publicações matemáticas

28º Colóquio Brasileiro de Matemática

Nonlinear Equations

Gregorio Malajovich

impa

Nonlinear Equations

Publicações Matemáticas

Nonlinear Equations

Gregorio Malajovich
UFRJ

impa

28º Colóquio Brasileiro de Matemática

Impresso no Brasil / Printed in Brazil

Capa: Noni Geiger / Sérgio R. Vaz

28º Colóquio Brasileiro de Matemática

- Cadeias de Markov e Teoria do Potencial - Johel Beltrán
- Cálculo e Estimação de Invariantes Geométricos: Uma Introdução às Geometrias Euclidiana e Afim - M. Andrade e T. Lewiner
- De Newton a Boltzmann: o Teorema de Lanford - Sérgio B. Volchan
- Extremal and Probabilistic Combinatorics - Robert Morris e Roberto Imbuzeiro Oliveira
- Fluxos Estrela - Alexander Arbieto, Bruno Santiago e Tatiana Sodero
- Geometria Aritmética em Retas e Cônicas - Rodrigo Gondim
- Hydrodynamical Methods in Last Passage Percolation Models - E. A. Cator e L. P. R. Pimentel
- Introduction to Optimal Transport: Theory and Applications - Nicola Gigli
- Introdução à Aproximação Numérica de Equações Diferenciais Parciais Via o Método de Elementos Finitos - Juan Galvis e Henrique Versieux
- Matrizes Especiais em Matemática Numérica - Licio Hernanes Bezerra
- Mecânica Quântica para Matemáticos em Formação - Bárbara Amaral, Alexandre Tavares Baraviera e Marcelo O. Terra Cunha
- Multiple Integrals and Modular Differential Equations - Hossein Movasati
- Nonlinear Equations - Gregorio Malajovich
- Partially Hyperbolic Dynamics - Federico Rodriguez Hertz, Jana Rodriguez Hertz e Raúl Ures
- Processos Aleatórios com Comprimento Variável - A. Toom, A. Ramos, A. Rocha e A. Simas
- Um Primeiro Contato com Bases de Gröbner - Marcelo Escudeiro Hernandes

ISBN: 978-85-244-329-3

Distribuição: IMPA
Estrada Dona Castorina, 110
22460-320 Rio de Janeiro, RJ
E-mail: ddic@impa.br
http://www.impa.br

To Beatriz

Foreword

I added together the ratio of the length to the width (and) the ratio of the width to the length. I multiplied (the result) by the sum of the length and the width. I multiplied the result which came out and the sum of the length and the width together, and (the result is) $1+30\times 60^{-1}+16\times 60^{-2}+40\times 60^{-3}$*. I returned. I added together the ratio of the length to the width (and) the ratio of the width to the length. I added (the result) to the 'inside' of two areas and of the square of the amount by which the length exceeded the width (and the result is)* $2+3(1\times 60^{-1}+40\times 60^{-2})$*. What are (the l)ength and the width ? (...)*

Susa mathematical text No. 12, as translated by Kazuo Muroi [64].

Since ancient times, problems reducing to nonlinear equations are recurrent in mathematics. The problem above reduces to

Gregorio Malajovich, *Nonlinear equations.*
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

solving

$$\begin{aligned}
\left(\frac{x}{y}+\frac{y}{x}\right)(x+y)^2 &= \frac{325}{216} \\
\left(\frac{x}{y}+\frac{y}{x}\right)+2xy+(x-y)^2 &= \frac{91}{36}.
\end{aligned}$$

It is believed to date from the end of the first dynasty of Babylon (16^{th} century BC). Yet, very little is known on how to efficiently *solve* nonlinear equations, and even *counting* the number of solutions of a specific nonlinear equation can be extremely challenging.

These notes

These notes correspond to a short course during the 28^{th} *Colóquio Brasileiro de Matemática*, held in Rio de Janeiro in July 2011. My plan is to let them grow into a book that can be used for a graduate course on the mathematics of nonlinear equation solving.

Several topics are not properly covered yet. Subjects such as univariate solving, modern elimination theory, straight line programs, random matrices, toric homotopy, finding start systems for homotopy, how to certify degenerate roots or curves of solutions [83], tropical geometry, Diophantine approximation, real solving and Khovanskii's theory of fewnomials [49] should certainly deserve extra chapters. Other topics may be a moving subject (see below).

At this time, these notes are *untested* and *unrefereed.* I will keep an errata in my page, `http://www.labma.ufrj.br/~gregorio`

Most of the material here is known, but some of it is new. To my knowledge, the systematic study of spaces of complex fewnomial spaces (nicknamed *fewspaces* in Definition 5.2) is not available in other books (though Theorem 5.11 was well known).

The theory of condition numbers for sparse polynomial systems (Chapter 8) presents clarifications over previous tentatives (to my knowledge only [58] and [59]). Theorem 8.23 is a strong improvement over known bounds.

Newton iteration and 'alpha theory' seem to be more mature topics, where sharp constants are known. However, I am unaware of another book with a systematic presentation that includes the sharp bounds (Chapters 7 and 9). Theorem 7.19 is new, and presents improvements over [56].

The last chapter contains novelties. The homotopy algorithm given there is a simplification of the one in [31], and allows to reduce Smale's 17$^{\text{th}}$ problem to a geometric problem. A big recent breakthrough is the construction of randomized (Vegas) algorithms that can approximate solutions of dense random polynomial systems in expected polynomial time. This is explained in Chapter 10.

Other recent books on the mathematics of polynomial/non-linear solving or with strong intersection are [20, 30], parts of [5] and a forthcoming book [26]. There is no superposition, as the subject is growing in breadth as well as in depth.

Acknowledgements

I would specially like to thank my friends Carlos Beltrán, Jean-Pierre Dedieu, Luis Miguel Pardo and Mike Shub for kindly providing the list of open problems at the end of this book. Diego Armentano, Felipe Cucker, Teresa Krick, Dinamérico Pombo and Mario Wschebor contributed with ideas and insight. I thank Tatiana Roque for explaining that the Babylonians did not think in terms of equations but arguably by completing squares, so that the opening problem may have been a geometric problem in its time.

The research program that resulted into this book was partially funded by CNPq, CAPES, FAPERJ, and by a MathAmSud cooperation grant. It was also previously funded by the Brazil-France agreement of Cooperation in Mathematics.

A warning to the reader

Problem F.1 (Algebraic equations over $\mathbb{F}_2$)**.** Given a system $\mathbf{f} = (f_1, \dots, f_s) \in \mathbb{F}_2[x_1, \dots, x_n]$, decide if there is $\mathbf{x} \in \mathbb{F}_2^n$ with $f_1(\mathbf{x}) = \cdots = f_s(\mathbf{x}) = 0$.

An instance $\mathbf{f}$ of the problem is said to have size S if the sum over all i of the sum of the degree of each monomial in f_i is equal to S.

The following is unknown:

Conjecture F.2 ($\mathbf{P} \neq \mathbf{NP}$)**.** *There cannot possibly exist an algorithm that decides problem F.1 in at most $O(S^r)$ operations, for any fixed $r > 1$.*

Above, an *algorithm* means a Turing machine, or a discrete RAM machine. For references, see [42]. Problem F.1 is AN9 p.251. It is still NP-hard if the degree of each monomial is ≤ 2.

In these notes we are mainly concerned about equations over the field of complex numbers. There is an analogous problem to 4-SAT (see [42]) or to Problem F.1, namely:

Problem F.3 (HN2, Hilbert Nullstellensatz for degree 2)**.** Given a system of complex polynomials $\mathbf{f} = (f_1, \ldots, f_s) \in \mathbb{C}[x_1, \ldots, x_n]$, each equation of degree 2, decide if there is $\mathbf{x} \in \mathbb{C}^n$ with $\mathbf{f}(\mathbf{x}) = 0$.

The polynomial above is said to have size $S = \sum S_i$ where S_i is the number of monomials of f_i. The following is also open (I personally believe it can be easier than the classical $\mathbf{P} \neq \mathbf{NP}$).

Conjecture F.4 ($\mathbf{P} \neq \mathbf{NP}$ over $\mathbb{C}$)**.** *There cannot possibly exist an algorithm that decides* HN2 *in at most $O(S^r)$ operations, for any fixed $r > 1$.*

Here, an *algorithm* means a *machine over* $\mathbb{C}$ and I refer to [20] for the precise definition.

We are not launching an attack to those hard problems here (see [63] for a credible attempt). Instead, we will be happy to obtain solution counts that are correct almost everywhere, or to look for algorithms that are efficient on average.

Contents

Chapter 1

Counting solutions of polynomial systems

In this notes, we will mostly look at equations over the field of complex numbers. The case of real equations is interesting but more difficult to handle. In many situations, it may be convenient to count or to solve over $\mathbb{C}$ rather than over $\mathbb{R}$, and then ignore non-real solutions.

Finding or even counting the solutions of specific systems of polynomials is *hard* in the complexity theory sense. Therefore, instead of looking at particular equations, we consider linear spaces of equations. Several bounds for the number of roots are known to be true *generically*. As many definitions of genericity are in use, we should be more specific.

Definition 1.1 (Zariski topology)**.** A set $V \subseteq \mathbb{C}^N$ is *Zariski closed*

Gregorio Malajovich, *Nonlinear equations.*
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

if and only if it is of the form

$$V = \{\mathbf{x} : f_1(x) = \cdots = f_s(x) = 0\}$$

for some finite (possibly empty) collection of polynomials $f_1, \ldots, f_s$. A set is *Zariski open* if it is the complementary of a Zariski closed set.

In particular, the empty set and the total space $\mathbb{C}^N$ are simultaneously open and closed.

Definition 1.2. We say that a property holds for a *generic* $\mathbf{y} \in \mathbb{C}^N$ (or more loosely for a generic choice of $y_1, \ldots, y_N$) when the set of $\mathbf{y}$ where this property holds contains a non-empty Zariski open set.

A property holding generically will also hold almost everywhere (in the measure-theory sense).

Exercise 1.1. Show that a finite union of Zariski closed sets is Zariski closed.

The proof that an arbitrary intersection of Zariski closed sets is Zariski closed (and hence the Zariski topology is a topology) is postponed to Corollary 2.7.

1.1 Bézout's theorem

Below is *the* classical theorem about root counting. The notation $\mathbf{x}^{\mathbf{a}}$ stands for

$$\mathbf{x}^{\mathbf{a}} = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}.$$

The *degree* of a multi-index $\mathbf{a}$ is $|\mathbf{a}| = a_1 + a_2 + \cdots + a_n$.

Theorem 1.3 (Étienne Bézout, 1730–1783)**.** *Let $n, d_1, \ldots, d_n \in \mathbb{N}$. For a generic choice of the coefficients $f_{i\mathbf{a}} \in \mathbb{C}$, the system of equa-*

tions

$$\begin{aligned} f_1(\mathbf{x}) &= \sum_{|\mathbf{a}| \le d_1} f_{1\mathbf{a}} \mathbf{x}^{\mathbf{a}} \\ &\vdots \\ f_n(\mathbf{x}) &= \sum_{|\mathbf{a}| \le d_n} f_{n\mathbf{a}} \mathbf{x}^{\mathbf{a}} \end{aligned}$$

has exactly $B = d_1 d_2 \dots d_n$ *roots* $\mathbf{x}$ *in* $\mathbb{C}^n$. *The number of isolated roots is never more than* B.

This can be restated in terms of homogeneous polynomials with roots in projective space $\mathbb{P}^n$. We introduce a new variable x_0 (the *homogenizing variable*) so that all monomials in the i-th equation have the same degree. We denote by f_i^{h} the homogenization of f_i,

$$f_i^{\mathrm{h}}(x_0, \dots, x_n) = x_0^{d_i} f_i \left(\frac{x_1}{x_0}, \dots, \frac{x_n}{x_0} \right)$$

Once this is done, if $(x_0, \cdots, x_n)$ is a simultaneous root of all f_i^{h}'s, so is $(\lambda x_0, \cdots, \lambda x_n)$ for all $\lambda \in \mathbb{C}$. Therefore, we count complex 'lines' through the origin instead of points in $\mathbb{C}^{n+1}$.

The space of complex lines through the origin is known as the projective space $\mathbb{P}^n$. More formally, $\mathbb{P}^n$ is the quotient of $(\mathbb{C}^{n+1})_{\neq 0}$ by the multiplicative group $\mathbb{C}_\times$.

A root $(z_1, \dots, z_n) \in \mathbb{C}^n$ of $\mathbf{f}$ corresponds to the line $(\lambda, \lambda z_1, \dots, \lambda z_n)$, also denoted by $(1 : z_1 : \cdots : z_n)$. That line is a root of $\mathbf{f}^{\mathrm{h}}$.

Roots $(z_0 : \cdots : z_n)$ of $\mathbf{f}^{\mathrm{h}}$ are of two types: if $z_0 \neq 0$, then z corresponds to the root $(z_1/z_0, \dots, z_n/z_0)$ of $\mathbf{f}$, and is said to be *finite*. Otherwise, z is said to be *at infinity*.

We will give below a short and sketchy proof of Bézout's theorem. It is based on four basic facts, not all of them proved here.

The first fact is that *Zariski open sets are path-connected.* Suppose that V is a Zariski closed set, and that $\mathbf{y}_1 \neq \mathbf{y}_2$ are not points of V. (This already implies $V \neq \mathbb{C}^n$). We claim that there is a path connecting $\mathbf{y}_1$ to $\mathbf{y}_2$ not cutting V. It suffices to exhibit a path in the complex 'line' L passing through $\mathbf{y}_1$ and $\mathbf{y}_2$, which can be parameterized by

$$(1-t)\mathbf{y}_1 + t\mathbf{y}_2, \quad t \in \mathbb{C}.$$

The set $L \cap V$ is the set of the simultaneous zeros of polynomials $f_i((1-t)\mathbf{y}_1 + t\mathbf{y}_2)$, where f_i are the defining polynomials of V. Hence $L \cap V$ is the zero set of the greater common divisor of those polynomials. It is a finite (possibly empty) set of points. Hence there is a path between $\mathbf{y}_1$ and $\mathbf{y}_2$ not crossing those points.

