_i^{\mathfrak{A}}) \ \wedge\ \theta^{\mathfrak{A}}_{i,2}(\underline{x}, \underline{w}, \underline{M}, M_i^{\mathfrak{A}}) \\ \\ (\mathfrak{B}, z_B) \models \theta^{\mathfrak{B}}_{i,1}(\underline{x}, \underline{w}, \underline{M}, c_i^{\mathfrak{B}}) \ \wedge\ \theta^{\mathfrak{B}}_{i,2}(\underline{x}, \underline{w}, \underline{M}, M_i^{\mathfrak{B}}) \end{cases}$$

(where either $c_i^{\mathfrak{A}}$ or $c_i^{\mathfrak{B}}$ might be a "dummy" constant according to the membership of $c_i^{\mathfrak{A}\oplus\mathfrak{B}}$ to $\widehat{A}, \widehat{B}$ or $\delta(A) = \delta(B)$.)

We define h_1 : let X_A be the set of all indices of Hintikka formulas in $H_{n_1,r-1}$ such that this index appears among the formulas $\theta^{\mathfrak{A}}_{i,1}$ in $(*)$. Let $\tilde{X}_A$ denote the corresponding set of indices appearing among one of the formulas $\theta^{\mathfrak{A}}_{i,2}$ in $(*)$. Then h_1 is defined to be built according to the general scheme (cf. equation 3 in the first subsection of this paragraph) and with X_A and $\tilde{X}_A$

as the chosen index sets. Formula h_2 is given in the same manner. It's now straightforward to check that h_1 and h_2 satisfy the requirements. □

The theorem together with finiteness of $H_{n,r}$ implies that there are only finitely many triples $\mathfrak{C} = \mathfrak{A} \oplus \mathfrak{B}$ with respect to the uniquely determined Hintikka formulas. The same is true if we allow parameters.

The Feferman-Vaught theorem generalizes the previous theorem by getting independent of the particular structures involved. It can be proved now as follows:

Proof of theorem 17:

According to theorem 20 ψ is equivalent to a disjunction $\theta_1 \vee \ldots \vee \theta_s$ of finitely many Hintikka formulas. For each θ_i there exist finitely many triples $(\theta_i, h_1^j, h_2^j), j \in J_i, |J_i|$ finite such that for given structures $\mathfrak{C} = \mathfrak{A} \oplus \mathfrak{B}$ theorem 21 is true for exactly one $j \in J_i$. If $\mathfrak{C}, \mathfrak{A}, \mathfrak{B}$ are varied then $\mathfrak{C} \models \theta_i$ iff $\bigvee_{j \in J_i} \mathfrak{A} \models h_1^j \ \wedge \ \mathfrak{B} \models h_2^j$. The formula ψ itself is equivalent to the disjunction $\bigvee_{i=1}^{s} \bigvee_{j \in J_i} \mathfrak{A} \models h_1^j \ \wedge \ \mathfrak{B} \models h_2^j$. This gives the formulas $h_{1,\alpha}$ and $h_{2,\alpha}$. □

8.5 The algorithm

The Feferman-Vaught theorem holds also true in a corresponding way for the other parsing operations introduced above. It is true as well for other classes of parameters like the clique-width of a graph and its corresponding graph operations, see [15].

It is then used for computational purposes as follows. Given a $MS_2(\tau)$ formula $\psi \in F_{n,r}$ we precompute a table of all Hintikka formulas in $H_{n,r}$ and how three of them fit together with respect to the Feferman-Vaught theorem for all the parsing operations. The formula ψ is written in its equivalent form as disjunction of Hintikka formulas, say $\psi = \theta_1 \vee \ldots \vee \theta_s$.

Now, given a structure of bounded tree-width we first construct its tree decomposition and the corresponding parse tree according to theorems 6 and 16. From the precomputed table of Hintikka formulas we obtain a corresponding decomposition of the θ_i 's. Then we evaluate bottom-up from the leaves of the parse tree to its root the Hintikka formulas in the decomposition (using again the precomputed table) and check whether ψ is true on the input. Since the parse tree has linear size in the size of the structure the running time is linear (though with large constants).

Some technicalities have to be taken care of. We demonstrate them again for the join operator $\oplus$.

Theorem 17 depends on the shape $(\underline{x}, \underline{y})$ an assignment z induces on the free first-order variables. Therefore, if the above algorithm is performed bottom up, we have to take into account all possible assignments for z, that is all possible patterns z induces on the free first-order variables. Suppose there are s many of them (where s only depends on the given formula). If we consider all decomposition patterns of these s many variables along the parse tree of a given structure of bounded tree-width we see that their number only depends on a function $f(s)$ in s. For every fixed pattern we perform the above algorithm in linear time, giving an $O(n)$ algorithm in total.

The interested reader might try to perform a proof of Feferman-Vaught theorem and the subsequent algorithm for the operator new_i!

The main contribution of the present paper is to combine the theory over finite structures presented in the Appendix with algebraic issues. To this aim a meta-finite monadic second order logic is defined carefully. It is constructed in such a way that on the one hand side a lot of important problems can be expressed by it. On the other hand side it is not too general in the sense that we can decompose a given meta-finite structure together with a $\exists$-MSO$_{\mathbb{R}}$ property in parallel to the decomposition of the underlying finite structure described in this Appendix.

Acknowledgments

The first author would like to thank D. Grigoriev, I. Shparlinski, M. Vardi and V. Weisspfenning for their interest and comments. The first author was partially supported by a Grant of the Fund for Promotion of Research of the Technion–Israeli Institute of Technology and the second author was supported in part by the IST Programme of the EU under contract number IST-1999-14186 ALCOM-FT. Both authors were partially supported by EU ESPRIT project NeuroCOLT 2.

References

1. S. Arnborg, J. Lagergren, and D. Seese, Easy problems for tree decomposable graphs, Journal of Algorithms **12**, 308–340 (1991).
2. A.I. Barvinok, Two algorithmic results for the traveling salesman problem, Mathematics of Operations Research **21**, 65–84 (1996).
3. S. Basu, New results on Quantifier Elimination Over Real Closed Fields and Applications to Constraint Databases, Journal of the ACM **46, No.** 4, 537–555 (1999).

4. P. Bürgisser, M. Clausen, and M.A. Shokrollahi, *Algebraic Complexity Theory*, (Volume 315 of *Grundlehren*, Springer Verlag, 1997).
5. L. Blum, F. Cucker, M. Shub, and S. Smale, *Complexity and Real Computation*, (Springer Verlag, 1998).
6. B. Bank, M. Giusti, J. Heintz, and G.M. Mbakop, Polar varieties, real equation solving, and data structures: The hypersurface case, Journal of Complexity **13**, 5–27 (1997).
7. H. Bodlaender, Treewidth: Algorithmic techniques and results, in *Proc. of the 22nd Symposium Mathematical Foundation of Computer Science*, I. Privara, P. Ruzicka, eds. (Volume 1295 of *Lecture Notes in Computer Science*, 19–36, Springer, 1997).
8. H. Bodlaender, A partial k-arboretum of graphs with bounded tree-width (tutorial), Theoretical Computer Science **208**, 1–45 (1998).
9. S. Basu, R. Pollack, and M.-F. Roy, On the combinatorial and algebraic complexity of quantifier elimination, Journal of the ACM **43(6)**, 1002–1045 (1996).
10. L. Blum, M. Shub, and S. Smale, On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines, Bull. Amer. Math. Soc. **21**, 1–46 (1989).
11. P. Bürgisser, *Completeness and Reduction in Algebraic Complexity Theory*, (Algorithms and Computation in Mathematics, vol. 7 Springer Verlag, 2000).
12. D. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, (Graduate Texts in Mathematics. Springer Verlag, 1997).
13. B. Courcelle and M. Mosbah, Monadic second–order evaluations on tree–decomposable graphs, Theoretical Computer Science **109**, 49–82 (1993).
14. B. Courcelle, J.A. Makowsky, and U. Rotics, Linear Time Solvable Optimization Problems on Graphs of Bounded Clique Width, Theory of Computing Systems **33, No.2**, 125–150 (2000).
15. B. Courcelle, J.A. Makowsky, and U. Rotics. On the fixed parameter complexity of graph enumeration problems definable in monadic second order logic, *Discrete and Applied Mathematics* **xx**, xx–yy (2000).
16. B. Courcelle, Graph rewriting: An algebraic approach, in *Handbook of Theoretical Computer Science*, J. van Leeuwen, ed. (Volume 2, Chapter 5. Elsevier Science Publishers, 1990).
17. Bruno Courcelle, The expression of graph properties and graph transformations in monadic second-order logic, in *Handbook of graph grammars and computing by graph transformations, Vol. 1: Foundations*, G. Rosenberg, ed. (pp. 313–400, World Scientific, 1997).
18. F. Cucker and M. Matamala, On digital nondeterminism, Mathematical

Systems Theory **29**, 635–647 (1996).

19. R.G. Downey and M.F Fellows, *Parametrized Complexity*, (Springer, 1999).
20. R. Diestel, *Graph Theory*, (Graduate Texts in Mathematics. Springer, 1996).
21. T. Feder and M. Vardi, The computational structure of monotone monadic SNP and constraint satisfaction, in *STOC'93*, (pp. 612–622, ACM, 1993).
22. T. Feder and M. Vardi, The computational structure of monotone monadic SNP and constraint satisfaction: A study through Datalog and group theory, SIAM Journal on Computing **28**, 57–104 (1999).
23. H.D. Ebbinhaus and J. Flum, *Finite Model Theory* (Perspectives in Mathematical Logic, Springer, 1995).
24. E. Grädel and Y. Gurevich, Metafinite model theory, Information and Computation **140**, 26–81 (1998). See also: Logic and Computational Complexity, D. Leivant ed. (Selected Papers, Springer, 1995, 313-366).
25. D.Y. Grigoriev and M. Karpinski, The matching problem for bipartite graphs with polynomial bounded permanents is in NC, in *28th Annual Symposium on Foundations of Computer Science*, pages 166–172, 1987.
26. I. Gelfand, M. Kapranov, and A. Zelevinsky, *Discriminants, Resultants and Multidimensional Determinants*, (Birkhäuser Verlag, 1994).
27. E. Grädel and K. Meer, Descriptive complexity theory over the real numbers, Lectures in Applied Mathematics **32**, 381–403 (1996). A preliminary version has been presented at the 27th ACM-Symposium on Theory of Computing, Las Vegas, 1995.
28. W. Hodges, Model theory, *Encyclopedia of Mathematics and its Applications*, (Vol. 42, Cambridge University Press, 1993).
29. M. Jerrum, Two–dimensional monomer–dimer systems are computationally intractable, Journal Stat. Phys. **48**, 121–134 (1987). Erratum in vol. 59, 1087–1088, 1990.
30. A.G. Khovanskii Fewnomials, *Translations of Mathematical Monographs*, (Vol. 88. American Mathematical Society, 1991).
31. P. Kolaitis and M. Vardi, Conjunctive query containement and constraint satisfaction, in *PODS'98*, (425–435, ACM, 1998).
32. Y.N. Lakshman and D. Lazard, On the complexity of zero-dimensional algebraic systems, in *Effective methods in Algebraic Geometry*, T. Mora and C. Traverso, eds. (Volume 94 of *Progress in Mathematics*, pp. 217–225, Birkhäuser, Basel, 1991).
33. J.A. Makowsky, Logical methods in graph algorithms, Lecture Notes of a course given at ESSLLI'99 in Utrecht, August, 1999.

34. J.A. Makowsky, Colored Tutte Polynomials and Kauffman Brackets for Graphs of Bounded Tree Width, Extended Abstract , submitted MFCS'00; revised further, March 31, 2000, submitted to Combinatorics, Probability and Computation.
35. J.A. Makowsky and K. Meer, On the Complexity of Combinatorial and Metafinite Generating Functions of Graph Properties in the Computational Model of Blum, Shub and Smale. Extended abstract, to appear in: *Proc. CSL 2000*, (LNCS, Springer).
36. K. Meer, On the complexity of quadratic programming in real number models of computation, Theoretical Computer Science **133**, 85–94 (1994).
37. K. Meer and C. Michaux, A survey on real structural complexity theory, Bulletin of the Belgian Math. Soc. **4**, 113–148 (1997).
38. B. Poizat, *Les Petits Cailloux: Une approche modèle-théorique de l'algorithmie*, Aléas, Paris, 1995.
39. P. Pedersen, M.-F. Roy, and A. Szpirglas, Counting real zeros in the multivariate case, in *Computational Algebraic Geometry*, F. Eysette and A. Galligo, eds. (Volume 109 of *Progress in Mathematics*, pp. 203–224, Birkhäuser, Basel, 1993).
40. M.-F. Roy, Basic algorithms in real algebraic geometry and their complexity: from Sturm's theorem to the existential theory of reals, in *Lectures in Real Geometry*, F. Broglia, ed. (Walter de Gruyter, 1996).
41. W. Thomas, Automata on infinite objects, in *Handbook of Theoretical Computer Science*, J. van Leeuwen, ed. (Volume 2, Chapter 4, Elsevier Science Publishers, 1990).
42. L.G. Valiant, The complexity of computing the permanent, *Theoretical Computer Science* **8**, 189–201 (1979).

POLYNOMIAL SYSTEMS AND THE MOMENTUM MAP

GREGORIO MALAJOVICH*
Departamento de Matemática Aplicada, Universidade Federal do Rio de Janeiro,
Caixa Postal 68530, CEP 21945-970, Rio de Janeiro, RJ, Brasil
http://www.labma.ufrj.br/~gregorio
e-mail: gregorio@labma.ufrj.br
On leave at the Department of Mathematics, City University of Hong Kong.

J. MAURICE ROJAS†
Department of Mathematics, City University of Hong Kong
and
Department of Mathematics
Texas A&M University
College Station, Texas 77843-3368, USA
http://math.tamu.edu/~rojas
e-mail: rojas@math.tamu.edu

Keywords: mixed volume, condition number, polynomial systems, sparse, random.
2000 Math Subject Classification: 65H10, 52A39.

1 Introduction

This paper outlines a Kähler-geometric approach to the study of random sparse polynomial systems, with a view toward understanding the distribution of solutions and their numerical conditioning. Section 2 develops the necessary mathematical framework while section 3 contains our main results for random sparse polynomial systems, including:

1. a new formula, in terms of a particular differential form, for the average number of complex roots in any measurable region

2. new bounds relating numerical conditioning of any given sparse polynomial system to the nearest ill-posed problem.

*G.M.'S VISIT TO CITY UNIVERSITY OF HONG KONG WAS SUPPORTED BY CERG GRANTS 9040393, 9040402, AND 9040188.

†PARTIALLY SUPPORTED BY HONG KONG UGC GRANT #9040402 AND A US NATIONAL SCIENCE FOUNDATION MATHEMATICAL SCIENCES POSTDOCTORAL FELLOWSHIP.

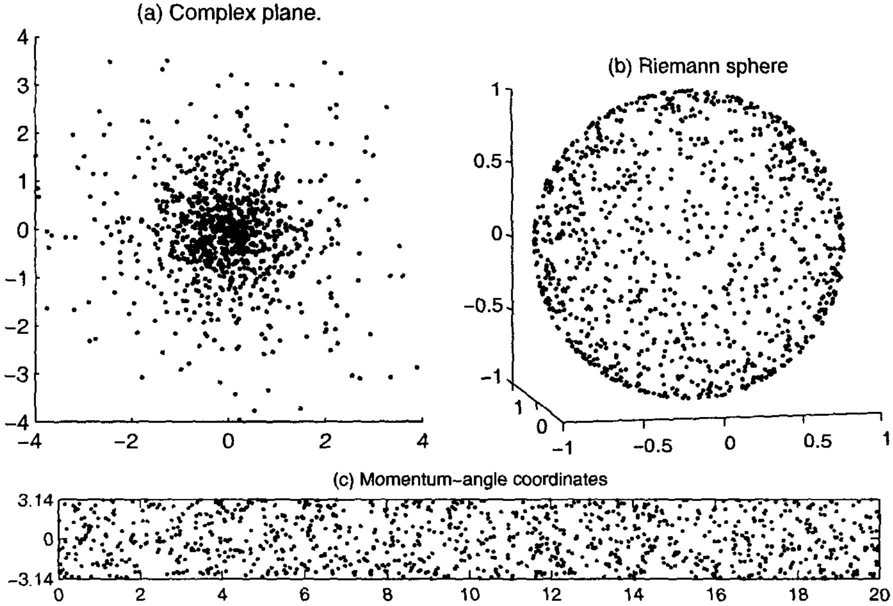


Figure 1. Root distribution when coefficient vector has a unitarily invariant Gaussian distribution

However, let us first give a simple preliminary illustration of some of the ideas we will develop.

Figures 1(a) through 2(c) feature one random root of each of 1000 random degree 20 univariate polynomials. The j-th coefficient of each polynomial is a pseudo-random Gaussian variable of variance $\binom{d}{j}$ in Figure 1(a–c), and of variance 1 in Figure 2(a–c). The first distribution is invariant under a naturally defined action of the 2 × 2 unitary matrix group on the complex projective line $\mathbb{P}^1$: the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ sends the point with projective coordinates $[x_0 : x_1]$ to $[ax_0 + bx_1 : cx_0 + dx_1]$. The second distribution lacks this "unitary" invariance.

The roots are plotted in three different "phase spaces". In Figures 1(a) and 2(a), the roots are plotted in the complex plane. We notice that in figure 2(a), the roots seem to accumulate on the unit circle. This is a well-known phenomenon which can be made more rigorous.

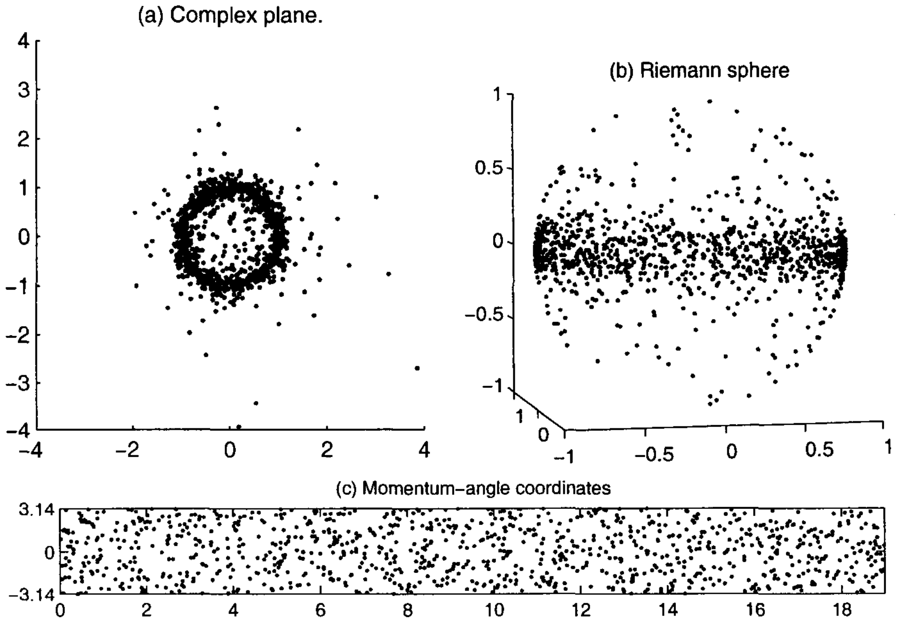


Figure 2. Root distribution when the coefficients are identically and independently distributed Gaussians

In Figures 1(b) and 2(b), we show the same roots as points in the Riemann Sphere. We can see that the selected roots of the unitarily invariant distributed polynomials are uniformly distributed on the Riemann Sphere, while those of the variance-one polynomials are not (this shows graphically that the distribution of the roots is not unitarily invariant).

There is strong evidence in [9] and [3] suggesting that unitarily invariant distribution would be the natural distribution for Gaussian random polynomials. While the unitarily invariant paradigm was fundamental to the development in [17,18,19,20,21], we will see here that a more general approach is viable and fruitful.

In figures 1(c) and 2(c), the same roots are plotted in a different "phase space", that we will construct below. The distribution seems uniform, it is indeed uniform. This suggests that there is another approach to random polynomials, that allows arbitrary variances through different coordinate systems.

This approach extends to systems of polynomial equations. While the unitary-invariant paradigma precluded the treatment of sparse polynomial systems, the approach we suggest extends naturally to systems of equations.

For instance, it is possible to bound the probability that the condition number of a random sparse polynomial system (arbitrary variances) is large.

The roots of a sparse polynomial system are known to belong to a certain toric variety. However, in order to obtain the theorems below, we needed to endow the toric variety with a certain geometrical structure, as explained below. The main insight comes from mechanics, and from symplectic and Kähler geometry. The main tool is the Momentum map.

The proofs of the results mentioned herein may be found in [10] and its references.

2 General setting

Let A be an $M \times n$ matrix, with non-negative integer entries. To the matrix A we associate the convex polytope $\mathrm{Conv}(A)$ given by the convex hull of all the rows, $\{A^\alpha\}_{\alpha \in \{1,\dots,M\}}$, of A:

$$\mathrm{Conv}(A) \overset{\mathrm{def}}{=} \left\{ \sum_{\alpha=1}^{M} t_\alpha A^\alpha \ : \ 0 \leq t_\alpha \leq 1, \ \sum_{\alpha=1}^{M} t_\alpha = 1 \right\} \subset (\mathbb{R}^n)^\vee \quad .$$

Here, we use the notation $X^\vee$ to denote the dual of a vector space X.

Assume that $\dim(\mathrm{Conv}(A)) = n$. Then we can associate to the matrix A the space $\mathcal{F}_A$ of polynomials with support contained in $\{A^\alpha : 1 \leq \alpha \leq M\}$. This is a linear space, and there are many reasonable choices of an inner product in $\mathcal{F}_A$.

Let C be a diagonal positive definite $M \times M$ matrix. Its inverse C^{-1} is also a diagonal positive definite $M \times M$ matrix. This inverse matrix defines the inner product:

$$\langle z^{A^\alpha}, z^{A^\beta} \rangle_{C^{-1}} = (C^{-1})_{\alpha,\beta} \quad .$$

The matrix C will be called the *variance matrix.* This terminology arises when we consider random normal polynomials in $\mathcal{F}_A$ with variance $C_{\alpha\alpha}$ for the α-th coefficient. We will refer to these randomly generated functions as *random normal polynomials*, for short.

We may also produce several objects associated to the matrix A (and to the variance matrix C). The most important one for this paper will be a Kähler manifold $(\mathcal{T}^n, \omega_A, J)$. This manifold is a natural "phase space" for the roots of polynomial systems with support in A. It is *the* natural phase space for the roots of systems of random normal polynomials in $(\mathcal{F}_A, \langle\cdot,\cdot\rangle_{C^{-1}})$.

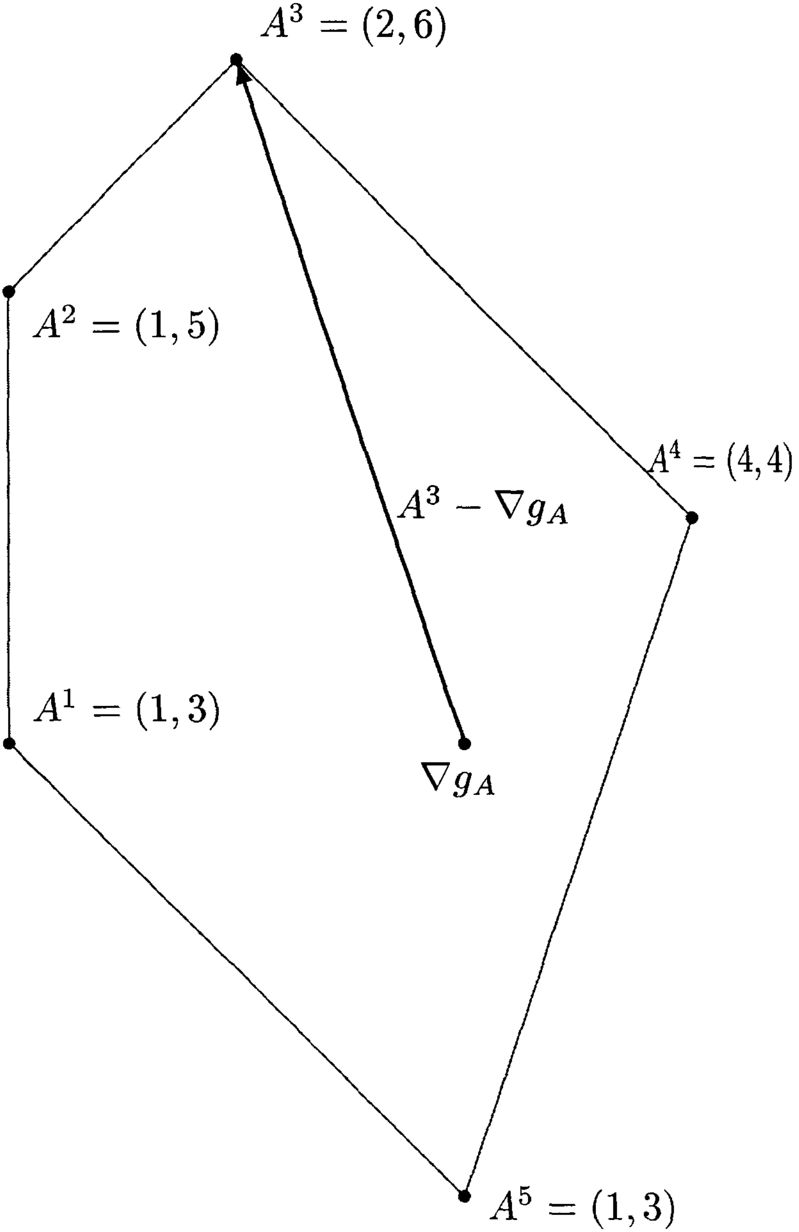


Figure 3. Geometric interpretation of the Momentum map ∇g_A. The momentum is related to the derivative of the Veronese embedding v_A by the formula: $Dv_A^\lambda = v_A^\lambda(A^\lambda - \nabla g_A)$

More explicitly, let $\mathcal{T}^n \overset{\text{def}}{=} \mathbb{C}^n \pmod{2\pi\sqrt{-1}\,\mathbb{Z}^n}$ (which, as a real manifold, happens to be an n-fold product of cylinders). Let $\exp : \mathcal{T}^n \to (\mathbb{C}^*)^n$ denote coordinatewise exponentiation. Then we will look at the preimages of the roots of a polynomial system by exp. We leave out roots that have one coordinate equal to zero and roots at infinity.

The Kähler manifold may be constructed as follows. The Complex Structure J is defined locally by

$$J_{(p,q)}(\dot{p}, \dot{q}) \overset{\text{def}}{=} (-\dot{q}, \dot{p})$$

The form ω_A is the pull-back of the Fubini-Study 2–form in $\mathbb{C}^M$ by the embedding

$$\begin{aligned} \hat{v}_A : \mathcal{T}^n \quad &\to \mathbb{C}^M \\ (p,q) &\mapsto C^{1/2} \cdot \exp(A(p + q\sqrt{-1})) \end{aligned}$$

If $v_A : \mathcal{T}^n \to \mathbb{P}^{M-1}$ is the composition of $\hat{v}_A$ with the canonical map $\mathbb{C}^M_* \to \mathbb{P}^{M-1}$, then ω_A is also the pull-back by v_A of the canonical symplectic form of $\mathbb{P}^{M-1}$.

Another important invariant is the real function:

$$\begin{aligned} g_A : \mathcal{T}^n &\to \mathbb{R} \\ p, q &\mapsto \log \|\hat{v}_A(p)\| \end{aligned}$$

This function is independent of p and we may treat g_A as a function of p alone. In that setting, the gradient $\nabla g_A : \mathbb{R}^n \to (\mathbb{R}^n)^\vee$ maps $\mathbb{R}^n$ diffeomorphically onto the interior of the convex hull of the rows of A (See Figure 3 and [10]). The mapping

$$\nabla g_A : \mathcal{T}^n \to (\mathbb{R}^n)^\vee$$

is called the *Momentum map.* It was introduced in its modern formulation by Smale [22] and Souriau [23]. The reader may consult one of the many textbooks in the subject (such as Abraham and Marsden [1] or McDuff and Salamon [13]) for a general exposition.

We can say now what the "phase space" of Figures 1(c) and 2(c) was. It was the image of the toric manifold $(\mathcal{T}^n, \omega_A, J)$ by the "volume-preserving" (up to a constant) map:

$$(p,q) \mapsto ((\nabla g_A)_{|p}, q)$$

The potential g_A and the Kähler form ω_A are related as follows:

$$\omega_A = -\frac{1}{2} dJ^* dg_A$$

where d stands for exterior differentiation. The Hermitian metric associated to A has also the expressions:

$$\langle a,b\rangle \overset{\text{def}}{=} \omega_A(a,Jb) = a^H Dv_A^H Dv_A b = \frac{1}{2}D^2 g_A(a,b)$$

3 Mixed and Unmixed systems

Systems where all the polynomials have the same support are called *unmixed.* The general situation (*mixed* polynomial systems), where the polynomials may have different supports, is of greater practical interest. It is also a much more challenging situation. We shall consider systems of n polynomials in n variables, each polynomial in some inner product space of the form $(\mathcal{F}_{A_i}, \langle\cdot,\cdot\rangle_{C_i^{-1}})$ (where $i = 1,\cdots,n$ and each A_i and each C_i are as above).

In this realm, a mathematical object (that we may call a *mixed manifold*) seems to arise naturally. A mixed manifold is an $(n+2)$–tuple $(\mathcal{T}^n, \omega_{A_1},\cdots,\omega_{A_n}, J)$ where for each i, $(\mathcal{T}^n, \omega_{A_i}, J)$ is a Kähler manifold. Mixed manifolds do **not** have a natural canonical Hermitian structure. They have n equally important Hermitian structures. However, they have one natural volume element, the *mixed volume form,* given by

$$d\mathcal{T}^n = \frac{(-1)^{n(n-1)/2}}{n!}\ \omega_{A_1}\wedge\cdots\wedge\omega_{A_n} \quad .$$

As explained in [7], the volume of $\mathcal{T}^n$ relative to the mixed volume form is (up to a constant) the mixed volume of the n–tuple of polytopes $(\mathrm{Conv}(A_1),\cdots,\mathrm{Conv}(A_n))$.

We extend the famous result by Bernshtein [4] on the number of roots of mixed systems of polynomials as follows:

Theorem 1. *Let $A_1,\cdots,A_n$ and $C_1,\cdots,C_n$ be as above. For each $i = 1,\cdots,n$, let f_i be an (independently distributed) normal random polynomial in $(\mathcal{F}_{A_i}, \langle\cdot,\cdot\rangle_{C_i^{-1}})$. Let U be a measurable region of $\mathcal{T}^n$. Then, the expected number of roots of the polynomial system $f(z) = 0$ in* $\exp U \subseteq (\mathbb{C}^*)^n$ *is*

$$\frac{n!}{\pi^n}\int_U d\mathcal{T}^n \quad .$$

Example 1. When each f_i is dense with a variance matrix C_i of the form:

$$C_i = \mathrm{Diag}\left(\frac{\deg f_i!}{I_1!I_2!\cdots,I_n!(\deg f_i - \sum_{j=1}^n I_j)!}\right) \quad ,$$

the volume element $d\mathcal{T}^n$ becomes the Bézout number $\prod \deg f_i$ times the pull-back to $\mathcal{T}^n$ of the Fubini-Study metric. We thus recover Shub and Smale's stochastic real version of Bézout's Theorem [18]. ■

The general unmixed case ($A_1 = \cdots = A_n$, $C_1 = \cdots = C_n$) is a particular case of Theorem 8.1 in [6]. This is the only overlap, since neither theorem generalizes the other.

On the other hand, when one sets $U = \mathcal{T}^n$, one recovers Bernshtein's first theorem. The quantity $\pi^{-n} \int_{\mathcal{T}^n} d\mathcal{T}^n$ is precisely the *mixed volume* of polytopes $A_1, \cdots, A_n$ (see [16] for the classical definition of Mixed Volume and main properties).

A version of Theorem 1 was known to Kazarnovskii [8] and Khovanskii. In [8], the supports A_i are allowed to have complex exponents. However, uniform variance ($C_i = I$) is assumed. His method may imply this special case of Theorem 1, but the indications given in [8] were insufficient for us to reconstruct a proof.

The idea of working with roots of polynomial systems in logarithmic coordinates seems to be extremely classical, yet it gives rise to interesting and surprising connections (see the discussions in [11,12,24]).

4 The Condition Number

Let $\mathcal{F} = \mathcal{F}_{A_1} \times \cdots \times \mathcal{F}_{A_n}$, and let $f \in \mathcal{F}$. A *root* of f will be represented by some $p + q\sqrt{-1} \in \mathcal{T}^n$. (Properly speaking, the root of f is $\exp(p + q\sqrt{-1})$).

In this discussion, we assume that the "root" $p+q\sqrt{-1}$ is non-degenerate. This means that the derivative of the *evaluation map*

$$\begin{array}{rrcl} ev: & \mathcal{F} \times \mathcal{T}^n & \rightarrow & \mathbb{C}^n \\ & (f, p + q\sqrt{-1}) & \mapsto & (f \circ \exp)(p + q\sqrt{-1}) \end{array}$$

with respect to the variable in $\mathcal{T}^n$ at the point $p+q\sqrt{-1}$ has rank $2n$. We are then in the situation of the implicit function theorem, and there is (locally) a smooth function $G : \mathcal{F} \rightarrow \mathcal{T}^n$ such that for $\hat{f}$ in a neighborhood of f, we have $ev(\hat{f}, G(\hat{f})) \equiv 0$ and $G(f) = p + q\sqrt{-1}$.

The condition number of f at $(p + q\sqrt{-1})$ is usually defined as

$$\mu(f; p + q\sqrt{-1}) = \|DG_f\| \quad .$$

This definition is sensitive to the norm used in the space of linear maps between tangent spaces $L(T_f\mathcal{F}, T_{(p,q)}\mathcal{T}^n)$. In general, one would like to use an operator norm, related to some natural Hermitian or Riemannian structure on $\mathcal{F}$ and $\mathcal{T}^n$.

In the previous section, we already defined an inner product in each coordinate subspace $\mathcal{F}_{A_i}$, given by the variance matrix C_i. Since the evaluation function is homogeneous in each coordinate, it makes sense to projectivize each of the coordinate spaces $\mathcal{F}_{A_i}$ (with respect to the inner product $\langle \cdot, \cdot \rangle_{C_i^{-1}}$). Alternatively, we can use the Fubini-Study metric in each of the $\mathcal{F}_{A_i}$'s. By doing so, we are endowing $\mathcal{F}$ with a Fubini-like metric that is scaling-invariant. We will treat $\mathcal{F}$ as a multiprojective space, and write $\mathbb{P}(\mathcal{F})$ for $\mathbb{P}(\mathcal{F}_{A_1}) \times \cdots \times \mathbb{P}(\mathcal{F}_{A_n})$.

Another useful metric in $\mathbb{P}(\mathcal{F})$ is given by

$$d_{\mathbb{P}}(f,g)^2 \overset{\text{def}}{=} \sum_{i=1}^{n} \left(\min_{\lambda \in \mathbb{C}^*} \frac{\|f^i - \lambda g^i\|}{\|f^i\|} \right)^2 \quad .$$

Each of the terms in the sum above corresponds to the square of the sine of the Fubini (or angular) distance between f^i and g^i. Therefore, $d_{\mathbb{P}}$ is never larger than the Hermitian distance between points in $\mathcal{F}$, but is a correct first-order aproximation of the distance when $g \to f$ in $\mathbb{P}(\mathcal{F})$. (Compare with [3]).

While $\mathcal{F}$ admits a natural Hermitian structure, the solution-space $\mathcal{T}^n$ admits n possibly different Hermitian structures, corresponding to each of the Kähler forms ω_{A_i}.

In order to elucidate what the natural definition of a condition number for mixed systems of polynomials is, we will interpret the condition number as the inverse of the distance to the *discriminant locus.* Given $p + q\sqrt{-1} \in \mathcal{T}^n$, we set:

$$\mathcal{F}_{(p,q)} = \{ f \in \mathcal{F} : ev(f;(p,q)) = 0 \}$$

and we set $\Sigma_{(p,q)}$ as the space of degenerate polynomial systems in $\mathcal{F}_{(p,q)}$. Since the fiber $\mathcal{F}_{(p,q)}$ inherits the metric structure of $\mathcal{F}$, we can speak of the distance to the discriminant locus along a fiber. In this setting, Theorem 3 in [3] becomes:

Theorem 2 (Condition number theorem). *Under the notations above, if (p,q) is a non-degenerate root of f,*

$$\max_{\|\dot{f}\| \leq 1} \min_i \|DG_f \dot{f}\|_{A_i} \leq \frac{1}{d_{\mathbb{P}}(f, \Sigma_{(p,q)})} \leq \max_{\|\dot{f}\| \leq 1} \max_i \|DG_f \dot{f}\|_{A_i} \quad .$$

There are two interesting particular cases. First of all, if $A_1 = \cdots = A_n$ and $C_1 = \cdots = C_n$, we obtain an equality:

Corollary 2.1 (Condition number theorem for unmixed systems). *Let $A_1 = \cdots = A_n$ and $C_1 = \cdots = C_n$, then under the hypotheses of Theorem 2,*

$$\mu(f;(p,q)) \overset{\text{def}}{=} \max_{\|\dot{f}\| \leq 1} \min_i \|DG_f \dot{f}\|_{A_i} = \max_i \max_{\|\dot{f}\| \leq 1} \|DG_f \dot{f}\|_{A_i} = \frac{1}{d_{\mathbb{P}}(f, \Sigma_{(p,q)})} \quad .$$

We can also obtain a version of Shub and Smale's condition number theorem (Theorem 3 in [3]) for dense systems as a particular case, once we choose the correct variance matrices:

Corollary 2.2 (Condition number theorem for dense systems). *Let $d_1, \cdots, d_n$ be positive integers, and let A_i be the n-columns matrix having all possible rows with non-negative entries adding up to at most d_i. Let*

$$C_i = \frac{1}{d_i} \operatorname{Diag} \left(\frac{d_i! \ \cdot}{(A_i)_1^\alpha!(A_i)_2^\alpha! \cdots (A_i)_n^\alpha!(d_i - \sum_{j=1}^n (A_i)_j^\alpha)!} \right) \quad .$$

Then,

$$\mu(f;(p,q)) \overset{\text{def}}{=} \max_{\|\dot{f}\| \leq 1} \min_i \|DG_f \dot{f}\|_{A_i} = \max_i \max_{\|\dot{f}\| \leq 1} \|DG_f \dot{f}\|_{A_i} = \frac{1}{d_{\mathbb{P}}(f, \Sigma_{(p,q)})} \quad .$$

The factor $\frac{1}{d_i}$ in the definition of the variance matrix C_i corresponds to the factor $\sqrt{d_i}$ in the definition of the normalized condition number in [3] .

In the general mixed case, we would like to interpret the two "minmax" bounds as condition numbers related to some natural Hermitian or Finslerian structures on $\mathcal{T}^n$. See [10] for a discussion

Theorem 2 is very similar to Theorem D in [5], but the philosophy here is radically different. Instead of changing the metric in the fiber $\mathcal{F}_{(p,q)}$, we consider the inner product in $\mathcal{F}$ as the starting point of our investigation. Theorem 2 gives us some insight about reasonable metric structures in $\mathcal{T}^n$.

As in Theorem 1, let U be a measurable set of $\mathcal{T}^n$. In view of Theorem 2, we define a restricted condition number (with respect to U) by:

$$\mu(f;U) \overset{\text{def}}{=} \frac{1}{\min_{(p,q) \in U} d_{\mathbb{P}}(f, \Sigma_{(p,q)})}$$

where the distance $d_{\mathbb{P}}$ is taken along the fiber $\mathcal{F}_{(p,q)} = \{f : (f \circ \exp)(p + q\sqrt{-1}) = 0\}$.

Although we do not know in general how to bound the expected value of $\mu(f; \mathcal{T}^n)$, we can give a convenient bound for $\mu(f;U)$ whenever U is compact and in some cases where U is not compact.

The group $GL(n)$ acts on $T_{(p,q)}\mathcal{T}^n$ by sending $(\dot{p},\dot{q})$ into $(L\dot{p},L\dot{q})$, for any $L \in GL(n)$. In more intrinsic terms, J and the $GL(n)$-action commute. With this convention, we can define an intrinsic invariant of the mixed structure $(\mathcal{T}^n, \omega_{A_1}, \cdots, \omega_{A_n}, J)$:

Definition 1. The *mixed dilation* of the tuple $(\omega_{A_1}, \cdots, \omega_{A_n})$ is:

$$\kappa(\omega_{A_1}, \cdots, \omega_{A_n}; (p,q)) \stackrel{\text{def}}{=} \min_{L \in GL(n)} \max_i \frac{\max_{\|u\|=1} (\omega_{A_i})_{(p,q)}(Lu, JLu)}{\min_{\|u\|=1} (\omega_{A_i})_{(p,q)}(Lu, JLu)} \quad .$$

Given a set U, we define:

$$\kappa_U \stackrel{\text{def}}{=} \sup_{(p,q) \in U} \kappa(\omega_{A_1}, \cdots, \omega_{A_n}; (p,q)) \quad ,$$

provided the supremum exists, and $\kappa_U = \infty$ otherwise.

We will bound the expected number of roots with condition number $\mu > \varepsilon^{-1}$ on U in terms of the mixed volume form, the mixed dilation κ_U and the expected number of ill-conditioned roots in the *linear case.* The linear case corresponds to polytopes and variances below:

$$A_i^{\text{Lin}} = \begin{bmatrix} 0 & \cdots & 0 \\ 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \qquad\qquad C_i^{\text{Lin}} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

Theorem 3 (Expected value of the condition number). *Let $\nu^{\text{Lin}}(n,\varepsilon)$ be the probability that a random n–variate linear complex polynomial has condition number larger than ε^{-1}. Let $\nu^A(U,\varepsilon)$ be the probability that $\mu(f,U) > \varepsilon^{-1}$ for a normal random polynomial system f with supports $A_1, \cdots, A_n$ and variance $C_1, \cdots, C_n$.*

Then,

$$\nu^A(U,\varepsilon) \le \frac{\int_U \bigwedge \omega_{A_i}}{\int_U \bigwedge \omega_{A_i^{\text{Lin}}}} \; \nu^{\text{Lin}}(n, \sqrt{\kappa_U}\varepsilon) \quad .$$

There are a few situations where we can assert that $\kappa_U = 1$. For instance,

Corollary 3.1. *Under the hypotheses of Theorem 3, if $A = A_1 = \cdots = A_n$ and $C = C_1 = \cdots = C_n$, then*

$$\nu^A(U,\varepsilon) \le \text{Vol}(U) \; \nu^{\text{Lin}}(n,\varepsilon) \quad .$$

The dense case (Theorem 1 p. 237 in [3]) is also a consequence of Theorem 3.

Remark 1. We interpret $\nu^{\mathrm{Lin}}(n,\varepsilon)$ as the probability that a random linear polynomial f is at multiprojective distance less than ε from the discriminant variety $\Sigma_{(p,q)}$. Let $g \in \Sigma_{(p,q)}$ be such that the following minimum is attained:

$$d_{\mathbb{P}}(f, \Sigma_{(p,q)})^2 = \inf_{\substack{g\in\Sigma_{(p,q)} \\ \lambda\in(\mathbb{C}^*)^n}} \sum_{i=1}^{n} \frac{\|f^i - \lambda_i g^i\|^2}{\|f^i\|^2} \quad .$$

Without loss of generality, we may scale g such that $\lambda_1 = \cdots = \lambda_n = 0$. In that case,

$$d_{\mathbb{P}}(f, \Sigma_{(p,q)})^2 = \sum_{i=1}^{n} \frac{\|f^i - g^i\|^2}{\|f^i\|^2} \geq \frac{\sum_{i=1}^{n} \|f^i - g^i\|^2}{\sum_{i=1}^{n} \|f^i\|^2} \quad .$$

The right hand term is the projective distance to the discriminant variety along the fiber, in the sense of [3]. Since we are in the linear case, this may be interpreted as the inverse of the condition number of f in the sense of [3] .

Recall that each f^i is an independent random normal linear polynomial of degree 1, and that C_i is the identity. Therefore, each f^i_α is an i.i.d. Gaussian variable. If we look at the system f as a random variable in $\mathbb{P}^{n(n+1)-1}$, then we obtain the same probability distribution as in [3]. Then, using Theorem 6 p. 254 *ibid*, we deduce that

$$\nu^{\mathrm{Lin}}(n,\varepsilon) \leq \frac{n^3(n+1)\Gamma(n^2+n)}{\Gamma(n^2+n-2)}\varepsilon^4 \quad . \blacksquare$$

5 Real Polynomials

Shub and Smale showed in [18] that the expected number of real roots, in the dense case (with unitarily invariant probability measure) is exactly the square root of the expected number of roots.

Unfortunately, this result seems to be very hard to generalize to the unmixed case. Under certain conditions, explicit formulæ for the unmixed case are available [15]. Also, less explicit bounds for the multi-homogeneous case were given by [14].

Here, we will give a very coarse estimate in terms of the square root of the mixed volume:

Theorem 4. *Let U be a measurable set in $\mathbb{R}^n$, with total Lebesgue volume $\lambda(U)$. Let $A_1, \cdots, A_n$ and $C_1, \cdots, C_n$ be as above. Let f be a normal random real polynomial system. Then the average number of real roots of f in* $\exp U \subset$

$(R_*^+)^n$ is bounded above by

$$(4\pi^2)^{-n/2}\sqrt{\lambda(U)}\sqrt{\int_{\substack{(p,q)\in\mathcal{T}^n\\ p\in U}} n!d\mathcal{T}^n} \quad .$$

This is of interest when n and U are fixed. In that case, the expected number of positive real roots (hence of real roots) grows as the square root of the mixed volume.

It is somewhat easier to investigate real random polynomials in the unmixed case.

Let $\nu_{\mathbb{R}}(n,\varepsilon)$ be the probability that a linear random real polynomial has condition number larger than ε^{-1}.

Theorem 5. *Let $A = A_1 = \cdots = A_n$ and $C = C_1 = \cdots = C_n$. Let $U \subseteq \mathbb{R}^n$ be measurable. Let f be a normal random real polynomial system. Then,*

$$\text{Prob}\left[\mu(f,U) > \varepsilon^{-1}\right] \leq E(U)\ \nu_{\mathbb{R}}(n,\varepsilon)$$

where $E(U)$ is the expected number of real roots on U.

Notice that $E(U)$ depends on C. Even if we make $U = \mathbb{R}^n$, we may still obtain a bound depending on C.

6 Mechanical Interpretation

The momentum map defined above is also the the momentum map (in the sense of [22]) associated to a certain Lie group action, namely the natural action of the n-torus on the *toric* manifold $\mathcal{T}^n$:

The n-torus $\mathbb{T}^n = \mathbb{R}^n \pmod{2\pi\,\mathbb{Z}^n}$ acts on $\mathcal{T}^n$ by

$$\rho : (p,q) \mapsto (p, q+\rho) \quad ,$$

where $\rho \in \mathbb{T}^n$.

This action preserves the symplectic structure, since it fixes the p-variables and translates the q-variables. Also, the Lie algebra of $\mathbb{T}^n$ is $\mathbb{R}^n$. An element ξ of $\mathbb{R}^n$ induces an *infinitesimal* action (i.e. a vector field) X_ξ in $\mathcal{T}^n$.

This vector field is the derivation that to any smooth function f associates:

$$(X_\xi)_{(p,q)}(f) = \iota_\xi(\frac{1}{2}\omega_A)_{(p,q)}(df) \overset{\text{def}}{=} (\frac{1}{2}\omega_A)_{(p,q)}(\xi, df) \quad .$$

If we write $df = d_p f dp + d_q f dq$, then this formula translates to:

$$(X_\xi)_{(p,q)}(f) = -\xi^T(D^2 g_A)_p d_p f$$

This vector field is Hamiltonian: if $(p(t), q(t))$ is a solution of the equation

$$(\dot{p}(t), \dot{q}(t)) = (X_\xi)_{p(t),q(t)}$$

then we can write

$$\begin{cases} \dot{p} = \frac{\partial H_\xi}{\partial q} \\ \dot{q} = -\frac{\partial H_\xi}{\partial p} \end{cases},$$

where $H_\xi = \frac{1}{2}\nabla g_A(p) \cdot \xi$.

This construction associates to every $\xi \in \mathbb{R}^n$, the Hamiltonian function $H_\xi = \nabla g_A(p) \cdot \xi$. The term $\nabla g_A(p)$ is a function of p, with values in $(\mathbb{R}^n)^\vee$ (the dual of $\mathbb{R}^n$). In more general Lie group actions, the momentum map takes values in the dual of the Lie algebra, so that the pairing $\nabla g_A(p) \cdot \xi$ always makes sense. A Lie group action with such an expression for the Hamiltonian is called *Hamiltonian* or *Strongly Hamiltonian*.

7 Acknowledgements

Steve Smale provided valuable inspiration for us to develop a theory similar to [17,18,19,20,21] for sparse polynomial systems. He also provided examples on how to eliminate the dependency upon unitary invariance in the dense case.

The paper by Gromov [7] was of foremost importance to this research. To the best of our knowledge, [7] is the only clear exposition available of mixed volume in terms of a wedge of differential forms. We thank Mike Shub for pointing out that reference, and for many suggestions.

We would like to thank Jean-Pierre Dedieu for sharing his thoughts with us on Newton iteration in Riemannian and quotient manifolds.

Also, we would like to thank Felipe Cucker, Alicia Dickenstein, Ioannis Emiris, Askold Khovanskii, Eric Kostlan, T.Y. Li, Martin Sombra, and Jorge P. Zubelli for their suggestions and support.

This paper was written while G.M. was visiting the Liu Bie Ju Center for Mathematics at the City University of Hong Kong. He wishes to thank CityU for their generous support.

References

1. Abraham, Ralph and Marsden, Jerrold E., *Foundations of mechanics*, Second edition, revised and enlarged, With the assistance of Tudor Ratiu and Richard Cushman, Benjamin/Cummings Publishing Co. Inc. Advanced Book Program, Reading, Mass., 1978.

2. Abreu, Miguel, *"Kähler geometry of toric manifolds in symplectic coordinates"*, April, 2000, Preprint, Mathematics ArXiv DG/0004122, `http://front.math.ucdavis.edu`.
3. Blum, Lenore; Cucker, Felipe; Shub, Michael; and Smale, Steve, *Complexity and real computation*, With a foreword by Richard M. Karp, Springer-Verlag, New York, 1998.
4. Bernstein, D. N., *"The number of roots of a system of equations"*, Functional Anal. Appl., Vol. 9, No. 2, (1975), pp. 183–185.
5. Dedieu, Jean-Pierre, *"Approximate solutions of numerical problems, condition number analysis and condition number theorem"*, The mathematics of numerical analysis (Park City, UT, 1995), pp. 263–283, Amer. Math. Soc., Providence, RI, 1996.
6. Edelman, Alan and Kostlan, Eric, *"How many zeros of a random polynomial are real?"*, Bull. Amer. Math. Soc. (N.S.), American Mathematical Society. Bulletin. New Series, vol. 32, 1995, no. 1, pp. 1–37.
7. Gromov, M., *"Convex sets and Kähler manifolds"*, Advances in differential geometry and topology, pp. 1–38, World Sci. Publishing, Teaneck, NJ, 1990.
8. Kazarnovskiĭ, B. Ja., *"On zeros of exponential sums"*, Soviet Math. Doklady, vol. 23, 1981, no. 2, pp. 347–351.
9. Eric Kostlan, *"Random Polynomials and the Statistical Fundamental Theorem of Algebra"*, Preprint, MSRI, 1987.
10. Gregorio Malajovich and J. Maurice Rojas, *"Random Sparse Polynomial Systems"*, Preprint, City University of Hong Kong and Math Archive, 2000.
11. Gregorio Malajovich and Jorge Zubelli, *"On the Geometry of Graeffe Iteration"*, Journal of Complexity (To appear).
12. Gregorio Malajovich and Jorge Zubelli, *"Tangent Graeffe Iteration"*, Numerische Mathematik (To appear).
13. McDuff, Dusa and Salamon, Dietmar, *Introduction to symplectic topology*, second edition, The Clarendon Press Oxford University Press, New York, 1998.
14. Andrew McLennan, *"The expected number of real roots of a multihomogeneous system of polynomial equations"*, American Journal of Mathematics, to appear.
15. J. Maurice Rojas, *"On the Average number of Real Roots of Certain Random Sparse Polynomial Systems"*, Lectures in Applied Mathematics, vol. 32, 1996, pp. 689–699.
16. Sangwine-Yager, J. R., *"Mixed volumes"*, Handbook of convex geometry, Vol. A, B, pp. 43–71, North-Holland, Amsterdam, 1993.

17. Shub, Michael and Smale, Steve, *"Complexity of Bézout's theorem I — Geometric aspects"*, Journal of the American Mathematical Society, vol. 6, 1993, no. 2, pp. 459–501.
18. Shub, Mike and Smale, Steve, *"Complexity of Bezout's theorem II — Volumes and probabilities"*, Computational algebraic geometry (Nice, 1992), pp. 267–285, Birkhäuser Boston, Boston, MA, 1993.
19. Shub, Michael and Smale, Steve, *"Complexity of Bezout's theorem III — Condition number and packing"*, Festschrift for Joseph F. Traub, Part I, Journal of Complexity, vol. 9, 1993, no. 1, pp. 4–14.
20. Shub, Michael and Smale, Steve, *"Complexity of Bezout's theorem IV — Probability of success; extensions"*, SIAM Journal on Numerical Analysis, vol. 33, 1996, no. 1, pp. 128–148.
21. Shub, Mike and Smale, Steve, *"Complexity of Bezout's theorem V — Polynomial time*, Selected papers of the Workshop on Continuous Algorithms and Complexity (Barcelona, 1993), Theoretical Computer Science, vol. 133, 1994, no. 1, pp. 141–164.
22. Smale, Steve, *"Topology and mechanics I"*, Invent. Math., vol. 10, 1970, pp. 305–331.
23. Souriau, J.-M., *"Structure des systèmes dynamiques"*, Maîtrises de mathématiques, Dunod, Paris, 1970.
24. Oleg Viro, *"Dequantization of real algebraic geometry on logarithmic paper"*, June, 2000, Preprint, Mathematics ArXiv AG/0005163, `http://front.math.ucdavis.edu`

ASYMPTOTIC ACCELERATION OF THE SOLUTION OF MULTIVARIATE POLYNOMIAL SYSTEMS OF EQUATIONS

B. MOURRAIN

INRIA, GALAAD, BP 93, 06902 Sophia-Antipolis, France
E-mail: mourrain@sophia.inria.fr

V.Y. PAN

Dept. of Mathematics and Computer Science, Lehman College, City University of New York, Bronx, NY 10468
E-mail: vpan@lehman.cuny.edu

O. RUATTA

INRIA, GALAAD, BP 93, 06902 Sophia-Antipolis, France
E-mail: oruatta@sophia.inria.fr

We propose new Las Vegas randomized algorithms for the solution of a square non-degenerate system of equations. The algorithms use $\mathcal{O}(\delta\, 3^n D^2 \log(D)\log(b))$ arithmetic operations to approximate all real roots of the system as well as all roots lying in a fixed n-dimensional box or disc. Here D is an upper bound on the number of all complex roots of the system, δ is the number of real roots or the roots lying in the box or disc, $\epsilon = 2^{-b}$ is the required upper bound on the output errors. We also yield the bound $\mathcal{O}(3^n D^2 \log(D))$ on the complexity of counting the numbers of all roots in a fixed box (disc) and all real roots. For a large class of inputs and typically in practical computations, the factor δ is much smaller than D, $\delta = o(D)$. This improves by order of magnitude the known complexity estimates of order at least $3^n D^3 \log(b)$ or $3^n D^3$, which so far are the record ones even for the approximation of a single root of a system and for each of the cited counting problems, respectively. Our progress relies on proposing several novel techniques. In particular, we exploit the structure of matrices associated to a given polynomial system and relate it to the associated linear operators, dual space of linear forms, and normal forms of polynomials in the quotient algebra; furthermore, our techniques support the new nontrivial extension of the matrix sign and quadratic inverse power iterations to the case of multivariate polynomial systems, where we emulate the recursive splitting of a univariate polynomial into factors of smaller degree.

1 Introduction.

The classical problem of solving a multivariate polynomial system of equations is presently the subject of intensive research and one of the central practical and theoretical problems in the area of algebraic computation (see some bibliography in [18], [4], [28], [13].) It has major applications, for instance, to robotics, computer modelling and graphics, molecular biology, and computational algebraic geometry.

The oldest approach to the solution is the elimination method, reducing the problem to the computation of the associated resultant or its multiples. This classical method evolved in the old works by Bezout, Dixon, and Macaulay (see e.g. [18], [41]), then remained largely ignored by the researchers and algorithm designers but it was resurrected by Canny in the 80s to become a very popular approach since then. One of the major further steps was the reduction of the solution of a multivariate polynomial system to matrix operations, in particular, by rational transformation of the original problem into a matrix eigenproblem (cf. [1], [14], [23], [21], [8]).

The approach has been explored and extended by many researchers, has been exploited in practice of algebraic computing, and also supported the record asymptotic upper bound $\mathcal{O}^*(D^3)$ on the arithmetic computational complexity of the solution of a polynomial system having a finite number of roots. Here and hereafter, $\mathcal{O}^*(s)$ stands for $\mathcal{O}(s \log^c s)$, c denoting a constant independent of s, and D is an upper bound on the number of roots of the given polynomial system. (For D, one may choose either the Bezout bound, $\prod_i d_i$, d_i denoting the maximum degree in the i-th variable in all monomials of the system, or the Bernstein bound, which is much smaller for sparse systems and equals the mixed volume of the associated Newton polytope, defined by the exponents of the monomials.) The cited record bound $\mathcal{O}^*(D^3)$ is due to [35] but also has several other derivations and has been staying as a stable landmark for the multivariate polynomial system solving, much like the complexity bound $\mathcal{O}(N^3)$ for solving nonsingular linear system of N equations, which was supported by Gaussian elimination and stayed as a landmark and a record until Strassen's celebrated result of 1969. In fact, even in the case of solving non-degenerate polynomial system as well as for many subproblems and related problems, no known algorithms support any better bound than $\mathcal{O}(D^3)$. This includes approximation of all real roots of a polynomial system (which is highly important due to applications to robotic and computer graphics), all its roots lying in a fixed n-dimensional box or disc, counting all roots in such a box or disc or all real roots, and even approximation of a single root. Some progress was achieved in [26], where a single root was approximated in $\mathcal{O}^*(3^n D^2)$ time, but under a strong restriction on the input polynomials.

Our new algorithms support the computational cost estimate of $\mathcal{O}^*(3^n D^2)$, for all the listed above subproblems, that is, for both of the counting problems, the computation of a single root, all real roots, and all roots in a fixed box or disc. More precisely, our bound is $\mathcal{O}^*(\delta\, 3^n D^2)$ in the latter two cases, where δ is the number of real roots or roots in the selected box or disc, respectively. In practical applications, such a number is typically much less than D. The number of real roots grows as $\sqrt{D}$ for a large class of input systems [37]. See also for the sparse case [36]. Thus, for all listed problems, we improve the known complexity estimates by an order of magnitude.

We have a reservation from a theoretical point of view, that is, our main algorithm relies on the known effective algorithms for the computation of the normal form of monomials on the boundary of the monomial basis (see section 4). These algorithms exploit structured matrices and in practice, apper to run faster than our subsequent computations (see [15], [29]), but their known theoretical cost bound are greater than the order of D^3 (see [19]).

Our paper addresses the problem of the asymptotic acceleration of the resolution stage where the structure of the quotient algebra $\mathcal{A}$ (associated with the polynomial system) is already described by using the minimal number of parameters, that is, via the normal form of the monomials on the boundary of the basis. From a purely theoretical point of view, we have an alternative approach that avoids the normal form algorithms at the price of using the order of $\mathcal{O}(12^n D^2)$ additional arithmetic operations [27]. This should be technically interesting because no other known approach yields such a bound, but in this paper, we prefer to stay with our present, practically superior version, referring the reader to [27] on the cited theoretical approach.

Our algorithms approximate the roots numerically, and in terms of the required upper bound 2^{-b} (b is the bit precsion) on the output errors of the computed solution, we obtain the estimate $\mathcal{O}(\log b)$. Within a constant factor, such an estimate matches the lower bound of [34] and enables us to yield a high output precision at relatively low cost; this gives us a substantial practical advantage versus the algorithms that only reach $\mathcal{O}(b)$, because the solution of a polynomial system is usually needed with a high precision. We achieve this by using the matrix sign and inverse quadratic iterations, which converge with quadratic rate right from the start. All techniques and results can be extended to the case of sparse input polynomials (see remark 3.16, section 3). In this case, the computation cost bounds become $\mathcal{O}(D\, C_{PolMult})$ where $C_{PolMult}$ is the cost of polynomial multiplication, which is small when the polynomials are sparse (this cost depends on the degree of the polynomials, not only on an upper bound D on the number of roots).

The factor 3^n is a substantial deficiency, of course, but it is still much less than D for the large and important class of input polynomials of degree higher than 3.

Our results require some other restrictions. First, we consider systems with simple roots or well separated roots. In the presence of a cluster, a specific analysis is needed [39] and deserves additional work, which is not in the scope of this paper. Secondly, we need the existence of a non-degenerate linear form, which implies that the quotient algebra $\mathcal{A}$ is a Gorenstein algebra [10,12]. This is the case where the solution set is 0-dimensional and is defined by n equations. If we have more than n equations defining a 0-dimensional variety, we may take their n-random linear combination (see, e.g. [11]), which yields the required Gorenstein property, but this may introduce extra solutions that we will have to remove at the end. Finally, for approximation, our algorithms converge quadratically (using $\mathcal{O}(\log(b))$ steps) but require certain nondegeneracy assumptions (such as uniqueness of the minimum of the value of $|h(\zeta)|$, where ζ is a root and $h(x)$ a polynomial). The latter assumptions can be ensured with a high probability by a random linear transformation of the variables. Even if these assumptions are barely satisfied, the slowdown of the converge is not dramatic, because the convergence is quadratic right from the start.

Similarly, we apply randomization to regularize the computations at the counting stages, and for the auxiliary computation of the non-degenerate linear form in the dual space $\widehat{\mathcal{A}}$. Then again, non-degeneracy is ensured probabilistically, and verified in the subsequent computation (that is, we stay under the LasVegas probabilistic model where failure may occur, with a small probability, but otherwise correctness of the output is ensured).

Some of our techniques should be of independent interest. In particular, we extend the theory of structured matrices to the ones associated to multivariate polynomials and show correlation among computations with such matrices and dual spaces of linear forms. We show some new non-trivial applications of the normal forms of polynomials of the quotient algebra. Furthermore, we establish new reduction from multivariate polynomial computations to some fundamental operations of linear algebra (such as the matrix sign iteration, the quadratic inverse power iteration, and the computation of Schur's complements).

Our progress has some technical similarity to the acceleration of the solution of linear systems of equations via fast matrix multiplication (in particular, we also rely on faster multiplication in the quotient algebra defined by the input polynomials), but even more so, with the recent progress in the univariate polynomial rootfinding via recursive splitting of the input polynomial into factors (cf. [5], [30], [31], [32]). Although recursive splitting into factors may be hard even to comprehend in the case of multivariate polynomial systems, this is exactly the basic step of our novel recursive process, which finally reduces our original problem to ones of small sizes. Of course, we could not achieve splitting in the original space of the variables, but we yield it in terms of idempotent elements of the associated quotient algebra (such elements represent the roots), and for this purpose we had to apply all our advanced techniques.

This approach generalizes the methods of [5] and [31] to the multivariate case. The only missing technical point of our extension of the univariate splitting construction of [31] is the balancing of the splitting, which was the most recent and elusive step in the univariate case (cf. [31], [32]). It is a major challenge to advance our approach to achieve balancing in our recursive splitting process even in the worst case (possibly by using the geometry of discriminant varieties) and, consequently, to approximate all the roots of any specific polynomial system in $\mathcal{O}^*(3^n D^2 \log b)$ arithmetic time. Another goal is the computations in the dual space, as well as with structured matrices. The latter subject is of independent interest too [40,28].

Let us conclude this section with a high level description of our approach. Our solution of polynomial systems consists of the following stages:

1. Compute a basic non-degenerate linear form on the quotient algebra $\mathcal{A}$ associated to a given system of polynomial equations.

2. Compute non-trivial idempotent elements of $\mathcal{A}$.

3. Recover the roots of the given polynomial system from the associated idempotents.

The quotient algebra $\mathcal{A}$ and the dual space of linear forms on it are defined and initially studied in section 2. Stage 1 is elaborated in section 4. Idempotents are computed by iterative algorithms of section 6. Section 7 shows how to recover or to count the roots efficiently when the idempotents are available. The computations are performed in the quotient algebra, and they are reduced to operations in the dual space by using the associated structured (quasi-Toeplitz and quasi-Hankel) matrices. In section 3 we define the classes of such matrices, show their correlation to polynomial computations, and exploit it to operate with such matrices faster. In section 5 we show how the combined power of the latter techniques and the ones developed for working in the dual space enables us to perform rapidly the basic operations in the quotient algebra and, consequently, the computations of sections 6 and 7.

Stage 1 contributes $O(3^n D^2 \log D)$ ops to the overall complexity bound, assuming that the normal form of the monomials on the boundary of a basis is known. The computation of a nontrivial idempotent at stage 2 has cost $O(3^n D^2 \log D \log b)$, which dominates the cost of the subsequent root counting or their recovery from the idempotents. The overall complexity depends on the number of idempotents that one has to compute, which in turn depends on the number δ of roots of interest. So far, we cannot utilize here the effective tools of balanced splitting, available in the similar situation for the univariate polynomial rootfinding. Thus, in the worst case, in each step we split out only a single root from the set of all roots, and then we need δ idempotents.

2 Definitions and preliminaries

Hereafter, $R = \mathbb{C}[x_1, \ldots, x_n]$ is the ring of multivariate polynomials in the variables $x_1, \ldots, x_n$, with coefficients in the complex field $\mathbb{C}$. $\mathbb{Z}$ is the set of integers, $\mathbb{N}$ is its subset of nonnegative integers, $L = \mathbb{C}[x_1^{\pm}, \ldots, x_n^{\pm}]$ is the set of Laurent polynomials with monomial exponents in $\mathbb{Z}^n$. For any $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{Z}^n$, $\mathbf{x}^{\mathbf{a}}$ is the monomial $\mathbf{x}^{\mathbf{a}} = x_1^{a_1} \cdots x_n^{a_n}$. $\lfloor E \rceil$ is the cardinality (that is, the number of elements) of a finite subset E of $\mathbb{Z}^n$. "ops" will stand for "arithmetic operations" in the underlying coefficient ring or field.

2.1 Quotient algebra

To motivate and to demonstrate our study, we will next consider the univariate case, where we have a fixed polynomial $f \in \mathbb{C}[x]$ of degree d with d simple roots: $f(x) = f_d \prod_{i=1}^{d} (x - \zeta_i)$. The quotient algebra of residue polynomials modulo f, denoted by $\mathcal{A} = \mathbb{C}[x]/(f)$, is a vector space of dimension d. Its basis is $(1, x, \ldots, x^{d-1})$. Consider the Lagrange polynomials

$$\mathbf{e}_i = \prod_{j \neq i} \frac{x - \zeta_j}{\zeta_i - \zeta_j}.$$

One immediately sees that $\sum_i \mathbf{e}_i = 1$ and $\mathbf{e}_i \mathbf{e}_j \equiv \mathbf{e}_i(\mathbf{e}_i - 1) \equiv 0$ (for these two polynomials vanish at the roots of f). In other words, the Lagrange polynomials $\mathbf{e}_i$ are orthogonal idempotents in $\mathcal{A}$, and we have $\mathcal{A} = \sum_i \mathbb{C}\mathbf{e}_i$. Moreover, for any polynomial $a \in \mathcal{A}$, we also have $(a - a(\zeta_i))\mathbf{e}_i \equiv 0$, so that $\mathbf{e}_i$ is an eigenvector for the operator of multiplication by a in $\mathcal{A}$, for the eigenvalue $a(\zeta_i)$. These multiplication operators have a diagonal form in the basis $(\mathbf{e}_i)$ of $\mathcal{A}$. According to a basic property of Lagrange polynomials, we have $a \equiv \sum_i a(\zeta_i)\, \mathbf{e}_i(x)$, for any $a \in \mathcal{A}$. Therefore, the dual basis of $(\mathbf{e}_i)$ (formed by the coefficients of the $\mathbf{e}_i$ in this decomposition) consists of the linear forms associating to a its values at the points ζ_i. We will extend this approach to the case of multivariate polynomial systems, which, of course, will require substantial further elaboration and algebraic formalism. We refer the reader to [22], [23], [28], [38] for further details.

Let $f_1, \ldots, f_m$ be m polynomials of R, defining the polynomial system $f_1(x) = 0, \ldots, f_m(x) = 0$. Let I be the ideal generated by these polynomials, that is, the set of polynomial combinations $\sum_i f_i q_i$ of these elements. $\mathcal{A} = R/I$ denotes the quotient ring (algebra) defined in R by I, and $\equiv$ denotes the equality in $\mathcal{A}$. We consider the case, where the quotient algebra $\mathcal{A} = R/I$ *is of finite dimension D over* $\mathbb{C}$. This implies that the set of roots or solutions

$\mathcal{Z}(I) = \{\zeta \in \mathbb{C}^n; f_1(\zeta) = \ldots = f_m(\zeta) = 0\}$ is finite: $\mathcal{Z}(I) = \{\zeta_1, \ldots, \zeta_d\}$ with $d \leq D$. Then we have a decomposition of the form

$$\mathcal{A} = \mathcal{A}_1 \oplus \cdots \oplus \mathcal{A}_d, \tag{1}$$

where $\mathcal{A}_i$ is a local algebra, for the maximal ideal $\mathbf{m}_{\zeta_i}$ defining the root ζ_i. From decomposition (1), we deduce that there exist orthogonal idempotents $\mathbf{e}_1, \ldots, \mathbf{e}_d$ satisfying

$$\mathbf{e}_1 + \cdots + \mathbf{e}_d \equiv 1, \text{ and } \mathbf{e}_i\, \mathbf{e}_j \equiv \begin{cases} 0 \text{ if } i \neq j, \\ \mathbf{e}_i \text{ if } i = j. \end{cases}$$

If $I = Q_1 \cap \cdots \cap Q_d$ is the minimal primary decomposition of I, we have $\mathbf{e}_i \mathcal{A} \sim R/Q_i$, where $\mathcal{A}_i = \mathbf{e}_i\, \mathcal{A}$ is a local algebra, for the maximal ideal $\mathbf{m}_{\zeta_i}$ defining the root ζ_i. Thus, to any root $\zeta \in \mathcal{Z}$, we associate an idempotent $\mathbf{e}_\zeta$.

2.2 Dual space

Let $\widehat{R}$ denote the dual of the $\mathbb{C}$-vector space R, that is, the space of linear forms

$$\begin{aligned} \Lambda : R &\to \mathbb{C} \\ p &\mapsto \Lambda(p). \end{aligned}$$

(R will be the primal space for $\widehat{R}$.) Let us recall two celebrated examples, that is, the *evaluation at a fixed point* ζ,

$$\begin{aligned} \mathbf{1}_\zeta \ : R &\to \mathbb{C} \\ p &\mapsto p(\zeta), \end{aligned}$$

and the map

$$\begin{aligned} (\mathbf{d}^{\mathbf{a}} = (\mathbf{d}_1)^{a_1} \cdots (\mathbf{d}_n)^{a_n}) : R &\to \mathbb{C} \\ p &\mapsto \frac{1}{\prod_{i=1}^n a_i!} \, (d_{x_1})^{a_1} \cdots (d_{x_n})^{a_n} \, (p)(0), \end{aligned} \tag{2}$$

where $\mathbf{a} = (a_1, \ldots, a_n)$ is any vector from $\mathbb{N}^n$, and d_{x_i} is the partial derivative with respect to the variable x_i. For any $\mathbf{b} = (b_1, \ldots, b_n) \in \mathbb{N}^n$, we have

$$\mathbf{d}^{\mathbf{a}}(\mathbf{x}^{\mathbf{b}}) = \begin{cases} 1 \text{ if } \forall i, a_i = b_i, \\ 0 \text{ otherwise.} \end{cases}$$

Therefore, $(\mathbf{d}^{\mathbf{a}})_{\mathbf{a}\in\mathbb{N}^n}$ is the dual basis of the primal monomial basis. Thus, we decompose any linear form $\Lambda \in \widehat{R}$ as

$$\Lambda = \sum_{\mathbf{a}\in\mathbb{N}^n} \Lambda(\mathbf{x}^{\mathbf{a}})\,\mathbf{d}^{\mathbf{a}}. \tag{3}$$

Hereafter, **we will identify $\widehat{R}$ with $\mathbb{C}[[\mathbf{d}_1,\dots,\mathbf{d}_n]]$.** The map $\Lambda \to \sum_{\mathbf{a}\in\mathbb{N}^n} \Lambda(\mathbf{x}^{\mathbf{a}})\,\mathbf{d}^{\mathbf{a}}$ defines a one-to-one correspondence between the set of linear forms Λ and the set $\mathbb{C}[[\mathbf{d}_1,\dots\mathbf{d}_n]] = \mathbb{C}[[\mathbf{d}]] = \{\sum_{\mathbf{a}\in\mathbb{N}^n} \lambda_{\mathbf{a}}\mathbf{d}_1^{a_1}\cdots\mathbf{d}_n^{a_n}\}$ of polynomials in the variables $\mathbf{d}_1,\dots,\mathbf{d}_n$.

The evaluation at 0 corresponds to the constant 1, under this definition. It will be also denoted by $\delta_0 = \mathbf{d}^0$.

We will denote by $\widehat{\mathcal{A}}$ and also by $I^{\perp}$ the subspace of $\widehat{R}$ made of those linear forms that vanish on the ideal I.

We now define multiplication of a linear form by a polynomial ($\widehat{R}$ is an R-module) as follows. For any $p \in R$ and $\Lambda \in \widehat{R}$, we write

$$\begin{aligned} p\star\Lambda : R &\to \mathbb{C} \\ q &\mapsto \Lambda(pq). \end{aligned}$$

For any pair of elements $p \in R$ and $a \in \mathbb{N}$, $a > 1$, we have

$$(d_{x_i})^a\,(x_i\,p)(0) = a\;(d_{x_i})^{a-1}\,p(0).$$

Consequently, for any pair $(p,\mathbf{a})$, $p \in R$, $\mathbf{a} = (a_1,\dots,a_n) \in \mathbb{N}^n$ (where $a_i \neq 0$ for a fixed i), we obtain

$$\begin{aligned} x_i\star\mathbf{d}^{\mathbf{a}}(p) &= \mathbf{d}^{\mathbf{a}}(x_i\,p) \\ &= \mathbf{d}_1^{a_1}\cdots\mathbf{d}_{i-1}^{a_{i-1}}\mathbf{d}_i^{a_i-1}\mathbf{d}_{i+1}^{a_{i+1}}\cdots\mathbf{d}_n^{a_n}(p), \end{aligned}$$

that is, x_i acts as the *inverse* of $\mathbf{d}_i$ in $\mathbb{C}[[\mathbf{d}]]$. For this reason such a representation is referred to as the *inverse systems* (see, for instance, [20]). If $a_i = 0$, then $x_i\star\mathbf{d}^{\mathbf{a}}(p) = 0$, which allows us to redefine the product $p\star\Lambda$ as follows:

Proposition 2.1 *For any pair $p,q \in R$ and any $\Lambda(\mathbf{d}) \in \mathbb{C}[[\mathbf{d}]]$, we have*

$$p\star\Lambda(q) = \Lambda(pq) = \pi_+(p(\mathbf{d}^{-1})\,\Lambda(\mathbf{d}))(q),$$

where π_+ is the projection mapping Laurent series onto the space generated by the monomials in $\mathbf{d}$ with positive exponents.

This yields the following algorithm:

Algorithm 2.2 FOR ANY POLYNOMIAL $p \in \langle\mathbf{x}^{\alpha}\rangle_{\alpha\in E}$ AND A VECTOR $[\Lambda(\mathbf{x}^{\beta})]_{\beta\in E+F}$, COMPUTE THE VECTOR $[p\star\Lambda(\mathbf{x}^{\beta})]_{\beta\in F}$ *as follows:*

- *Write* $\tilde{\Lambda}(\mathbf{d}) = \sum_{\beta\in E+F} \Lambda(\mathbf{x}^{\beta})\,\mathbf{d}^{\beta}$.

- *Compute the product* $\rho(\mathbf{d}) = p(\mathbf{d}^{-1})\tilde{\Lambda}(\mathbf{d})$ *in* $\mathbb{C}[\mathbf{d}, \mathbf{d}^{-1}]$, *and*
- *keep the coefficients* ρ_α *of* $\mathbf{d}^\alpha$ *for* $\alpha \in F$.

3 Quasi-Toeplitz and quasi-Hankel matrices

In this section we describe the structure of the matrices and some tools that we will use for our algorithm design.

Let us recall first the known arithmetic complexity bounds for polynomial multiplication (see [2], pp. 56-64), which is the basic step of our subsequent algorithms. Let $C_{PolMult}(E,F)$ denote the number of ops (that is, of arithmetic operations) required for the multiplication of a polynomial with support in E by a polynomial with support in F.

Theorem 3.1 *Let* $E{+}F = \{\alpha^i = (\alpha_1^{(i)}, \ldots, \alpha_n^{(i)}),\ i = 1, \ldots, N\}$ *with* $|\alpha^{(i)}| = \sum_j \alpha_j^{(i)} = d_i$ *for* $i = 1, \ldots, N$ *and* $d = max_i(d_i)$. *Let* $C_{K;Eval}(G)$ *ops suffice to evaluate a polynomial with a support* G *on a set of* K *points. Then we have*

$$C_{PolMut}(E,F) = \mathcal{O}\left(C_{N;Eval}(E) + C_{N;Eval}(F) + N\left(\log^2(N) + \log(d)\right)\right).$$

Proof. Apply the evaluation-interpolation techniques to multiply the two polynomials (cf. [2]). That is, first evaluate the input polynomials on a fixed set of N points, then multiply pairwise the computed values to obtain the values of the product on the same set, and finally interpolate from these values and compute the coefficients of the product by applying the (sparse) polynomial interpolation algorithm (cf. [2]). By summarizing the computational cost estimates, we obtain the theorem. □

For special sets E and F, we have better bounds.

Theorem 3.2 *Let* $E_d = [0, \ldots, d-1] \subset \mathbb{N}$. *Then*

$$C_{PolMult}(E_d, E_d) = \mathcal{O}(d \log(d)).$$

Theorem 3.3 *Let* $E_\mathbf{c} = \{(\alpha_1, \ldots, \alpha_n)\ ;\ 0 \le \alpha_i \le c_i - 1\}$, $E_\mathbf{d} = \{(\beta_1, \ldots, \beta_n)\ ;\ 0 \le \beta_i \le d_i - 1\}$, $c = max\{c_1, \ldots, c_n\}$, *and* $d = max\{d_1, \ldots, d_n\}$. *Then we have*

$$C_{PolMult}(E_\mathbf{c}, E_\mathbf{d}) = \mathcal{O}(M \log(M)),$$

where $M = f^n$, *and* $f = c + d + 1$.

Theorem 3.4 *Let* $E_{f,n}$ *be the set of exponents having total degree at most* f *in* n *variables. Then*

$$C_{PolMult}(E_{c,n}, E_{d,n}) = \mathcal{O}(T \log^2(T)),$$

where $T = \binom{n+c+d}{n}$ *is the number of monomials of degree at most* $c + d$ *in* n *variables.*

Remark 3.5 *Theorems 3.1 and 3.3 correspondly respectively to lattice points in a product of intervals or in the scaled standard simplex, can be extended to the computations over any ring of constants (rather than over the complex field) at the expense of increasing their complexity bounds by at most the factors of* $\log\log(N)$ *or* $\log\log(M)$*, respectively* [2]. *Theorem 3.4 can be extended similarly to any field of constants having characteristic* 0.

Next, by following [28]-[27], we will extend the definitions of Toeplitz and Hankel matrices to the multivariate case. As we will see, these structures are omnipresent, when we solve polynomial systems.

Definition 3.6 *Let E and F be two finite subsets of $\mathbb{N}^n$ and let $M = (m_{\alpha,\beta})_{\alpha\in E,\beta\in F}$ be a matrix whose rows are indexed by the elements of E and columns by the elements of F. Let $\mathtt{i}$ denote the i^{th} basis coordinate vector of $\mathbb{N}^n$.*

- *$M = [m_{\alpha,\beta}]_{\alpha\in E,\beta\in F}$ is an (E,F)* quasi-Toeplitz *matrix if and only if, for all $\alpha \in E, \beta \in F$, the entries $m_{\alpha,\beta} = t_{\alpha-\beta}$ depend only on $\alpha - \beta$, that is, if and only if, for $i = 1,\dots,n$, we have $m_{\alpha+\mathtt{i},\beta+\mathtt{i}} = m_{\alpha,\beta}$, provided that $\alpha, \alpha + \mathtt{i} \in E; \beta, \beta + \mathtt{i} \in F$; such a matrix M is associated with the polynomial $T_M(\mathbf{x}) = \sum_{\mathbf{u}\in E+F} t_{\mathbf{u}}\,\mathbf{x}^{\mathbf{u}}$.*

- *M is an (E,F)* quasi-Hankel *matrix if and only if, for all $\alpha \in E, \beta \in F$, the entries $m_{\alpha,\beta} = h_{\alpha+\beta}$ depend only on $\alpha + \beta$, that is, if and only if, for $i = 1,\dots,n$, we have $m_{\alpha-\mathtt{i},\beta+\mathtt{i}} = m_{\alpha,\beta}$ provided that $\alpha, \alpha - i \in E; \beta, \beta + \mathtt{i} \in F$; such a matrix M is associated with the Laurent polynomial $H_M(\mathbf{d}) = \sum_{\mathbf{u}\in E-F} h_{\mathbf{u}}\mathbf{d}^{\mathbf{u}}$.*

For $E = [0,\dots,m-1]$ and $F = [0,\dots,n-1]$ (resp. $F = [-n+1,\dots,0]$), definition 3.6 turns into the usual definition of Toeplitz (resp. Hankel) matrices (see [2]). Quasi-Toeplitz matrices have also been studied under the name of Multilevel Toeplitz matrices (see, e.g., [40]), in the restricted special case where the sets E and F are rectangular (ie. a product of intervals). For our study of polynomial systems of equations, using the latter restricted case is not sufficient, and our more general definitions are required.

The definitions can be immediately extended to all subsets E, F of $\mathbb{Z}^n$, if we work with the Laurent polynomials.

The classes of quasi-Toeplitz and quasi-Hankel matrices can be transformed into each other by means of multiplication by the reflection matrix, having ones on its antidiagonal and zeros elsewhere.

Definition 3.7 *Let $\pi_E : L \to L$ be the projection map such that $\pi_E(\mathbf{x}^\alpha) = \mathbf{x}^\alpha$ if $\alpha \in E$ and $\pi_E(\mathbf{x}^\alpha) = 0$ otherwise. Also let $\pi_E : \mathbb{C}[[\mathbf{d}]] \to \mathbb{C}[[\mathbf{d}]]$*

denote the projection map such that $\pi_E(\mathbf{d}^\alpha) = \mathbf{d}^\alpha$ *if* $\alpha \in E$ *and* $\pi_E(\mathbf{d}^\alpha) = 0$ *otherwise.*

We can describe the quasi-Toeplitz and quasi-Hankel operators in terms of polynomial multiplication (see [26], [25]), and the next proposition reduces multiplication of an (E, F) quasi-Toeplitz (resp. quasi-Hankel) matrix by a vector $\mathbf{v} = [v_\beta] \in \mathbb{C}^F$ to (Laurent) polynomial multiplication.

Proposition 3.8 *The matrix* M *is an* (E, F) *quasi-Toeplitz (resp. an* (E, F) *quasi-Hankel) matrix, if and only if it is the matrix of the operator* $\pi_E \circ \mu_{T_M} \circ \pi_F$ *(resp.* $\pi_E \circ \mu_{H_M} \circ \pi_F$*), where for any* $p \in L$, $\mu_p : q \mapsto pq$ *is the operator of multiplication by* p *in* L.

Proof. (See [25].) We will give a proof only for an (E, F) quasi-Toeplitz matrix $M = (M_{\alpha,\beta})_{\alpha\in E,\beta\in F}$. (The proof is similar for a quasi-Hankel matrix.) The associated polynomial is $T_M(\mathbf{x}) = \sum_{\mathbf{u}\in E+F} t_\mathbf{u}\mathbf{x}^\mathbf{u}$. For any vector $\mathbf{v} = [v_\beta] \in \mathbb{C}^F$, let $v(\mathbf{x})$ denote the polynomial $\sum_{\beta\in F} v_\beta \mathbf{x}^\beta$. Then

$$
\begin{aligned}
T_M(\mathbf{x})\, v(\mathbf{x}) &= \sum_{\mathbf{u}\in E+F,\beta\in F} \mathbf{x}^{\mathbf{u}+\beta}\, t_\mathbf{u}\, v_\beta \\
&= \sum_{\alpha=\mathbf{u}+\beta\in E+2\,F} \mathbf{x}^\alpha \left(\sum_{\beta\in F} t_{\alpha-\beta}\, v_\beta \right),
\end{aligned}
$$

where we assume that $t_\mathbf{u} = 0$ if $\mathbf{u} \notin E + F$. Therefore, for $\alpha \in E$, the coefficient of $\mathbf{x}^\alpha$ equals

$$
\sum_{\beta\in F} t_{\alpha-\beta}\, v_\beta = \sum_{\beta\in F} M_{\alpha,\beta}\, v_\beta,
$$

which is precisely the coefficient α of $M\mathbf{v}$. □

Algorithm 3.9 MULTIPLICATION OF THE (E, F) QUASI-TOEPLITZ (RESP. QUASI-HANKEL) MATRIX $M = (M_{\alpha,\beta})_{\alpha\in E,\beta\in F}$ BY A VECTOR $\mathbf{v} = [v_\beta] \in \mathbb{C}^F$:

- *multiply the polynomials* $T_M = \sum_{\mathbf{u}\in E+F} t_\mathbf{u}\, \mathbf{x}^\mathbf{u}$ *(resp.* $H_M(\mathbf{d}) = \sum_{\mathbf{u}\in E-F} h_\mathbf{u}\mathbf{d}^\mathbf{u}$*) by* $v(\mathbf{x}) = \sum_{\beta\in F} v_\beta \mathbf{x}^\beta$ *(resp.* $v(\mathbf{d}^{-1}) = \sum_{\beta\in F} v_\beta \mathbf{d}^{-\beta}$*)*

- *and output the projection of the product on* $\mathbf{x}^E$ *(resp.* $\mathbf{d}^E$*).*

Definition 3.10 $C_{PolMult}(E, F)$ *denotes the number of ops required to multiply a polynomial with a support in* E *by a polynomial with a support in* F.

Clearly, algorithm 3.9 uses $C_{PolMult}(E + F, F)$, resp. $C_{PolMult}(E - F, -F)$, ops.

Proposition 3.11

a) An (E,F) quasi-Hankel (resp. an (E,F) quasi-Toeplitz) matrix M can be multiplied by a vector by using $\mathcal{O}(N\log^2(N) + N\log(d) + C_{M,N})$ ops, where $d = \deg H_M$ (resp. $\deg T_M$) , $N = \lfloor E - 2F \rceil$ (resp. $\lfloor E + 2F \rceil$), and $C_{M,N}$ denotes the cost of the evaluation of all monomials of the polynomial H_M (resp. T_M) on a fixed set of N points.

b) In particular, the ops bound becomes $\mathcal{O}(M\log(M))$ where $E + F = E_{\mathbf{c}}, F = E_{\mathbf{d}}$ and $E_{\mathbf{c}}, E_{\mathbf{d}}$ and $M = (c + d + 1)^n$ are defined as in theorem 3.3, whereas

c) the bound turns into $\mathcal{O}(T\log^2(T))$ where $E + F = E_{c,n}, F = E_{d,n}$ and $E_{c,n}, E_{d,n}$, and $T = \binom{n+c+d}{n}$ are defined as in theorem 3.4.

Proof. Reduce the problem to computing the product of the two polynomials $H_M(\mathbf{x})$ (resp. $T_M(\mathbf{x})$) and $V(\mathbf{x})$ and then apply theorems 3.1-3.4. □
Applying these results, we can bound the number of ops in algorithm 2.2 as follows

Proposition 3.12 *For any polynomial $p \in R$ with support in E, for any vector $[\Lambda(\mathbf{x}^\alpha)]_{\alpha \in E+F}$ (with $\Lambda \in \widehat{R}$), the vector $[p \star \Lambda(\mathbf{x}^\beta)]_{\beta \in F}$ can be computed in $\mathcal{O}(\lfloor E + F \rceil \log^2(\lfloor E + F \rceil))$ ops.*

Once we have a fast matrix-by-vector multiplication, a nonsingular linear system of equations can be also solved fast by means of the conjugate gradient algorithm, which is based on the following theorem ([16], sect. 10.2).

Theorem 3.13 *Let $W\,\mathbf{v} = \mathbf{w}$ be a nonsingular linear system of N equations. Then N multiplications of each of the matrices W and W^T by vectors and $\mathcal{O}(N^2)$ additional ops suffice to compute the solution $\mathbf{v}$ to this linear system.*

Note that W^T is a quasi-Toeplitz (resp. quasi-Hankel) matrix if so is W, and then both matrices can be multiplied by a vector quickly (see proposition 3.11). Therefore, in the cases of quasi-Toeplitz and quasi-Hankel matrices W, theorem 3.13, yields a fast algorithm for solving the linear system $W\,\mathbf{v} = \mathbf{w}$. We will also need the following related result.

Theorem 3.14 [28]. *Let W be an N-by-N real symmetric or Hermitian matrix. Let S be a fixed finite set of complex numbers. Then there is a randomized algorithm that selects N random parameters from the set S independently of each other (under uniform probability distribution on S) and either fails with a probability at most $\frac{(N+1)N}{2\lfloor S \rceil}$ or performs $\mathcal{O}(N)$ multiplications of the matrix W by vectors and $\mathcal{O}(N^2\log(N))$ other ops to compute the rank and the signature of W.*

Hereafter, random selection of elements of a set S as in theorem 3.14, will be called *sampling.*

Proof. To support the claimed estimate, we first tridiagonalize the matrix W by the Lanczos randomized algorithm ([2], pp.118-119), which involves an initial vector of dimension N and fails with a probability at $\frac{(N+1)N}{2\lfloor S\rceil}$ if the N coordinates of the vector have been sampled at random from the set S. The above bound on the failure probability and the cost bound of $\mathcal{O}(N)$ multiplications of the matrix W by vectors and $\mathcal{O}(N^2\log(N))$ other ops of this stage have been proved in [33]. Then, in $\mathcal{O}(N)$ ops, we compute the Sturm sequence of the N values of the determinants of all the $k \times k$ northwestern (leading principal) submatrices of W for $k = 1, \dots, N$ and obtain the numbers N_+ and N_- of positive and negative eigenvalues of W from the Sturm sequence (cf., e.g., [3]). These two numbers immediately define the rank and the signature of W. □

Combining proposition 3.11 with theorems 3.13 and 3.14 gives us the next corollary.

Corollary 3.15 *For an $N \times N$ quasi-Toeplitz or quasi-Hankel matrix W, the estimates of Theorems 3.13 and 3.14 turn into $O(N^2\log(N))$ ops if the matrix has a maximal (c,d) support where $c+d=N$. They turn into $\mathcal{O}(N^2\log^2(N))$ ops if the matrix has a total degree (c,d) support where $c+d=O(N)$ and into $\mathcal{O}((\log^2(N)+\log(d))N^2+C_{W,N})$ otherwise where d and $C_{W,N}$ are defined as in proposition 3.11 (a) for $M=W$.*

Remark 3.16 *Hereafter, we will refer to the matrices of case (b) in proposition 3.11 as the matrices with support of the maximal degree (c,d) and to the matrices of case (c) as the ones with support of the total degree (c,d). Furthermore, stating our estimates for the arithmetic complexity of computations, we will assume that the input polynomials have the maximal degree (c,d) support. That is, we will rely on theorem 3.3 and proposition 3.11 (b) and we will express the estimates in terms of the cardinality of the supports E and/or F or in terms of an upper bound D of the number of common roots of the input polynomials. The estimates can be easily extended to the other cases based on theorems 3.1 or 3.4 and proposition 3.11 (a) or (c) instead of theorem 3.3 and proposition 3.11 (b). In the latter case (theorem 3.4 and proposition 3.11 (c)), the cost estimates increase by the factors $\log(D)$, $\log(\lfloor E\rceil)$ or $\log(\lfloor F\rceil)$, respectively. In case of using theorems 3.1 and proposition 3.11 (a), the estimates are expressed in terms of the bounds $C_{PolMult}(G,H)$ or $C_{M,N}$ for appropriate sets G and H, matrix M and integer N. The latter case covers sparse input polynomials for which the respective bounds $C_{PolMult}(G,H)$ and $C_{M,N}$ are smaller than for the general (or dense) input, though they are not expressed solely in terms of the cardinality D (they also depend on the degree* *of the monomials or the cardinality of the supports of the input polynomial system).*

4 Computation of a non-degenerate linear form

In this section, we will compute a non-degenerate linear form on $\mathcal{A}$ provided that we are given a basis $(\mathbf{x}^\alpha)_{\alpha \in E}$ of $\mathcal{A}$ and the normal form of the elements on the boundary of this basis. This is the case, for instance, when we have computed a Gröbner basis of our ideal I for any monomial ordering [7] or when we apply any other normal form algorithm [24,29].

Definition 4.1

- *Let $v_i = (\delta_{i,1}, \dots, \delta_{i,n}) \in \mathbb{N}^n$, where $\delta_{i,j}$ is the Kronecker symbol.*

- *For all $A \subset \mathbb{N}^n$, $\Omega(A) = \{\alpha \in \mathbb{N}^n : \alpha \in A \text{ or } \exists i \in \{1, \dots, n\}, \alpha - v_i \in A\}$.*

- *N_α, for $\alpha \in \Omega(E)$ is the normal form of the monomial $\mathbf{x}^\alpha \mathrm{mod} I$, i.e. the canonical representative of its class modulo the ideal I. $N_\alpha = \mathbf{x}^\alpha$ if $\alpha \in E$ and*

  $$N_\alpha = \sum_{\beta \in E} n_{\alpha,\beta} \mathbf{x}^\beta,$$

  if $\alpha \in \Omega(E) - E$.

Our goal is to obtain the coefficients $\tau(\mathbf{x}^\alpha)$ for $\alpha \in E{+}E{+}E$ where $\tau \in \widehat{\mathcal{A}} = I^\perp$ is a generic linear form. We will compute them, by induction, under the following hypotheses:

Hypothesis 4.2

- *$(x^\alpha)_{\alpha \in E}$ is stable under derivation, that is, $\alpha = \alpha' + v_i \in E$ implies that $\alpha' \in E$.*

- *N_α, the normal form of $\mathbf{x}^\alpha$ is available for every $\alpha \in \Omega(E)$.*

- *The values $\tau_\alpha = \tau(x^\alpha)$ are available for all $\alpha \in E$, where τ is not degenerate $\in \widehat{\mathcal{A}} = I^\perp$.*

For the third part, we can remark that a random choice of $\tau(\mathbf{x}^\alpha)$ will imply with a high probability that τ does not degenerate. Our procedure is based on the following property:

Proposition 4.3 *For each $\alpha \in \Omega(E)$, we have $\tau_\alpha = \tau(N_\alpha) = \sum_{\beta \in E} n_{\alpha,\beta}\tau_\beta$. This value can be computed by applying $\mathcal{O}(D)$ ops, where $D = \lfloor E \rceil$. More generally, $\forall \gamma \in E$ we have the following inductive relation :*

$$\tau_{\alpha+\gamma} = \sum_{\beta \in E} n_{\alpha,\beta}\tau_{\beta+\gamma}.$$

Now assume that we have computed all the values τ_β, for $\beta \in \Omega(E)$, and let $\alpha = \alpha_0 + v_i \in \Omega(\Omega(E))$ with $\alpha_0 \in \Omega(E)$. Then

$$\tau(\mathbf{x}^\alpha) = \tau(x_i N_{\alpha_0}) = \sum_{\beta \in E} n_{\alpha_0,\beta}\tau\left(x_i \mathbf{x}^\beta\right).$$

We know all the $n_{\alpha_0,\beta}$ and all the $\tau\left(x_i\mathbf{x}^\beta\right)$, because $\beta + v_i \in \Omega(E)$. Therefore, we obtain $\tau_\alpha = \sum_{\beta \in E} n_{\alpha_0,\beta}\tau_{\beta+v_i}$ by computing a scalar product. Recursively, this leads us to the following inductive definition of the "levels" Ω_i.

Definition 4.4 *Write $\Omega_0 = E$, $\Omega_1 = \Omega(E)$ and $\Omega_i = \Omega(\Omega_{i-1}) \cap (E + E + E)$, $i = 2, 3, \ldots$, and write $h = \max\{|\alpha| : \alpha \in E\}$ so that $E + E + E = \Omega_{2h}$.*

Proposition 4.5 *For every $\alpha \in \Omega_i$, there is $\alpha' \in \mathbb{N}^n$ and $\alpha_1 \in \Omega_1 - \Omega_0$ such that $\alpha = \alpha_1 + \alpha'$ with $|\alpha'| \leq i - 1$ and for all $\beta \in E$ we have $\beta + \alpha' \in \Omega_{i-1}$.*

Proof. Assume that $i > 0$. Let $\alpha \in \Omega_i \subset E + E + E$. Then α can be decomposed as follows: $\alpha = \gamma_0 + \gamma_1 + \gamma_2$ with $\gamma_0, \gamma_1, \gamma_2 \in E$ and $|\gamma_1 + \gamma_2| = i$. As $i > 1$ there exists $\alpha' = \gamma_1 + \gamma_2 - v_j \in \mathbb{N}^n$, and because $(\mathbf{x}^\alpha)_{\alpha \in E}$ is stable by hypothesis 4.2, we have $\alpha' \in E + E$. It follows that $\alpha = \alpha_1 + \alpha'$ where $\alpha_1 = \gamma_0 + v_j \in \Omega_1$ and $|\alpha'| \leq i - 1$. Therefore, $\forall \beta \in E$, $\beta + \alpha' \in \Omega_{i-1}$, which completes the proof. □

Assume now that we have already computed all the values τ_β for $\beta \in \Omega_{i-1}$. Then, according to proposition 4.5, for any $\alpha \in \Omega_i$, we have $\alpha = \alpha_1 + \alpha'$, with $\alpha_1 \in \Omega_1$ and $|\alpha'| \leq i - 1$. Thus, if $\alpha_1 \in \Omega_1 - \Omega_0$, we have

$$\tau(\mathbf{x}^\alpha) = \tau(\mathbf{x}^{\alpha_1}\mathbf{x}^{\alpha'}) = \sum_{\beta \in E} n_{\alpha_1,\beta}\tau(\mathbf{x}^{\beta+\alpha'})$$

with $\beta + \alpha' \in \Omega_{i-1}$; otherwise if $\alpha_1 \in \Omega_0$, we have $\alpha = \alpha_1 + \alpha' \in \Omega_{i-1}$. In other words, we can compute by induction the values of τ on Ω_i from its values on Ω_{i-1}. This yields the following recursive algorithm for the computation of $\tau(\mathbf{x}^\alpha)$ with $\alpha \in E + E + E$.

Algorithm 4.6 Compute the first coefficients of the series associated with a linear form τ of $I^\perp$ *as follows:*

1. *For i from 1 to 2h do*

```
    for each $\alpha = \alpha_0 + \alpha_1 \in \Omega_i$ with $\alpha_0$ and $\alpha_1$ as in proposition 4.5 compute
    $\tau_\alpha = \sum_{\beta \in E} n_{\alpha_1, \beta} \tau_{\alpha_0 + \beta}$
    End for
2. Compute and output the polynomial $S = \sum_{\alpha \in E+E+E} \tau_\alpha \mathrm{d}^\alpha$.
```

Proposition 4.7 *The arithmetic complexity of algorithm 4.6 is $\mathcal{O}\left(3^n D^2\right)$.*

Proof. For each element $\alpha \in E+E+E$, we compute τ_α in $\mathcal{O}\left(D\right)$ arithmetic operations, and there are at most $\mathcal{O}\left(3^n D\right)$ elements in $E+E+E$, which gives us the claimed arithmetic complexity estimate. □

5 Arithmetic in the algebra $\mathcal{A}$

Our algorithms of the next sections perform computations in $\mathcal{A}$ efficiently based on the knowledge of a certain linear form on $\mathcal{A}$ (such as the one computed in the previous section), which induces a non-degenerate inner product. More precisely, we assume the following items available:

Basic Set of Items.

- *a linear form $\tau \in \widehat{\mathcal{A}} = I^\perp$, such that the bilinear form $\tau(a\,b)$ from $\mathcal{A} \times \mathcal{A}$ to $\mathbb{C}$ is non-degenerate,*
- *a monomial basis $(\mathbf{x}^\alpha)_{\alpha \in E}$ of $\mathcal{A}$,*
- *the coefficients $(\tau(\mathbf{x}^\alpha))_{\alpha \in F}$ where $F = E+E+E$.*

The number of elements in E is the dimension D of $\mathcal{A}$ over $\mathbb{C}$. We describe basic operations in the quotient ring $\mathcal{A}$, in terms of the following quasi-Hankel matrix:

Definition 5.1 *For any Λ in $\widehat{\mathcal{A}}$ and for any subset F of $\mathbb{N}^n$, let H_Λ^F denote the quasi-Hankel matrix, $\mathtt{H}_\Lambda^F = (\Lambda(\mathbf{x}^{\alpha+\beta}))_{\alpha,\beta \in F}$.*

By default we will assume dealing with the maximal degree support whenever we state our arithmetic complexity estimates (see remark 3.16).

Proposition 5.2 *The matrix $\mathtt{H}_\Lambda^F$ can be multiplied by a vector by using $\mathcal{O}(3^n \lfloor F \rceil \log(3^n \lfloor F \rceil))$ ops.*

Proof. Apply proposition 3.11 (b) to the (F, F) quasi-Hankel matrix $\mathtt{H}_\Lambda^F$ and observe that $\lfloor F+F+F \rceil = 3^n \lfloor F \rceil$. □

Combining corollary 3.15 and proposition 5.2 implies the following result:

Proposition 5.3 *Checking if the linear system $\mathtt{H}_\Lambda^F \mathbf{u} = \mathbf{v}$ has a unique solution and if so computing the solution requires $\mathcal{O}(3^n \lfloor F \rceil^2 \log(3^n \lfloor F \rceil))$ ops. The same cost estimate applies to the computation of the rank of the matrix*

H_{Λ}^{F}, *which involves randomization with* $\lceil F \rceil$ *random parameters and has failure probability of at most* $(\lfloor F \rceil + 1) \lfloor F \rfloor / (2 \lfloor S \rceil)$ *provided that the parameters have been sampled from a fixed finite set* S.

5.1 Dual basis

As τ defines a non-degenerate bilinear form, there exists a family of polynomials $(\mathbf{w}_\alpha)_{\alpha \in E}$ such that $\tau(\mathbf{x}^\alpha \, \mathbf{w}_\beta) = \delta_{\alpha,\beta}$, $\delta_{\alpha,\beta}$ being Kronecker's symbol, $\delta_{\alpha,\alpha} = 1$, $\delta_{\alpha,\beta} = 0$ if $\alpha \neq \beta$. The family $(\mathbf{w}_\alpha)_{\alpha \in E}$ is called the *dual basis* of $(\mathbf{x}^\alpha)_{\alpha \in E}$ for τ.

Proposition 5.4 (Projection formula). *For any* $p \in R$, *we have*

$$p \equiv \sum_{\alpha \in E} \tau(p \, \mathbf{w}_\alpha) \mathbf{x}^\alpha \equiv \sum_{\alpha \in E} \tau(p \, \mathbf{x}^\alpha) \mathbf{w}_\alpha. \tag{4}$$

Proof. See [6], [9]. □

Definition 5.5 *For any* $p \in \mathcal{A}$, *denote by* $[p]_{\mathbf{x}}$ *and* $[p]_{\mathbf{w}}$ *the coordinate vectors of* p *in the bases* $(\mathbf{x}^\alpha)_{\alpha \in E}$ *and* $(\mathbf{w}_\alpha)_{\alpha \in E}$, *respectively.*

Let $\mathbf{w}_\alpha = \sum_{\beta \in E} w_{\beta,\alpha} \, \mathbf{x}^\beta$, let $\mathtt{W}_\tau = (w_{\alpha,\beta})_{\alpha,\beta \in E}$ be the coefficient matrix. By the definition of the dual basis,

$$\tau(\mathbf{w}_\alpha \, \mathbf{x}^\gamma) = \sum_{\beta \in E} w_{\alpha,\beta} \, \tau(\mathbf{x}^{\beta+\gamma}) \tag{5}$$

is 1 if $\alpha = \gamma$ and 0 elsewhere. In terms of matrices, equation (5) implies that

$$\mathtt{H}_\tau \, \mathtt{W}_\tau = \mathbb{I}_D \tag{6}$$

where $\mathtt{H}_\tau = \mathtt{H}_\tau^E = (\tau(\mathbf{x}^{\beta+\gamma}))_{\beta,\gamma \in E}$. From the definition of $\mathtt{W}_\tau$ and equation (6), we deduce that

$$[p]_{\mathbf{x}} = \mathtt{W}_\tau \, [p]_{\mathbf{w}}, \;\; [p]_{\mathbf{w}} = \mathtt{H}_\tau \, [p]_{\mathbf{x}}. \tag{7}$$

The next result follows from proposition 5.3.

Proposition 5.6 *For any* $p \in \mathcal{A}$, *the coordinates* $[p]_{\mathbf{x}}$ *of* p *in the monomial basis can be computed from its coordinates* $[p]_{\mathbf{w}}$ *in the dual basis by using* $\mathcal{O}(3^n D^2 \log(3^n D))$ *ops.*

5.2 Product in $\mathcal{A}$

We apply projection formula (4) and for any $f \in R$ deduce that $f \equiv \sum_{\alpha \in E} \tau(f \, \mathbf{x}^\alpha) \, \mathbf{w}_\alpha = \sum_{\alpha \in E} f \star \tau(\mathbf{x}^\alpha) \, \mathbf{w}_\alpha$ in $\mathcal{A}$. Furthermore, by expressing the linear form $f \star \tau$ as a formal power series, we obtain $f \star \tau = \sum_{\alpha \in \mathbb{N}^n} f \star \tau(\mathbf{x}^\alpha) \, \mathbf{d}^\alpha$, so that the coefficients of $(\mathbf{d}^{\alpha})_{\alpha\in E}$ in the expansion of $f\star\tau$ are the coefficients $[f]_{\mathbf{w}}$ of f in the dual basis $(\mathbf{w}_{\alpha})_{\alpha\in E}$.

Similarly, for any $f, g \in \mathcal{A}$, the coefficients of $(\mathbf{d}^{\alpha})_{\alpha\in E}$ in $fg\star\tau$ are the coefficients $[fg]_{\mathbf{w}}$ of $f\,g$ in the dual basis $(\mathbf{w}_{\alpha})_{\alpha\in E}$. This leads to the following algorithm for computing the product in $\mathcal{A}$ as follow:

Algorithm 5.7 FOR ANY PAIR $f, g \in \langle \mathbf{x}^{\alpha}\rangle_{\alpha\in E}$, COMPUTE THE PRODUCT fg IN THE BASIS $\langle \mathbf{x}^{\alpha}\rangle_{\alpha\in E}$ OF $\mathcal{A}$ *as follows:*

1. *Compute the coefficients of* $(\mathbf{d}^{\alpha})_{\alpha\in E}$ *in the product* $f\,g\star\tau$.
2. *Obtain the coefficients* $[f\,g]_{\mathbf{w}}$ *from the first coefficients of* $fg\star\tau$.
3. *Solve in* $\mathbf{u}$ *the linear system* $[f\,g]_{\mathbf{w}} = \mathtt{H}_{\tau}\,\mathbf{u}$.

Output the vector $\mathbf{u}$*, which is the coordinate vector* $[f\,g]_{\mathbf{x}}$ *of* $f\,g$ *in the monomial basis of* $\mathcal{A}$.

Proposition 5.8 *The product* $f\,g$ *can be computed in* $\mathcal{O}(3^n D^2 \log(3^n D))$ *ops.*

Proof. $f\,g\star\tau$ is the product of polynomials with supports in $-E$ or $E+E+E$. Such a product can be computed in $\mathcal{O}(3^n D \log^2(3^n D))$ ops (see proposition 3.11 and remark 3.16 and observe that $\lfloor E+E+E\rceil = \mathcal{O}\left(3^n \lfloor E\rceil\right)$). The complexity of the third step is bounded according to proposition 5.3 (with $F = E$). □

5.3 Inversion in $\mathcal{A}$

The projection formula of proposition 5.4 implies that $f\mathbf{x}^{\alpha} = \sum_{\beta\in E} f\star\tau(\mathbf{x}^{\alpha+\beta})\,\mathbf{w}_{\beta}$, which means that $[f\,\mathbf{x}^{\alpha}]_{\mathbf{w}}$ is the coordinate vector $[f\star\tau(\mathbf{x}^{\alpha+\beta})]_{\beta\in E}$, that is, the column of the matrix $\mathtt{H}_{f\star\tau}$ indexed by α. In other words, $[f\,\mathbf{x}^{\alpha}]_{\mathbf{w}} = \mathtt{H}_{f\star\tau}\,[\mathbf{x}^{\alpha}]_{\mathbf{x}}$. By linearity, for any $g\in\mathcal{A}$, we have

$$[f\,g]_{\mathbf{w}} = \mathtt{H}_{f\star\tau}[g]_{\mathbf{x}} = \mathtt{H}_{\tau}\,[f\,g]_{\mathbf{x}},$$

according to (7). Thus, if $fg = 1$, that is, if $g = f^{-1}$, we have $\mathtt{H}_{f\star\tau}[g]_{\mathbf{x}} = \mathtt{H}_{\tau}\,[1]_{\mathbf{x}}$. This leads to the following algorithm for computing the inverses (reciprocals) in $\mathcal{A}$:

Algorithm 5.9 FOR ANY $f \in \langle \mathbf{x}^{\alpha}\rangle_{\alpha\in E}$, VERIFY WHETHER THERE EXISTS THE INVERSE (RECIPROCAL) OF $f\in\mathcal{A}$ AND IF SO COMPUTE IT.

1. *Compute* $\mathbf{v} = \mathtt{H}_{\tau}\,[1]_{\mathbf{x}}$.
2. *Solve in* $\mathbf{u}$ *the linear system* $\mathtt{H}_{f\star\tau}\mathbf{u} = \mathbf{v}$ *or output* FAILURE *if the matrix* $\mathtt{H}_{\tau}$ *is not invertible.*

Output the vector $\mathbf{u}$, *which is the coordinate vector* $[f^{-1}]_{\mathbf{x}}$ *of* f^{-1} *in the monomial basis of* $\mathcal{A}$.

By combining propositions 5.2, 5.3, and remark 3.16, we obtain

Proposition 5.10 *The inverse (reciprocal)* f^{-1} *of an element* f *of* $\mathcal{A}$ *can be computed by using* $\mathcal{O}(3^n D^2 \log(3^n D))$ *ops.*

6 Iterative methods

Our algorithms for the root approximation will essentially amount to computing non-trivial idempotents in the quotient algebra $\mathcal{A}$ by iterative processes with the subsequent simple recovery of the roots from the idempotents. The algorithms work in $\mathbb{C}^D$, and we write will $\mathbf{i} = \sqrt{-1}$. More rudimentary univariate versions of such algorithms were studied in [5]. We will use the basic operations in the quotient algebra $\mathcal{A}$ in order to devise two iterative methods, which will converge to non-trivial idempotents. We will first consider iteration associated to a slight modification of the so-called *Joukovski* map (see [17],[5]): $z \mapsto \frac{1}{2}(z + \frac{1}{z})$ and its variant $z \mapsto \frac{1}{2}(z - \frac{1}{z})$. The two attractive fixed points of this map are 1 and -1; for its variant, they turn into $\mathbf{i}$ and $-\mathbf{i}$.

Algorithm 6.1 SIGN ITERATION. *Choose* $u_0 = h \in \langle \mathbf{x}^\alpha \rangle_{\alpha \in E}$ *and recursively compute* $u_{k+1} \equiv \frac{1}{2}(u_k - \frac{1}{u_k}) \in \mathcal{A}$, $k = 0, 1, \ldots$.

By applying proposition 5.10 and remark 3.16, we obtain the following result.

Proposition 6.2 *Each iteration of algorithm 6.1 requires* $\mathcal{O}(3^n D^2 \log(3^n D))$ *ops.*

Proof. Apply proposition 5.3 and remark 3.16 to estimate the arithmetic cost of the computation of the inverse (reciprocal) of an element of $\mathcal{A}$. To yield the claimed cost bound of proposition 6.2, it remains to compute a linear combination of u_n and u_n^{-1} in $\mathcal{O}(D)$ ops, by direct operations on vectors of size D. □

Hereafter, $\Re(h)$ and $\Im(h)$ denote the real and the imaginary parts of a complex number h, respectively. Recall that we write ζ to denote the common roots $\zeta \in \mathcal{Z}(I)$ of given polynomials $f_1, \ldots, f_m$.

Remark 6.3 *In proposition 6.4 we will assume that* $J(h(\zeta)) \neq 0$ *for all* $\zeta \in \mathcal{Z}(I)$ *and in proposition 6.6 that* $|h(\zeta)|$ *is minimized for a unique root* $\zeta \in \mathcal{Z}(I)$. *These assumptions are satisfied for a generic system of polynomials or a generic polynomial* h.

Proposition 6.4 *The sequence* $(u_0, u_1, \ldots)$ *of algorithm 6.1 converges quadratically to* $\sigma = \sum_{\Im(h(\zeta))>0} \mathbf{e}_\zeta - \sum_{\Im(h(\zeta))<0} \mathbf{e}_\zeta$, *and we have*

$$||u_n - \sigma|| \leq K \times \rho^{2^n}$$

(for some constant K), where

$$\rho^{+} = max_{\Im(h(\zeta))>0,\zeta\in\mathcal{Z}(I)} \left|\frac{h(\zeta)-\mathbf{i}}{h(\zeta)+\mathbf{i}}\right|,$$

$$\rho^{-} = max_{\Im(h(\zeta))<0,\zeta\in\mathcal{Z}(I)} \left|\frac{h(\zeta)+\mathbf{i}}{h(\zeta)-\mathbf{i}}\right|,$$

$\mathbf{i} = \sqrt{-1}$, *and* $\rho = \max\{\rho^{+}, \rho^{-}\}$.

Proof. Apply the classical convergence analysis of the Joukovski map (see [17]) to the matrices of multiplication by u_n in $\mathcal{A}$, whose eigenvalues are $\{u_n(\zeta), \zeta \in \mathcal{Z}(I)\}$. □

Let

$$\mathbf{e}^{+} = \sum_{\Im(h(\zeta))>0} \mathbf{e}_{\zeta} = \frac{1}{2}(1+\sigma), \quad \mathbf{e}^{-} = \sum_{\Im(h(\zeta))\leq 0} \mathbf{e}_{\zeta} = \frac{1}{2}(1-\sigma)$$

denote the two sums of the idempotents associated to the roots $\zeta \in \mathcal{Z}$ such that $\Im(h(\zeta)) > 0$ and $\Im(h(\zeta)) < 0$, respectively.

If $h(\mathbf{x})$ is a linear function in $\mathbf{x}$, then each of the idempotents $\mathbf{e}^{+}$ and $\mathbf{e}^{-}$ is associated with all the roots lying in a fixed half-space of $\mathbb{C}^n$ defined by the inequalities $\Im(h(\zeta)) > 0$ or $\Im(h(\zeta)) < 0$. Conversely, an appropriate linear function $h(\mathbf{x})$ defines the idempotents $\mathbf{e}^{+}$ and $\mathbf{e}^{-}$ associated with any fixed half-space of $\mathbb{C}^n$. Furthermore, for any fixed polytope in $\mathbb{C}^n$ defined as the intersection of half-spaces, we may compute the family of the associated idempotents whose product will be associated with the polytope. In particular, any bounded box is the intersection of $4n$ half-spaces, and the associated idempotent can be computed in $4n$ applications of algorithm 6.1. Let us specify the case where the polytope is the almost flat unbounded box approximating the real manifold $R^n = \{\mathbf{x} : \Im(x_i) = 0, i = 1, \ldots, n\}$. In this case, the choices of $h = x_i - \epsilon$ and $h = x_i + \epsilon$ allow us to approximate the two idempotents,

$$\mathbf{e}_{i,\epsilon}^{-} = \sum_{\Im(\zeta_i)<\epsilon} \mathbf{e}_{\zeta}, \; \mathbf{e}_{i,\epsilon}^{+} = \sum_{\Im(\zeta_i)>-\epsilon} \mathbf{e}_{\zeta}.$$

Their product can be computed in $\mathcal{O}(3^n D^2 \log(3^n D))$ ops to yield $\mathbf{r}_{i,\epsilon} = \sum_{|\Im(\zeta_i)|<\epsilon} \mathbf{e}_{\zeta}$, and the product $\mathbf{r}_{\epsilon} \equiv \mathbf{r}_{1,\epsilon} \cdots \mathbf{r}_{n,\epsilon}$ can be computed in $\mathcal{O}(3^n D^2 \log(3^n D))$ ops, to yield the sum of the fundamental idempotents whose associated roots of the polynomial system are nearly real.

Algorithm 6.5 COMPUTING THE SUM OF THE FUNDAMENTAL (NEARLY REAL) IDEMPOTENTS.

```
• for i from 1 to n do
    $u_0 = x_i \pm \epsilon$; $u_1 :\equiv \frac{1}{2}(u_0 - \frac{1}{u_0})$ in $\mathcal{A}$; $k := 1$;
    while $\|u_k - u_{k-1}\| < 2^{-b}$ do { $u_{k+1} := \frac{1}{2}(u_k - \frac{1}{u_k})$; $k := k+1$ }
    Compute $\mathbf{e}_{i,e}^{\pm}$ and $\mathbf{r}_{i,\epsilon}$.
• Compute and output the product $\mathbf{r}_\epsilon \equiv \mathbf{r}_{1,\epsilon} \cdots \mathbf{r}_{n,\epsilon}$ in $\mathcal{A}$.
```

According to propositions 6.2 and 6.4 and remark 3.16, we have

Proposition 6.6 *An approximation of $\mathbf{r}_\epsilon$ (within the error bound $\epsilon = 2^{-b}$) can be computed in $\mathcal{O}(\mu\, 3^n D^2 \;\log(3^n D))$ ops, where*

$$\mu = \mu(b, \rho) = \log |b/\log(\rho)\,| \tag{8}$$

and

$$\rho = max_i\{\; max_{\Im(\zeta_i)>0, \zeta \in \mathcal{Z}(I)} |\tfrac{\zeta_i - \mathbf{i}}{\zeta_i + \mathbf{i}}|, \\ max_{\Im(\zeta_i)<0, \zeta \in \mathcal{Z}(I)} |\tfrac{\zeta_i + \mathbf{i}}{\zeta_i - \mathbf{i}}| \;\}. \tag{9}$$

The second iterative method is the quadratic power method:

Algorithm 6.7 Quadratic power iteration. *Choose $u_0 = h \in \langle \mathbf{x}^\alpha \rangle_{\alpha \in E}$ and recursively compute $u_{n+1} \equiv u_n^2 \in \mathcal{A}$, $n = 0, 1, \ldots$.*

Each step of this iteration requires at most $\mathcal{O}\left(3^n D^2 \log(3^n D)\right)$ ops, and we have the following property:

Proposition 6.8 *An approximation (within the error bound $\epsilon = 2^{-b}$) of the idempotent $\mathbf{e}_\zeta$ such that a unique simple root ζ minimizes $|h|$ on $\mathcal{Z}(I)$ can be computed in $\mathcal{O}\left(\nu 3^n D^2 \log(3^n D)\right)$ ops, where*

$$\nu = \nu\,(b, \gamma) = \log\left(b/|\log(\gamma)\,|\right), \tag{10}$$

$$\gamma = |\frac{h(\zeta)}{h(\zeta')}| \tag{11}$$

and $|h(\zeta')|$ is the second smallest value of $|h|$ over $\mathcal{Z}(I)$.

Proof. We rely on the convergence analysis of the quadratic power method applied to the matrices of multiplication by u_n in $\mathcal{A}$, whose eigenvalues are $\{u_n(\zeta), \zeta \in \mathcal{Z}(I)\}$. □

7 Counting and approximating the roots and the real roots

In this section we will apply the techniques and algorithms of the previous sections to the problems of counting and approximation of the roots of the system $\mathbf{p} = \mathbf{0}$.

In the algorithms for counting roots, we will use the randomization required to apply theorem 3.13. The resulting randomized algorithms and the computational complexity estimates for counting (excluding preprocessing stage of subsection 7.5) will apply to any zero-dimensional polynomial system.

In the approximation algorithms we do not need randomization except for the ensurance of the assumption of propositions 6.4 (cf. remark 6.3), but the estimates for the computational cost depend on the parameters ρ and γ of the two latter propositions (cf. equations 8, 11) and remain meaningful unless these parameters are extremely close to 1.

7.1 Counting the roots and the real roots

Theorem 7.1 [25]. *The number of the roots (resp. real roots) of the system* $\mathbf{p} = \mathbf{0}$ *is given by the rank (resp. the signature) of the quasi-Hankel matrix* H_{τ}^{E}.

Theorem 7.1, corollary 3.15, and remark 3.16 together imply the following result.

Corollary 7.2 *The numbers of the roots and of the real roots of the polynomial system* $\mathbf{p} = \mathbf{0}$ *can be computed by a randomized algorithm that generates* D *random parameters and in addition performs* $O(3^n D^2 \log(3^n D))$ *ops. If the random parameters are sampled from a fixed finite set* S*, then the algorithm may fail with a probability at most* $(3^n D + 1)\, 3^n D / \, (2\lfloor S \rceil)$.

7.2 Approximation of a root

Application of algorithm 6.7 in $\mathcal{A}$ yields the following theorem.

Theorem 7.3 *The idempotent corresponding to a root* ζ *that maximizes the absolute values* $|h(\zeta)|$ *of a fixed polynomial* $h(\mathbf{x})$ *can be approximated (within an error bound* $\epsilon = 2^{-b}$*) by using* $O(3^n D^2 \nu \log(3^n D))$ *ops where* ν *is defined in equations (10) and (11) of Proposition 6.8.*

The latter cost bound dominates the cost of the subsequent transition from the idempotent to a root.

Theorem 7.4 *The* n *coordinates of a simple root* ζ *can be determined from the idempotent* $\mathbf{e}_{\zeta}$ *in* $O(3^n D^2 \log(3^n D))$ *ops. This bound increases by the factor of* n *if the root is multiple.*

Proof. We compute $J\mathbf{e}_\zeta$ in $\mathcal{A}$ (where J is the Jacobian of the n equations) by algorithm 5.7. According to [25], [28], in the case of a simple root, we have

$$\mathrm{H}_\tau^E\,[J\,\mathbf{e}_\zeta]_{\mathbf{x}} = \lambda\,[\zeta^\alpha]_{\alpha\in E}, \lambda \in \mathbb{C}.$$

This vector is computed at the arithmetic cost within the complexity bound of the proposition 5.2 (cf. [28]), and this immediately gives us the coordinates of the root ζ if $\mathbf{x}^E$ contains $1, x_1, \ldots, x_n$, which is generically the case. If the root is not simple, then, according to the relation

$$x_i\,J\,\mathbf{e}_\zeta \equiv \zeta_i\,J\,\mathbf{e}_\zeta$$

(see [25], [28], [9]), we recover the coordinates of ζ, by computing $n+1$ products in $\mathcal{A}$ (by algorithm 5.7). □

7.3 Approximation of a selected root

In view of theorem 7.4, it is sufficient to approximate the idempotents associated to the roots.

Suppose that we seek a root of the system $\mathbf{p} = \mathbf{0}$ whose coordinate x_1 is the closest to a given value $u \in \mathbb{C}$. Let us assume that u is not a projection of any root of the system $\mathbf{p} = \mathbf{0}$, so that $x_1 - u$ has the inverse (reciprocal) in $\mathcal{A}$. Let $h(\mathbf{x})$ denote such an inverse (reciprocal). We have $h\,(\mathbf{x})(x_1 - u) \equiv 1$ and $h(\zeta) = \frac{1}{\zeta_1 - u}$. Therefore, a root whose coordinate x_1 is the closest to u_1 is a root for which $|h(\zeta)|$ is the largest. Consequently, iterative squaring of $h = h(\mathbf{x})$ shall converge to this root.

The polynomial h can be computed by using $O(3^n D^2 \nu \log(3^n D))$ ops for ν of equations (10) and (11) (see [28], section 3.3.4).

One may compute several roots of the polynomial system by applying the latter computation (successively or concurrently) to several initial values u.

7.4 Counting nearly real roots and the roots in a polytope

As long as we have (a close approximation to) the idempotent $\mathbf{r}$ associated with a fixed polytope, we may restrict our counting and approximation algorithms to such a polytope simply by moving from the basic nondegenerate linear form τ to the form $\mathbf{r}\star\tau$ (by using $O(3^n D^2 \log(3^n D))$ ops). Let us specify this in the case where the polytope is the nearly flat box approximating the real space $\mathbb{R}^n$ (cf. algorithm 6.5 and proposition 6.6).

Let $\mathcal{A}_\epsilon^{\mathbb{R}} = \mathbf{r}_\epsilon \mathcal{A}$ denote the subalgebra of $\mathcal{A}$ corresponding to the (nearly) real idempotents for a fixed $\epsilon = 2^{-b}$.

We may restrict our computation on $\mathcal{A}_\epsilon^{\mathbb{R}}$ by computing the linear form $\tau' = \mathbf{r}_\epsilon \star \tau$ (in $\mathcal{O}(3^n D^2 \log(D))$ ops, according to proposition 3.12), and we have the following properties:

Proposition 7.5

- *The linear form $\tau' = \mathbf{r}_\epsilon \star \tau$ defines a non-degenerate inner product on $\mathcal{A}_\epsilon^{\mathbb{R}}$.*
- *The number of nearly real roots (counted with their multiplicities) is the rank of the matrix $H^E_{\mathbf{r}_\epsilon \star \tau} = (\mathbf{r}_\epsilon \star \tau(\mathbf{x}^{\beta+\gamma}))_{\beta,\gamma} \in F$.*
- *Let E' be a subset of E such that the submatrix $H^{E'}_{\tau'}$ is of the maximal rank. Then E' is a basis of $\mathcal{A}_\epsilon$.*

Proof. See [28]. □

This leads to an algorithm for computing the rank of $H^E_{\tau'}$. Assuming (8) and (9), we deduce the following result from theorem 3.14.

Proposition 7.6 *The number of all nearly real roots can be computed by using $\mathcal{O}(\mu\, 3^n\, D^2 \log(3^n D))$ ops (for μ of (8) and (9)).*

7.5 Approximation of nearly real roots and the roots in a box.

To compute a nearly real root as well as a root lying in a fixed box in $\mathbb{C}^n$ maximizing a given function $|h|$, we may apply algorithm 6.7 in $\mathcal{A}$ (or $\mathcal{A}_\epsilon^{\mathbb{R}}$) and proposition 6.8 and obtain the following theorem:

Theorem 7.7 *A nearly real root (as well as a root lying in a fixed box) that maximizes a function $|h|$ can be computed (up to an error $\epsilon = 2^{-b}$) by using $\mathcal{O}((\mu + \nu)\ 3^n D^2 \log(3^n D))$ ops for μ and ν of equations (8)-(11).*

This process can be extended to compute the other roots via deflation. That is, we replace $\mathbf{r}_\epsilon$ by $\mathbf{r}'_\epsilon = \mathbf{r}_\epsilon - \mathbf{e}_\zeta$, compute $\tau'' = \mathbf{r}'_\epsilon \star \tau$ and apply the same iteration to compute the next (real) root, where $|h|$ takes on its second smallest value over $\mathcal{Z}(I)$. We can also restrict our computation to a fixed box by using the algorithm of subsection 7.4 to compute the sum of the idempotents corresponding to the roots lying inside the box. The complexity of each step being bounded in theorem 7.7, this leads to the following result for δ (real) roots in a given box:

Theorem 7.8 *The δ (real) roots ζ lying in a given box can be computed (up to an error $\epsilon = 2^{-b}$) by using $\mathcal{O}((\mu + \nu)\, n 3^n \delta\, D^2\ \log(D) \log(b))$ ops for μ and ν of equations (8)-(11).*

Acknowledgments

Some results of this paper were presented at the 30th Annual ACM Symposium on Theory of Computing, in May 1998. The second author was supported by NSF Grant CCR 9732206 and PSC CUNY Award 62435-0031.

References

1. W. Auzinger and H. J. Stetter, An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations, in *Proc. Intern. Conf. on Numerical Math.*, (Volume 86 of *Int. Series of Numerical Math*, pp. 12–30, Birkhäuser, 1988).
2. D. Bini and V. Y. Pan, *Polynomial and matrix computations, Vol 1 : Fundamental Algorithms*, Birkhäuser, Boston, 1994.
3. D. Bini and V. Y. Pan, Computing matrix eigenvalues and polynomial zeros where the output is real, SIAM J. on Computing **27(4)**, 1099–1115 (1998).
4. J. Canny and I. Emiris, An efficient algorithm for the sparse mixed resultant, in *Proc. Intern. Symp. on Applied Algebra, Algebraic Algorithms and Error-Corr. Codes (Puerto Rico)*, G. Cohen, T. Mora, and O. Moreno, eds. (Volume 673 of *Lect. Notes in Comp. Science*, pp. 89–104. Springer, 1993).
5. J. P. Cardinal, On two iterative methods for approximating the roots of a polynomial, in *Proc. AMS-SIAM Summer Seminar on Math. of Numerical Analysis, (Park City, Utah, 1995)*, J. Renegar, M. Shub, and S. Smale, eds. (Volume 32 of *Lectures in Applied Math.*, pp. 165–188, American Mathematical Society Press, Providence, 1996).
6. J. P. Cardinal and B. Mourrain, Algebraic approach of residues and applications, in *Proc. AMS-SIAM Summer Seminar on Math. of Numerical Analysis, (Park City, Utah, 1995)*, J. Renegar, M. Shub, and S. Smale, eds. (Volume 32 of *Lectures in Applied Math.*, pp. 189–210, American Mathematical Society Press, Providence, 1996).
7. D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, (Undergraduate Texts in Mathematics. Springer, New York, 1992).
8. D. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1998.
9. M. Elkadi and B. Mourrain, Approche effective des résidus algébriques, Rapport de Recherche 2884, INRIA, Sophia Antipolis, 1996.
10. M. Elkadi and B. Mourrain, Some applications of bezoutians in effective

algebraic geometry, Rapport de Recherche 3572, INRIA, Sophia Antipolis, 1998.

11. M. Elkadi and B. Mourrain, A new algorithm for the geometric decomposition of a variety, in *Proc. Intern. Symp. on Symbolic and Algebraic Computation*, S. Dooley, ed. (ACM Press, pp. 9–16, New York, 1999).
12. M. Elkadi and B. Mourrain, Algorithms for residues and Lojasiewicz exponents, J. of Pure and Applied Algebra **153**, 27–44 (2000).
13. I. Z. Emiris and V. Y. Pan, The structure of sparse resultant matrices. in *Proc. Intern. Symp. on Symbolic and Algebraic Computation*, (ACM Press, pp. 189–196, New York, 1997).
14. I. Z. Emiris and A. Rege, Monomial bases and polynomial system solving, in *Proc. Intern. Symp. on Symbolic and Algebraic Computation*, (ACM Press, pp. 114–122, New York, 1994).
15. J. C. Faugère, A new efficient algorithm for computing Gröbner Basis (F4), *J. of Pure and Applied Algebra* **139**, 61–88 (1999).
16. G. H. Golub and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, Maryland, 3rd edition, 1996.
17. P. Henrici, *Applied and Computational Complex Analysis*, Volume I. Wiley, 1988.
18. D. Kapur and Y. N. Lakshman, Elimination methods: an introduction, in *Symbolic and Numerical Computation for Artifitial Intellingence*, B. Donald, D. Kapur, and J. Mundy, eds. (Academic Press, pp. 45–89, New York, 1992).
19. Y. N. Lakshman and D. Lazard, On the complexity of zero-dimensional algebraic systems, in *Effective Methods in Algebraic Geometry (MEGA'90)*, (Volume 94 of *Progress in Math.*, pages 217–225, Castiglioncello Italy, Birkhäuser, 1991).
20. F. S. Macaulay, *The Algebraic Theory of Modular Systems*, Cambridge Univ. Press, 1916.
21. D. Manocha, *Systems Using Matrix Computations*, (Advances in Computational Mathematics (New Delhi), pp. 99–129, Ser. Approx. Decompos., 4, World Sci. Publishing, River Edge, NJ, 1994).
22. B. Mourrain, Isolated points, duality and residues, J. of Pure and Applied Algebra **117** & **118**, 469–493 (1996). Special issue for the *Proc. of the 4th Int. Symp. on Effective Methods in Algebraic Geometry (MEGA)*.
23. B. Mourrain, Computing isolated polynomial roots by matrix methods, J. of Symbolic Computation, Special Issue on Symbolic-Numeric Algebra for Polynomials **26(6)**, 715–738 (Dec. 1998).
24. B. Mourrain, A new criterion for normal form algorithms, in *Proc. Intern. Symp. on Applied Algebra, Algebraic Algorithms and Error-Corr. Codes*

(Puerto Rico), M. Fossorier, H. Imai, Shu Lin, and A. Poli, eds. (Volume 1719 of *LNCS*, pp. 430–443, Springer, Berlin, 1999).

25. B. Mourrain and V. Y. Pan, Multidimensional structured matrices and polynomial systems, Calcolo, Special Issue for the workshop: Structure, Algorithms and Applications **33**, 389–401 (1997).
26. B. Mourrain and V. Y. Pan, Solving special polynomial systems by using structured matrices and algebraic residues, in *Foundations of Computational Mathematics (Rio de Janeiro)*, F. Cucker and M. Shub, eds. (pp. 87–304, Springer, 1997).
27. B. Mourrain and V. Y. Pan, Asymptotic acceleration of solving multivariate polynomial systems of equations, in *Proc. 30th Annual Symp. on Theory of Computing (STOC'98)*, (ACM Press, pp. 488–496, New York, 1998).
28. B. Mourrain and V. Y. Pan, Multivariate polynomials, duality and structured matrices, J. of Complexity **16(1)**, 110–180 (2000).
29. B. Mourrain and Ph. Trébuchet, Solving projective complete intersection faster, *Proc. Intern. Symp. on Symbolic and Algebraic Computation*, (ACM Press, pp. 231–238, New York, 2000).
30. V. Y. Pan, Optimal (up to polylog factors) sequential and parallel algorithms for approximating complex polynomial zeros, in *Proceedings, 27th Annual Symp. on Theory of Computing (STOC'95)*, (ACM Press, pp. 741–750, New York, 1995).
31. V. Y. Pan, Optimal and nearly optimal algorithms for approximating complex polynomial zeros, *Computers and Math. with Applications* **31(12)**, 97–138 (1996).
32. V. Y. Pan, Solving a polynomial equation: some history and recent progress, SIAM Review **39(2)**, 187–220 (1997).
33. V. Y. Pan and Z. Chen, The complexity of matrix eigenproblem, in *Proc. 31st Annual Symposium on Theory of Computing (STOC'99)*, (ACM Press, pp. 507–516, New York, 1999).
34. J. Renegar, On the worst-case complexity of approximating zeros of polynomials, J. of Complexity **3(2)**, 90–113 (1987).
35. J. Renegar, On the worst-case arithmetic complexity of approximating zeros of system of polynomials, SIAM J. of Computing **18**, 350–370 (1989).
36. J. M. Rojas, On the Average Number of Real Roots of Certain Random Sparse Polynomial Systems, in *Proc. AMS-SIAM Summer Seminar on Math. of Numerical Analysis, (Park City, Utah, 1995)*, J. Renegar, M. Shub, and S. Smale, eds. (Volume 32 of *Lectures in Applied Math.*, American Mathematical Society Press, pp. 689–699, Providence, 1996).
37. M. Shub and S. Smale, On the complexity of Bezout's theorem I – geo-

metric aspects, J. AMS **6(2)**, 459–501 (1993).

38. H. J. Stetter, Eigenproblems are at the heart of polynomial system solving, SIGSAM Bulletin **30(4)**, 22–25 (1996).
39. H. J. Stetter, Analysis of zero clusters in multivariate polynomial systems, in *Proc. Intern. Symp. on Symbolic and Algebraic Computation,* (ACM Press, pp. 127–135, New York, 1996).
40. E. E. Tyrtyshnikov, A unifying approach to some old and new theorems on distribution and clustering, Linear Alg. and Its Applications **232**, 1–43 (1996).
41. B. L. Van der Waerden, *Modern algebra, Vol. II*, Frederick Ungar Publishing Co. New-York, 1948.

IBC-PROBLEMS RELATED TO STEVE SMALE

ERICH NOVAK
Mathematisches Institut, Universität Jena, Ernst-Abbe-Platz 4, D-07740 Jena, Germany
E-mail: novak@mathematik.uni-jena.de

HENRYK WOŹNIAKOWSKI
Department of Computer Science, Columbia University, New York, NY 10027, USA, and Institute of Applied Mathematics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland
E-mail: henryk@cs.columbia.edu

The first problem deals with topological complexity; that is, with the minimal number of comparisons which have to be performed to solve certain numerical problems. We present recent results on solving scalar nonlinear equations to within ε for different classes F of continuous functions. Depending on F and the set of permissible arithmetic operations, the topological complexity may be zero or roughly $\log\log\varepsilon^{-1}$ or even $\log\varepsilon^{-1}$.

The second problem concerns uniform versus nonuniform algorithms. In particular, we discuss the role of the error threshold ε. If ε is regarded as a fixed parameter we permit built-in constants of algorithms depending on ε, and this is the case of nonuniform algorithms. On the other hand, if ε is treated as a varying parameter and as one of the inputs, the built-in constants of algorithms are independent of ε, and this is the case of uniform algorithms. For some problems, the minimal costs of nonuniform and uniform algorithms may be quite different. We present conditions under which they are roughly the same.

We started these two projects during the Park City conference organized by Steve Smale and the discussions with Steve in Park City and later in Hong Kong, Dagstuhl and by e-mail were very helpful for us.

We end this paper by discussing tractability which is a central theme of discrete and continuous computational complexity. This problem has been also studied by Steve Smale. We want to know for which problems there exists an algorithm which computes an ε-approximation with cost bounded by a polynomial in ε^{-1} and in the input size. In contrast to many areas of complexity, we *prove* intractability or tractability of some problems using information-based arguments. We illustrate this in the worst case setting for numerical integration and discrepancy. We report on recent results that numerical integration is intractable in the L_2 norm and tractable in the L_1 norm. The latter is done jointly with S. Heinrich and G. W. Wasilkowski and follows from the fact that the inverse of the star-discrepancy depends linearly on the dimension.

1 Introduction

Information-based complexity, IBC, is part of continuous computational complexity. IBC is usually developed over abstract infinite-dimensional linear spaces such as Hilbert or Banach spaces. The applications are typically multivariate problems sometimes in hundreds or thousands of variables. Examples of such problems include multivariate integration or approximation, ordinary or partial differential equations, integral equations, optimization, and nonlinear equations. For instance, consider a partial differential equation whose boundary conditions and coefficients are specified by multivariate functions. Since functions cannot be entered into a digital computer, we may only compute a finite number of functionals, such as function values. These functionals are sometimes called oracles or subroutines. Thus only partial information about the problem is available in the computer and the original problem can be only approximately solved. Furthermore, this partial information is often contaminated with noise. Contaminated information is studied by Plaskota (1996); for simplicity we assume here that the partial information is not contaminated.

The goal of IBC is to compute an approximation of the original problem at minimal cost. The error and the cost of approximation can be defined in different settings. In this paper we concentrate on the worst case setting. The ε-complexity is defined as the minimal cost of computing an approximation with error at most ε.

We use the *real number model with an oracle.* Roughly speaking, we add an oracle (or subroutine) for the computation of function values to the BSS-model over the reals, see Blum, Shub, Smale (1989), Blum, Cucker, Shub, Smale (1998), Meer, Michaux (1997), Novak (1995), as well as Traub, Woźniakowski (1980), and Traub, Wasilkowski, Woźniakowski (1988). In this model we assume that we can exactly compute arithmetic operations over reals, comparisons of real numbers and function values by an oracle. This model is typically used for numerical and scientific computations since it is an abstraction of floating point arithmetic in fixed precision. More about this model of computation can be found in Novak, Woźniakowski (1996, 1999a, 2000a), and in Woźniakowski (1998).

Usually we are interested in the total cost of an algorithm which is given by a weighted sum of all operations and often it turns out that sharp complexity bounds are obtained by studying only the information cost. But sometimes it is interesting to count only one kind of admissible operations. Section 2 deals with topological complexity or the question of how many comparisons have to be performed to solve certain numerical problems. We present recent results on solving scalar nonlinear equations to within ε for different classes F of continuous functions. Depending on F and the set of permissible arithmetic operations, the topological complexity may be zero or roughly $\log\log\varepsilon^{-1}$ or even $\log\varepsilon^{-1}$.

In Section 3 we discuss the problem of uniform versus nonuniform algorithms. In particular, we discuss the role of the error threshold ε. If ε is regarded as a fixed parameter we may have built-in constants of algorithms depending on ε, and this is the case of nonuniform algorithms. On the other hand, if ε is treated as a varying parameter and as one of the inputs, the built-in constants of algorithms are independent of ε, and this is the case of uniform algorithms. For some problems, the minimal costs of nonuniform and uniform algorithms may be quite different. We present conditions under which they are roughly the same.

We started our research concerning topological complexity and uniform algorithms during the Park City conference organized by Steve Smale. The discussions with Steve in Park City and later in Hong Kong, Dagstuhl and by e-mail were very helpful for us.

We end this paper by discussing tractability which is a central theme of discrete and continuous computational complexity. This problem has been also studied by Steve Smale. We want to know for which problems there exists an algorithm which computes an ε-approximation with cost bounded by a polynomial in ε^{-1} and in the input size. In contrast to many areas of complexity, we *prove* intractability or tractability of some problems using information-based arguments. We illustrate this in the worst case setting for numerical integration and discrepancy. We report on recent results that numerical integration is intractable in the L_2 norm and tractable in the L_1 norm. The latter is done jointly with S. Heinrich and G. W. Wasilkowski and follows from the fact that the inverse of the star-discrepancy depends linearly on the dimension.