The second fact is a classical result in Elimination Theory. *Given a system of homogeneous polynomials* $\mathbf{g}(\mathbf{x})$ *with indeterminate coefficients, the coefficient values such that there is a common solution in* $\mathbb{P}^n$ *are a Zariski closed set.* This will be Theorem 2.33.

The third fact is that *the set of polynomial systems with a root at infinity is Zariski closed.* A system $\mathbf{g}$ has a root $\mathbf{x}$ at infinity if and only if for each i,

$$G_i(x_1, \ldots, x_n) \stackrel{\text{def}}{=} g_i^{\text{h}}(0, x_1, \ldots, x_n) = 0$$

for some choice of the $x_1, \ldots, x_n$. Now, each G_i is homogeneous of degree d_i in n variables. By the fact #2, this happens only for the G_i (hence the g_i) in some Zariski-closed set.

The fourth fact is that *the number of isolated roots is lower semi-continuous* as a function of the coefficients of the polynomial system $\mathbf{f}$. This is a topological fact about systems of complex analytic equations (Corollary 3.9). It is not true for real analytic equations.

Sketch: Proof of Bézout's Theorem. We consider first the polynomial system

$$\begin{array}{rcl} f_1^{\text{ini}}(\mathbf{x}) & = & x_1^{d_1} - 1 \\ & \vdots & \\ f_n^{\text{ini}}(\mathbf{x}) & = & x_n^{d_n} - 1. \end{array}$$

This polynomial has exactly $d_1 d_2 \cdots d_n$ roots in $\mathbb{C}^n$ and no root at infinity. The derivative $D\mathbf{f}(z)$ is non-degenerate at any root z.

The derivative of the evaluation function $\mathrm{ev} : \mathbf{f}, \mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ is

$$\dot{\mathbf{f}}, \dot{\mathbf{x}} \mapsto D\mathbf{f}(\mathbf{x})\dot{\mathbf{x}} + \dot{\mathbf{f}}(\mathbf{x}).$$

Assume that $\mathbf{f}_0(\mathbf{x}_0) = 0$ with $D\mathbf{f}_0(\mathbf{x}_0)\dot{\mathbf{x}}$ non-degenerate. Then the derivative of ev with respect to the $\mathbf{x}$ variables is an isomorphism. By the implicit function theorem, there is a neighborhood $U \ni \mathbf{f}_0$ and a function $\mathbf{x}(\mathbf{f}) : U \to \mathbb{C}^n$ so that $\mathbf{f}(\mathbf{x}_0) = \mathbf{f}_0$ and

$$\mathrm{ev}(\mathbf{f}(\mathbf{x}), \mathbf{x}) \equiv 0.$$

Now, let

$$\Sigma = \left\{ \mathbf{f} : \exists \mathbf{x} \in \mathbb{P}^{n+1} : \mathbf{f}^{\mathrm{h}}(1, \mathbf{x}) = 0 \text{ and } \left(\det D\mathbf{f}(\cdot)\right)^{\mathrm{h}}(1, \mathbf{x}) = 0 \right\}.$$

By elimination theory, Σ is a Zariski closed set. It does not contain $\mathbf{f}^{\mathrm{ini}}$, so its complement is not empty.

Let $\mathbf{g}$ be a polynomial system not in Σ and without roots at infinity. (Fact 3 says that this is true for a *generic* $\mathbf{g}$). We claim that $\mathbf{g}$ has the same number of roots as $\mathbf{f}^{\mathrm{ini}}$.

Since Σ and the set of polynomials with roots at infinity are Zariski closed, there is a smooth path (or *homotopy*) between $\mathbf{f}^{\mathrm{ini}}$ and $\mathbf{g}$ avoiding those sets. Along this path, locally, the root count is constant. Indeed, let $I \subseteq [0, 1]$ be the maximal interval so that the implicit function $\mathbf{x}_t$ for $\mathbf{f}_t(\mathbf{x}_t) \equiv 0$ can be defined. Let $t_0 = \sup I$. If $1 \neq t_0 \in I$, then (by the implicit function theorem) the implicit function $\mathbf{x}_t$ can be extended to some interval $(0, t_0 + \epsilon)$ contradicting that $t_0 = \sup I$. So let's suppose that $t_0 \notin I$. The fact that $\mathbf{f}_{t_0}$ has no root at infinity makes $\mathbf{x}_t$ convergent when $t \to t_0 \pm \epsilon$. Hence $\mathbf{x}_t$ can be extended to the closed interval $[0, t_0]$, another contradiction. Therefore $I = [0, 1]$.

Thus, $\mathbf{f}^{\mathrm{ini}}$ and $\mathbf{g}$ have the same number of roots.

Until now we counted roots of systems outside Σ. Suppose that $\mathbf{f} \in \Sigma$ has more roots than the Bézout bound. By lower semi-continuity of the root count, there is a neighborhood of $\mathbf{f}$ (in the usual topology) where there are at least as many roots as in $\mathbf{f}$. However, this neighborhood is not contained in Σ, contradiction.

$\square$

1.2 Shortcomings of Bézout's Theorem

The example below (which I learned long ago from T.Y. Li) illustrates one of the major shortcomings of Bézout's theorem:

Example 1.4. Let A be a $n \times n$ matrix, and we consider the eigenvalue problem

$$A\mathbf{x} - \lambda \mathbf{x} = 0.$$

Eigenvectors are defined up to a multiplicative constant, so let us fix $x_n = 1$. We have $n-1$ equations of degree 2 and one linear equation. The Bézout bound is $B = 2^{n-1}$.

Of course there should be (generically) n eigenvalues with a corresponding eigenvector. The other solutions given by Bézout bound lie *at infinity*: if one homogenizes the system, say

$$\begin{array}{rcl} \sum_{j=1}^{n-1} a_{1j}\mu x_j + a_{1n}\mu^2 - \lambda x_1 & = & 0 \\ & \vdots & \\ \sum_{j=1}^{n-1} a_{n-1,j}\mu x_j + a_{n-1,n}\mu^2 - \lambda x_{n-1} & = & 0 \\ \sum_{j=1}^{n-1} a_{nj} x_j + a_{n,n}\mu - \lambda & = & 0 \end{array}$$

where μ is the homogenizing variable, and then set $\mu = 0$, one gets:

$$\begin{array}{rcl} -\lambda x_1 & = & 0 \\ & \vdots & \\ -\lambda x_{n-1} & = & 0 \\ \sum_{j=1}^{n-1} a_{nj} x_j - \lambda & = & 0 \end{array}$$

This defines an $n-2$-dimensional space of solutions at infinity for $\lambda = 0$ and $a_{n1}x_1 + \cdots + a_{n,n-1}x_{n-1} = 0$.

Here is what happened: when $n \geq 2$, no system of the form $A\mathbf{x} - \lambda \mathbf{x} = 0$ can be generic in the space of polynomials systems of degree $(2, 2, \cdots, 2, 1)$. This situation is quite common, and it pays off to refine Bézout's bound.

One can think of the system above as a bi-linear homogeneous system, of degree 1 in the variables $x_1, \dots, x_{n-1}, x_n$ and degree 1 in variables λ, μ. The equations are now

$$\mu A\mathbf{x} - \lambda \mathbf{x} = 0.$$

The eigenvectors $\mathbf{x}$ are elements of projective space $\mathbb{P}^n$ and the eigenvalue is $(\lambda : \mu) \in \mathbb{P} = \mathbb{P}^1$. Examples of "ghost" roots in $\mathbb{P}^{n+1}$ but not in $\mathbb{P}^{n-1} \times \mathbb{P}$ are, for instance, the codimension 2 subspace $\lambda = \mu = 0$.

In general, let $n = n_1 + \cdots + n_s$ be a partition of n. We will divide variables $x_1, \dots, x_n$ into s sets, and write $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_s)$ for $\mathbf{x}_i \in \mathbb{C}^{n_i}$. The same convention will hold for multi-indices.

Theorem 1.5 (Multihomogeneous Bézout). *Let $n = n_1 + \cdots + n_s$, with $n_1, \dots, n_s \in \mathbb{N}$. Let $d_{ij} \in F_{\geq 0}$ be given for $1 \leq i \leq n$ and $1 \leq j \leq s$.*

Let B denote the coefficient of $\omega_1^{n_1} \omega_2^{n_2} \cdots \omega_s^{n_s}$ in

$$\prod_{i=1}^{n} (d_{i1}\omega_1 + \cdots + d_{is}\omega_s) .$$

Then, for a generic choice of coefficients $f_{i\mathbf{a}} \in \mathbb{C}$, the system of equations

$$\begin{array}{rcl} f_1(\mathbf{x}) & = & \displaystyle\sum_{\substack{|\mathbf{a}_1| \leq d_{11} \\ \vdots \\ |\mathbf{a}_s| \leq d_{1s}}} f_{1\mathbf{a}} \mathbf{x}_1^{\mathbf{a}_1} \cdots \mathbf{x}_s^{\mathbf{a}_s} \\ & \vdots & \\ f_n(\mathbf{x}) & = & \displaystyle\sum_{\substack{|\mathbf{a}_1| \leq d_{n1} \\ \vdots \\ |\mathbf{a}_s| \leq d_{ns}}} f_{n\mathbf{a}} \mathbf{x}_1^{\mathbf{a}_1} \cdots \mathbf{x}_s^{\mathbf{a}_s} \end{array}$$

has exactly B roots $\mathbf{x}$ in $\mathbb{C}^n$. The number of isolated roots is never more than the number above.

This can also be formulated in terms of homogeneous polynomials and roots in multi-projective space $\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_s}$. The above theorem is quite convenient when the partition of variables is given.

The reader should be aware that it is **NP**-hard to find, given a system, the best partition of variables [57]. Even computing an approximation of the minimal Bézout B is **NP**-hard.

A formal proof of Theorem 1.5 is postponed to Section 5.5.

Exercise 1.2. Prove Theorem 1.5, assuming the same basic facts as in the proof of Bézout's Theorem.

1.3 Sparse polynomial systems

The following theorems will be proved in chapter 6.

Theorem 1.6 (Kushnirenko [52]). *Let $A \subset \mathbb{Z}^n$ be finite. Let $\mathcal{A}$ be the convex hull of A. Then, for a generic choice of coefficients $f_{i\mathbf{a}} \in \mathbb{C}$, the system of equations*

$$\begin{aligned} f_1(\mathbf{x}) &= \sum_{\mathbf{a}\in A} f_{1\mathbf{a}} \mathbf{x}^{\mathbf{a}} \\ &\vdots \\ f_n(\mathbf{x}) &= \sum_{\mathbf{a}\in A} f_{n\mathbf{a}} \mathbf{x}^{\mathbf{a}} \end{aligned}$$

has exactly $B = n!\mathrm{Vol}(\mathcal{A})$ roots $\mathbf{x}$ in $(\mathbb{C} \setminus \{0\})^n$. The number of isolated roots is never more than B.

The case $n = 1$ was known to Newton, and $n = 2$ was published by Minding [62] in 1841.

We call A the *support* of equations $f_1, \ldots, f_n$. When each equation has a different support, root counting requires a more subtle statement.

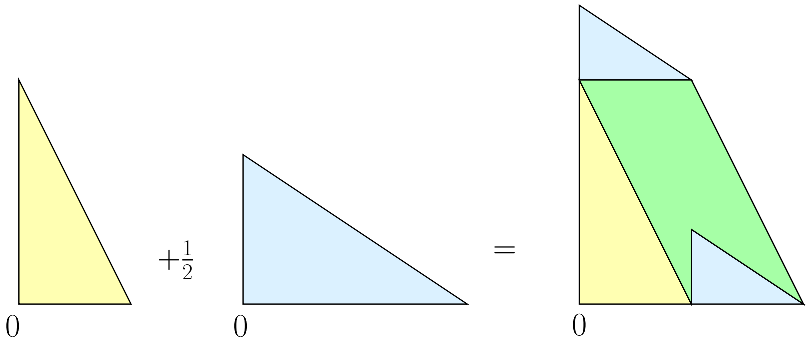

Figure 1.1: Minkowski linear combination.

Definition 1.7 (Minkowski linear combinations)**.** (See fig.1.1) Given convex sets $\mathcal{A}_1, \dots, \mathcal{A}_n$ and fixed coefficients $\lambda_1, \dots, \lambda_n$, the linear combination $\lambda_1 \mathcal{A}_1 + \cdots + \lambda_n \mathcal{A}_n$ is the set of all

$$\lambda_1 \mathbf{a}_1 + \cdots + \lambda_n \mathbf{a}_n$$

where $\mathbf{a}_i \in \mathcal{A}_i$.

The reader will show in the exercises that

Proposition 1.8. *Let $\mathcal{A}_1, \dots, \mathcal{A}_s$ be compact convex subsets of $\mathbb{R}^n$. Let $\lambda_1, \dots, \lambda_s > 0$. Then,*

$$\mathrm{Vol}(\lambda_1 \mathcal{A}_1 + \cdots + \lambda_s \mathcal{A}_s)$$

is a homogeneous polynomial of degree s in $\lambda_1, \dots, \lambda_s$.