2 Topological Complexity

Our interest in the topological complexity is motivated by the work of Smale (1987) and Vassiliev (1992, 1996). In particular, they consider zero-finding for monic univariate polynomials of degree d with complex coefficients whose absolute values are at most one. They prove that the minimal total number of comparison nodes in the computational graph is roughly d for small ε.

We deal with zero-finding for univariate functions defined on the interval $[0,1]$, and by the topological complexity we mean the minimal depth of the computational graph. We present some results from our paper Novak, Woźniakowski (1996) where the proofs and more details can be found. We analyze different classes F of functions that are subsets of the class

$$F^* = \{f \in C[0,1] \mid f(0) < 0 \text{ and } f(1) > 0\}. \qquad (1)$$

We consider the *root error* criterion[a], which was used in Smale and Vassiliev's papers. For this criterion, x is an ε-approximation of a zero of the function f if

$$|x - x^*| \le \varepsilon \quad \text{for some } x^* \text{ such that } f(x^*) = 0.$$

We now comment on arithmetic operations. Usually the standard operations of addition, subtraction, multiplication and division are called arithmetic operations. As in many papers in this area, we understand arithmetic operations in a more general sense. Namely, by an *arithmetic operation* we mean an operation from a given set ARI. We consider various sets ARI. We always assume that

$$\mathrm{ARI}_{\min} = \{+, -, *\},$$

i.e., addition, subtraction and multiplication, is a subset of the set ARI. As we shall see, division plays a special role and the results depend on whether division belongs to the set ARI.

Since we are primarily interested in the topological complexity, we assume that the cost of each comparison is taken as unity, whereas information and arithmetic operations are free of charge. By

$$\mathrm{comp}^{\mathrm{TOP}}(F, \mathrm{ARI}, \varepsilon)$$

we denote the topological complexity, which is defined as the minimal number of comparisons necessary to compute an ε-approximation for any function from the class F by using finitely many information operations (function values), and finitely many arithmetic operations from the set ARI.

Observe that we define the topological complexity here as the depth of the computation graph which corresponds to the maximal number of comparisons used in one particular computation. The total number of comparison nodes in the computation graph can be much bigger. This latter quantity is also called topological complexity, see Smale (1987).

We will see that the topological complexity depends crucially on the class F of functions and on the choice of arithmetic operations ARI. Sometimes apparently innocent changes of F or ARI lead to completely different topological complexities.

[a] In our paper Novak, Woźniakowski (1996) we also consider the *residual* error criterion, where x is an ε-approximation of a zero of the function f if $|f(x)| \le \varepsilon$. The results for these two error criteria can be very different.

For all F and ARI studied in this paper we have

$$0 \leq \mathrm{comp}^{\mathrm{TOP}}(F, \mathrm{ARI}, \varepsilon) \leq \lceil -\log_2 \varepsilon - 1 \rceil. \tag{2}$$

The lower bound is trivial, and the upper bound follows from the bisection algorithm which is well defined for any $f \in F^*$ and only uses operations from $\mathrm{ARI}_{\mathrm{min}}$. For $\varepsilon \geq 1/2$, bisection gives $x = 1/2$ with $0 = \lceil -\log_2 \varepsilon - 1 \rceil$ comparisons. For a positive $\varepsilon < 1/2$, bisection uses $k = \lceil -\log_2 \varepsilon - 1 \rceil$ function values and comparisons. The number of additions and multiplications is $2k$.

It is known, see Kung (1976), that at least k function values must be used by any algorithm which computes an ε-approximation for all functions from F^*. Hence, bisection is optimal with respect to the number of function values. Optimality of bisection is preserved also for subsets of F^* which consist of smooth functions, see Sikorski (1985). On the other hand, bisection is *not* optimal in an average case setting (with a Brownian bridge on smooth functions as the probability measure), see Novak, Ritter, Woźniakowski (1995).

We stress that for most classes F of functions and sets ARI studied in this paper, the total complexity, i.e., when all information, arithmetic and comparison operations cost unity, is insensitive to these changes and is always of order $-\log_2 \varepsilon$. Hence, the sensitivity of the topological complexity usually does not correspond to the sensitivity of the total complexity.

We ask whether the upper bound of (2) is sharp, or equivalently, is bisection also optimal with respect to the number of comparisons? The answer depends on F and ARI. The bound is sharp even for small subsets F of F^* if we only allow arithmetic operations that are Hölder on bounded domains. More precisely, let

$$\mathrm{F}_{\mathrm{lin}} = \{ f \in F^* : \ f(x) = ax + b, \forall\, x \in [0,1], \text{ for some } a, b \ \}.$$

That is, $\mathrm{F}_{\mathrm{lin}}$ is a subclass of the class of linear functions. Observe that $f \in F^*$ implies that $f(0) = b < 0$ and $f(1) = a + b > 0$. This implies $a > 0$, however, $|a|$ can be arbitrarily small. The only zero of f is

$$x^*(f) = \frac{-b}{a} = \frac{-f(0)}{f(1) - f(0)} \in (0,1).$$

If the set ARI contains division then the exact solution may be computed without comparisons, and therefore the topological complexity for $\mathrm{F}_{\mathrm{lin}}$ is zero.

On the other hand, if division is *not* in ARI then, as we shall see, comparisons are needed and the upper bound (2) is sharp. This holds for the set $\mathrm{ARI}_{\mathrm{hol}}$ which is defined as the set of all operations that are Hölder on bounded domains. That is, the operation op belongs to $\mathrm{ARI}_{\mathrm{hol}}$ iff $\mathrm{op} : D_{\mathrm{op}} \to \mathbb{R}$, where the domain $D_{\text{op}} \subset \mathbb{R}^j$ for some j, and for any bounded set $M \subset D_{\text{op}}$ there are constants $\beta, \gamma > 0$ such that

$$|\text{op}(x) - \text{op}(y)| \leq \gamma \cdot \|x - y\|^{\beta} \qquad \forall\, x, y \in M. \tag{3}$$

The set ARI_{hol} is very large. It obviously contains ARI_{min} as well as many of the standard functions. Note that the operation $\text{op}(x,y) = x/y$ with the domain $D_{\text{op}} \subset \mathbb{R}^2$ belongs to ARI_{hol} if there exists a positive δ such that the domain D_{op} satisfies the condition

$$(x,y) \in D_{\text{op}} \quad \text{implies} \quad |y| \geq \delta.$$

Of course, the division $\text{op}(x,y) = x/y$ with the domain $D_{\text{op}} = \{(x,y) : y \neq 0\}$ does *not* belong to ARI_{hol}. Similarly, the natural logarithm $\text{op}(x) = \log(x)$ with the domain $D_{\text{op}} = \{x : x > \delta\}$ belongs to ARI_{hol} if $\delta > 0$, but not for $\delta = 0$.

Theorem 1.

$$\text{comp}^{\text{TOP}}(\text{F}_{\text{lin}}, \text{ARI}_{\text{hol}}, \varepsilon) \;=\; \lceil -\log_2 \varepsilon \,-\, 1 \rceil.$$

Observe that, contrary to the case of Smale and Vassiliev, the topological complexity depends on ε and goes to infinity as ε approaches zero.

Assume that we want to approximate $-b/a$ for pairs $a, b \in \mathbb{R}$ such that $b < 0$ and $a + b > 0$. If we allow Hölder operations, i.e., no division, then comparisons are necessary and bisection is optimal. Hence, one division can be avoided at the expense of $\lceil -\log_2 \varepsilon \,-\, 1 \rceil$ comparisons to compute an ε-approximation of $-b/a$. This result can be compared with other results in complexity theory, see e.g., Strassen (1973).

We now show that the assumptions of Theorem 1 are essential. First we replace the set ARI_{hol} by $\text{ARI}_{\text{sgn}} = \text{ARI}_{\text{min}} \cup \{\text{sgn}\}$, where $\text{sgn}(y) = 1$ for $y > 0$, $\text{sgn}(0) = 0$, and $\text{sgn}(y) = -1$ for $y < 0$. That is, we can now compute signs of the function values, $\text{sgn}(f(x))$ for $x \in [0,1]$. How many comparisons do we now need? It turns out that no comparisons are needed. Indeed, this follows from a modification of the bisection algorithm presented in Novak and Woźniakowski (1996).

Hence, the signum function makes the topological complexity zero. Obviously, the signum function is discontinuous. The use of this particular discontinuous operation allows us to eliminate comparisons.

How about continuous operations which are *not* Hölder? Once more, Theorem 1 is not true for the class F_{lin} and the set ARI of continuous operations since division is continuous. But what will happen if we enlarge the class $\mathrm{F_{lin}}$? Again, Theorem 1 does not hold for the class of polynomials of bounded degree. This follows from the result of Renegar (1987) who proved that bisection is far from being optimal for polynomials of bounded degree assuming that the four standard arithmetic operations are used, i.e., $\mathrm{ARI_{min}}$ and division.

We now show that Theorem 1 does not hold and, in fact, the topological complexity is zero for the class of strictly increasing functions,

$$\mathrm{F_{inc}} = \{f \in F^* |\ f \text{ is strictly increasing }\}$$

and for the set $\mathrm{ARI_{ext}}$ of the four standard arithmetic operations, the exponential function $\exp(x)$, $\forall\, x \in \mathbb{R}$, and the natural logarithm $\log(x)$, $\forall x > 0$,

$$\mathrm{ARI_{ext}} = \mathrm{ARI_{min}} \cup \{\,/\,,\, \exp\,,\, \log\,\}.$$

Observe that all operations in $\mathrm{ARI_{ext}}$ are continuous.

Theorem 2. *For every $\varepsilon > 0$,*

$$\mathrm{comp^{TOP}}(\mathrm{F_{inc}}, \mathrm{ARI_{ext}}, \varepsilon) = 0.$$

We present an algorithm (machine) P_ε which computes an ε-approximation and which does not use comparisons. At the computation nodes, the algorithm P_ε uses function values and operations from $\mathrm{ARI_{ext}}$. Let

$$n = \lceil 1 + 9/(4\varepsilon) \rceil.$$

The idea of the algorithm is to compute a suitable weighted mean of the equally spaced $x_i = (i-1)/(n-1)$, $i = 1, 2, \ldots, n$. The weights depend on the function values $f(x_i)$ and are chosen such that the weighted mean approximates the zero of the function f. Hence, the weights are now used instead of comparisons to localize the position of the zero. It is convenient to assume that

$$f(0)^2 + f(1)^2 = 1. \tag{4}$$

This can be done without loss of generality since, in what follows, we can replace the $f(x_i)$ by $y_i = f(x_i)/(f(x_1)^2 + f(x_n)^2)$. The algorithm $P_\varepsilon(f)$ is defined as follows:

Step 1: Compute $f(x_1), \ldots, f(x_n)$

Step 2: Compute the numbers

$$d_{jk} := (f(x_j) - f(x_k))^2 > 0$$

for all $j, k = 1, \ldots, n$ with $j \neq k$.

It is important that each d_{jk} is *strictly* positive. This holds since $f \in F_{\text{inc}}$. We will see later that Theorem 2 does not hold for the slightly larger class of nondecreasing functions.

Step 3: Compute

$$w_{ijk} := \left(\frac{1}{f(x_i)^2 + d_{jk}} \right)^{2n^4/d_{jk}}$$

for $j \neq k$, and

$$w_i := \sum_{j \neq k} w_{ijk},$$

and finally,

$$P_\varepsilon(f) := \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

All quantities are well defined since $d_{jk} > 0$ and division by 0 does not occur in the definition of w_{ijk}. Since we use ARI_{ext}, the computation of powers is possible using log and exp, i.e., by using $x^y = \exp(y \log(x))$. We stress that no comparison is used during the computation of $P_\varepsilon(f)$. One can show that $P_\varepsilon(f)$ is an ε-approximation of the zero $x^* = x^*(f)$ for f from F_{inc}, see Novak, Woźniakowski (1996) for details.

Comparing Theorem 1 and 2 we see that Theorem 1 holds for all subsets of F^* that contain the class F_{lin} of linear functions, while Theorem 2 is only proved for the class F_{inc} of strictly increasing functions. Actually, Theorem 2 is not true for the slightly larger class F_{nd} of nondecreasing functions,

$$\text{F}_{\text{nd}} = \{ f \in F^* \mid f \text{ is nondecreasing}, f(0) = -1, f(1) = 1 \}.$$

We close this section with some results of Peter Hertling. In his paper Hertling (1996) proved results about the power of the set of arithmetic operations

$$\text{ARI}_{\text{abs}} = \{ +, -, *, /, |\cdot| \}.$$

That is, he permits the four standard arithmetic operations and the absolute value. First of all, Hertling (1996) modified Theorem 2 and proved that

$$\text{comp}^{\text{TOP}}(\text{F}_{\text{inc}}, \text{ARI}_{\text{abs}}, \varepsilon) = 0.$$

Hertling also proved that the arithmetic operations ARI_{abs} are exponentially better than ARI_{hol} for all classes F between F_{nd} and F^*.

Theorem 3 (Hertling 1996). *Assume that* $\mathrm{F}_{\mathrm{nd}} \subset F \subset F^*$. *Then*

$$\mathrm{comp}^{\mathrm{TOP}}(F, \mathrm{ARI}_{\mathrm{abs}}, \varepsilon) = \left\lceil \log_2 \left(\left\lceil \log_2(\varepsilon^{-1} + 2) \right\rceil - 1 \right) \right\rceil .$$

Hence, the topological complexity is roughly $\log_2 \log_2 \varepsilon^{-1}$ which is exponentially smaller than $\log_2 \varepsilon^{-1}$. This also shows that bisection is not optimal for the set $\mathrm{ARI}_{\mathrm{abs}}$. Hertling (1996) also proved that adding more continuous operations to the set $\mathrm{ARI}_{\mathrm{abs}}$ does not help. Namely, if $\mathrm{ARI}_{\mathrm{con}}$ denotes the set of all continuous operations then

$$\mathrm{comp}^{\mathrm{TOP}}(\mathrm{F}_{\mathrm{nd}}, \mathrm{ARI}_{\mathrm{con}}, \varepsilon) = \mathrm{comp}^{\mathrm{TOP}}(\mathrm{F}_{\mathrm{nd}}, \mathrm{ARI}_{\mathrm{abs}}, \varepsilon).$$

One may argue that the most interesting set of arithmetic operations is the *standard set*

$$\mathrm{ARI}_{\mathrm{std}} = \{ +, -, *, / \}.$$

Until recently, not much was known about this set. The following results are from Hertling (2000).

Theorem 4 (Hertling 2000).

- *Bisection is optimal for* F^* *and* $\mathrm{ARI}_{\mathrm{std}}$.
- *For the class* F_{nd}*, the topological complexity for the sets* $\mathrm{ARI}_{\mathrm{abs}}$ *and* $\mathrm{ARI}_{\mathrm{std}}$ *differs at most by 1.*
- *For the class* $\mathrm{F}_{\mathrm{inc}}$ *and* $\mathrm{ARI}_{\mathrm{std}}$*, the topological complexity is 1 for all* $0 < \varepsilon < 1/2$.

It is interesting to note that the addition of the operation $x \mapsto |x|$ to the set $\mathrm{ARI}_{\mathrm{std}}$ does not change much the topological complexity for the class F_{nd} but gives an exponential improvement for the class F^*.

3 Uniform versus Nonuniform Algorithms

Different notions of computation over the reals were considered before 1989 in algebraic complexity theory and in IBC. These approaches were nonuniform, however, and only problems with a fixed input dimension or a fixed error threshold could be studied. The BSS-model, introduced in Blum, Shub, Smale (1989), is a uniform model which permits varying input dimension or error threshold and generalizes many ideas of discrete complexity theory to computation over more general rings.

Similar as in algebraic complexity theory, most work in IBC assumes the *nonuniform* model. A typical question is the following: given a class F_d of functions $f : [0,1]^d \to \mathbb{R}$ and given a positive ε, what is the complexity (minimal cost) of computing values $S_d(f)$ of a given functional $S_d : F_d \to \mathbb{R}$ up to some error ε? We will illustrate such a problem by the *integration* problem for which $S_d(f) = \int_{[0,1]^d} f(x)\,dx$, see also Section 4. Observe that usually the dimension d, the class F_d, as well as S_d and ε are *given and fixed.* Any algorithm to solve this problem may use this a priori information and the only input of such an algorithm is f from F_d.

In this section we study the "uniform versus nonuniform" alternatives for the BSS-model with an oracle, i.e., for the IBC-model of computation. An example of an uniform algorithm for the integration problem is an algorithm that works for different classes F_d and for different positive ε. Input for such an algorithm might consist of (d, ε, f), where $f \in F_d$. As in Section 2 we assume that $f \in F_d$ is given by an oracle (black box, subroutine) for function values.

In particular, we discuss the role of the error threshold ε. If ε is regarded as a fixed parameter we may have built-in constants of algorithms depending on ε, and this is the case of nonuniform algorithms. On the other hand, if ε is treated as a varying parameter and as one of the inputs, the built-in constants of algorithms are independent of ε, and this is the case of uniform algorithms. We compare the costs of uniform and nonuniform algorithms for approximate solutions of continuous problems assuming the real number model. By the complexity we mean the minimal cost of nonuniform algorithms for computing an ε-approximation. In our paper Novak, Woźniakowski (1999a) we show that, in general, anything can happen:

1. For some problems, the class of uniform algorithms that compute an ε-approximation is empty even though the class of nonuniform algorithms is non-empty and the complexity is relatively small.

2. For some problems, the cost of any uniform algorithm is arbitrarily larger than the complexity.

3. For some problems, there exist uniform algorithms with essentially the same cost as the complexity.

Our examples for problems for which the negative results (1) and (2) hold are rather artificial. Therefore we concentrate here on problems for which the positive result (3) hold. Our first result guarantees the existence of uniform algorithms if we assume the existence of *continuous* or *weakly continuous* nonuniform algorithms. To define the concept of continuity we proceed as follows.

Let $S : F \to G$ be a operator, possibly nonlinear. Here F is a given set and G is assumed to be a metric space with the metric ϱ. We assume that G is a subset of

$$\mathbb{R}^* = \bigcup_{i=1}^{\infty} \mathbb{R}^i$$

consisting of finite sequences of real numbers. We want to approximate $S(f)$ for all f from F by means of oracles $L(f)$ for some functionals L, i.e., $L : F \to \mathbb{R}$. Assume that for all positive ε, there exists a nonuniform algorithm P_ε which computes an ε-approximation. That is,

$$\sup_{f \in F} \varrho\,(S(f), P_\varepsilon(f)) \leq \varepsilon.$$

We assume that the built-in constants of the nonuniform algorithm P_ε

$$g_{1,\varepsilon}, \ldots, g_{k(\varepsilon),\varepsilon}$$

are real numbers.

By continuity, we mean that a small change of the numbers $g_{i,\varepsilon}$ does not cause much change in the computed result $P_\varepsilon(f)$. More precisely, for $\delta > 0$ we define a nonuniform algorithm $P_{\varepsilon,\delta}$ in the following way. The algorithms P_ε and $P_{\varepsilon,\delta}$ are identical except that the built-in real numbers $g_{i,\varepsilon}$ of P_ε are replaced by built-in *rational* numbers $g_{i,\varepsilon}^{\delta}$ for $P_{\varepsilon,\delta}$ such that

$$|g_{i,\varepsilon} - g_{i,\varepsilon}^{\delta}| \leq \delta, \qquad \forall\, i = 1, 2, \ldots, k(\varepsilon). \tag{5}$$

We say that the nonuniform algorithm P_ε is *continuous* if for any positive ε there exists a positive δ such that for *all* rationals $g_{i,\varepsilon}^{\delta}$ satisfying (5) we have

$$\sup_{f \in F} \varrho\,(S(f), P_{\varepsilon,\delta}(f)) \leq 2\varepsilon. \tag{6}$$

We say that the nonuniform algorithm P_ε is *weakly continuous* if for any positive ε there exists a positive δ such that for *some* rationals $g_{i,\varepsilon}^{\delta}$ satisfying (5) we have (6). We now illustrate continuous and weakly continuous algorithms by an example.

Example 1. We discuss these concepts for integration and quadrature formulas. We show that continuity and weak continuity depend on how the quadrature formulas are implemented. Consider the class

$$F = \{f \in C^1([0,1]) \mid \|f'\|_\infty \leq 1\}$$

with $S(f) = \int_0^1 f(t)\,dt$, $\varrho(a,b) = |a-b|$, and the quadrature formulas

$$P_\varepsilon(f) = Q_n(f) = \frac{1}{n}\sum_{i=1}^{n} f\left(\frac{2i-1}{2n}\right).$$

It is well known that for $n = \lceil 1/(4\varepsilon)\rceil$, the error of P_ε is at most ε.

Observe that any linear algorithm A_ε with a finite worst case error must exactly integrate the constant functions $f_c(\cdot) = c$. This means, $A_\varepsilon(f_c) = S(f_c) = c$, $\forall\, c \in \mathbb{R}$. This condition, obviously, holds for Q_n.

Assume for a moment that $\varepsilon = 1/8$ so that $n = 2$. An obvious way to compute $Q_2(f)$ is by the formula

$$Q_2(f) = g_1\, f(g_2) + g_3\, f(g_4)$$

with the four constants $g_1 = g_3 = \frac{1}{2}$ and $g_2 = \frac{1}{4}$ and $g_4 = \frac{3}{4}$. If programmed this way then Q_2 is *not* continuous. Indeed, if we use g_i^δ such that $g_1^\delta + g_3^\delta \neq 1$ then we do not integrate f_c exactly and the error is infinity. Hence (6) does not hold.

Let us now return to an arbitrary ε. We have to guarantee that the sum of the weights is 1. Clearly, for *some* rational approximations g_i^δ we have $g_1^\delta + g_3^\delta + \cdots + g_{2n-1}^\delta = 1$. In fact we can now even take $g_{2i-1}^\delta = 1/n$. So P_ε is *weakly* continuous.

To obtain continuity we can eliminate one of the constants g_i. For instance, we take $g_{2i-1} = 1/n$ for $i = 2, 3, \ldots, n$ and $g_{2i} = (2i-1)/(2n)$ for $i = 1, 2, \ldots, n$. Then we compute $g_1 = 1 - g_3 - g_5 - \cdots - g_{2n-1}$. A small error in the constants $g_2, g_3, \ldots, g_{2n}$ now leads to a small error in the result, which yields continuity of P_ε.

To illustrate further this point, let us consider integration for the class

$$F = F_k = \{f \in C^k[0,1] \mid \|f^{(k)}\|_\infty \leq 1\}.$$

Then

$$Q_n(f) = \sum_{i=1}^{n} g_i f(t_i)$$

has finite worst case error iff Q_n is exact for all polynomials of degree less than k. Furthermore, for $n = \Theta(\varepsilon^{-1/k})$ there exist $g_i = g_{i,\varepsilon}$ and $t_i = t_{i,\varepsilon}$ for which $P_\varepsilon = Q_n$ has error at most ε.

Hence, for distinct sample points t_i, the constants g_i must satisfy the linear equations

$$Q_n(t^j) = \sum_{i=1}^{n} g_i t_i^j = S(t^j) = \frac{1}{j+1}, \qquad j = 0, 1, \ldots, k-1.$$

To obtain continuity of P_ε, we have to guarantee that the perturbed knots $t^\delta_{i,\varepsilon}$ and weights $g^\delta_{i,\varepsilon}$ satisfy the same system of linear equations, i.e.,

$$\sum_{i=1}^{n} g^\delta_{i,\varepsilon}\,(t^\delta_{i,\varepsilon})^j = \frac{1}{j+1}, \qquad j = 0, 1, \dots, k-1. \tag{7}$$

Hence, if all $g_{i,\varepsilon}$ and $t_{i,\varepsilon}$ are regarded as built-in constants of P_ε then P_ε is *not* continuous since not all small rational perturbations $g^\delta_{i,\varepsilon}$ and $t^\delta_{i,\varepsilon}$ satisfy (7). On the other hand, P_ε *is* weakly continuous since some small rational perturbations $g^\delta_{i,\varepsilon}$ and $t^\delta_{i,\varepsilon}$ satisfy (7). As before, we can obtain continuity of P_ε by eliminating, say, the first k constants $g_1, g_2, \dots, g_k$ and compute them as the solution of the linear equations

$$\sum_{i=1}^{k} g_i (t^\delta_{i,\varepsilon})^j = \frac{1}{j+1} - \sum_{i=k+1}^{n} g^\delta_{i,\varepsilon}\,(t^\delta_{i,\varepsilon})^j, \qquad j = 0, 1, \dots, k-1.$$

We stress that such elimination of built-in constants is *not* always needed. For instance, if F_k is replaced by the smaller class

$$\widetilde{F}_k = \{f \in C^k[0,1] \mid \|f^{(r)}\|_\infty \le 1 \text{ for } r = 0, \dots, k\}$$

then we can implement quadrature formulas in the obvious way and still get continuous algorithms.

We are ready to prove that weak continuity implies the existence of uniform algorithms. This will be shown by coding and a magic constant.

Theorem 5. *Assume that for each positive ε we can approximate S to within ε by weakly continuous nonuniform algorithms P_ε. Then there exists a uniform algorithm which computes an ε-approximation of S.*

To obtain a uniform algorithm, we consider the sequence of weakly continuous nonuniform algorithms $\widetilde{P}_{2^{-k}}$ with error $2 \cdot 2^{-k}$. Since each $\widetilde{P}_{2^{-k}}$ needs only finitely many rational built-in numbers, we can code these numbers as well as the whole program $\widetilde{P}_{2^{-k}}$ into a single (natural or rational) number G_k. As in the classical theory of computable functions, this number G_k can be called a *Gödel number.* All these Gödel numbers can be coded into one (magic) real number G and a *universal BSS-machine* can run (simulate) a computation $\widetilde{P}_{2^{-k}}(f)$ on an input G, f and ε with $k = \lceil \log_2 2/\varepsilon \rceil$. See Blum, Shub, Smale (1989) for the construction of a universal machine over the reals.

Theorem 5 does not address the cost of uniform algorithms. We now analyze the cost of uniform algorithms in terms of the minimal cost of nonuniform algorithms. This will be done for linear functionals. That is, $S : F_1 \to \mathbb{R}$ is a linear functional and F_1 is a linear normed space of real functions. We want to approximate $S(f)$ for all f from a given subset F of F_1. We assume that F is convex and balanced. By oracles, we now mean function values.

Let the (nonuniform) complexity be

$$\text{comp}(\varepsilon) = O\left(\varepsilon^{-1/q} \cdot (-\log \varepsilon)^{\alpha}\right), \quad 0 < \varepsilon < 1/2 \tag{8}$$

for some positive q and nonnegative α.

It is known that the complexity in the nonuniform case is obtained by *linear* algorithms using nonadaptive function values, see Traub, Wasilkowski, Woźniakowski (1988). That is, there exists an integer $n = n(\varepsilon) = O(\varepsilon^{-1/q}(-\log \varepsilon)^{\alpha})$ and a linear P_ε such that

$$P_\varepsilon(f) = \sum_{i=1}^{n} g_{2i-1,\varepsilon}\, f(g_{2i,\varepsilon}) \tag{9}$$

has worst case error $\sup_{f \in F} |S(f) - P_\varepsilon(f)| \leq \varepsilon$.

We assume that the built-in constants $g_{i,\varepsilon}$ of P_ε are uniformly bounded,

$$|g_{i,\varepsilon}| \leq M, \qquad \forall i = 1, 2, \ldots, 2n(\varepsilon),\ \forall \varepsilon > 0, \tag{10}$$

for some $M \geq 2$.

We assume that P_ε is weakly continuous. That is, for a positive ε there exists a positive δ such that the linear algorithm

$$P_{\varepsilon,\delta}(f) = \sum_{i=1}^{n} g^{\delta}_{2i-1,\varepsilon}\, f(g^{\delta}_{2i,\varepsilon})$$

with some rational $g^{\delta}_{i,\varepsilon}$ satisfying $|g_{i,\varepsilon} - g^{\delta}_{i,\varepsilon}| \leq \delta$ has error at most 2ε,

$$\sup_{f \in F} |S(f) - P_{\varepsilon,\delta}(f)| \leq 2\varepsilon. \tag{11}$$

Without loss of generality we may assume that the perturbed constants $g^{\delta}_{i,\varepsilon}$ are bounded by M, i.e., $|g^{\delta}_{i,\varepsilon}| \leq M$. We additionally assume that

$$O(\log M + \log n) \tag{12}$$

bits are sufficient to represent each $g^{\delta}_{i,\varepsilon}$ exactly.

We add in passing that for the classes F_k discussed above it is relatively easy to check that (12) holds.

Theorem 6. *Under the above assumptions (8-12) there exists a uniform algorithm P which computes an ε-approximation to S with cost*

$$\operatorname{cost}(\varepsilon) = O\left(\varepsilon^{-1/q} \cdot (-\log \varepsilon)^{\alpha+1}\right).$$

That is, the cost of P is at most $\log 1/\varepsilon$ *larger than the (nonuniform) complexity.*

We do not claim that the positive results of Theorem 5 and Theorem 6 (both are from Novak, Woźniakowski (1999a)) are always practical. All these results are based on some kind of coding and to use such uniform algorithms one has, in general, to find out and store at least one *magic* real number. This magic number is needed *with arbitrary precision*, i.e., exactly. Such an assumption is not realistic in practical computation where usually floating point arithmetic is used, and all numbers are rounded and only their approximations are used. This suggests that one should also study algorithms where only integer or rational numbers are allowed as built-in constants, see Hemmerling (1998).

To address the practical side, we note that the difference between uniform and nonuniform algorithms basically disappears if we do not let ε go to zero, and instead consider ε in the interval $[\varepsilon_{\text{low}}, \varepsilon_{\text{upp}}]$ with, say, $\varepsilon_{\text{low}} = 10^{-8}$ and $\varepsilon_{\text{upp}} = 10^{-2}$. Then no magic numbers are needed to construct a uniform algorithm which works well for all $\varepsilon \in [\varepsilon_{\text{low}}, \varepsilon_{\text{upp}}]$, see Novak, Woźniakowski (1999a) for more details.

4 When are Integration and Discrepancy Tractable?

One of the most important problems in computer science and mathematics concerns the tractability of problems. Usually this problem, and the famous $P = NP$ problem, is posed for discrete decision problems and the Turing machine. The same problem was generalized to the real number model in Blum, Shub, Smale (1989).

One can also study the class of tractable problems in the case of partial information, i.e., in the IBC model, see Woźniakowski (1994a, 1994b). The basic question is the following:

For which problems does there exist an algorithm which computes an ε-approximation with cost bounded by a polynomial in ε^{-1} and the input size?

By "input size" we mean here the number d of variables of a function. Tractability means that there is an algorithm that approximates the solution with error ε using $n = n(\varepsilon, d)$ samples of the function, where $n(\varepsilon, d)$ is polynomially bounded in ε^{-1} and d. The number $n(\varepsilon, d)$ of samples is usually directly related to the cost of an algorithm, and therefore tractability means that we can solve the problem to within ε in cost polynomially dependent on ε^{-1} and d.

In this section we present some recent results for numerical integration in the worst case setting, see Novak, Woźniakowski (2000b) for more details and a survey of recent results. For some classes of functions integration is related to various notions of *discrepancy*, such as the L_2-discrepancy and the $*$-discrepancy. Discrepancy is widely used and studied in many areas of mathematics.

Let F_d be a normed space of integrable functions $f : D_d \to \mathbb{R}$ where $D_d = [0,1]^d$. For $f \in F_d$ we want to approximate the multivariate integral

$$I_d(f) = \int_{D_d} f(t)\, dt. \tag{13}$$

We approximate $I_d(f)$ by algorithms that use finitely many function values. We consider a number of classes of algorithms:

- The class QMC of *quasi-Monte Carlo* algorithms, which is widely used for financial problems for which d is often very large, say in the hundreds or thousands. QMC algorithms are of the form

  $$Q_{n,d}(f) = \frac{1}{n} \sum_{i=1}^{n} f(t_i). \tag{14}$$

  The sample points t_i are deterministic, belong to D_d, and may depend on n and d as well as on the space F_d. The sample points t_i are *nonadaptive.* That is, they are given a priori and do not depend on the integrand f.

- The class POS of algorithms for which the weights $1/n$ of QMC algorithms are replaced by *positive or non-negative* weights a_i. The POS algorithms are of the form

  $$Q_{n,d}(f) = \sum_{i=1}^{n} a_i f(t_i), \qquad a_i \geq 0. \tag{15}$$

  The coefficients a_i and the sample points t_i are deterministic and may depend on n, d and F_d.

- The class LIN is the class of *linear* algorithms of the form (15) with arbitrary real weights a_i. In this case, some a_i's may be negative. The sample points t_i for this class satisfy the same conditions as for the classes QMC and LIN, i.e., they are nonadaptive.

As long as the class F_d is convex and symmetric, it is not necessary to study more general algorithms, see Traub, Wasilkowski, Woźniakowski (1988). That is, the worst case error is minimized by linear algorithms using non-adaptive sample points. This important result is due to Bakhvalov and Smolyak.

We are ready to define the error of an algorithm $Q_{n,d}$ for any of these classes. In this paper we restrict ourselves to the worst case error, which is defined by the worst case performance of $Q_{n,d}$ over the unit ball of F_d,

$$e(Q_{n,d}) = \sup_{f \in F_d,\, \|f\|_d \le 1} \left| I_d(f) - Q_{n,d}(f) \right|. \tag{16}$$

For $n = 0$ we do not sample the function and we[b] set $Q_{0,d} = 0$. Then

$$e(Q_{0,d}) = \sup_{f \in F_d,\, \|f\|_d \le 1} |I_d(f)| = \|I_d\|$$

is the *initial error.* This is the a priori error in multivariate integration without sampling the function. We call $e(Q_{n,d})$ the *(absolute) error* of $Q_{n,d}$ and $e(Q_{n,d})/e(Q_{0,d})$ its *normalized error.*

For each class of algorithms we want to find $Q_{n,d}$ having minimal error. Let A be one of the class QMC, POS or LIN, and let

$$e(n, F_d, \mathrm{A}) = \inf \{ e(Q_{n,d}) \,:\, Q_{n,d} \in \mathrm{A} \} \tag{17}$$

be the minimal error of algorithms from the class A when we use n function values. Clearly,

$$e(0, F_d, \mathrm{A}) = e(Q_{0,d}) = \|I_d\|, \quad \mathrm{A} \in \{\mathrm{QMC},\ \mathrm{POS},\ \mathrm{LIN}\}.$$

Since QMC $\subset$ POS $\subset$ LIN we also have

$$e(n, F_d, \mathrm{LIN}) \le e(n, F_d, \mathrm{POS}) \le e(n, F_d, \mathrm{QMC}).$$

We now formally define tractability for the class A$\in$ {QMC, POS, LIN}. We would like to reduce the initial error by a factor ε, where $\varepsilon \in (0,1)$. We are looking for the smallest $n = n(\varepsilon, F_d, A)$ for which there exists an algorithm from the class A such that $e(Q_{n,d}) \le \varepsilon\, e(Q_{0,d})$. That is,

$$n(\varepsilon, F_d, A) = \min \{ n \,:\, e(n, F_d, A) \le \varepsilon\, e(Q_{0,d}) \}. \tag{18}$$

[b] For $n = 0$ we could take $Q_{0,d} = c$ for some number c. It is easy to see that $c = 0$ minimizes the error of $Q_{0,d}$.

We say that integration for a sequence $\{F_d\}$ of spaces is *tractable* (with respect to the normalized error) in the class A iff there exist nonnegative C, q and p such that

$$n(\varepsilon, F_d, A) \le C\, d^q\, \varepsilon^{-p} \qquad \forall\, d = 1, 2, \dots, \forall\, \varepsilon \in (0,1). \tag{19}$$

Tractability means that we can reduce the initial error by a factor ε by using a number of function values which is polynomial in d and ε^{-1}. If $q = 0$ in the bound above we say that the problem is *strongly* tractable, and the infimum of p satisfying the bound above is called the *strong exponent.*

We stress that the minimal number $n(\varepsilon, F_d, \mathrm{LIN})$ of function samples is directly related to the complexity, which is the minimal cost of computing an approximation with error ε.

Discrepancy is a quantitative measure of the lack of uniformity of the points in the d-dimensional unit cube. Today we have various notions of discrepancy, and there are literally thousands of papers studying different aspects of discrepancy. Research on discrepancy is very intensive, and the reader is referred to the recent books Drmota, Tichy (1997), Matoušek (1999), Niederreiter (1992), and Tezuka (1995).

We study the classical L_2-discrepancy, which is sometimes also called L_2-star discrepancy or L_2-star discrepancy with boundary condition. We first recall the definition of the L_2-discrepancy and then present some bounds. Then we discuss the L_p-star discrepancy for finite p and the $*$-discrepancy, where $p = \infty$.

Let $x = [x_1, \dots, x_d] \in [0,1]^d$. By the box $[0, x)$ we mean the set $[0, x_1) \times \cdots \times [0, x_d)$ whose (Lebesgue) volume is clearly $x_1 \cdots x_d$. For given points $t_1, \dots, t_n \in [0,1]^d$, we approximate the volume of $[0, x)$ by the fraction of the points t_i which are in the box $[0, x)$. The error of such an approximation is

$$x_1 \cdots x_d - \frac{1}{n} \sum_{i=1}^{n} 1_{[0,x)}(t_i),$$

where $1_{[0,x)}(t_i)$ is the indicator (characteristic) function, which is equal to 1 if $t_i \in [0, x)$, and to 0 otherwise.

Observe that we use equal coefficients n^{-1} in the previous approximation scheme. As we shall see, this corresponds to QMC algorithms for integration. We generalize this approach by allowing arbitrary coefficients a_i instead of n^{-1}. That is, we approximate the volume of $[0, x)$ by the weighted sum

$$\mathrm{disc}(x) := x_1 \cdots x_d - \sum_{i=1}^{n} a_i 1_{[0,x)}(t_i). \tag{20}$$

The L_2-discrepancy of points $t_1, \dots, t_n$ and coefficients $a_1, \dots, a_n$ is just the L_2-norm of the error function (20),

$$\mathrm{disc}_2(\{t_i\}, \{a_i\}) = \left(\int_{[0,1]^d} \left(x_1 \cdots x_d - \sum_{i=1}^{n} a_i 1_{[0,x)}(t_i) \right)^2 dx \right)^{1/2} . \tag{21}$$

By direct integration, we have the explicit formula

$$\mathrm{disc}_2^2(\{t_i\}, \{a_i\}) = \tag{22}$$
$$\frac{1}{3^d} - \frac{1}{2^{d-1}} \sum_{i=1}^{n} a_i \prod_{k=1}^{d} (1 - t_{i,k}^2) + \sum_{i,j=1}^{n} a_i a_j \prod_{k=1}^{d} (1 - \max(t_{i,k}, t_{j,k})) ,$$

for the L_2-discrepancy, where $t_i = [t_{i,1}, \dots, t_{i,d}]$.

The major problem of L_2-discrepancy is to find points $t_1, \dots, t_n$ and coefficients $a_1, \dots, a_n$ for which $\mathrm{disc}_2(\{t_i\}, \{a_i\})$ is minimized. Let

$$\overline{\mathrm{disc}}_2(n, d) = \inf_{t_1, \dots, t_n} \mathrm{disc}_2(\{t_i\}, \{n^{-1}\})$$

and

$$\mathrm{disc}_2(n, d) = \inf_{t_1, \dots, t_n, \, a_1, \dots, a_n} \mathrm{disc}_2(\{t_i\}, \{a_i\})$$

denote the minimal L_2-discrepancy when we use n points in dimension d. For the minimal L_2-discrepancy $\overline{\mathrm{disc}}_2(n, d)$ we choose optimal t_i for coefficients $a_i = n^{-1}$ whereas for $\mathrm{disc}_2(n, d)$ we also choose optimal a_i.

Observe that for $n = 0$ we do not use the points t_i or the coefficients a_i, and obtain the initial L_2-discrepancy

$$\overline{\mathrm{disc}}_2(0, d) = \mathrm{disc}_2(0, d) = \left(\int_{[0,1]^d} x_1^2 \cdots x_d^2 \, dx \right)^{1/2} = 3^{-d/2}. \tag{23}$$

Hence, the initial L_2-discrepancy is exponentially small in d.

We now discuss bounds on the normalized L_2-discrepancy. By the normalized L_2-discrepancy we mean $\mathrm{disc}_2(\{t_i\}, \{a_i\}) / \mathrm{disc}_2(0, d)$. That is, we normalize by the initial value of the L_2-discrepancy, which is $3^{-d/2}$. As we shall see, this case is directly related to tractability of integration for some spaces F_d when we reduce the initial error by a factor of ε. Similarly to (18), we define

$$\overline{n}(\varepsilon, d) = \min\{ n : \overline{\mathrm{disc}}_2(n, d) \le \varepsilon \, \mathrm{disc}_2(0, d) \}, \tag{24}$$
$$n(\varepsilon, d) = \min\{ n : \mathrm{disc}_2(n, d) \le \varepsilon \, \mathrm{disc}_2(0, d) \} \tag{25}$$

and ask whether $\overline{n}(\varepsilon, d)$ and $n(\varepsilon, d)$ are polynomial in ε^{-1} and d. Recent results show that the normalized L_2-discrepancy is intractable.

Theorem 7.

$$\overline{n}(\varepsilon, d) \geq (9/8)^d (1 - \varepsilon^2). \tag{26}$$

The bound (26) is also valid if $a_i \geq 0$, which corresponds to the assumption defining the class POS. For arbitrary a_i,

$$n(\varepsilon, d) \geq 1.0628^d (1 + o(1)) \qquad \textit{as } d \to \infty, \tag{27}$$

the lower bound holding for any fixed $\varepsilon < 1$.

The bound (26) was proven in Woźniakowski (1999), see also Sloan, Woźniakowski (1998). The bound (27) is from Novak, Woźniakowski (1999b).

The L_2-discrepancy is related to uniform integration for Sobolev spaces and the Wiener sheet measure. Consider the classical Sobolev space $W_2^1([0,1])$ of absolutely continuous univariate functions $f : [0,1] \to \mathbb{R}$ for which $f' \in L_2([0,1])$ and $f(0) = 0$. The space $W_2^1([0,1])$ is equipped with the inner product

$$\langle f, g \rangle = \int_0^1 f'(x) g'(x)\, dx.$$

The Sobolev space $W_2^1([0,1])$ is a reproducing kernel Hilbert space with the reproducing kernel

$$K_1(x,t) = \min(x,t).$$

Take now F_d as the tensor product of $W_2^1([0,1])$,

$$F_d = W_2^{1,\dots,1}([0,1]^d) := W_2^1([0,1]) \otimes \cdots \otimes W_2^1([0,1]), \quad (d \,\text{times}),$$

i.e., the completion of the algebraic tensor product. Then F_d is the Sobolev space of multivariate functions $f : [0,1]^d \to \mathbb{R}$ that are differentiable once with respect to each variable, and for which $f(x) = 0$ if at least one component of x is zero. It is a reproducing kernel Hilbert space with the reproducing kernel

$$K_d(x,t) = \prod_{j=1}^{d} \min(x_j, t_j). \tag{28}$$

Let $Q_{n,d}(f) = \sum_{i=1}^{n} a_i f(t_i)$ be an algorithm for approximating $I_d(f)$. Consider now the worst case error $e(Q_{n,d})$ of $Q_{n,d}$. It is known that

$$e^2(Q_{n,d}) = \frac{1}{3^d} - 2 \sum_{i=1}^{n} a_i \prod_{k=1}^{d} (t_{i,k} - t_{i,k}^2/2) + \sum_{i,j=1}^{n} a_i a_j \prod_{k=1}^{d} \min(t_{i,k}, t_{j,k}).$$

Since for any numbers a, b we have

$$\min(a,b) = 1 - \max(1-a, 1-b) \quad \text{and} \quad a - a^2/2 = \left(1-(1-a)^2\right)/2,$$

comparing the last formula with the L_2-discrepancy formula (20), we immediately obtain

$$e(Q_{n,d}, F_d) = \operatorname{disc}_2(\{1-t_i\}, \{a_i\}),$$

where $1 - t_i = [1-t_{i,1}, \ldots, 1-t_{i,d}]$.

From the normalized bounds on the L_2-discrepancy we conclude that integration for $F_d = W_2^{1,\ldots,1}([0,1]^d)$ is *intractable* for the normalized error, i.e., for reduction of the initial error.

Recall that the L_p-star discrepancy of points $t_1, \ldots, t_n \in [0,1]^d$ is defined by

$$\operatorname{disc}_p^*(t_1,\ldots,t_n) = \left(\int_{[0,1]^d} \left| x_1 \cdots x_d - \frac{1}{n}\sum_{i=1}^{n} 1_{[0,x)}(t_i)\right|^p dx\right)^{1/p}, \tag{29}$$

for $1 \le p < \infty$, and

$$\operatorname{disc}_\infty^*(t_1,\ldots,t_n) = \sup_{x\in[0,1]^d} \left| x_1 \cdots x_d - \frac{1}{n}\sum_{i=1}^{n} 1_{[0,x)}(t_i)\right| \tag{30}$$

for $p = \infty$. It is customary to denote the L_∞-star discrepancy as the $*$-discrepancy. Let

$$\operatorname{disc}_p^*(n,d) = \inf_{t_1,\ldots,t_n} \operatorname{disc}_p^*(t_1,\ldots,t_n)$$

denote the minimal L_p-discrepancy for n points. Note that for $n = 0$ we get the initial discrepancy $\operatorname{disc}_p^*(0,d) = (1/(p+1))^{d/p}$. This shows that for $p < \infty$, the initial discrepancy in the L_p-norm goes exponentially fast to zero as d approaches infinity. For $p = \infty$, we have $\operatorname{disc}_\infty^*(0,d) = 1$ and the initial discrepancy is properly normalized. Finally, let

$$n_p^*(\varepsilon, d) = \min\{\, n : \operatorname{disc}_p^*(n,d) \le \varepsilon \,\}.$$

The usual bounds on the L_p-star discrepancy are for fixed dimension d and large n. The asymptotic behavior of $\operatorname{disc}_p^*(n,d)$ with respect to n is known, see once more Drmota, Tichy (1997), Matoušek (1999), and Niederreiter (1992). This yields to

$$\begin{aligned} n_p^*(\varepsilon,d) &= \Theta\left(\varepsilon^{-1}(\log 1/\varepsilon)^{(d-1)/2}\right) && \text{for } p \in (1,\infty), \\ n_p^*(\varepsilon,d) &= O\left(\varepsilon^{-1}(\log 1/\varepsilon)^{(d-1)a_p}\right) && \text{for } p \in \{1,\infty\}, \end{aligned}$$

where $a_1 = 1/2$ and $a_\infty = 1$, as ε tends to zero. Here O- and Θ-factors depend on d.

The question of dependence on d for $p = \infty$ was raised by Larcher who asked whether there exists an $a > 1$ such that $\text{disc}^*_\infty(\lceil a^d \rceil, d)$ tends to 1 as d goes to infinity, and also asked whether, in particular, $\text{disc}^*_\infty(2^d, d)$ goes to 1 as d goes to infinity. Based on the results for $p = 2$ one might have been inclined to believe that the answer to at least one of these questions is affirmative.

It was surprising for us that this is *not* the case and that a positive result holds. In Heinrich, Novak, Wasilkowski, Woźniakowski (1999), it is shown that $n^*_\infty(\varepsilon, d)$ depends only *polynomially* on d and n^{-1}.

Theorem 8. *There exists a positive number C such that*

$$n^*_\infty(\varepsilon, d) \leq C\, d\, \varepsilon^{-2} \qquad \forall n, d = 1, 2, \ldots . \tag{31}$$

Hence, the inverse of the $*$-discrepancy depends at most linearly on d. This dependence on d cannot be improved. The proof of (31) follows from deep results from the theory of empirical processes. The proof is non-constructive, and we do not know for which points the bound (31) holds.

Acknowledgment

We wish to thank Joseph Traub for valuable comments.

References

1. L. Blum, M. Shub, and S. Smale, On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines.=, Bull. of the AMS **21**, 1–46 (1989).
2. L. Blum, F. Cucker, M. Shub, and S. Smale, *Complexity and Real Computation*, (Springer, New York, 1998).
3. M. Drmota and R.F. Tichy, *Sequences, Discrepancies and Applications*, (Lecture Notes in Mathematics, **1651** Springer, 1997).
4. S. Heinrich, E. Novak, G.W. Wasilkowski, and H. Woźniakowski, The inverse of the star-discrepancy depends linearly on the dimension, *ACTA Arithmetica* (2000), to appear.
5. A. Hemmerling, Computability of string functions over algebraic structures, Math. Logic Quarterly **44**, 1–44 (1998).
6. P. Hertling, Topological complexity with continuous operations, J. Complexity **12**, 315–338 (1996).

7. P. Hertling, Topological complexity of zero finding with algebraic operations, (2000), submitted.
8. H.T. Kung, The complexity of obtaining starting points for solving operator equations by Newton's method, in *Analytical Computational Complexity* J.F. Traub, ed. (Academic Press, New York, pp. 35-75, 1976).
9. J. Matoušek, *Geometric Discrepancy*, (Springer-Verlag, Berlin, 1999).
10. K. Meer and C. Michaux, A survey on real structural complexity theory, Bull. Belg. Math. Soc. **4**, 113–148 (1997).
11. H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, (SIAM, Philadelphia, 1992).
12. E. Novak, The real number model in numerical analysis, J. Complexity **11**, 57–73 (1995).
13. E. Novak, K. Ritter, and H. Woźniakowski, Average case optimality of a hybrid secant-bisection method, Math. Computation **64**, 1517–1539 (1995).
14. E. Novak and H. Woźniakowski, Topological complexity of zero finding, J. Complexity **12**, 380–400 (1996).
15. E. Novak and H. Woźniakowski, On the cost of uniform and nonuniform algorithms, Theoretical Computer Science **219**, 301–318 (1999a).
16. E. Novak and H. Woźniakowski, Intractability results for integration and discrepancy, *J. Complexity*, (1999b), to appear.
17. E. Novak and H. Woźniakowski, Complexity of linear problems with a fixed output basis, J. Complexity **16**, 333–362 (2000a).
18. E. Novak and H. Woźniakowski, When are integration and discrepancy tractable? To appear in the *FOCM Proceedings*, (Oxford, 1999) (2000b).
19. L. Plaskota, *Noisy Information and Computational Complexity*, (Cambridge Univ. Press, Cambrigde, 1996).
20. J. Renegar, On the worst-case arithmetic complexity of approximating zeros of polynomials, J. Complexity **3**, 90–113 (1987).
21. K. Sikorski, Optimal solution of nonlinear equations, J. Complexity **1**, 197-209 (1985).
22. I.H. Sloan and H. Woźniakowski, When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? J. Complexity **14**, 1–33 (1998).
23. S. Smale, On the topology of algorithms, J. Complexity **3**, 81–89 (1987).
24. V. Strassen, Vermeidung von divisionen, J. Reine Angew. Math. **264**, 184–202 (1973).
25. S. Tezuka, *Uniform Random Numbers: Theory and Practice*, (Kluwer, Boston, 1995).
26. J.F. Traub, G.W. Wasilkowski, and H. Woźniakowski, *Information-Based Complexity*, (Academic Press, New York, 1988).

27. J.F. Traub and H. Woźniakowski, *A general theory of optimal algorithms*, (Academic Press, New York, 1980).

28. V.A. Vassiliev, Complements of discriminants of smooth maps: topology and applications, Transl. of Math. Monographs 98, 1992, revised 1994, (Amer. Math. Soc., Providence, R.I., 1992).

29. V.A. Vassiliev, Topological complexity of root-finding algorithms, in *The Mathematics of Numerical Analysis*, J. Renegar, M. Shub, and S. Smale, eds., (Lectures in Applied Mathematics, **32** AMS, pp. 831–856, 1996).

30. H. Woźniakowski, Tractability and strong tractability of linear multivariate problems, J. Complexity **10**, 96–128 (1994).

31. H. Woźniakowski, Tractability and strong tractability of multivariate tensor product problems, J. of Computing and Information **4**, 1–19 (1994b).

32. H. Woźniakowski, Why does information-based complexity use the real number model? Theoretical Computer Science **219**, 451–466 (1998).

33. H. Woźniakowski, Efficiency of quasi-Monte Carlo algorithms for high dimensional integrals, in *Monte Carlo and Quasi-Monte Carlo Methods 1998*, H. Niederreiter and J. Spanier, eds. (Springer Verlag, Berlin, pp. 114–136, 1999).

ON SAMPLING INTEGER POINTS IN POLYHEDRA

IGOR PAK
Department of Mathematics, Yale University, New Haven, CT 06520
E-mail: paki@math.yale.edu

We investigate the problem of sampling integer points in rational polyhedra provided an oracle for counting these integer points. When dimension is bounded, this assumption is justified in view of a recent algorithm due to Barvinok [1,2,3]. We show that the exactly uniform sampling is possible in full generality, when the oracle is called polynomial number of times. Further, when Barvinok's algorithm is used, poly-log number of calls suffices.

Introduction

Let $\mathrm{P} \subset \mathbb{R}^d$ be a rational polyhedron of dimension d, where d is a fixed constant. Let $B = \mathrm{P} \cap \mathbb{Z}^d$ be the set of integer points in P. In a pioneering paper [2], Barvinok presented an algorithm for computing $|B|$ in time polynomial in the size of the input. In sharp contrast with various approximation algorithms (see [7,10]), Barvinok's algorithm is algebraic, and by itself insufficient for sampling from B, i.e. picking a uniformly random integer point in P. In this paper we show how one can efficiently utilize advantages of this algorithm for uniform sampling from B.

The problem of uniform sampling of integer points in polyhedra is of interest in computational geometry as well as in enumerative combinatorics, algebraic geometry, and Applied Statistics (see [4,6,9,13,14]). There are numerous algorithms for uniform sampling of combinatorial objects (see e.g. [11,15]), which often can be viewed as integer points in very special rational polyhedra. In statistics, one often need to obtain many independent uniform samples of the integer points in certain polyhedra (e. g. the set of contingency tables) to approximate a certain distribution on them (e. g. χ^2 distribution). We refer to [5,6] for references and details.

Let us note that Monte Carlo algorithms for *nearly uniform* sampling, based on a Markov chain approach, have been of interest for some time. Remarkable polynomial time algorithms (polynomial in even the dimension!) have been discovered (see [7,10]). These algorithms, however, work under certain "roundness" assumptions on polytopes and miss some "hard to reach" points. Theoretical results (see [9]) show hardness of uniform sampling in time polynomial in dimension. While the dimension of polytopes often grows quickly in cases of practical interest, it still remains to be seen what can be done when the dimension is bounded.

By L everywhere below we denote the bit size of the input, and d will denote the dimension (cf. [12]).

Theorem 1. *Let* $\mathrm{P} \subset \mathbb{R}^d$ *be a rational polytope, and let* $B = \mathrm{P} \cap \mathbb{Z}^d$. *Assume an oracle can compute* $|B|$ *for any* P *as above. Then there exists a polynomial time algorithm for sampling uniformly from* B *which calls this oracle* $O(d^2\, L^2)$ *times.*

Theorem 2. *In conditions of Theorem 1, there exists a polynomial time algorithm for sampling uniformly from* B *which calls Barvinok's algorithm* $O(d^2 \log L)$ *times.*

1 Uniform sampling

First we shall prove Theorem 1. Here is a general strategy. We will find a hyperplane H such that $\alpha = |B \cap H_+|/|B|$ and $\beta = |B \cap H_-|/|B| \leq \frac{1}{2}$, where H_-, H_+ are the two halfspaces of $\mathbb{R}^d \setminus H$. Note that we can have $\gamma = |B \cap H|/|B| \geq \frac{1}{2}$, $\alpha + \beta + \gamma = 1$. Then sample a random variable with three outcomes,with respective probabilities α, β, γ. Depending on the outcome, reduce the overall problem to the smaller subproblem. Observe that either dimension drops, or the the number of integer points is reduced by a factor ≥ 2. On the other hand, the dimension can be decreased at most d times. Since the number of integer points is $\exp\bigl(O(d\,L)\bigr)$, we need $O(d\,L)$ times to halve it.

To find the hyperplane as above, consider all level hyperplanes $x_1 = C$, where x_i are coordinates in $V = \mathbb{R}^d$. Clearly, for some integer C this defines H as above. Now determine the constant C by binary search. Recall that C is bounded by $c_1 \leq C \leq c_2$, where c_1, c_2 are polynomial in $\exp(d\,L)$. Checking whether conditions $\alpha, \beta \leq \frac{1}{2}$ are satisfied requires two calls of an oracle for each constant to be tested, the total number of calls to half the polytope is $O(d\,L)$. Combining with the previous observation, this completes the proof of Theorem 1. □

2 Using Barvinok's algorithm

The strategy is similar, but we will choose a desired constant C in a "smarter way", by utilizing the full power of Barvinok's algorithm.

Recall the idea of the algorithm in [2] (see also [1,3]). Given a presentation of P by equations and inequalities, Barvinok computed $F(x) = F(x_1, \ldots, x_d; \mathrm{P})$

defined as

$$F(x_1, \ldots, x_d) = \sum_{m=(m_1,\ldots,m_d)\in B} x^m,$$

where $x^m = x_1^{m_1} \cdot \cdots \cdot x_d^{m_d}$. The solution is given in the form

$$(*) \qquad F(x) = \sum_{j\in J} \epsilon_j \frac{x^{a_j}}{(1 - x^{b_{1,j}}) \cdot \cdots \cdot (1 - x^{b_{d,j}})},$$

where $\epsilon_j \in \{\pm 1\}$, $J = \{1, \ldots, r\}$, and $a_j, b_{i,j} \in \mathbb{Z}^d$, are of size $L^{O(d)}$, polynomial in the size of the input. Now $|B| = F(1, \ldots, 1)$, where the substitution is taken with care (cf. [8]).

The real meaning of $(*)$ is that F is presented as a short alternating sum of the integer points of unimodular cones (with $det = \pm 1$). These cones originate in the vertices a_j of the polytope P. It is crucial that the number of cones $r = |J| = L^{O(d)}$, and was shown in [2] that this bound can be achieved.

Now we can present our algorithm which proves Theorem 2. For simplicity assume that $\mathrm{P} \in \mathbb{R}^d_+$, and has no facets parallel to $H = \{x_1 = 0\}$ (otherwise, one can always find a unimodular transformation of V which places P in general position).

Let us orient all unimodular cones "upward", i.e. to not intersect H. Simply, for each $b_{i,j} \in H_-$ make a substitution $b'_{i,j} = -b'_{i,j}$, $\epsilon'_j = -\epsilon_j$, $a'_j = a_j - b_{i,j}$. Geometrically, this corresponds to flipping a cone in an appropriate cone with the same defining hyperplanes but different orientation. This is possible since the function $F(x) \equiv 0$ for sets containing lines (see part 4) of Theorem 3.1 in [3]). Algebraically, this corresponds to substitution

$$\frac{1}{1 - z^{-1}} = \frac{-z}{1 - z}$$

for every $z = x^{b_{i,j}}$, $b_{i,j} \in H_-$.

Now observe that the volume $\mathrm{vol}(\mathrm{P} \cap \{x_1 \leq C\})$ is piecewise polynomial in C, with the polynomial changing at first coordinate of vertices. Use binary search as in the previous section to determine between which of these the desired C lies (such that $\alpha, \beta \leq \frac{1}{2}$ as in section 1.) The number of vertices is at most L^d, so $O(\log L)$ calls of an oracle suffices. One can simply pick random vertices, use oracle to determine the probabilities of restricting the polytope to either half, etc. With probability $\geq 1/2$ at most $3/4$ fraction of the points will remain in the half, so it will take $O(d \log L)$ iterations. At the end we obtain that the desired "random" point has been sampled uniformly from a polytope $\mathrm{Q} = \mathrm{P} \cap \{c_1 \leq x_1 \leq c_2\}$.

Consider the structure of the polytope Q. Let $Q_C = Q \cap \{x_1 \leq C\}$. From above, the volume $\mathrm{vol}(Q_C)$ is polynomial in C degree C. Recall that we have presented all integer points in Q as an alternating sum of the integer points in the unimodular cones R_j, $j \in J$, since each cone R_j is chosen to have a compact intersection with a plane $\{x_1 = C\}$.

Fix one cone $R = \{a + \mu_1 b_1 + \cdots + \mu_d b_d \,|\, \mu_i \in \mathbb{R}_+\}$, where $a, b_i \in \mathbb{Z}^d$. For simplicity, assume $a = 0$. Denote by M the sum of the first coordinates of b_i (all positive, from above). Observe that every integer point in R_{C-M} corresponds to a block of volume 1 in Q_c, which implies that

$$\left|R_{C-M} \cap \mathbb{Z}^d\right| \leq \mathrm{vol}(R_C) \leq \left|R_{C+M} \cap \mathbb{Z}^d\right|.$$

By linearity, the above inequality holds for Q_C as well.

Now, the volume $\mathrm{vol}(R_C)$ as a polynomial of degree d in C can be explicitly computed from a, b_i and c_1. Thus we obtain an explicit polynomial $f(C)$ for the volume of Q_C. Let $N = |Q \cap \mathbb{Z}^d|$, and pick a random number $n \in \{1, \ldots, N\}$. Estimate the unique solution C_0 of the equation $f(C) = n$ (up to the nearest integer). Then use binary search to determine the desired $C \in \{C_0 - M, \ldots, C_0 + M\}$ (i.e. such that $\alpha, \beta \leq \frac{1}{2}$). This will require $O(\log L)$ oracle calls. Then proceed as in section 1.

Adding up the number of calls for Barvinok's algorithm, we conclude that for each of the d directions we need to call it $O(d \log L)$ times. This completes the proof of Theorem 2. □

3 Concluding remarks

It remains to be seen if Barvinok's algorithm is efficient in practice. In theory, it has $L^{O(d)}$ cost, which is perhaps excessive unless general assumptions are made. In particular, recall that one needs to calculate all vertices of the polyhedron when running Barvinok's algorithm. The main point of this note is to show that at a small additional cost one can use the algorithm for sampling of integer points in the convex hull as well.

Let us give a few simple observations to show that the performance of our algorithm is somewhat better than we showed. First, recall that in section 2 all polytopes Q_C have the same combinatorial structure and thus covered by the second part of Theorem 4.4 in [3]. Also, the estimate $O(d^2 \log L)$ is too conservative. One can make an argument that $O(d \log L)$ is enough when the hyperplane H is chosen appropriately. Roughly, one can choose hyperplanes in general position and avoid paying the "dimension price". Additional analysis of our simple algorithm is unnecessary since the dominating term - cost of Barvinok's algorithm - grows exponentially with the dimension.

Note that when faster approximation algorithms are available, one can use them in place of a counting oracle everywhere when determining which hyperplane to use. But the probabilities must be determined by the precise counting oracle since the errors will blow up otherwise.

Finally, when the function to be approximated on integer points is polynomial or exponential, one can use Barvinok's algorithm to obtain the exact result. In general, however, our approach can be effective.

Acknowledgments

We would like to thank A. Barvinok for explaining his algorithm and for helpful conversations. The author was partially supported by an NSF Postdoctoral Research Fellowship.

References

1. A. Barvinok, Computing the volume, counting integer points, and exponential sums, Discrete and Computational Geometry, **10** (1993), 123–141.
2. A. Barvinok, A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed, Mathematics of Operations Research, **19** (1994), 769–779.
3. A. Barvinok, J. Pommersheim, An algorithmic theory of lattice points in polyhedra, in New perspectives in algebraic combinatorics, Math. Sci. Res. Inst. Publ., **38** Cambridge Univ. Press, Cambridge, 1999, 91–147.
4. M. Brion, Points entiers dans les polytopes convexes, (Séminaire Bourbaki, Vol. 1993/94) Astérisque No. 227, (1995), Exp. No. 780, **4**, 145–169.
5. P. Diaconis, B. Efron, Testing for independence in a two-way table: new interpretations of the chi-square statistic. With discussions and with a reply by the authors, Ann. Statist. **13** (1985), 845–913.
6. P. Diaconis, A. Gangolli, Rectangular arrays with fixed margins, IMA series, Springer **72** (1995), 15–41.
7. M. Dyer, A. Frieze, R. Kannan, A random polynomial-time algorithm for approximating the volume of convex bodies, J. ACM, **38** (1991), 1–17.
8. M. Dyer, R. Kannan, On Barvinok's algorithm for counting lattice points in fixed dimension, Math. Oper. Res., **22** (1997), 545–549.
9. M. Dyer, R. Kannan, J. Mount, Sampling contingency tables, Random Structures and Algorithms, **10** (1997), 487–506.
10. R. Kannan, L. Lovsz, M. Simonovits, Random walks and an $O^*(n^5)$ volume algorithm for convex bodies, Random Structures Algorithms, **11**

(1997), 1–50.

11. J. Propp, D. B. Wilson, How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph, J. Algorithms, **27** (1998), 170–217.
12. A. Schrijer, *Theory of Integer and Linear Programming*, John Wiley, New York, NY, 1988.
13. R. Stanley, *Combinatorics and Commutative Algebra*, Birkhouser, Boston, 1996.
14. B. Sturmfels, Equations defining toric varieties, A.M.S. Proceeding of Symposia in Pure Mathematics, **62** Providence, RI, (1998), 447–449.
15. H. Wilf, Combinatorial algorithms: an update, CBMS-NSF Regional Conference Series in Applied Mathematics, 55. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.
16. G. Ziegler, Lectures on Polytopes, Graduate Texts in Mathematics 152, Springer, New York, 1995.

NEARLY OPTIMAL POLYNOMIAL FACTORIZATION AND ROOTFINDING I: SPLITTING A UNIVARIATE POLYNOMIAL INTO FACTORS OVER AN ANNULUS

VICTOR Y. PAN
Mathematics and Computer Science Department, Lehman College, CUNY, Bronx, NY 10468
E-mail: vpan@alpha.lehman.cuny.edu

Summary

We improve substantially our lifting/descending algorithms that enable numerical splitting of a univariate polynomial of degree n into the product of two nonlinear factors over a fixed thin zero-free annulus on the complex plane. In combination with our improved computation of some basic annuli for splitting, this improves the known algorithms for complete numerical factorizaton of a polynomial into the product of linear factors, so the estimated computational precision and the Boolean (bit-operation) cost bound decrease by roughly the factor of n. The computational complexity estimates supported by the resulting algorithms are optimal (up to polylogarithmic factors) under both arithmetic and Boolean models of computing. The algorithms are immediately extended to yield nearly optimal polynomial rootfinders for the input polynomial with both well conditioned (pairwise isolated) and ill conditioned (clustered or multiple) zeros. The resulting upper estimates for the computational precision and the Boolean (bit-operation) complexity are also nearly optimal and also improve by roughly the factor of n the known estimates for polynomial rootfinding in the case where all the zeros are well conditioned (well isolated). The same algorithm remains nearly optimal (under both Boolean and arithmetic models of computing) where the roots are allowed to be ill conditioned, that is, for the worst case input. All our algorithms can be implemented in polylogarithmic parallel time still using arithmetic and Boolean work which is optimal up to polylog factors. For some classes of input polynomials the presented algorithms are practically promising. The auxiliary analysis of the perturbation of Padé approximation and the converse of the Graeffe lifting may be of independent interest.

Keywords: univariate polynomials, rootfinding, numerical factorization, Padé approximation, Graeffe's lifting, computational complexity

2000 AMS Math. Subject Classification

68Q40, 68Q25, 65D99, 12Y05

1 Introduction

1.1 The subject, some background, and new progress

Univariate polynomial rootfinding is a classical subject, both four millennia old (see [2], [3] on its study since the time of Babylonia and ancient Egypt) and still fundamental for algebraic and numerical computing. The study of this subject is related to various areas of pure and applied mathematics as well as the theory and practice of computing and has huge bibliography [18], [19], [29], [7]. We focus on one important aspect of this study, that is, the computational complexity of the solution under the arithmetic and Boolean bit-operation) models. The modern interest to this aspect of the study is due to [39], [40], and [37], and major progress was obtained quite recently. Nearly optimal solution algorithms were devised in [26], [28]. They rely on the recursive balanced splitting of an input polynomial $p = p(x)$ into the product of two factors of balanced degrees (that is, neither the ratio of the degrees nor its resiprocal can exeed a fixed constant).

More precisely, the recursive balanced splitting finally ends with complete numerical factorization of an input polynomial $p = p(x)$ of degree n into the product of n linear factors, each of which defines a root (zero) of the polynomial. For both complete factorization and rootfinding, the cited algorithms of [26], [28] support the optimal arithmetic time bound $O(n)$ (up to a polylogarithmic factor). Estimating the Boolean time-cost and computational precision, we assume that $p = \sum_{i=0}^{n} p_i x^i$, p_i are real, with the order of b_n bits querried for each coefficient p_i, b being the required output precision.

In the case of complex coefficients p_i, one may routinely shift to the real coefficient polynomial

$$p\bar{p} = \left(\sum_i p_i x^i\right)\left(\sum_i \bar{p}_i x^i\right),$$

$\bar{p}_i$ denoting the complex conjugates of p_i. Arithmetic time and precision bounds combined immediately imply Boolean cost bounds under the customary model of Boolean circuits [11].

We recall that the $\lceil n/2 \rceil$ leading coefficients of p must be represented with the precision of the order of bn to ensure the output precision of b bits (see Fact 1.1 in both [26] and [28]). This implies the Boolean time bound of the order of at least bn^2. That is, the entire input representation of $n+1$ coefficients requires at least the order of nb^2 bits to ensure the output precision of b bits. The algorithms in [26], [28] yield (up to polylog factors) these nearly optimal precision and Boolean cost bounds for the worst case polynomial rootfinding.

The argument of Fact 1.1 in [26], [28] leading to the cited lower bounds on the precision and Boolean cost, however, extends neither to the polynomial factorization problem nor to the special but important case of the problem of rootfinding where all the zeros of a polynomial are well conditioned, that is, simple and well isolated from each other. The lower bounds on the computational complexity of these problems are smaller by the factor of n. That is, the factor of n gap between the lower and upper bounds was left here by the algorithms in [26] and [28].

In the present paper and in [32], we fill this gap. We revisit the construction in [26], [28], modify and rearrange its techniques, and improve the algorithms to yield near optimality for both complete factorization and well conditioned rootfinding.

In the present paper, we revisit the splitting stage provided that the zero sets of the two factors are isolated from each other by a fixed and not extremely thin zero-free annulus on the complex plane [26], [28], [29]. This enables a decisive improvement. In [31], we describe an algorithm that computes the desired basic annuli for splitting by extending the earlier techniques of [23], [26], and [28]. Thus, our work is naturally partitioned into two parts.

Presently we cover splitting stage. We focus on the computation of a sufficiently close initial approximation to the splitting, which is rapidly improved by Newton's iteration elaborated upon in [37] and [16]. Note, however, that the perfection at the improvement stage alone is not sufficient for our final goal. Indeed, the overall arithmetic and Boolean cost bound of the algorithms in [37] and [16] for rootfinding are inferior by roughly factor n to the ones in [26], [28] except that [16] reaches the same Boolean cost bounds as [26], [28] in the pathological case where the output precision has the order of $(1+\frac{1}{\gamma \log n})n^2$ bits, γ being the minimum distance between the distinct roots. We refer the reader to [31], [7], and our Appendix D on some further comparison with related works, to [17], [4], [29], [21], [30], [34], [22], [7], and [32] on the applications to solving polynomial systems of equations, the algebraic eigenproblem, and the computation of approximate polynomial gcd, and to [29], [31] and [7] on further study of polynomial rootfinding and some related subjects. It is important that our algorithm handles the splitting of a polynomial over narrow (zero free basic) annuli. If the annuli are wide, various known techniques can handle splitting as well [37], [31], [16].

More rudimentary version of splitting (based on [37] and not including our lifting/descending process) was implemented by X. Gourdon in 1993 and 1996. This implementation is now a part of the PARI and MAGMA packages. Splitting itself is a major part of Gourdon's implementation, whose power should be substantially extended when our block of lifting/descending process is added because with this block we may utilize more narrow basic annuli for splitting. The implementation of our more advanced algorithms requires further effort, and our ability to utilize narrow basic annuli as well as our improvement of the estimated computational cost should motivate this effort.

Our analysis of the auxiliary stages of reversing Graeffe's lifting and computing Padé approximation may be of independent technical interest—we bound the perturbation of Padé approximants caused by the input perturbation where the zeros of the approximants are weakly isolated from its poles and apply these results to support our descending process, which reverses Graeffe's lifting process.

1.2 The problem and some known results

Let us write

$$p = p(x) = \sum_{i=0}^{n} p_i x^i = p_n \prod_{j=1}^{n} (x - z_j), \quad p_n \neq 0, \tag{1.1}$$

$$A = A(X, r_-, r^+) = \{ \; x \; : \; r_- \leq |x - X| \leq r^+ \; \}, \tag{1.2}$$

$$|u| = ||u(x)|| = \sum_i |u_i| \quad \text{for } u = u(x) = \sum_i u_i x^i, \tag{1.3}$$

$$\mu(b) = O((b \log b) \log\log b), \tag{1.4}$$

$\mu(b)$ denoting the number of bit-operations required to multiply two integers modulo $2^b + 1$. The norm in (1.3) is the 1-norm of the coefficient vector of the polynomial $u = u(x)$. We use both definitions $|u|$ and $||u(x)||$. Assume by default that a polynomial is given with its coefficients and assume w.l.o.g. (cf. [26], [28], [29], [16]) that all its (unknown) zeros satisfy the bounds

$$|z_j| \leq 1, \quad j = 1, \ldots, n. \tag{1.5}$$

We call by "op" each arithmetic operation as well as a comparison of two real numbers and the computation of the values $|z|$ and $|z|^{1/k}$ for a complex number z and a positive integer k. We say that the ratio $\psi = r^+/r_-$ is *the* *relative width* of the annulus A of (1.2). We also call ψ *the isolation ratio* of the internal disc

$$D = D(X, r_-) = \{\ x\ :\ |x| \leq r_-\ \}, \tag{1.6}$$

of the annulus A, and we call this disc ψ -isolated [25], [29].

Technically, the nearly optimal rootfinders of [26], [28] rely on some *preprocessing algorithms*, which compute the basic annulus A for balanced splitting with relative width

$$\psi = \frac{r^+}{r_-} \geq 1 + \frac{c}{n^d} \tag{1.7}$$

for two real constants $c > 0$ and d independent of n and have their latest version in [31], and *splitting algorithms*, which consist of two stages: computing a *crude initial splitting* and its *refinement* by nearly optimal Newton's iterative process. The latter process has been elaborated upon in several papers (see [18], [19], [9], [13], [37], [16]). Hereafter, log stands for $\log_2$.

Theorem 1.1. [37], [16]. *Given a polynomial p of (1.1), (1.5), a positive integer k, $k < n$, real $c > 0$, d,*

$$N = N(n, d) = \begin{cases} n & \text{for } d \leq 0, \\ n \log n & \text{for } d > 0, \end{cases} \tag{1.8}$$

and $b \geq N$, an annulus $A = A(X, r_-, r^+)$ of (1.2), (1.7) such that

$$|z_j| \leq r_- \text{ for } j \leq k, \qquad |z_j| \geq r^+ \text{ for } j > k, \tag{1.9}$$

and two polynomials $\widetilde{F}$ (monic, of degree k, with all zeros lying in the disc $D = D(X, r_-)$) and $\widetilde{G}$ (of degree $n - k$, with all zeros lying outside the disc $D(X, r^+)$), satisfying

$$|p - \widetilde{F}\widetilde{G}| \leq 2^{-\tilde{c}N}|p| \tag{1.10}$$

for a fixed and sufficiently large constant $\tilde{c}$ independent of n, it is sufficient to perform $O((n \log n) \log b)$ ops with $O(b)$-bit precision, that is, $O(\mu(b)n)$ bit-operations for $\mu(b)$ of (1.4), to compute the coefficients of two polynomials $F^* = F^*(x)$ (monic, of degree k, and having all its zeros lying in the disc $D = D(X, r_-)$) and $G^* = G^*(x)$ (of degree $n - k$ and having all its zeros lying outside the disc D) such that

$$|p - F^*G^*| \leq 2^{-b}|p|. \tag{1.11}$$

Theorems 1.1 above and 1.2 below are implicit in [37], [16] although the stated assumptions are slightly different and ops count is not stated in these two papers (see [7] on details). We use an equivalent version of the theorem where we relax assumption (1.5) and make linear transformation of the variable x (shifting X into the origin) to ensure that

$$X = 0, \quad qr_- = 1, \quad q = r^+, \quad \psi = q^2 \tag{1.12}$$

for some $q > 1$. In this case, we say that all concentric annuli $A(0, r_1, r_2)$ for $r_- \leq r_1 \leq r_2 \leq r^+$ as well as the unit circle $C(0,1) = \{\ x,\ |x| = 1\ \}$ are *basic for splitting the polynomial p into factors,* and we call the computation of the factors F^* and G^* of Theorem 1.1 satisfying (1.11) *splitting the polynomial p over the unit circle.* Under these assumptions, an algorithm in [37] (which utilizes the construction in [9]) supports the following result on the initial splitting computation (cf. our Appendix C):

Theorem 1.2. [37] *(cf.* [9], [18], [19], *and* [27] *). Given a polynomial p of (1.1), a real N and an annulus $A = A(0, r_-, r^+)$ such that (1.2), (1.7)-(1.9), (1.12) hold, it is sufficient to perform $O(M \log M)$ ops with $O(N)$-bit precision, that is, $O((M \log M)\mu(N))$ bit-operations, to compute the initial splitting polynomials $\widetilde{F}$ (monic, of degree k, and with all zeros lying in the disc $D = D(0,1)$) and $\widetilde{G}$ (of degree $n - k$ and with all zeros lying outside the disc $D(0,1)$) satisfying (1.10). Here, we have (cf. (1.7)):*

$$M = n + N/(\psi - 1) = \begin{cases} O(n) & \textit{for } d \leq 0, \\ O(n^{1+d} \log n) & \textit{for } d > 0. \end{cases}$$

Based on this theorem the initial splitting can be computed in nearly optimal time (up to a polylog factor) if $d \leq 0$ but not so if $d > 0$.

The bit-operation cost bounds of Theorems 1.1, 1.2, and apparently also 1.3 (below) can be improved by roughly logarithmic factor if one applies fast integer arithmetic based on the binary segmentation techniques (cf. [36], [37], [16], and [6]). Indeed, these techniques are slightly superior to the FFT-based arithmetic, on which we rely to extend the ops and precision bounds of these theorems to the bounds on the bit-operation cost.

1.3 Our results

As in [26], [28], we rely on a lifting/descending process to reduce the case of bound (1.7) for a positive d to the case of $d = 0$ but now yield a substantially stronger result, that is, we obtain the factor of n improvement of the resulting bounds on the computational precision and the Boolean cost versus the bounds in [26], [28]. The next theorem states the complexity estimates, under the same assumption that $b \geq N$ required in the splitting algorithms of [37] and [16], which we use as the basis.

Theorem 1.3. *Under the assumptions of Theorem 1.2, it is sufficient to perform $O((n \log n)(\log^2 n + \log b))$ ops with $O(b)$ precision, that is, to perform $O((\mu(b)n \log n)(\log^2 n + \log b))$ bit-operations, to compute the coefficients of the two polynomials F^* and G^* of Theorem 1.1 satisfying (1.11).*

Combined with the algorithms in [32] for the computation of the zero-free annuli of relative width ψ (for ψ of (1.7)) that support the balanced splitting of a polynomial p, we yield a similar improvement of the known estimates for the computational precision and the Boolean cost of complete numerical factorization of a polynomial into the product of its linear factors and, consequently, of rootfinding for polynomials with well conditioned zeros. As well as the algorithms of [P95a], [P96a], our algorithm supports the nearly optimal arithmetic and Boolean complexity estimates for the rootfinding for the worst case input and allows processor efficient parallel implementation that uses polylogarithmic arithmetic and Boolean parallel time. Note that the upper bound δ on $||p^* - p||$ implies an upper bound $\delta\mu_j$ on the perturbation of the zero z_j of p in the transition to p^* where μ_j is the condition number of the zero z_j under the same norm $|| \cdot ||_1$.

Technically, we focus on the refined analysis of the lifting/descending process, which, in spite of its crucial role in the design of nearly optimal polynomial rootfinders, remains essentially unknown to the computer algebra community. For instance, even the most serious and comprehensive treatise of the splitting of a polynomial in [16] apparently overlooks the glaring flaw in the variation of this process presented in [24], even though this process is a centerpiece of the paper [24], whose main result was invalidated by the flaw (see Appendix D).

Our analysis of this process is technically involved but finally reveals surprising numerical stability (in terms of the asymptotic relative errors of the order 2^{-cn}) of Padé approximation (provided that the zeros of the input analytic function are isolated from its poles) and of Graeffe's lifting process, and this is a springboard for our current progress in polynomial factorization and rootfinding.

1.4 Organization of the paper

In the next section we define our lifting/descending process, which splits a polynomial into two factors over a fixed zero-free annulus. We also estimate the arithmetic cost of the performance of this process and state the bound on the precision of its computation. The correctness of the algorithm under this precision bound is shown in Sections 3 – 5. The analysis includes the error estimates for the perturbation of the Padé approximation involved. In the appendix, we cover the extensions of our splitting over the unit circle to any basic circle (in part A) and to complete numerical factorization of a polynomial (in part B) as well as the computation of an initial splitting (in part C) and comparison with some related works (in part D). The computation of the basic annuli for splitting is covered in [31].

2 Initial Splitting via a Lifting / Descending Process

Here is our algorithm supporting Theorem 1.3:

Algorithm 2.1. Recursive lifting, splitting, and descending.

INPUT: *positive c, r_-, r^+, real $\tilde{c}$ and d, and the coefficients of a polynomial p satisfying (1.1), (1.7), (1.9), and (1.12).*

OUTPUT: *polynomials F^* (monic and of degree k) and G^* (of degree $n-k$), split by the unit circle $C(0,1)$ and satisfying bound (1.10) for $\epsilon = 2^{-\tilde{c}N}$ and N of (1.8).*

COMPUTATION:

*Stage 1 (*recursive lifting*). Write $q_0 = p/p_n$, compute the integer*

$$u = \lceil d \log n + \log(2/c) \rceil, \tag{2.1}$$

and apply root-squaring Graeffe's steps:

$$q_{l+1}(x) = (-1)^n q_l(-\sqrt{x}) q_l(\sqrt{x}), \quad l = 0, 1, \ldots, u-1. \tag{2.2}$$

(Note that $q_l = \prod_{i=1}^{n} (x - z_i^{2^l}), l = 0, 1, \ldots, u$, so $D(0,1)$ is a ψ^{2^l}-isolated disc for q_l, for all l.)

*Stage 2 (*splitting q_u*). Deduce from (2.1) that $\psi^{2^u} > 4$ and apply the algorithms supporting Theorem 1.2 to split numerically the polynomial $p_{\underset{\sim}{u}} = q_u/|q_u|$ over the unit circle. Denote the two computed factors by F_u^* and $\widetilde{G}_u$. Obtain numerical factorization of the polynomial q_u into the product $F_u^* G_u^*$, where $G_u^* = |q_u|\widetilde{G}_u$,*

$$|q_u - F_u^* G_u^*| = \epsilon_u |q_u|, \quad \epsilon_u \leq 2^{-CN} \tag{2.3}$$

for a sufficiently large constant $C = C(c, d)$.

*Stage 3 (*recursive descending*). Based on the latter splitting of q_u, proceed recursively to recover some approximations to the factors F_{u-j} and G_{u-j} that split the polynomials q_{u-j} of (2.2) over the unit circle, for $j = 1, \ldots, u$. Output the computed approximations $F^* = F_0^*$ and $G^* = p_n G_0^*$ to the two* *factors of the polynomial $p = p_n q_0 = FG$. (The approximation error bounds are specified later on.)*

Remark 2.1. *The presented algorithm applies Theorem 1.2 only at Stage 2, where its supporting computations are not costly because we have sufficient isolation of the zeros of the input polynomial p_u from the unit circle, that is, we satisfy relations (1.7), (1.9), and (1.12) with $1/(\psi - 1) = O(1)$, for p replaced by p_u.*

Let us specify Stage 3 of the *recursive descending.*

Stage 3 (*recursive descending*). Step j, $j = 1, 2, \ldots, u$. Stop where $j = u$; for $j < u$, go to the $(j+1)$-st step.

INPUT: the polynomial q_{u-j} (computed at Stage 1) and the computed approximations F^*_{u-j+1} and G^*_{u-j+1} to the factors F_{u-j+1} and G_{u-j+1} of the polynomial q_{u-j+1}, which is split over the unit circle. (The approximations are computed at Stage 2 for $j = 1$ and at the preceding, $(j-1)$-st, descending step of Stage 3 for $j > 1$.)

COMPUTATION: approximate the pair of polynomials $F_{u-j}(x)$ and $G_{u-j}(-x)$ as the pair filling the $(k, n-k)$-entry of the Padé approximation table for the meromorphic function

$$\begin{aligned} M_{u-j}(x) &= q_{u-j}(x)/G_{u-j+1}(x^2) \\ &= (-1)^{n-k} F_{u-j}(x)/G_{u-j}(-x). \end{aligned} \tag{2.4}$$

That is, given polynomials q_{u-j} and G^*_{u-j+1} (the latter one approximating the factor G_{u-j+1} of q_{u-j+1}), first approximate the polynomial $M_{u-j}(x)$ mod x^{n+1}. Then solve the Padé approximation problem (cf. Problem 5.2b (PADÉ) in Chapter 1, or Problem 2.9.2 in [33]) where the computed approximation to the polynomial $M_{u-j}(x)$ mod x^{n+1} is used as the input and approximations $F^*_{u-j}(x)$ and $G^*_{u-j}(-x)$ to the polynomials $F_{u-j}(x)$ and $G_{u-j}(-x)$ are output.

OUTPUT OF STEP j: polynomials $F^*_{u-j} = F^*_{u-j}(x)$ (approximating F_{u-j}) and $G^*_{u-j} = G^*_{u-j}(x)$ (approximating G_{u-j}) such that

$$|F^*_{u-j} G^*_{u-j} - q_{u-j}| = \epsilon_{u-j} |q_{u-j}|, \quad \epsilon_{u-j} \leq 2^{-\bar{c}N}, \tag{2.5}$$

for $\bar{c}$ of (1.10), where $q_{u-j} = F_{u-j} G_{u-j}$, $\deg F^*_{u-j} = k$, the polynomial F^*_{u-j} is monic, and $\deg G^*_{u-j} \leq n - k$.

Bound (2.5) enables us to improve the approximations of F_{u-j} by F^*_{u-j} and of G_{u-j} by G^*_{u-j}, by applying the algorithm supporting Theorem 1.1 where p is replaced by q_{u-j}, F^* by F^*_{u-j}, and G^* by G^*_{u-j}. In the refinement, ϵ_{u-j} remains the value of the order of $1/2^{O(n \log n)}$ for $j < u$, whereas the bound $\epsilon_0 < 2^{-b}$ is ensured at the last (u-th) step.

Of the two computed factors, F^*_{u-j} and G^*_{u-j}, only the latter one is used at the subsequent descending step, although at the last step, both F^* and G^* are output.

The polynomial equations $\gcd(F_{u-j}(x), G_{u-j}(-x)) = 1$ and $G_{u-j+1}(x^2) = (-1)^{n-k} G_{u-j}(x) G_{u-j}(-x)$ together with the ones of (2.4) immediately imply the correctness of Algorithm 2.1 under the assumptions that it is performed with infinite precision and with no rounding errors and that bound (2.5) holds true for $\epsilon_{u-j} = 0$ (that is, that $F^*_{u-j} = F_{u-j}, G^*_{u-j} = G_{u-j}$) for all j.

Let us next estimate the arithmetic complexity of Algorithm 2.1.

Stage 1. $O(un \log n) = O(n \log^2 n)$ ops are used at the $u = O(\log n)$ lifting steps, each amounts to a polynomial multiplication (we use the FFT based algorithms).

Stage 2 (for $\epsilon_u = 1/2^{O(n \log n)}$). a total of $O(n \log^2 n)$ ops are sufficient, by Theorems 1.1 and 1.2.

Stage 3 $O(n \log^2 n)$ ops are used for the computation of the polynomials $M_{u-j+1}(x) \bmod x^{n+1}$ for all j,$j = 1, \dots, u$ (this is polynomial division modulo x^{n+1} for each j) and $O(n \log^3 n)$ ops for the computation of the $(k, n-k)$-th entries of the Padé approximation tables for the polynomials $M_{u-j+1}(x) \bmod x^{n+1}$ for $j = 1, \dots, u$.

Let us specify the latter computation. For every j, this computation is reduced to solving the associated nonsingular Toeplitz or Hankel linear system of $n-k$ equations (see, e.g., [6] , equation (5.6) for $z_0 = 1$ or Proposition 9.4 where $s(x) = 1$); this entry is filled with the nondegenerating pair of polynomials $(F_{u-j}(x), G_{u-j}(-x))$. (Nonsingularity and nondegeneration follow because the polynomials $F_{u-j}(x)$ and $G_{u-j}(-x)$ have no common zeros and, therefore, have only constant common divisors; we extend this property to their approximations in the next section.) Moreover, the input coefficients of the auxiliary nonsingular Toeplitz linear systems (each of $n-k$ equations) are exactly the coefficients of the input polynomial $M_{u-j}(x) \bmod x^{n+1}$ of the Padé approximation problem.

To solve the u nonsingular Toeplitz linear systems (where $u = O(\log n)$), we first symmetrize these systems and then apply the fast MBA algorithm, by Morf and by Bitmead and Anderson (cf. [6] or [33]). The symmetrization ensures positive definiteness of the resulting linear system of equations, and this implies numerical stability of the algorithm (cf. [8]). This approach to the solution of Padé problems enables us to perform the u steps of Stage 3 in $O(n \log^3 n)$ ops. Summarizing, we arrive at the aritmetic cost estimates of Theorem 1.3.

We perform all computations by Algorithm 2.1 with the precision of

$O(n \log n)$ bits, except for the refinement of the approximate initial splitting of the polynomial $q_0(x)$. At the latter stage, we require (1.11) for a fixed $\epsilon = 2^{-b}$, $b \geq N$, and use computations with the b-bit precision. Now to prove Theorem 1.3, it remains to show that under the cited precision bounds, Algorithm 2.1 remains correct, that is, bound (2.5) holds for a fixed and sufficiently large $\tilde{c}$. We show this in the next section.

3 Padé Approximation and Polynomial Splitting: Precision and Complexity Estimates

Our goal to prove that the computational precision of $O(N)$ bits and the bounds of order 2^{-cN} on the values ϵ_{u-j} of (2.5) for $j = 0, 1, \ldots, u$ are sufficient to support Algorithm 2.1. We first recall

Theorem 3.1. [38] . *Let*

$$p = p_n \prod_{j=1}^{n} (x - z_j), \quad p^* = p_n^* \prod_{j=1}^{n} (x - z_j^*),$$
$$|p^* - p| \leq \nu |p|, \quad \nu < 2^{-7n},$$
$$|z_j| \leq 1, \ j = 1, \ldots, k; \quad |z_j| \geq 1, \ j = k+1, \ldots, n.$$

Then after appropriate reordering of z_j^*, we have

$$|z_j^* - z_j| < 9 \sqrt[n]{\nu}, \quad j = 1, \ldots, k;$$
$$|1/z_j^* - 1/z_j| < 9 \sqrt[n]{\nu}, \quad j = k+1, \ldots, n.$$

We easily deduce from the latter theorem that even where ϵ_{u-j} is as large as $2^{-\tilde{c}N}$, we have the desired equation
$\gcd(F_{u-j}^*(x), G_{u-j}^*(-x)) = 1$ for all j:

Corollary 3.1. *Let relations (1.1), (1.9), (1.12), (2.1), and (2.5) hold and let $\epsilon_{u-j} < \min\{2^{-7n}, ((\psi - 1)\theta/9)^n\}$ for all j and a fixed θ, $0 \leq \theta < 1$. Then for all j, $j = 0, 1, \ldots, u$, all zeros of the polynomials $F_{u-j}^*(x)$ and the reciprocals of all zeros of the polynomials $G_{u-j}^*(x)$ lie inside the disc $D(0, \theta + (1-\theta)/\psi)$. For $\psi - 1 \geq c/n^d$, $c > 0$, the latter properties of the zeros are ensured already where $\epsilon_{u-j} \leq 1/n^{O(N)}$ for all j.*

We next estimate the error of splitting the polynomial $q_{u-j}(x)$ in terms of the approximation error bound for splitting the polynomial $q_{u-j+1}(x)$.

Proposition 3.1. *Suppose that a polynomial G_{u-j+1}^* approximates the factor G_{u-j+1} of q_{u-j+1} such that*

$$|F_{u-j+1}^* G_{u-j+1}^* - q_{u-j+1}| \leq \epsilon_{u-j+1} |q_{j-j+1}|$$

for some real ϵ_{u-j+1} and a monic polynomial F^*_{u-j+1} of degree k. Let F^*_{u-j}, G^*_{u-j} denote the solution to the Padé approximation problem solved exactly (with infinite precision) for the input polynomial

$$M^*_{u-j}(x) = (q_{u-j}(x)/G^*_{u-j+1}(x^2)) \bmod x^{n+1},$$

and let ϵ_{u-j} be defined by the equation of (2.5). Then

$$\epsilon_{u-j} = \epsilon_{u-j+1} 2^{O(n \log n)}.$$

Due to this proposition applied for the value ϵ_{u-j} of (2.5) of the order of $\epsilon_{u-j+1}2^{-\tilde{c}N}$ for a large positive $\tilde{c}$, we ensure the splitting of the polynomial q_{u-j} within an error bound (1.10), that is, small enough to allow the subsequent refinement of the splitting based on Theorem 1.1.

The next theorem of independent interest is used in the proof of Proposition 3.1; it shows upper estimates for the perturbation error of the Padé approximation problem. Generally, the input perturbation for this problem causes unbounded output errors but not so in our special case where the zeros of the output pair of polynomials are separated by a fixed annulus containing the unit circle.

Theorem 3.2. *Let us be given two integers, k and n, $n > k > 0$, three positive constants C_0 (to be specified by relations (5.6) and (5.7) of Section 5), γ, and ψ,*

$$\psi > 1, \tag{3.1}$$

and six polynomials F, f, G, g, M and m. Let the following relations hold:

$$F = \prod_{i=1}^{k}(x - \hat{z}_i), \qquad |\hat{z}_i| \leq 1/\psi,\ i = 1, \ldots, k, \tag{3.2}$$

$$G = \prod_{i=k+1}^{n} (1 - x/\hat{z}_i), \quad |\hat{z}_i| \geq \psi,\ i = k+1, \ldots, n \tag{3.3}$$

(compare (1.9), (1.12)),

$$F = MG \bmod x^{n+1}, \tag{3.4}$$

$$F + f = (M + m)(G + g) \bmod x^{n+1}, \tag{3.5}$$

$$\deg f \leq k, \tag{3.6}$$

$$\deg g \leq n - k, \tag{3.7}$$

$$|m| \leq \gamma^n (2 + 1/(\psi - 1))^{-C_0 n}, \tag{3.8}$$

$$\gamma < \min\{1/128, (1 - 1/\psi)/9\}.$$

Then there exist two positive constants C and C^* independent of n and such that if $|m| \le (2+1/(\psi-1))^{-Cn}$, then

$$|f|+|g| \le |m|(2+1/(\psi-1))^{C^*n}. \tag{3.9}$$

The proof of Theorem 3.2 is elementary but quite long and is covered in the next two sections.

PROOF OF PROPOSITION 3.1. The relative error norms ϵ_{u-j} and ϵ_{u-j+1} are invariant in the scaling of the considered polynomials. For convenience, we will use the scaling that makes the polynomials F, F^*, $G_{rev} = x^{n-k}G(1/x)$, and $G^*_{rev} = x^{n-k}G^*(1/x)$ monic, that is, $F = \prod_{j=1}^{k}(x-z_j)$, $F^* = \prod_{j=1}^{k}(x-z_j^*)$, $G = \prod_{j=k+1}^{n}(1-x/z_j^*)$, $G^* = \prod_{j=k+1}^{n}(1-x/z_j^*)$, $q = FG$, $q^* = F^*G^*$, where for simplicity we drop all the subscripts of F, F^*, G, q and q^*. (Note that the polynomials q and q^* are not assumed monic anymore and compare Remark 3.1 at the end of this section.) Furthermore, by (3.1)–(3.3) and Corollary 3.1, we may assume that $|z_j| < 1$, $|z_j^*| < 1$, for $j \le k$, whereas $|z_j^*| > 1$, $|z_j| > 1$, for $j > k$. Therefore, $1 \le |F| < 2^k$, $1 \le |F^*| < 2^k$, $1 \le |G| < 2^{n-k}$, $1 \le |G^*| < 2^{n-k}$, $1 < |q| < 2^n$, $1 < |q^*| < 2^n$.

Let us deduce that

$$\left\|\frac{1}{G_{u-j+1}(x)} \bmod x^{r+1}\right\| \le \left\|(1-x)^{k-n} \bmod x^{r+1}\right\|$$
$$= \sum_{i=0}^{r}\binom{n-k+i-1}{n-k-1} = \binom{n-k+r}{r} < 2^{n-k+r}, \tag{3.10}$$

for any positive r. Indeed, write $(-x)^{n-k}/G_{n-k}(x) = \sum_{i=0}^{\infty} g_i/x_i$. Observe for each i that $|g_i|$ reaches its maximum where $z_i = 1$, that is, where $(-x)^{n-k}/G_{n-k}(x) = x^{n-k}/(1-x)^{n-k}$, and (3.10) follows.

Likewise, we have

$$\|(1/G^*_{u-j+1}(x)) \bmod x^r\| < 2^{n-k+r}.$$

We also apply a bound of Section 10 of [37] to obtain that

$$|G^*_{u-j+1} - G_{u-j+1}| \le \epsilon_{u-j+1}2^{O(N)}.$$

Now, we write

$$\Delta_{u-j+1} = \left(\frac{1}{G^*_{u-j+1}} - \frac{1}{G_{u-j+1}}\right) = \frac{G_{u-j+1}-G^*_{u-j+1}}{G_{u-j+1}G^*_{u-j+1}},$$

summarize the above estimates, and obtain that

$$||\Delta_{u-j+1}(x) \bmod x^r|| \leq \epsilon_{u-j+1} 2^{O(n\log n)}$$

for $r = O(n)$.

Next, let us write $m_{u-j} = m_{u-j}(x) = (M^*_{u-j}(x) - M_{u-j}(x)) \bmod x^{n+1}$ and combine our latter bound with (2.4) and with the bound $|q_{u-j}| \leq 2^n$ to obtain that $|m_{u-j}| \leq \epsilon_{u-j+1}2^{O(N)}$. By combining this estimate with the ones of Theorem 3.2, we obtain that

$$\Delta_{F,G} = |F^*_{u-j} - F_{u-j}| + |G^*_{u-j} - G_{u-j}| \leq \epsilon_{u-j+1}2^{O(N)}.$$

Now, we deduce that

$$\begin{aligned}\epsilon_{u-j} &= |F^*_{u-j}G^*_{u_j} - F_{u-j}G_{u-j}| \\ &\leq |F^*_{u-j}(G^*_{u-j} - G_{u-j}) + (F^*_{u-j} - F_{u-j})G_{u-j}| \\ &\leq |F^*_{u-j}| \cdot |G^*_{u-j} - G_{u-j}| + |F^*_{u-j} - F_{u-j}| \cdot |G_{u-j}| \\ &\leq \max\left\{|F^*_{u-j}|, |G_{u-j}|\right\} \Delta_{F,G} \leq \epsilon_{u-j+1}2^{O(N)}.\end{aligned}$$

□

Similarly to Proposition 3.1, we may prove that any perturbation of the coefficients of the polynomial q_{u-j} within the relative norm bound of the order $1/2^{O(N)}$ causes a perturbation of the factors of q_{u-j} within the relative error norm of the order of $1/2^{O(N)}$ as well.

Proposition 3.1 and Theorem 3.2 together show that the relative errors of the order of $O(N)$ bits do not propagate in the descending process of Stage 3 of Algorithm 2.1. To complete the proof of Theorem 1.3, it remains to prove Theorem 3.2 (see the next sections) and to show that the relative precision of $O(N)$ bits for the output of the descending process of Algorithm 2.1 can be supported by the computations with rounding to the precision of $O(N)$ bits. To yield this goal one may apply the elaborate but tedious techniques of [36] (cf. also [37] and [16]). Alternatively, one may apply the backward error analysis to all the polynomial multiplications and divisions involved, to simulate the effect of rounding errors of such operations by the input perturbation errors. This will lead us to the desired estimates simply via the invocation of Theorem 3.2 and Proposition 3.1, except that we need some distinct techniques at the stages of the solution of Toeplitz or Hankel linear systems of equations associated with the Padé problem.

To extend our analysis to these linear systems, we recall that they are nonsingular because the Padé problem does not degenerate in our case. Moreover, Theorem 3.2 bounds the condition number of the problem. Furthermore, we solve the Pedé problem by applying the cited MBA algorithm to the symmetrized linear systems. (The symmetrization squares the condition number, which may require doubling the precision of the computation, but this is not substantial for proving our estimate of $O(N)$ bits.) We then recall that the algorithm only operates with some displacement generators defined by the entries of the Padé input, $M^*_{u-j}(x) \bmod x^{n+1}$, and is proved to be numerically stable [8]. It follows that $O(N)$-bit precision of the computation is sufficient at the stages of solving Padé problems too, and we finally arrive at Theorem 1.3. □

Remark 3.1. *One could have expected even a greater increase of the precision required at the lifting steps of (2.2). Indeed, such steps generally cause rapid growth of the ratio of the absolutely largest and the absolutely smallest coefficients of the input polynomial. Such a growth, however, does not affect the precision of computing because all our error norm bounds are relative to the norms of the polynomials. Technically, to control the output errors, we apply scaling, to make the polynomials F, F^*, G_{rev} and G^*_{rev} monic, and then proceed as in the proof of Proposition 3.1, where the properties (1.9) of the zeros of the input polynomials are extended to the approximations to the zeros, due to Corollary 3.1.*

4 Perturbation Error Bounds for Padé Approximation

Corollary 4.1, which we will prove in this section, implies Theorem 3.2 in the case where assumption (3.6) is replaced by the following inequality:

$$\deg f < k. \tag{4.1}$$

We need some auxiliary estimates.

Proposition 4.1. [20]. *If $p = p(x) = \prod_{i=1}^{l} f_i$, $\deg p \le n$, and all f_i are polynomials, then $\prod_{i=1}^{l} |f_i| \le 2^n \max_{|x|=1} |p(x)| \le 2^n |p|_2$.*

The next two results extend ones of [37].

Proposition 4.2. *For a fixed pair of scalars, $\psi \ge 1$ and β, let*

$$p = \beta \prod_{i=1}^{k} (x - z_i) \prod_{i=k+1}^{n} (1 - x/z_i),$$

where $|z_i| \le 1/\psi$ for $i \le k$; $|z_i| \ge \psi$ for $i > k$ (cf. (1.9), (1.12)). Then

$$|\beta| \ge |p|/(1 + 1/\psi)^n.$$

Proof. The assumed factorization of the polynomial p yields the inequality

$$|p|/|\beta| \leq (\prod_{i=1}^{k} |x - z_i|) \prod_{i=k+1}^{n} ||1 - x/z_i||,$$

where neither of the n factors on the right-hand side exceeds $1 + 1/\psi$. □

Proposition 4.3. *Let (1.9) hold for some $\psi \geq 1$. Then*

$$|p| \left(\frac{\psi-1}{\psi+1}\right)^n \leq \min_{|x|=1} |p(x)| \leq |p|.$$

Proof. The upper bound on $\min_{|x|=1} |p(x)|$ is obvious. To prove the lower bound, recall the equation of Proposition 4.2 and deduce that

$$|p(x)| \geq |\beta| \prod_{i=1}^{k} |x - z_i| \prod_{i=k+1}^{n} |1 - x/z_i| \text{ for all } x.$$

Substitute the bounds $|x| = 1$, (1.9) and (1.12) and obtain that

$$|p(x)| \geq (1 - 1/\psi)^n |\beta|.$$

Now substitute the bound on $|\beta|$ of Proposition 4.2 and arrive at Proposition 4.3. □

Proposition 4.4. *Let $f(x)$ and $F(x)$ be two polynomials having degrees at most $k - 1$ and k, respectively. Let $R(x)$ be a rational function having no poles in the disc $D(0, 1) = \{x, |x| \leq 1\}$. Then, for any complex x, we have*

$$\int_{|x|=1} R(t) \frac{F(t) - F(x)}{t - x} dt = 0,$$

and if $F(x) \neq 0$ for $|x| = 1$, then

$$f(x) = \frac{1}{2\pi\sqrt{-1}} \int_{|x|=1} \frac{f(t)}{F(t)} \cdot \frac{F(x) - F(t)}{x - t} dt.$$

Proof. (compare [35], III, Ch.4, No.163; [15], proof of Lemma 4.6). The first equation of Proposition 4.4 immediately follows from the Cauchy theorem on complex contour integrals of analytic functions [1]. Cauchy's integral formula [1] implies the second equation of Proposition 4.4 for every x equal to a zero of $F(x)$. If $F(x)$ has k distinct zeros, then the second equation is extended identically in x, since $f(x)$ has a degree less than k. The confluence argument enables us to extend the result to the case of a polynomial $F(x)$ having multiple zeros. □

Now we are prepared to start estimating from above the norms $|f|$ and $|g|$.

Proposition 4.5. *Let a constant ψ and six polynomials F, f, G, g, M and m satisfy relations (3.1)–(3.7),(4.1). Let*

$$v(x) = (G(x) + g(x))G(x)m(x) \bmod x^{n+1}, \deg v \leq n. \tag{4.2}$$

Then we have

$$|f| \leq k\tau|F|, \quad \tau = \max_{|x|=1}\left|\frac{v(x)}{F(x)G(x)}\right|.$$

Proof. Subtract (3.4) from (3.5) and obtain that

$$f(x) = (M(x) + m(x))g(x) + m(x)G(x) \bmod x^{n+1}.$$

Multiply this equation by the polynomial G and substitute

$$F(x) = G(x)M(x) \bmod x^{n+1}$$

into the resulting equation to arrive at the equation

$$G(x)f(x) = F(x)g(x) + (G(x) + g(x))G(x)m(x) \bmod x^{n+1}.$$

Observe that $\deg(Gf - Fg) \leq n$, due to (3.2), (3.3), (3.6) and (3.7), and deduce that

$$Gf = Fg + v, \tag{4.3}$$

for the polynomial v of (4.2). It follows that

$$f = \frac{gF}{G} + \frac{v}{G}.$$

Combine the latter equation with Proposition 4.4 for $R(t) = g(t)F(t)/G(t)$ and deduce that

$$f = \frac{1}{2\pi\sqrt{-1}}\int_{\Gamma}\frac{v(t)}{F(t)G(t)} \cdot \frac{F(x) - F(t)}{x - t}dt.$$

Proposition 4.5 follows from this equation applied to the polynomial f coefficient-wise. □

Let us further refine our bound on $|f|$. Combine (3.2) and (3.3) with Proposition 4.3 and obtain that $\min_{|x|=1}|F(x)G(x)| \geq \frac{(\psi-1)^n}{(\psi+1)^n}|p|$. Now, because $\max_{|x|=1}|v(x)| \leq |v|$, obtain from Proposition 4.5 that

$$\begin{aligned} &|f| \leq k|F| \cdot |v|/\phi_-^n|p|, \\ &\phi_- = (\psi - 1)/(\psi + 1) = 1 - 2/(\psi + 1). \end{aligned} \tag{4.4}$$

Let us bound the norm $|g|$ from above.

Proposition 4.6. *Assume relations (3.1)-(3.7) and (4.1)-(4.3). Then*

$$|g| \leq 2^n \phi_+^{n-k}(|f| + \phi_+^{n-k}|m|)/(1 - 2^n\phi_+^{n-k}|m|),$$

where $\phi_+ = 1 + 1/\psi < 2$.

Proof. Combine the relations $\deg g \leq n - k$ and $\deg F = k$ (cf. (3.2) and (3.7)) with Proposition 4.1 for $l = 2$ and obtain the bound $|F| \cdot |g| \leq 2^n|Fg|$. Therefore, $|g| \leq 2^n|Fg|$ because $|F| \geq 1$ (see (3.2)). On the other hand, (4.3) implies that $|Fg| \leq |G| \cdot |f| + |v|$. Combine the two latter bounds to obtain that $|g| \leq 2^n(|G| \cdot |f| + |v|)$. Deduce from (4.2) that $|v| \leq |G + g| \cdot |G| \cdot |m|$. Substitute the bound $|G| \leq \phi_+^{n-k}$, $\phi_+ = 1+1/\psi$, implied by (3.3), and deduce that

$$|v| \leq (\phi_+^{n-k} + |g|)\phi_+^{n-k}|m|, \tag{4.5}$$
$$|g| \leq 2^n\phi_+^{n-k}(|f| + (\phi_+^{n-k} + |g|)|m|).$$

Therefore, we have

$$(1 - 2^n\phi_+^{n-k}|m|)|g| \leq 2^n\phi_+^{n-k}(|f| + \phi_+^{n-k}|m|),$$

and Proposition 4.6 follows. □

Corollary 4.1. *Assume relations (3.1)-(3.7), (4.1) and let*

$$2^n\phi_+^{n-k}|m| \leq 1/2, \tag{4.6}$$
$$k2^n(\phi_+/\phi_-)^n\phi_+^{n-k}|m| \leq |p|/4. \tag{4.7}$$

Then we have

$$|f| \leq 4k(\phi_+/\phi_-)^n\phi_+^{n-k}|m|/|p|, \tag{4.8}$$
$$|g| \leq 2^{n+1}(1 + 4k(\phi_+/\phi_-)^n/|p|)\phi_+^{2n-2k}|m|, \tag{4.9}$$
$$1 \leq |p| \leq \phi_+^n \tag{4.10}$$

for $\phi_- = 1 - 2/(\psi + 1)$ *of (4.4) and for*

$$\phi_+ = 1 + 1/\psi < 2, \quad \phi_+/\phi_- = (\psi + 1)^2/((\psi - 1)\psi). \tag{4.11}$$

Proof. Combine Proposition 4.6 with inequality and obtain that

$$|g| \leq 2^{n+1}(|f| + \phi_+^{n-k}|m|)\phi_+^{n-k}. \tag{4.12}$$

Combine (4.4), (4.5), and the bound $|F| \leq \phi_+^k$, implied by (3.2), and obtain that

$$|f| \leq k(\phi_+/\phi_-)^n(\phi_+^{n-k} + |g|)|m|/|p|.$$

Combining the latter inequality with (4.12) implies that

$$|p|\cdot|f|\leq k(\phi_+/\phi_-)^n(1+2^{n+1}(|f|+\phi_+^{n-k}|m|))\phi_+^{n-k}|m|.$$

Therefore,

$$\begin{aligned}&|f|\cdot(|p|-k2^{n+1}(\phi_+/\phi_-)^n|m|\phi_+^{n-k})\\ \leq\ & k(\phi_+/\phi_-)^n(1+2^{n+1}\phi_+^{n-k}|m|)\phi_+^{n-k}|m|.\end{aligned}$$

Substitute (4.6) on the right-hand side and (4.7) on the left-hand side and obtain (4.8). Combine (4.8) and (4.12) and obtain (4.10). Combine (3.2) and (3.3) and obtain (4.9). □

5 Local Nonsingularity of Padé Approximations

In this section, we prove Theorem 3.2 by using the following immediate consequence of Corollary 4.1:

Corollary 5.1. *Let all the assumptions of Theorem 3.2 hold, except possibly for (3.6), and let relations (4.1), (4.6) and (4.7) hold. Then bound (3.9) holds for a sufficiently large constant C^*.*

Due to Corollary 5.1, it remains to prove (4.1) under (3.8) in order to complete the proof of Theorem 3.2.

By the Frobenius theorem [14], there exists a unique rational function F/G satisfying (3.4) for any given polynomial M and any pair of integers k and n such that $0\leq k\leq n, \deg F\leq k, \deg G\leq n-k$. Assuming further that the polynomials F and G have no common nonconstant factors and that the polynomial F is monic, we uniquely define the pair of the polynomials F and G (unless M is identically 0), which we call *the normalized pair filling the $(k,n-k)$-th entry* of the Padé table for a polynomial M.

Now, suppose that equations (3.1)–(3.7) hold and let (F,G) and $(F+f,G+g)$ be two normalized pairs filling the $(k,n-k)$-th entry of the Padé table for the meromorphic functions M and $M+m$, respectively, where $\deg F=k$. Then, clearly, we have (4.1) if and only if

$$\deg(F+f)=k. \tag{5.1}$$

Let (F_δ,G_δ) denote the normalized pair filling the $(k,n-k)$-th entry of the Padé table for $M+m+\delta$, where δ is a perturbation polynomial. Even if (5.1) does not hold, there always exists a sequence of polynomials $\{\delta_h\}$, $h=1,2,\ldots,$ such that $|\delta_h|\to 0$ as $h\to\infty$ and

$$\deg F_{\delta_h}=k \text{ for } h=1,2,\ldots. \tag{5.2}$$

(Indeed, the coefficient vectors of polynomials δ for which $\deg F_\delta < k$ form an algebraic variety of dimension k in the space of the $(k+1)$-st dimensional coefficient vectors of all polynomials of degree at most k.)

Due to (5.2), we have $\deg f_{\delta_h} < k$, and therefore, we may apply Corollary 5.1 to the polynomials $M + m + \delta_h$ and obtain that the coefficient vectors of all polynomials F_{δ_n} and G_{δ_h} are uniformly bounded as follows:

$$|F_{\delta_h} - F| + |G_{\delta_h} - G| \le (2 + \tfrac{1}{\psi-1})^{C_1 n}|m + \delta_h| \tag{5.3}$$

provided that $|m + \delta_h| \le (2 + 1/(\psi - 1))^{-C_0 n}$. Because of this bound, there exists a subsequence $\{h(i),\ i = 1, 2, \ldots\}$ of the sequence $h = 1, 2, \ldots$, for which the coefficient vectors $(\mathbf{F}^T_{\delta_{h(i)}}, \mathbf{G}^T_{\delta_{h(i)}})^T$ of the polynomials $F_{\delta_{h(i)}}$, $G_{\delta_{h(i)}}$ converge to some $(n+2)$-nd dimensional vector $(\mathbf{F}^{*T}, \mathbf{G}^{*T})^T$. Let F^*, G^* denote the associated polynomials and let us write

$$F + f = F^*, \quad G + g = G^*. \tag{5.4}$$

Because $\delta_{h(i)} \to 0$, we immediately extend (5.3) and obtain that

$$F^*(x) = (M(x) + m(x))G^*(x) \bmod x^{n+1}$$

and

$$|f| + |g| = |F^* - F| + |G^* - G| \le (2 + \tfrac{1}{\psi-1})^{C_1 n}|m| \tag{5.5}$$

provided that

$$|m| \le (2 + \tfrac{1}{\psi-1})^{-C_0 n}.$$

To complete the proofs of Theorems 3.2 and 1.3, it remains to show that $\deg f < k$, that is, that $\deg F^* = k$ and that the polynomials F^* and G^* of (5.4) have only constant common factors. We do this by applying Theorem 3.1. First combine the bounds (4.8) and (3.8) (where C_0 satisfies the bound

$$(2 + \tfrac{1}{\psi-1})^{C_0 n} \ge 4k(\phi_+)^{n-k}\left(\tfrac{\phi_+}{\phi_-}\right)^n |p|/|F| \tag{5.6}$$

for ϕ_- and ϕ_+ of (4.4) and (4.11)) with Theorem 3.1, for p and p^* replaced by F and F^*, respectively, and deduce that the zeros of the polynomial $F + f$ deviate from the respective zeros of the polynomial F by less than $1 - 1/\psi$, so the polynomial $F + f$ has exactly k zeros all lying strictly inside the unit disc $D(0, 1)$. Similarly, obtain that $\deg(G + g) = n - k$ and all the zeros of the polynomial $G + g$ lie outside this disc (provided that the constant C_0 of (3.8) satisfies the inequality

$$(2 + \tfrac{1}{\psi-1})^{C_0 n} \ge 2^{n+1}\psi_+^{2n-2k}\left(1 + \tfrac{4k}{|p|}\left(\tfrac{\phi_+}{\phi_-}\right)^n /\right)\frac{|m|}{|G|} \tag{5.7}$$

(cf. (4.9))), so the polynomial $G + g$ has only constant common factors with $F + f$. This completes the proof of (4.1) and, therefore, also the proofs of Theorems 3.2 and 1.3. □

Acknowledgment

Thanks go to a referee for helpful comments. Supported by NSF Grant CCR 9732206 and PSC-CUNY Awards 669363 and 62435-0031

References

1. L. Ahlfors, *Complex analysis*, McGraw-Hill, New York, 1979.
2. E. T. Bell, *The Development of Mathematics*, McGraw-Hill, New York, 1940.
3. C. A. Boyer, *A History of Mathematics*, Wiley, New York, 1968.
4. L. Blum, F. Cucker, M. Shub, S. Smale, *Complexity and Real Computations*, Springer, New York, 1997.
5. R. P. Brent, F. G. Gustavson, D. Y. Y. Yun, Fast solution of Toeplitz systems of equations and computation of Padé approximations, *J. Algorithms*, **1**, 259–295, 1980.
6. D. Bini and V. Y. Pan, *Polynomial and Matrix Computations, Vol.1: Fundamental Algorithms*, Birkhäuser, Boston, 1994.
7. D. Bini and V. Y. Pan, *Polynomial and Matrix Computations, Vol.2: Fundamental and Practical Algorithms*, Birkhäuser, Boston, to appear.
8. J. R. Bunch, Stability of Methods for Solving Toeplitz Systems of Equations, *SIAM J. Sci. Stat. Comput.*, **6**, **2**, 349–364, 1985.
9. L. M. Delves, J. N. Lyness, A numerical method for locating zeros of an analytic functions, *Math. Comp.*, **21**, 543–560, 1967.
10. I. Z. Emiris, A. Galligo, H. Lombardi, Numerical Univeriate Polynomial GCD, *Proc. of AMS-SIAM Summer Seminar: Mathematics of Numerical Analysis: Real Number Algorithms, (Park City, Utah, 1995), Lectures in Applied Math.*, **32**, 323–343, Amer. Math. Society, Providence, Rhode Island, 1996.
11. J. von zur Gathen, Parallel Arithmetic Computations: A Survey, *Proc. Math. Foundation of Computer Science, Lecture Notes in Computer Science,* **233,** 93–112, Springer, Berlin, 1986.
12. I. Z. Emiris, A. Galligo, H. Lombardi, Certified approximate polynomial gcds, *J. Pure and Applied Algebra*, **117/118**, 229–251, 1997.
13. A. A. Grau, The simultaneous improvement of a complete set of approximate factors of a polynomial, *SIAM J. of Numer. Analysis*, **8**, 425–438,

1971.

14. W. B. Gragg, The Padé table and its relation to certain algorithms of numerical analysis, *SIAM Review*, **14, 1**, 1–62, 1972.
15. P. Kirrinnis, Fast computation of numerical partial fraction decompositions and contour integrals of rational functions, *Proc. Inter. Symp. on Symb. and Algebraic Comput. (ISSAC 92)*, (Paul S. Wang editor), 16–26, ACM Press, New York, 1992.
16. P. Kirrinnis, Polynomial factorization and partial fraction decomposition by simultaneous Newton's iteration, *J. of Complexity*, **14, 3**, 378–444, 1998.
17. D. Kapur and Y. N. Lakshman, Elimination methods: An introduction, in *Symbolic and Numerical Computation for Artificial Intelligence* (B. Donald, D. Kapur, and J. Mundy, editors), pp. 45–89, Academic Press, New York, 1992.
18. J. M. McNamee, bibliography on roots of polynomials, *J. Comp. Appl. Math.*, **47**, 391–394, 1993.
19. J. M. McNamee, A supplementary bibliography on roots of polynomials, *J. Computational Applied Mathematics*, **78, 1**, 1997, also http://www.elsevier. nl/homepage/sac/cam/mcnamee/index.html.
20. M. Mignotte, An inequality about factors of polynomials, *Math. Comp.*, **28**, 1153–1157, 1974.
21. B. Mourrain, V. Y. Pan, Asymptotic acceleration of solving polynomial systems, *Proc. 27th Ann. ACM Symp. on Theory of Computing*, 488–496, ACM Press, New York, May 1998.
22. B. Mourrain, V. Y. Pan, Multivariate Polynomials, Duality, and Structured Matrices, *J. of Complexity*, **16, 1**, 110–180, 2000.
23. C. A. Neff, J. H. Reif, An $O(n^{l+\epsilon})$ algorithm for the complex root problem, in *Proc. 35th Ann. IEEE Symp. on Foundations of Computer Science*, 540–547, IEEE Computer Society Press, Los Alamitos, California, 1994.
24. C. A. Neff, J. H. Reif, An efficient algorithm for the complex roots problem, *J. of Complexity*, **12**, 81–115, 1996.
25. V. Y. Pan, Sequential and parallel complexity of approximate evaluation of polynomial zeros, *Computers & Math. (with Applications)*, **14, 8**, 591–622, 1987.
26. V. Y. Pan, Optimal (up to polylog factors) sequential and parallel algorithms for approximating complex polynomial zeros, *Proc. 27th Ann. ACM Symp. on Theory of Computing*, 741–750, ACM Press, New York, May, 1995.
27. V. Y. Pan, Deterministic improvement of complex polynomial factoriza-

tion based on the properties of the associated resultant, *Computers & Math. (with Applications)*, **30**, **2**, 71–94, 1995.

28. V. Y. Pan, Optimal and nearly optimal algorithms for approximating polynomial zeros, *Computers & Math. (with Applications)*, **31**, **12**, 97–138, 1996.
29. V. Y. Pan, Solving a polynomial equation: Some history and recent progress, *SIAM Review*, **39**, **2**, 187–220, 1997.
30. V. Y. Pan, Approximate polynomial gcds, Padé approximation, polynomial zeros, and bipartite graphs, *Proc. 9th Ann. ACM-SIAM Symp. on Discrete Algorithms*, 68–77, ACM Press, New York, and SIAM Publications, Philadelphia, 1998.
31. V. Y. Pan, Univariate polynomials: nearly optimal algorithms for factorization and rootfinding, *ACM-SIGSAM International Symposium on Symbolic and Algebraic Computation (ISSAC' 2001)*, ACM Press, New York, 2001.
32. V. Y. Pan, Approximate polynomial gcd, *Information and Computation,* in press.
33. V. Y. Pan, *Structured matrices and polynomials: Unified superfast Algorithms,* Birkhäuser/Springer, Boston, 2001.
34. V. Y. Pan, Z. Q. Chen, The complexity of the matrix eigenproblem, *Proc. 31st Annual ACM Symposium on Theory of Computing*, 507–516, ACM Press, New York, 1999.
35. G. Pólya, G. Szegö, *Aufgaben und Lehrsätze aus der Analysis*, Verlag Von Julius Springer, Berlin, 1925.
36. A. Schönhage, Asymptotically fast algorithms for the numerical multiplication and division of polynomials with complex coefficients, *Proc. EUROCAM, Marseille, Lecture Notes in Computer Science*, **144**, 3–15, Springer, Berlin, 1982.
37. A. Schönhage, The fundamental theorem of algebra in terms of computational complexity, Math. Dept., University of Tübingen, Tübingen, Germany, 1982.
38. A. Schönhage, Quasi-gcd computation, *J. of Complexity*, **1**, 118–137, 1985.
39. S. Smale, The fundamental theorem of algebra and complexity theory, *Bull. Amer. Math. Soc.*, **4**, **1**, 1–36, 1981.
40. S. Smale, On the efficiency of algorithms of analysis, *Bulletin of the American Mathematical Society*, **13**, **2**, 87–121, 1985.

Appendix A: Extension to Splitting over Any Circle

By the initial scaling of the variable, we may move the zeros of a given polynomial into the unit disc $D(0,1)$. Therefore, it is sufficient to consider splitting of a polynomial p of (1.1) (within a fixed error tolerance ϵ) over any disc $D(X,r)$, with X and r satisfying the bounds $r > 0$ and

$$r + |X| \leq 1. \tag{A.1}$$

To reduce such a splitting to the normalized case of splitting over the unit circle $C(0,1)$, we will shift and scale the variable x and estimate the new relative error norm bound $\tilde{\epsilon}$ as a function in ϵ, X and r. To relate ϵ and $\tilde{\epsilon}$, we will prove the following result:

Proposition A.1. *Let relations (1.11) and (A.1) hold. Write*

$$y = rx + X, \tag{A.2}$$

$$\begin{aligned} \tilde{p}(y) &= \sum_{i=0}^{n} \tilde{p}_i y^i = \tilde{p}(rx+X) = q(x), \\ p(x) &= q(x)/||q(x)||, \end{aligned} \tag{A.3}$$

$$\begin{aligned} \widetilde{F}^*(y) &= \widetilde{F}^*(rx+X) = F^*(x)r^k, \\ \widetilde{G}^*(y) &= \widetilde{G}^*(rx+X) = G^*(x)/(||q(x)||r^k), \\ \Delta(x) &= p(x) - F^*(x)G^*(x), \\ \widetilde{\Delta}(y) &= \tilde{p}(y) - \widetilde{F}^*(y)\widetilde{G}^*(y). \end{aligned}$$

Then (A.2) maps the disc $D(0,1) = \{x : |x| \leq 1\}$ onto the disc $D(X,r) = \{y : |y - X| \leq r\}$; moreover,

$$\begin{aligned} ||\widetilde{\Delta}(y)|| &\leq ||\Delta(x)|| \cdot ((1+|X|)/r)^n \cdot ||\tilde{p}(y)|| \\ &\leq ||\Delta(x)|| \cdot ((2-r)/r)^n \cdot ||\tilde{p}(y)||. \end{aligned} \tag{A.4}$$

Proof. Clearly, (A.2) maps the disc $D(X,r)$ as we stated. To prove (A.4), first observe that $\Delta(x) = \Delta(\frac{y-X}{r}) = \widetilde{\Delta}(y)/||q(x)||$. Therefore,

$$||\widetilde{\Delta}(y)|| = ||\Delta(\tfrac{y-X}{r})|| \cdot ||q(x)||. \tag{A.5}$$

By combining the relations $1 \le ||(y-X)^i/r^i|| = (1+|X|)^i/r^i$, for $i = 0, 1, \ldots,$ with the one of (A.1), we deduce that

$$\begin{aligned} ||\Delta\left(\tfrac{y-X}{r}\right)|| &\le ||\Delta(x)|| \cdot \max_i \left(\tfrac{||(y-X)^i||}{r^i}\right) \\ &= ||\Delta(x)|| \left(\tfrac{1+|X|}{r}\right)^n \\ &\le ||\Delta(x)|| \left(\tfrac{2-r}{r}\right)^n . \end{aligned} \tag{A.6}$$

On the other hand, having $q(x) = \tilde{p}(rx+X)$ and $||(rx+X)^i|| = (r+|X|)^i$ for $i = 0, 1, \ldots,$ we deduce that

$$||q(x)|| = ||\tilde{p}(rx+X)|| = \left\| \sum_{i=0}^{n} \tilde{p}_i (rx+X)^i \right\| \le \sum_{i=0}^{n} |\tilde{p}_i| (r+|X|)^i .$$

Due to (A.1), it follows that

$$||q(x)|| \le \sum_{i=0}^{n} |\tilde{p}_i| = ||\tilde{p}(y)||.$$

Combine the latter bound with (A.5) and (A.6) to obtain (A.4). □

Appendix B: Error Estimates for Recursive Splitting

Suppose that we recursively split each approximate factor of p over the boundary circle of some well isolated disc and continue this process until we arrive at the factors of the form $(ux+v)^d$. This gives us a desired approximate factorization

$$p^* = p^*(x) = \prod_{j=1}^{n} (u_j x + v_j), \tag{B.1}$$

and we next estimate the norm of the residual polynomial

$$\Delta^* = p^* - p. \tag{B.2}$$

Note that the perturbation of the coefficients of p such that $|p^* - p| \le \delta$ implies that a zero z_j of p is perturbed by at most $\delta \mu_j$ where μ_j is the condition number of this zero [7]. We are going to estimate the perturbation of the zeros without involving the condition numbers μ_j. We begin with an auxiliary result from [37] .

Proposition B.1. *Let*

$$\Delta_k = |p - f_1 \dots f_k| \leq k\epsilon|p|/n, \tag{B.3}$$

$$\Delta = |f_1 - fg| \leq \epsilon_k|f_1|, \tag{B.4}$$

for some nonconstant polynomials $f_1, \dots, f_k, f$ *and* g *and for*

$$\epsilon_k \leq \epsilon|p|/(n\prod_{i=1}^{k}|f_i|). \tag{B.5}$$

Then

$$|\Delta_{k+1}| = |p - fgf_2 \dots f_k| \leq (k+1)\epsilon|p|/n. \tag{B.6}$$

Proof. $\Delta_{k+1} = |p - f_1 \dots f_k + (f_1 - fg)f_2 \dots f_k| \leq \Delta_k + \Delta|f_2 \dots f_k|$. Substitute (B.3)–(B.5) and deduce (B.6). □

If we write $f_1 = f, f_{k+1} = g$, then (B.6) will turn into (B.3) for k replaced by $k+1$. If we now split one of the factors f_i, as in (B.4), we may apply Proposition B.1 and then recursively continue splitting p into factors of smaller degrees until we arrive at factorization (B.1), with

$$|\Delta^*| \leq \epsilon|p| \tag{B.7}$$

for Δ^* of (B.2). Let us call this computation **Recursive Splitting Process** provided that it starts with $k = 1$ and $f_1 = p$ and ends with $k = n$.

Proposition B.2. [37]. *Performing Recursive Splitting Process for a positive* $\epsilon \leq 1$, *it is sufficient to choose* ϵ_k *in (B.4) satisfying*

$$\epsilon_k \leq \epsilon/(n2^{n+1}) \tag{B.8}$$

for all k *in order to support (B.3) for all* $k = 1, 2, \dots, n$.

Proof. We will prove the bound (B.3) for all k by induction on k. Clearly, the bound holds for $k = 1$. Therefore, it remains to deduce bound (B.6) from bounds (B.3) and (B.8) for any k. By first applying Proposition 4.1 and then the bound of (B.3), we obtain that

$$\prod_{i=1}^{k}|f_i| \leq 2^n \left|\prod_{i=1}^{k} f_i\right| \leq 2^n(1 + k\epsilon/n)|p|,$$

which cannot exceed $2^{n+1}|p|$ for $k \leq n, \epsilon \leq 1$. Consequently, (B.8) ensures (B.5), and then (B.6) follows by Proposition B.1. □

Appendix C: Initial Approximate Splitting via FFT

In this section, we follow [9] and [37] (compare also Appendix A of [27]) to describe an algorithm that supports Theorem 1.2, that is, given a polynomial p satisfying (1.1), (1.9) and (1.12) computes its initial approximate splitting over the unit circle $C(0,1)$ into the product of two factors F and G . Furthermore, the algorithm performs within the cost bounds of Theorem 1.2.

The algorithm first computes sufficiently close approximations

$$s_m^* = \frac{1}{Q}\sum_{q=0}^{Q-1} \omega^{(m+1)q}\frac{p'(\omega^q)}{p(\omega^q)}, \tag{C.1}$$

$m = 1, \ldots, Q$; $\omega = \exp(2\pi\sqrt{-1}/Q)$, to the power sums,

$$s_m = \sum_{j=1}^{k} z_j^m, \quad m = 1, \ldots, 2k-1,$$

and then approximates (within the error bounds 2^{-cn} for two fixed constants $c = c_F$ and $c = c_G$) the coefficients of the two factors F and G.

Let us estimate the errors of tha computed approximations and the computational cost. By [37], we have

$$|s_m^* - s_m| \le (kz^{Q+m} + (n-k)z^{Q-m})/(1-z^Q), \tag{C.2}$$

$$z = \max_{1\le j\le n} \min(|z_j|,\ 1/|z_j|). \tag{C.3}$$

The computation of the values $s_1^*, \ldots, s_{2k-1}^*$ costs $O(Q\log Q)$ ops for $Q \ge n$ because it is reduced to performing three discrete Fourier transforms (DFT's) on the set of the Q-th roots of 1. Due to (C.2) and (C.3), it is sufficient to choose Q of the order of $N(n)/(\psi - 1)$ to ensure the error bound

$$|s_m^* - s_m| < 2^{-cN(n)} \tag{C.4}$$

for any function $N(n) \ge n$, for all $m < 2k$, and for any fixed constant c. Under such a choice, the bound $O(Q\log Q)$ turns into $O(\frac{N(n)}{\psi-1}\log\frac{N(n)}{\psi-1})$, and it is sufficient to perform the computations with the precision of $O(N(n))$ bits (cf. Corollary 4.1 of [6], Chapter 3). Then, the algorithm from Section 4 of Chapter 1 of [6] for Problem 4.8 ($I \cdot POWER \cdot SUMS$) which uses $O(n\log n) ops$(performed with $O(N(n))$-bit precision) computes an approximation F^* to the factor $F = \prod_{j=1}^{k}(x - z_j)$ of the polynomial p, within the error norm bound

$$\epsilon_F|F| = |F^* - F|, \quad \epsilon_F \le 2^{-c_F N(n)}, \tag{C.5}$$

for some fixed constant c_F (provided that the exponent c in (C.4) is chosen sufficiently large). Then again, it is sufficient to perform the computations by this algorithm with $O(N(n))$-bit precision (cf. [37] .

Similarly we compute an approximation G^*_{rev} to the factor G_{rev} of the reverse polynomial $p_{rev} = F_{rev}G_{rev}$, where we write $w_{rev}(x) = x^m w(1/x) = \sum_{i=0}^{m} w_i x^{m-i}$ for a polynomial $w(x) = \sum_{i=0}^{m} w_i x^i$ of a degree m. Observe that, the sets of the coefficients as well as the norms of any pair of polynomials w and w_{rev} coincide with each other. On the other hand, all zeros of the reverse polynomial G_{rev} lie in the disc $D(0, 1/f)$. Therefore, the same techniques that we applied previously enable us to approximate the polynomial G_{rev}, which gives us a polynomial G^* of degree $n - k$ satisfying

$$\epsilon_G|G| = |G^* - G|, \quad \epsilon_G \le 2^{-c_G N(n)} \tag{C.6}$$

for some fixed constant c_G. (See [37] for the alternative ways of the computation of the approximate factor G^* via polynomial division.)

Let us now deduce the bound 1.10. With no loss of generality, we may assume that $|p| = 1$. Write $p^* = F^*G^*$ and recall that $|F| \le 2^n$ (by (1.9)), $|F^*| \ge 1$ (because F^* is a monic polynomial), $|G^*| \le 2^n|p^*|/|F^*|$ (by Proposition 4.1), and therefore, $|G^*| \le 2^n|p^*| \le 2^n(1 + |p^* - p|)$. Observe that $p^* - p = F^*G^* - FG = (F^* - F)G^* + F(G^* - G)$, write $\epsilon_p = |p^* - p|$, and deduce that

$$\epsilon_p \le \epsilon_F|G^*| + \epsilon_G|F| \le 2^n(\epsilon_F(1 + \epsilon_p) + \epsilon_G) \le 2^{-c_p N(n)} \tag{C.7}$$

for $c_p < \min(c_F, c_G) - n - 2$, provided that $c_F \ge 1$.

It is sufficient to perform the entire algorithm for the above computation of F^* and G^* with the precision of $O(N(n))$ bits to arrive at the error norm bounds of (C.5)–(C.7) (apply the estimates of [36], equation (12.6) of [37], and Corollary 4.1 of [6], Chapter 3). By summarizing our analysis, we obtain Theorem 1.2. □

Remark C.1. *By choosing sufficiently large constants c_p (or c_F and c_G), one may ensure that the unit circle $C(0,1)$ splits the polynomials F^* and G^* (see Theorem 3.1).*

Appendix D: Modifications of the Descending Process

Consider modifications of the descending stage of Algorithm 2.1 based on either or both of the two following equations applied for all j:

$$F_{u-j}(x) = \gcd(q_{u-j}(x), F_{u-j+1}(x^2)),$$
$$G_{u-j}(x) = \gcd(q_{u-j}(x), G_{u-j+1}(x^2)), \quad j = 1, \dots, u.$$

Here and hereafter, $\gcd(u(x), v(x))$ denotes the monic greatest common divisor (gcd) of the two polynomials $u(x)$ and $v(x)$.

In this modification of Algorithm 2.1, Padé computation is replaced by the polynomial gcd computation. This produces the same output as in Algorithm 2.1 if we assume infinite precision of computing. The approach was originally introduced in the proceedings paper [26] but in its journal version [28] was replaced by the one based on Padé computation. This enabled more direct control over the propagation of the perturbation errors (cf. Theorem 3.2), although both approaches can be made computationally equivalent because both Padé and gcd computations can be reduced to the same Toeplitz linear system of equations (cf. [5], [6]).

The gcd approach, however, may lead into a trap if one tries to solve the gcd problems based on the fast Euclidean algorithm (cf. Algorithm 5.1a of [6] or [33]). In this case, each descending step (2.4) is replaced by a recursive Euclidean process, known to be prone to the severe problems of numerical stability involved in it (cf. [38], [10], and [12]) and to possible blow up of the precision of the computations and their Boolean cost as a result. In particular, this is the case with the paper [24] where the fast Euclidean algorithm in the gcd version of the descending process is applied reproduced from [26], but unfortunately their analysis of its Boolean cost has been invalidated by a technical flaw. Namely, the analysis hinges on the invalid assumption that the value $\delta = \psi - 1$ exceeds a fixed positive constant (ψ^2 being the relative width of the basic annulus for splitting a polynomial q_{u-j}). This assumption is satisfied only for the polynomials q_{u-j} computed at a few last lifting steps, that is, for $j = u - O(1)$ but not for $j = 0, 1, \dots, u/2$ (say). Thus, the analysis presented in [24] applies only to a few first descending steps, and the Boolean cost of performing all other steps remains unbounded. Furthermore, this flaw is not easy to fix; clearly it cannot be fixed based on the techniques of the paper [24]. Fortunately, the distinct construction of [28] achieves the same result, and now we have its extension to the complete factorization problem as well.

COMPLEXITY ISSUES IN DYNAMIC GEOMETRY

JÜRGEN RICHTER-GEBERT
TU Müchen, Zentrum Mathematik, SB4, D-80290 München, Germany
E-mail: richter@ma.tum.de

ULRICH H. KORTENKAMP
FU Berlin, Institut fr Informatik, Takustrae 9, D-14195 Berlin, Germany
E-mail: kortenkamp@inf.fu-berlin.de

This article deals with the intrinsic complexity of tracing and reachability questions in the context of elementary geometric constructions. We consider constructions from elementary geometry as *dynamic entities*: while the free points of a construction perform a continuous motion the dependent points should move consistently and continuously. We focus on constructions that are entirely built up from *join*, *meet* and *angular bisector* operations. In particular the last operation introduces an intrinsic ambiguity: Two intersecting lines have *two* different angular bisectors. Under the requirement of continuity it is a fundamental algorithmic problem to resolve this ambiguity properly during motions of the free elements.
After formalizing this intuitive setup we prove the following main results of this article:

- It is NP-hard to trace the dependent elements in such a construction.
- It is NP-hard to decide whether two instances of the same construction lie in the same component of the configuration space.
- The last problem becomes PSPACE-hard if we allow one additional sidedness test which has to be satisfied during the entire motion.

On the one hand the results have practical relevance for the implementations of Dynamic Geometry Systems. On the other hand the results can be interpreted as statements concerning the intrinsic complexity of analytic continuation.

1 Introduction

1.1 What is Dynamic Geometry

Imagine any construction of elementary geometry – for instance, a ruler and compass construction of the midpoint of two points A and B. It consists of certain *free* elements (the points A and B) and certain *dependent* elements whose positions are determined by the positions of the free elements. Each specific drawing of such a construction is a snapshot that belongs to the whole continuum of all possible drawings for all possible locations of the free

Keywords: Dynamic Geometry, analytic continuations, NP, PSPACE, complexity, reachability, ruler and compass, linkages, warehouseman's problem

elements. If we move the free elements we can walk continuously from one *instance* (i.e. snapshot) of the construction to another one. During such a walk a continuous motion of the free elements should result in a continuous movement of the dependent elements.

This article deals with those effects and problems that genuinely arise from such a dynamic and continuous setup of geometry. The research that led to the results presented in this article was motivated by the desire (and the actual work) of implementing a software package for doing Dynamic Geometry on a computer [22,23]. With such a program one should be able to do constructions of elementary geometry with a few mouse clicks, and after this pick the free elements with the mouse – drag them around – while the whole construction follows accordingly. The unsuspicious looking requirement of *continuity of dependent elements* turned out to be fundamentally hard to fulfill. In fact, one has to rely on notions of complex function theory and Riemann surfaces to get a mathematically sound treatment of these effects [11,12]. While this is no problem in theory, we prove here that from a complexity theoretic point of view most algorithmic questions related to that context are provably intractable (unless P=NP, of course). The complexity classes that arise here range from NP-hard problems via PSPACE-hard problems up to even undecidable problems. In particular we prove that ...

... it is NP-hard to calculate the positions of the dependent elements after a specific move of a free element (Sec. 5),

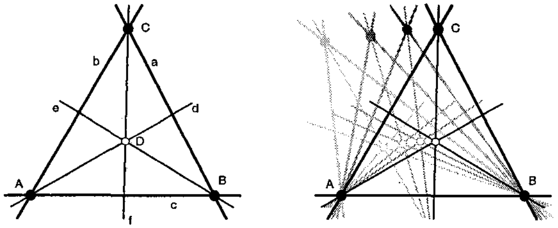
... in general, it is PSPACE-hard to decide whether two instances of the same construction can be continuously deformed into each other if all free and dependent elements must have real coordinates (Sec. 6),

... this reachability problem is still NP-hard if only *join*, *meet*, and *angular bisector* operations occur (Sec. 4),

... it is undecidable whether two instances of a construction involving "wheels" – devices that transfer angles to distances – can be continuously deformed into each other by moving the base elements (Sec. 7).