Theorem 1.9 (Bernstein [17])**.** *Let $A_1, \dots, A_n \subset \mathbb{Z}^n$ be finite sets. Let $\mathcal{A}_i$ be the convex hull of A_i. Let B be the coefficient of $\lambda_1 \dots \lambda_n$ in the polynomial*

$$\mathrm{Vol}(\lambda_1 \mathcal{A}_1 + \cdots + \lambda_n \mathcal{A}_n).$$

Then, for a generic choice of coefficients $f_{i\mathbf{a}} \in \mathbb{C}$, the system of *equations*

$$
\begin{aligned}
f_1(\mathbf{x}) &= \sum_{\mathbf{a}\in A_1} f_{1\mathbf{a}}\mathbf{x}^{\mathbf{a}} \\
&\vdots \\
f_n(\mathbf{x}) &= \sum_{\mathbf{a}\in A_n} f_{n\mathbf{a}}\mathbf{x}^{\mathbf{a}}
\end{aligned}
$$

has exactly B *roots* $\mathbf{x}$ *in* $(\mathbb{C} \setminus \{0\})^n$. *The number of isolated roots is never more than* B.

The number $B/n!$ is known as the *mixed volume* of $\mathcal{A}_1, \ldots, \mathcal{A}_n$. The generic root number B is also known as the BKK bound, after Bernstein, Kushnirenko and Khovanskii [18].

The objective of the Exercises below is to show Proposition 1.8. We will show it first for $s = 2$. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be compact convex subsets of $\mathbb{R}^n$. Let E_i denote the linear hull of A_i, and assume without loss of generality that 0 is in the interior of A_i as a subset of E_i.

For any point $x \in \mathcal{A}_1$, define the cone x^C as the set of all $y \in E_2$ with the following property: for all $x' \in A_1$, $\langle y, x - x' \rangle \geq 0$.

Exercise 1.3. Let $\lambda_1, \lambda_2 > 0$ and $\mathcal{A} = \lambda_1\mathcal{A}_1 + \lambda_2\mathcal{A}_2$. Show that for all $z \in A$, there are $x \in \mathcal{A}_1$, $y \in x^C \cap \mathcal{A}_2$ such that $z = \lambda_1 x + \lambda_2 y$.

Exercise 1.4. Show that this decomposition is unique.

Exercise 1.5. Assume that λ_1 and λ_2 are fixed. Show that the map $z \mapsto (x, y)$ given by the decomposition above is Lipschitz.

At this point you need to believe the following fact.

Theorem 1.10 (Rademacher). *Let* U *be an open subset of* $\mathbb{R}^n$. *Let* $f : U \to \mathbb{R}^m$ *be Lipschitz. Then* f *is smooth, except possibly on a measure zero subset.*

Exercise 1.6. Use Rademacher's theorem to show that $z \mapsto (x, y)$ is smooth almost everywhere. Can you give a description of the non-smoothness set?

Exercise 1.7. Conclude the proof of Proposition 1.8 with $s = 2$.

Exercise 1.8. Generalize for all values of s.

1.4 Smale's 17$^{\text{th}}$ problem

Theorems like Bézout's or Bernstein's give precise information on the solution of systems of polynomial equations. Proofs of those theorems (such as in Chapters 2, 5 or 6) give a hint on how to find those roots. They do not necessarily help us to find those roots in an efficient way.

In this aspect, nonlinear equation solving is radically different from the subject of linear equation solving, where algorithms have running time typically bounded by a small degree polynomial on the input size. Here the number of roots is already exponential, and even finding one root can be a desperate task.

As in numerical linear algebra, nonlinear systems of equations may have solutions that are extremely sensitive to the value of the coefficients. Instances with such behavior are said to be *poorly conditioned*, and their 'hardness' is measured by an invariant known as the *condition number*. It is known that the condition number of random polynomial systems is small with high probability (See Chapter 8). Smale 17$^{\text{th}}$ problem was introduced in [78] as:

Open Problem 1.11 (Smale). *Can a zero of n complex polynomial equations in n unknowns be found approximately , on the average, in polynomial time with a uniform algorithm?*

The precise probability space referred in [78] is what we call $(\mathcal{H}_{\mathbf{d}}, \mathrm{d}\mathcal{H}_{\mathbf{d}})$ in Chapter 5. Zero means a zero in projective space $\mathbb{P}^n$, and the notion of approximate zero is discussed in Chapter 7. Polynomial time means that the running time of the algorithm should be bound by a polynomial in the input size, that we can take as $N = \dim \mathcal{H}_{\mathbf{d}}$. The precise model of computation will not be discussed in this book, and we refer to [20]. However, the algorithm should be **uniform** in the sense that the same algorithm should work for all inputs. The number n of variables and degrees $\mathbf{d} = (d_1, \dots, d_n)$ are part of the input.

Exercise 1.9. Show that $N = \sum_{i=1}^{n} \binom{d_i + n}{n}$. Conclude that there cannot exist an algorithm that approximates all the roots of a random homogeneous polynomial system in polynomial time.

Chapter 2

The Nullstellensatz

The study of polynomial equations motivated a huge and profound subject, *algebraic geometry*. This chapter covers some very basic and shallow algebraic geometry. Our point of view is closer to classical elimination theory rather than to modern commutative algebra. This does not replace a formal course in the subject.

Through this chapter, $\mathbf{k}$ denotes an algebraic closed field. The main example is $\mathbb{C}$. Custom and convenience mandate to state results in greater generality.

2.1 Sylvester's resultant

We start with a classical result of elimination theory. Let $\mathcal{P}_d$ denote the space of univariate polynomials of degree at most d, with coefficients in $\mathbf{k}$.

Theorem 2.1 (Sylvester's resultant)**.** *Let $f \in \mathcal{P}_d$ and $g \in \mathcal{P}_e$ for $d, e \in \mathbb{N}$. Assume that the higher coefficients f_d and g_e are not both*

Gregorio Malajovich, *Nonlinear equations*.
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

zero. The polynomials f and g have a common root if and only if the linear map $M_{f,g} : \mathcal{P}_{e-1} \times \mathcal{P}_{d-1} \rightarrow \mathcal{P}_{d+e-1}$ defined by

$$a, b \mapsto af + bg$$

is degenerate.

If we assimilate each $\mathcal{P}_d$ to $\mathbf{k}^{d+1}$ by associating each $a(x) = a_d x^d + \cdots + a_0$ to $[a_d, \cdots, a_0]^T \in \mathbf{k}^{d+1}$, the linear map $M_{f,g}$ corresponds to the *Sylvester matrix*

$$\mathrm{Syl}(f,g) = \left[\begin{array}{ccccccc}
f_d & & & & g_e & & \\
f_{d-1} & f_d & & & g_{e-1} & \ddots & \\
 & f_{d-1} & \ddots & & g_{e-2} & \ddots & g_e \\
\vdots & & \ddots & f_d & & \ddots & g_{e-1} \\
 & \vdots & & f_{d-1} & \vdots & & g_{e-2} \\
f_1 & & & & & & \\
f_0 & f_1 & & \vdots & g_1 & & \vdots \\
 & f_0 & \ddots & & g_0 & \ddots & \\
 & & \ddots & f_1 & & \ddots & g_1 \\
 & & & f_0 & & & g_0
\end{array}\right].$$

The *Sylvester resultant* is usually defined as

$$\mathrm{Res}_x(f(x), g(x)) \stackrel{\text{def}}{=} \det \mathrm{Syl}(f,g).$$

Proof of Theorem 2.1. Assume that $z \in \mathbf{k}$ is a common root for f and g. Then,

$$[z^{d+e} \ z^{d+e-1} \cdots z \ 1] \ \mathrm{Syl}(f,g) = a(z)f(z) + b(z)g(z) = 0.$$

Therefore the determinant of $\mathrm{Syl}(f,g)$ must vanish. Hence $M_{f,g}$ is degenerate.

Reciprocally, assume that $M_{f,g}$ is degenerate. Then there are $a \in \mathcal{P}_{e-1}, b \in \mathcal{P}_{d-1}$ so that $af + bg \equiv 0$. Assume for simplicity that $d \leq e$ and $g_e \neq 0$. By the Fundamental Theorem of Algebra, g admits e roots $z_1, \ldots, z_e$ (counted with multiplicity). By the pigeonhole principle, those cannot be all roots of a. Hence, at least one of them is also root of f.

If $g_e = 0$, the polynomial g may admit $r \geq 1$ roots at infinity. Hence the top r coefficients of bg will vanish, and the same for af. But f is monic, so the top r coefficients of a will vanish. We may proceed as before, with $g \in \mathcal{P}_{e-r}$ and $a \in \mathcal{P}_{e-r-1}$. □

As for complex projective space, we define $\mathbb{P}(\mathbf{k}^2)$ as the space of $\mathbf{k}$-lines through the origin.

Corollary 2.2. *Let* $\mathbf{k}$ *be an algebraic closed field. Two homogeneous polynomials* $f(x_0, x_1)$ *and* $g(x_0, x_1)$ *over* $\mathbf{k}$ *of respective degree* d *and* e *have a common zero on* $\mathbb{P}(\mathbf{k}^2)$ *if and only if*

$$\mathrm{Res}(f, g) \stackrel{def}{=} \mathrm{Res}_{x_1}(f(1, x_1), g(1, x_1)) = 0.$$

Corollary 2.3. *A polynomial* f *over an algebraic closed field is irreducible if and only if its* discriminant*, defined by*

$$\mathrm{Discr}_x(f(x)) \stackrel{def}{=} \mathrm{Res}_x(f(x), f'(x)),$$

vanishes.

(Convention: If f has degree exactly d, we assume that $f \in \mathcal{P}_d$ and compute the resultant accordingly).

Example 2.4. The following expressions should remind the reader about some familiar formulæ:

$$\begin{aligned}
\mathrm{Discr}_x(ax^2 + bx + c) &= a(4ac - b^2) \\
\mathrm{Discr}_x(ax^3 + bx + c) &= a^2(27ac^2 + 4b^3)
\end{aligned}$$

Exercise 2.1. Let $\mathbf{R} \subset \mathbf{S} \subset \mathbf{T} \subset \mathbf{k}$ be rings. Let $s \in \mathbf{S}$ be integral over $\mathbf{R}$, meaning that there is a monic polynomial $0 \neq f \in \mathbf{R}[x]$ with $f(s) = 0$. Let t be integral over $\mathbf{S}$. Show that t is integral over $\mathbf{R}$. (Hint: use Sylvester's resultant. Then open an algebra book, and compare its proof to your solution).

Exercise 2.2. Let x, y be integral over the ring $\mathbf{R}$. Show that $x + y$ is integral over $\mathbf{R}$.

Exercise 2.3. Same exercise for xy.

Exercise 2.4. Let s be integral over $\mathbf{R}$, show that there is $d \in \mathbb{N}$ such that every element of $\mathbf{S}$ can be represented uniquely by a degree d polynomial with coefficients in $\mathbf{R}$. What is d?

Remark 2.5. The same holds for algebraic extensions. Computer algebra systems represent algebraic integers or algebraic numbers through a primitive element s and the polynomial of Exercise 2.4. The primitive element is represented by its defining polynomial, and a numeric approximation that makes it unique.

2.2 Ideals

Let $\mathbf{R}$ be a ring (commutative, with unity and no divisors of zero). Recall from undergraduate algebra that an *ideal* in $\mathbf{R}$ is a subset $\mathfrak{J} \subseteq \mathbf{R}$ such that, for all $f, g \in \mathfrak{J}$ and all $u \in \mathbf{R}$,

$$f + g \in \mathfrak{J} \text{ and } uf \in \mathfrak{J}.$$

Let $\mathbf{R} = \mathbf{k}[x_1, \ldots, x_n]$ be the ring of n-variate polynomials over $\mathbf{k}$. Polynomial equations are elements of $\mathbf{R}$. Given $f_1, \ldots, f_s \in \mathbf{R}$, the ideal generated by them, denoted by $(f_1, \ldots, f_s)$, is the set of polynomials of the form

$$f_1 g_1 + \cdots + f_s g_s$$

where $g_j \in \mathbf{R}$. Every ideal of polynomials is of this form.

Theorem 2.6 (Hilbert's basis Theorem)**.** *Let* $\mathbf{k}$ *be a field. Then any ideal* $\mathfrak{J} \subseteq \mathbf{k}[x_1, \ldots, x_n]$ *is finitely generated.*

The following consequence is immediate, settling a point left open in Chapter 1:

Corollary 2.7. *The arbitrary intersection of Zariski closed sets is Zariski closed. Hence, the set of Zariski open sets constitutes a topology.*

Before proving Theorem 2.6, we need a preliminary result. The set $(\mathbb{Z}_{\geq 0})^n$ can be well-ordered lexicographically. When $n = 1$, set $\mathbf{a} \prec \mathbf{b}$ if and only if $\mathbf{a} < \mathbf{b}$. Inductively, $\mathbf{a} \prec \mathbf{b}$ if and only if

$$a_1 < b_1$$

or

$$a_1 = b_1 \quad \text{and} \quad (a_2, \dots, a_n) \prec (b_2, \dots, b_n).$$

Note that $\mathbf{0} \preceq \mathbf{a}$ for all $\mathbf{a}$.

Given $f = \sum_{\mathbf{a} \in A} f_{\mathbf{a}} \mathbf{x}^{\mathbf{a}} \in k[x_1, \dots, x_n]$, its *leading term* (with respect to the $\prec$ ordering) is the non-zero monomial $f_{\mathbf{a}} x^{\mathbf{a}}$ such that $\mathbf{a}$ is maximal with respect to $\prec$.

We will also say that $\mathbf{a} \leq \mathbf{b}$ if and only if $a_i \leq b_i$ for all i. The ordering $\leq$ is a partial ordering, and $\mathbf{a} \leq \mathbf{b}$ implies $\mathbf{a} \preceq \mathbf{b}$.