Although the results of this article arose from the study of configuration spaces of elementary geometric constructions they are naturally related to many other setups in the area of geometry. Among those are the study of configuration spaces of mechanical linkages [6,9,10], realization spaces of oriented matroids [16,4,20,25] and polytopes [21], and the warehouseman's problem [7,24]. The results of this article are partially generalizations and strengthenings of known complexity results in these areas. Besides the context of Dynamic

```
1:  A=FreePoint;
2:  B=FreePoint;
3:  C=FreePoint;
4:  a=Join(B,C);
5:  b=Join(A,C);
6:  c=Join(A,B);
7:  d=AngularBisector(b,c);
8:  e=AngularBisector(a,c);
9:  f=AngularBisector(a,b);
10:  D=Meet(d,e);
```



Construction sequence static picture idea of the dynamic picture

Fig. 1: Dynamic behavior of the angular bisector theorem.

Geometry our results are relevant for all areas where geometric objects are moved around under certain geometric constraints, like robotics, parametric CAD [5], virtual reality, or computational kinematics. Our results imply that many problems of these areas are computationally difficult (like the *persistent naming problem* of parametric CAD [5] or the *navigation problem* of computational kinematics). Also one can interpret the results of this paper as statements on the complexity of analytic continuation (all coordinate functions in our setup turn out to be analytic). In particular this gives intrinsic complexity bounds on homotopy methods for solving polynomial equations as they were discussed in [26,27,28,29,30]. This article is complemented by [11,12] were we give conceptual approaches to handle a dynamic setup of geometry at all.

1.2 Constructions, Forbidden Situations and Ambiguities

In a typical setup for this article we will study construction sequences in which each single construction step is of very elementary nature like taking the *join of two points*, the *meet of two lines*, the *angular bisector of two lines*, or the *intersection of a line and a circle*, etc. A construction sequence starts with some free points and generates new elements by performing elementary operations on already existing elements one at a time. It may happen that an operation cannot be carried out (for instance, if one wants to construct the join of two identical points, the meet of two identical or parallel lines, or the intersection of a line and a circle that do not meet). In order to avoid such situations let us assume that the input points are in suitable positions such that each step of the construction sequence can be done. In that case we will call the input point position *admissible*, otherwise we call it *forbidden*.

The *join* and *meet* operations are *deterministic* construction steps in the

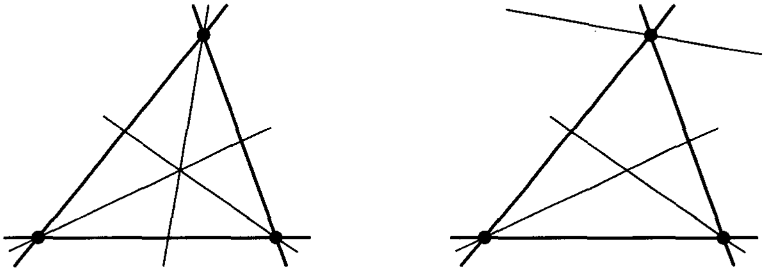


Fig. 2: Two instances of the angular bisectors of a triangle.

sense that for each admissible input there exists exactly one corresponding output element (for instance two distinct, non-parallel lines have exactly one point of intersection). Construction sequences that exclusively involve join and meet operations are easy to handle: If for a certain position of the free elements each construction step is admissible then the positions of the dependent elements are uniquely determined.

The situation is substantially different for operations like *intersection of a circle and a line*, or *angular bisector of two lines.* For these operations one has a binary choice of what the output of an operation should be (two lines have two angular bisectors, a line and a circle have in general two points of intersection). For a construction involving such operations the positions of the dependent elements are no longer uniquely determined by the positions of the free elements. This kind of *non-determinism* will be captured by the concept of a *geometric straight line program*, which is formalized in Sec. 2 (see also [11]).

The intrinsic ambiguities of these operations together with continuity requirements are the fundamental sources that make the algorithmic problems studied in this article difficult. These intrinsic ambiguities even touch the very heart of the notion of *"What is a geometric theorem?"* Consider the theorem stating that *the angular bisectors of the sides of a triangle meet in a point.* Due to the intrinsic ambiguity of the angular bisector operation this sentence stated as such is not true. Consider the drawing in Fig. 2. It shows two valid instances of the construction: Take three points – form the three joins of any pair of them – draw the three angular bisectors of any pair of lines. In the left drawing the chosen angular bisectors meet, in the right drawing they do not.

Having these ambiguities in mind, in the context of Dynamic Geometry two natural questions arise:

- **Reachability problem:** Is it possible to move the free points such that a first instance is smoothly deformed into a specific second one?

- **Tracing problem:** How can a Dynamic Geometry program decide after a motion of the free elements what instance to draw for the new position?

After a suitable formalization, we will show that the reachability problem is in general PSPACE-hard. It is still NP-hard if one restricts oneself to constructions that only use *join*, *meet*, and *angular bisector* operations. The tracing problem turns out to be (at least) NP-hard.

1.3 Restricting the Operations

We try to formulate our statements as strongly as possible and restrict the allowed elementary operations to a minimum. The only operations we will use are *join*, *meet*, *angular bisectors* and *intersection of circle and line.* Furthermore, we assume that initially four fixed *constant* base points $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$ are given. In fact, we try to use the *intersection of circle with line* operation as sparsely as possible. The reason for this is that the possible non-existence of such an intersection includes the possibility to encode sidedness conditions and "cutting holes" in the admissible range of input parameters. By explicitly excluding operations like *intersection of circle and line* we substantially strengthen our results. In fact the NP-hardness results for the tracing and reachability problems can be exclusively stated in terms of angular bisectors, join and meet. The PSPACE-hardness results need just one single intersection-of-circle-and-line operation. We also try to introduce as few free points as possible into or constructions. This complements many other related complexity results since there usually many free variables are needed. The following table summarizes the complexity results covered in this article:

Problem	**Complexity**	#free points	#angular bisectors	# int. circle/lines	# wheels
Tracing	NP-hard	1	many	—	—
Reachability	NP-hard	many	3	—	—
Reachability	NP-hard	1	many	—	—
Reachability	PSPACE-hard	1	many	1	—
Reachability	PSPACE-hard	many	—	1	—
Reachability	Undecidable	1	2	—	11

In order to exclude unnecessary technicalities arising from special cases we also assume that the plane is extended by elements at infinity to the usual projective plane.

The results that use exclusively angular bisectors (and join and meet) are in a sense generalizations of corresponding results in other setups in the following sense. While with mechanical linkages or with ruler and compass constructions it is easily possible to construct angular bisectors, the converse is impossible. A complexity theoretic lower bound for a setup that uses only angular bisectors is therefore a stronger result than a corresponding one for linkages or ruler and compass constructions.

1.4 Related Results

There are other related areas of geometry where similar complexity results arise. In this section we want to briefly discuss the relations – similarities and differences – to these results.

1.4.1 Oriented Matroids and Polytopes

Research over the last few decades showed that for oriented matroids and polytopes so called *universality theorems* can be proved. These theorems show that the corresponding realization spaces can essentially be (stably equivalent to) any solution space of a system of (finitely many) polynomial equations and inequalities [16,4,20,21,25]. These results are usually derived by a direct translation procedure, which starts from a system of polynomials and ends up with a configuration of the desired category (an oriented matroid or a polytope). In the realization space of the oriented matroid (or polytope) the original variables of the algebraic equations can be rediscovered from the coordinates of certain points. The constructions of universality theorems deeply rely on the generation of large loops that feedback the result of an evaluation of a polynomial to a initially chosen constant (for instance a point whose coordinates represent the "1".) Deciding whether the realization space of an oriented matroid or polytope is empty or not turns out to be NP-hard. Deciding whether two realizations of an oriented matroid (or polytope) are in the same connected component of the realization space turns out to be PSPACE-hard.

In Sec. 4 although we derive a similar result for elementary geometric constructions where we rely on very different effects. The complex behavior is generated by a strictly forward oriented construction without any feedback of information. Orientation information cannot and is not included in any way

IFor the PSPACE-hardness result in Sec. 6. the significant difference to the constructions for oriented matroids or polytopes is that we use only *one* free point instead of *many* free points.

1.4.2 Mechanical Linkages

A similar comparison holds for mechanical linkages, for which universality theorems are known [6,9,10] that prove that arbitrary primary semialgebraic sets can show up as components of configuration spaces. The corresponding reachability problem ("Do two instances of a linkage lie in the same component of the configuration space?") is also PSPACE-hard.

The results there are obtained by making use of construction loops and inequality relations. The inequality relations arise naturally in that context, since the bars of a linkage are only of finite length. In our setup only weaker construction primitives are allowed. Furthermore also the linkage results rely on the introduction of *many* free elements, in contrast to our results.

1.4.3 Warehouseman's Problem

Moving an object with a non-restricted number of degrees of freedom through a world of geometric obstacles leads to another PSPACE-hard reachability problem [7,24]. Again, the use of inequality relations is already inherent in the statement of the problem, and many free elements are needed.

1.5 Acknowledgements

We want to thank Alexander Below, Vanessa Krummeck and Jesus de Loera for careful proofreading and many valuable discussions. We want to thank Maurice Rojas for drawing our attention to Plaisted's Theorem that served as a paradigm for the proof in Sec. 5 and simplified our original construction. We also thank Yuri Matiyasevich who supplied us with the proper reference for the best known bound for the number of variables that is needed to prove the undecidability of Hiberts 10th problem over the integers (see Sec. 7).

2 Geometric Constructions

2.1 Geometric Straight Line Programs

We now start to formalize the concept of a geometric construction. We take special care to have a setup that allows the results of an operation to be non-existent or ambiguous. For this we first define the notion of a *relational instruction set.* Here, instead of giving an algorithm or formula for the operation, only a relation is specified that enables us to check the validity of a certain input and output pair. A slightly more general approach can be found in [11].

Definition 2.1. A *relational instruction set (RIS)* is a pair $(\mathcal{O}, \Omega)$ of *objects* $\mathcal{O}$ and *primitive operations* Ω with the following properties: $\mathcal{O} = (\mathcal{O}_1, \ldots, \mathcal{O}_k)$ is a family of sets $\mathcal{O}_i$. These sets partition the objects into classes of the same *type*. The primitive operations $\Omega = (\omega_1, \ldots, \omega_l)$ are relations

$$\omega_i \subset (\mathcal{O}_{x_1^i} \times \cdots \times \mathcal{O}_{x_{s_i}^i}) \times \mathcal{O}_{x_{s_i+1}^i}$$

with *input size (arity)* $\mathrm{ar}(\omega_i) = s_i$ and *type* $\mathrm{type}(\omega_i) := \mathcal{O}_{x_{s_i+1}^i}$. An element $(o_1, \ldots, o_{\mathrm{ar}(\omega)}) \in \mathcal{O}_{x_1^i} \times \cdots \times \mathcal{O}_{x_{s_i}^i}$ is called an *input* and an element $o \in \mathcal{O}_{x_{s_i+1}}$ is called an *output* of ω_i.

Remark 2.2. For the relation $(o_1, \ldots, o_{\mathrm{ar}(\omega)}, o) \in \omega_i$ we will also use the more intuitive notation

$$o \leftarrow \omega_i(o_1, \ldots, o_{\mathrm{ar}(\omega)}).$$

This notion may be considered as a non-deterministic assignment operation. It assigns to an input $(o_1, \ldots, o_{\mathrm{ar}(\omega)})$ one of the potential outputs o of ω. However, one should have in mind that this notion still represents a *relation* that can be *true* or *false*. It is true if the input is admissible for the operation and if the output is one of the proper evaluations of ω on this input.

In our geometric setup the different classes of objects will correspond to points, lines, circles, etc. Each primitive operation will represent a certain type of geometric primitive construction like join, meet, angular bisectors, etc. In addition, we will allow special operations to create free points which will play the role of the "input" of our constructions. Observe that relational instruction sets are general enough to describe not only geometric, but also arithmetic operations (see [11]).

We now describe the specific objects and operations used in this article. Although we will make use of Euclidean operations, we will describe the purely incidence geometric part for points and lines in terms of projective geometry. This will exclude unnecessary special cases and helps in defining the right concept of continuity later on. We embed everything in the real projective plane $\mathbb{RP}^2$. In the usual way we can represent points and lines (in homogeneous coordinates) by vectors in $\mathbb{R}^3 \setminus \{(0,0,0)\}$. Vectors that only differ by a scalar multiple are identified and represent the same point (or line). A point (x, y, z) is on a line (a, b, c) if and only if $ax + by + cz = 0$. Meet and join can then be simply expressed as cross-products of such vectors (see for instance [11]).

Since we also want to deal with objects and operations of Euclidean geometry like circles and angular bisectors, we have to embed the usual Euclidean plane (equipped with a Euclidean metric) in $\mathbb{RP}^2$. A finite point $(x, y) \in \mathbb{R}^2$ will be represented by the point $(x, y, 1)$ of $\mathbb{RP}^2$. With this standard embedding a line $ax + by + c = 0$ of $\mathbb{R}^2$ is represented by (a, b, c), $\ell_\infty = (0, 0, 1)$ represents the line at infinity, and two lines $l_1 = (a_1, b_1, c_1)$ and $l_2 = (a_2, b_2, c_2)$ are orthogonal if $a_1 b_2 - b_2 a_2 = 0$. In order to simplify the notation later on we will also identify a finite point $(x, y, 1)$ with a complex number $x + iy$. By this we identify the finite part of the projective plane with $\mathbb{C}$.

We restrict the use of angular bisectors to those lines that pass through the origin $(0, 0)$ of $\mathbb{R}^2$. An angular bisector of two lines l_1 and l_2 through the origin is a line ℓ through the origin such that $\angle(l_1, \ell) = \angle(\ell, l_2)$. For a pair of lines there are two angular bisectors, which are orthogonal to each other. Restricting the use of angular bisectors to lines through the origin reduces the occurrence of non-admissible situations to a minimum. Formally, we will make use of the following primitive operations. For the sets P of points and L of lines, we define:

$$\begin{aligned}
\text{JOIN} &:= \{(p_1, p_2, l) \mid l \text{ is the line through } p_1 \text{ and } p_2 \text{ and } p_1 \neq p_2\} \\
&\subset (P \times P) \times L \\
\text{MEET} &:= \{(l_1, l_2, p) \mid p \text{ is the intersection of } l_1 \text{ and } l_2 \text{ and } l_1 \neq l_2\} \\
&\subset (L \times L) \times P \\
\text{BISECT} &:= \{(l_1, l_2, l) \mid l \text{ is an angular bisector of } l_1 \text{ and } l_2 \text{ and} \\
&\quad l_1, l_2, l \text{ pass through the origin}\} \\
&\subset (L \times L) \times L
\end{aligned}$$

Furthermore, we define the following four *constants* (i.e. primitives with input size zero):

$$P^{(a,b)} := \{(a, b, 1)\} \subset P; \text{ for } a, b \in \{0, 1\}.$$

These constants will be used to fix a coordinate system. For the generation of *free points* we define a special instruction that has no input elements and allows the output to be any point of P:

$$\text{FREE} := P.$$

We will deal with the following relational instruction set:

$$\text{JMB} := ((P, L), (\text{JOIN}, \text{MEET}, \text{BISECT}, \text{FREE}, P^{(0,0)}, P^{(1,0)}, P^{(0,1)}, P^{(1,1)})).$$

Remark 2.3. Here are three comments on the choice of the primitive operations:

(i) The only cases where the two operations JOIN and MEET are not admissible is when the two input elements are identical. For all other cases they are well-behaved.

(ii) The only operation that introduces an ambiguity is BISECT. The primitives JOIN and MEET are "deterministic" in the sense that each admissible input has exactly one possible output.

(iii) The operation BISECT has been chosen for our investigations since it isolates the effect of generating an ambiguity. Unlike the *intersection circle with line* operation it has no open region of the input parameters where it is not admissible. Such effects (which we want to exclude here) would allow the possibility to construct some kind of "sidedness test", which are at the core of of complexity results for oriented matroids, polytopes, mechanical linkages or the warehouseman's problem. In Sec. 6 when we prove the PSPACE-hardness result we will make a very selected use of one such additional operation.

A construction sequence is formalized by the concept of a *geometric straight line program* (GSP).

Definition 2.4. A *geometric straight-line program* on a relational instruction set $(\mathcal{O}, \Omega)$ is defined by a sequence of *statements* $\Gamma = (\Gamma_1, \dots, \Gamma_m)$. Each Γ_j has the form $\Gamma_j = (\omega, i_1, \dots, i_{\mathrm{ar}(\omega)})$ where

(i) ω is an operation from the instruction set Ω,

(ii) the type of Γ_j is defined to be the type of ω,

(iii) for each $k \in \{1, \dots, \mathrm{ar}(\omega)\}$ we have $i_k < j$,

(iv) for each $k \in \{1, \dots, \mathrm{ar}(\omega)\}$ the type of Γ_{i_k} matches the type of the k-th input of ω.

After a suitable set of primitive operations is given it is straightforward to describe construction sequences by a GSP. Each statement $\Gamma_j = (\omega, i_1, \dots, i_{\mathrm{ar}(\omega)})$ of a GSP describes the generation of a new element by means of a primitive operation ω whose input is given by the output of the statements $\Gamma_{i_1}, \dots, \Gamma_{i_{\mathrm{ar}(\omega)}}$. Item (iii) of the above definition ensures that only elements are used as input that have been already constructed. Item (iv) ensures a correct typing. The concept of a GSP emphasizes the constructive step-by-step nature, however it allows for a certain "non-determinism" during a construction, since it does not specify which output of an (ambiguous) operation to take. To make GSPs more readable we also use the " $\leftarrow$ " notation of Remark 2.2 to encode each statement. A statement $\Gamma_j = (\omega, i_1, \dots, i_{\text{ar}(\omega)})$ will then be written as $j \leftarrow \omega(i_1, \dots, i_{\text{ar}(\omega)})$. Furthermore, we allow to exchange the references $j, i_1, \dots, i_{\text{ar}(\omega)}$ by meaningful variable names.

We may consider a certain set of primitive operations as a kind of programming language. Each GSP is a certain program. In what follows we are mainly interested in the constructions/programs that can be described by the operations in JMB.

Example 2.5. *The following sequence of instructions is a simple GSP over the* JMB *instruction set. It takes two free points p and q, joins them to the origin o, and constructs the angular bisector of the two resulting lines.*

$$\begin{aligned}
p &\leftarrow \text{FREE}\\
q &\leftarrow \text{FREE}\\
o &\leftarrow P^{(0,0)}\\
l_1 &\leftarrow \text{JOIN}(a, o)\\
l_2 &\leftarrow \text{JOIN}(b, o)\\
b &\leftarrow \text{BISECT}(l_1, l_2)
\end{aligned}$$

We will still simplify the notions by assuming that points that do not occur explicitly on the left of any assignment are automatically initialized by a FREE operation. Furthermore, if the output of an operation is unique and used only once we allow that it is used directly (without intermediate variable) as an input of another operation. In particular this convention applies to the constants in JMB. With these conventions the above GSP can simply be written as

$$b \leftarrow \text{BISECT}(\text{JOIN}(p, P^{(0,0)}), \text{JOIN}(q, P^{(0,0)})).$$

Closely related to the concept of a GSP $(\Gamma_1, \dots, \Gamma_m)$ is the notion of an instance of the GSP. Roughly speaking an instance of a GSP is an assignment of a concrete object to each of the statements Γ_i such that all corresponding relations are satisfied.

Definition 2.6. An *instance* of a geometric straight-line program $(\Gamma_1, \dots, \Gamma_m)$ is an assignment of objects $\tilde{X} = \tilde{X}_1, \dots, \tilde{X}_m$ such that all primitives are *satisfied*, that is, for every statement $\Gamma_j = (\omega_j, i_1, \dots, i_{\text{ar}(\omega_j)})$ the relation $(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_{\text{ar}(\omega_j)}}, \tilde{X}_j) \in \omega_j$ holds.

Example 2.7. *For the GSP given in Example 2.5. we have in particular the* *following instance (in homogeneous coordinates):*

$$\begin{array}{l} \tilde{p} = (1,0,1) \ \tilde{q} = (0,1,1) \ \tilde{o} = (0,0,1) \\ \tilde{l}_1 = (1,0,0) \ \tilde{l}_2 = (0,1,0) \ \tilde{b} = (1,1,0) \end{array}$$

It is important that for the same choice of free elements there also exists another possible instance that exactly differs in the choice of the angular bisector:

$$\begin{array}{l} \tilde{p} = (1,0,1) \ \tilde{q} = (0,1,1) \ \tilde{o} = (0,0,1) \\ \tilde{l}_1 = (1,0,0) \ \tilde{l}_2 = (0,1,0) \ \tilde{b} = (-1,1,0) \end{array}$$

Remark 2.8. By our definition of an instance we implicitly assume that for any specific instance the positions of the elements are *admissible* in the sense that each primitive operation can be executed.

Remark 2.9. A more formal treatment of RIS's and GSPs would include a careful separation of syntax and semantics of GSPs, a separation of references to objects and the objects themselves, and many other subtleties that are present whenever the aim is to formalize the concept of computing. However, we hope that the slightly informal treatment used in this article satisfies the needs of the reader as long as only complexity issues are concerned. A more elaborated treatment of GSPs can be found in [11].

2.2 Continuity

Along with the notion of GSPs and their instances comes a natural notion of continuity. For this we will split a specific GSP $\mathcal{P} = (\Gamma_1, \ldots, \Gamma_m)$ over the instruction set JMB into input variables and dependent variables. We consider each point in $\mathcal{P}$ that comes from a FREE operation as an *input* to $\mathcal{P}$. W.l.o.g. we may assume that the definition of the input points are the first k statements $\mathcal{P}$. Each of the operations JOIN,MEET, and BISECT has only a finite number of possible output values. This is the case since if we prescribe the positions of the input points all other objects of this instance are determined up to a finite number of possible binary choices. Each choice that has to be made comes from one application of a BISECT operation.

Now assume that $p_1, \ldots, p_k$ are the input points of $\mathcal{P}$. Furthermore, assume that we are given continuous functions

$$p_i(t)\colon [0,1] \to \mathbb{R}^3 \setminus \{(0,0,0)\}$$

for each $i \in \{1, \ldots, k\}$. These functions describe a continuous movement of the input points (in homogeneous coordinates).

Definition 2.10. A *continuous evaluation* of the GSP $\mathcal{P}$ over the JMB instruction set under the movement $p_i(t)$ is an assignment of continuous functions

$$o_i(t)\colon [0,1] \to \mathbb{R}^3 \setminus \{(0,0,0)\}$$

for each $i \in \{k+1, \ldots, m\}$ such that for all $t \in [0,1]$ the objects

$$(p_1(t), \ldots, p_k(t), o_{k+1}(t), \ldots, o_m(t))$$

form an (admissible) instance of $\mathcal{P}$.

This concept formalizes the intuitive requirement that under a continuous movement of free elements the dependent elements should move continuously as well. For instance, if we have the simple GSP of Example 2.5 and move from one instance to another by changing the positions of the free elements a and b, a continuous evaluation makes sure that we do not jump spontaneously from one choice of the angular bisector to the other one.

Observe that the way we define continuity leaves room for the necessary indeterminism: Usually one would require that the output elements are given by continuous functions in the input, but here both the path of the input and the path of the output are given by continuous functions on the interval [0,1].

The following property of continuous evaluations is crucial:

Lemma 2.11. *If there exists a continuous evaluation of the GSP $\mathcal{P}$ over the* JMB *for a continuous movement $p_i(t)$ then it is unique.*

Proof. We can prove this lemma by induction on the length of P. Assuming that the statement holds for all programs of length $m-1$ we prove that it also holds for programs of length m. Assume that for such a program $\mathcal{P}$ the functions $p_i(t)$ describe a continuous movement for which a continuous evaluation exists. If the last operation of $\mathcal{P}$ is one of the constant points then the statement holds trivially. If the last operation of $\mathcal{P}$ is one of the deterministic operations JOIN or MEET, then the statement holds by the continuity of these operations. If the last operation is BISECT then we can argue as follows: The two possible outputs of BISECT are two lines that are orthogonal to each other. If there was a way to continuously get from one branch to the other there must be a position in which these two lines coincide. This is impossible since the two angular bisectors are orthogonal. □

Remark 2.12. This Lemma shows the importance of non-admissible positions: At these *singularities* the different branches coincide, both angular bisectors degenerate to the zero vector. It is not possible to extend the projective setting by this additional line $(0,0,0)$ (and a corresponding point) without destroying the uniqueness of continuous evaluations, even though it is possible to extend the JMB instruction set to include them.

2.3 Fundamental Problems in Dynamic Geometry

After formalizing the concept of GSPs and continuity we are finally in the position of formalizing the main questions of this work. The first problem formalizes the most fundamental operation of a Dynamic Geometry program: After you pick a free point of a construction and move it to another position, how did the rest of the construction change?

Definition 2.13. (Tracing problem): Let $\mathcal{P}$ be a GSP and let $p_i(t)$ describe a continuous movement for which a continuous evaluation $(p_1(t), \dots, p_k(t), o_{k+1}(t), \dots, o_m(t))$ exists. Furthermore, let $(p_1, \dots, p_k, o_{k+1}, \dots, o_m)$ be an instance of $\mathcal{P}$ with free points $p_i = p_i(1)$ for all $i \in \{1, \dots, k\}$. Decide whether $o_i = o_i(1)$ for all $i \in \{k+1, \dots, m\}$.

The second problem asks for the mere existence of a path from one instance to another.

Definition 2.14. (Reachability problem): Let $P^0 = (p_1^0, \dots, p_k^0, o_{k+1}^0, \dots, o_m^0)$ and $P^1 = (p_1^1, \dots, p_k^1, o_{k+1}^1, \dots, o_m^1)$. Decide whether there exists a continuous evaluation that starts at P^0 and ends at P^1.

We will see that both problems turn out to be (at least) NP-hard. If we allow one single use of a sidedness test to constrain admissible regions the reachability problem even turns out to be PSPACE-hard.

3 Useful Gadgets

This section will describe small constructions that are helpful to compose the more complicated constructions that we need later.

3.1 More Primitives

Since our set of primitive operations is very restricted we first show that other useful primitive operations can be easily composed from these primitives.

3.1.1 The Line at Infinity

By a simple sequence of join and meet operations we can construct the line at infinity:

$$\begin{aligned} a &\leftarrow \textsc{meet}(\textsc{join}(P^{(0,0)}, P^{(0,1)}), \textsc{join}(P^{(1,0)}, P^{(1,1)})) \\ b &\leftarrow \textsc{meet}(\textsc{join}(P^{(0,0)}, P^{(1,0)}), \textsc{join}(P^{(0,1)}, P^{(1,1)})) \\ \ell_\infty &\leftarrow \textsc{join}(a, b) \end{aligned}$$

By construction, a and b are two distinct points on the line at infinity and hence ℓ_∞ is the line at infinity with homogeneous coordinates $(0, 0, 1)$.

3.1.2 Parallel Lines

For a line l and a point p we can calculate the parallel to l through p by

$$\textsc{join}(\textsc{meet}(l, \ell_\infty), p).$$

If p lies on l this formula produces l itself. If $l = \ell_\infty$ this formula is not admissible. We will refer to this "macro" by $\textsc{parallel}(l, p)$.

3.1.3 Perpendicular

A bit less trivial is the construction of a perpendicular to l trough p. We can only do such a construction since the choice of our constant points provides us with a sample of two perpendicular lines. This right angle can then be transferred to another line. Since we already have a parallel operation w.l.o.g. we may assume that l passes through $P^{(0,0)}$ and that $p = P^{(0,0)}$. The construction is given in Fig. 3. We have

$$\begin{aligned} a &\leftarrow \textsc{meet}(\textsc{join}(P^{(1,0)}, P^{(1,1)}), l), \\ b &\leftarrow \textsc{meet}(\textsc{join}(P^{(0,0)}, P^{(0,1)}), \textsc{parallel}(a, \textsc{join}(P^{(1,0)}, P^{(0,1)}))), \\ c &\leftarrow \textsc{meet}(\textsc{join}(P^{(0,1)}, P^{(1,1)}), \textsc{parallel}(b, \textsc{join}(P^{(0,0)}, P^{(1,1)}))), \\ perp &\leftarrow \textsc{join}(c, P^{(0,0)}). \end{aligned}$$

This construction is admissible for all situations where l passes through $P^{(0,0)}$. We will refer to the general construction for perpendiculars by $\textsc{perpendicular}(l, p)$.

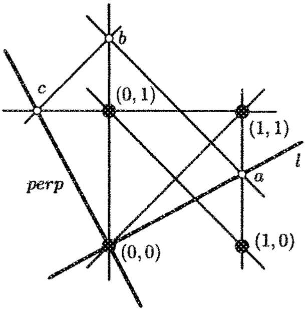


Fig. 3: Construction of a perpendicular.

3.2 Arithmetics

An essential part of our constructions will be the evaluation of certain polynomial expressions. For this we single out one particular line l on which we perform the evaluation. On this line we fix two points that play the roles of "0" and "1" and therefore fix an origin and a scale. To every point x on this line we can assign a unique value with respect to this scale. This value is given by the ratio $\frac{|0x|}{|01|}$ of oriented segment lengths. Sometimes we will abuse notation and use the name of the point as name for the value.

3.2.4 Von Staudt Constructions

The evaluation of arbitrary polynomials can be done if we are able to perform an elementary addition $z = x + y$ and an elementary multiplication $z = x \cdot y$. This can be done by the classical *von Staudt constructions.* They are shown in Fig. 4. In these pictures lines that seem to be parallel are *really* parallel. The desired arithmetic relations follow immediately from the similarities of

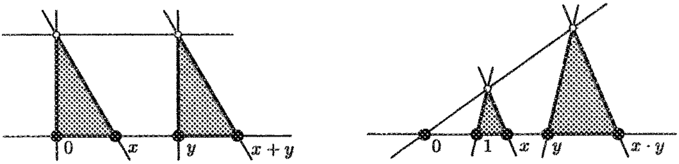


Fig. 4: Von Staudt constructions for addition and multiplication.

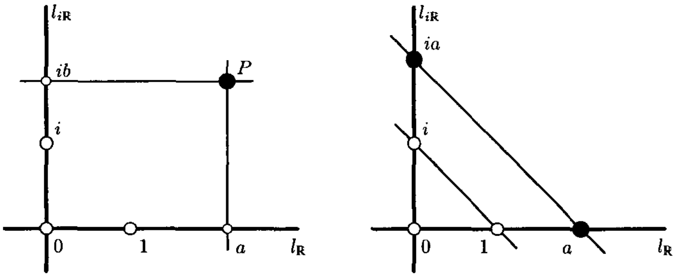


Fig. 5: Coordinate extraction for "complex" points.

the darkened triangles. Both constructions can be easily decomposed into a sequence of JOIN, MEET and PARALLEL operations, that start from the points 0, 1, x, y and one auxiliary point p not on l. In particular we get

$$\begin{aligned} x+y &\leftarrow \textsc{meet}(\textsc{join}(0,x),\\ &\qquad\quad \textsc{parallel}(\textsc{join}(x,p),\\ &\qquad\qquad\qquad \textsc{meet}(\textsc{parallel}(\textsc{join}(0,p),y),\\ &\qquad\qquad\qquad\qquad \textsc{parallel}(\textsc{join}(0,x),p)))),\\ x\cdot y &\leftarrow \textsc{meet}(\textsc{join}(0,x),\\ &\qquad\quad \textsc{parallel}(\textsc{join}(x,p),\\ &\qquad\qquad\qquad \textsc{meet}(\textsc{parallel}(\textsc{join}(1,p),y),\textsc{join}(0,p)))). \end{aligned}$$

These construction sequences are chosen with care such that as long as the auxiliary point p is not on l the only non-admissible situations arise when in the addition both points x and y are at infinity or in the multiplication one of the points is at 0 and the other is at infinity.

3.2.5 Complex Arithmetics

As well as calculations over the real numbers we can also do calculations over *complex numbers*. For this we fix points "0", "1" and "i" in the plane, such that the lines $l_{\mathbb{R}} \leftarrow \textsc{join}(0,1)$ and $l_{i\mathbb{R}} \leftarrow \textsc{join}(0,i)$ are perpendicular and such that the distance from 0 to 1 is the same as the distance from 0 to i. For convenience we take $0 \leftarrow P^{(0,0)}$, $1 \leftarrow P^{(1,0)}$, $i \leftarrow P^{(0,1)}$. The lines $l_{\mathbb{R}}$ and $l_{i\mathbb{R}}$ play the roles of the real and imaginary axes of the complex plane. The points 0 and 1 define a scale on $l_{\mathbb{R}}$. The points 0 and i define a scale on $l_{i\mathbb{R}}$. For each point p in the plane we can (after orthogonal projection to these two axes) assign two coordinates, the real and the imaginary part of a complex number $a+ib$. If no confusion can arise we simply denote the points with homogeneous coordinates $(a, b, 1)$ by the corresponding complex number $a+ib$. By a parallel projection along the direction of JOIN$(1, i)$ we can transfer any number in $l_{i\mathbb{R}}$ to the corresponding number on $l_{\mathbb{R}}$, and vice versa. Let $z_1 = a_1 + ib_1$ and $z_2 = a_2 + ib_2$ be two complex numbers. With respect to our coordinate system we can model complex addition and complex multiplication of the points z_1 and z_2 by first transferring the real and imaginary parts to the line $l_{\mathbb{R}}$, then modeling the formulas

$$\begin{aligned} z_1 + z_2 &= (a_1 + a_2) + i(b_1 + b_2), \\ z_1 \cdot z_2 &= (a_1a_2 - b_1b_2) + i(a_1b_2 + b_1a_2), \end{aligned}$$

by a sequence of von Staudt constructions and finally construct a new point from the resulting real and imaginary part. The complex addition can be used for *addition of vectors* as well.

Remark 3.1. One might think that using von Staudt constructions for vector addition is more than necessary. The simple construction sequence

$$z_1 + z_2 \leftarrow \text{MEET}(\text{PARALLEL}(0, z_1), \text{PARALLEL}(0, z_2))$$

seems to work as well. However, this construction has the disadvantage that it is non-admissible whenever 0, z_1 and z_2 are collinear. For the complexity issues that we consider later the actual length of these elementary operations is irrelevant as long as it is constant.

3.2.6 Integer and Rational Points

By being able to add and multiply via von Staudt constructions we are also able to construct points $a + ib$ for arbitrary integers a and b with respect to our coordinate system. We simply have to find a sequence of additions and multiplications that computes the numbers a and b starting from 0 and 1. In particular, using the binary representation any integer $n > 0$ can be constructed in $O(\log(n))$ construction steps.

It is also easy to construct numbers of the form $\frac{1}{2^n}$. The construction in Fig. 6 shows that this can be done in $O(n)$ steps.

3.3 Points on Circles and Intervals

In our relational instruction set JMB we do not have direct access to circles. However, by Thales' theorem we can freely generate points on circles that are given by two diameter points (see Fig. 7 left). Let a and b be the two endpoints of a diameter of the desired circle. We take a free point p and construct

$$q \leftarrow \text{MEET}(\text{JOIN}(a, p), \text{PERPENDICULAR}(\text{JOIN}(a, p), b)).$$

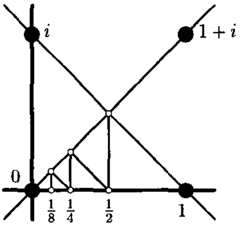


Fig. 6: Constructions for $\frac{1}{2^n}$.

Using Thales' theorem it is immediate that the point p is on the circle with the segment $[ab]$ as diameter. We will abbreviate this construction by

$$q \leftarrow \text{ONCIRCLE}(a, b, p).$$

If we furthermore project the resulting point orthogonally to the line a, b by

$$x \leftarrow \text{MEET}(\text{JOIN}(a, b), \text{PERPENDICULAR}(\text{JOIN}(a, b), \text{ONCIRCLE}(a, b, p))),$$

we get a point x that is constrained to lie in the closed segment from a to b (see Fig. 7 right). We abbreviate this by

$$x \leftarrow \text{ONINTERVAL}(a, b, p).$$

Only if p and a coincide these two operations are not admissible.

Remark 3.2. This construction has the side effect that while point p cycles once around point a, the derived point q makes *two* full cycles on the circle.

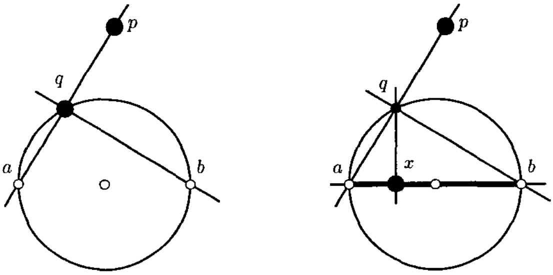


Fig. 7: Constructing points on circles and on segments.

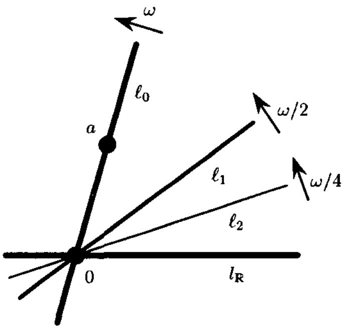


Fig. 8: Detection of a winding number.

Remark 3.3. Although by this construction we can generate a point freely on the boundary of a circle, this circle is not available for further constructions like intersecting it with a line.

3.4 Detecting a Winding Number

So far the constructions used in our gadgets did not contain any BISECT operations and therefore no non-determinism occurred. The basic functionality for which we will use BISECT operations is the generation of *monodromy effects*: "One can start with an instance A of a GSP and continuously make a round-trip with the free elements and end up in a different instance B." The smallest device for which such an effect occurs is given by the following GSP:

$$\begin{aligned} \ell_0 &\leftarrow \text{JOIN}(a, 0) \\ \ell_1 &\leftarrow \text{BISECT}(l_\mathbb{R}, \ell_0) \\ \ell_2 &\leftarrow \text{BISECT}(l_\mathbb{R}, \ell_1) \end{aligned}$$

It takes a free point a, joins it with the origin and constructs an angular "quad-sector" of this line and $l_\mathbb{R}$. Assume that a is in a certain position ($\neq 0$) and from there makes a round-trip with continuous speed around the origin. While ℓ_0 moves with an angular velocity ω the line ℓ_1 moves with angular velocity $\omega/2$ and the line ℓ_2 moves with angular velocity $\omega/4$. Thus after a has performed a full cycle around the origin the line ℓ_2 has made a quarter turn. If a moves along an arbitrary path (that avoids point 0) and returns to its original position then the resulting situation reflects the parity of the winding number of a around the origin. If ℓ_2 returned to its original position, the winding number was even, if ℓ_2 moved to the orthogonal of its original position the winding number was odd.

We can furthermore iterate this construction by adding more statements of the form

$$\ell_i \leftarrow \text{BISECT}(l_{\mathbb{R}}, \ell_{i-1})$$

for $i \in \{3, \dots, k\}$. The resulting line ℓ_k moves with angular velocity $\omega/2^k$. By this construction we can determine the winding number of a round-trip of a modulo the exponential number 2^{k-1}. If the line ℓ_k made a total turn of $i \cdot \pi/2^{k-1}$ the winding number w satisfies

$$w \equiv i \mod 2^{k-1}.$$

The situation for the first two iterations is shown in Fig. 8.

4 Reachability Problems

This chapter is dedicated to our first theorem. We will prove:

Theorem 4.1. *The following decision problem is NP-hard: Given a GSP $\mathcal{P}$ over the* JMB *instruction set that uses at most three* BISECT *operations. Furthermore, given two instances A and B of $\mathcal{P}$. Decide whether there is an admissible real path from A to B.*

We will prove this theorem by giving a reduction from the well known 3-SAT decision problem.

4.1 From 3-SAT to Algebra

The following problem is one of the standard NP-complete decision problems [2].

Decision Problem 4.2 (3-SAT). *Let $B = (b_1, \dots, b_n)$ be boolean variables, and let the literals over B be $\widetilde{B} = (b_1, \dots, b_n, \neg b_1, \dots, \neg b_n)$. Furthermore, let $C_1, \dots, C_k$ be clauses formed by disjunction of three literals from $\widetilde{B}$. Decide whether there is a truth assignment for B that satisfies all clauses $C_1, \dots, C_k$ simultaneously.*

W.l.o.g. we may assume that each variable occurs at most once in each clause. We first give a (polynomial time) procedure that transfers each instance of 3-SAT into a corresponding problem concerning the roots of a multivariate polynomial. Let $b_1, \dots, b_n$ be the boolean variables and let $C_1, \dots, C_k$ be the clauses of a given 3-SAT S. To each b_i we assign a formal variable x_i.

For a literal $l_i \in \{b_i, \neg b_i\}$ we set

$$f(x_i) := \begin{cases} x_i & \text{if } l_i = b_i, \\ 1 - x_i & \text{if } l_i = \neg b_i. \end{cases}$$

Assume that for each $j = 1, \dots, k$ the clause C_j is of the form $l_r^j \vee l_s^j \vee l_t^j$ where the literal l_i^j is either b_i or $\neg b_i$. We set

$$F_j := f(l_r^j) \cdot f(l_s^j) \cdot f(l_t^j).$$

Finally we set

$$F_S = \sum_{j=1}^{k} F_j.$$

By this translation for instance the 3-SAT formula $(b_1 \vee \neg b_3 \vee b_5) \wedge (\neg b_2 \vee b_4 \vee \neg b_5)$ is translated to $(x_1 \cdot (1 - x_3) \cdot x_5) + ((1 - x_2) \cdot x_4 \cdot (1 - x_5))$. The satisfying truth assignments for S and the roots of F_S in $[0,1]^n$ are related by the following lemma (here $[0,1]$ denotes the closed interval between 0 and 1).

Lemma 4.3. *S has a satisfying truth assignment if and only if there are $(x_1, \dots, x_n) \in [0,1]^n$ with $F_S(x_1, \dots, x_n) = 0$.*

Proof. If S has a satisfying truth assignment $(b_1, \dots, b_n) \in \{\text{TRUE}, \text{FALSE}\}^n$ we set

$$x_i := \begin{cases} 0 \text{ if } b_i = \text{TRUE}, \\ 1 \text{ if } b_i = \text{FALSE}. \end{cases}$$

Since every clause contains at least one true literal we the get that all $f_1, \dots, f_k$ are zero. This yields that F_S is zero as well. Conversely, assume that there are values $(x_1, \dots, x_n) \in [0,1]^n$ such that $F_S(x_1, \dots, x_n) = 0$. If the x_i are chosen in the interval $[0,1]$ all f_j are non-negative. Thus $\sum_{j=1}^{k} f_j = 0$ implies that all f_j are zero. However, each f_i can only be zero if at least one of its factors is zero. By setting

$$b_i := \begin{cases} \text{TRUE if } x_i = 0, \\ \text{FALSE if } x_i \neq 0, \end{cases}$$

we get a satisfying truth assignment for S. □

Using the structure of the polynomial $F_S(x_1, \dots, x_n)$ we can derive a simple gap theorem in the case that S is not satisfiable.

Lemma 4.4. *If S is not satisfiable then $F_S(x_1, \dots, x_n) \geq 1$ for all $(x_1, \dots, x_n) \in [0,1]^n$.*

Proof. This is true since F_S is a multilinear form and $F_S(x_1, \dots, x_n)$ is an integer that is greater or equal to zero for all vertices $(x_1, \dots, x_n) \in \{0,1\}^n$ of the unit cube. □

4.2 From Algebra to Geometry

Our next step is to transfer the algebraic situation in F_S to a geometric construction using exclusively JOIN and MEET operations and the constant points 0, 1, i, and $1+i$. This construction has the following properties: It contains freely movable points $p_1, \dots, p_n$ (one for each boolean variable in S), and a dependent point q that is constrained to lie on $l_{\mathbb{R}}$. There will be admissible positions for $p_1, \dots, p_n$ such that 0 and q coincide if and only if S is satisfiable.

Using the gadgets from Sec. 3 the construction is straightforward. We construct n points $x_1, \dots, x_n$ according to

$$x_i \leftarrow \text{ONINTERVAL}(0, 1, p_i).$$

The construction constrains each of the points x_i to the segment $[0,1]$ (see Sec. 3.3). Except for this there is no restriction to the positions of the points $x_1, \dots, x_n$. These points model the input variables $x_1, \dots, x_n$ of the equation F_S whose values should be chosen in the interval $[0,1]$.

Using von Staudt constructions we now encode the polynomial $F_S(x_1, \dots, x_n)$ geometrically. All calculations are carried out on the line $l_{\mathbb{R}}$. The point that finally represents the result of the calculation is called q. This point q lies on $l_{\mathbb{R}}$ by its construction and Lemma 4.3. It can coincide with 0 if and only if S was satisfiable. We call the whole construction C_S. Altogether we obtain:

Lemma 4.5. (i) *In C_S the point q lies on $l_{\mathbb{R}}$.*

(ii) *There is an admissible position for $p_1, \dots, p_n$ in C_S such that q and* 0 *coincide if and only if S has a satisfying truth assignment.*

(iii) *If S is not satisfiable then $q \geq 1$ for all admissible positions of $p_1, \dots, p_n$.*

Proof. (i) The point q lies on $l_{\mathbb{R}}$ by construction. (ii) is a consequence of the construction and Lemma 4.3. (iii) is a consequence of the construction and Lemma 4.4. □

4.3 A Geometric Combination Lock

Our final task for proving Thm. 4.1 is to transfer the construction C_S into a construction that can be used for proving NP-hardness of a reachability prob-

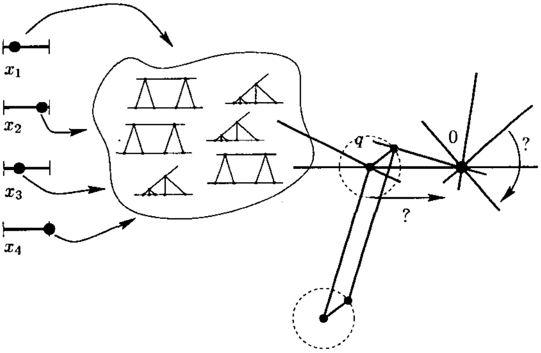


Fig. 9: Schematic view of the construction of a "geometric combination lock".

lem. The construction of $\mathcal{C}_S$ so far included only JOIN and MEET operations without non-determinism. The idea now is to conclude the construction by linking it to the "winding number gadget" presented in Sec. 3.4. We will do this in such a way such that a certain angular bisector can be rotated by $\pi/2$ if and only if the original 3-SAT problem was satisfiable.

For this we add a new free point p from which we construct a derived point v on the circle with diameter $[-\frac{1}{2} + 0i, \frac{1}{2} + 0i]$. The construction is standard using the point on circle gadget from Sec. 3.3:

$$v \leftarrow \text{ONCIRCLE}(-\frac{1}{2}, \frac{1}{2}, p).$$

Then we take the construction for $\mathcal{C}_S$ and use our gadget for complex addition to construct $w = q + v$ and the line $\ell_0 \leftarrow \text{JOIN}(0, w)$. Finally, we add three non-deterministic statements

$$\ell_1 \leftarrow \text{BISECT}(l_\mathbb{R}, \ell_0), \quad \ell_2 \leftarrow \text{BISECT}(l_\mathbb{R}, \ell_1), \quad \ell_3 \leftarrow \text{BISECT}(l_\mathbb{R}, \ell_2).$$

The line ℓ_3 is a three times iterated angular bisector of $l_\mathbb{R}$ and ℓ_0. By construction the lines ℓ_0, ℓ_1, and ℓ_2 pass through the origin. Hence the bisector operations are admissible. If the positions of $p, p_1, \ldots, p_n$ are fixed then the construction is completely determined up to the actual position of ℓ_1, ℓ_2, and ℓ_3. The final construction is called $\mathcal{R}_S$.

For the positions $\widehat{p} = \widehat{p_1} = \ldots = \widehat{p_n} = 1$ the line ℓ_0 coincides with $l_\mathbb{R}$ and the choice

$$\ell_1 = \ell_2 = \ell_3 = l_\mathbb{R}$$

is a proper instance A of $\mathcal{R}_S$. For these positions of $p, p_1, \ldots, p_n$ also the choice

$$\ell_1 = \ell_2 = l_{\mathbb{R}}, \quad \ell_3 = l_{i\mathbb{R}}$$

is a proper instance B. The only distinction between the instances A and B is the position of ℓ_3.

Lemma 4.6. *There is a continuous admissible path from A to B (induced by a movement of the free points $p, p_1, \ldots, p_n$) if and only if S is satisfiable.*

Proof. The only way to get from A to B is that the point $w = q + v$ turns an odd number of cycles around the origin. Assume for a moment that $p_1, \ldots, p_n$ are fixed. Then the point q has a certain position on the line $l_{\mathbb{R}}$. The point w is then constrained to lie on a circle of radius $\frac{1}{2}$ around q. By moving p we can freely influence the position of w on this circle. The only way to let w cycle around the origin is to move p to a position that has less than distance $\frac{1}{2}$ to the origin, and then move p to achieve a full cycle of w around the origin. However, Lemma 4.5 shows that q can only come so close to the origin if and only if S was satisfiable. This proves the claim. □

We may think of the whole construction as a "geometric combination lock:" The points $p_1, \ldots, p_n$ play the role of the code dials. The point p plays the role of an opening wheel. The angular bisector is the bolt of the combination lock. The reachability problem translates to the question whether one can open the lock. Initially the dials and the wheel are in some position. If we want to open the combination lock we first have to move the dials into the correct position (this can only be done if we know the solution to the 3-SAT problem S). If the dials are in the correct position we can turn the opening wheel and open the lock. After opening the lock we move all dials and the opening wheel again to the initial position. Nothing has changed except for the fact that the lock is now open.

A schematic picture of the whole situation is shown in Fig. 9. Points on an interval are used for von Staudt constructions. The result of this computation is used for the opening wheel.

Finally, observing that the whole translation from the original 3-SAT to the construction $\mathcal{R}_S$ can be carried out in polynomial (even linear) time in the length of the 3-SAT problem proves Thm. 4.1.

5 Computing a Specific Trace

The goal of this section is to prove our next main theorem. It describes the complexity of the basic situation in a Dynamic Geometry system: You pick a point and move it from one position to another. It will turn out to be NP-hard to decide whether a continuous evaluation of the situation ends up in a specific situation.

Theorem 5.1. *Given a GSP $\mathcal{P}$ over the* JMB *instruction set that contains exactly one free point p. Furthermore, given two instances A and B such that p is at position a in A and p is at position b in B. Let $p(t) : [0,1] \to [a,b]$ be a (straight) movement of p with $p(0) = a$ and $p(1) = b$. It is NP-hard to decide whether a continuous evaluation of $\mathcal{P}$ under this movement that starts at instance A ends up at the instance B.*

Here is an overview over the ingredients of our proof: First, we map the moving point p to the unit circle. Then we construct a set of polynomials $B_j(z)$ that correspond to the variables of a given 3-SAT problem in a way that all possible 0-1 combinations are represented by the values of the B_j on the unit circle. Finally, another polynomial $F_s(z)$ encodes the boolean formula of the 3-SAT problem and controls a point q, that will cycle around the origin. The winding number of this point can be used to read off the satisfiability of the 3-SAT.

A similar polynomial construction has been used by Plaisted in [17,18,19]. He used it to prove that it is NP-hard to decide for a sparse univariate polynomial whether it has a complex root of modulus 1. Our constructions differ from Plaisted's work by being more focused on evaluations of polynomials over the real numbers. One of the direct consequences of our construction is that it is NP-hard to decide whether a real polynomial encoded by a straight line program has a root over the real numbers (see Sec. 5.7). This fact can also be derived as a consequence of Plaisted's Theorem by a Moebius transformation argument. The alert reader will find out that we could have used the binary counter construction of Sec. 6 to prove Thm. 5.1, but the additional results for real polynomial roots (and some additional insight) would not have been possible then. We are convinced that the additional effort pays off very well.

5.1 A Point on the Unit Circle

We will start our construction with a little gadget that maps a certain line segment to a point on the unit-circle. For this we first use the point-on-circle gadget of Sec. 3.3. and set

$$w \leftarrow \textsc{oncircle}(-1, 1, p).$$

If p is located at 2 the point w is located at 1. While p moves on a straight vertical path to the point $2 + 3i$ point w makes a quarter turn on the unit circle. We set

$$z \leftarrow w^4.$$

This point is constructible by the gadgets for complex arithmetics from Sec. 3.2. While p moves along the segment from 2 to $2+3i$ the point $z =: z(p)$ makes exactly one full cycle on the unit circle.

5.2 Complex Polynomials for Variables

From now on we fix a specific instance S of a 3-SAT problem with variables $b_1, \dots, b_n$ and clauses $C_1, \dots, C_k$. We will encode S into an instance of the decision problem of Thm. 5.1.

Let P_j be the j-th prime number, and let $M = \prod_{j=1}^{n} P_j$ be the product of the first n primes. The size of the j-th prime is less than $j \log(j)$. Hence the size of M is less than $n^{n \log n}$. The polynomial $z^M - 1$ has altogether M single roots, the M-th roots of unity, equally spaced on the unit circle at $z = e^{2i\pi(r/M)}$, for $r \in \{1, 2, \dots, M\}$. We abbreviate $\epsilon_M(r) = e^{2i\pi(r/M)}$. We consider two classes of polynomials for which the sets of roots are subsets of the roots of $z^M - 1$:

$$\begin{aligned} B_j(z) &= 1 - z^{M/P_j}, \\ \overline{B}_j(z) &= 1 + z^{M/P_j} + z^{2M/P_j} + z^{3M/P_j} + \dots + z^{(P_j-1)M/P_j}. \end{aligned}$$

For each $j \in \{1, \dots, n\}$ we set $A_j = \{P_j, 2P_j, 3P_j, \dots, M\}$ and $\overline{A}_j = \{1, 2, \dots, M\} - A_j$. The following relations are immediate.

Lemma 5.2. *With the notation as set above we have:*

(i) *For each $j \in \{1, \dots, n\}$ we have $B_j(z) \cdot \overline{B}_j(z) = z^M - 1$.*

(ii) *The roots of $B_j(z)$ are at $z = \epsilon_M(r)$, for $r \in A_j$.*

(iii) *The roots of $\overline{B}_j(z)$ are at $z = \epsilon_M(r)$, for $r \in \overline{A}_j$.*

Proof. Claim (i) can be directly proved by expansion of the product. Claim (ii) is trivial. Claim (iii) is a consequence of (i) and (ii) and the fact that there are no multiple roots in $z^M - 1$. □

Later on we will associate to each number $r \in \{1, 2, \dots, M\}$ a certain evaluation of the boolean variables $b_1, \dots, b_n$. For a number r the boolean variable b_j will be considered TRUE if $\epsilon_M(r)$ is a root of B_j and FALSE otherwise. The above lemma proves that under this correspondence b_j is FALSE (at r) if and only if $\epsilon_M(r)$ is a root of $\overline{B}_j$. We set

$$b_j(r) := \begin{cases} \text{TRUE if } B_j(\epsilon_M(r)) = 0, \\ \text{FALSE if } B_j(\epsilon_M(r)) \neq 0. \end{cases}$$

Lemma 5.3. *For each truth assignment* $(b_1, \dots, b_n) \in \{\text{TRUE}, \text{FALSE}\}^n$ *there is at least one number* $r \in \{1, \dots, M\}$ *such that* $b_j(r) = b_j$ *for all* $j \in \{1, \dots, n\}$.

Proof. For a given assignment $(b_1, \dots, b_n) \in \{\text{TRUE}, \text{FALSE}\}^n$ we are looking for an integer $r \leq M$ that has P_j as a prime factor if and only if $b_j(r) = \text{TRUE}$. For this we can simply take the number

$$\prod_{\{j \mid b_j(r)=\text{TRUE}\}} P_j$$

which has this property. □

We will explicitly calculate the above polynomials $B_j(z)$, $\overline{B}_j(z)$ and $z^M - 1$ by suitable geometric constructions. We will have to take care that the effort of doing this is no more than polynomial in the coding length of the original 3-SAT S. For this we need:

Lemma 5.4. *For each* n *the polynomial* z^n *can be evaluated by a straight line program using at most* $O(\log_2(n))$ *multiplications. Similarly for each* n *and each divisor* m *of* n *the polynomial* $1 + z^m + z^{2m} + \dots + z^n$ *can be evaluated by a straight line program using at most* $O(\log(n)^2)$ *additions or multiplications.*

Proof. Let $n = \sigma_0 2^0 + \sigma_1 2^1 + \sigma_2 2^2 + \dots + \sigma_k 2^k$ with $k \leq \log_2(n)$ and $\sigma_i \in \{0, 1\}$ be the binary expansion of n. We can write $z^n = \prod_{\{i \mid \sigma_j = 1\}} z^{2^j}$. This product has at most $\log_2(n)$ terms. The polynomial $z^{2^j} = z^{2^{j-1}} \cdot z^{2^{j-1}}$ uses only one additional multiplication if we already have $z^{2^{j-1}}$. This proves the first claim.

For the second claim we first set $f_k(z) = 1 + z^1 + z^2 + \dots + z^k$. We have $1 + z^m + z^{2m} + \dots + z^n = f_{n/m}(z^m)$. Thus after having used $2\log_2(m)$ multiplications for computing z^m we just have to care about $f_k(z)$ for $k = n/m$. If k is even we get $f_k(z) = f_{k/2}(z)(1 + z^{k/2})$, if k is odd we get $f_k(z) = f_{k/2-1}(z)(1 + z^{k/2}) + z^{k/2}$, and a simple recursion on k proves the claim. □

The last lemma together with the observation that M is less than $n^{n \log n}$ shows that all polynomials $B_j(z)$, $\overline{B}_j(z)$ and $z^M - 1$ can be encoded by straight line programs whose length is polynomial in n.

5.3 Evaluating a 3-SAT

We now proceed by encoding the original 3-SAT instance S into our construction. For a complex number $z = a + ib$ we set $||z|| = a^2 + b^2$. For all $j \in \{1, \dots, n\}$ we consider $L_j(z) = ||B_j(z)||$ and $\overline{L}_j(z) = ||\overline{B}_j(z)||$. These two functions are *real-valued* and even non-negative. The only way for these functions to be zero is that the corresponding functions $B_j(z)$ and $\overline{B}_j(z)$ become zero. For a literal $l_j \in \{b_j, \neg b_j\}$ we set

$$f^{l_j}(z) := \begin{cases} L_j(z) \text{ if } l_j = b_j, \\ \overline{L}_j(z) \text{ if } l_i = \neg b_i. \end{cases}$$

Assume that for each $j = 1, \dots, k$ the clause C_j is of the form $l^j_r \vee l^j_s \vee l^j_t$ where the literal l^j_k is either b_k or $\neg b_k$. We set

$$F_j(z) := f^{l^j_r}(z) \cdot f^{l^j_s}(z) \cdot f^{l^j_t}(z).$$

Finally we set

$$F_S(z) = \sum_{j=1}^{k} F_j(z).$$

If $z = a + ib$ the function $F_S(z)$ is a real polynomial in a and b that can be realized by a straight line program with length polynomial in the size of the 3-SAT instance S. It is important that $F_S(z)$ is not an element of the polynomial ring $\mathbb{C}[Z]$, because otherwise the following lemma could not be true.

Lemma 5.5. *It is NP-hard to decide whether there is a $z \in \mathbb{C}$ with $F_S(z) = 0$.*

Proof. The only way for $F_S(z)$ to become zero is that all its summands are zero. This however can only be the case if z is of the form $\epsilon_M(r)$ for an r that corresponds to a satisfying truth assignment of S. □

Example 5.6. Let us consider a specific satisfiability problem S and the associated function F_S. In order for the example to have a reasonable size we consider a 2-SAT instead of an actual 3-SAT instance:

$$S = (b_1 \vee b_3) \wedge (b_2 \vee b_1) \wedge (\neg b_2 \vee b_3).$$

The satisfying truth assignments are $(1, 0, 0)$, $(1, 0, 1)$ and $(1, 1, 1)$. We associate b_1 with the prime number 2, b_2 with 3, and b_3 with 5. The resulting graph of $F_S(\cos(2\pi t/30) + i \sin(2\pi t/30))$ together with the corresponding bit patterns is shown in Fig. 10. The ticks mark the corresponding 30th roots of unity. Whenever we have a bit pattern corresponding to a satisfying assignment the function is zero, else it is greater than zero.

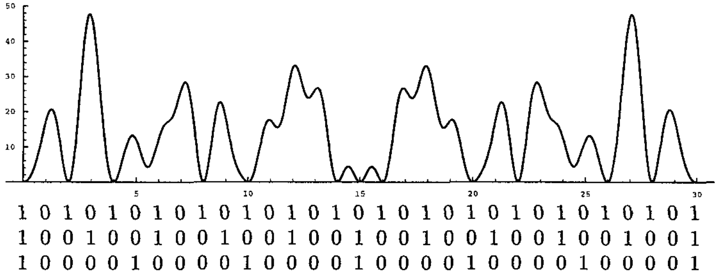


Fig. 10: The graph of the function $F_S(\cos(2\pi t/30) + i\sin(2\pi t/30))$.

5.4 Another Gap-Theorem

Again we need an estimate how small the function $F_S(z)$ can become if S is not satisfiable. In fact we will need this minimum only for points $\epsilon_M(r)$ with $r \in \{1, 2, \ldots, M\}$. We can get a lower bound for this value by calculating the smallest possible non-zero summand of $F_S(\epsilon_M(r))$. This value in turn can be bounded by the cube of the smallest non-zero value α that one of the functions $L_j(\epsilon_M(r))$, $\overline{L}_j(\epsilon_M(r))$ can take for $r \in \{1, 2, \ldots, M\}$. For this it is useful to observe that $L_j(\epsilon_M(r)) = 2 - 2\cos(2\pi \cdot r/P_j)$. This desired value α is taken at $L_n(\epsilon_M(1))$. Thus we have $\alpha = 2 - 2\cos(2\pi/P_n)$. We obtain the following lemma.

Lemma 5.7. *If $F_S(\epsilon_M(r))$ is non-zero for some $r \in \{1, \ldots, M\}$, then we have*

$$F_S(\epsilon_M(r)) \geq (2 - 2\cos(2\pi/P_n))^3 > (2 - 2\cos(2\pi 2^{-\log(n)^2})^3.$$

Proof. The first inequality follows from our considerations above. The second inequality is a very rough estimate following from the monotonicity of $\cos(t)$ in $[0, \pi/2]$ and the fact that $P_n < 2^{\log(n)^2}$. □

We set $\beta_S = (2 - 2\cos(2\pi \cdot 2^{-\log(n)^2}))^3$. This number can be constructed geometrically using an iterative sequence of $\log(n)^2$ BISECT operations starting with the right angle followed by a constant number of JOIN and MEET operations.

5.5 Tracing the Flight of a Bumble Bee

Now we are done with the algebraic part of our construction. We come back to the geometric part that started with the construction of a point z that moves once around the unit circle while the free point p moves from a to b (Sec. 5.1).

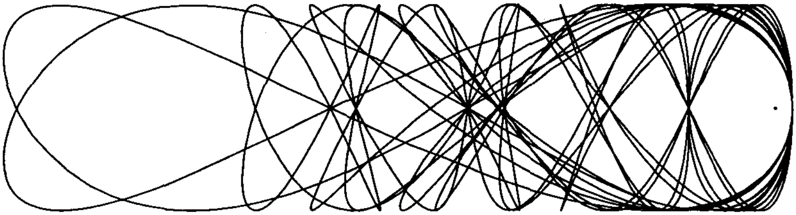


Fig. 11: The path of the dependent point $g_S(z)$.

We take z as the input of $F_S(z)$ and model the evaluation of $F_S(z)$ by von Staudt constructions as described in Sec. 3.2. The result is a non-negative point q on the real axis $l_{\mathbb{R}}$. This point can coincide with the origin whenever z corresponds to a satisfying truth assignment of S. In particular, this can only happen if $z = \epsilon_M(r)$ for a suitable r. We now consider the function

$$g_S(z) := z^M - 1 - F_S(z) + \beta_S.$$

Lemma 5.8. *Let G be the path of $g_S(z)$ while z makes one full cycle around the unit circle. The winding number of G with respect to the origin is zero if and only if S has no satisfying truth assignment.*

Proof. Let $z = e^{2i\pi t}$ for $t \in [0,1]$ move with constant speed on the unit circle. We can get the corresponding winding number around the origin in the following way: We take the ray of all positive real numbers. We calculate two numbers: w^+ counts how often the point $g_S(z)$ crosses this ray moving from the lower to the upper half plane, and w^- counts how often the point $g_S(z)$ crosses this ray moving from the upper to the lower half plane. The number $w^+ - w^-$ gives the winding number.

The real part of $g_S(z)$ is by construction and by Lemma 5.7 at most β_S. For $g_S(z)$ being real and positive the numbers $z^M - 1$ and $F_S(z)$ must both vanish. Thus we obtain

$$w^+ = |\{r \in \{1, \ldots, M\} \mid F_S(\epsilon_M(r)) = 0\}| \quad \textit{and} \quad w^- = 0.$$

The winding number counts exactly the number of possible $r \in \{0, \ldots, M\}$ for which $b_j(r)$ is a satisfying truth assignment of S. □

Fig. 11 shows the trace of $g_S(z)$ for the 2-SAT formula of Example 5.6. The origin lies on the symmetry axis of the figure very close to the right boundary. The winding number will be exactly 12 corresponding to the 12 zeros of $F_S(z)$ shown in Fig. 10.

5.6 NP-hardness of Tracing

Now we are in principle done. We do a geometric construction that calculates $g_S(z(p))$ using the free point p as input parameter. This can be done by a number of JOIN, MEET, and BISECT operations that is polynomial in the parameter n and k of S (by Lemma 5.4 and the considerations in Sec. 5.4). We construct the line $\ell_0 = \text{JOIN}(0, g_S(z(p)))$. Finally, we add n^3 non-deterministic statements:

$$\begin{aligned}\ell_1 &\leftarrow \text{BISECT}(l_{\mathbb{R}}, \ell_0)\\ \ell_2 &\leftarrow \text{BISECT}(l_{\mathbb{R}}, \ell_1)\\ &\vdots\\ \ell_{n^3} &\leftarrow \text{BISECT}(l_{\mathbb{R}}, \ell_{n^3-1})\end{aligned}$$

The winding number w of Lemma 5.8 satisfies

$$w < M < n^{n\log(n)} < n^{n^2} < (2^n)^{n^2} = 2^{n^3}.$$

Thus we obtain the following lemma.

Lemma 5.9. *Let A be an initial position of our entire construction for $p = 2$ and let B be the corresponding position with identical choice of the angular bisectors for $p = 2 + 3i$. B is the result of a continuous evaluation under the straight movement of p if and only if S was not satisfiable.*

Proof. If S is not satisfiable, then the winding number of $g_S(z(p))$ around the origin is zero and the position of the angular bisectors remains unchanged. If S is satisfiable, then the winding number lies between 1 and 2^{n^3}. Thus a movement of p causes a change of the positions of the angular bisectors. □

This concludes our proof of Thm. 5.1, which is a direct consequence of the above Lemma and the fact that the construction was polynomial in the size of n and k.

5.7 Roots of Univariate Polynomials

We will close this chapter with a little side remark. Assume that the free point p of our construction is parameterized by $(2, x, 1)$ (in homogeneous coordinates with $x \in \mathbb{R}$). The construction of the function $F_S(z(p))$ could be exclusively done with JOIN and MEET operations. The coordinates of the resulting point $F_S(z(p)) = (\alpha(x), \beta(x), \gamma(x))$ are then polynomials in the single variable x. These polynomials can be calculated by straight line programs whose length is polynomial in the size of S, by translating the GSP into an equivalent SLP (see [11]). The 3-SAT S is satisfiable if and only if there is an x with $\alpha(x) = 0$. Thus we obtain

Theorem 5.10. *It is NP-hard to decide whether a univariate polynomial encoded by a straight line program has a real root.*

6 PSPACE-Hard Problems

This section focuses on proving that certain reachability problems are PSPACE-hard do decide. Compared to the previous sections there is one important difference. Every construction done so far only needed *constant points*, *meet*, *join*, and *bisect* operations. For the proof of the following PSPACE-hardness results for real reachability we need one additional ingredient, a semialgebraic constraint on the configuration space of the geometric configuration. We will demonstrate several variants of the result with different such constraints:

- the condition that a certain point is always on the left of a certain line,
- the condition that a certain point is always inside a certain circle,
- the condition that the intersection of a line and a circle is always real,
- the condition that the total length of the path of a freely movable point stays below a certain threshold.

Note that the first three variants can be transformed into each other. In fact there are many other variants of the result since the necessary restrictions that come from the additional inequality can be formalized in a very weak way. Nevertheless we have not been able to derive a comparable result without the additional condition. Our proof will be entirely constructive and self contained. It just relies on the well known PSPACE hardness of *quantified boolean formulas.* Moreover we will be very restrictive in the use of free points: the final construction has only *one* free point.

Let us first formulate one of the natural version of the main result of this section which is very similar to Thm. 4.1, except for the additional inequality constraint (for which we choose an *incircle test* here).

Theorem 6.1. *The following decision problem is PSPACE-hard: Given a GSP $\mathcal{P}$ over the* JMB *instruction set that has exactly one free point and a certain dependent point d. Furthermore, given two instances A and B of $\mathcal{P}$. Decide whether there is an admissible real path from A to B, such that along the path we always have $|d| < 2$.*

The proof of this result will be done by a reduction to the PSPACE-hardness of *Quantified Boolean Formulas* (QBF). Formally the PSPACE-hardness of QBF can be stated as "it is PSPACE-hard to decide whether the formula

$$\forall x_1 \exists y_1 \forall x_2 \exists y_2 \ldots \forall x_n \exists y_n f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$$

is true, where f is a boolean expression." In a sense this formula resembles a two player game with players X and Y. It asks for a winning strategy for player Y. The formula f encodes the winning positions of Y: "For each move x_1 of X there is a move y_1 for Y such that for each move x_2 of X there is a move y_2 for Y such that ... such that there is a final move y_n for Y such that Y wins the game.

For the proof we will geometrically construct a binary counter that counts through all possibilities for the $x_1, \ldots, x_n$. The entire construction is such that in order to get from position A to B in the reachability problem one has (at certain positions) to *set* the values of the $y_1, \ldots, y_n$ properly which can only be done if one knows the complete strategy for player Y.

Remark 6.2. Before we start with the proof let us contemplate for a moment the value of the following constructions. It is a remarkable fact that a similar result can be obtained even without using the BISECT operation at all – however for the price of an unbounded number of free points. The idea for this is to use one of the well known PSPACE-hard semialgebraic reachability problems (like the warehouseman's problem [7,24]) as the starting point of the reduction. All involved equations and inequalities can be condensed into one big inequality (this new inequality describes an ε-approximation of the original problem). This translation can only be done with the help of additional slack-variables (one variable per original inequality). The information of a certain state of the construction is "stored" in the actual values of the slack variables. Particular technical difficulties arise from the right choice of the involved ε-sizes.

Compared to this approach the construction presented on the following pages is much more direct. Its "computational power" is more or less distributed among the monodromy behavior of several BISECT operations. Each of these angular bisectors contributes one bit of information to the "storage" of the device.

Our construction allows for variants and extensions that are not possible in the other approach. In particular, the results can be strengthened further to have only *one* free complex input variable. A streamlined variant of this result for the case of analytic continuation will be presented in [14].

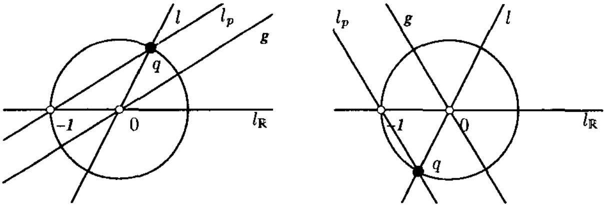


Fig. 12: Two instances of an INT_UNITCIRCLE(l) gadget.

6.1 Another Gadget

Before starting the crucial construction of a binary counter we will introduce another gadget that simplifies this construction. We show that over the JMB construction set, we can intersect a line through the origin with the unit circle (note that there are no circles available in JMB). Let l be a line through the origin and $l_\mathbb{R}$ be the real axis. We consider the following GSP:

$$\begin{aligned} g &\leftarrow \text{BISECT}(l, l_\mathbb{R}) \\ q &\leftarrow \text{MEET}(l, \text{PARALLEL}(g, -1)) \end{aligned}$$

The output point q is one of the intersections of l and the unit circle. Which of the two intersections we get, depends on the choice of angular bisector. If the line l makes a half turn (and by this comes back to its original position), the intersection moves continuously from one possibility to the other, see Fig. 12. We will encapsulate this construction within a (non-deterministic) macro INT_UNITCIRCLE(l) that produces one of the two intersections.

6.2 A Binary Counter

Our first sub-goal is to construct a binary counter, that drives the construction through an exponential number of different stages. For this we again start with a point z, which is given by

$$\begin{aligned} w &\leftarrow \text{ONCIRCLE}(-1, 1, p), \\ z &\leftarrow w^4. \end{aligned}$$

As described in Sec. 5.1, while p moves on a straight vertical path from $a = 2$ to the point $b = 2+3i$, the point z makes one full counterclockwise cycle on the unit circle, starting and ending at 1. W.l.o.g. for our considerations we may assume that p is bound to lie on the line that connects a and b.

We will consider the point z directly as a driving input point that is bound to lie on the unit circle. We now have a look at the following functions:

$$\begin{array}{lll} z_1 & \leftarrow & \mathrm{Re}(z^2) + z^{2^n} - 2 \\ z_2 & \leftarrow & \mathrm{Re}(z^4) + z^{2^n} - 2 \\ z_3 & \leftarrow & \mathrm{Re}(z^8) + z^{2^n} - 2 \\ \vdots & & \vdots \qquad\quad \vdots \\ z_n & \leftarrow & \mathrm{Re}(z^{2^n}) + z^{2^n} - 2 \end{array}$$

The Re(...) operation (that extracts the real part of complex number) can be carried out geometrically by a projection to the real axis. All the $z_1, \ldots, z_n$ can be constructed by a GSP whose total number of construction steps is linear in n. No angular bisectors are needed.

All the real parts of the functions $z_1, \ldots, z_n$ are in the interval $[-2, 0]$, since the numbers z^k all lie on the unit circle. While z moves along the unit circle the function z_k is zero exactly if z is a 2^k-th root of unity. Each of the functions z_k is real iff z is a 2^{n+1}-st root of unity. For $j = 0, \ldots, 2^n$ we set

$$a_j = e^{2i\pi(j/2^n)} \quad \text{and} \quad b_j = e^{2i\pi(j/2^n + 1/2^{n+1})}.$$

We will refer to regions on the unit circle by the term *circular interval.* For two points a, b on the unit circle the open circular interval (a, b) is the (open) arc on the unit circle that arises by traveling counterclockwise from a to b.

The imaginary parts of the functions z_k only depend on the function z^{2^n}. $\mathrm{Im}(z_k)$ is positive in the circular intervals (a_j, b_j) and negative in the circular intervals (b_j, a_{j+1}) for $j = 0, \ldots, 2^n$. The largest purely real value that occurs for the function z_k is 0, the second largest real value is $\cos(2\pi/2^k) - 1$. We set $\varepsilon = -\cos(2\pi/2^{n+1}) + 1$, a number that can be constructed geometrically by n successive BISECT operations of a right angle, followed by a projection.

Now let $l_\mathbb{R}$ be the real axis and we add for each $k = 1, \ldots, n$ the following instructions to our GSP:

$$\begin{array}{lll} l_k & \leftarrow & \mathrm{BISECT}(\mathrm{JOIN}(0, z_k + \varepsilon), l_\mathbb{R}) \\ q_k & \leftarrow & \mathrm{INT\_UNITCIRCLE}(l_k) \end{array}$$

First observe that the operation $\mathrm{JOIN}(0, z_k + \varepsilon)$ is always admissible, since $z_k + \varepsilon$ is never 0. The line $\mathrm{JOIN}(0, z_k + \varepsilon)$ is identical to $l_\mathbb{R}$ whenever z is either a_j or b_j for some $i \in \{0, \ldots, 2^n\}$. Thus at these places the line l_k can either be the real or the imaginary axis. Note that the only freely movable point in the whole construction so far is z (indirectly controlled by p). As a starting instance of the GSP we set $z = 1$. All lines $\mathrm{JOIN}(0, z_k + \varepsilon)$ are then identical to $l_\mathbb{R}$. We get an admissible instance of the above operations by setting all $l_k = l_\mathbb{R}$ and setting all $q_k = 1$.

From this admissible starting instance A the behavior of the entire construction is determined (by analytic continuation) while z travels along the unit circle.

Lemma 6.3. *With all settings as above (starting at A) the values of the q_k are uniquely determined for all z (with $|z| = 1$). In particular, we have that for all $j \in \{0, \dots, 2^n - 1\}$ the value $z = b_j$ implies that*

$$q_k = \begin{cases} i & \textit{for} \quad \sigma_{n-k} = 0, \\ -i & \textit{for} \quad \sigma_{n-k} \neq 0. \end{cases}$$

Here $j = \sigma_0 2^0 + \sigma_1 2^1 + \sigma_2 2^2 + \cdots + \sigma_{n-1} 2^{n-1}$ with $\sigma_k \in \{0, 1\}$ is the binary expansion.

Proof. For the proof let us investigate what happens during a full counterclockwise cycle of the driving point z. At the beginning ($z = 1 = a_0$) all the l_k are by definition of A identical to the real axis and all q_k are by definition equal to 1. Furthermore, all z_k are positive (namely equal to ε). While z travels from a_0 to b_0 all z_k move through the upper halfplane to a negative value. Thus all lines JOIN$(0, z_k + \varepsilon)$ make a counterclockwise half turn. Consequently all l_k make a counterclockwise quarter turn and all q_k move to the value i as stated in the theorem.

Now let us investigate what happens when z moves from position b_j on the shortest possible path via a_{j+1} to position b_{j+1}. During such a move the position of q_k will make a half turn if and only if z_k makes a cycle around the origin. This in turn exactly happens if for $z = a_{j+1}$ the value of z_k is positive. However, this is only the case if a_{j+1} is a 2^k-th root of unity. This gives exactly the desired counting behavior. Finally after z has completed one full cycle, all elements are back to their initial positions. Hence no global monodromy occurs and the behavior is globally determined. □

The whole construction behaves like a binary counter. For each position $z = b_j$ all q_k are either i or $-i$. The positions exactly resemble the behavior of the binary expansion of j with i playing the role of the 0 and $-i$ playing the role of the 1.

6.3 A Register

The output of the counter we constructed so far will later on play the role of the x_k that occur in the quantified boolean formula

$$\forall x_1 \exists y_1 \forall x_2 \exists y_2 \dots \forall x_n \exists y_n f(x_1, y_1, x_2, y_2, \dots, x_n, y_n).$$

We now explain how to model the y_k. For this we construct a *register* with dependent points $r_1, \ldots, r_n$. Whenever the driving point z is at a position b_j the r_k will either be i or $-i$. However, which of the two values will be taken depends on the position of certain free points $p_1, \ldots, p_n$ during z being at a position a_j. We will call the points b_j *evaluation points* and call the points a_j *setting points.* For each $k \in \{1, \ldots, n\}$ we add the following four lines to our GSP constructed so far:

$$\begin{aligned}
p'_k &\leftarrow \text{ONINTERVAL}(0, 1, p_k)\\
z'_k &\leftarrow p'_k - 1 + z_k + \varepsilon\\
l'_k &\leftarrow \text{BISECT}(\text{JOIN}(0, z'_k), l_{\mathbb{R}})\\
r_k &\leftarrow \text{INT\_UNITCIRCLE}(l_k)
\end{aligned}$$

We extend our initial position A (remember there we had $z = 1$) first by setting $p_1 = p_2 = \ldots = p_n = 2$. This forces the lines $\text{JOIN}(0, z'_k)$ to be $l_{\mathbb{R}}$. As we did for the counter we set all $l'_k = l_{\mathbb{R}}$ and all $r_k = 1$. The following Lemma summarizes the properties of the register.

Lemma 6.4. *For any admissible move of the free input points $p, p_1, \ldots, p_n$ starting from instance A we have the following properties:*

(i) *Whenever z is at an evaluation point each of the r_k is either i or $-i$.*

(ii) *For each $k \in \{1, \ldots, n\}$ and each $j \in \{0, \ldots, 2^k - 1\}$ we have: Whenever z stays in the circular interval $I_{k,j} = (a_{j \cdot 2^{n-k}}, a_{(j+1) \cdot 2^{n-k}})$, all values of r_k when z is at an evaluation point in $I_{k,j}$ are identical.*

(iii) *Except conditions* (i) *and* (ii) *there are no other restrictions to the values of r_k when z is at an evaluation point.*

Proof. The proof is very similar to the proof of Lemma 6.3. The initial situation of A ensures that condition (i) is satisfied. For each $j \in \{0, \ldots, 2^k - 1\}$ the function z'_k is always negative within the entire circular interval $I_{k,j}$. Hence, as long as z does not leave this interval for all evaluation points the values of the r_k must be identical (since z'_k cannot cycle around the origin.) This proves (ii). The function z'_k *can* be positive whenever z takes one of the values $a_{j \cdot 2^{n-k}}$ with $j \in \{0, \ldots, 2^k - 1\}$ (these are the 2^k-th roots of unity). Whether it actually *is* positive depends on the position of the free point p_k. Thus whenever z makes a transition between the intervals $I_{k,j}$ and $I_{k,j+1}$ (in either direction) we can (by moving p_k accordingly) control whether the r_k at z being at an evaluation point in $I_{k,j}$ and $I_{k,j+1}$ are identical for both intervals or not. This proves (iii). □

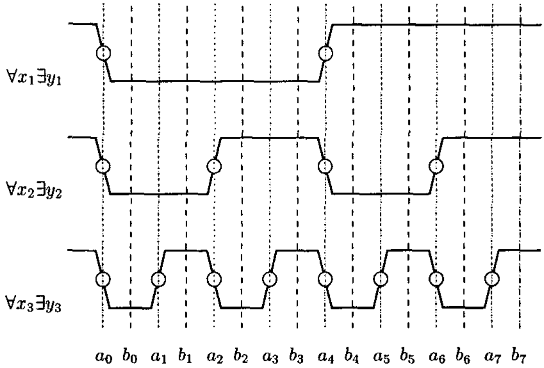

$\forall x_1 \exists y_1$
$\forall x_2 \exists y_2$
$\forall x_3 \exists y_3$
a_0 b_0 a_1 b_1 a_2 b_2 a_3 b_3 a_4 b_4 a_5 b_5 a_6 b_6 a_7 b_7

Fig. 13: The timing of the counter and the register

6.4 Encoding the QBF

Let us now step back and see what we have achieved so far. We have constructed a GSP together with an initial position A. The free points of the GSP are $p, p_1, , \ldots, p_n$. We have output points $q_1, \ldots, q_n$ for the counter and $r_1, \ldots, r_n$ for the register. Whenever z is at an evaluation point all these output points are either i or $-i$. For $n = 3$ the "timing" of the whole construction is shown in Fig. 13. The horizontal axis shows the positions for z on the unit circle. Each of the oscillating curves roughly represents the values of the counting points q_1, q_2, q_3. The dots mark those positions that are relevant for possibly changing the values of the register points r_1, r_2, r_3. The bottom row represents the point pair (q_3, r_3). The middle row represents (q_2, r_2), and the top row represents (q_1, r_1). Between two of the positions marked with a dot the corresponding r-point always has the same value at the evaluation points. If for instance z passes a_0 in clockwise direction the position of p_1 determines which values r_1 can take at the evaluation points b_0, b_1, b_2, b_3. These values can only be changed if z passes once more one of the setting points a_0 or a_4.

We are now going to link a given QBF formula to our construction. For this consider the QBF

$$\forall x_1 \exists y_1 \forall x_2 \exists y_2 \ldots \forall x_n \exists y_n f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$$

where we assume that f is given in conjunctive normal form. As in Sec. 4.1 we first translate the boolean formula f into a polynomial F_f by replacing each positive literal x_k by a real variable x_k and each negative literal $\neg x_k$

by $(1 - x_k)$. We do so similarly for the y_k. Then each *and*-operation is replaced by a multiplication and each *or*-operation is replaced by an addition. The resulting polynomial is called $F_f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$. Similar to Lemma 4.4 we obtain

Lemma 6.5. *We have $F_f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n) = 0$ for some $(x_1, y_1, \ldots, x_n, y_n) \in [0,1]^{2n}$ if and only if all variables are either 0 or 1, and $f(b(x_1), b(y_1), \ldots, b(x_n), b(y_n)) = \text{TRUE}$ with $b(0) = \text{TRUE}$ and $b(1) = \text{FALSE}$. For all other choices of the variables in $[0,1]^{2n}$ the value of F_f is strictly positive. Furthermore, at each corner of the cube $[0,1]^{2n}$ the polynomial F_f evaluates to an integer number.*

Proof. The proof is straightforward. It follows exactly the considerations in Sec. 4.1. □

We now (constructively) identify the values $-i$ and i (of the q_k and r_k) with the boolean values TRUE and FALSE, respectively by adding the statements

$$x_k \leftarrow \tfrac{1}{2}(iq_k + 1),$$
$$y_k \leftarrow \tfrac{1}{2}(ir_k + 1)$$

for $k = 1, \ldots, n$ to our GSP. Furthermore, we set with these newly constructed points

$$F \leftarrow F_f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n).$$

The next lemma brings us close to the complexity result we are aiming for. Recall that the point p was w.l.o.g bound to the line connecting 2 and $2 + 3i$, and that while p moves straight from 2 to $2 + 3i$ the path of the point z is a full clockwise cycle on the unit circle. by

Lemma 6.6. *The QBF $\forall x_1 \exists y_1 \forall x_2 \exists y_2 \ldots \forall x_n \exists y_n f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ is true if and only if the following holds: In our GSP starting with instance A there is an admissible path that moves point p from 2 to $2 + 3i$ such that the value of F is 0 whenever z is at an evaluation point.*

Proof. First assume that $\forall x_1 \exists y_1 \forall x_2 \exists y_2 \ldots \forall x_n \exists y_n f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ is true. We can skolemize the variables y_k of the $\exists$-quantors by introducing Skolem functions

$$s_1(x_1), s_2(x_1, x_2), \ldots, s_n(x_1, \ldots, x_n)$$

such that

$$f(x_1, s_1(x_1), x_2, s_2(x_1, x_2), \ldots, x_n, s_n(x_1, \ldots, x_n))$$

is a tautology. These Skolem functions are exactly the "strategy" of the two player game associated to the QBF, which tell player Y how to move. We can derive a path as desired in the theorem as follows. We first choose δ to be a sufficiently small positive number such that for $p = 2 - \delta i$ the point z lies in the circular interval (b_{2^n-1}, a_0) and move point p straight from 2 to $2 - \delta i$ while leaving the p_k unchanged.

After that we move point p straight from $2 - \delta i$ to $2 + 3i$. Whenever z passes one of the setting points we make sure that the positions of the points $p_1, \dots, p_n$ were adjusted such that for the forthcoming evaluation point the y_k take the values of the corresponding Skolem functions (this is possible by Lemma 6.4). The construction allows exactly enough freedom to possibly change the value of s_k whenever the variable x_k changes its value. By this choice and by Lemma 6.5 the derived point F is 0 at each evaluation point.

Conversely assume that we know how to move the points p_k such that point p can move from 2 to $2+3i$ in a way that whenever z is at an evaluation point the dependent point F is 0. We call such a path *correct*. We show how to reconstruct the Skolem functions from the situations at the evaluation points. One technical difficulty arises from the fact that it may happen that p changes its moving direction while traveling from 2 to $2+3i$ arbitrarily often. If this is the case the point z possibly meets several of the points a_j or b_j more than once.

Assume that the movement of $p = 2 + \phi(t)i$ is given by the function $\phi(t) : [0,1] \to \mathbb{R}$ with $\phi(0) = 0$ and $\phi(1) = 3$. The induced movement of z will be denoted by $z(t)$. Furthermore, we set $b(i) = \text{FALSE}$ and $b(-i) = \text{TRUE}$. Let us abbreviate TRUE by T and FALSE by F. We describe step by step how to derive the functions $s(x_1), s(x_1, x_2), \dots$.

Since $z(t)$ (continuously) describes one full cycle there is an (open) interval $I_1^{\mathrm{F}} \subset (0,1)$ such that $z(I_1^{\mathrm{F}})$ equals the (open) circular interval (a_0, a_{2^n-1}). Lemma 6.5(ii) tells us, that within this interval the value of point r_1 must be the same for all evaluation points. Lemma 6.5(i) shows that this value r_1^{F} must be either i or $-i$. We set $s_1(\mathrm{F}) = b(r_1^{\mathrm{F}})$. Similarly there is an open interval $I_1^{\mathrm{T}} \subset (0,1)$ such that $z(I_1^{\mathrm{T}}) = (a_{2^n-1}, a_{2^n})$. Also there the value r_1^{T} at the evaluation points is uniquely either i or $-i$. We set $s_1(\mathrm{T}) = b(r_1^{\mathrm{T}})$.

We now proceed inductively. Within the interval I_1^{F} there are subintervals $I_2^{\mathrm{F,F}}$ and $I_2^{\mathrm{F,T}}$ such that $z(I_2^{\mathrm{F,F}}) = (a_0, a_{2^n-2})$ and $z(I_2^{\mathrm{F,F}}) = (a_{2^n-2}, a_{2^n-1})$. We let $s_2(\mathrm{F},\mathrm{F}) = b(r_2^{\mathrm{F,F}})$ and $s_2(\mathrm{F},\mathrm{T}) = b(r_2^{\mathrm{F,T}})$. Similarly we define the values $s_2(\mathrm{T},\mathrm{F})$ and $s_2(\mathrm{T},\mathrm{T})$ by considering suitable subintervals of I_1^{T}. The values of the Skolem functions $s_3, \dots, s_n$ are defined similarly each one by looking at a suitable subintervals of the intervals considered for the previous function. Now, by our initial assumption the value of F was 0 at each evaluation point.

Lemma 6.5 thus ensures that

$$f(x_1, s_1(x_1), x_2, s_2(x_1, x_2), \dots, x_n, s_n(x_1, \dots, x_n))$$

is a tautology. Hence the original QBF was true. □

In the current construction the number of free points depends on the problem size. However, the character of the construction allows for an easy alternative that has just one free input point. We can strengthen Lemma 6.6 to the following version.

Lemma 6.7. *The GSP of Lemma 6.6 can be assumed to have only one free point p.*

Proof. For this note that in order to get a correct path we have to set the points $p'_1, \dots, p'_n$ to suitable values in $\{0, 1\}$ whenever z is at a setting point. We construct a second counter controlled by a free point p' on the segment from 2 to $2 + 3i$. This new counter is just an identical copy of the decvice described in Sec. 6.2. We connect the outputs $q'_1, \dots, q'_n$ of this new counter directly to the p'_k by setting (i.e. redefining)

$$p'_k \leftarrow \mathrm{Re}((iq'_k + 1)/2).$$

Now the position of the p'_k can be controlled by the position of p'. In particular every 0/1-combination of the p'_k can be achieved by placing p' to a suitable position. This construction still behaves like the construction of Lemma 6.6 but it has only two input points p and p'. We now "rename" our input point p to p'', then we introduce a new free point p, and add the following instructions $p'' \leftarrow 2 + \mathrm{Im}(p)$ and $p' \leftarrow 2 + \mathrm{Re}(p - 2) \cdot i$ to our GSP. This controls the points p' and p'' by the x and y parameter of just one single input point p. Still all necessary freedom that makes Lemma 6.6 work is maintained. □

6.5 The Inequality Condition

Our final task is now to transfer the "existence of a correct path"-property of Lemma 6.7 to a suitable geometric condition like the incircle condition in Thm. 6.1. For this we add the following statements to our GSP:

$$\begin{aligned} G &\leftarrow F + z^{2^n} + 1/2 \\ z' &\leftarrow \textsc{int\_unitcircle}(\textsc{join}(G, 0)) \\ d &\leftarrow z_n - z' \end{aligned}$$

We resolve the nondeterminisms by setting $h = l_{\mathbb{R}}$ at our initial position A and $z' = 1$.

Lemma 6.8. *In our GSP starting with instance A consider an admissible path that moves point p from 2 to $2 + 3i$. For such a path the following conditions are equivalent:*

(i) *If z is at an evaluation point the value of F is 0.*

(ii) *We have $|d| < 2$ throughout the path.*

Proof. Since F is positive and stays on the real axis the value of G is real if and only if z^{2^n} is real. This is exactly the case if z is either at a setting point a_j (then $z^{2^n} = 1$) or at an evaluation point b_j (then $z^{2^n} = -1$). This implies that G is positive at all setting points. Remember that F assumes only integer values at the evaluation points. Hence G is negative at an evaluation point if and only if $F = 0$ at this evaluation point. Point z' is the image of this point G mapped by a central projection to the unit circle.

Now assume that the path has the property as claimed in (i). We prove that this path automatically satisfies (ii). In this path $z' = 1$ at each setting point, $z' = -1$ at each evaluation point. Furthermore, $\mathrm{sign}(\mathrm{Im}(z')) = \mathrm{sign}(\mathrm{Im}(z^{2^n}))$. Hence there is always a line through the origin that has z' and z^{2^n} on the same side and therefore we have $|d| < 2$ throughout the path.

Conversely, assume that there is no path that satisfies condition (i). This means that for every possible path along which p is moved from 2 to $2 + 3i$ the total winding number of z' with respect to the origin is less than 2^n. This is the case since z' can cross the negative half line only for all the evaluation points. On the other hand the total winding number of z^{2^n} with respect to the origin is 2^n. This implies that there is at least one position along the path where z and z^{2^n} are antipodals on the circle. At this position we have $d = 2$. □

Now we obtain Thm. 6.1 as a direct consequence of the PSPACE-hardness of QBF, Lemma 6.5 and Lemma 6.6. This finishes the proof of Thm. 6.1.
Remark 6.9. Without formal proof we mention a few possible alterations of Thm. 6.1.

- *Non-Strict inequalities:* The construction we gave really needed a *strict* inequality as additional obstruction (two points on the unit circle cannot be further apart than 2 units). It is easy to obtain the same result also with a non strict inequality. For instance one could introduce an additional BISECT operation of JOIN$(G, 0)$ and $l_{\mathbb{R}}$ intersect the result with the unit circle and compare the resulting point with $z^{2^{n-1}}$.

- *Sidedness vs. incircle test:* One can turn an incircle condition $|d| < 1$ as used in Thm. 6.1 into a sidedness test of a point w.r.t. a line. One possibility for this is to construct the point

$$d' \leftarrow (\text{MEET}(\text{PERPENDICULAR}(\text{JOIN}(0,d),d),l_{\mathbb{R}}))^2.$$

The incircle condition then translates to $d' < 1$.

- *Admissibility of circle-line intersections:* We may also introduce the operation of intersecting the unit circle with a line. We may restrict the range of admissible situations to those where such an intersection properly exists. By this we can also express the inequality conditions that are needed in Thm. 6.1.

- *Restriction on total length of the path of p:* One can also perturb Thm. 6.1 in a way such that the inequality becomes a restriction on the total length of the path of p in a reachability problem. For this we first replace our assignment $p' \leftarrow 2+\text{Re}(p-2)i$ by $p' \leftarrow 2+N\cdot\text{Re}(p-2)i$ for a very large number N. This operation rescales the imaginary part of p such that the control of the p_i' does not really contribute significantly to the overall length of the path p takes. Then we take the $(n+1)$-times iterated angular bisector h_{n+1} of $\text{JOIN}(G,0)$ and $l_{\mathbb{R}}$ and ask the following reachability problem. "Is it possibly to move p from 2 to $2+3i$ such that h_{n+1} make a 90°-turn into its other alternative such that the total length of the path described by p does not exceed $3+\delta$ (for sufficiently small $\delta > 0$)?" The only way to do this is to move straight from 2 to $2+3i$ by passing every setting point at most once. At each evaluation point the function F must be 0 in order to end up with a rotation of the desired amount. This is by Lemma 6.5 equivalent to knowing the Skolem functions for the QBF. This variant is particularly important, since then no additional sidedness or incircle test is needed.

- *Games against external forces:* The last statement can reformulated also in another way. Redefine p' and p'' as free points again. Assume that an exterior force moves p'' from $2-\delta i$ to $2+3i$. Can you simultaneously move the points p' such that h_{n+1} makes a 90°-turn? This is PSPACE-hard to decide.

7 Undecidable Problems

In this section we enlarge the set of our possible primitive operations. We add one non-deterministic operation that models the mechanical behavior of a *wheel* that rolls along a road. In principle wheels have the ability to transfer angles to distances. If the wheel was rotated for a certain angle it has traveled along the road for a certain distance. If we (as usual) denote angles modulo 2π this introduces a new kind of monodromy behavior to our context. For the same angle as input there is an infinity of possible output values. All these output lengths are integer multiples of the length that is generated by rotating the wheel once by 2π.

This new type of monodromy introduces a drastic change in the complexity behavior of the reachability problem. We will see that we can translate the solvability of Diophantine equations into a reachability problem for a construction involving several wheels. By the undecidability of Hibert's 10th problem this induces the undecidability for this reachability problem.

7.1 Wheels

Let us first formalize the concept of a wheel to fit into our setup of geometric straight line programs. The right algebraic function that models the behavior of wheels is the logarithm function applied to points on the unit circle. Our WHEEL-primitive will take a point $p = r \cdot e^{i\varphi} \neq 0$ and map it non-deterministically to a point on the real axis that represents the possible angles. We define the relation

$$\text{WHEEL} := \{(p, q) \mid q = \varphi/2\pi; q = r \cdot e^{i\varphi}\} \subset P \times P.$$

As usual we allow ourselves to write $q \leftarrow \text{WHEEL}(p)$ if $(p, q) \in \text{WHEEL}$. The operation is not admissible for $p = 0$. If $(p, q) \in \text{WHEEL}$ then we have also $(p, q + k) \in \text{WHEEL}$ for all $k \in \mathbb{Z}$. In particular we have $(1, k) \in \text{WHEEL}$ for all $k \in \mathbb{Z}$. If we start with an admissible instance $p = 1; q = 0$ of $q \leftarrow \text{WHEEL}(p)$ then we can continuously reach the situation $p = 1; q = k$ by letting p spin around the origin k times. In our picture of an actual mechanical wheel the operation WHEEL is designed to model the properties of a wheel of circumference 1 (and thus of irrational (!) radius $\frac{1}{2\pi}$). The resulting RIS is called JMBW.

7.2 Diophantine Equations

The following theorem states one version of the famous undecidability of Diophantine equations.

Theorem 7.1. *Let $N = 11$ and $f \in \mathbb{Z}[x_1, \ldots, x_N]$ be a polynomial. It is algorithmically undecidable whether f has a zero in $\mathbb{Z}^N$.*

The number N depends on the actual state of research in the area around Hilbert's 10th problem [31,32]. We keep it fixed for the following considerations.

We will prove the following theorem by reduction to the above statement:

Theorem 7.2. *Let $\mathcal{P}$ be a GSP over the* JMBW *instruction set with at least N* WHEEL*-operations and two* BISECT*-operations. Let A and B be two admissible instances of $\mathcal{P}$. It is undecidable whether there is an admissible real path from A to B.*

Proof. Step by step we will construct the translation from the polynomial in Thm. 7.1 to the GSP in Thm 7.2. We start with a free point p and a point z given by

$$z \leftarrow \text{ONCIRCLE}(-1, 1, p).$$

Then we add for $i = 1, \dots, N$ the instructions:

$$\begin{aligned} q_i &\leftarrow \text{ONINTERVAL}(0, 2, p_i) \\ x_i &\leftarrow \text{WHEEL}(q_i + z) \end{aligned}$$

Assume that in the instances A and B we have $p = p_1 = \dots = p_N = (3, 0)$ and thus $z = 1$, and $q_1 = \dots = q_n = 2$. Any admissible instance with $z = 1$ that is reachable from A by an admissible path satisfies $x_i \in \mathbb{Z}$ for all $i = 1, \dots, N$, since whenever $z = 1$ the $q_i + z$ are positive. Additionally, for every $(y_i, \dots, y_N) \in \mathbb{Z}^N$ there is an admissible path starting at A and ending at a position with $z = 1$ and $x_i = y_i$ for all $i = 1, \dots, N$. For this claim we only have to prove that each variable x_i can be changed by ±1 independently from the others. In order to obtain such an elementary change simply set $q_i = 0$ and $q_j = 2$ for $i \neq j$, and do a full clockwise or counterclockwise turn with point z.

Now let $f(x_1, \dots, x_N)$ be an instance of the polynomial used in Thm. 7.1. We finish our construction by adding the following statements to our GSP:

$$\begin{aligned} F &\leftarrow -(f(x_1, \dots, x_N))^2 - 3/2 \\ q_{N+1} &\leftarrow \text{ONINTERVAL}(0, 1, p_{N+1}) \\ l_0 &\leftarrow \text{JOIN}(q_{N+1} + F + z, 0) \\ l_1 &\leftarrow \text{BISECT}(l_{\mathbb{R}}, l_0) \\ l_2 &\leftarrow \text{BISECT}(l_{\mathbb{R}}, l_1) \end{aligned}$$

For the starting instance A we assume that we have $p_{N+1} = -1$ and hence $q_{N+1} = 0$. This implies that $q_{N+1} + F$ is real. Thus in A we have $l_0 = l_{\mathbb{R}}$. For A we resolve the ambiguities of the BISECT-operations by setting $l_1 = l_2 = l_{\mathbb{R}}$. The only point in which instance B differs from A is that in B we set $l_2 = l_{i\mathbb{R}}$.

We now claim that it is undecidable whether instance B is reachable from instance A: Observe that the only way in which A can be transformed into B is that the point $q_{N+1} + F + z$ makes an odd number of cycles around the origin.

If there are $(y_i, \dots, y_N) \in \mathbb{Z}^N$ with $f(y_1, \dots, y_N) = 0$ we can achieve such a movement as follows. We set $q_{N+1} = 0$. By the procedure described above we follow a path that puts the x_i to the values of the y_i. Then we set all $q_1 = \dots q_N = 0$ and $q_{N+1} = 1$, do another full cycle with z, set $q_{N+1} = 0$ again, and reset to $x_i = 0$ again by a suitable movement. The resulting situation is exactly instance B.

Conversely assume that there is an admissible path from A to B. During this path we must have at least one position where $q_{N+1} + F + z$ is real and positive. This can only happen if $z = 1$, since $q_{N+1} + F$ is always real and by definition at most $-1/2$. However, if $z = 1$ the $f(x_1, \dots, x_N)$ must have an integer value. The only way to get $q_{N+1} + F + z > 0$ is to have $f(x_1, \dots, x_N) = 0$. Since all x_i were integral this means that a solution of the Diophantine equation exists. □

8 Open Problems

We end our considerations by stating at least some of the open problems in decision complexity of tracing and reachability.

Problem 8.1. *Determine upper bounds for the decision complexity in the various setups described in this paper.*

Problem 8.2. *Extend the* JMB *instruction set by a new non-deterministic operation that intersects a circle and a line. Furthermore, enlarge the setup such that also complex coordinates for points lines and circles are allowed. By this a circle and a line always have two or one intersections. What is the decision complexity of the reachability problem in this context?*

This problem is of fundamental importance, since if the intrinsic complexity would turn out not to be too big this might yield good algorithms for randomized theorem proving for ruler and compass theorems. The structure of this problem seems to be fundamentally different from the problems discussed in this paper. It is not unlikely that for this problem there are effective randomized methods. However, we are pessimistic about fast deterministic methods, since we can prove that it is at least as hard as zero testing for polynomials [13].

Closely related to the above problem is the following:

Problem 8.3. *Extend the* JMB *instruction set by a new non-deterministic operation that intersects a circle and a line. Furthermore, enlarge the setup such that also complex coordinates for points, lines, and circles are allowed. Let A be an instance of a construction and let B' be a partial instance that just defines the positions of the free elements. How difficult is it to complete B' to an instance B that is reachable from A?*

Our last problem forms another approach that may single out certain tracing problems to be more easy than others. Since it asks for a new concept, the formulation is kept a little vague on purpose.

Problem 8.4. *Define the "right" concept of* output sensitivity *for the tracing problem that allows statements like "if the elements move not too wildly we can trace them easily".*

Finally we ask for a slightly stronger version of Thm. 7.2 that gets rid of the constant π hidden in the WHEEL-operation.

Problem 8.5. *Redefine the relation* WHEEL *by*

$$\text{WHEEL} := \{(p,q) \mid q = \varphi; q = r \cdot e^{i\varphi}\} \subset P \times P.$$

Is Thm. 7.2 still valid in this setup?

References

1. J. Culberson, Sokoban is PSPACE-complete, Proceedings in *Informatics 4, Fun With Algorithms*, E. Lodi, L. Pagli and N. Santoro, eds. (Carleton Scientific, Waterloo, pp. 65-76, 1999).
2. M.J. Garey amd D. S. Johnson, *Computers and Intractability*, (W.H. Freeman and Company, New York, 1979).
3. J.R. Gilbert, T. Lengauer and R.E. Tarjan, The pebbling problem is complete in polynomial space, SIAM J. Comput. **9**, 513–524 (1980).
4. H. Günzel, The universal partition theorem for oriented matroids, Discrete Comput. Geom. **19**, 521–551 (1998).
5. Ch.M. Hoffmann, Solid modeling, in *Handbook of Discrete and Computational Geometry*, J.E. Goodman & J. O'Rourke eds. (Lecture Notes in Mathematics 1346, CRC Press, Boca Raton, New York, pp. 863–880, 1997).
6. J. Hopcroft, D. Joseph and S. Whitesides, Movement problems for 2-dimensional Linkages, SIAM J. Comput. **13**, 610–629 (1984).
7. J. Hopcroft, J.T. Schwarz and M. Sharir, On the complexity of motion planning for multiple independent objects; PSPACE-Hardness of the "Warehouseman's Problem", Intern. J. Robotics Research **3**, 76–87 (1984).
8. N. Jackiw, *The Geometer's Sketchpad*, (Key Curriculum Press, Berkeley, 1991–1995).
9. D. Jordan and M. Steiner, Configuration Spaces of Mechanical Linkages, Discrete Comput. Geom. **22**, 297–315 (1999).

10. M. Kapovich and J. Millson, On the moduli spaces of polygons in the Euclidean plane, J. of Differential Geometry **42**, 133–164 (1995).
11. U. Kortenkamp, *Foundations of Dynamic Geometry*, (PhD-thesis, ETH Zürich, 1999) `http://www.inf.fu-berlin.de/~kortenka/Papers/diss.pdf`.
12. U. Kortenkamp and J. Richter-Gebert, Grundlagen dynamischer geometrie, in *Zeichnung - Figur - Zugfigur*, H.-J. Elschenbroich, Th. Gawlick and H.-W. Henn, eds. (Franzbecker, 2001).
13. U. Kortenkamp and J. Richter-Gebert, Decision complexity in dynamic geometry, in *Automated Deduction in Geometry (ADG 2000)*, J. Richter-Gebert, D. Wang, eds. (LNAI **2061**, Springer, Heidelberg, pp. 193–198, 2001).
14. U. Kortenkamp and J. Richter-Gebert, The intrinsic complexity of analytic continuation, in preparation.
15. Jean-Marie Laborde and Franck Bellemain, *Cabri-Geometry II*, (Texas Instruments, 1993–1998).
16. N.E. Mnëv, The universality theorems on the classification problem of configuration varieties and convex polytopes varieties, in *Topology and Geometry – Rohlin Seminar*, O.Ya. Viro, ed. (Lecture Notes in Mathematics, **1346** Springer, Heidelberg, pp. 527–544, 1988).
17. D.A. Plaisted, Sparse complex polynomials and polynomial reducibility, Journal of Computers and System Sciences **14**, 210–221 (1977).
18. D.A. Plaisted, Some polynomial and integer divisibility problems are NP-hard, SIAM J. Comput. **7**, 458–464 (1978).
19. D.A. Plaisted, New NP-Hard and NP-Complete Polynomial and Integer Divisibility Problems, Theoretical Computer Science **31**, 125–138 (1984).
20. J. Richter-Gebert, The Universality theorems for oriented matroids and polytopes, Contemporary Mathematics **223**, 269–292 (1999).
21. J. Richter-Gebert, *Realization Spaces of Polytopes*, (Lecture Notes in Mathematics, **1643** Springer, Heidelberg, 1996).
22. J. Richter-Gebert and U. Kortenkamp, *Cinderella - The Interactive Geometry Software*, (Springer 1999); see also `http://www.cinderella.de`.
23. J. Richter-Gebert and U. Kortenkamp, *Cinderella - Die Interaktive Geometriesoftware*, (HEUREKA Klett, 2000).
24. J. Reif, *Complexity of the Movers' Problem and Generalizations*, (Proc. 20th IEEE Conf. on Foundations of Comp. Sci., Long Beach, Calif.: IEEE Computer Society, pp. 421–427, 1979).
25. P. Shor, Stretchability of pseudolines is NP–hard, in: *Applied Geometry and Discrete Mathematics – The Victor Klee Festschrift*, P. Gritzmann and B. Sturmfels, eds. (DIMACS Series in Discrete Mathematics and

Theoretical Computer Science, Amer. Math. Soc., Providence, RI, **4** pp. 531–554, 1991).

26. M. Shub and S. Smale, Complexity of Bezout's theorem I: Geometric aspects, J. Amer. Math. Soc. **6**, 459–501 (1993).
27. M. Shub and S. Smale, Complexity of Bezout's theorem II: Volumes and probabilities, in: *Computational Algebraic Geometry*, F. Eyssette and A. Galligo, eds. (Progress in Mathematics, Birkhauser, **109** pp. 267–285, 1993).
28. M. Shub and S. Smale, Complexity of Bezout's theorem III: Condition number and packing, J. Complexity **9**, 4–14 (1993).
29. M. Shub and S. Smale, Complexity of Bezout's theorem IV: Probability of success, SIAM Jour. of Numerical Analysis **33**, 128–148 (1996).
30. M. Shub and S. Smale, Complexity of Bezout's theorem V: Polynomial time, Theoretical Computer Science **133**, 141–164 (1994).
31. Z.W. Sun, Reduction of unknowns in Diophantine representations (English Summary), Sci. Schina Ser A **35**, 257–269 (1992).
32. Z.W. Sun, J.P. Jones' work on Hilbert's tenth problem and related topics, Adv. in Math. (China) **22**, 312–331 (1993).

GRACE-LIKE POLYNOMIALS

DAVID RUELLE
IHES, 91440 Bures sur Yvette, France
E-mail: ruelle@ihes.fr

Results of somewhat mysterious nature are known on the location of zeros of certain polynomials associated with statistical mechanics (Lee-Yang circle theorem) and also with graph counting. In an attempt at clarifying the situation we introduce and discuss here a natural class of polynomials. Let $P(z_1, \ldots, z_m, w_1, \ldots, w_n)$ be separately of degree 1 in each of its $m+n$ arguments. We say that P is a Grace-like polynomial if $P(z_1, \ldots, w_n) \neq 0$ whenever there is a circle in $\mathbf{C}$ separating $z_1, \ldots, z_m$ from $w_1, \ldots, w_n$. A number of properties and characterizations of these polynomials are obtained. *I had the luck to meet Steve Smale early in my scientific career, and I have read his 1967 article in the Bulletin of the AMS more times than any other scientific paper. It took me a while to realize that Steve had worked successively on a variety of subjects, of which "differentiable dynamical systems" was only one. Progressively also I came to appreciate his independence of mind, expressed in such revolutionary notions as that the beaches of Copacabana are a good place to do mathematics. Turning away from scientific nostalgy, I shall now discuss a problem which is not very close to Steve's work, but has relations to his interests in recent years: finding where zeros of polynomials are located in the complex plane.*

1 Introduction

One form of the Lee-Yang circle theorem [3] states that if $|a_{ij}| \leq 1$ for $i, j = 1, \ldots, n$, and $a_{ij} = a_{ji}^*$, the polynomial

$$\sum_{X \subset \{1, \ldots, n\}} z^{\mathrm{card} X} \prod_{i \in X} \prod_{j \notin X} a_{ij}$$

has all its zeros on the unit circle $\{z : |z| = 1\}$.

Let now Γ be a finite graph. We denote by Γ' the set of dimer subgraphs γ (at most one edge of γ meets any vertex of Γ), and by Γ'' the set of unbranched subgraphs γ (no more than two edges of γ meet any vertex of Γ). Writing $|\gamma|$ for the number of edges in γ, on proves that

$$\sum_{\gamma \in \Gamma'} z^{|\gamma|}$$

has all its zeros on the negative real axis (Heilmann-Lieb [2]) and

$$\sum_{\gamma \in \Gamma''} z^{|\gamma|}$$

has all its zeros in the left-hand half plane $\{z : \text{Im}z < 0\}$ (Ruelle [6]).

The above results can all be obtained in a uniform manner by studying the zeros of polynomials

$$P(z_1, \dots, z_n)$$

which are *multiaffine* (separately of degree 1 in their n variables), and then taking $z_1 = \dots = z_n = z$. The multiaffine polynomials corresponding to the three examples above are obtained by multiplying factors for which the location of zeros is known and performing *Asano contractions*:

$$Auv + Bu + Cv + D \qquad \to \qquad Az + D$$

The *key lemma* (see [5]) is that if K, L are closed subsets of $\mathbf{C}\backslash\{0\}$ and if

$$u \notin K, v \notin L \qquad \Rightarrow \qquad Auv + Bu + Cv + D \neq 0$$

then

$$z \notin -K * L \qquad \Rightarrow \qquad Az + D \neq 0$$

where we have written $K * L = \{uv : u \in K, v \in L\}$.

To get started, let $P(z_1, \dots, z_n)$ be a multiaffine *symmetric* polynomial. If $W_1, \dots, W_n$ are the roots of $P(z, \dots, z) = 0$, we have

$$P(z_1, \dots, z_n) = \text{const.} \sum_{\pi} \prod_{j=1}^{n} (z_j - W_{\pi(j)})$$

where the sum is over all permutations π of n objects. *Grace's theorem* asserts that if $Z_1, \dots, Z_n$ are separated from $W_1, \dots, W_n$ by a circle of the Riemann sphere, then $P(Z_1 \dots, Z_n) \neq 0$. For example, if a is real and $-1 \leq a \leq 1$, the roots of $z^2 + 2az + 1$ are on the unit circle, and therefore

$$uv + au + av + 1$$

cannot vanish when $|u| < 1$, $|v| < 1$; from this one can get the Lee-Yang theorem.

In view of the above, it is natural to consider multiaffine polynomials

$$P(z_1, \dots, z_m, w_1, \dots, w_n)$$

which cannot vanish when $z_1, \dots, z_m$ are separated from $w_1, \dots, w_n$ by a circle. We call these polynomials Grace-like, and the purpose of this note is to study and characterize them.

2 General theory

We say that a complex polynomial $P(z_1, z_2, \ldots)$ in the variables z_1, $z_2, \ldots$ is a Multi-Affine Polynomial (*MA-nomial* for short) if it is separately of degree 1 in z_1, $z_2, \ldots$. We say that a circle $\Gamma \subset \mathbf{C}$ *separates* the sets A', $A'' \subset \mathbf{C}$ if $\mathbf{C}\backslash\Gamma = C' \cup C''$, where C', C'' are open, $C' \cap C'' = \emptyset$ and $A' \subset C'$, $A'' \subset C''$. We say that the MA-nomial $P(z_1, \ldots, z_m, w_1, \ldots, w_n)$ is Grace-like (or a G-nomial for short) if it satisfies the following condition

(G) *Whenever there is a circle* Γ *separating* $\{Z_1, \ldots, Z_m\}$, $\{W_1, \ldots, W_n\}$, *then*

$$P(Z_1, \ldots, W_n) \neq 0$$

[Note that we call circle either a straight line $\Gamma \subset \mathbf{R}$ or a *proper circle* $\Gamma = \{z \in \mathbf{C} : |z - a| = R\}$ with $a \in \mathbf{C}$, $0 < R < \infty$].

Lemma 1 (homogeneity). *The G-nomial P is homogeneous of degree* $k \leq$ $\min(m, n)$.

If there is a circle Γ separating $\{z_1, \ldots, z_m\}$, $\{w_1, \ldots, w_n\}$, then the polynomial $\lambda \mapsto P(\lambda z_1, \ldots, \lambda w_n)$ does not vanish when $\lambda \neq 0$, hence it is of the form $C\lambda^k$, where $C = P(z_1, \ldots, w_n)$. Thus

$$P(\lambda z_1, \ldots, \lambda w_n) = \lambda^k P(z_1, \ldots, w_n)$$

on an open set of $\mathbf{C}^{m+n}$, hence identically, *i.e.*, P is homogeneous of degree k.

Assuming $k > n$, each monomial in P would have a factor z_i, hence

$$P(0, \ldots, 0, 1, \ldots, 1) = 0$$

in contradiction with the fact that $\{0, \ldots, 0\}$, $\{1, \ldots, 1\}$ are separated by a circle. Thus $k \leq n$, and similarly $k \leq m$. □

Lemma 2 (degree). *If all the variables* $z_1, \ldots, w_n$ *effectively occur in the G-nomial P, then* $m = n$ *and P has degree* $k = n$.

By assumption

$$(\prod_{i=1}^{m} z_i)(\prod_{j=1}^{n} w_j)P(z_1^{-1}, \ldots, w_n^{-1})$$

is a homogeneous MA-nomial $\tilde{P}(z_1, \ldots, w_n)$ of degree $m+n-k$. If $Z_1, \ldots, W_n$ are all $\neq 0$ and $\{Z_1, \ldots, Z_m\}$, $\{W_1, \ldots, W_n\}$ are separated by a circle Γ, we may assume that Γ does not pass through 0. Then $\{Z_1^{-1}, \ldots, Z_m^{-1}\}$, $\{W_1^{-1}, \ldots, W_n^{-1}\}$ are separated by Γ^{-1}, hence $\tilde{P}(Z_1, \ldots, W_n) \neq 0$. Let $\mathcal{V}$ be the variety of zeros of $\tilde{P}$ and

$$\mathcal{Z}_i = \{(z_1, \ldots, w_n) : z_i = 0\} \quad , \quad \mathcal{W}_j = \{(z_1, \ldots, w_n) : w_j = 0\}$$

Then

$$\mathcal{V} \subset (\mathcal{V} \setminus \cup_{i,j} (\mathcal{Z}_i \cup \mathcal{W}_j)) \cup \cup_{i,j} (\mathcal{Z}_i \cup \mathcal{W}_j)$$

Since all the variables $z_1, \ldots, w_n$ effectively occur in $P(z_1, \ldots, w_n)$, none of the hyperplanes $\mathcal{Z}_i$, $\mathcal{W}_j$ is contained in $\mathcal{V}$, and therefore

$$\mathcal{V} \subset \text{closure}(\mathcal{V} \setminus \cup_{i,j} (\mathcal{Z}_i \cup \mathcal{W}_j))$$

We have seen that the points $(Z_1, \ldots, W_n)$ in $\mathcal{V} \setminus \cup_{i,j} (\mathcal{Z}_i \cup \mathcal{W}_j)$ are such that $\{Z_1, \ldots, Z_m\}$, $\{W_1, \ldots, W_m\}$ cannot be separated by a circle Γ, and the same applies to their limits. Therefore $\tilde{P}$ again satisfies (G). Applying Lemma 1 to P and $\tilde{P}$ we see that $k \leq \min(m,n)$, $m + n - k \leq \min(m,n)$. Therefore $m + n \leq 2\min(m,n)$, thus $m = n$, and also $k = n$. □

Proposition 3 (reduced G-nomials). *If $P(z_1, \ldots, z_m, w_1, \ldots, w_n)$ is a G-nomial, then P depends effectively on a subset of variables which may be relabelled $z_1, \ldots, z_k, w_1, \ldots, w_k$ so that*

$$P(z_1, \ldots, z_m, w_1, \ldots, w_n) = \alpha R(z_1, \ldots, z_k, w_1, \ldots, w_k)$$

where $\alpha \neq 0$, the G-nomial R is homogeneous of degree k, and the coefficient of $z_1 \cdots z_k$ in R is 1.

This follows directly from Lemma 2. □

We call a G-nomial R as above a *reduced* G-nomial.

Lemma 4 (translation invariance). *If $P(z_1, \ldots, w_n)$ is a G-nomial, then*

$$P(z_1 + s, \ldots, w_n + s) = P(z_1, \ldots, w_n)$$

i.e., P is translation invariant.

If there is a circle Γ separating $\{z_1, \ldots, z_m\}$, $\{w_1, \ldots, w_n\}$, then the polynomial

$$p(s) = P(z_1 + s, \ldots, w_n + s)$$

satisfies $p(s) \neq 0$ for all $s \in \mathbf{C}$. This implies that $p(s)$ is constant, or $dp/ds = 0$, for $(z_1, \ldots, w_n)$ in a nonempty open subset of $\mathbf{C}^{2n}$. Therefore $dp/ds = 0$ identically, and p depends only on $(z_1, \ldots, w_n)$. From this the lemma follows. □

Proposition 5 (properties of reduced G-nomials). *If $P(z_1, \ldots, w_n)$ is a reduced G-nomial, the following properties hold:* (reduced form) *there are constants C_π such that P has the reduced form*

$$P(z_1, \ldots, w_n) = \sum_\pi C_\pi \prod_{j=1}^{n} (z_j - w_{\pi(j)})$$

where the sum is over all permutations π *of* $(1, \ldots, n)$ (conformal invariance) *if* $ad - bc \neq 0$, *then*

$$P\left(\frac{az_1 + b}{cz_1 + d}, \ldots, \frac{aw_n + b}{cw_n + d}\right) = P(z_1, \ldots, w_n) \prod_{j=1}^{n} \frac{ad - bc}{(cz_j + d)(cw_j + d)}$$

in particular we have the identity

$$\left(\prod_{i=1}^{k} z_i\right)\left(\prod_{j=1}^{k} w_j\right) R(z_1^{-1}, \ldots, w_k^{-1}) = (-1)^k R(z_1, \ldots, w_k)$$

(roots) *the polynomial*

$$\hat{P}(z) = P(z, \ldots, z, w_1, \ldots, w_n)$$

is equal to $\prod_{k=1}^{n}(z - w_k)$, *so that its roots are the* w_k *(repeated according to multiplicity).*

Using Proposition 3 and Lemma 4, the above properties follow from Proposition A2 and Corollary A3 in Appendix A. □

Proposition 6 (compactness). *The space of MA-nomials in* $2n$ *variables which are homogeneous of degree* n *may be identified with* $\mathbf{C}^{\binom{2n}{n}}$. *The set* G_n *of reduced G-nomials of degree* n *is then a compact subset of* $\mathbf{C}^{\binom{2n}{n}}$. *We shall see later (Corollary 15) that* G_n *is also contractible.*

Let $P_k \in G_n$ and $P_k \to P_\infty$. In particular P_∞ is homogeneous of degree n, and the monomial $z_1 \cdots z_n$ occurs with coefficient 1. Suppose now that

$$P_\infty(Z_1, \ldots, Z_m, W_1, \ldots, W_n) = 0$$

with $\{Z_1, \ldots, Z_m\}$, $\{W_1, \ldots, W_n\}$ separated by a circle Γ. One can then choose discs $D_1, \ldots, D_{2n}$ containing $\{Z_1, \ldots, W_n\}$ and not intersecting Γ. Lemma A1 in Appendix A would then imply that P_∞ vanishes identically in contradiction with the fact that P_∞ contains the term $z_1 \cdots z_n$. Therefore $P_\infty \in G_n$, *i.e.*, G_n is closed.

Suppose now that G_n were unbounded. There would then be P_k such that the largest coefficient (in modulus) c_k in P_k tends to ∞. Going to a subsequence we may assume that

$$c_k^{-1} P_k \to P_\infty$$

where P_∞ is a homogeneous MA-nomial of degree n, and does not vanish identically. The same argument as above shows that P_∞ is a G-nomial, hence (by Proposition 3) the coefficient α of $z_1 \cdots z_n$ does not vanish, but since $\alpha = \lim c_k^{-1}$, c_k cannot tend to infinity as supposed. G_n is thus bounded, hence compact. □

Proposition 7 (the cases $n = 1, 2$). *The reduced G-nomials with $n = 1, 2$ are as follows:*

For $n = 1$: $P = z_1 - w_1$.

For $n = 2$: $P = (1-\theta)(z_1 - w_1)(z_2 - w_2) + \theta(z_1 - w_2)(z_2 - w_1)$ with real $\theta \in [0, 1]$.

We use Proposition 5 to write P in reduced form. In the case $n = 1$, we have $P = C(z_1 - w_1)$, and $C = 1$ by normalization.

In the case $n = 2$, we have

$$P = C'(z_1 - w_1)(z_2 - w_2) + C''(z_1 - w_2)(z_2 - w_1)$$

In view of (G), C', C'' are not both 0. Assume $C' \neq 0$, then (G) says that

$$\frac{z_1 - w_1}{z_1 - w_2} : \frac{z_2 - w_1}{z_2 - w_2} + \frac{C''}{C'} \neq 0 \tag{1}$$

when $\{z_1, z_2\}$, $\{w_1, w_2\}$ are separated by a circle. If C''/C' were not real, we could find z_1, z_2, w_1, w_2 such that

$$\frac{z_1 - w_1}{z_1 - w_2} : \frac{z_2 - w_1}{z_2 - w_2} = -\frac{C''}{C'} \tag{2}$$

but the fact that the cross-ratio in the left hand side of (2) is not real means that z_1, z_2, w_1, w_2 are not on the same circle, and this implies that there is a circle separating $\{z_1, z_2\}$, $\{w_1, w_2\}$. Therefore (1) and (2) both hold, which is impossible. We must therefore assume C''/C' real, and it suffices to check (1) for z_1, z_2, w_1, w_2 on a circle. The condition that $\{z_1, z_2\}$, $\{w_1, w_2\}$ are separated by a circle is now equivalent to the cross-ratio being > 0, and therefore (G) is equivalent to $C''/C' \geq 0$. If we assume $C'' \neq 0$, the argument is similar and gives $C'/C'' \geq 0$. The normalization condition yields then $C' = 1 - \theta$, $C'' = \theta$ with $\theta \in [0, 1]$ □

Proposition 8 (determinants). *Let Δ_z be the diagonal $n \times n$ matrix where the j-th diagonal element is z_j and similarly for Δ_w. Also let U be a unitary $n \times n$ matrix ($U\Delta_w U^{-1}$ is thus an arbitrary normal matrix with eigenvalues $w_1, \ldots, w_n$). Then*

$$P(z_1, \ldots, z_n, w_1, \ldots, w_n) = \det(\Delta_z - U\Delta_w U^{-1})$$

is a reduced G-nomial. We may assume that $\det U = 1$ *and write*

$$\det(\Delta_z - U\Delta_w U^{-1}) = \det((U_{ij}(z_i - w_j)))$$

Let $\{z_1, \ldots, z_n\}$, $\{w_1, \ldots, w_n\}$ be separated by a circle Γ. We may assume that Γ is a proper circle. Suppose first that the z_j are inside the circle Γ and the w_j outside. We want to prove that $\det(\Delta_z - U\Delta_w U^{-1}) \neq 0$. By translation we may assume that Γ is centered at the origin, say $\Gamma = \{z : |z| = R\}$; then, by assumption, using the operator norm,

$$||\Delta_z|| < R \quad , \quad ||\Delta_w^{-1}|| < R^{-1}$$

Therefore

$$||\Delta_z(U\Delta_w U^{-1})^{-1}|| < 1$$

so that

$$\det(\Delta_z - U\Delta_w U^{-1}) = \det(-U\Delta_w U^{-1})\det(1 - \Delta_z(U\Delta_w U^{-1})^{-1}) \neq 0$$

as announced. The case where the w_j are inside Γ and the z_j outside is similar (consider $\det(\Delta_w - U^{-1}\Delta_z U)$. □

Proposition 9 (Grace's theorem). *The polynomial*

$$P_\Sigma(z_1, \ldots, z_n, w_1, \ldots, w_n) = \frac{1}{n!}\sum_\pi \prod_{j=1}^n (z_j - w_{\pi(j)}) \tag{3}$$

where the sum is over all permutations of $(1, \ldots, n)$ *is a reduced G-nomial.*

See Polya and Szegö [4] Exercise V 145. □

This result will also follow from our proof of Corollary 15 below.

Proposition 10 (permanence properties). (Permutations) *If* $P(z_1, \ldots, z_n, w_1, \ldots, w_n)$ *is a reduced G-nomial, permutation of the* z_i, *or the* w_j, *or interchange of* $(z_1, \ldots, z_n)$ *and* $(w_1, \ldots, w_n)$ *and multiplication by* $(-1)^n$ *produces again a reduced G-nomial.*

(Products) If $P'(z_1', \ldots, w_{n'}')$, $P''(z_1'', \ldots, w_{n''}'')$ *are reduced G-nomials, then their product* $P' \otimes P''(z_1', \ldots, z_{n''}'', w_1', \ldots, w_{n''}'')$ *is a reduced G-nomial.*

(Symmetrization) *Let* $P(z_1, \ldots, z_n, w_1, \ldots, w_n)$ *be a reduced G-nomial, and*

$$P_S(z_1, \ldots, z_n, w_1, \ldots, w_n)$$

be obtained by symmetrization with respect to a subset S *of the variables* $z_1, \ldots, z_n$, *then* P_S *is again a reduced G-nomial. Symmetrization with respect to all variables* $z_1, \ldots, z_n$ *produces the polynomial* P_Σ *given by (3).*

The part of the proposition relative to permutations and products follows readily from the definitions. To prove the symmetrization property we may relabel variables so that S consists of $z_1, \ldots, z_s$. We denote by $\hat{P}(z)$ the polynomial obtained by replacing $z_1, \ldots, z_s$ by z in P (the dependence on

$z_{s+1}, \ldots, w_n$ is not made explicit). With this notation P_S is the only MA-nomial symmetric with respect to $z_1, \ldots, z_n$ and such that $\hat{P}_S(z) = \hat{P}(z)$. We may write

$$\hat{P}(z) = \alpha(z - a_1) \cdots (z - a_s) \tag{4}$$

where $\alpha, a_1, \ldots, a_s$ may depend on $z_{s+1}, \ldots, w_n$. If Γ is a circle separating the regions C', C'', and $z_{s+1}, \ldots, z_n \in C'$, $w_1, \ldots, w_n \in C''$, (G) implies that $\alpha \neq 0$ and $a_1, \ldots, a_s \notin C'$. Grace's theorem implies that P_S does not vanish when $z_1, \ldots, z_s$ are separated by a circle from $a_1, \ldots, a_s$. Therefore P_S does not vanish when $z_1, \ldots, z_s \in C'$, hence P_S is a G-nomial, which is easily seen to be reduced. If $s = n$, (4) becomes

$$\hat{P}(z) = (z - w_1) \cdots (z - w_n)$$

in view of Proposition 5, hence symmetrisation of P gives P_Σ.

3 Further results

We define now G$_0$-nomials as a class of MA-nomials satisfying a new condition (G$_0$) weaker than (G). It will turn out later that G$_0$-nomials and G-nomials are in fact the same (Proposition 12). The new condition is (G$_0$) *If there are two proper circles, or a proper circle and a straight line Γ', $\Gamma'' \subset \mathbf{C}$ such that $z_1, \ldots, z_m \in \Gamma'$, $w_1, \ldots, w_n \in \Gamma''$, and $\Gamma' \cap \Gamma'' = \emptyset$, then*

$$P(z_1, \ldots, z_m, w_1, \ldots, w_n) \neq 0$$

Remember that a proper circle is of the form $\{z : |z - a| = R\}$, with $0 < R < \infty$. For the purposes of (G$_0$) we may allow $R = 0$ (because a circle Γ' or Γ'' reduced to a point a' or a'' can be replaced by a small circle through a' or a'').

Lemma 11. *Let $P(z_1, \ldots, w_n)$ be a G$_0$-nomial, and define*

$$\tilde{P}(z_1, \ldots, w_n) = \left(\prod_{i=1}^{m} z_i\right)\left(\prod_{j=1}^{n} w_j\right) P(z_1^{-1}, \ldots, w_n^{-1}) \tag{5}$$

(a) P is translation invariant.

(b) If P depends effectively on $z_1, \ldots, w_n$, then $\tilde{P}$ is translation invariant, and therefore a G$_0$-nomial.

The polynomial $a \mapsto p(a) = P(z_1 + a, \dots, w_n + a)$ does not vanish, and is therefore constant if $z_1, \dots, z_m \in \Gamma'$, $w_1, \dots, w_n \in \Gamma''$, and $\Gamma' \cap \Gamma'' = \emptyset$. But this means $dp/da = 0$ under the same conditions, and therefore dp/da vanishes identically. This proves (a). From the fact that P depends effectively on $z_1, \dots, w_n$, we obtain that none of the $m + n$ polynomials

$$\tilde{P}(0, z_2 - z_1, \dots, w_n - z_1)$$

$$\dots$$

$$\tilde{P}(z_1 - w_n, \dots, w_{n-1}, 0)$$

vanishes identically. The union $\mathcal{Z}$ of their zeros has thus a dense complement in $\mathbf{C}^{m+n}$. Let now Γ', Γ'' be disjoint proper circles in $\mathbf{C}$. If $z_1, \dots, z_m \in \Gamma'$, $w_1, \dots, w_n \in \Gamma''$, the polynomial

$$a \mapsto \tilde{p}(a) = \tilde{P}(z_1 + a, \dots, w_n + a)$$

can vanish only if $a \in \{-z_1, \dots, -w_n\}$. [This follows from (G_0) and the fact that $(a + \Gamma')^{-1}$, $(a + \Gamma'')^{-1}$ are disjoint and are proper circles or a proper circle and a straight line]. To summarize, $\tilde{p}(a)$ can vanish only if

$$a \in \{-z_1, \dots, -w_n\} \qquad \text{and} \qquad (z_1, \dots, w_n) \in \mathcal{Z}$$

Since a polynomial vanishing on a nonempty open set of $\Gamma'^m \times \Gamma''^n$ must vanish identically on $\mathbf{C}^{m+n}$, we have

$$(\mathbf{C}^{m+n} \backslash \mathcal{Z}) \cap (\Gamma'^m \times \Gamma''^n) \neq \emptyset$$

There is thus a nonempty open set $U \subset (\Gamma'^m \times \Gamma''^n) \backslash \mathcal{Z}$. For $(z_1, \dots, w_n) \in U$, $\tilde{p}(\cdot)$ never vanishes, and is thus constant, *i.e.*, $d\tilde{p}(a)/da = 0$. Therefore $d\tilde{p}(a)/da = 0$ for all $(z_1, \dots, w_n) \in \mathbf{C}^{m+n}$. In conclusion, $\tilde{P}$ is translation invariant. This implies immediately that $\tilde{P}$ is a G_0-nomial. □

Proposition 12. *If the MA-nomial $P(z_1, \dots, z_m, w_1, \dots, w_n)$ satisfies (G_0), it also satisfies (G).*

If the sets $\{z_1, \dots, z_m\}$ and $\{w_1, \dots, w_n\}$ are separated by a circle Γ, we can find two disjoint proper circles Γ' and Γ'' close to Γ and separating them. By a transformation $\Phi : z \mapsto (z + a)^{-1}$, we may assume that $\Phi z_1, \dots, \Phi z_m$ are *inside* of the circle $\Phi\Gamma'$, and $\Phi w_1, \dots, \Phi w_n$ *inside* of the circle $\Phi\Gamma''$. We may write

$$\Phi\Gamma' = \{z : |z - u| = r'\} \qquad , \qquad \Phi\Gamma'' = \{w : |w - v| = r''\}$$

The assumption that P is a G_0-nomial and Lemma 11 imply that $\tilde{P}$ (defined by (5)) satisfies $\tilde{P}(z_1, \dots, w_n) \neq 0$ if

$$z_1, \dots, z_m \in \{z : |z - u| = \rho'\}, \qquad , \qquad w_1, \dots, w_n \in \{w : |w - v| = \rho''\}$$

whenever $0 \leq \rho' \leq r'$ and $0 \leq \rho'' \leq r''$. Considered as a function of the $\xi_i = \log(z_i - u)$ and $\eta_j = \log(w_j - v)$, $\tilde{P}$ has no zero, and $1/\tilde{P}$ is thus analytic in a region

$$\{\mathrm{Re}\,\xi_i < c \text{ for } i = 1, \dots, m \text{ and } \mathrm{Re}\,\eta_j < c \text{ for } j = 1, \dots, n\}$$
$$\cup\{\mathrm{Re}\,\xi_1 = \dots = \mathrm{Re}\,\xi_m < \log r' \text{ and } \mathrm{Re}\,\eta_1 = \dots = \mathrm{Re}\,\eta_n < \log r''\}$$

for suitable (large negative) c. This is a tube and by the Tube Theorem[a] $1/\tilde{P}$ is analytic in

$$\{\mathrm{Re}\,\xi_i < \log r' \text{ for } i = 1, \dots, m \text{ and } \mathrm{Re}\,\eta_j < \log r'' \text{ for } j = 1, \dots, n\}$$

and therefore $\tilde{P}$ does not vanish when $z_1, \dots, z_m$ are inside of $\Phi\Gamma'$ and $w_1, \dots, w_n$ inside $\Phi\Gamma''$. Going back to the polynomial P, we see that it cannot vanish when $\{z_1, \dots, z_m\}$ and $\{w_1, \dots, w_n\}$ are separated by Γ' and Γ''. □

Proposition 13. *Suppose that $P(z_1, \dots, z_n, w_1, \dots, w_n)$ satisfies the conditions of Proposition A2 and that*

$$P(z_1, \dots, z_n, w_1, \dots, w_n) \neq 0$$

when $|z_1| = \dots = |z_n| = a$, $|w_1| = \dots = |w_n| = b$ and $0 < a \neq b$. Then P is a G-nomial.

Taking $z_1, \dots, z_n = 3/2$, $w_1, \dots, w_n = 1/2$, we have $0 \neq P(3/2, \dots, 1/2) = P(1, \dots, 0) = \alpha$, *i.e.*, the coefficient α of the monomial $z_1 \dots z_n$ in P is different from 0 . Therefore we have

$$P(z_1, \dots, z_n, w_1, \dots, w_n) \neq 0 \tag{6}$$

if $|z_1| = \dots = |z_n| = a$, $|w_1| = \dots = |w_n| = b$ and $0 \leq a < b$; (6) also holds if $|w_1| = \dots = |w_n| = b$ provided $|z_1|, \dots, |z_n| < e^c$ for suitable (large negative) c. Applying the Tube Theorem as in the proof of Proposition 12 we find thus that (6) holds when

$$|z_1|, \dots, |z_n| < b \qquad , \qquad |w_1| = \dots = |w_n| = b.$$

In particular, $P(z_1, \dots, w_n) \neq 0$ if $z_1, \dots, z_n \in \Gamma'$, $w_1, \dots, w_n \in \Gamma''$ where Γ', Γ'' are proper circles such that Γ' is entirely inside Γ'' and Γ'' is centered at 0. But by conformal invariance (Corollary A3) we can replace these conditions by $\Gamma' \cap \Gamma'' = \emptyset$. Proposition 12 then implies that P is a G-nomial. □

[a] For the standard Tube Theorem see for instance [7] Theorem 2.5.10. Here we need a variant, the Flattened Tube Theorem, for which see Epstein [1]

Proposition 14. *Suppose that $P_0(z_1,\dots,w_n)$ and $P_1(z_1,\dots,w_n)$ are reduced G-nomials which become equal when $z_1 = z_2$:*

$$P_0(z,z,z_3,\dots,w_n) = P_1(z,z,z_3,\dots,w_n)$$

Then, for $0 \leq \alpha \leq 1$

$$P_\alpha(z_1,\dots,w_n) = (1-\alpha)P_0(z_1,\dots,w_n) + \alpha P_1(z_1,\dots,w_n)$$

is again a reduced G-nomial.

[Note that instead of the pair (z_1,z_2) one could take any pair (z_i,z_j)]. We have to prove that if the proper circle Γ separates $\{z_1,\dots,z_n\}$, $\{w_1,\dots,w_n\}$, then $P_\alpha(z_1,\dots,w_n) \neq 0$. Let $p_\alpha(z_1,z_2)$ be obtained from $P_\alpha(z_1,\dots,w_n)$ by fixing $z_3,\dots,z_n$ on one side of Γ and $w_1,\dots,w_n$ on the other side. By assumption

$$p_\alpha(z_1,z_2) = az_1z_2 + b_\alpha z_1 + c_\alpha z_2 + d \tag{7}$$

where $b_\alpha = (1-\alpha)b_0 + \alpha b_1$, $c_\alpha = (1-\alpha)c_0 + \alpha c_1$, and $b_0 + c_0 = b_1 + c_1$. We have to prove: (A) *If $z_1, z_2 \in \Delta$ where Δ is the region bounded by Γ and not containing $w_1,\dots,w_n$, then $p_\alpha(z_1,z_2) \neq 0$.* We remark now that, as functions of $z_3,\dots,w_n$, the expressions

$$a \quad , \quad -\frac{(b_0+c_0)^2}{4a} + d$$

cannot vanish identically. For a this is because the coefficient of $z_1 \cdots z_n$ in (7) is 1. Note now that if we decompose a in prime factors, these cannot occur with an exponent > 1 because a is of degree ≤ 1 in each variable $z_3,\dots,w_n$. Therefore if $-(b_0+c_0)^2/4a + d = 0$, *i.e.*, if a divides $(b_0+c_0)^2$, then a divides (b_0+c_0) and the quotient is homogeneous of degree 1. But then $(b_0+c_0)^2/4a$ contains some variables with an exponent 2, in contradiction with the fact that in d all variables occur with an exponent ≤ 1. In conclusion $-(b_0+c_0)^2/4a+d$ cannot vanish identically. By a small change of $z_3,\dots,w_n$ we can thus assume that

$$a \neq 0 \quad , \quad -\frac{(b_0+c_0)^2}{4a} + d \neq 0 \tag{8}$$

We shall first consider this case and later use a limit argument to prove (A) when (8) does not hold. By the change of coordinates

$$z_1 = u_1 - \frac{b_0+c_0}{2a} \quad , \quad z_2 = u_2 - \frac{b_0+c_0}{2a}$$

(linear in z_1, z_2) we obtain

$$p_\alpha = au_1u_2 + \frac{1}{2}(b_\alpha - c_\alpha)(u_1 - u_2) - \frac{(b_0 + c_0)^2}{4a} + d$$

(Note that $b_\alpha + c_\alpha = b_0 + c_0$). Write

$$A = (b_0 + c_0)^2 - 4ad \quad , \quad \beta = \frac{\sqrt{A}}{2a} \quad , \quad \lambda(\alpha) = \frac{b_\alpha - c_\alpha}{\sqrt{A}}$$

for some choice of the square root of A, and

$$u_1 = \beta v_1 \quad , \quad u_2 = \beta v_2$$

then

$$p_\alpha = \frac{A}{4a}(v_1v_2 + \lambda(\alpha)(v_1 - v_2) - 1)$$

If we write $v_1 = (\zeta_1 + 1)/(\zeta_1 - 1)$, $v_2 = (\zeta_2 + 1)/(\zeta_2 - 1)$, the condition $p_\alpha \neq 0$ becomes

$$\zeta_1(1 - \lambda(\alpha)) + \zeta_2(1 + \lambda(\alpha)) \neq 0 \tag{9}$$

Note that $\lambda(\alpha) = \pm 1$ means $(b_\alpha - c_\alpha)^2 - A = 0$, *i.e.*, $ad - b_\alpha c_\alpha = 0$ and

$$p_\alpha = a(z_1 - S_\alpha)(z_2 - T_\alpha)$$

By assumption $p_0(z, z) \neq 0$ when $z \in \Delta$. Therefore, the image Δ_v of Δ in the v-variable does not contain $+1$, -1, and the image Δ_ζ in the ζ-variable does not contain 0, ∞. In particular Δ_ζ is a circular disc or a half-plane, and thus *convex*. If $\lambda(\alpha)$ is real and $-1 \leq \lambda(\alpha) \leq 1$, (9) holds when $\zeta_1, \zeta_2 \in \Delta_\zeta$.