The long division algorithm applies as follows: if f and g have leading terms $f_{\mathbf{a}} \mathbf{x}^{\mathbf{a}}$ and $f_{\mathbf{b}} \mathbf{x}^{\mathbf{b}}$ respectively, and $\mathbf{b} \leq \mathbf{a}$ then there are q, r with leading terms $\frac{f_{\mathbf{a}}}{f_{\mathbf{b}}} \mathbf{x}^{\mathbf{a}-\mathbf{b}}$ and $r_{\mathbf{c}} \mathbf{x}^{\mathbf{c}}$ such that $f = qg + r$ and $\neg(\mathbf{b} \leq \mathbf{c})$. In particular $\mathbf{c} \prec \mathbf{a}$.

Theorem 2.6 follows from the following fact.

Lemma 2.8 (Dickson)**.** *Let $\mathbf{a}_i$ be a sequence in $(\mathbb{Z}_{\geq 0})^n$, such that*

$$i < j \Rightarrow \neg\, \mathbf{a}_i \leq \mathbf{a}_j. \tag{2.1}$$

Then this sequence is finite.

Proof. The case $n = 1$ is trivial, for the sequence is strictly decreasing.

Assume that in dimension n, there is an infinite sequence $\mathbf{a}_i$ satisfying (2.1). Then there is an infinite subsequence $\mathbf{a}_{i_j}$, with last coordinate $a_{i_j n}$ non-decreasing We set $\mathbf{b}_j = (a_{i_j 1}, \dots, a_{i_j n-1})$. The sequence $\mathbf{b}_j$ satisfies (2.1). Hence by induction it should be finite. □

Proof of Theorem 2.6. Let $f_1 \in \mathfrak{J}$ be the polynomial with minimal leading term. As it is defined up to a multiplicative constant in $\mathbf{k}$, we take it monic. Inductively, choose f_j as the monic polynomial with minimal leading term in $\mathfrak{J}$ that does not belong to $(f_1, \dots, f_{j-1})$. We claim this process is finite.

Let $\mathbf{x}^{\mathbf{a}_i}$ be the leading term of f_i. The long division algorithm implies that, for $i < j$, we cannot have $\mathbf{a}_i \leq \mathbf{a}_j$ or f_j would not be minimal.

By Dickson's Lemma, the sequence $\mathbf{a}_i$ is finite. □

Remark 2.9. The basis we obtained is a particular example of a *Gröbner basis* for the ideal $\mathfrak{J}$. In general, $\prec$ can be any well-ordering of $(\mathbb{Z}_{\geq 0})^n$ such that $\mathbf{a} \prec \mathbf{b} \Rightarrow \mathbf{a} + \mathbf{c} \prec \mathbf{b} + \mathbf{c}$. (When comparing monomials, this is called a *monomial ordering*). A Gröbner basis for $\mathfrak{J}$ is a finite set $(f_1, \ldots, f_s) \in \mathfrak{J}$ such that for any $g \in \mathfrak{J}$, the leading term of g is divisible by the leading term of some f_i. In particular, $\mathfrak{J} = (f_1, \ldots, f_s)$. It is possible to use Gröbner basis representation to answer many questions about ideals, see [27]. Since no complexity results are known, those should be considered as a *method* for specific tasks rather than a reliable algorithm. Modern elimination algorithms are available, see for instance [43] for algebraic geometry based elimination, and [39] for fast linear algebra based elimination. A numerical algorithm is given in chapter 10. References for practical numerical applications are, for instance, [80] and of course [53] and [54].

2.3 The coordinate ring

Let $X \subseteq \mathbf{k}^n$ be a Zariski closed set, and denote by $I(X)$ the ideal of polynomials vanishing on all of X.

Example 2.10. Let $X = \{\mathbf{a}\}$. Then $I(X)$ is $(x_1 - a_1, \ldots, x_n - a_n)$.

Polynomials in $\mathbf{k}[x_1, \ldots, x_n]$ restrict to functions of X. Two of those functions are equal on X if and only if they differ by some element of $I(X)$.

This leads us to study the *coordinate ring* $\mathbf{k}[x_1, \ldots, x_n]/I(X)$ of X, or more generally the quotient of $\mathbf{k}[x_1, \ldots, x_n]$ by an arbitrary ideal $\mathfrak{J}$.

Note that we can look at $\mathbf{A} = k[x_1, \ldots, x_n]/\mathfrak{J}$ as a ring or as an algebra, whatever is more convenient. We start by the simplest case, namely the ring of coordinates of a hypersurface in 'normal form':

Proposition 2.11. *Assume that* $f \in \mathbf{k}[x_1, \ldots, x_n]$ *is of the form* $f(\mathbf{x}) = x_n^d + f_1(x_1, \ldots, x_n)$ *and no monomial of* f_1 *has degree* $\geq d$ *in* x_n*. Let* $\mathbf{A} = \mathbf{k}[x_1, \ldots, x_n]/(f)$ *and* $\mathbf{R} = \mathbf{k}[x_1, \ldots, x_{n-1}]$*. Then,*

1. $\mathbf{A}$ *is a finite integral extension of* $\mathbf{R}$ *of degree* d*.*

2. $\mathbf{A} = \mathbf{R}[h]$ *where* $h = x_n + (f)$*.*

3. *The projection* $\pi : \mathbf{k}^n \to \mathbf{k}^{n-1}$ *onto the first* $n-1$ *coordinates maps the zero-set of* f *onto* $\mathbf{k}^{n-1}$.

4. *The point* $(x_1, \dots, x_{n-1})$ *has exactly* d *distinct preimages by* π *in the zero-set of* f *if and only if*

$$\mathrm{Discr}_{x_n} f(x_1, \dots, x_{n-1}, x_n) \neq 0.$$

   The notation above stands for the discriminant with respect to x_n*, the other variables treated as parameters.*

5. *In case* f *is irreducible, the condition of item 4 holds for* $\mathbf{x} = (x_1, \dots, x_{n-1})$ *in a non-empty Zariski open set.*

Proof. 1 and 2: The homomorphism $i : \mathbf{R} \to \mathbf{A}$ given by $i(g) = g + (f)$ has trivial kernel, making $\mathbf{R}$ a subring of $\mathbf{A}$.

We need to prove now that for any $a \in \mathbf{A}$, there are $g_0, \dots, g_{d-1} \in \mathbf{R}$ such that

$$a^d + g_{d-1} a^{d-1} + \cdots + g_0 \equiv 0. \tag{2.2}$$

For any $\mathbf{y} = (y_1, \dots, y_{n-1}) \in \mathbf{k}^{n-1}$, define

$$g_j(\mathbf{y}) = (-1)^j \ \sigma_{d-j} \left(a(\mathbf{y}, t_1), \dots, a(\mathbf{y}, t_d) \right) \tag{2.3}$$

where σ_j is the j-th symmetric function and $t_1, \dots, t_d$ are the roots (with multiplicity) of the polynomial $t \mapsto f(\mathbf{y}, t) = 0$.

The right-hand-side of (2.3) is a polynomial in $\mathbf{y}, t_1, \dots, t_d$. It is symmetric in $t_1, \dots, t_d$ hence it depends only on the coefficients with respect to t of the polynomial $t \mapsto f(\mathbf{y}, t)$. Those are polynomials in $\mathbf{y}$, whence g_j is a polynomial in $\mathbf{y}$.

Once we fixed an arbitrary value for $\mathbf{y}$, (2.2) specializes to

$$\prod_{j=1}^{d} a(\mathbf{y}, t) - a(\mathbf{y}, t_j)$$

and therefore vanishes uniformly on the zero-set of f.

We need to prove that $\mathbf{A}$ has degree exactly d over $\mathbf{R}$. Since $\mathbf{k}[x_1, \dots, x_n] = \mathbf{R}[x_n]$, the coset $h = x_n + (f)$ of x_n is a primitive element for $\mathbf{A}$.

It cannot have a degree smaller than d, for otherwise there would be $e < d$, $\alpha \in k$ and $G_0, \ldots G_{e-1} \in \mathbf{R}$ with

$$x_n^e + G_{e-1}(\mathbf{y})x_n^{e-1} + \cdots + G_0(\mathbf{y}) = \alpha f(\mathbf{y}, x_n).$$

To see this is impossible, just specialize $\mathbf{y} = 0$.

3: Fix an arbitrary $\mathbf{y}$ in k^{n-1} and solve $f(y_1, \cdots, y_{n-1}, x) = x^d + f_1(y_1, \ldots, y_{n-1}, x)$.

4: this is just Corollary 2.3.

5: In case f is irreducible, the discriminant in item 4 is not uniformly zero. Hence in this case, for $x_1, \ldots, x_{n-1}$ generic (in a Zariski-open set), there are d possible distinct values of x_n for $f(\mathbf{x}) = 0$. □

The result above gives us a pretty good description of of hypersurfaces in special position. Geometrically, we may say that when f is irreducible, *a generic 'vertical' line intersect the hypersurface in exactly d distinct points.* Moreover, generic n-variate polynomials are irreducible when $n \geq 2$.

2.4 Group action and normalization

The special position hypothesis $f(\mathbf{x}) = x_n^d$+(low order terms) is quite restrictive, and can be removed by a change of coordinates.

Recall that a group $\mathcal{G}$ *acts ('on the left') on a set S* if there is a function $a : \mathcal{G} \times S \to S$ such that $a(gh, s) = a(g, a(h, s))$ and $a(1, s) = s$. This makes $\mathcal{G}$ into a subset of invertible mappings of S. When S is a linear space, the linear group of S (denoted by $GL(S)$) is the group of invertible linear maps.

We consider changes of coordinates in linear space $\mathbf{k}^n$ that are elements of the group $GL(\mathbf{k}^n)$ of invertible linear transformations of $\mathbf{k}^n$. This action induces a left-action on $\mathbf{k}[x_1, \ldots, x_n]$, so that $(f \circ L^{-1})(L(\mathbf{x})) = f(\mathbf{x})$. If $L \in GL(k^n)$, we summarize those actions as

$$\mathbf{x} \overset{L}{\rightsquigarrow} a(L, \mathbf{x}) \overset{\text{def}}{=} L(\mathbf{x}) \qquad \text{and} \qquad f \overset{L}{\rightsquigarrow} f \circ L^{-1}.$$

This action extends to ideals and quotient rings,

$$\mathfrak{J} \overset{L}{\rightsquigarrow} \mathfrak{J}^L \overset{\text{def}}{=} \{f \circ L^{-1} : f \in \mathfrak{J}\}$$

and

$$\mathbf{A} = \mathbf{k}[x_1, \dots, x_n]/\mathfrak{J} \overset{L}{\rightsquigarrow} \mathbf{A}^L \overset{\text{def}}{=} \mathbf{k}[x_1, \dots, x_n]/\mathfrak{J}^L.$$

Lemma 2.12. *Let $\mathbf{A} = \mathbf{k}[x_1, \dots, x_n]/\mathfrak{J}$ and let $\mathbf{R}$ be a subring of $\mathbf{k}[x_1, \dots, x_n]$. Let $L \in GL(\mathbf{k}^n)$. Then, $\mathbf{A}$ is an integral extension of $\mathbf{R}$ of degree d if and only if $\mathbf{A}^L$ is an integral extension of $\mathbf{R}^L$ of degree d. If $\mathbf{A} = \mathbf{R}[h]$, then $\mathbf{A}^L = \mathbf{R}^L[h \circ L^{-1}]$.*

Proof. Let $h \in \mathbf{A}$ be the primitive element with respect to $\mathbf{R}$:

$$h^d + g_{d-1}h^{d-1} + \cdots + g_0 = 0_{\mathbf{A}}.$$

Then

$$(h \circ L^{-1})^d + (g_{d-1} \circ L^{-1})(h \circ L^{-1})^{d-1} + \cdots + g_0 \circ L^{-1} = 0_{\mathbf{A}}$$

and $h^L = h \circ L^{-1}$ is a primitive element of $\mathbf{A}^L$ over $\mathbf{R}^L$. The same works in the opposite direction. $\square$

We say that a sub-group $\mathcal{G}$ of $GL(\mathbf{k}^n)$ acts *transitively* on $\mathbf{k}^n$ if and only if, for all pairs $\mathbf{x}, \mathbf{y} \in \mathbf{k}^n$, there is $G \in \mathcal{G}$ with $\mathbf{y} = G\mathbf{x} (= a(G, \mathbf{x}))$.

Example 2.13. The unitary group $U(\mathbb{C}^n) = \{Q \in GL(\mathbb{C}^n) : Q^*Q = I\}$ acts transitively on the unit sphere $\|\mathbf{z}\| = 1$ of $\mathbb{C}^n$. The 'conformal' group $U(\mathbb{C}^n) \times \mathbb{C}_\times$ acts transitively on $\mathbb{C}^n$.

We restate Proposition 2.11, so we have a description of the ring of coordinates for an arbitrary hypersurface. A *generic* element of $\mathcal{G} \subseteq GL(\mathbf{k}^n)$ means an element of a non-empty set of the form $U \cap \mathcal{G}$, where U is Zariski-open in $\mathbf{k}^{n^2}$.