[This is because Δ_ζ is convex and $\Delta_\zeta \not\ni 0$]. Therefore in that case (A) holds. We may thus exclude the values of α such that $-1 \leq \lambda(\alpha) \leq 1$, and reduce

the proof of the proposition to the case when at most one of $\lambda(0)$, $\lambda(1)$ is in $[-1, 1]$, and the other $\lambda(\alpha) \notin [-1, 1]$. Exchanging possibly P_0, P_1, we may assume that all $\lambda(\alpha) \notin [-1, 1]$ except $\lambda(1)$. Exchanging possibly z_1, z_2, (*i.e.*, replacing λ by $-\lambda$) we may assume that $\lambda(1) \neq 1$. We may finally assume that

$$|\lambda(0) + 1| + |\lambda(0) - 1| \geq |\lambda(1) + 1| + |\lambda(1) - 1| \tag{10}$$

where the left hand side is > 2 while the right hand side is $=2$ if $\lambda(1) \in [-1, 1]$. For $\alpha \in [0, 1]$ we define the map

$$f_\alpha : \zeta \mapsto \frac{\lambda(\alpha) + 1}{\lambda(\alpha) - 1}\zeta$$

When $\alpha = 0, 1$ the inequality (9) holds by assumption for $\zeta_1, \zeta_2 \in \Delta_\zeta$. [Note that the point $v = \infty$, *i.e.*, $\zeta = 1$ does not make a problem: if $\lambda \neq \pm 1$ this is seen by continuity; if $\lambda = \pm 1$ this follows from $\Delta_\zeta \not\ni 0$]. Therefore

$$\Delta_\zeta \cap f_0 \Delta_\zeta = \emptyset \quad , \quad \Delta_\zeta \cap f_1 \Delta_\zeta = \emptyset$$

We want to show that $\Delta_\zeta \cap f_\alpha \Delta_\zeta = \emptyset$ for $0 < \alpha < 1$. In fact, it suffices to prove

$$\Delta'_\zeta \cap f_\alpha \Delta'_\zeta = \emptyset$$

for slightly smaller $\Delta' \subset \Delta_\zeta$, *viz*, the inside of a proper circle Γ' such that 0 is outside of Γ'. Since we may replace Δ' by any $c\Delta'$ where $c \in \mathbf{C}\backslash\{0\}$, we assume that Δ' is the interior of a circle centered at $\lambda(0) - 1$ and with radius $r_-^0 < |\lambda(0) - 1|$. Then $f_0\Delta'$ is the interior of a circle centered at $\lambda(0) + 1$ and with radius r_+^0. The above two circles are disjoint, but we may increase r_-^0 until they touch, obtaining

$$r_-^0 + r_+^0 = 2 \quad , \quad r_+^0 = |\frac{\lambda(0) + 1}{\lambda(0) - 1}| r_-^0$$

i.e.,

$$r_-^0 = \frac{2|\lambda(0) - 1|}{|\lambda(0) + 1| + |\lambda(0) - 1|} \quad , \quad r_+^0 = \frac{2|\lambda(0) + 1|}{|\lambda(0) + 1| + |\lambda(0) - 1|}$$

We define r_-^α and r_+^α similarly, with $\lambda(0)$ replaced by $\lambda(\alpha)$ for $\alpha \in [0, 1]$. To prove that $\Delta' \cap f_\alpha \Delta' = \emptyset$ for $0 < \alpha < 1$, we may replace Δ' by $\frac{\lambda(\alpha)-1}{\lambda(0)-1}\Delta'$ (which is a disc centered at $\lambda(\alpha) - 1$) and it suffices to prove that the radius $|\frac{\lambda(\alpha)-1}{\lambda(0)-1}| r_-^0$ of this disc is $\leq r_-^\alpha$, *i.e.*,

$$\frac{2|\lambda(\alpha) - 1|}{|\lambda(0) + 1| + |\lambda(0) - 1|} \leq \frac{2|\lambda(\alpha) - 1|}{|\lambda(\alpha) + 1| + |\lambda(\alpha) - 1|}$$

or

$$|\lambda(0) + 1| + |\lambda(0) - 1| \geq |\lambda(\alpha) + 1| + |\lambda(\alpha) - 1| \tag{11}$$

Note now that $\{\lambda \in \mathbf{C} : |\lambda + 1| + |\lambda - 1| = \mathit{const.}\}$ is an ellipse with foci ± 1, and since $\lambda(\alpha)$ is affine in α, the maximum value of $|\lambda(\alpha) + 1| + |\lambda(\alpha) - 1|$ for $\alpha \in [0, 1]$ is reached at 0 or 1, and in fact at 0 by (10). This proves (11). This

concludes the proof of (A) under the assumption (8). Consider now a limiting case when (8) fails and suppose that (A) does not hold. Then, by Lemma A1, p_α vanishes identically. In particular this would imply $p_\alpha(z, z) = 0$, in contradiction with the assumption that P_0 is a G-nomial. We have thus shown that P_α is a G-nomial, and since it is homogeneous of degree n in the $2n$ variables $z_1, \dots, w_n$, and contains $z_1 \cdots z_n$ with coefficient 1, P_α is a reduced G-nomial. $\square$

Corollary 15 (contractibility). *The set G_n of reduced G-nomials is contractible.*

In the linear space of MA-nomials $P(z_1, \dots, w_n)$ satisfying the conditions of Proposition A2 we define a flow by

$$\frac{dP}{dt} = -P + \binom{n}{2}^{-1} \sum{}^* \pi P \tag{12}$$

where Σ^* is the sum over the $\binom{n}{2}$ transpositions π, *i.e.* interchanges of two of the variables $z_1, \dots, z_n$ of P. In view of Proposition 14, the positive semiflow defined by (12) preserves the set G_n of reduced G-nomials. Condition $(b)_n$ of Proposition A2 shows that the only fixed point of (12) is, up to a normalizing factor, Grace's polynomial P_Σ. We have thus a contraction of G_n to $\{P_\Sigma\}$, and G_n is therefore contractible. $\square$

Appendix

Lemma 15 (limits). *Let $D_1, \dots, D_r$ be open discs, and assume that the MA-nomials $P_k(z_1, \dots, z_r)$ do not vanish when $z_1 \in D_1, \dots, z_r \in D_r$. If the P_k have a limit P_∞ when $k \to \infty$, and if $P_\infty(\hat{z}_1, \dots, \hat{z}_r) = 0$ for some $\hat{z}_1 \in D_1, \dots, \hat{z}_r \in D_r$, then $P_\infty = 0$ identically.*

There is no loss of generality in assuming that $\hat{z}_1 = \ldots = \hat{z}_r = 0$. We prove the lemma by induction on r. For $r = 1$, if the affine function P_∞ vanishes at 0 but not identically, the implicit function theorem shows that P_k vanishes for large k at some point close to 0, contrary to assumption. For $r > 1$, the induction assumption implies that putting any one of the variables $z_1, \dots, z_r$ equal to 0 in P_∞ gives the zero polynomial. Therefore $P_\infty(z_1, \dots, z_r) = \alpha z_1 \cdots z_r$. Fix now $z_j = a_i \in D_i \backslash \{0\}$ for $j = 1, \dots, r-1$. Then $P_k(a_1, \dots, a_{r-1}, z_r) \neq 0$ for $z_r \in D_r$, but the limit $P_\infty(a_1, \dots, a_{r-1}, z_r) = \alpha a_1 \cdots a_{r-1} z_r$ vanishes at $z_r = 0$ and therefore identically, *i.e.*, $\alpha = 0$, which proves the lemma. $\square$

Proposition 16 (reduced forms). *For $n \geq 1$, the following conditions on a MA-nomial $P(z_1, \dots, z_n, w_1, \dots, w_n)$ not identically zero are equivalent:*

$(a)_n$ P satisfies

$$P(z_1 + \xi, \dots, w_n + \xi) = P(z_1, \dots, w_n) \qquad \textit{(translation invariance)}$$

$$P(\lambda z_1, \dots, \lambda w_n) = \lambda^n P(z_1, \dots, w_n) \qquad \textit{(homogeneity of degree n)}$$

$(b)_n$ There are constants C_π such that

$$P(z_1,\dots,w_n)=\sum_\pi C_\pi\prod_{j=1}^{n}(z_j-w_{\pi(j)})$$

where the sum is over all permutations π of $(1,\dots,n)$.

We say that $(\mathrm{b})_n$ gives a *reduced form* of P (it need not be unique). Clearly $(\mathrm{b})_n \Rightarrow (\mathrm{a})_n$. We shall prove $(\mathrm{a})_n \Rightarrow (\mathrm{b})_n$ by induction on n, and obtain at the same time a bound $\sum|C_\pi|\le k_n.||P||$ for some norm $||P||$ (the space of P's is finite dimensional, so all norms are equivalent). Clearly, $(\mathrm{a})_1$ implies that $P(z_1,w_1)=C(z_1-w_1)$, so that $(\mathrm{b})_1$ holds. Let us now assume that P satisfies $(\mathrm{a})_n$ for some $n>1$. If X is an n-element subset of $\{z_1,\dots,w_n\}$, let $A(X)$ denote the coefficient of the corresponding monomial in P. We have

$$\sum_X A(X)=P(1,\dots,1)=P(0,\dots,0)=0$$

In particular

$$\max_{X',X''}|A(X')-A(X'')|\ge\max_X A(X)$$

Note also that one can go from X' to X'' in a bounded number of steps exchanging a z_j and a w_k. Therefore one can choose z_j, w_k, Z containing z_j and not w_k, and W obtained by replacing z_j by w_k in Z so that

$$|A(Z)-A(W)|\ge\alpha(\sum_X|A(X)|^2)^{1/2}$$

where α depends only on n. Write now

$$P=az_jw_k+bz_j+cw_k+d$$

where the polynomials a, b, c, d do not contain z_j, w_k. We have thus

$$P=P_1+\frac{1}{2}(b-c)(z_j-w_k)$$

where

$$P_1=az_jw_k+\frac{1}{2}(b+c)(z_j+w_k)+d$$

Let $\tilde a$, $\tilde b$, $\tilde c$, $\tilde d$ be obtained by adding ξ to all the arguments of a, b, c, d. By translation invariance we have thus

$$\begin{aligned}az_jw_k+bz_j+cw_k+d&=\tilde a(z_j+\xi)(w_k+\xi)+\tilde b(z_j+\xi)+\tilde c(w_k+\xi)+\tilde d\\&=\tilde az_jw_k+(\tilde a\xi+\tilde b)z_j+(\tilde a\xi+\tilde c)w_k+\tilde a\xi^2+(\tilde b+\tilde c)\xi+\tilde d\end{aligned}$$

hence $\tilde{b} - \tilde{c} = b - c$. Therefore $b - c$ satisfies $(\mathrm{a})_{n-1}$ and, using the induction assumption we see that

$$\frac{1}{2}(b-c)(z_j - w_k)$$

has the form given by $(\mathrm{b})_n$. In particular P_1 again satisfies $(\mathrm{a})_n$. We compare now the coefficients $A_1(X)$ for P_1 and $A(X)$ for P:

$$\sum_X |A(X)|^2 - \sum_X |A_1(X)|^2 \geq |A(Z)|^2 + |A(W)|^2 - \frac{1}{2}|A(Z) + A(W)|^2$$
$$= \frac{1}{2}|A(Z) - A(W)|^2 \geq \frac{\alpha^2}{2}\sum_X |A(X)|^2$$

so that

$$\sum |A_1(X)|^2 \leq (1 - \frac{\alpha^2}{2})\sum_X |A(X)|^2$$

We have thus a geometrically convergent approximation of P by expressions satisfying $(\mathrm{b})_n$, and an estimate of $\sum |C_\pi|$ as desired. □

Corollary 17. *If the MA-nomial $P(z_1, \ldots, w_n)$ satisfies the conditions of Proposition A2, the following properties hold:* (conformal invariance) *if $ad - bc \neq 0$, then*

$$P(\frac{az_1 + b}{cz_1 + d}, \ldots, \frac{aw_n + b}{cw_n + d}) = P(z_1, \ldots, w_n)\prod_{j=1}^{n}\frac{ad - bc}{(cz_j + d)(cw_j + d)}$$

(roots) *the polynomial*

$$\hat{P}(z) = P(z, \ldots, z, w_1, \ldots, w_n)$$

has exactly the roots $w_1, \ldots, w_n$ (repeated according to multiplicity).

These properties follow directly if one writes P in reduced form. □

References

1. H. Epstein. "Some analytic properties of scattering amplitudes in quantum field theory." in M. Chretien and S. Deser eds. *Axiomatic Field Theory.* Gordon and Breach, New York, 1966.
2. O.J. Heilmann and E.H. Lieb. "Theory of monomer-dimer systems." Commun. Math. Phys. **25**, (1972), 190–232; **27**, (1972), 166.

3. T. D. Lee and C. N. Yang. “Statistical theory of equations of state and phase relations. II. Lattice gas and Ising model.” Phys. Rev. **87**, (1952), 410–419.
4. G. Polya and G. Szegö. *Problems and theorems in analysis II.* Springer, Berlin, 1976.
5. D. Ruelle. “Extension of the Lee-Yang circle theorem.” Phys. Rev. Letters **26**, (1971), 303–304.
6. D. Ruelle. “Counting unbranched subgraphs.” J. Algebraic Combinatorics **9**,157-160(1999); “Zeros of graph-counting polynomials.” Commun. Math. Phys. **200**, (1999), 43–56.
7. L. Hörmander. *An introduction to complex analysis in several variables.* D. Van Nostrand, Princeton, 1966.

FROM DYNAMICS TO COMPUTATION AND BACK?

MICHAEL SHUB
IBM T. J. Watson Research Center, P. O. Box 218, Yorktown Heights, NY 10598
E-mail: mshub@us.ibm.com

Felicitations

It is a pleasure for me to be celebrating Steve's seventieth birthday. Twenty years ago I sent him a note congratulating him on his fiftieth birthday and wishing him another half century as productive as the first. Twenty years later I can say, so far so good. Real computation and complexity and now learning theory are added to the tremendous influence he has had on twentieth (now twenty first)century mathematics. Steve is an impossible act for his students to follow and there is no end in sight. Last week I was in Paris where I visited the Monet museum. Monet was doing his best work at eighty-five. I expect that Steve will be doing the same.

The title of my talk echoes the title of the last Smalefest for Steve's sixtieth birthday "From Topolology to Computation". I met Steve in 1962. He had by this time finished his immersion theorem-turning the sphere inside out, the generalized Poincaré conjecture and the H-cobordism theorem. He had found a horseshoe on the beaches of Rio and had begun his modern restructuring of the geometric theory of dynamical systems, focusing on the global stable and unstable manifolds and their intersections. The period 1958 to 1962 had been incredibly creative for Steve. The number of Steve's remarkable accomplishments still boggles my imagination. And they are not of a whole, one following from the other, but rather disparate independent inventions. By 1962 Steve had already left finite dimensional differential topology. So I missed this wonderful part of his career. But luckily for the subject and for me he had not left dynamical systems. So that is where my story begins. My work in dynamics has been tremendously influenced by Steve. Now after years of collaboration with Steve, Felipe and Lenore on computation, I find that some of the techniques that Steve and I used in our sequence of papers on Bezout's theorem and complexity may be useful again back in dynamics.

What I am reporting on today is joint work with Keith Burns, Charles Pugh, Amie Wilkinson, and Jean-Pierre Dedieu see [6],[7]. Much of what I am saying is taken from these two papers without specific attribution.The material may be found in much expanded form in these two references.

1 Introduction

In his 1967 survey paper on dynamical systems [15] Steve asked for stable properties which hold for most dynamical systems and which in some sense describe the orbit structure of the system. The concepts under study at the time were structural stability and Ω stability, which roughly require that at least on the most dynamically interesting sets the orbit structure of the dynamical system be locally constant under perturbation of the system up to continuous change of coordinates. By work beginning with Steve's work on the horseshoe, Anosov's [3] structural stability theorem and Steve's Ω stability theorem [15], uniform hyperbolicity of the dynamics is known to imply Ω or structural stability. A remarkable feature of these new results, which set them apart from previous work on structural stability by Andronov-Pontryagin and Peixoto, is the complexity of the dynamics encompassed. The horseshoe, strange attractors and globally hyperbolic dynamics are chaotic. They exhibit exponentially sensitive dependence on initial conditions. Thus, while in some sense the future history of a particular orbit may be too difficult to predict, the ensemble of orbits in these stable systems is topologically rigid in its behaviour.

One of the major achievements of the uniformly hyperbolic theory of dynamical systems is the work of Anosov, Sinai, Ruelle and Bowen on the ergodic theory of uniformly hyperbolic systems. Anosov proved that smooth volume preserving globally hyperbolic systems are ergodic. Sinai, Ruelle and Bowen extended this work to specifically constructed invariant measures for general uniformly hyperbolic systems now called SRB measures. The ergodicity of these measures asserts that although particular histories are difficult to compute the statistics of these histories, the probability that a point is in a given region at a given time, is captured by the measure.

Steve's program is however not accomplished, since structurally stable, Ω stable and uniformly hyperbolic systems are not dense in the space of dynamical sysytems [14,2]. Much of the work in dynamical systems in recent years has been an attempt to extend the results of the uniformly hyperbolic theory to more general systems. One theme is to relax the notion of uniform hyperbolicity to non-uniform or partial hyperbolicity and then to conclude the existence of measures sharing ergodic properties of the SRB measures. The Proceedings of the Seattle AMS Summer Symposium on Smooth Ergodic Theory will surely contain much along these lines. In particular, you can find the survey of recent progress on ergodicity of partially hyperbolic systems [6] included there. There is much more available concerning the quadratic family, Henon and Lorenz attractors and more, but I will not try to reference that work here. It is my feeling that much of the work proving the presence of non-uniform hyperbolicity or non-zero Lyapunov exponents (which is the same) is too particular to low dimensions to be able to apply in general.

This paper reports on a result in the theory of random matrices which is an analogue in linear algebra of a mechanism we may hope to use to find non-zero Lyapunov exponents for general dynamical systems.

2 Rich Families

In rich enough families individual members generally inherit family properties.

This sentence which a truism in ordinary language sometimes also applies in mathematics. The first theorem I learned from Steve in his 1962 course on infinite dimensional topology, the Abraham transversality theorem is an example. Let us recall that a smooth map $F : M \to N$ between differentiable manifolds is transversal to the submanifold W of N if $TF(x)(T_xM)$ contains a vector space complement to $T_{F(x)}W$ in $T_{F(x)}N$ for every $x \in M$ such that $F(x) \in W$.

We give a simple finite dimensional version of the Abraham transversality theorem which is valid in infinite dimensions [1]. Let $\mathcal{P}$ be a finite dimensional smooth manifold which we will think of as a space of parameters for a space of maps . Suppose $\Phi : \mathcal{P} \times M \to N$ is a smooth map. For $p \in \mathcal{P}$ let $\Phi_p = \Phi(p, -)$ which is a smooth mapping from M to N. Suppose that Φ is transversal to W. Then $V = \Phi^{-1}(W)$ is a smooth submanifold of $\mathcal{P} \times M$. Let $\Pi_1 : V \to \mathcal{P}$ be the projection of $\mathcal{P} \times M$ onto the first factor restricted to V. The following proposition is then an exercise in counting dimensions of vector spaces.

Proposition 2.1. *With Φ,M,$\mathcal{P}$, N,W and V as above and $p \in \mathcal{P}$; Φ_p is transversal to W on M if and only if p is a regular value of Π_1.*

Now by Sard's Theorem it follows that almost every $p \in \mathcal{P}$ is a regular value of Π_1. So we have proven a version of Abraham's theorem.

Theorem 2.2. *If $\Phi : \mathcal{P} \times M \to N$ is a smooth map transversal to W then Φ_p is transversal to W for almost every $p \in \mathcal{P}$.*

Thus almost every member of the family Φ_p inherits the transversality property from the transversality of the whole family. The richness of the family is expressed by the transversality of the mapping Φ.

Here is another, more dynamical, example of our truism in ergodic theory [11].

Theorem 2.3. *If $\Phi : \mathbb{R}^n \times X \to X$ is an ergodic action of $\mathbb{R}^n$ on a probability space X then for almost every $r \in \mathbb{R}^n$, $\Phi_r : X \to X$ is ergodic.*

The ergodicity of the family is inherited by almost all elements. Further examples of our truism in ergodic theory are provided by the Mautner phenomenon [9],[5]. Both Theorem 2.3 and the Mautner phenomenon are proven via representation theory. The richness in the family comes from the Lie group structure and the ergodicity of the group.

We would like to have a notion of *richness of a family of dynamical systems* and *Lyapunov exponent of the family* so as to be able *to conclude that most or at least many of the elements of the family have some non-zero exponents when the family does.* For the notion of Lyapunov exponent of the family we shall use the exponents of random products of elements of the family with respect to a probability measure on the space of systems.

We begin in the next section with linear maps where we use as a notion of richness the unitary invariance of the probability distribution on the space of matrices.

3 Unitarily invariant measures on $\mathbb{GL}_n(\mathbb{C})$

Let L_i be a sequence of linear maps mapping finite dimensional normed vector spaces V_i to V_{i+1} for $i \in \mathbb{N}$. Let $v \in V_0\backslash\{0\}$. If the limit $\lim \frac{1}{k} \log ||L_{k-1} \ldots L_0(v)||$ exists it is called a Lyapunov exponent of the sequence. It is easy to see that if two vectors have the same exponent then so does every vector in the space spanned by them. It follows that there are at most $dim(V_0)$ exponents. We denote them λ_j where $j \leq k \leq dim(V_0)$. We order the λ_i so that $\lambda_i \geq \lambda_{i+1}$ Thus it makes sense to talk about the Lyapunov exponents of a diffeomorphism f of a compact manifold M at a point $m \in M$, $\lambda_j(f, m)$ by choosing the sequence L_i equal to $Tf(f^i(m))$.

Given a probability measure μ on $\mathbb{GL}_n(\mathbb{C})$ the space of invertible $n \times n$ complex matrices we may form infinite sequences of elements chosen at random from μ by taking the product measure on $\mathbb{GL}_n(\mathbb{C})^{\mathbb{N}}$. Thus we may also talk about the Lyapunov exponents of sequences or almost all sequences in $\mathbb{GL}_n(\mathbb{C})^{\mathbb{N}}$.

Oseledec's Theorem applies in our two contexts.

For diffeomorphisms f Oseldec's theorem says that for any f invariant measure ν, for ν almost all $m \in M$, f has $dim(M)$ Lyapunov exponents at

m, $\lambda_j(f,m)$ for $1 \le j \le dim(M)$.

For measures μ on $\mathbb{GL}_n(\mathbb{C})$ satisfying a mild integrability condition, we have n Lyapunov exponents $r_1 \ge r_2 \ge \ldots \ge r_n \ge -\infty$ such that for almost every sequence $\ldots g_k \ldots g_1 \in \mathbb{GL}_n(\mathbb{C})$ the limit $\lim \frac{1}{k} \log \|g_k \ldots g_1 v\|$ exists for every $v \in \mathbb{C}^n \setminus \{0\}$ and equals one of the r_i, $i = 1 \ldots n$, see Gol'dsheid and Margulis [8] or Ruelle [12] or Oseledec [10]. We may call the numbers $r_1, \ldots, r_n$ random Lyapunov exponents or even just random exponents. If the measure is concentrated on a point A, these numbers $\lim \frac{1}{n} \log \|A^n v\|$ are $\log|\lambda_1|, \ldots,$ $\log|\lambda_n|$ where $\lambda_i(A) = \lambda_i$, $i = 1 \ldots n$, are the eigenvalues of A written with multiplicity and $|\lambda_1| \ge |\lambda_2| \ge \ldots \ge |\lambda_n|$.

The integrability condition for Oseledec's Theorem is

$$g \in \mathbb{GL}_n(\mathbb{C}) \to \log^+(\|g\|) \text{ is } \mu - \text{integrable}$$

where for a real valued function f, $f^+ = \max[0, f]$. Here we will assume more so that all our integrals are defined and finite, namely:

$$(*) \qquad g \in \mathbb{GL}_n(\mathbb{C}) \to \log^+(\|g\|) \text{ and } \log^+(\|g^{-1}\|) \text{ are } \mu\text{-integrable.}$$

In [7] we prove:

Theorem 3.1. *If μ is a unitarily invariant measure on $\mathbb{GL}_n(\mathbb{C})$ satisfying $(*)$ then, for $k = 1, \ldots, n$,*

$$\int_{A \in \mathbb{GL}_n(\mathbb{C})} \sum_{i=1}^{k} \log|\lambda_i(A)| d\mu(A) \ge \sum_{i=1}^{k} r_i.$$

By unitary invariance we mean $\mu(U(X)) = \mu(X)$ for all unitary transformations $U \in \mathbb{U}_n(\mathbb{C})$ and all μ-measurable $X \in \mathbb{GL}_n(\mathbb{C})$.

Thus non-zero Lyapunov exponents for the family, i.e. non-zero random exponents, implies that at least some of the individual linear maps have non-zero exponents i.e eigenvalues. The notion of richness here is unitary invariance of the measure. For complex matrices we have achieved part of our goal. Later we will suggest a way in which these results may be extended to dynamical systems.

Corollary 3.2.

$$\int_{A \in \mathbb{GL}_n(\mathbb{C})} \sum_{i=1}^{n} \log^+|\lambda_i(A)| d\mu(A) \ge \sum_{i=1}^{n} r_i{}^+.$$

Theorem 3.1 is not true for general measures on $\mathbb{GL}_n(\mathbb{C})$ or $\mathbb{GL}_n(\mathbb{R})$ even for $n = 2$. Consider

$$A_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

and give probability 1/2 to each. The left hand integral is zero but as is easily seen the right hand sum is positive. So, in this case the inequality goes the other way. We do not know a characterization of measures which make Theorem 3.1 valid.

In order to prove 3.1 we first identify the the right hand summation in terms of an integral. Let $\mathbb{G}_{n,k}(\mathbb{C})$ denote the Grassmannian manifold of k dimensional vector subspaces in $\mathbb{C}^n$. If $A \in \mathbb{GL}_n(\mathbb{C})$ and $G_{n,k} \in \mathbb{G}_{n,k}(\mathbb{C})$, $A|G_{n,k}$ the restriction of A to the subspace $G_{n,k}$. Let ν be the natural unitarily invariant probability measure on $\mathbb{G}_{n,k}(\mathbb{C})$. The next theorem is a fairly standard fact.

Theorem 3.3. *If μ is a unitarily invariant probability measure on $\mathbb{GL}_n(\mathbb{C})$ satisfying $(*)$ then,*

$$\sum_{i=1}^{k} r_i = \int_{A\in\mathbb{GL}_n(\mathbb{C})} \int_{G_{n,k}\in\mathbb{G}_{n,k}(\mathbb{C})} \log|\det(A|G_{n,k})| d\nu(G_{n,k}) d\mu(A).$$

We may then restate Theorem 3.1.

Theorem 3.4. *If μ is a unitarily invariant probability measure on $\mathbb{GL}_n(\mathbb{C})$ satisfying $(*)$ then, for $k = 1, \ldots, n$*

$$\int_{A\in\mathbb{GL}_n(\mathbb{C})} \sum_{i=1}^{k} \log|\lambda_i(A)| d\mu(A) \geq$$

$$\int_{A\in\mathbb{GL}_n(\mathbb{C})} \int_{G_{n,k}\in\mathbb{G}_{n,k}(\mathbb{C})} \log|\det(A|G_{n,k})| d\nu(G_{n,k}) d\mu(A).$$

Theorems 3.4 reduces to a special case.

Let $A \in \mathbb{GL}_n(\mathbb{C})$ and μ be the Haar measure on $\mathbb{U}_n(\mathbb{C})$ (the unitary subgroup of $\mathbb{GL}_n(\mathbb{C})$) normalized to be a probability measure. In this case Theorem 3.4 becomes:

Theorem 3.5. *Let $A \in \mathbb{GL}_n(\mathbb{C})$. Then, for $1 \leq k \leq n$,*

$$\int_{U\in\mathbb{U}_n(\mathbb{C})} \sum_{i=1}^{k} \log|\lambda_i(UA)| d\mu(U) \geq \int_{G_{n,k}\in\mathbb{G}_{n,k}(\mathbb{C})} \log|\det(A|G_{n,k})| d\nu(G_{n,k})$$

We expect similar results for orthogonally invariant probability measures on $\mathbb{GL}_n(\mathbb{R})$ but we have not proven it except in dimension 2.

Theorem 3.6. *Let μ be a probability measure on $\mathbb{GL}_2(\mathbb{R})$ satisfying*

$$g \in \mathbb{GL}_2(\mathbb{R}) \to \log^+(\|g\|) \textit{ and } \log^+(\|g^{-1}\|) \textit{ are } \mu-\textit{integrable.}$$

a. If μ is a $\mathbb{SO}_2(\mathbb{R})$ invariant measure on $\mathbb{GL}_2^+(\mathbb{R})$ then,

$$\int_{A\in\mathbb{GL}_2^+(\mathbb{R})} \log|\lambda_1(A)| d\mu(A) = \int_{A\in\mathbb{GL}_2^+(\mathbb{R})} \int_{x\in\mathbb{S}^1} \log\|Ax\| d\mathbb{S}^1(x) d\mu(A).$$

b. If μ is a $\mathbb{SO}_2(\mathbb{R})$ invariant measure on $\mathbb{GL}_2^-(\mathbb{R})$, whose support is not contained in $\mathbb{RO}_2(\mathbb{R})$ i.e. in the set of scalar multiples of orthogonal matrices, then

$$\int_{A\in\mathbb{GL}_2^-(\mathbb{R})} \log|\lambda_1(A)|d\mu(A) > \int_{A\in\mathbb{GL}_2^-(\mathbb{R})}\int_{x\in\mathbb{S}^1} \log\|Ax\|d\mathbb{S}^1(x)d\mu(A).$$

Here $\mathbb{GL}_2^+(\mathbb{R})$ (resp. $\mathbb{GL}_2^-(\mathbb{R})$) is the set of invertible matrices with positive (resp. negative) determinant.

4 Proofs and the Complexity of Bezout's Theorem

In our series of papers on complexity and Bezout's theorem, Steve and I concentrated on the manifold of solutions $V = \{(P,z) \in \mathbb{P}(H_{(D)}) \times \mathbb{P}(\mathbb{C}^n) | P(z) = 0\}$ and the two projections

$$\begin{tikzcd} & V \arrow[dl, "\Pi_1"'] \arrow[dr, "\Pi_2"] & \\ \mathbb{P}(H_{(D)}) & & \mathbb{P}(\mathbb{C}^n) \end{tikzcd}$$

in order to transfer integrals over $\mathbb{P}(H_{(D)})$ to integrals over $\mathbb{P}(\mathbb{C}^n)$. See [4]

Here $(D) = (d_1, \cdots, d_{n-1})$ and $H_{(D)}$ is the vector space of homogeneous polynomials systems $P = (P_1, \ldots, P_{n-1})$ where each P_i is a homogeneous polynomial of degree d_i in n complex variables. For a vector space V, $\mathbb{P}(V)$ denotes the projective space of V.

Our proof of 3.5 relies heavily on this technique, but with respect to a manifold of fixed points.

A flag F in $\mathbb{C}^n$ is a sequence of vector subspaces of $\mathbb{C}^n$: $F = (F_1, F_2, \ldots, F_n)$, with $F_i \subset F_{i+1}$ and $\dim F_i = i$. The space of flags is called the flag manifold and we denote it by $\mathbb{F}_n(\mathbb{C})$. An invertible linear map $A : \mathbb{C}^n \to \mathbb{C}^n$ naturally induces a map $A_\sharp$ on flags by

$$A_\sharp(F_1, F_2, \ldots, F_n) = (AF_1, AF_2, \ldots, AF_n).$$

The flag manifold and the action of a linear map A on $\mathbb{F}_n(\mathbb{C})$ is closely related to the QR algorithm, see [13] for a discussion of this. In particular if F is a fixed flag for A i.e. $A_\sharp F = F$, then A is upper triangular in a basis corresponding to the flag F, with the eigenvalues of A appearing on the diagonal in some order: $\lambda_1(A,F)$, ..., $\lambda_n(A,F)$.

Let

$$\mathbb{V}_A = \{(U,F) \in \mathbb{U}_n(\mathbb{C}) \times \mathbb{F}_n(\mathbb{C}) \ : \ (UA)_\sharp F = F\}.$$

We denote by Π_1 and Π_2 the restrictions to $\mathbb{V}_A$ of the projections $\mathbb{U}_n(\mathbb{C}) \times \mathbb{F}_n(\mathbb{C}) \to \mathbb{U}_n(\mathbb{C})$ and $\mathbb{U}_n(\mathbb{C}) \times \mathbb{F}_n(\mathbb{C}) \to \mathbb{F}_n(\mathbb{C})$. $\mathbb{V}_A$ is a manifold of fixed points. We use the diagram

$$\begin{tikzcd}
& \mathbb{V}_A \arrow[dl, "\Pi_1"'] \arrow[dr, "\Pi_2"] & \\
\mathbb{U}_n(\mathbb{C}) & & \mathbb{F}_n(\mathbb{C})
\end{tikzcd}$$

in order to transfer the right hand integral in 3.5 over $\mathbb{F}_n(\mathbb{C})$ to an integral over $\mathbb{U}_n(\mathbb{C})$.

5 A dynamical systems analogue

Is there a notion of richness for a family $\mathcal{P}$ of diffeomorphisms of a compact manifold M which would allow us to conclude that at least some members of the family have non-zero exponents?

We introduce now a notion of richness of $\mathcal{P}$ which might, in some situations, be sufficient to deduce properties of the exponents of elements of $\mathcal{P}$ from those of the random exponents. This notion was suggested to us by some preliminary numerical experiments and by the results in the setting of random matrix products in section 3.

We focus on the problem for $M = S^n$, the n-sphere. Let μ be Lebesgue measure on S^n normalized to be a probability measure, and let m be Liouville measure on $T_1(S^n)$, the unit tangent bundle of S^n, similarly normalized to be a probability measure. The orthogonal group $O(n+1)$ acts by isometries on the n-sphere and so induces an action on the space of μ-preserving diffeomorphisms by

$$f \mapsto O \circ f, \qquad \text{for } O \in O(n+1).$$

Let ν be a probability measure supported on $\mathcal{P} \subset \mathrm{Diff}^r_\mu(S^n)$. We say that ν is *orthogonally invariant* if ν is preserved by every element of $O(n+1)$ under the action described above.

For example, let

$$\mathbb{F}_n(\mathbb{C}) = O(n+1)f = \{O \circ f \mid O \in O(n+1)\},$$

for a fixed $f \in \mathrm{Diff}^r_\mu(S^n)$ Defining ν by transporting Haar measure on $O(n+1)$ to $\mathcal{P}$, we obtain an orthogonally-invariant measure. Because $O(n+1)$ acts transitively on $T_1(S^n)$, a random product of elements of $\mathcal{P}$ will pick up the behavior of f in almost all tangent directions — the family is reasonably rich in that sense.

Let ν be an orthogonally invariant measure on $\mathcal{P}$. The largest random Lyapunov exponent for $\mathcal{P}$, which we will denote by $R(\nu)$, can be expressed as an integral:

$$R(\nu) = \int R(\nu, x)\, d\mu = \int_{\mathrm{Diff}^r_\mu(S^n)} \int_{T_1(S^n)} \ln \|Df(x)v\|\, dm\, d\nu.$$

We define the *mean largest Lyapunov exponent* to be

$$\Lambda(\nu) = \int_{\mathrm{Diff}^r_\mu(S^n)} \int_{S^n} \lambda_1(f, x)\, d\mu\, d\nu$$

where $\lambda_1(f, x)$ is the largest Lyapunov exponent of f at x.

Question 5.1. Is there a positive constant $C(n)$ — perhaps 1 — depending on n alone such that $\Lambda(\nu) \geq C(n)R(\nu)$?

If the answer to Question 5.1 were affirmative, then a positive measure set of elements of $\mathcal{P}$ would have areas of positive exponents, (assuming a mild nondegeneracy condition on ν). We add here that this type of question has been asked before and has been the subject of a lot of research. What is new is the notion of richness which allows us to express the relation between exponents as an inequality of integrals.

The question is already interesting for S^2. Express S^2 as the sphere of radius 1/2 centered at $(1/2, 0)$ in $\mathbf{R} \times \mathbf{C}$, so that the coordinates $(r, z) \in S^2$ satisfy the equation

$$|r - 1/2|^2 + |z|^2 = 1/4.$$

In these coordinates define a twist map $f_\epsilon : S^2 \to S^2$, for $\epsilon > 0$, by

$$f_\epsilon(r, z) = (r, \exp(2\pi i r \epsilon) z).$$

Let $\mathcal{P}$ be the orbit $O(3)f$ and let ν be the push forward of Haar measure on $O(3)$. A very small and inconclusive numerical experiment seemed to indicate that for ϵ close to 0 the inequality may hold with $C(n) = 1$. It seemed the constant may decrease as the twist increases speed.

Michel Herman thinks Question 5.1 has a negative answer, precisely for the twist map example f_ϵ, for ϵ very small due to references cited in section 6 of [6]. Perhaps more and better experiments would shed some light on the question. Whether or not Herman is correct, it would be interesting to know if other lower bound estimates are available with an appropriate concept of richness of the family.

Acknowledgment

Partially supported by NSF Grant #DMS-9616120.

References

1. Abraham, R. and J. Robbin, *Transversal Mappings and Flows*, W. A. Benjamin, Inc., New York-Amsterdam 1967.
2. Abraham, R. and S. Smale, *Nongenericity of Ω-stability*, 1970 Global Analysis (Proc. Sympos. Pure Math., Vol. XIV, Berkeley, Calif., 1968) pp. 5–8 Amer. Math. Soc., Providence,R.I.
3. Anosov, D. V., *Geodesic flows on closed Riemannian manifolds of negative curvature*, Proc. Steklov. Inst. Math. **90** (1967).
4. Blum L., F. Cucker, M. Shub, S. Smale *Complexity and Real Computation*, Springer, 1998.
5. Brezin, J. and C. Moore, *Flows on homogeneous spaces: a new look.* Amer. J. Math. **103** (1981), 571–613.
6. Burns K., C. Pugh, M. Shub and A. Wilkinson, *Recent Results about Stable Ergodicity*, to appear in: Proceedings on Symposia in Pure Mathematics, the Seattle Conference of Smooth Ergodic Theory, AMS.
7. Dedieu, J.P. and M. Shub, *On random and mean exponents for unitarily invariant probability measures on $GL(n, \mathbf{C})$*, preprint.
8. Gol'shied I. Ya. and G. A. Margulis, *Lyapunov Indices of a Product of Random Matrices*, Russian Math. Surveys **44:5** (1989), pp. 11-71.
9. Moore, C.C. *Ergodicity of flows on homogeneous spaces*, Amer. J. Math. **88** (1966), 154–178.
10. Oseledec, V. I., *A Multiplicative Ergodic Theorem. Lyapunov Characteristic Numbers for Dynamical Systems*, Trans. Moscow Math. Soc., 19, (1968), pp. 197-231.
11. Pugh, C. and M. Shub, *Ergodic elements of ergodic actions*, Compositio Math. Vol. 23 (1971), 115–121.
12. Ruelle, D., *Ergodic Theory of Differentiable Dynamical Systems*, Publications Mathématiques de l'IHES, Volume 50, (1979), pp. 27-58.
13. Shub M. and A. Vasquez, *Some Linearly Induced Morse-Smale Systems, the QR Algorithm and the Toda Lattice*, in: The Legacy of Sonia Kovalevskaya, Linda Keen Ed., Contemporary Mathematics, Vol. 64, AMS, (1987), pp. 181-193.
14. Smale, S., *Structurally stable systems are not dense*, Amer. J. Math. **88** (1966), 491–496.
15. Smale, S., *Differentiable dynamical systems*, Bull.A.M.S. **73** (1967), 747-817.

SIMULTANEOUS COMPUTATION OF ALL THE ZERO-CLUSTERS OF A UNIVARIATE POLYNOMIAL

JEAN-CLAUDE YAKOUBSOHN
Laboratoire MIP
Université Paul Sabatier
118, Route de Narbonne
31062 Toulouse Cedex, France
e-mail: yak@mip.ups-tlse.fr

We first prove a conjecture concerning the local behaviour of the Weierstrass method to approximate simultaneously all the zeroes of a univariate polynomial. This conjecture states that the convergence of each coordinate of the Weierstrass sequence depends only on the multiplicity of each root. There is quadratic convergence toward the simple roots. We prove that if there is a root of multiplicity m then there are m coordinates of the Weierstrass sequence which converge geometrically with a limit ratio $\frac{m-1}{m}$. We reformulate this numerical result in terms of the notion of zero-clusters, which is a more natural way to study this problem. We next combine numerical path-following with Weierstrass method to approximate all the zero-clusters. These theoretical results are illustrated by some numerical experiments. We also apply this process to approximate all the clusters of the eigenvalues of a matrix, without the computation of the characteristic polynomial.

1 Introduction

The purpose of this paper is to study the computation of all the zero-clusters of a univariate polynomial by the Weierstrass method. This method was introduced in [21] to give a constructive proof of the fundamental theorem of algebra. The Weierstrass method consists of applying Newton's method to the system of Viete symmetric functions associated to a polynomial $f(x) = \sum_{k=0}^{d} a_k z^k \in \mathbb{C}[z]$. For $x = (x_1, \ldots, x_d) \in \mathbb{C}^d$ denote by $S(x)$ the n-tuple $(S_1(x), \ldots, S_d(x))$ where $S_i(x) := \left(\sum_{1 \leq i_1 < \ldots < i_k \leq d} x_{i_1} \cdots x_{i_k}\right) - (-1)^k \frac{a_{d-k}}{a_d}$ for all i. The number of zeroes of this system is $d!$ and the set of zeroes of $S(x)$ is equal to $Z(S) = \{w = (w_1, \ldots, w_d) \in \mathbb{C}^d \mid f(w_i) = 0,\ 1 \leq i \leq d\}$. The Weierstrass function is the function defined by

$$x \in \mathbb{C}^d \to W(f, x) = x - (DS(x))^{-1} S(x) \in \mathbb{C}^d.$$

The Weierstrass sequence is the sequence of points in $\mathbb{C}^d$ defined by

$$x^0 \in \mathbb{C}^d, \ldots, x^{k+1} = W(f, x^k), \quad k \geq 0.$$

We say that the Weierstrass method converges iff the sequence $\left(x^k\right)_{k \geq 0}$ converges towards some point $w \in Z(S)$. Clearly, the Weierstrass method can be used to approximate all the roots of f simultaneously.

In this context, it is fundamental to know a condition for the quadratic convergence of the Weierstrass sequence. In terms of the modern analysis of the Newton method by the alpha-theory of S. Smale [4], we can state the following theorem.

Theorem 1 *Let $f \in \mathbb{C}[x]$ be a polynomial whose roots are all simple. Let us consider $w \in Z(S)$ and the quantity*

$$\gamma(S,w) = \max_{k\geq 2}\left(\frac{||DS^{-1}S(w)D^kS(w)||}{k!}\right)^{\frac{1}{k-1}}.$$

Then for all $x \in \mathbb{C}^d$ satisfying the inequality

$$||x-w|| \leq \frac{3-\sqrt{7}}{2\gamma(S,w)}$$

the Weierstrass sequence initialized at $x^0 = x$ converges towards w. Moreover

$$||x^k - w|| \leq \left(\frac{1}{2}\right)^{2^k-1} ||x^0 - w||.$$

As it is mentioned, this result holds in the case of simple roots. The purpose of this paper is to investigate the general case of zero-clusters. In fact, the case of multiple zeroes is meaningless, from a numerical analysis point of view, see [24]. For that we will use the explicit formulas for the Weierstrass function. It is well-known how to calculate formally the inverse of $DS(x)$ (see [3]). We have:

$x \in \mathbb{C}^d \to W(f,x) = \left(x_1 - \frac{f(x_1)}{a_d\prod_{j\neq 1}(x_1-x_j)}, \dots, x_d - \frac{f(x_d)}{a_d\prod_{j\neq d}(x_d-x_j)}\right) \in \mathbb{C}^d.$

A very good survey of the Weierstrass method and possible extensions is done by Sendov, Andreev, and Kjurkchiev in [18]. We recall the result obtained by Kjurkchiev and Markov in 1983.

Theorem 18.1 (See page 700 of [18].) *Let $f \in \mathbb{C}[x]$ be a polynomial of degree d with roots $w_1, \dots, w_d$. Let $0 < q < 1$, $s = \min_{i\neq j}|w_i - w_j|$ and $0 < c < \frac{s}{1+d\alpha}$ where $\alpha = 1.7632283...$ is determined from the equality $\alpha = e^{\frac{1}{\alpha}}$. If the initial approximations $x_i^0, \dots, x_d^0$ of the roots $w_1, \dots, w_d$ of $f(x) = 0$ satisfy the inequalities*

$$|x_i^0 - w_i| \leq cq, \quad i = 1, \dots, d,$$

then the sequence $\{x^k\}_{k\geq 0} = \{(x_1^k, \dots, x_d^k)\}_{k\geq 0}$ satisfies the inequalities

$$|x_i^k - w_i| \leq cq^{2^k}, \quad i = 1, \dots, d.$$

We will state an alternative version of the above result where knowledge of the spacings of the roots is unneeded: we instead use the quantity $\gamma(f)$ defined below and introduced by Smale [4] (see theorem 2 of the next section). Indeed, the hypotheses of our result will ensure that the quantity $s = \min_{i \neq j} |w_i - w_j|$ is positive. This is done through the point estimate $s > \frac{1}{2\gamma(f)}$ (see Lemma 1).

To understand the convergence of Weierstrass method in the general case, consider the polynomial $f(z) = z^m$. Then $W(f,x) = \frac{m-1}{m}x$. Each coordinates, say x_i^k, of the sequence $x^{k+1} = W(z^m, x^k)$ converges towards 0 with a geometric rate of convergence $\frac{m-1}{m}$ following the straight line $[x_i^0, 0]$.

The main result of this paper is that the previous example describes the general case, in the following sense: the behaviour of the Weierstrass sequence x^k towards a zero-cluster depends of the multiplicity of this cluster. There is a quadratic convergence if the cluster is a simple root and a geometric convergence if the cluster contains m zeroes: see Theorem 3.

In the second part of this paper, we combine a path-following method with the Weierstrass method in order to follow a curve which finishes to a point of $Z(S)$. The complexity of this homotopy method is given in theorem 4. We next give some numerical experiments and we show an application of this process to the computation of all the eigenvalues of a matrix without computation of coefficients of characteristic polynomial.

We conclude this introduction by some short remarks and historical comments to explain the context and the new results of this paper. Globally the ideas of this study are those developed by Weierstrass in [21]. In fact, Weierstrass has used the Newton iteration applied to the Viete symmetric functions system to give a constructive proof of the Fundamental Theorem of Algebra: in the case of simple roots, he first proves the local quadratic convergence of the method. He then introduced a classical linear homotopy with rational subdivision of the time interval to conclude. Many authors have re-discovered this method which is also known as the Durand-Kerner's method [7], [12], [13]. Most of the earlier literature only studied the local behaviour of this method in the context of circular arithmetic: see [17] and [16]. The use of classical linear homotopy or other can be found in [20] and [6]. The cited authors were primarily interested in polynomials with simple roots. The dependence of the rate of convergence of the Durand-Kerner method on the multiplicity of the root has been numerically observed in [9],[14],[10],[2]. However, the proof of a precise numerical result along these lines has not appeared before in the literature.

2 Main Results

Denote by

$$f(z) = \sum_{k=0}^{d} a_k z^k = a_d \prod_{i=1}^{d} (z - w_i)$$

a complex polynomial of degree $d \geq 2$. We will let $w = (w_1, \dots, w_d)$. The analysis of the convergence will be done with respect to the quantities

$$\beta_m(f,z) = \max_{0 \leq k \leq m-1} \left| \frac{m! f^{(k)}(z)}{k! f^{(m)}(z)} \right|^{\frac{1}{m-k}}, \quad z \in \mathbb{C}, m \geq 1,$$

$$\gamma_m(f,z) = \max_{k \geq m+1} \left| \frac{m! f^{(k)}(z)}{k! f^{(m)}(z)} \right|^{\frac{1}{k-m}}, \quad z \in \mathbb{C}, m \geq 1.$$

These quantities were first introduced in [24]. We first treat the case where all the roots of polynomial f are simple and state a γ-theorem, see [4] Theorem 1 page 156. For that we introduce the quantity:

$$\gamma(f) = \max_{1 \leq i \leq d} \gamma_1(f, w_i).$$

Theorem 2 *Let us consider $x = (x_1, \dots, x_d) \in \mathbb{C}^d$ be such that*

$$|x_i - w_i| \leq \frac{1}{(5d-2)\gamma(f)}, \quad 1 \leq i \leq d.$$

Then the Weierstrass sequence defined by

$$x^0 = x, \quad x^{k+1} = W(f, x^k), \quad k \geq 0$$

converge towards w with error bound:

$$|x_i^k - w_i| \leq \left(\frac{1}{2}\right)^{2^k - 1} |x_i^0 - w_i|, \quad 1 \leq i \leq d.$$

This result is based on the following lemma which appears in [24].

Lemma 1 *Following the notation above, if $w_i \neq w_j$ then $|w_i - w_j| \geq \frac{1}{2\gamma(f)}$.*

We now consider the case of zeroes clusters. Recall that an m-cluster of f is a disk which contains m roots of f. We will denote by $D(x,r)$ any disk around x of radius r and $cD(x,r)$ the disk $D(cx, cr)$. We assume without loss of generality that the roots of f can be partitioned to lie in p disjoint clusters where, for all $i \in \{1, \dots, p\}$, the $i^{\underline{\text{th}}}$ cluster is an m_i-cluster $D(z_i, r)$. Denote the roots in the m_i-cluster $D(z_i, r)$ by $w_{i1}, \dots, w_{im_i}$ and let $M_i =$

$\{1, \ldots, m_i\}$. Clearly $\sum_{i=1}^{p} m_i = d$. In this context, we will write $x \in \mathbb{C}^d$ as $x = \left(x_{11}, \ldots, x_{1m_1}, \ldots, x_{p1}, \ldots, x_{pm_p}\right)$ and $W(f, x)$ becomes

$$W(f,x) = \left(W_{11}(f,x), \ldots W_{1m_1}(f,x), \ldots, W_{p1}(f,x), \ldots W_{pm_p}(f,x)\right).$$

Consequently the coordinates of $x^{k+1} = W(f, x^k)$ will be $x_{ij}^{k+1} = W_{ij}(x^k)$, $1 \leq j \leq m_i$ and $1 \leq i \leq p$. The quantities $\beta(f)$ and $\gamma(f)$ above are now define as:

$$\beta(f) = \max_{1 \leq i \leq p} \beta_{m_i}(f, z_i), \quad \gamma(f) = \max_{1 \leq i \leq p} \gamma_{m_i}(f, z_i).$$

Also, introduce

$$\alpha_{m_i}(f, z_i) = \beta_{m_i}(f, z_i)\gamma_{m_i}(f, z_i), \quad \alpha(f) = \beta(f)\gamma(f).$$

We will also define the function

$$\varphi(m,u) = \begin{cases} \left(\frac{m-1}{m} + \frac{2(m^2-1)u}{m(2-(m-2)u)} + \frac{2(d-m)(1+u)^{m-1}v}{m(2-(d-m-1)v)} + \frac{(1+u)^{m-1}(1+v)^{d-m}u}{m}\right) \frac{1}{1-\frac{2(m-1)u}{2-(m-2)u}} & (m > 1) \\ \dfrac{2(d-1)u}{1-(d+2)u} & (m = 1), \end{cases}$$

with $v = \frac{2u(1+u)}{1-a-4u-2u^2}$ and $a = 3\alpha(f)$.

Theorem 3 *Let r, R, and u be positive real number such that $u = \gamma(f)R$ and*

$$\frac{r}{R\sqrt{1 - \left(\frac{r}{R}\right)^2 \left(\frac{2}{m} - \frac{r}{R}\right)}} \leq u.$$

Suppose also that $9\alpha(f) \leq 1$, $\varphi(m_i, u) < 1$ for all $i = 1, \ldots, p$, and that the radius r of the m_i-cluster satisfies

$$r > 3\beta(f)$$

for all i.

Choose p complex numbers $y_{11}, \ldots y_{p1}$ lying respectively on the circles $S(z_1, R), \ldots S(z_p, R)$. Next, for $i = 1, \ldots p$, consider respectively the roots $y_{i1}, \ldots y_{im_i}$ of the equation in z:

$$(z - z_i)^{m_i} = (y_{i1} - z_i)^{m_i}.$$

Then the Weierstrass sequence

$$x^0 = y, \quad x^{k+1} = W(f, x^k), \quad k \geq 0,$$

is well defined and converges toward w as follows:

1. *For all i such that $m_i > 1$ and for all x_{ij}^k, $j \in M_i$, which satisfies $r < u|x_{ij}^k - z_i|$, we have:*

$$|x_{ij}^k - z_i| \leq \varphi(m_i, u)^k (R + r), 1 \leq j \leq m_i, \quad k \geq 0.$$

Moreover each x_{ij}^k, $j \in M_i$, lies in the disk $\varphi(m_i, u)^k D(x_{ij}^0, r)$.

2. *On other hand, if w_i is a simple root ($m_i = 1$), we have*

$$|x_{i1}^k - w_i| \leq \left(\frac{2(d-1)u}{1-(d+2)u}\right)^{2^k - 1} R, \quad k \geq 0.$$

We now say how to simultaneously find all the clusters. For that, we introduce the homotopy studied in [4],[23], and [24]. Denote by $\Sigma = \{x = (x_1, \ldots, x_d) \in \mathbb{C}^d \mid \exists i \neq j, \quad x_i = x_j\}$. Let $z^0 \in \mathbb{C}^d - \Sigma$. For $t \in [0, 1]$, let us consider the family of map S_t defined by:

$$x \in \mathbb{C}^d \to S_t(x) = S(x) - tS(z^0).$$

A straightforward computation shows that the polynomial f_t associated to S_t is equal to

$$f_t(z) = (1-t)f(z) + tg(z),$$

with $g(z) = a_d z^d + \sum_{k=0}^{d-1} (-1)^{d-k} a_d S_{d-k}(z^0) z^k$. Hence the homotopy defined with the function S_t induces a linear homotopy with the polynomials $f(z)$ and $g(z)$. Let $w^t = (w_1^t, \ldots, w_d^t)$ the curve of $\mathbb{C}^d$ such that $S_t(w^t) = 0$. The linear map $DS_t(w_t)$ is invertible for all $t \in]0, 1]$. To follow numerically the curve w^t, consider a positive real number $M < 1$, the sequences $(t_k)_{k \geq 0}$ and $(z^k)_{k \geq 0}$ defined respectively by

$$t_0 = 1, t_k = M^k,$$
$$z^{k+1} = W(f_{k+1}, z^k)$$

where $f_k = f_{t_k}$. The purpose is to quantify the computational complexity of the preceding numerical path-following method. More precisely, we estimate the number of steps k such that each coordinates z_i^k of z^k is a point which is closed to a zero of f. Toward this end, let us introduce the following quantities where $u > 0$ and i are given a priori:

- $D(x_i, r)$ is an m_i-cluster which contains the root $w_i = w_i^0$ of f. If $m_i = 1$ then $x_i = w_i$ and $r = 0$.

- Let R be such that $u = \gamma(f)R$.
- $t_i^+ = \sup\{t \in]0,1] \quad | \quad \left|w_i^{t_i^+} - w_i\right| = R\}$.
- $g = \max_i \max_{t_i^+ \leq t \leq 1} \gamma(f_t)$.
- $b = \max_i \left(\max_{t_i^+ \leq t \leq 1} \left|\frac{f(w_i^t) - g(w_i^t)}{f_t'(w_i^t)}\right|\right), \quad a = bg$.
- $b_i = \max_{0 \leq t \leq 1} \left|\frac{m_i! g(w_i^t)}{f^{(m_i)}(w_i^t)}\right|, \quad a_i = b_i \gamma(f)$.
- $M = 1 - \frac{u(1-2u)}{a(1-u)} > 0$.
- $h(u) = \frac{(1-3u)u}{1-u}$.

Theorem 4 *Suppose $u \leq \frac{1}{5d-2}$, $4ug < 1$ and $\beta(f) \leq uR$. Let i be given and denote by k_i the index satisfying $t_{k_i} \leq t_i^+ \leq t_{k_i - 1}$. Then, the following assertions hold:*

1. *For all k such that $t_k \geq t_i^+$, the points z_i^k's are well defined and are approximate zeroes of w_i^k.*
2. *The value t_i^+ is bounded by*
$$\frac{h(u)R^{m_i - 1}}{a_i + h(u)R^{m_i-1}} \leq t_i^+$$
3. $|z^{k_i} - w_i| \leq R + \frac{u}{g}$.

3 Practical Algorithms and Numerical Experiments

Our first numerical experiment will be performed on the polynomial $f(z)$ given by $z^{15} + (-6.30211 - 6.51486\,i)\,z^{14} + (-4.602131368 + 34.33661782\,i)\,z^{13} + (63.59234833 - 17.54183753\,i)\,z^{12}$ $+ (7.872508933 - 100.8166243\,i)\,z^{11} + (-222.2761285 - 64.53993948\,i)\,z^{10} + (-167.1645824 + 289.8629099\,i)\,z^{9}$ $+ (388.8936599 + 490.8785827\,i)\,z^{8} + (642.5739956 - 425.4499699\,i)\,z^{7} + (78.15261237 - 888.3687140\,i)\,z^{6}$ $+ (-1127.773435 - 316.6287328\,i)\,z^{5} + (-750.6198536 + 605.8898563\,i)\,z^{4} + (441.7642441 + 974.1649811\,i)\,z^{3}$ $+ (422.0315160 + 163.1681751\,i)\,z^{2} + (416.5648287 - 448.9257391\,i)\,z + 8.41840810 - 48.40497033\,i$
This polynomial has the five i-clusters $D(z_i, 0.05)$ with:
$z_1 = -0.071 + 0.043i \;,\; z_2 = -1.227 + 0.179i \;,\; z_3 = -0.686 - 0.855i$
$z_4 = 0.388 + 1.825i \;,\; z_5 = 1.866 + 0.275i.$

We first show the behaviour of the Weierstrass iteration around the 5-cluster $D(z_5, 0.05)$. For that the sequence x^k is initialized with the coordinates of x^0 specified as follows: $x^0_{kj} = z_i + 0.1e^{i\frac{2j\pi}{k}}, \quad 1 \leq j \leq k, \quad 1 \leq k \leq 5.$ Figures 1 and 2 illustrate theorem 3.

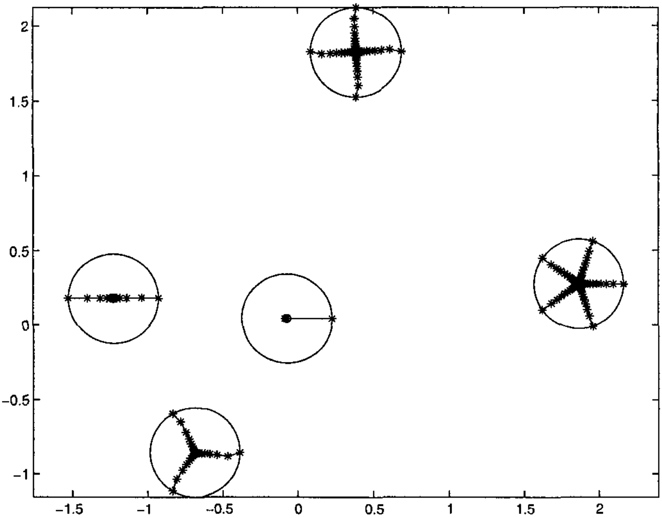


Figure 1: Local behaviour of Weierstrass iteration.

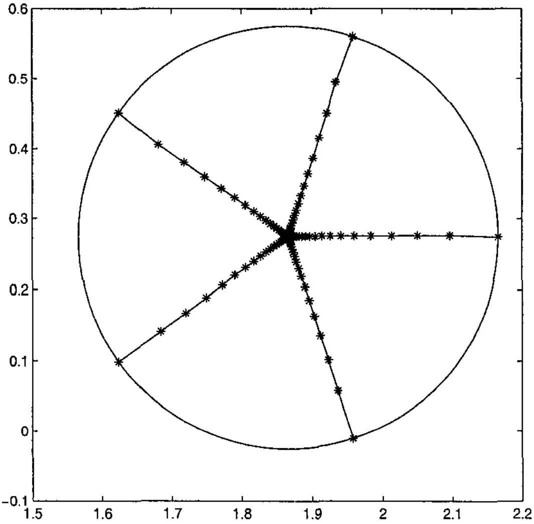


Figure 2: Enlargement around the 5-cluster $D(z_5, 0.05)$.

We next illustrate the numerical path-following using Weierstrass method via the algorithm described below. Recall that $f_t(z) = (1-t)f(z) + tg(z)$ is defined previouly and W_{f_t} is the Weierstrass operator associated to f_t. For a given integer $n_{it} \geq 2$ we will denote $W_{f_t}^{n_{it}}$ the repeated composition of the operator W_{f_t} n_{it} times.

Weierstrass Path-Following Algorithm

```
Inputs: $f$ a polynomial of degree $d$, $z^0 \in \mathbb{C}^d$, $\epsilon$ a positive real number, $n_{it} \geq 2$ an integer and $0 < M < 1$.
$\beta = 2\epsilon, \quad t_0 = 1, \quad t_1 = 1 - M$.
while $\beta > \epsilon$ or $t_0 \geq 0$
Compute the point $z^1 = W_{f_{t_1}}^{n_{it}}(z^0)$.
$\beta = \left| \frac{f_{t_1}(z^1)}{\prod_{j \neq i} z_i^1 - z_j^1} \right|$ if $\beta \leq \epsilon$ and $t_1 = 0$ Determine the clusters and stop.
if $\beta > \epsilon$ and $t_1 > 0$ replace $t_1$ by $\frac{t_1 + t_0}{2}$.
If $\beta \leq \epsilon$ and $t_1 > 0$ replace $t_0$ by $t_1$, $t_1$ by $\max(3t_1 - 2t_0, 0)$ and $z^0$ by $z^1$.
end Outputs: the set of all the clusters.
```

The proof of the convergence of this algorithm is given in [23]. It also has been used in [24] to compute only one cluster. The figure below was derived by taking $z_k^0 = e^{\frac{2i\pi}{k}}$, $1 \leq k \leq d$.

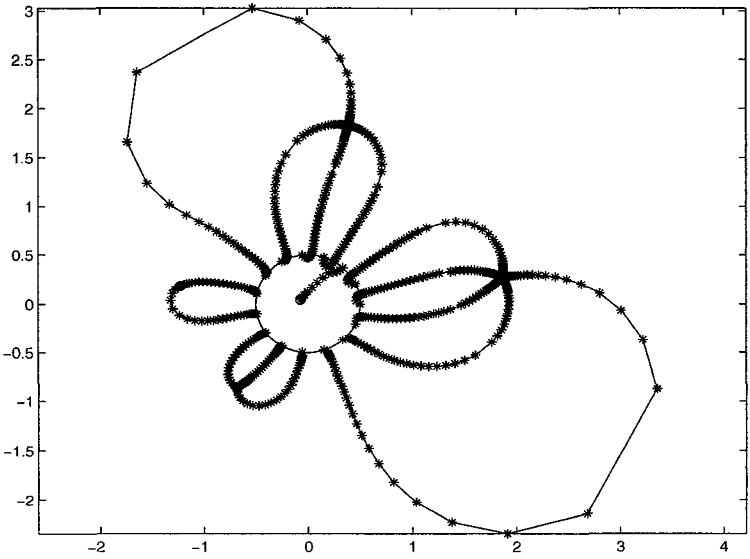


Figure 3: Simultaneous computation of all the zeroes-clusters.

3.1 Computing Eigenvalues

We consider any $d \times d$ matrix A whose eigenvalues are the roots of the polynomial f above. To compute the clusters of the eigenvalues we use the Weierstrass function with

$$h_t(z) = (1-t)det(zI-A) + t(z^d + \sum_{k=1}^{d}(-1)^{d-k+1}S_{d-k+1}(x^0)z^{k-1}).$$

Consequently the process only requires evaluation of a determinant. Obviously the coefficients of the characteristic polynomial are not calculated.

To illustrate this approach, take any 15×15 matrix whose eigenvalues are the roots of the polynomial f above. Setting $z_k^0 = e^{\frac{2i\pi}{k}}$, $1 \leq k \leq d$ we then obtain the following behavior for our sequence of approximations.

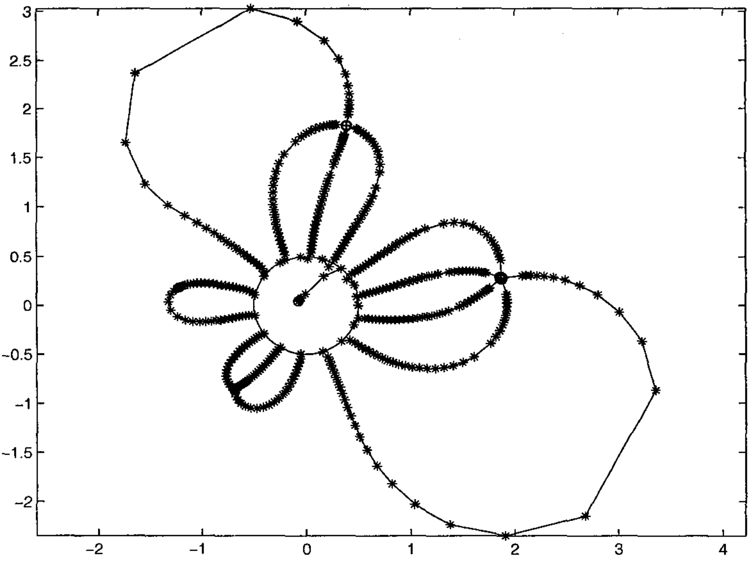


Figure 4: Simultaneous computation of all the eigenvalues-clusters.

4 Proof of Theorem 2

To prove this theorem we need some lemmas. We first state a similar lemma given in [4] Lemma 10 page 269:

Lemma 2

$$\sup_{k\geq 2}\left(\frac{1}{p}\binom{p}{k}\right)^{\frac{1}{k-1}}=\frac{p-1}{2}.$$

Proof: It is sufficient to prove that $\left(\frac{1}{p}\binom{p}{k}\right)^{\frac{1}{k-1}}$ is a decreasing function of k. For that verify

$$\frac{1}{p}\binom{p}{k+1}\leq\left(\frac{1}{p}\binom{p}{k}\right)^{\frac{k}{k-1}}.$$

This inequality is equivalent to

$$\frac{(p-1)\cdots(p-k)}{(k+1)!}\leq\left(\frac{(p-1)\cdots(p-k+1)}{k!}\right)^{1+\frac{1}{k-1}}.$$

We first have $\left(\frac{(p-1)\cdots(p-k+1)}{k!}\right)^{\frac{1}{k-1}}\geq\frac{p-k+1}{k}$. On the other hand, we have $\frac{p-k}{k+1}\leq\frac{p-k+1}{k}$. Hence

$$\begin{aligned}\left(\frac{(p-1)\cdots(p-k+1)}{k!}\right)^{1+\frac{1}{k-1}}&\geq\frac{(p-1)\cdots(p-k+1)}{k!}\frac{p-k+1}{k}\\&\geq\frac{(p-1)\cdots(p-k+1)}{k!}\frac{p-k}{k+1}.\end{aligned}$$

We are done.■

We now give a point estimate for $|\prod_{j=1}^{p}(1+z_j)-1|$ improving one given by Higham on page 75 of [11].

Lemma 3 *Let $p\geq 1$ an integer and ϵ a real number be such that $0\leq\epsilon<\frac{2}{p-1}$. Consider p complex numbers $z_1,\ldots,z_p$ satisfying $|z_j|\leq\epsilon$. Then,*

$$|\prod_{j=1}^{p}(1+z_j)-1|\leq\frac{2p\epsilon}{2-(p-1)\epsilon}.$$

Moreover $|\prod_{j=1}^{p}(1+z_j)|\geq 1-\frac{2p\epsilon}{2-(p-1)\epsilon}$.

Proof: We have $\prod_{j=1}^{p}(1+z_j)-1=\sum_{k=1}^{p}\sigma_k(z_1,\ldots,z_p)$ where the σ_k's are the Viete symmetric functions of $z_1,\ldots,z_p$. We know $|\sigma_k(z_1,\ldots,z_p)|\leq$

$\binom{p}{k}\epsilon^k$. From lemma 2 with $u = \frac{p-1}{2}\epsilon$, it follows that

$$\begin{aligned}\left|\prod_{j=1}^{p}(1+z_j)-1\right| &\leq p\epsilon\left(1+\sum_{k\geq 2}\frac{1}{p}\binom{p}{k}\epsilon^{k-1}\right)\\ &\leq p\epsilon\left(1+\sum_{k\geq 2}u^{k-1}\right)\\ &\leq \frac{p\epsilon}{1-\frac{p-1}{2}\epsilon}.\end{aligned}$$

So we are done.■

Lemma 4 *Let $x = (x_1, \ldots, x_d)$ be such that $x_i \in D(w_i, r)$, $1 \leq i \leq d$. Let $u = \gamma(f)r < \frac{1}{3d}$. Then the $W_i(f,x)'s$ are well defined and we have*

$$|W_i(f,x) - w_i| \leq \frac{2(d-1)\gamma(f)|x_i - w_i|}{1-(d+2)u}|x_i - w_i|, \quad 1 \leq i \leq d.$$

Moreover, $W_i(f,x) \neq W_j(f,x)$ for all $i \neq j$.

Proof: The distance between two distinct disks $D(w_i, r)$ and $D(w_j, r)$ is equal to $max(0, |w_i - w_j| - 2r)$. From lemma 1 and since $u = \gamma(f)r < \dfrac{1}{3d} \leq \dfrac{1}{6}$, $(d \geq 2)$, we get

$$|w_i - w_j| - 2r \geq \frac{1}{2\gamma(f)} - \frac{1}{6\gamma(f)} = \frac{1}{3\gamma(f)} > 0.$$

Consequently, the two disks $D(w_i, r)$ and $D(w_j, r)$ are disjoint and $x_i \neq x_j$ for all $i \neq j$. Hence the $W_i(f,x)$'s are well defined. A straightforward calculation shows that

$$W_i(f,x) - w_i = \left(1 - \prod_{j\neq i}\left(1 - \frac{x_j - w_j}{x_i - x_j}\right)\right)(x_i - w_i), \quad 1 \leq i \leq d.$$

From lemma 1 we have the point estimate with $u = \gamma(f)r$,

$$\left|\frac{x_j - w_j}{x_i - x_j}\right| \leq \frac{|x_j - w_j|}{|w_j - w_i| - |x_i - w_i| - |x_j - w_j|} \leq \frac{|x_j - w_j|}{\frac{1}{2\gamma(f)} - 2r} \leq \frac{2u}{1-4u}$$

where $i \neq j$ and $1 \leq i, j \leq d$. Applying the lemma 3 with $p = d-1$, $\epsilon = \frac{2u}{1-4u}$ and $z_j = \frac{x_j - w_j}{x_i - x_j}$, for $1 \leq j \leq d$, $j \neq i$, we obtain:

$$\begin{aligned}|W_i(f,x) - w_i| &\leq \frac{2(d-1)\frac{2u}{1-4u}}{2-(d-2)\frac{2u}{1-4u}}|x_i - w_i| \\ &\leq \frac{2\,(d-1)\,u}{1-(d+2)u}|x_i - w_i|, \quad 1 \leq i \leq d.\end{aligned}$$

We now show that the $W_i(f,x)$'s are distinct real numbers: The condition $u < \frac{1}{3d}$ implies $\frac{2(d-1)u}{1-(d+2)u} < 1$. Consequently $|W_i((f,x) - w_i| < |x_i - w_i| \leq r$. But the disks $D(w_i, r)$ are distinct. Hence $W_i(f,x) \neq W_j(f,x)$ if $i \neq j$. We are done.■

Proof of Theorem 2. We proceed by induction. The condition $u \leq \frac{1}{5d-2}$ and $d \geq 2$ implies $u \leq \frac{1}{3d}$. From lemma 4, if $i \neq j$ then $x_i \neq x_j$ and $x^1 = W(f,x)$ is well defined. Obviously the inequalities

$$|x_i^k - w_i| \leq \left(\frac{1}{2}\right)^{2^k - 1} r, \quad 1 \leq i \leq d,$$

hold for $k = 0$. Suppose now that for a given k we have that x^k is well defined and that the previous inequalities are satisfied. From lemma 4 we have

$$|x_i^{k+1} - w_i| \leq \frac{2\,(d-1)\,\gamma(f)|x_i^k - w_i|^2}{1-(d+2)\gamma(f)|x_i^k - w_i|}, \quad 1 \leq i \leq d.$$

On the other hand, the condition $u \leq \frac{1}{5d-2}$ implies $\frac{2(d-1)u}{1-(d+2)u} \leq \frac{1}{2}$. We then get by induction that

$$\begin{aligned}|x_i^{k+1} - w_i| &\leq \frac{2\,(d-1)\,\gamma(f)}{1-(d+2)u}\left(\frac{1}{2}\right)^{2^{k+1}-2}|x_i - w_i|^2 \\ &\leq \left(\frac{1}{2}\right)^{2^{k+1}-1}|x_i - w_i|.\end{aligned}$$

Moreover $x_i^{k+1} \neq x_j^{k+1}$ for $i \neq j$. Then the theorem follows.■

5 Proof of Theorem 3

Denote by $x = \left(x_{11}, \ldots, x_{1m_1}, \ldots, x_{p1}, \ldots, x_{pm_p}\right) \in \mathbb{C}^d$. Remember also $u = \gamma(f)R$. The radius r of each cluster $D(z_i, r)$ satisfies $r \leq uR$.

Lemma 5 *Let i and $j \in M_i$ be given. Let $\bar{x}_{ik}$, $1 \leq k \leq m_i$ the roots of the equation in z*

$$(z - z_i)^{m_i} = (x_{ij} - z_i)^{m_i},$$

such that $\bar{x}_{i1} = x_{ij}$. Let us suppose

(a) $|x_{ik} - \bar{x}_{ik}| \leq u|x_{ij} - \bar{x}_{ik}|, \quad 1 \leq k \leq m_i, \quad k \neq j$

(b) $|z_i - w_{ik}| \leq u|x_{ij} - z_i|, \quad 1 \leq k \leq m_i$, *and* $k \neq j$.

Then

1. $\left|1 - \prod_{k \in M_i, k \neq j} \frac{x_{ij} - w_{ik}}{x_{ij} - x_{ik}}\right| \leq \frac{m_i - 1 + \frac{2(m_i^2 - 1)u}{2 - (m_i - 2)u}}{m_i\left(1 - \frac{2(m_i - 1)u}{2 - (m_i - 2)u}\right)}$

2. $\left|\prod_{k \in M_i, k \neq j} \frac{x_{ij} - w_{ik}}{x_{ij} - x_{ik}}\right| \leq \frac{(1+u)^{m_i - 1}}{m_i\left(1 - \frac{2(m_i - 1)u}{2 - (m_i - 2)u}\right)}$.

Proof: Since the $\bar{x}_{ik}$ are the roots of the equation given in this lemma, we have $\prod_{k \in M_i, k \neq j}(x_{ij} - \bar{x}_{ik}) = m_i(x_{ij} - z_i)^{m_i - 1}$. We can then write

$$1 - \prod_{k \in M_i, k \neq j} \frac{x_{ij} - w_{ik}}{x_{ij} - x_{ik}} = 1 - \frac{(x_{ij} - z_i)^{m_i - 1}}{\prod_{k \in M_i, k \neq j}(x_{ij} - \bar{x}_{ik})} \frac{\prod_{k \in M_i, k \neq j}\left(1 + \frac{z_i - w_{ik}}{x_{ij} - z_i}\right)}{\prod_{k \in M_i, k \neq j}\left(1 + \frac{\bar{x}_{ik} - x_{ik}}{x_{ij} - \bar{x}_{ik}}\right)}$$

$$= 1 - \frac{1}{m_i \prod_{k \in M_i, k \neq j}\left(1 + \frac{\bar{x}_{ik} - x_{ik}}{x_{ij} - \bar{x}_{ik}}\right)} \prod_{k \in M_i, k \neq j}\left(1 + \frac{z_i - w_{ik}}{x_{ij} - z_i}\right)$$

$$= \frac{m_i - 1 + m_i\left(\prod_{k \in M_i, k \neq j}\left(1 + \frac{\bar{x}_{ik} - x_{ik}}{x_{ij} - \bar{x}_{ik}}\right) - 1\right) - \left(\prod_{k \in M_i, k \neq j}\left(1 + \frac{z_i - w_{ik}}{x_{ij} - z_i}\right) - 1\right)}{m_i \prod_{k \in M_i, k \neq j}\left(1 + \frac{\bar{x}_{ik} - x_{ik}}{x_{ij} - \bar{x}_{ik}}\right)}.$$

Using lemma 3 with $p = m_i - 1$, $\epsilon = u$, and the assumptions 1 and 2, a straightforward computation gives successively

$$\left|1 - \prod_{k \in M_i, k \neq j} \frac{x_{ij} - w_{ik}}{x_{ij} - x_{ik}}\right| \leq \frac{m_i - 1 + \frac{2m_i(m_i - 1)u}{2 - (m_i - 2)u} + \frac{2(m_i - 1)u}{2 - (m_i - 2)u}}{m_i\left(1 - \frac{2(m_i - 1)u}{2 - (m_i - 2)u}\right)}$$

$$\leq \frac{m_i - 1 + \frac{2(m_i^2 - 1)u}{2 - (m_i - 2)u}}{m_i\left(1 - \frac{2(m_i - 1)u}{2 - (m_i - 2)u}\right)}.$$

Just as before, we have $\left|\prod_{k \in M_i, k \neq j} \frac{x_{ij} - w_{ik}}{x_{ij} - x_{ik}}\right| = \frac{1}{m_i \prod_{k \in M_i, k \neq j}\left(1 + \frac{\bar{x}_{ik} - x_{ik}}{x_{ij} - \bar{x}_{ik}}\right)} \prod_{k \in M_i, k \neq j}\left(1 + \frac{z_i - w_{ik}}{x_{ij} - z_i}\right)$, and upon using lemma 3, part 2 is directly obtained. ■

Lemma 6 *Let i be given. Let $x_{kl} \in D(z_k, r)$, $l = 1, \dots, m_k$ and $k = 1, \dots p$. Suppose $9\alpha(f) \leq 1$, $r \geq 3\beta(f)$ and $r \leq uR$. Then*

1. *For $k \neq i$ we have $|z_i - w_{kl}| \geq \frac{1}{2\gamma(f)} - \frac{3}{2}\beta(f)$.*

2. *For $(k, l) \neq (i, j)$ we have $\left|\frac{x_{kl} - w_{kl}}{x_{ij} - x_{kl}}\right| \leq \frac{2u(1+u)}{1-a-4u-2u^2} = v$, with $a = 3\alpha(f)$.*

3. *Moreover*

$$B = \left| \prod_{k \neq i, l \in M_k} \left(1 + \frac{x_{kl} - w_{kl}}{x_{ij} - x_{kl}}\right) - 1 \right| \leq \frac{2(d - m_i)v}{2 - (d - m_i - 1)v}$$

$$B + 1 \leq (1 + v)^{d - m_i}.$$

Proof:
(1): Let us consider w a zero of f which don't lies in $D(z_i, r)$. We have successively:
$|f(w)| = 0 = \left|\sum_{k=0}^{d} \frac{f^{(k)}(z_i)}{k!}(w - z_i)^k\right|$
$\geq \frac{|f^{(m_i)}(z_i)|}{m_i!}|w - z_i|^{m_i} - \sum_{k=0}^{m_i-1} \frac{|f^{(k)}(z_i)|}{k!}|w - z_i|^k - \sum_{k \geq m_i+1} \frac{|f^{(k)}(z_i)|}{k!}|w - z_i|^k$
$\geq \frac{|f^{(m_i)}(z_i)|}{m_i!}|w - z_i|^{m_i}\left(1 - \sum_{k=0}^{m_i-1} \frac{m_i!|f^{(k)}(z_i)|}{k!|f^{(m_i)}(z_i)|}|w - z_i|^{k-m_i} - \sum_{k \geq m_i+1} \frac{m_i!|f^{(k)}(z_i)|}{k!|f^{(m_i)}(z_i)}|w - z_i|^{k-m_i}\right)$
$\geq \frac{|f^{(m_i)}(z_i)|}{m_i!}|w - z_i|^{m_i}\left(1 - \sum_{k=0}^{m_i-1}\left(\frac{\beta(f)}{|w-z_i|}\right)^{m_i-k} - \sum_{k \geq m_i+1}(\gamma_{m_i}(f)|w - z_i|)^{m_i-k}\right).$
Via the inequality $\gamma_{m_i}(f) \leq \gamma(f)$ we finally get:

$$0 \geq 1 - \frac{\frac{\beta(f)}{|w-z_i|}}{1 - \frac{\beta(f)}{|w-z_i|}} - \frac{\gamma(f)|w - z_i|}{1 - \gamma(f)|w - z_i|}$$

$$\geq \frac{-2\gamma(f)|w - z_i|^2 + (1 + 3\gamma(f)\beta(f))|w - z_i| - 2\beta(f)}{(1 - \gamma(f)|w - z_i|)(|w - z_i| - \beta(f))}.$$

Under the condition $9\alpha(f) \leq 1$, the polynomial

$$-2\gamma(f)t^2 + (1 + 3\gamma(f)\beta(f))t - 2\beta(f)$$

has two real roots: $r_1 = \frac{1+3\alpha(f)-\sqrt{1-10\alpha(f)+9\alpha(f)^2}}{4\gamma(f)}$ and $r_2 = \frac{1+3\alpha(f)+\sqrt{1-10\alpha(f)+9\alpha(f)^2}}{4\gamma(f)}$. It is then easy to see that the assumption $r > 3\beta(f)$ implies $r > r_1$. Since $|w - z_i| > r$, the inequality

$$-2\gamma(f)|w - z_i|^2 + (1 + 3\gamma(f)\beta(f))|w - z_i| - 2\beta(f) \leq 0$$

holds when $|w - z_i| \geq r_2$. A lower bound for r_2 is $\frac{1}{2\gamma(f)} - \frac{3}{2}\beta(f)$. In fact, the study of the function $t \in [0, \frac{1}{9}] \to \frac{1+3t+\sqrt{1-10t+9t^2}}{4}$ shows that $\frac{1+3t+\sqrt{1-10t+9t^2}}{4} \geq \frac{1}{2} - \frac{3}{2}t$. Replace t by α and divide the previous inequality by $\gamma(f)$, we find the lower bound announced for r_2. By definition of $\beta(f)$ and $\gamma(f)$ we finally have

$$|w - z_i| \geq \frac{1}{2\gamma(f)} - \frac{3}{2}\beta(f).$$

(2): From part (1) we have

$$\begin{aligned} |x_{ij} - x_{kl}| &= |x_{ij} - z_i + z_i - w_{kl} + w_{kl} - z_k + z_k - x_{kl}| \\ &\geq |z_i - w_{kl}| - |x_{ij} - z_i| - |w_{kl} - z_k| - |z_k - x_{kl}| \\ &\geq \frac{1}{2\gamma(f)} - \frac{3}{2}\beta(f) - 2R - r. \end{aligned}$$

Since $r \leq uR$, it follows

$$\begin{aligned} \left|\frac{x_{kl} - w_{kl}}{x_{ij} - x_{kl}}\right| &\leq \frac{R + r}{\frac{1}{2\gamma(f)} - \frac{3}{2}\beta(f) - 2R - r} \\ &\leq \frac{2u(1+u)}{1 - a - 4u - 2u^2} = v, \end{aligned}$$

with $a = 3\alpha(f)$.

(3): There are $d - m_i$ factors in the product of the quantity B. From lemma 3 and part (2) we get

$$B \leq (1+v)^{d-m_i} - 1 \leq \frac{2(d - m_i)v}{2 - (d - m_i - 1)v}.$$

The quantity $B + 1$ is obviously bounded by: $B + 1 \leq (1+v)^{d-m_i}$ We are done.■

Lemma 7 *Let R, r and be two positive real numbers which such that $\frac{r}{R} < \frac{2}{m}$. Introduce the point $y = (y_1, \ldots y_m)$ such that $y_j^m = R^m$, $1 \leq j \leq m$. For $c < 1$, consider the sequence of disks*

$$D_j^0 = D(y_j, r), \quad D_j^k = c^k D\left(y_j, r\right), 1 \leq j \leq m, k \geq 1.$$

Then for k be fixed, we have $D_j^k \cap D_l^k = \emptyset$ and the distance between two distinct disks is greater than

$$2R\sqrt{1 - \left(\frac{r}{R}\right)^2}\left(\frac{2}{m} - \frac{r}{R}\right)c^k.$$

Proof: Each disk D_j^k is contained in a wedge of angle 2θ such that $\sin\theta = \frac{r}{R}$, see figure. It is obvious that $D_j^k \cap D_l^k = \emptyset$ for $j \neq l$ if $\frac{2\pi}{m} - 2\theta > 0$. But $\frac{r}{R} = \sin\theta \geq \frac{2}{\pi}\theta$. Hence, using the assumption $\frac{r}{R} < \frac{2}{m}$, the inequality

$$\frac{2\pi}{m} - 2\theta > \frac{2\pi}{m} - \frac{\pi r}{R} > 0$$

holds.

Elementary geometric considerations in the triangle OAB (see figure) show that the distance between two distinct disks D_j^0 and D_l^0 is greater than

$$2R\sqrt{1-\left(\frac{r}{R}\right)^2}\sin\left(\frac{\pi}{m}-\theta\right).$$

Since $\sin\left(\frac{\pi}{m}-\theta\right) \geq \frac{2}{m} - \frac{r}{R}$, the result follows.■

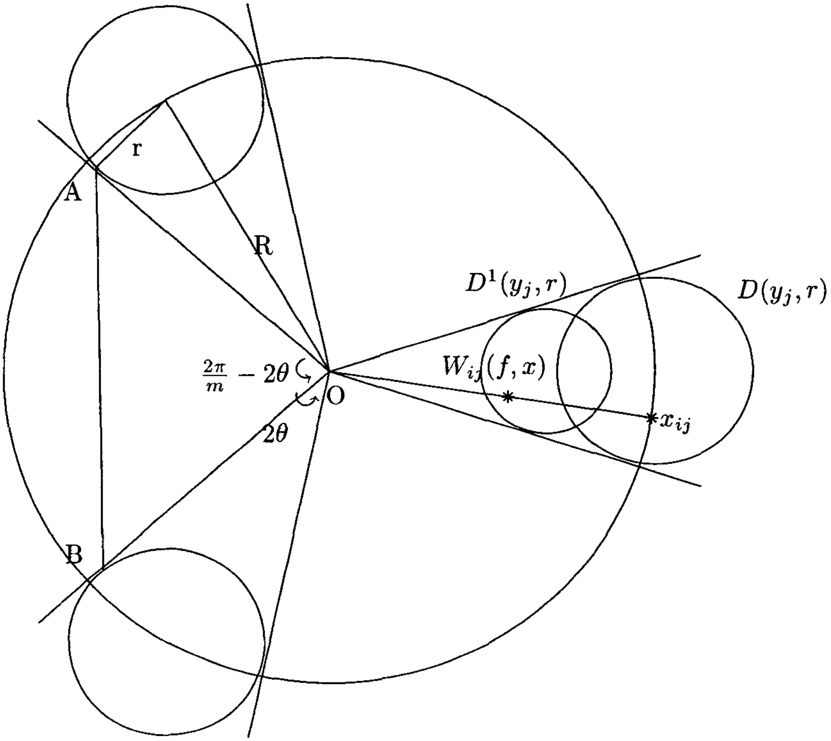


Figure 5: Illustration of the proof of lemma 7.

Lemma 8 *Let us consider positive real numbers r, R, and u such that*

$$\frac{r}{R\sqrt{1-\left(\frac{r}{R}\right)^2\left(\frac{2}{m}-\frac{r}{R}\right)}} < u.$$

Suppose also that $9\alpha(f) \le 1$ and $r > 3\beta(f)$. Introduce p complex numbers $y_{11}, \dots, y_{p1}$ be such that $|z_i - y_{i1}| \le R$.

For all i, $1 \le i \le p$, consider respectively the roots $y_{i1}, \dots y_{im_i}$ of the equations

$$(z-z_i)^{m_i} = (y_{i1}-z_i)^{m_i}.$$

Let x be such that $x_{ij} \in D(y_{ij}, r)$, $j \in M_i$, $1 \le i \le p$.

Suppose for all i, $\varphi(m_i, u) < 1$, and $|z_i - w_{ij} \le u|x_{ij} - z_i|$. We then have:

1. *For all i, $1 \le i \le p$ such that $m_i > 1$ we have:*

   $$|W_{ij}(f,x) - z_i| \le \varphi(m_i,u)|x_{ij}-z_i|, \quad 1 \le j \le m_i.$$

   The $W_{ij}(f,x)$ respectively belong to the disks $\phi(m_i,u)D(y_{ij},r)$, $j \in M_i$, $1 \le i \le p$. Moreover, these disks are pair-wise disjoint.

2. *For all i, $1 \le i \le p$ such that $m_i = 1$ we have:*

   $$|W_{i1}(f,x) - w_i| \le \frac{2(d-1)u}{1-(d+2)u}|x_{i1} - w_i|.$$

Proof: Let us first consider an index i such that $m_i > 1$.

A straightforward computation shows that:

$$W_{ij}(f,x) - z_i = \left(1 - \prod_{(k,l)\neq(i,j)} \frac{x_{ij}-w_{kl}}{x_{ij}-x_{kl}}\right)(x_{ij}-z_i) - \prod_{(k,l)\neq(i,j)} \frac{x_{ij}-w_{kl}}{x_{ij}-x_{kl}}(z_i - w_{ij}).$$

We first bound $\left|1 - \prod_{(k,l)\neq(i,j)} \frac{x_{ij}-w_{kl}}{x_{ij}-x_{kl}}\right|$. For that we write

$$\begin{aligned} 1 - \prod_{(k,l)\neq(i,j)} \frac{x_{ij}-w_{kl}}{x_{ij}-x_{kl}} &= 1 - \prod_{l\neq j} \frac{x_{ij}-w_{il}}{x_{ij}-x_{il}} \prod_{k\neq i} \frac{x_{ij}-w_{kl}}{x_{ij}-x_{kl}} \\ &= 1 - \prod_{l\neq j} \frac{x_{ij}-w_{il}}{x_{ij}-x_{il}} \prod_{k\neq i}\left(1 + \frac{x_{kl}-w_{kl}}{x_{ij}-x_{kl}}\right) \\ &= 1 - A - AB \end{aligned}$$

with $A = \prod_{l \neq j} \frac{x_{ij} - w_{il}}{x_{ij} - x_{il}}$ and $B = \prod_{k \neq i} \left(1 + \frac{x_{kl} - w_{kl}}{x_{ij} - x_{kl}}\right) - 1$. Hence, using the assumtion $|z_i - w_{ij}| \leq u|x_{ij} - z_i|$, we get

$$|W_{ij}(f,x) - z_i| \leq \left(|1 - A| + |AB| + |A(B+1)|u\right)|x_{ij} - z_i|.$$

Verify the assumption 2 of lemma 5 to bound $|1 - A|$. With the point x_{ij}, construct the roots $\bar{x}_{ik}$ of the equation $(z - z_i)^{m_i} = (x_{ij} - z_i)^{m_i}$ as in lemma 5. By definition of y_{ij} and since $x_{ij} \in D(y_{ij}, r)$, we have $\bar{x}_{ik} \in D(y_{ij}, r)$. Hence $|x_{ik} - \bar{x}_{ik}| \leq 2r$. Using lemma 7, the distance between two distinct disks $D(y_{ij}, r)$ and $D(y_{ik}, r)$ is greater than

$$2R\sqrt{1 - \left(\frac{r}{R}\right)^2}\left(\frac{2}{m} - \frac{r}{R}\right).$$

Hence, for all $k \neq j$, $1 \leq k \leq m_i$, we have

$$\frac{|x_{ik} - \bar{x}_{ik}|}{|x_{ij} - \bar{x}_{ik}|} \leq \frac{r}{R\sqrt{1 - \left(\frac{r}{R}\right)^2}\left(\frac{2}{m} - \frac{r}{R}\right)} < u.$$

The assumptions of lemma 5 hold. Consequently $|1 - A|$ is bounded by

$$|1 - A| \leq \frac{m_i - 1 + \frac{2(m_i^2 - 1)u}{2 - (m_i - 2)u}}{m_i\left(1 - \frac{2(m_i - 1)u}{2 - (m_i - 2)u}\right)}.$$

We bound the quantities B and $B + 1$ using the lemma 6:

$$B \leq \frac{2(d - m_i)v}{2 - (d - m_i - 1)v}$$

$$B + 1 \leq (1 + v)^{d - m_i},$$

where v is defined in the introduction. From part 2 of lemma 5 we bound A by $\frac{(1+u)^{m_i - 1}}{m_i\left(1 - \frac{2(m_i - 1)u}{2 - (m_i - 2)u}\right)}$.

With all these point estimates, we obtain the following upper bound on $|W_{ij}(f,x) - z_i|$: $\left(\frac{m_i - 1}{m_i} + \frac{2(m_i^2 - 1)u}{m_i(2 - (m_i - 2)u)} + \frac{2(d - m_i)(1+u)^{m_i - 1}v}{m_i(2 - (d - m_i - 1)v)} + \frac{(1+u)^{m_i - 1}(1+v)^{d - m_i}u}{m_i}\right)\frac{1}{1 - \frac{2(m_i - 1)u}{2 - (m_i - 2)u}}|x_{ij} - z_i|$, which in turn is bounded above by $\varphi(m_i, u)|x_{ij} - z_i|$. Hence the point $W_{ij}(f,x)$ lies in the disk $\varphi(m_i, u)D\left(y_{ij}, |x_{ij} - z_i|\right)$. Lemma 7 shows that the W_{ij}'s are distinct points. We have thus proved the first part.

Let us now suppose there exists an index i be such that $m_i = 1$. In this case, the point estimate is a direct consequence of lemma 4. So we are done. ■

Proof of Theorem 3: We proceed by induction. The case $k = 0$ is obvious. Suppose the sequence $(x^k)_{k\geq 0}$ is well defined and that the x_{ij}^k's are distinct points satisfying the point estimates of our theorem. For the indices i such that $m_i > 1$, the assumptions of lemma 8 are satisfied. Then for $j \neq l$ we have $x_{ij}^{k+1} \neq x_{il}^{k+1}$. Remember also that $x_{ij}^k \in D_{ij}^k$ and from inductive assumption

$$\begin{aligned}|x_{ij}^{k+1} - z_i| &\leq \varphi(m_i, u)|x_{ij}^k - z_i| \\ &\leq \varphi(m_i, u)^{k+1}(R + r)\end{aligned}$$

In the case $m_i = 1$, suppose $|x_{i1}^k - w_i| \leq \varphi(1, u)^{2^k - 1}$. By lemma 8, we have

$$\begin{aligned}|x_{i1}^{k+1} - w_i| &\leq \varphi(1, u)|x_{i1}^k - w_i| \\ &\leq \frac{\varphi(1, u)}{R}|x_{i1}^k - w_i|^2 \\ &\leq \frac{\varphi(1, u)}{R}\left(\varphi(1, v)\right)^{2^{k+1}-2} R^2 \\ &\leq \left(\varphi(1, u)\right)^{2^{k+1}-1} R.\end{aligned}$$

We are done.■

6 Proof of Theorem 4

Remember that we consider the curve w_i^t for an index i be given, $1 \leq i \leq d$ defined by $f_t(w_i^t) = 0$. We first prove the following

Lemma 9 *Let $t_k \geq t_i^+$. For all t and s lying in $[t_{k+1}, t_k]$, we have $|w_i^t - w_i^s| \leq u$.*

Proof: We obviously have $0 = f_t(w_i^t) = f_s(w_i^s) = f_t(w_i^s) + (s - t)(g(w_i^s) - f(w_i^s))$. From Taylor's formula, we get

$$f_t^{'}(w_i^s)(w_i^t - w_i^s) + \sum_{k\geq 2} \frac{f_t^{(k)}(w_i^s)}{k!}(w_i^t - w_i^s)^k = (t - s)(g(w_i^s) - f(w_i^s)).$$

Hence, using $g = \max_i \max_{t_i^+ \leq t \leq 1} \gamma(f_t)$ and $|t - s| \leq t_k - t_{k+1} = M^k(1 - M) \leq 1 - M$, we obtain, for all $t, s \in [t_{k+1}, t_k]$

$$|w_i^t - w_i^s|\left(1 - \sum_{k\geq 2}(g|w_i^t - w_i^s|)^{k-1}\right) \leq |s - t| \max_i \max_{t_i^+ \leq s \leq 1} \frac{|g(w_i^s) - f(w_i^s)|}{|f_t^{'}(w_i^s)|}$$

$$\frac{g|w_i^t - w_i^s|(1 - 2g|w_i^t - w_i^s|)}{1 - g|w_i^t - w_i^s|} \leq (1 - M)a = \frac{u(1 - 2u)}{1 - u}.$$

Since $u \to \frac{u(1-2u)}{1-u}$ is a decreasing function, we conclude $g|w_i^t - w_i^s| \leq u$.■

Lemma 10 *Let* $h(u) = \frac{(1-3u)u}{1-u}$. *Then* $t_i^+ := \sup\{t \in]0,1] \mid |w_i^t - w_i| = R\}$ *satisfies*

$$\frac{h(u)R^{m_i-1}}{a_i + h(u)R^{m_i-1}} \leq t_i^+$$

Proof: Let $D(z_i, r)$ be a m_i cluster and $w_i \in D(x_i, r)$ be a zero of f. Denote $t = t_i^+$. We have $0 = f_t(w_i^t) = (1-t)f(w_i^t) + tg(w_i^t)$. Consequently $(1-t)\left(\sum_{k=1}^{m_i-1} \frac{f^{(k)}(w_i)}{k!}(w_i^t - w_i)^k + \frac{f^{(m_i)}(w_i)}{m_i!}(w_i^t - w_i)^{m_i} + \sum_{k \geq m_i+1} \frac{f^{(k)}(w_i)}{k!}(w_i^t - w_i)^k\right) = tg(w_i^t)$. As in the proof of lemma 6, we obtain from $|w_i^t - w_i| = R$ and $\beta(f) \leq Ru$ that... $(1-t)\left(1 - \sum_{k=1}^{m_i-1}\left(\frac{\beta(f,w_i)}{R}\right)^{m_i-k} - \sum_{k \geq m_i+1}(\gamma(f,w_i)R)^{k-m_i}\right) R^{m_i} \leq \max_{0 \leq t \leq 1}\left|\frac{m_i! g(w_i^t)}{f^{(m_i)}(w_i)}\right| t$

So $(1-t)\left(1 - \frac{\frac{\beta(f)}{R}}{1-\frac{\beta(f)}{R}} - \frac{\gamma(f)R}{1-\gamma(f)R}\right)\gamma(f)R^{m_i} \leq a_i t$ and thus $(1-t)\frac{(1-3u)u}{1-u}R^{m_i-1} \leq a_i t$. The inequality of our lemma follows. ■

Proof of Theorem 4: Let an index i and t_i^+ be such that $|w_i^{t^+} - w_i| = R$. For all $t \in [t_i^+, 1]$, the polynomial f_t only has simple roots. The quantity $g = \max_i \max_{t_i^+ \leq t \leq 1} \gamma(f_t)$ is bounded and we have, from lemma 1

$$|w_i^t - w_j^t| \geq \frac{1}{2g}.$$

(1): We first prove by induction that the z_i^k's are approximate zeroes of f_k associated to w_i^k for all index k such that $t_k \in [t_i^+, 1]$. It is obvious for $k = 0$. Suppose the z_i^k are distinct points and we have $g|z_i^k - w_i^k| \leq u$.Then prove $g|z_i^{k+1} - w_i^{k+1}| \leq u$. Applying lemma 9, we have

$$g|z_i^k - w_i^{k+1}| \leq g|z_i^k - w_i^k| + g|w_i^k - w_i^{k+1}| \leq u + u = 2u.$$

Since $u \leq \frac{1}{5d-2}$, we have from theorem 2 for all k be such that $t_k \in [t_i^+, 1]$:

$$g|z_i^{k+1} - w_i^{k+1}| \leq \frac{1}{2}g|z_i^k - w_i^{k+1}| \leq u.$$

Prove now that the z_i^k are distinct points. For $i \neq j$ we have from the assumption $4ug < 1$:

$$|z_i^{k+1} - z_j^{k+1}| \geq |w_i^{k+1} - w_j^{k+1}| - |z_i^{k+1} - w_i^{k+1}| - |w_j^{k+1} - z_j^{k+1}| \geq \frac{1}{2g} - 2u > 0.$$

Hence the z_i^{k+1}'s are distinct points and $W(f_t, z^{k+1})$ is well defined.
(2): Part 2 follows from lemma 10.

(3): From lemma 9 and the part 1 of this theorem,the point $z_i^{k_i}$ is an approximate zero of $w_i^{t_i^+}$. Hence

$$|z_i^{k_i} - w_i| \leq |w_i^{t_i^+} - w_i| + |z_i^{k_i} - w_i^{t_i^+}| \leq R + \frac{u}{g}.$$

We are done.■

References

1. ABERTH, O., *Iteration methods for finding all zeroes of a polynomial simultaneously,* Mathematics of Computation, Vol.27, 1973, pp. 39–344.
2. ALEFED, G., HERZBERGER, J., *On the convergence speed of some algorithms for simultaneous approximation of polynomial roots,* SIAM Journal Numerical Analysis, 11, pp. 237–243, 1974.
3. BELLIDO,A.,*Construction de fonctions d'itérations pour le calcul simultané des solutions d'équations et de systèmes d'équations algébriques,* Thèse Université Paul Sabatier,1992.
4. BLUM, L., CUCKER, F., SHUB, M., SMALE, S., *Complexity and Real Computation,* Springer-Verlag, 1998.
5. CARSTENSEN, C., PETROVIĆ, M.C.,TRAJKOVIĆ, M., *Weierstrass formula and zero finding methods,* Numerische Mathematik, 69, 1995, pp. 353–372.
6. DEREN, W., FENGGUANG, Z., *The Globalization of Durand-Kerner Algorithm,* preprint 1995.
7. DURAND, E. *Solution numérique des équations algébriques,* Tome 1, Masson, Paris, 1968.
8. GREEN, M.W., KORSACK, A.J., PEASE, M.C., *Simultaneous iteration towards all roots of a complex polynomial,* SIAM Review, Vol.18, pp. 501–502, 1976.
9. FARMER, M.R., LOIZOU, G., *A clazs of iterations functions for improving, simultaneously, approximations to the zeroes of a polynomial,* BIT 15, pp. 250–258, 1975.
10. FRAIGNIAUD, P, *The Durand-Kerner's Polynomial Roots-Finding in Case of Multiple Roots,* BIT 31, pp. 112–123, 1991.
11. HIGHAM N.J., *Accuracy and Stability of Numerical Algorithm,* SIAM, 1996.
12. KERNER, I.O. *Ein Gesamtschrittverfarhen zur Berechnung der Nullstellen von Polynomem,* Numerische Mathematik, Vol. 8, pp. 290–294, 1966.

13. KERNER, I.O. *Algorithm 283,* Communications ACM, Vol.9, P.273, 1966.
14. KJELLBERG, G.,*Two Observations on Durand-Kerner's Root-Finding Methods,* BIT 24, pp. 556–559, 1984.
15. PASQUINI, L., TRIGIANTE, D.,*A Globally Convergent Method for Simultaneously Finding Polynomials Roots*, Mathematics of Computation, Vol.44,169, pp. 135–149, 1985.
16. PETKOVIĆ, M.S., STEFANOVIĆ, L.V., *On some iterations functions for the simultaneous computation of multiple complex polynomial zeroes,* BIT, 27, pp. 111–122, 1987.
17. PETKOVIĆ, M.S., *Iterative Methods for Simultaneous Inclusion of Polynomial Zeros,* Springer, Berlin, 1989.
18. SENDOV, BL., ANDREEV, A., KJURKCHIEV, N.,*Numerical solution of polynomial equations,* in Handbook of Numerical Analysis, Vol 3, Ciarlet-Lions Editors, 1994.
19. TILLI, P. *Convergence conditions of some methods for the simultaneous computations of polynomial zeroes,* University of Pisa, preprint 1995.
20. TILLI, P., *Polynomial root finding by means of continuation,* University of Pisa, preprint 1995.
21. WEIERSTRASS, K., *Neuer beweis des satzes, dass jede ganze rationale function einer veränderlichen dargestellt werden kann als ein product aus linearen functionen derselben veränderlichen,* Mathematische Werke, tome 3, Mayer u. Müller, Berlin, pp. 251–269, 1903.
22. WERNER, W., *On the simultaneous determination of polynomials roots,* in Iterative Solutions of Nonlinear Systems of Equations, Lecture Notes in Math., Vol. 43, 167, pp. 205–217, 1984.
23. YAKOUBSOHN, J.C., *Contraction, Robustness and Numerical Path-Following Using Secant Maps,* Journal of Complexity, 16, 1, 286-310, (2000).
24. YAKOUBSOHN, J.C., *Finding a Cluster of Zeros of Univariate Polynomials,* Journal of Complexity, 16, 3, 603-638, (2000).

CROSS-CONSTRAINED VARIATIONAL PROBLEM AND NONLINEAR SCHRÖDINGER EQUATION

JIAN ZHANG

Department of Mathematics, Sichuan Normal University, Chengdu, 610066, China[*]
E-mail: jianzhan@mail.sc.cninfo.net

By constructing a type of cross constrained variational problem and establishing so-called cross-invariant manifolds of the evolution flow, we derive a sharp criterion for global existence and blowup of the solutions to the nonlinear Schrödinger equation. The instability of the standing waves in the equation is also shown.

1 Introduction

We are concerned with the following nonlinear Schrödinger equation

$$i\varphi_t + \Delta\varphi + |\varphi|^{p-1}\varphi = 0, \qquad t \geq 0, x \in \mathbb{R}^N, \tag{1.1}$$

where $1 < p < \frac{N+2}{(N-2)^+}$ (we use the convention: $\frac{N+2}{(N-2)^+} = \infty$ when $N = 1, 2$ and $(n-2)^+ = N-2$ when $N \geq 3$).

Ginibre and Vero [5] established the local existence of (1.1) and the global existence of (1.1) for $1 < p < 1 + \frac{4}{N}$, as well as $1 + \frac{4}{N} \leq p < \frac{N+2}{(N-2)^+}$ with small initial data. Glassey [6], Ogawa and Tsutsumi [9,10] studied the blowup properties of (1.1) for some initial data. Berestycki and Cazenave [1] as well as Weinstein [16] showed some interesting sharp criteria for blowup and global existence of (1.1), as well as the strong instability of the standing waves in (1.1) by variational arguments. The related work also sees [15] and [19] etc.

In the present paper, we construct a type of cross-constrained variational problem and establish it's property, then apply it to the nonlinear Schrödinger equation (1.1). By studying the corresponding cross-invariant manifolds under the flow of equation (1.1), we establish the sharp criterion for global existence and blowup of the solutions. By this criterion and the property of the cross-constrained variational problem, we also show the strong instability of the standing waves in equation (1.1). Berestycki and Cazenave [1] as well as Weinstein [16] have studied the similar problems. But in [1] and [16], the related variational problems have to be solved, the Schwarz symmetrization and complicated variational computation have to be conducted. By our new variational argument, we can refrain from solving the attaching variational

[*] Former address: Department of Mathematical Sciences, The University of Tokyo, 3 - 8 -1 Komaba, Meguro-Ku, Tokyo 153, Japan

problems, and directly establish the sharp criterion for global existence and blowup of equation (1.1), which is different from [1] and [16]. Furthermore By using our sharp criterion for blowup, the strong instability of the standing waves in equation (1.1) is also shown. Moreover we see that the argument proposed here may be developed to treat nonlinear Schrödinger equation with potentials as well as systems. For these equations, in solving the attaching variational problems, we often meet some essential difficulties (for example see [12] and [11]).

In the following, we first state some preliminaries for nonlinear Schrödinger equation in section 2. Next we establish the cross-constrained variational problem and the invariant manifolds in section 3. Then we derive the sharp criterion for global existence and blowup in section 4. Lastly we show the strong instability of the standing waves in section 5.

2 Preliminaries

We impose the initial data of (1.1) as follows.

$$\varphi(0,x) = \varphi_0(x), \qquad x \in \mathbb{R}^N. \tag{2.1}$$

From Ginibre and Velo [5], we have the following local well-posedness for the Cauchy problem (1.1)–(2.1).

Theorem 2.1 *Let $\varphi_0 \in H^1(\mathbb{R}^N)$. Then there exists a unique solution $\varphi(t,x)$ of the Cauchy problem (1.1) - (2.1) in $C([0,T);H^1(\mathbb{R}^N))$ for some $T \in (0,\infty)$ (maximal existence time), either $T=\infty$, or else $T<\infty$ and*

$$\lim_{t\to T-} ||\varphi(t,\cdot)||_{H^1(\mathbb{R}^N)} = \infty.$$

Furthermore for $\forall t \in [0,T), \varphi(t,x)$ satisfies

$$\int_{\mathbb{R}^N} |\varphi(t,x)|^2 dx = \int_{\mathbb{R}^N} |\varphi_0(x)|^2 dx, \tag{2.2}$$

$$E(\varphi) := \int_{\mathbb{R}^N} \left[\frac{1}{2}|\nabla\varphi(t,x)|^2 - \frac{1}{p+1}|\varphi(t,x)|^{p+1}\right] dx = E(\varphi_0). \tag{2.3}$$

From Glassey [6] and Cazenave [3], we have the following result.

Theorem 2.2 *Let $\varphi_0 \in H^1(\mathbb{R}^N), |\cdot|\varphi_0(\cdot) \in L^2(\mathbb{R}^N)$ and $\varphi(t,x)$ be a solution of the Cauchy problem (1.1)–(2.1). Put $J(t) := \int_{\mathbb{R}^N} \frac{1}{2}|x|^2|\varphi(t,x)|^2 dx$. Then*

$$J''(t) = \int_{\mathbb{R}^N} 4\left(|\nabla\varphi|^2 - \frac{N}{2}\frac{p-1}{p+1}|\varphi|^{p+1}\right) dx. \tag{2.4}$$

3 The cross-constrained variational problem and invariant manifolds

For $u \in H^1(\mathbb{R}^N)$ and $1 < p < \frac{N+2}{(N-2)^+}$, we define the following functionals.

$$I(u) := \int_{\mathbb{R}^N} \left(\frac{1}{2}|u|^2 + \frac{1}{2}|\nabla u|^2 - \frac{1}{p+1}|u|^{p+1} \right) dx. \tag{3.1}$$

$$S(u) := \int_{\mathbb{R}^N} (|u|^2 + |\nabla u|^2 - |u|^{p+1}) dx. \tag{3.2}$$

$$Q(u) := \int_{\mathbb{R}^N} \left(|\nabla u|^2 - \frac{N}{2}\frac{p-1}{p+1}|u|^{p+1} \right) dx. \tag{3.3}$$

From the Sobolev's embedding theorem, the above functionals are well defined. In addition, we define a manifold as follows.

$$M := \{u \in H^1(\mathbb{R}^N), \quad Q(u) = 0, \ \ S(u) < 0\}. \tag{3.4}$$

In section 5, we will give a remark to explain that M is not empty.

Now we consider the following two constrained variational problems.

$$d := \inf_{\{u \in H^1(\mathbb{R}^N) \setminus \{0\}, S(u)=0\}} I(u). \tag{3.5}$$

$$d_M := \inf_M I(u). \tag{3.6}$$

First from (3.5) we have the result.

Lemma 3.1 $\quad d > 0$.

Proof: From $S(u) = 0$ and the Sobolev's embedding inequality, we have

$$\int_{\mathbb{R}^N} |u|^{p+1} dx \le c \left(\int_{\mathbb{R}^N} |u|^{p+1} dx \right)^{\frac{p+1}{2}}. \tag{3.7}$$

Here and hereafter c denotes various positive constants. From $p > 1$ and $u \neq 0$, (3.7) implies that

$$\int_{\mathbb{R}^N} |u|^{p+1} dx \ge c > 0. \tag{3.8}$$

By (3.5), we have

$$I(u) = \left(\frac{1}{2} - \frac{1}{p+1} \right) \int_{\mathbb{R}^N} |u|^{p+1} dx. \tag{3.9}$$

Thus we get $d \ge c > 0$.

Next from both (3.5) and (3.6) we have the result.

Lemma 3.2 $d_M \geq d$ *provided* $1+\frac{4}{N} \leq p < \frac{N+2}{(N-2)^+}$.
Proof. Let $u \in M$ and

$$u_\lambda = \lambda^{\frac{2}{p-1}} u(\lambda x) \qquad for \ \ \lambda > 0. \tag{3.10}$$

Put $\alpha = \frac{N+2-p(N-2)}{p-1}$, $\beta = \frac{N+4-pN}{p-1}$. From $1+\frac{4}{N} \leq p < \frac{N+2}{(N-2)^+}$, we have

$$\alpha > 0; \qquad \beta \leq 0; \qquad \alpha = \beta + 2. \tag{3.11}$$

Moreover

$$S(u_\lambda) = \lambda^\alpha \int_{\mathbb{R}^N} (|\nabla u|^2 - |u|^{p+1})dx + \lambda^\beta \int_{\mathbb{R}^N} |u|^2 dx. \tag{3.12}$$

$$Q(u_\lambda) = \lambda^\alpha \int_{\mathbb{R}^N} (|\nabla u|^2 - \frac{N}{2}\frac{p-1}{p+1}|u|^{p+1})dx. \tag{3.13}$$

Thus $S(u) < 0$ implies that there exists a unique $0 < \lambda^* < 1$ such that $S(u_{\lambda^*}) = 0$. It is clear that $u \neq 0$ and $u_{\lambda^*} \neq 0$. By (3.5) it follows that

$$I(u_{\lambda^*}) \geq d. \tag{3.14}$$

At the same time, $Q(u) = 0$ implies that for any $\lambda > 0, Q(u_\lambda) = 0$. It follows that

$$I(u_\lambda) = \int_{\mathbb{R}^N} \frac{Np-N-4}{4(p+1)}|u_\lambda|^{p+1}dx + \int_{\mathbb{R}^N} \frac{1}{2}|u_\lambda|^2 dx. \tag{3.15}$$

$$S(u_\lambda) = \int_{\mathbb{R}^N} \frac{Np-2p-N-2}{2(p+1)}|u_\lambda|^{p+1}dx + \int_{\mathbb{R}^N} |u_\lambda|^2 dx. \tag{3.16}$$

By (3.15), we have

$$\lambda\frac{d}{d\lambda}I(u_\lambda) = \alpha\lambda^\alpha \int_{\mathbb{R}^N} \frac{Np-N-4}{4(p+1)}|u|^{p+1}dx + \beta\lambda^\beta \int_{\mathbb{R}^N} \frac{1}{2}|u|^2 dx. \tag{3.17}$$

Note that $Np - N - 4 = \beta(1-p)$ and $Np - 2p - N - 2 = \alpha(1-p)$. Then (3.16) and (3.17) imply that

$$\lambda\frac{d}{d\lambda}I(u_\lambda) = \frac{1}{2}\beta S(u_\lambda). \tag{3.18}$$

So $I(u_\lambda)$ takes the minimal value at $\lambda = \lambda^*$. Thus for $\lambda = 1 > \lambda^*$, we have $I(u) = I(u_\lambda) \geq I(u_{\lambda^*})$. Recall (3.14), we get $I(u) \geq d$. Therefore $d_M \geq d$.
Remark 3.1 *We call the variational problem (3.6) cross-constrained variational problem since there are two constrained conditions in (3.6). The following corresponding invariant manifold will be called cross-invariant manifold.*

Theorem 3.1 *Define*

$$K := \{\phi \in H^1(\mathbb{R}^N), \quad I(\phi) < d, Q(\phi) < 0, S(\phi) < 0\}.$$

If $1 + \frac{4}{N} \leq p < \frac{N+2}{(N-2)^+}$, *then* K *is an invariant manifold of (1.1). More precisely, from* $\varphi_0 \in K$ *it follows that the solution* $\varphi(t,x)$ *of the Cauchy problem (1.1)–(2.1) satisfies* $\varphi(t,\cdot) \in K$ *for any* $t \in [0,T)$.

Proof. Let $\varphi_0 \in K$. By Theorem 2.1, there exists a unique $\varphi(t,\cdot) \in C([0,T); H^1(\mathbb{R}^N))$ with $T \leq \infty$ such that $\varphi(t,x)$ is a solution of the Cauchy problem (1.1)–(2.1) . From (2.2), (2.3), we have

$$I(\varphi(t,\cdot)) = I(\varphi_0), \qquad t \in [0,T). \tag{3.19}$$

Thus $I(\varphi_0) < d$ implies that $I(\varphi(t,\cdot)) < d$ for any $t \in [0,T)$.

Now we show $S(\varphi(t,\cdot)) < 0$ for $t \in [0,T)$. If otherwise, from the continuity, there were a $t_0 \in (0,T)$ such that $S(\varphi(t_0,\cdot)) = 0$. By (3.19), $\varphi(t_0,\cdot) \neq 0$. From (3.5) it follows that $I(\varphi(t_0,\cdot)) \geq d$. This is contradictory with $I(\varphi(t,\cdot)) < d$ for $t \in [0,T)$. Therefore $S(\varphi(t,\cdot)) < 0$ for all $t \in [0,T)$.

At last we show $Q(\varphi(t,\cdot)) < 0$ for $t \in [0,T)$. If otherwise , from the continuity, there were a $t_1 \in (0,T)$ such that $Q(\varphi(t_1,\cdot)) = 0$. Because we have showed $S(\varphi(t_1,\cdot)) < 0$, it follows that $\varphi(t_1,\cdot) \in M$. Thus (3.6) and Lemma 3.2 imply that $I(\varphi(t_1,\cdot)) \geq d_M \geq d$. This is contradictory with $I(\varphi(t,\cdot)) < d$ for $t \in [0,T)$. Therefore $Q(\varphi(t,\cdot)) < 0$ for all $t \in [0,T)$.

From the above we proved $\varphi(t,\cdot) \in K$ for any $t \in [0,T)$.

This completes the proof of this theorem.

By the same argument as Theorem 3.1, we get the following results.

Theorem 3.2 *Define*

$$K_+ := \{\phi \in H^1(\mathbb{R}^N), I(\phi) < d, Q(\phi) > 0, S(\phi) < 0\},$$

$$R_- := \{\phi \in H^1(\mathbb{R}^N), I(\phi) < d, S(\phi) < 0\},$$

$$R_+ := \{\phi \in H^1(\mathbb{R}^N), I(\phi) < d, S(\phi) > 0\}.$$

If $1 + \frac{4}{N} \leq p < \frac{N+2}{(N-2)^+}$,*then* K_+, R_- *and* R_+ *are all invariant manifolds of (1.1).*

4 Sharp criterion for global existence and blow up

Theorem 4.1 *If* $\varphi_0 \in K_+ \cup R_+$, *then the solution* $\varphi(t,x)$ *of the Cauchy problem (1.1)–(2.1) exists globally in* $t \in (0,\infty)$.

Proof: Firstly we let $\varphi_0 \in K_+$. Thus Theorem 3.2 implies that the solution $\varphi(t,x)$ of the Cauchy problem (1.1)–(2.1) satisfies that $\varphi(t,\cdot) \in K_+$ for $t \in [0,T)$. For fixed $t \in [0,T)$, denote $\varphi(t,\cdot) = \varphi$. Thus we have $I(\varphi) < d, Q(\varphi) > 0$. It follows that from (3.1) and (3.3)

$$\int_{\mathbb{R}^N} \left[\frac{Np-N-4}{2N(p-1)} |\nabla\varphi|^2 dx + \frac{1}{2}|\varphi|^2 \right] dx < d. \tag{4.1}$$

First we treat with the critical case $p = 1 + \frac{4}{N}$. In this case, by (4.1), we have

$$\int_{\mathbb{R}^N} \frac{1}{2}|\varphi|^2 dx < d. \tag{4.2}$$

We put $\varphi^\mu = \mu^{\frac{N}{p+1}}\varphi(\mu x)$. Noting that $p = 1 + \frac{4}{N}$, we get

$$Q(\varphi^\mu) = \mu^{\frac{4}{N+2}} \int_{\mathbb{R}^N} |\nabla\varphi|^2 dx - \int_{\mathbb{R}^N} \frac{N}{2}\frac{p-1}{p+1}|\varphi|^{p+1} dx. \tag{4.3}$$

Thus $Q(\varphi) > 0$ implies that there exists a $0 < \mu^* < 1$ such that $Q(\varphi^{\mu^*}) = 0$ By (3.1), (3.3) , we have

$$I(\varphi^{\mu^*}) = \int_{\mathbb{R}^N} \frac{1}{2}|\varphi^{\mu^*}|^2 dx = \mu^{*\frac{-2N}{N+2}} \int_{\mathbb{R}^N} \frac{1}{2}|\varphi|^2 dx. \tag{4.4}$$

From (4.2), it follows that

$$I(\varphi^{\mu^*}) < \mu^{*\frac{-2N}{N+2}} d. \tag{4.5}$$

Now we see $S(\varphi^{\mu^*})$, which has two possibilities. One is $S(\varphi^{\mu^*}) < 0$. In this case, note that $Q(\varphi^{\mu^*}) = 0$, then Lemma 3.2 implies that

$$I(\varphi^{\mu^*}) \geq d_M \geq d > I(\varphi). \tag{4.6}$$

It follows that

$$I(\varphi) - I(\varphi^{\mu^*}) < 0. \tag{4.7}$$

That is

$$(1 - \mu^{*\frac{4}{N+2}}) \int_{\mathbb{R}^N} \frac{1}{2}|\nabla\varphi|^2 dx + (1 - \mu^{*\frac{-2N}{N+2}}) \int_{\mathbb{R}^N} \frac{1}{2}|\varphi|^2 dx < 0. \tag{4.8}$$

It follows that

$$\int_{\mathbb{R}^N} |\nabla\varphi|^2 dx < c \int_{\mathbb{R}^N} |\varphi|^2 dx. \tag{4.9}$$

By (4.2), we get

$$\int_{\mathbb{R}^N} |\nabla\varphi|^2 dx < c. \tag{4.10}$$

For $S(\varphi^{\mu^*})$, the other possible case is $S(\varphi^{\mu^*}) \geq 0$. In this case, from (4.5), we have

$$I(\varphi^{\mu^*}) - \frac{1}{p+1} S(\varphi^{\mu^*}) < \mu^{*\frac{-2N}{N+2}} d. \tag{4.11}$$

It follows that

$$\left(\frac{1}{2} - \frac{1}{p+1}\right) \int_{\mathbb{R}^N} (|\nabla\varphi^{\mu^*}|^2 + |\varphi^{\mu^*}|^2) dx < \mu^{*\frac{-2N}{N+2}} d. \tag{4.12}$$

That is

$$\mu^{*\frac{4}{N+2}} \int_{\mathbb{R}^N} |\nabla\varphi|^2 dx + \mu^{*\frac{-2N}{N+2}} \int_{\mathbb{R}^N} |\varphi|^2 dx < \frac{2(p+1)}{p-1} \mu^{*\frac{-2N}{N+2}} d. \tag{4.13}$$

It follows that

$$\int_{\mathbb{R}^N} |\nabla\varphi|^2 dx < c. \tag{4.14}$$

(4.10) and (4.14) show that in the critical case $p = 1 + \frac{4}{N}$, we always get $\int_{\mathbb{R}^N} |\nabla\varphi|^2 dx$ is bounded for any $t \in [0, T)$. Thus by Theorem 2.1, we get $\varphi(t,x)$ exists globally in $t \in [0, \infty)$.

For the subcritical case $1 < p < 1 + \frac{4}{N}$, from [5], for any $\varphi_0 \in H^1(\mathbb{R}^N), \varphi(t,x)$ exists globally in $t \in [0, \infty)$.

For the supercritical case $1 + \frac{4}{N} < p < \frac{N+2}{(N-2)^+}$, from (4.1), we always get

$$\int_{\mathbb{R}^N} |\nabla\varphi|^2 dx < c. \tag{4.15}$$

Therefore Theorem 2.1 implies that $\varphi(t,x)$ exists globally in $t \in [0, T)$.

Thus for $\varphi_0 \in K_+$ we proved the solution $\varphi(t,x)$ of the Cauchy problem (1.1)–(2.1) exists globally in $t \in [0, \infty)$.

Now we see $\varphi_0 \in R_+$. This case is simple. By $\varphi_0 \in R_+$, Theorem 3.2 implies that the solution $\varphi(t,x)$ of the Cauchy problem (1.1)–(2.1) satisfies that $\varphi(t,\cdot) \in R_+$ for $t \in [0,T)$. Thus we have $I(\varphi(t,\cdot)) \in R_+$ for $t \in [0,T)$. Then we have $I(\varphi) < d, S(\varphi) > 0$. It follows that

$$\left(\frac{1}{2} - \frac{1}{p+1}\right) \int_{\mathbb{R}^N} (|\nabla\varphi|^2 + |\varphi|^2) dx < d. \tag{4.16}$$

Thus Theorem 2.1 implies that $\varphi(t,x)$ exists globally in $t \in [0, \infty)$.

Theorem 4.2 *Let* $1+\frac{4}{N}\leq p<\frac{N+2}{(N-2)^+}$. *If* $\varphi_0\in K$ *and satisfies* $|\cdot|\varphi_0(\cdot)\in L^2(\mathbb{R}^N)$, *then the solution* $\varphi(t,x)$ *of the Cauchy problem (1.1)-(2.1) blows up in a finite time.*

Proof: From $\varphi_0\in K$, Theorem 3.1 implies that $\varphi(t,\cdot)\in K$ with $t\in[0,T)$. For $J(t)=\int_{\mathbb{R}^N}\frac{1}{2}|x|^2|\varphi(t,x)|^2dx$, (2.4) and (3.3) imply that

$$J''(t)=4Q(\varphi(t,\cdot)),\qquad t\in[0,T). \tag{4.17}$$

Fix $t\in[0,T)$, and denote $\varphi(t,\cdot)=\varphi$. Thus φ satisfies that $Q(\varphi)<0,S(\varphi)<0$. For $\lambda>0$, we let $\varphi_\lambda=\lambda^{\frac{N}{2}}\varphi(\lambda x)$. Thus

$$S(\varphi_\lambda)=\lambda^2\int_{\mathbb{R}^N}|\nabla\varphi|^2dx+\int_{\mathbb{R}^N}|\varphi|^2dx-\lambda^{\frac{N(p-1)}{2}}\int_{\mathbb{R}^N}|\varphi|^{p+1}dx. \tag{4.18}$$

$$Q(\varphi_\lambda)=\lambda^2\int_{\mathbb{R}^N}|\nabla\varphi|^2dx-\lambda^{\frac{N(p-1)}{2}}\int_{\mathbb{R}^N}\frac{N}{2}\frac{p-1}{p+1}|\varphi|^{p+1}dx. \tag{4.19}$$

Since $1+\frac{4}{N}\leq p<\frac{N+2}{(N-2)^+}$, $S(\varphi)<0$, it yields that there exists $0<\lambda^*<1$ such that $S(\varphi_{\lambda^*})=0$, and when $\lambda\in(\lambda^*,1],S(\varphi_\lambda)<0$. For $\lambda\in[\lambda^*,1],Q(\varphi_\lambda)$ has the following three possibilities.

i). $Q(\varphi_\lambda)<0$ *for* $\lambda\in[\lambda^*,1]$.

ii). $Q(\varphi^*_\lambda)=0$.

iii). There exists $\lambda^*<\mu<1$ such that $Q(\varphi_\mu)=0$.

For the case i) and ii), we all have $S(\varphi_{\lambda^*})=0$ and $Q(\varphi_{\lambda^*})\leq 0$. It follows that $I(\varphi_{\lambda^*})\geq d$. Moreover by

$$I(\varphi)-I(\varphi_{\lambda^*})=\frac{1}{2}(1-\lambda^{*2})\int_{\mathbb{R}^N}|\nabla\varphi|^2dx-\frac{1}{p+1}[1-\lambda^{*\frac{N(p-1)}{2}}]\int_{\mathbb{R}^N}|\varphi|^{p+1}dx, \tag{4.20}$$

$$Q(\varphi)-Q(\varphi_{\lambda^*})=(1-\lambda^{*2})\int_{\mathbb{R}^N}|\nabla\varphi|^2dx-\frac{N}{2}\frac{p-1}{p+1}[1-\lambda^{*\frac{N(p-1)}{2}}]\int_{\mathbb{R}^N}|\varphi|^{p+1}dx, \tag{4.21}$$

$0<\lambda^*<1$ and $1+\frac{4}{N}\leq p<\frac{N+2}{(N-2)^+}$, we have

$$I(\varphi)-I(\varphi_{\lambda^*})\geq\frac{1}{2}Q(\varphi)-\frac{1}{2}Q(\varphi_{\lambda^*})\geq\frac{1}{2}Q(\varphi). \tag{4.22}$$

For the case iii), we have $Q(\varphi_\mu)=0$ and $S(\varphi_\mu)<0$. Thus Lemma 3.2 implies that $I(\varphi_\mu)\geq d_M\geq d$. And

$$I(\varphi)-I(\varphi_\mu)\geq\frac{1}{2}Q(\varphi)-\frac{1}{2}Q(\varphi_\mu)\geq\frac{1}{2}Q(\varphi). \tag{4.23}$$

Since $I(\varphi_{\lambda^*}) \geq d, I(\varphi_\mu) \geq d$, from (4.22), (4.23), we all get

$$Q(\varphi) < 2[I(\varphi) - d]. \tag{4.24}$$

From (2.2), (2.3), and (3.1), $I(\varphi) = I(\varphi_0)$. Thus by $\varphi_0 \in K$ and (4.17), we have

$$J''(t) = 4Q(\varphi) < 8[I(\varphi_0) - d] < 0. \tag{4.25}$$

Obviously $J(t)$ can not verify (4.25) for all time t. Therefore from Theorem 2.1, it must be the case that $T < \infty$, which implies

$$\lim_{t \to T} ||\varphi(t, \cdot)||_{H^1(\mathbb{R}^N)} = \infty.$$

Remark 4.1 *It is clear that*

$$\{\phi \in H^1(\mathbb{R}^N)\backslash\{0\}, \quad I(\phi) < d\} = R_+ \cup K_+ \cup K.$$

Thus Theorem 4.2 shows that Theorem 4.1 is sharp.

Corollary 4.1 *Let $\varphi_0 \in H^1(\mathbb{R}^N)$ and satisfy $||\varphi_0||^2_{H^1(\mathbb{R}^N)} < 2d$. Then the solution $\varphi(t, x)$ of the Cauchy problem (1.1)–(2.1) exists globally in $t \in [0, \infty)$.*

Proof. From $||\varphi_0||^2_{H^1(\mathbb{R}^N)} < 2d$, we have $I(\varphi_0) < d$. Moreover we claim that $S(\varphi_0) > 0$. If otherwise, there were a $0 < \lambda \leq 1$ such that $S(\lambda\varphi_0) = 0$. Thus $I(\lambda\varphi_0) \geq d$. On the other hand

$$||\lambda\varphi_0||^2_{H^1(\mathbb{R}^N)} = \lambda^2||\varphi_0||^2_{H^1(\mathbb{R}^N)} < 2\lambda^2 d < 2d.$$

It follows that $I(\lambda\varphi_0) < d$. This is a contradiction. Therefore we have $\varphi_0 \in R_+$. Thus Theorem 4.1 implies this corollary.

5 Instability of the standing waves

Let u be a solution of (3.5), that is we have

$$d = \min_{\{u \in H^1(\mathbb{R}^N)\backslash\{0\}, S(u)=0\}} I(u). \tag{5.1}$$

Then by a standard variational computation, we have that u is a solution of the following nonlinear Euclidean scalar equation

$$-\Delta u + u - u|u|^{p-1} = 0, \quad u \in H^1(\mathbb{R}^N)\backslash\{0\}. \tag{5.2}$$

Thus $\varphi(t, x) = e^{it}u(x)$ is a standing wave solution of (1.1). Since u is a minimizer of (5.1) , we call $u(x)$ to be a ground state solution of (5.2). [13] and [2] all provided the existence of the minimizer of (5.1) for $N \geq 2$. Berestycki and Cazenave [1] have proved the strong instability of the standing wave. But their proof has to rely on the solvability of the following variational problem

$$d_Q := \inf_{\{u \in H^1(\mathbb{R}^N)\setminus\{0\}, Q(u)=0\}} I(u). \tag{5.3}$$

In general this is difficult even though we have got the solvability of (5.1). Now by Lemma 3.2 and Theorem 4.2, we can refrain from solving the problem (5.3), and show the instability directly. Firstly we state two lemmas.

Lemma 5.1 *Let $\phi \in H^1(\mathbb{R}^N)\setminus\{0\}$. Then there exists a unique $\mu > 0$ such that $S(\mu\phi) = 0$ and $I(\mu\phi) > I(\lambda\phi)$ for any $\lambda > 0$ and $\lambda \neq \mu$.*

Proof. For $\lambda > 0$, we have

$$S(\lambda\varphi) = \lambda^2 \int_{\mathbb{R}^N} (|\phi|^2 + |\nabla\varphi|^2)dx - \lambda^{p+1} \int_{\mathbb{R}^N} |\varphi|^{p+1}dx. \tag{5.4}$$

$$\frac{d}{d\lambda} I(\lambda\phi) = \lambda^{-1} S(\lambda\phi). \tag{5.5}$$

From (5.4) and (5.5), Lemma 5.1 is obtained immediately.

Lemma 5.2 *Let u be a minimizer of (5.1). Then $Q(u) = 0$.*

Proof: Since u is a minimizer of (5.1), u is also a solution of (5.2). Thus we have Pohzaev identity

$$\int_{\mathbb{R}^N} \left(|u|^2 + \frac{N-2}{N}|\nabla u|^2 - \frac{2}{p+1}\right) dx = 0, \tag{5.6}$$

which is obtained from multiplying (5.2) by $x \cdot \nabla u$. Note that $S(u) = 0$. Thus $Q(u) = 0$.

Remark 5.1 *From (5.1), (5.2) and Lemma 5.2, we know that there exists $u \in H^1(\mathbb{R}^N)\setminus\{0\}$ such that both $S(u) = 0$ and $Q(u) = 0$. As (3.10), we denote $u_\lambda = \lambda^{\frac{2}{p-1}} u(\lambda x)$ for $\lambda > 0$. Then from (3.11), (3.12) and (3.13), for $\lambda > 1$ we always have $S(u_\lambda) < 0$ and $Q(u_\lambda) = 0$. This shows that M in (3.4) is not empty.*

Now we give the following theorem which originates in Berestycki and Cazenave [1] as well as Weinstein [16]. As stated in the above, our argument in terms of Theorem 4.2 is more direct and simple than ones in [1] and [16].

Theorem 5.1 *For $1 + \frac{4}{N} \leq p < \frac{N+2}{(N-2)^+}$, let u be a minimizer of (5.1). Then for any $\varepsilon > 0$, there exists $\varphi_0 \in H^1(\mathbb{R}^N)$ with $||\varphi_0 - u||_{H^1(\mathbb{R}^N)} < \varepsilon$ such that the solution $\varphi(t, x)$ of the Cauchy problem (1.1)-(2.1) blows up in a finite time.*

Proof. From Lemma 5.2, $Q(u) = 0$. Thus we have $S(u) = 0; Q(u) = 0$. It follows that for any $\lambda > 1$, we have

$$S(\lambda u) < 0, \qquad Q(\lambda u) < 0, \qquad \lambda > 1. \tag{5.7}$$

On the other hand from Lemma 5.1, $S(u) = 0$ implies that $I(\lambda u) < I(u)$ for any $\lambda > 1$. Note that $I(u) = d$. Thus for any $\lambda > 1$ we have $\lambda u \in K$. Furthermore since u has an exponential fall-off at infinity (see e. g. Strauss [13] or Berestycki and Lions [2]), it is clear that $| \cdot |u(\cdot) \in L^2(\mathbb{R}^N)$. Thus $\lambda | \cdot |u(\cdot) \in L^2(\mathbb{R}^N)$. Now we take $\lambda > 1$, and λ is sufficiently close to 1 such that

$$||\lambda u - u||_{H^1(\mathbb{R}^N)} = (\lambda - 1)||u||_{H^1(\mathbb{R}^N)} < \varepsilon. \tag{5.8}$$

Then take $\varphi_0 = \lambda u(x)$. From Theorem 4.2, the solution $\varphi(t,x)$ of the Cauchy problem (1.1)–(2.1) blows up in a finite time.

This completed the proof of this theorem.

In addition, for the above u, if we put

$$u_\omega(x) = \omega^{\frac{1}{p-1}} u(\omega^{-\frac{1}{2}} x), \quad \textit{with} \quad \omega > 0, \tag{5.9}$$

then by a scaling argument , u_ω is a solution of the variational problem

$$d_\omega := \min_{\{v \in H^1(\mathbb{R}^N) \setminus \{0\}, S_\omega(v) = 0\}} I_\omega(v), \tag{5.10}$$

where

$$I_\omega(v) := \int_{\mathbb{R}^N} \left(\frac{1}{2}\omega|v|^2 + \frac{1}{2}|\nabla v|^2 - \frac{1}{p+1}|v|^{p+1} \right) dx. \tag{5.11}$$

$$S_\omega(v) := \int_{\mathbb{R}^N} (\omega|v|^2 + |\nabla v|^2 - |v|^{p+1}) dx. \tag{5.12}$$

Moreover u_ω is also a solution of the equation

$$-\Delta v + \omega v - v|v|^{p-1} = 0, \quad v \in H^1(\mathbb{R}^N) \setminus \{0\}. \tag{5.13}$$

Thus for every $\omega > 0$,

$$\varphi(t,x) = e^{i\omega t} u_\omega(x) \tag{5.14}$$

is a standing wave solution of (1.1). By a completely analogous process, we may get

Theorem 5.2 *For* $1 + \frac{4}{N} \le p < \frac{N+2}{(N-2)^+}$, *let* u *be a minimizer of (5.1). Then for any* $\omega > 0$ *and* $\varepsilon > 0$, *there exists* $\varphi_0 \in H^1(\mathbb{R}^N)$ *with* $||\varphi_0 - u_\omega||_{H^1(\mathbb{R}^N)} < \varepsilon$ *such that the solution* $\varphi(t,x)$ *of the Cauchy problem (1.1)–(2.1) blows up in a finite time.*

Acknowledgment

I am grateful to Professor H. Matano and Professor Y. Tsutsumi for their helpful comments, and Professor M. Ohta for valuable discussions. This work was partly Supported by the JSPS Research Fellowship and the Monbusho's Grant No. P-98029.

References

1. H. Berestycki and T. Cazenave, Instabilité des états stationnaires dans les équations de Schrödinger et de Klein-Gordon non linéarires, C. R. Acad. Sci. Paris, Seire I **293**, 489–492 (1981).
2. H. Berestycki and P.-L. Lions, Nonlinear scalar field equations, Arch. Rat. Mech. Anal. **82**, 313–375 (1983).
3. T. Cazenave, *An Introduction to Nonlinear Schrödinger Equations*, (Textos de Metodos Matematicos, Vol. 22, Rio de Janeiro, 1989).
4. T. Cazenave and P.-L. Lions, Orbital stability of standing waves for some nonlinear Schödinger equations, Commun. Math. Phys. **85**, 549–561 (1982).
5. J. Ginibre and G. Velo, On a class of nonlinear Schrödinger equations, J. Funct. Anal. **32**, 1–71 (1979).
6. R. T. Glassey, On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations, J. Math. Phys. **18(9)**, 1794–1797 (1977).
7. T. Kato, On nonlinear Schrödinger equations, Ann. Inst. H. Poincaré , Phys. Théor. **46**, 113–129 (1987).
8. M. K. Kwong, Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in R^N, Arch. Rat. Mech. Anal. **105**, 243–266 (1989).
9. T. Ogawa and Y. Tsutsumi, Blow-up of H^1 solution for the nonlinear Schrödinger equation , J. Differential Equations **92(2)**, 317–330 (1991).
10. T. Ogawa and Y. Tsutsumi, Blow-up of H^1 solution for the nonlinear Schrödinger equation with critical power nonlinearity, Proceedings of the American Mathematical Society **111(2)**, 487–496 (1991).
11. G. Ribeiro, Instability of symmetric stationary states for some nonlinear Schrödinger equations with an external magnetic field, Ann. Inst. Henri Poincare, Phys. Theor. **54**, 403–433 (1991).
12. G. Ribeiro, Finite-time blow-up for some nonlinear Schrödinger equations with an external magnetic field, Nonlinear Anal., T. M. A. **16**, 941–948 (1991).
13. W. A. Straus, Existence of Solitary waves in high dimensions, Commun.

Math. Phys. **55**, 149–162 (1977).

14. W. Strauss, *Nonlinear Wave Equations*, (C.B.M.S. No. 73, Amer. Math. Soc., Providence, 1989).
15. C. Sulem, and P. L. Sulem, *The nonlinear Schrödinger Equation, Self-focusing and Wave Collapse*, (Springer, 1999).
16. M. I. Weinstein, Nonlinear Schrödinger equations and sharp interpolations estimates, Commun. Math. Phys. **87**, 567–576 (1993).
17. W. E. Zakharov, Collapse of Langmuir waves, Sov. Phys. JETP **23**, 1025-1033 (1966).
18. J. Zhang, On the finite-time behaviour for nonlinear Schrödinger equations, Commun. Math. Phys. **162**, 249–260 (1994).
19. J. Zhang, Sharp conditions of global existence for nonlinear Schrödinger and Klein-Gordon equations, to appear in *Nonlinear Anal.: Theory, Methods and Applications.*