Proposition 2.14. *Let $\mathbf{k}$ be an algebraically closed field. Let $f \in k[x_1, \dots, x_n]$ have degree d. Let $\mathbf{A} = \mathbf{k}[x_1, \dots, x_n]/(f)$. Then,*

1. *The ring $\mathbf{A}$ is a finite integral extension of $\mathbf{R}$ of degree d, where $\mathbf{R}$ is isomorphic to $\simeq \mathbf{k}[y_1, \dots, y_{n-1}]$.*
2. *Let $\mathcal{G} \subseteq GL(\mathbf{k}^n)$ act transitively on $\mathbf{k}^n$. For L generic in $\mathcal{G}$, item 1 holds for the linear forms y_j in the variables x_j given by $x_i = \sum_{j=1}^n L_{ij} y_j$. Then, $\mathbf{k}[y_1, \dots, y_n] = \mathbf{k}[x_1, \dots, x_n]^L$ and $\mathbf{A} = \mathbf{R}[h]$ where $h = y_n + (f \circ L)$.*

3. *Let E the hyperplane $y_n = 0$. The canonical projection $\pi : \mathbf{k}^n \rightarrow E$ maps the zero-set of (f) onto E.*
4. *Furthermore, $(y_1, \ldots, y_{n-1})$ has exactly d preimages by π in the zero-set of f if and only if*

$$\mathrm{Discr}_{y_n} f(y_1, \ldots, y_{n-1}, y_n) \neq 0$$

.

Again, when f is irreducible, for L in a Zariski-open set, the polynomial in item 5 is not uniformly zero. Hence, we may say that for f irreducible, a *generic line* intersects the zero-set of f in exactly d points.

Proof of Proposition 2.14. The coefficient of y_n^d in $(f \circ L)(\mathbf{y})$ is a polynomial in the coefficients of L. We will show that this polynomial is not uniformly zero. Then, for generic L, it suffices to multiply f by a non-zero constant to recover the situation of Proposition 2.11. The other items of this Proposition follow immediately.

Let $f = F_0 + \cdots + F_d$ where each F_i is homogeneous of degree d.

The field $\mathbf{k}$ is algebraically closed, hence infinite, so there are $\alpha_1, \cdots, \alpha_{d-1}$ so that $F_d(\alpha_1, \cdots, \alpha_{d-1}, 1) \neq 0$. Then there is $L \in \mathcal{G}$ that takes $\mathbf{e}_n$ into $c[\alpha_1, \cdots, \alpha_{n-1}, 1]$ for $c \neq 0$.

Then up to a non-zero multiplicative constant,

$$f \circ L = x_n^d + (\text{low order terms in } x_n)$$

□

We may extend the construction above to quotient by arbitrary ideals. Let $\mathfrak{J}$ be an ideal in $\mathbf{k}[x_1, \ldots, x_n]$. Then the quotient $\mathbf{A} = \mathbf{k}[x_1, \ldots, x_n]/\mathfrak{J}$ is finitely generated. (For instance, by the cosets $x_i + \mathfrak{J}$).

We say that an ideal $\mathfrak{p}$ of a ring $\mathbf{R}$ is *prime* if and only if, for all $f, g \in \mathbf{R}$ with $fg \in \mathfrak{p}$, $f \in \mathfrak{p}$ or $g \in \mathfrak{p}$.

Given an ideal $\mathfrak{J}$, let $Z(\mathfrak{J}) = \{\mathbf{x} \in k^n : f(\mathbf{x}) = 0 \forall f \in \mathfrak{J}\}$ denote its zero-set.

Lemma 2.15 (Noether's normalization)**.** *Let $\mathbf{k}$ be an algebraically closed field, and let $\mathbf{A} \neq \{0\}$ be a finitely generated $\mathbf{k}$-algebra. Then:*

1. *There are $y_1, \dots, y_r \in \mathbf{A}$, $r \geq 0$, algebraically independent over $\mathbf{k}$, such that $\mathbf{A}$ is integral over $\mathbf{k}[y_1, \dots, y_r]$.*
2. *Assume that $\mathbf{A} = \mathbf{k}[x_1, \dots, x_n]/\mathfrak{J}$. Let $\mathcal{G} \subseteq GL(k^n)$ act transitively on $\mathbf{k}^n$. Then for L generic in $\mathcal{G}$, item 1 holds for the linear forms y_j in the variables x_j, given by $x_i = \sum_{j=1}^n L_{ij} y_j$. Furthermore, $\mathbf{k}[y_1, \dots, y_n] = \mathbf{k}[x_1, \dots, x_n]^L$ and $\mathbf{A} = \mathbf{R}[h_{r+1}, \dots, h_n]$ where $h_j = y_j + \mathfrak{J}^{L^{-1}}$.*
3. *Let E the linear space $y_{r+1} = \cdots = y_n = 0$. The canonical projection $\pi : \mathbf{k}^n \to E$ maps the zero-set of $\mathfrak{J}$ onto E.*
4. *If $\mathfrak{J}$ is prime, then for L generic, the set of points of E with $d = [\mathbf{A} : \mathbf{R}]$ distinct preimages by π is Zariski-open.*

In other words, when $\mathfrak{J}$ is prime, a generic affine space of the complementary dimension intersects $Z(\mathfrak{J})$ in exactly d distinct points.

Remark 2.16. Effective versions of Lemma 2.15 play a foundamental rôle in modern elimination theory, see for instance [41] and references.

Proof of Lemma 2.15. Let $y_1, \dots, y_n$ generate $\mathbf{A}$ over $\mathbf{k}$. We renumber the y_i, so that $y_1, \dots, y_r$ are algebraically independent over $\mathbf{k}$ and each y_j, $r < j \leq n$, is algebraic over $\mathbf{k}[y_1, \dots, y_{j-1}]$. Proposition 2.14 says that y_j is integral over $\mathbf{k}[y_1, \dots, y_{j-1}]$. From Exercise 2.4, it follows by induction that $\mathbf{k}[y_1, \dots, y_n]$ is integral over $\mathbf{k}[y_1, \dots, y_r]$.

For the second item, choose as generators the cosets $y_1 + \mathfrak{J}, \cdots, y_n + \mathfrak{J}$. After reordering, the first item tells us that there are polynomials $f_{r+1}, \dots, f_n$ with

$$f_j(y_1, \dots, y_j) \in \mathfrak{J}.$$

and $\mathfrak{J} = (f_j, \dots, f_n)$. Moreover, if $\mathfrak{J}$ is prime then we can take $f_1, \dots, f_n$ irreducible. The projection π into the r first coordinates maps the zero-set set of $\mathfrak{J}$ into $\mathbf{k}^r$. It is onto, because fixing the values of $y_1, \dots, y_r$, one can solve successively for $y_{r+1}, \dots, y_n$. □

Lemma 2.17. *Let $\mathbf{A} = \mathbf{k}[x_1, \dots, x_n]/\mathfrak{J}$. Then $\mathbf{A}$ is finite dimensional as a vector space over $\mathbf{k}$ if and only if $Z(\mathfrak{J})$ is finite.*

Proof. Both conditions are equivalent to $r = 0$ in Lemma 2.15. □

In this situation, $\#Z(\mathfrak{J})$ is not larger than the degree of A with respect to k.

Example 2.18. $n = 1$, $\mathfrak{J} = (x^2)$. In this case $\mathbf{A} = \mathbf{k}^2$ so $r = 2$. Note however that $\#Z(\mathfrak{J}) = 1$.

However, if we require $\mathfrak{J}$ to be prime, the number of zeros is precisely the degree $[\mathbf{A} : \mathbf{k}]$. The same principle holds for $\mathfrak{J} = (f_1, \dots, f_n)$ for *generic* polynomials. We can prove now a version of Bézout's theorem:

Theorem 2.19 (Bézout's Theorem, generic case)**.** *Let $d_1, \dots, d_n \geq 1$. Let $B = d_1 d_2 \cdots d_n$. Then generically, $\mathbf{f} \in \mathcal{P}_{d_1} \times \cdots \times \mathcal{P}_{d_n}$ has B isolated zeros in $\mathbf{k}^n$.*

Proof. Let $\mathfrak{J}_r = (f_{r+1}, \dots, f_n)$ and $\mathbf{A}_r = \mathbf{k}[x_1, \dots, x_n]/\mathfrak{J}_r$.

Our induction hypothesis (in $n - r$) is:

$$[\mathbf{A}_r : \mathbf{k}[x_1, \dots, x_{r-1}]] = d_{r+1} d_{r+2} \dots d_n$$

When $r = n$, this is Proposition 2.11.

For $r < n$, $\mathbf{A}_r$ is integral of degree d_r over $\mathbf{A}_{r+1}$. The integral equation (in x_r) is, up to a multiplicative factor,

$$f_r(x_1, \dots, x_r, y_{r+1}, \dots, y_n) = 0$$

where $y_{r+1}, \dots, y_n$ are elements of $\mathbf{A}_{r+1}$ (hence constants). Hence,

$$[\mathbf{A} : \mathbf{k}] = d_1 d_2 \cdots d_n.$$

$\square$

Noether normalization provides information about the ring $\mathbf{R} = \mathbf{k}[x_1, \dots, x_n]$.

Definition 2.20. A ring $\mathbf{R}$ is *noetherian* if and only if, there cannot be an infinite ascending chain $\mathfrak{J}_1 \subsetneq \mathfrak{J}_2 \subsetneq \cdots$ of ideals in $\mathbf{R}$.

Theorem 2.21. *Let $\mathbf{k}$ be algebraically closed. Then $\mathbf{R} = \mathbf{k}[x_1, \dots, x_n]$ is Noetherian.*

Proof. Let $\mathbf{A}_i = \mathbf{R}/\mathfrak{J}_i$. Then $\mathbf{A}_1 \supsetneq \mathbf{A}_2 \supsetneq \cdots$. However, since $\mathbf{A}_i \neq \mathbf{A}_{i+1}$, they cannot have the same transcendence degree r and the same degree over $\mathbf{k}[y_1, \ldots, y_r]$. Therefore at least one of those quantities decreases, and the chain must be finite. □

Exercise 2.5. Consider the ideal $\mathfrak{J} = (x_2^2 - x_2, x_1 x_2)$. Describe the algebra $\mathbf{A} = \mathbf{k}[x_1, x_2]/\mathfrak{J}$.

2.5 Irreducibility

A Zariski closed set X is irreducible if and only if it cannot be written in the form $X = X_1 \cup X_2$, with both X_1 and X_2 Zariski closed, and $X \neq X_1$, $X \neq X_2$.

Recall that an ideal $\mathfrak{p} \subset \mathbf{R}$ is *prime* if for any $f, g \in \mathfrak{p}$, whenever $fg \in \mathfrak{p}$ we have $f \in \mathfrak{p}$ or $g \in \mathfrak{p}$.

Lemma 2.22. *X is irreducible if and only if $I(X)$ is prime.*

Proof. Assume that X is irreducible, and $fg \in I(X)$. Suppose that $f, g \notin I(X)$. Then set $X_1 = X \cap Z(f)$ and $X_2 = X \cap Z(g)$, contradiction.

Now, assume that X is the union of X_1 and X_2, with $X_1 \neq X$ and $X_2 \neq X$. Then, there are $f \in I(X_1), f \notin I(X)$ and $g \in I(X_2), g \notin I(X)$. So neither f or g belong to $I(X)$. However, fg vanishes for all X. □

Now we move to general ideals. The definition is analogous. An ideal $\mathfrak{J}$ is said to be *irreducible* if it cannot be written as $\mathfrak{J} = \mathfrak{J}_1 \cap \mathfrak{J}_2$ with $\mathfrak{J} \neq \mathfrak{J}_1$ and $\mathfrak{J} \neq \mathfrak{J}_2$. At this time, we can say more that in the case of closed sets:

Lemma 2.23. *In a Noetherian ring R, every ideal $\mathfrak{J}$ is the intersection of finitely many irreducible ideals.*

Proof. Assume that the Lemma is false. Let $\mathbf{J}$ be the set of ideals of R that are not the intersection of finitely many irreducible ideals.

Assume by contradiction that $\mathbf{J}$ is not empty. By the Noetherian condition, there cannot be an infinite chain

$$\mathfrak{J}_1 \subsetneq \mathfrak{J}_2 \subsetneq \cdots$$

of ideals in $\mathbf{J}$. Therefore, there must be an element $\mathfrak{J} \in \mathbf{J}$ that is maximal with respect to the inclusion.

But $\mathfrak{J}$ is not irreducible itself, so there are $\mathfrak{J}_1$, $\mathfrak{J}_2$ with $\mathfrak{J} = \mathfrak{J}_1 \cap \mathfrak{J}_2$, $\mathfrak{J} \neq \mathfrak{J}_1$, $\mathfrak{J} \neq \mathfrak{J}_2$.

If $\mathfrak{J}_1$ and $\mathfrak{J}_2$ are intersections of finitely many irreducible ideals, then so does $\mathfrak{J} = \mathfrak{J}_1 \cap \mathfrak{J}_2$ and hence $\mathfrak{J} \notin \mathbf{J}$, contradiction. If however one of them (say $\mathfrak{J}_1$) is not the intersection of finitely many irreducible ideals, then $\mathfrak{J} \subseteq \mathfrak{J}_1$ with $\mathfrak{J}_1$ in $\mathbf{J}$. Then $\mathfrak{J}$ is not maximal with respect to the inclusion, contradicting the definition.

Thus, $\mathbf{J}$ must be empty. □

An ideal $\mathfrak{p}$ in R is *primary* if and only if, for any $x, y \in R$,

$$xy \in \mathfrak{p} \Longrightarrow x \in \mathfrak{p} \text{ or } \exists n \in \mathbb{N} : y^n \in \mathfrak{p}$$

For instance, $(4) \subset \mathbb{Z}$ and $(x^2) \subset k[x]$ are primary ideals, but (12) is not. Prime ideals are primary but the converse is not always true. The reader will show a famous theorem:

Theorem 2.24 (Primary Decomposition Theorem)**.** *If* $\mathbf{R}$ *is Noetherian, then every ideal in* $\mathbf{R}$ *is the intersection of finitely many primary ideals.*

Exercise 2.6. Let $\mathbf{R}$ be Noetherian. Assume the zero ideal is irreducible. Show then that the zero ideal $(0) = \{0\}$ is primary. Hint: assume that $xy = 0$ with $x \neq 0$. Set $\mathfrak{J}_n = \{z : zy^n = 0\}$. Using Noether's condition, show that there is n such that $y^n = 0$.

Exercise 2.7. Let $\mathfrak{J}$ be irreducible in $\mathbf{R}$. Show that the zero ideal in $\mathbf{R}/\mathfrak{J}$ is irreducible.

Exercise 2.8. Let $\mathfrak{J}$ be and ideal of $\mathbf{R}$, such that $\mathbf{R}/\mathfrak{J}$ is primary. Show that $\mathfrak{J}$ is primary. This finishes the proof of Theorem 2.24

2.6 The Nullstellensatz

To each subset $X \subseteq \mathbf{k}^n$, we associated the ideal of polynomials vanishing in X:

$$I(X) = \{f \in \mathbf{k}[x_1, \dots, x_n] : \forall \mathbf{x} \in X, f(\mathbf{x}) = 0\}.$$

To each ideal $\mathfrak{J}$ of polynomials, we associated its zero set

$$Z(\mathfrak{J}) = \{\mathbf{x} \in \mathbf{k}^n : \forall f \in \mathfrak{J}, f(\mathbf{x}) = 0\}.$$

Those two operators are inclusion reversing:

$$\text{If } X \subseteq Y \text{ then } I(Y) \subseteq I(X).$$

$$\text{If } \mathfrak{J} \subseteq \mathfrak{K} \text{ then } Z(\mathfrak{K}) \subseteq Z(\mathfrak{J}).$$

Hence, compositions $Z \circ I$ and $I \circ Z$ are inclusion preserving:

$$\text{If } X \subseteq Y \text{ then } (Z \circ I)(X) \subseteq (Z \circ I)(Y).$$

$$\text{If } \mathfrak{J} \subseteq \mathfrak{K} \text{ then } (I \circ Z)(\mathfrak{J}) \subseteq (I \circ Z)(\mathfrak{K}).$$

By construction, compositions are nondecreasing:

$$X \subseteq (Z \circ I)(X) \text{ and } \mathfrak{J} \subseteq (I \circ Z)(\mathfrak{J}).$$

The operation $Z \circ I$ is called Zariski closure. It has the following property. Suppose that X is Zariski closed, that is $X = Z(\mathfrak{J})$ for some $\mathfrak{J}$. Then

$$(Z \circ I)(X) = X.$$

Indeed, assume that $x \in (Z \circ I)(X)$. Then for all $f \in I(X)$, $f(x) = 0$. In particular, this holds for $f \in \mathfrak{J}$. Thus $x \in X$.

The opposite is also true. Suppose that $\mathfrak{J} = I(X)$. We claim that

$$I(Z(\mathfrak{J})) = \mathfrak{J}.$$

Indeed, let $f \in I(Z(\mathfrak{J}))$. This means that f vanishes in all of $Z(\mathfrak{J})$. In particular it vanishes in $X \subseteq Z(\mathfrak{J})$. So $f \in \mathfrak{J} = I(X)$.

The operation $I \circ Z$ is akin to the closure of a set, but more subtle.

Example 2.25. Let $n = 1$ and $a \in \mathbf{k}$. Let $\mathfrak{J} = ((x-a)^3)$ be the ideal of polynomials vanishing at a with multiplicity ≥ 3. Then, $Z(\mathfrak{J}) = \{a\}$ and $I(Z(\mathfrak{J})) = ((x-a))$ the polynomials vanishing at a (no multiplicity assumed).

In general, the *radical* of an ideal $\mathfrak{J}$ is defined as

$$\sqrt{\mathfrak{J}} = \{f \in \mathbf{k}[x_1, \dots, x_n] : \exists r \in \mathbb{N}, f^r \in \mathfrak{J}\}.$$

The reader shall check as an exercise that $\sqrt{\mathfrak{J}}$ is an ideal.

Theorem 2.26 (Hilbert Nullstellensatz)**.** *Let* $\mathbf{k}$ *be an algebraically closed field. Then, for all ideal* $\mathfrak{J}$ *in* $\mathbf{k}[x_1, \dots, x_n]$*,*

$$I(Z(\mathfrak{J})) = \sqrt{\mathfrak{J}}.$$

We will derive this theorem from a weaker version.

Theorem 2.27 (weak Nullstellensatz)**.** *Assume that* $f_1, \dots, f_s \in \mathbf{k}[x_1, \dots, x_n]$ *have no common root. Then, there are* $g_1, \dots, g_s \in \mathbf{k}[x_1, \dots, x_n]$ *such that*

$$f_1 g_1 + \cdots + f_s g_s \equiv 1.$$

Proof. Let $\mathfrak{J} = (f_1, \cdots, f_s)$ and assume that $1 \not\in \mathfrak{J}$. In that case, the algebra

$$A = \mathbf{k}[x_1, \dots, x_n]/\mathfrak{J}$$

is not the zero algebra. By Lemma 2.15, there is a surjective projection from the zero-set of $\mathfrak{J}$ onto some r-dimensional subspace of $\mathbf{k}^n$, $r \geq 0$. Thus the f_i have a common root. □

Proof of Theorem 2.26(Hilbert Nullstellensatz).
The inclusion $I(Z(\mathfrak{J})) \supseteq \sqrt{\mathfrak{J}}$ is easy, so let $h \in I(Z(\mathfrak{J}))$.

Let $(f_1, \dots, f_s)$ be a basis of $\mathfrak{J}$ (Theorem 2.6). Assume that $(f_1, \dots, f_s) \not\ni 1$ (or else $h \in \mathfrak{J} \subseteq \sqrt{\mathfrak{J}}$ and we are done).

Consider now the ideal $\mathfrak{K} = (f_1, \dots, f_s, (1 - x_{n+1}h)) \in \mathbf{k}[x_1, \dots, x_{n+1}]$. The set $Z(\mathfrak{K})$ is empty. Otherwise, there would be $(x_1, \dots, x_{n+1}) \in \mathbf{k}^{n+1}$ so that $f_i(x_1, \dots, x_n)$ would vanish for all i. But then by hypothesis $h(x_1, \dots, x_n) = 0$ and $1 - x_{n+1}h \neq 0$.

By the weak Nullstellensatz (Theorem 2.27), $1 \in \mathfrak{K}$. Thus, there are polynomials $G_1, \dots, G_{n+1}$ with

$$1 = f_1 G_1 + \cdots + f_n G_n + (1 - x_{n+1}h) G_{n+1}$$

Specializing $x_{n+1} = 1/h$ and clearing denominators, we get

$$h^r = f_1 g_1 + \cdots + f_n g_n$$

for

$$g_i(x_1, \dots, x_n) = h(x_1, \dots, x_n)^r G_i(x_1, \dots, x_n, 1/h(x_1, \dots, x_n))$$

and r the maximal degree of the g_i's in the variable x_n. $\square$

The Nullstellensatz is is rich in consequences, and we should discuss some of them.

Suppose that a bound for the degree of the g_i is available in function of the degree of the f_i. One can solve the system $f_1(\mathbf{x}) = \cdots = f_n(\mathbf{x})$ by setting $f_{n+1}(\mathbf{x}) = 1 - \langle \mathbf{u}, \mathbf{x} \rangle$, where v and the coordinates of $\mathbf{u}$ will be treated as parameters. $\mathbf{x}$ is a common root for $f_1, \dots, f_n$ if and only if there is $\mathbf{u}, v$ such that $\mathbf{x}$ is a common root of $f_1, \dots, f_{n+1}$.

This means in particular that the operator

$$M(\mathbf{u}, v) : g_1, \cdots, g_{n+1} \mapsto f_1 g_1 + \cdots + f_{n+1} g_{n+1}$$

is not surjective. Using the available bound on the degree of the g_i, this means that the subdeterminants of the matrix associated to M vanish. This matrix has coordinates that may be zero, coefficients of $f_1, \dots, f_n$, or coordinates of $\mathbf{u}$, or v.

By fixing a generic value for $\mathbf{u}$, those determinants become polynomials in v. Their solutions can be used to eliminate one of the variables $x_1, \dots, x_n$.

Finding bounds for the degree of the g_i in function of the degree of the f_i became an active and competitive subject since the pioneering paper by Brownawell [24]. See [3, 51] and references for more recent developments.

Now we move to other applications of the Nullstellensatz. An ideal $\mathfrak{m}$ over a ring R is *maximal* if and only if, $\mathfrak{m} \neq R$ and for all ideal J with $\mathfrak{m} \subseteq J \subseteq R$, either $J = \mathfrak{m}$ or $J = R$.

Example 2.28. For every $\mathbf{a} \in \mathbf{k}^n$, define $\mathfrak{m} = I(\mathbf{a}) = (x_1 - a_1, \dots, x_n - a_n)$. Then $\mathfrak{m}$ is maximal in $\mathbf{k}[x_1, \dots, x_n]$. Indeed, any polynomial vanishing in $\mathbf{a}$ may be expanded in powers of $x_i - a_i$, so it belongs to $\mathfrak{m}$. Let $\mathfrak{m} \subsetneq R$. Then R must contain a polynomial not vanishing in $\mathbf{a}$. Therefore it must contain 1, and $R = \mathbf{k}[x_1, \dots, x_n]$.

Corollary 2.29. *If $\mathfrak{m}$ is a maximal ideal then $Z(\mathfrak{m})$ is a point.*

Proof. Let $\mathfrak{m}$ be a maximal ideal. Would $Z(\mathfrak{m})$ be empty, J would contain 1, contradiction. So $Z(\mathfrak{m})$ contains at least one point $\mathbf{a}$. Assume now that it contains a second point $\mathbf{b} \neq a$. They differ in at least one coordinate, say $a_1 \neq b_1$. Let J be the ideal generated by the elements of $\mathfrak{m}$ and by $x_1 - a_1$. Then $\mathbf{a} \in Z(J)$ but $b \neq Z(J)$. Hence $\mathfrak{m} \subsetneq J \subsetneq R$. □

Thus, I induces a bijection between points of $\mathbf{k}^n$ and maximal ideals of $\mathbf{k}[x_1, \dots, x_n]$.

Corollary 2.30. *Every non-empty Zariski-closed set can be written as a finite union of irreducible Zariski-closed sets.*

Proof. Let X be Zariski closed. By Theorem 2.24, $I(X)$ is a finite intersection of primary ideals:

$$I(X) = \mathfrak{J}_1 \cap \cdots \cap \mathfrak{J}_r.$$

Let $X_i = Z(\mathfrak{J}_i)$, for $i = 1, \dots, r$. By the Nullstellensatz, $I(X_i) = \sqrt{\mathfrak{J}_i}$. An ideal that is radical and primary is prime. Hence (Proposition 2.22) X_i is irreducible. □

An irreducible Zariski-closed set X is called an (affine) *algebraic variety*.i Its *dimension* r is the transcendence degree of $A = \mathbf{k}[x_1, \dots, x_n]$ over the prime ideal $Z(X)$. Its degree is the degree of A as an extension of $\mathbf{k}[x_1, \dots, x_r]$.

We restate an important consequence of Lemma 2.15 in the new language.

Lemma 2.31. *Let X be a variety of dimension r and degree d. Then, the number of isolated intersections of X with an affine hyperplane of codimension r is at most d. This number is attained for a generic choice of the hyperplane.*

Exercise 2.9. Let $\mathfrak{J}$ be an ideal. Show that $\sqrt{\mathfrak{J}}$ is an ideal.

Exercise 2.10. Prove that $\mathfrak{m}$ is a maximal ideal in $\mathbf{k}[x_1, \dots, x_n]$ if and only if, $A = \mathbf{k}[x_1, \dots, x_n]/\mathfrak{m}$ is a field.

2.7 Projective geometry

Corollary 2.32 (Projective Nullstellensatz)**.** *Let*

$$f_1, \dots, f_s \in \mathbf{k}[x_0, \dots, x_n]$$

be homogeneous polynomials. Assume they have no common root in $\mathbb{P}^n$. Then, there is $D \in \mathbb{N}$ such that $(x_0, \dots, x_n)^D \subseteq (f_1, \dots, f_s)$.

Proof. We first claim that for all i, there is $D_i \in \mathbb{N}$ so that $x_i^{D_i} \in (f_1, \dots, f_s)$. By reordering variables we may assume that $i = 0$. Specialize

$$F_j(x_1, \dots, x_n) = f_j(1, x_1, \dots, x_n).$$

Polynomials $F_1, \dots, F_s$ cannot have a common root, so Theorem 2.27 implies the existence of $G_1, \dots, G_s \in \mathbf{k}[x_1, \dots, x_n]$ with

$$F_1 G_1 + \cdots + F_s G_s = 1.$$

Let g_i denote the homogenization of G_i. We can homogenize so that all the $f_i g_i$ have the same degree D_0. In that case,

$$f_1 g_1 + \cdots + f_s g_s = x_0^{D_0}.$$

Now, set $D = D_0 + \cdots + D_n - n$. For any monomial $\mathbf{x}^{\mathbf{a}}$ of degree D, there is i such that $a_i \geq D_i$. Therefore, $\mathbf{x}^{\mathbf{a}}$ can be written as a linear combination of the f_i. $\square$

Let $d_1, \dots, d_s$ be fixed. By using the canonical monomial basis, we will consider $\mathcal{H}_{\mathbf{d}} = \mathcal{H}_{d_1} \times \cdots \times \mathcal{H}_{d_s}$ as a copy of $\mathbf{k}^S$, for $S = \sum_{i=1}^s \binom{d_i + n}{n}$. Elements of $\mathcal{H}_{\mathbf{d}}$ may be interpreted as systems of homogeneous polynomial equations.

Theorem 2.33 (Main theorem of elimination theory)**.** *Let $\mathbf{k}$ be an algebraically closed field. The set of $\mathbf{f} \in \mathcal{H}_{\mathbf{d}}$ with a common root in $\mathbb{P}(\mathbf{k}^{n+1})$ is a Zariski-closed set.*

Proof. Let X be the set of all $\mathbf{f} \in \mathcal{H}_{\mathbf{d}}$ with a common projective root. By the projective Nullstellensatz (Corollary 2.32), the condition $\mathbf{f} \in X$ is equivalent to:

$$\forall D, (x_0, \dots, x_n)^D \not\subseteq (f_1, \dots, f_s)$$

Denote by $M_{\mathbf{f}}^D : \mathcal{H}_{D-d_1} \times \cdots \mathcal{H}_{D-d_s} \mapsto \mathcal{H}_D$ the map

$$M_{\mathbf{f}}^D : g_1, \dots, g_s \mapsto f_1 g_1 + \cdots + f_s g_s.$$

Let X_D be the set of all $\mathbf{f}$ so that $M_{\mathbf{f}}^D$ fails to be surjective. The ideal $I(X_D)$ is either (1) or the zero-set of the ideal of maximal sub-determinants of $M_{\mathbf{f}}^D$. So it is always a Zariski closed set.

By Corollary 2.7 $X = \cap X_D$ is Zariski closed. □

We can use the Main Theorem of Elimination to deduce that for a larger class of polynomial systems, the number of zeros is *generically* independent of the value of the coefficients. We first will count roots in $\mathbb{P}^n$.

Corollary 2.34. *Let $\mathbf{k} = \mathbb{C}$. Let $\mathcal{F}$ be a subspace of $\mathcal{H} = \mathcal{H}_{d_1} \times \cdots \times \mathcal{H}_{d_n}$. Let $\mathcal{V} = \{(\mathbf{f}, \mathbf{x}) \in \mathcal{F} \times \mathbb{P}^n : \mathbf{f}(\mathbf{x}) = 0\}$ be the solution variety. Let $\pi_1 : \mathcal{V} \to \mathcal{F}$ and $\pi_2 : \mathcal{V} \to \mathbb{P}^n$ denote the canonical projections.*

Then, the critical values of π_1 are a strict Zariski closed subset of $\mathcal{F}$.

In particular, when $\mathbf{f} \in \mathcal{F}$ is a regular value for π_1,

$$n_{\mathbb{P}^n}(f) = \# \left(\pi_2 \circ \pi_1^{-1} \right) (\mathbf{f})$$

is independent of $\mathbf{f}$.

Proof. The critical values of π_1 are the systems $\mathbf{f} \in \mathcal{F}$ such that there is $0 \neq \mathbf{x} \in \mathbb{C}^{n+1}$ with

$$\mathbf{f}(\mathbf{x}) = 0 \quad \text{and} \quad \operatorname{rank}(D\mathbf{f}(\mathbf{x})) < n.$$

The rank of a $n \times n+1$ matrix is $< n$ if and only if all the $n \times n$ sub-matrices obtained by removing a column from $D\mathbf{f}(\mathbf{x})$ have zero determinant. By Theorem 2.33, the critical values of π_1 are then the intersection of $n+1$ Zariski-closed sets, hence in a Zariski-closed set.

Because of Sard's Theorem, the set of singular values has zero measure. Hence, it is a strict Zariski subset of $\mathcal{F}$.

Let $\mathbf{f}_0$ and $\mathbf{f}_1 \in \mathcal{F}$ be regular values of π_1. Because Zariski open sets are path-connected, there is a path joining $\mathbf{f}_0$ and $\mathbf{f}_1$ avoiding singular values. If $\mathbf{x}_0$ is a root of $\mathbf{f}_0$, then (by the implicit function theorem) the path $\mathbf{f}_t$ can be lifted to a path $(\mathbf{f}_t, \mathbf{x}_t) \in \mathcal{V}$. This implies that f_0 and f_1 have the same number of roots in $\mathbb{P}^n$. □

Corollary 2.35. *Let* $\mathbf{k} = \mathbb{C}$. *Let* $\mathcal{F}$ *be a subspace of* $\mathcal{H} = \mathcal{H}_{d_1} \times \cdots \times \mathcal{H}_{d_n}$. *Let* $U \subseteq \mathbb{P}^n$ *be Zariski open. Let* $\mathcal{V}_U = \{(\mathbf{f}, \mathbf{x}) \in \mathcal{F} \times U : \mathbf{f}(\mathbf{x}) = 0\}$ *be the incidence variety. Let* $\pi_1 : \mathcal{V} \to \mathcal{F}$ *and* $\pi_2 : \mathcal{V}_U \to \mathbb{P}^n$ *denote the canonical projections.*

Then, the critical values of π_1 *are a Zariski closed subset of* $\mathcal{F}$.

In particular, when $\mathbf{f} \in \mathcal{F}$ *is generic,*

$$n_U(\mathbf{f}) = \# \left(\pi_2 \circ \pi_1^{-1} \right) (\mathbf{f})$$

is independent of $\mathbf{f}$.

Proof. Let

$$\hat{V} = \{(\mathbf{f}, \mathbf{x}) \in \mathcal{F} \times \mathbb{P}^n : \mathbf{f}(\mathbf{x}) = 0\} = \cup_{\lambda \in \Lambda} \mathcal{V}_\lambda$$

where the V_λ are irreducible components. Let $\Lambda_\infty = \{\lambda \in \Lambda : V_\lambda \subseteq \pi_2^{-1}(\mathbb{P}^n \setminus U)\}$ be the components 'at infinity'.

Let $\Lambda_0 = \Lambda \setminus \Lambda_\infty$. Then $\mathcal{V}_U$ is an open subset of $\cup_{\lambda \in \Lambda_0} \mathcal{V}_\lambda$. Let

$$\mathcal{V}_{U,\infty} \stackrel{\text{def}}{=} \cup_{\lambda \in \Lambda_0} \mathcal{V}_\lambda \setminus \mathcal{V}_U.$$

This is a Zariski-closed set. Let W be the set of regular values of $(\pi_1)_{|\mathcal{V}_U}$ that are not in the projection of $\mathcal{V}_{U,\infty}$. W is Zariski-open.

Let $\mathbf{f}_0, \mathbf{f}_1 \in W$. Then there is a path $\mathbf{f}_t \in W$ connecting them. For each root $\mathbf{x}_0$ of $\mathbf{f}_0$, we can lift $\mathbf{f}_t$ to $(\mathbf{f}_t, \mathbf{x}_t) \subset \mathcal{V}_U$ as in the previous Corollary. $\square$

Chapter 3

Topology and zero counting

Arbitrarily small perturbations can obliterate zeros of smooth, even analytic real functions. For instance, $x^2 = 0$ admits a (double) root, but $x^2 = \epsilon$ admits no root for $\epsilon < 0$.

This cannot happen for complex analytic mappings. Recall that a real function φ from a metric space is *lower semi-continuous* at x if and only if,

$$\forall \delta > 0, \exists \epsilon > 0 \text{ s.t.} (d(x, y) < \epsilon) \Rightarrow \varphi(y) \geq \varphi(x) - \delta.$$

We will prove in Theorem 3.9) that the number of isolated roots of an analytic mapping is lower semi-continuous. As the local root count $n_U(f) = \#\{x \in U : f(x) = 0\}$ is a discrete function, this just means that

$$\exists \epsilon > 0 \text{ s.t.} \sup_{x \in U} \|f(x) - g(x)\| < \epsilon) \Rightarrow n_U(y) \geq n_U(x).$$

Gregorio Malajovich, *Nonlinear equations.*
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

As a side reference, I strongly recommend Milnor's book [61].

3.1 Manifolds

Definition 3.1 (Embedded manifold)**.** A *smooth (resp.* $\mathcal{C}^k$ *for* $k \geq 1$*, resp. analytic)* m*-dimensional real manifold* M *embedded in* $\mathbb{R}^n$ is a subset $M \subseteq \mathbb{R}^n$ with the following property: for any $\mathbf{p} \in M$, there are open sets $U \subseteq \mathbb{R}^m$, $\mathbf{p} \in V \subseteq \mathbb{R}^n$, and a smooth (resp. $\mathcal{C}^k$, resp. analytic) diffeomorphism $X : U \to M \cap V$. The map X is called a *parameterization* or a *chart*.

Recall that a *regular point* $\mathbf{x} \in \mathbb{R}^n$ of a $\mathcal{C}^1$ mapping $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^l$ is a point $\mathbf{x}$ such that the rank of $D\mathbf{f}(\mathbf{x})$ is $\min(n, l)$. A *regular value* $\mathbf{y} \in \mathbb{R}^l$ is a point such that $\mathbf{f}^{-1}(\mathbf{y})$ contains only regular points. A point that is not regular is said to be a *critical point*. Any $\mathbf{y} \in \mathbb{R}^l$ that is the image of a critical point is said to be a *critical value* for $\mathbf{f}$. Here is a canonical way to construct manifolds:

Proposition 3.2. *Let* $\Phi : \mathbb{R}^n \to \mathbb{R}^{n-m}$ *be a smooth (resp.* $\mathcal{C}^k$ *for* $k \geq 1$*, resp. analytic) function. If* 0 *is a regular value for* Φ*, then* $M = \Phi^{-1}(0)$ *is a smooth (resp.* $\mathcal{C}^k$*, resp. analytic)* m*-dimensional manifold.*

Proof. Let $\mathbf{p} \in M$. Because 0 is a regular value for Φ, we can apply the implicit function theorem to Φ in a neighborhood of $\mathbf{p}$. More precisely, we consider the orthogonal splitting $\mathbb{R}^n = \ker D\Phi(\mathbf{p}) \oplus \ker D\Phi(\mathbf{p})^\perp$. Locally at $\mathbf{p}$, we write Φ as

$$\mathbf{x}, \mathbf{y} \mapsto \Phi(\mathbf{p} + (\mathbf{x} \oplus \mathbf{y})).$$

Since $\mathbf{y} \mapsto D\Phi(\mathbf{p})\mathbf{y}$ is an isomorphism, the Implicit Function Theorem asserts that there is an open set $0 \in U \in \ker D\Phi(\mathbf{p}) \simeq \mathbb{R}^m$, and a an implicit function $y : U \to \ker D\Phi(\mathbf{p})^\perp$ such that

$$\Phi(\mathbf{p} + (\mathbf{x} \oplus y(\mathbf{x})) \equiv 0.$$

The function $y(\mathbf{x})$ has the same differentiability class as Φ.

By choosing an arbitrary basis for $\ker D\Phi(\mathbf{p})$, we obtain the 'local chart' $X : U \subseteq \mathbb{R}^m \to M$, given by $X(\mathbf{x}) = \mathbf{p} + (\mathbf{x} \oplus y(\mathbf{x}))$. ☐

Note that if $X : U \to M$ and $Y : V \to M$ are two local charts and domains $X(U) \cap Y(V) \neq \emptyset$, then $Y^{-1} \circ X$ is a diffeomorphism, of the same class as Φ.

A smooth (resp. $\mathcal{C}^k$, resp. analytic) m-dimensional *abstract manifold* is a topological space M such that, for every $p \in M$, there is a neighborhood of p in M that is smoothly (resp. $\mathcal{C}^k$, resp. analytically) diffeomorphic to an embedded m-dimensional manifold of the same differentiability class. Whitney's embedding theorem guarantees that a smooth abstract m-dimensional manifold can be embedded in $\mathbb{R}^{2m}$.

Let $\mathbb{H}^m_+$ (resp.) $\mathbb{H}^m_-$ be the closed half-space in $\mathbb{R}^m$ defined by the inequation $x_m \geq 0$ (resp. $x_m \leq 0$).

Definition 3.3 (Embedded manifold with boundary)**.** A *smooth (resp. $\mathcal{C}^k$ for $k \geq 1$, resp. analytic) m-dimensional real manifold M with boundary, embedded in* $\mathbb{R}^n$ is a subset $M \subseteq \mathbb{R}^n$ with the following property: for any $p \in M$, there are open sets $U \subseteq \mathbb{H}^m_+$ or $\mathbb{H}^m_-$, $p \in V \subseteq \mathbb{R}^n$, and a smooth (resp. $\mathcal{C}^k$, resp. analytic) diffeomorphism $X : U \to M \cap V$. The map X is called a *parameterization* or a *chart.*

The *boundary* ∂M of an embedded manifold M is the union of the images of the $X(U \cap [x_m = 0])$. It is also a smooth (resp. $\mathcal{C}^k$ resp. analytic) manifold (without boundary) of dimension $m - 1$.

Note the linguistic trap: every manifold is a manifold with boundary, while a manifold with boundary does not need to have a nonempty boundary.

Let $\mathbb{E}$ be a finite-dimensional real linear space. We say that two bases $(\alpha_1, \dots, \alpha_m)$ and $(\beta_1, \dots, \beta_m)$ of $\mathbb{E}$ have the same orientation if and only if $\det A > 0$, where A is the matrix relating those two bases:

$$\alpha_i = \sum_j A_{ij} \beta_j .$$

There are two possible orientations for a linear space. The canonical orientation of $\mathbb{R}^m$ is given by the canonical basis $(\mathrm{e}_1, \dots, \mathrm{e}_m)$.

The tangent space of M at p, denoted by T_pM, is the image of $DX_p \subseteq \mathbb{R}^n$. An *orientation* for an m-dimensional manifold M with boundary (this includes ordinary manifolds !) when $m \geq 1$ is a class of charts $X_\alpha : U_\alpha \to M$ covering M, such that whenever $V_\alpha \cap V_\beta \neq \emptyset$, $\det(D\left(X_\alpha^{-1} X_\beta\right)_{\mathbf{x}}) > 0$ for all $\mathbf{x} \in U_\beta \cap X_\beta^{-1}(V_\alpha)$.

An orientation of M defines orientations in each T_pM. A manifold admitting an orientation is said to be *orientable*. If M is orientable and connected, an orientation in one T_pM defines an orientation in all M.

A 0-dimensional manifold is just a union of disjoint points. An An orientation for a zero-manifold is an assignment of ± 1 to each point.

If M is an oriented manifold and ∂M is non-empty, the boundary ∂M is oriented by the following rule: let $p \in \partial M$ and assume a parameterization $X : U \cap \mathbb{H}^m_- \to M$. With this convention we choose the sign so that $u = \pm \frac{\partial X}{\partial x_n}$ is an outward pointing vector. We say that $X_{|U\cap[x_m=0]}$ is positively oriented if and only if X is positively oriented.

The following result will be used:

Proposition 3.4. *A smooth connected 1-dimensional manifold (possibly) with boundary is diffeomorphic either to the circle S^1 or to a connected subset of $\mathbb{R}$.*

Proof. A parameterization by arc-length is a parameterization $X : U \to M$ with

$$\left\| \frac{\partial X}{\partial x_1} \right\| = 1.$$

Step 1: For each interior point $p \in M$, there is a parameterization $X : U \to V \in M$ by arc-length.

Indeed, we know that there is a parameterization $Y : (a, b) \to V \ni p$, $Y(0) = p$.

For each $q = Y(c) \in V$, let

$$t(q) = \begin{cases} \int_0^c \|Y'(t)\| \mathrm{d}t & \text{if } c \geq 0 \\ -\int_c^0 \|Y'(t)\| \mathrm{d}t & \text{if } c \leq 0 \end{cases}$$

The map $t : V \to \mathbb{R}$ is a diffeomorphism of V into some interval $(d, e) \subset \mathbb{R}$. Let $U = (d, e)$ and $X = Y \circ t^{-1}$. Then $X : U \to M$ is a parameterization by arc length.

Step 2: Let p be a fixed interior point of M. Let q be an arbitrary point of M. Because M is connected, there is a path $\gamma(t)$ linking p

to q. Each point of $\gamma(t)$ admits an arc-length parameterization for a neighborhood of it. As the path is compact, we can pick a finite subcovering of those neighborhoods.

By patching together the parameterizations, we obtain one by arc length $X' : (a', b') \to M$ with $X'(a') = p$, $X'(b') = q$.

Step 3: Two parameterizations by arc length with $X(0) = Y(0)$ are equal in the overlap of their domains, or differ by time reversal.

Step 4: Let $p \in M$ be an arbitrary interior point. Then, let $X : W \to M$ be the maximal parameterization by arc length with $X(0) \to M$. The domain W is connected. Now we distinguish two cases.

Step 4, case 1: X is injective. In that case, X is a diffeomorphism between M and a connected subset of $\mathbb{R}$

Step 4, case 2: Let r have minimal modulus so that $X(0) = X(r)$. Unicity of the path-length parameterization implies that for all $k \in \mathbb{Z}$, $X(kr) = X(r)$. In that case, X is a diffeomorphism of the topological circle $\mathbb{R} \mod r$ into M. □

Exercise 3.1. Give an example of embedded manifold in $\mathbb{R}^n$ that is not the preimage of a regular value of a function. (This does not mean it cannot be embedded into some $\mathbb{R}^N$!).

3.2 Brouwer degree

Through this section, let $\mathcal{B}$ be an open ball in $\mathbb{R}^n$, $\overline{\mathcal{B}}$ denotes its topological closure, and $\partial\mathcal{B}$ its boundary.

Lemma 3.5. *Let $f : \mathcal{B} \to \mathbb{R}^n$ be a smooth map, extending to a $\mathcal{C}^1$ map $\bar{f}$ from $\overline{\mathcal{B}}$ to $\mathbb{R}^n$. Let $\mathcal{Y}_f \subset \mathbb{R}^n$ be the set of regular values of f, not in $f(\partial\mathcal{B})$. Then, $\mathcal{Y}_f$ has full measure and any $\mathbf{y} \in \mathcal{Y}_f$ has at most finitely many preimages in $\mathcal{B}$.*

Proof. By Sard's theorem, the set of regular values of f has full measure. Moreover, $\partial\mathcal{B}$ has finite volume, hence it can be covered by a finite union of balls of arbitrarily small total volume. Its image $f(\partial\mathcal{B})$ is contained in the image of this union of balls. Since f is $\mathcal{C}^1$ on $\mathcal{B}$, we can make the volume of the image of the union of balls arbitrarily small. Hence, $f(\partial\mathcal{B})$ has zero measure. Therefore, $\mathcal{Y}_f$ has full measure. □

For $y \in \mathcal{Y}_f$, we define:

$$\deg(f, y) = \sum_{x \in f^{-1}(y)} \operatorname{sign} \det Df(x).$$

Theorem 3.6. *Under the conditions of Lemma 3.5,* $\deg(f, y)$ *does not depend on the choice of* $y \in \mathcal{Y}_f$.

We define the *Brouwer degree* $\deg(f)$ *of* f as $\deg(f, y)$ for $y \in \mathcal{Y}_f$.

Before proving theorem 3.6, we need a few preliminary definitions. Let $\mathcal{F}$ be the space of mappings satisfying the conditions of Lemma 3.5, namely the smooth maps $f : \mathcal{B} \to \mathbb{R}^n$ extending to a $\mathcal{C}^1$ map $\bar{f} : \overline{\mathcal{B}} \to \mathbb{R}^n$.

A *smooth homotopy* on $\mathcal{F}$ is a smooth map $f : [0, 1] \times B \to \mathbb{R}^n$, extending to a $\mathcal{C}^1$ map $\bar{f}$ on $[0, 1] \times \overline{B}$. We say that f and $g \in \mathcal{F}$ are smoothly homotopic if and only if there is a smooth homotopy $H : [a, b] \times \mathcal{B} \to \mathbb{R}^n$ with $H(a, \mathbf{x}) \equiv f(\mathbf{x})$ and $H(b, \mathbf{x}) \equiv g(\mathbf{x})$.

Lemma 3.7. *Assume that* f *and* $g \in \mathcal{F}$ *are smoothly homotopic, and that* $\mathbf{y} \in \mathcal{Y}_f \cap \mathcal{Y}_g$. *Then,*

$$\deg(f; y) = \deg(g; y).$$

Proof. Let $H : [a, b] \times \mathcal{B} \to \mathcal{R}^n$ be the smooth homotopy between f and g. Let $\mathcal{Y}$ be the set of regular values of H, not in $H([a, b] \times \partial B)$. Then $\mathcal{Y}$ has full measure in $\mathbb{R}^n$.

Consider the manifold $M = [a, b] \times \mathcal{B}$. It admits an obvious orientation as a subset of $\mathbb{R}^{n+1}$. Its boundary is

$$\partial M = (\{a\} \times B) \cup (\{b\} \times B) \cup ([a, b] \times \partial B)$$

Now, $\overline{H}_{|\{a,b\} \times B}$ is smooth and admits y as a regular value. Therefore, there is an open neighborhood $U \ni y$ so that all $\tilde{y} \in U$ is a regular value for $\overline{H}_{|\{a,b\} \times B}$.

Because $\overline{B}$ is compact, we can take U small enough so that the number of preimages of $\tilde{y}$ in $\{a\} \times B$ (and also on $\{b\} \times B$) is constant. Since $\mathcal{Y}$ has full measure, there is $\tilde{y} \in U$ regular value for H, and also for $\overline{H}_{|\{a,b\} \times B}$.

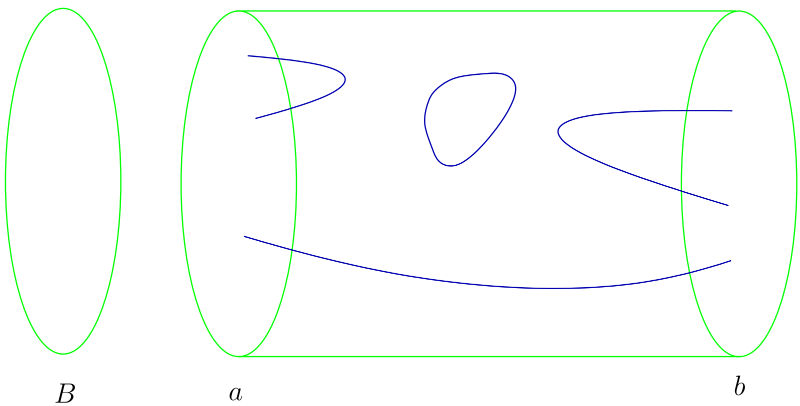

Figure 3.1: The four possible cases.

Let $X = \bar{H}^{-1}(\tilde{y})$. Then X is a one-dimensional manifold. Its boundary belongs to ∂M. But by construction, it cannot intersect $[a,b] \times \partial B$. Therefore, if we set $\hat{H}(t,x) = (t, H(t,x))$, we can interpret

$$\deg(g,y) - \deg(f,y) = \sum_{(b,x)\in\partial X} \operatorname{sign}\det D\hat{H}(b,x) - \sum_{(a,x)\in\partial X} \operatorname{sign}\det D\hat{H}(a,x).$$

By Proposition 3.4, each of the connected components X_i is diffeomorphic to either the circle S^1, or a connected subset of the real line. We claim that each ∂X_i has a zero contribution to the sum above.

There are four possibilities (fig. 3.1) for each connected component X_i: both boundary points in $\{a\} \times B$, in $\{b\} \times B$, one in each, or the component is isomorphic to S^1 (no boundary).

In the first case, let $s \mapsto (t(s), x(s)), s_0 \le s \le s_1$ be a (regular) parameterization of X_i.

Because $\hat{y}$ is a regular value of H, $\ker DH(x,t)$ is always one-dimensional.

$$D(s) = \det \begin{bmatrix} \frac{\partial}{\partial s} t(s) & \frac{\partial}{\partial s} \mathbf{x}(s)^* \\ D_t H(t(s), x(x)) & D_x H(t(s), x(s)) \end{bmatrix} \neq 0$$

and in particular this determinant has the same sign at the boundaries of X_i.

Again, because $\tilde{y}$ is a regular value of f, the tangent vector of X_i at s_0 is of the form

$$\begin{bmatrix} v \\ -v D_x H(t,x)^{-1}(g(x) - f(x)) \end{bmatrix}$$

Thus,

$$D(s_0) = \det \left(\begin{bmatrix} v & 0 \\ 0 & Df(x) \end{bmatrix} \begin{bmatrix} 1 & -\mathbf{w}^* \\ \mathbf{w} & I \end{bmatrix} \right)$$

with $\mathbf{w} = Df(x)^{-1}(g(x) - f(x))$ and $x = x(s_0)$. The reader shall check that the rightmost term has always strictly positive determinant $1 + \|\mathbf{w}\|^2$. Therefore, $\det D(s_0)$ has the same sign of $\det Df(x)$.

When $s = s_1$, we have exactly the same situation with $v < 0$. Thus,

$$\operatorname{sign} \det Df(x(s_0)) + \operatorname{sign} \det Df(x(s_1)) = 0$$

The second case $t(s_0) = t(s_1) = b$ is identical with signs of v reversed. In the third case, we assume that $t(s_0) = a$ and $t(s_1) = b$, and hence $v > 0$ in both extremities. There we have

$$\operatorname{sign} \det Df(x(s_0)) - \operatorname{sign} \det Df(x(s_1)) = 0$$

The fourth case is trivial.

We conclude that

$$\deg(g, y) - \deg(f, y) = \sum_i \left(\sum_{(b,x) \in \partial X_i} \operatorname{sign} \det DH(b, x) - \right.$$
$$\left. - \sum_{(a,x) \in \partial X_i} \operatorname{sign} \det DH(a, x) \right) = 0.$$

□

Proof of Theorem 3.6. Let y, z be regular values of f. Since M is connected, they belong to the same component of M. Let $h_t(x) = x + t(z - y)$, $t \in [0, 1]$.

Then, f and $f \circ h(1, \cdot)$ are smoothly homotopic, and admit y as a common regular value. Using the chain rule, we deduce that the degree of f in y is equal to the degree of f in z. □

3.3 Complex manifolds and equations

Let M be a complex manifold. In a neighborhood U of some $p \in M$, pick a bi-holomorphic function f from U to $f(U) \subseteq \mathbb{C}^n$. The pull-back of the canonical orientation of $\mathbb{C}^n$ by f defines an orientation on $T_q M$ for all $q \in U$. This orientation does not depend on the choice of f. We call this orientation *the canonical orientation* of M. We proved:

Theorem 3.8. *Complex manifolds are orientable.*

Theorem 3.9. *Let M be an n-dimensional complex manifold, without boundary. Let $\mathcal{F}$ be a space of holomorphic functions $M \to \mathbb{C}^n$. Given $f \in \mathcal{F}$ and U open in M, let $n_U(f) = \#f^{-1}(0) \cap U$ be the number of isolated zeros of f in U, counted without multiplicity. Then, $n_U : \mathcal{F} \to \mathbb{Z}_{\geq 0}$ is lower semi-continuous at all f where $n_U(f) < \infty$.*

Proof. In order to prove lower semi-continuity of n_U, it suffices to prove that for any isolated zero ζ of f, for any $\delta > 0$ small enough, there is $\epsilon > 0$ such that if $\|g - f\| < \epsilon$, then g has a root in $B(\zeta, \delta)$. Then pick δ such that two isolated roots of f are always at distance $> 2\delta$.

Because complex manifolds admit a canonical orientation, the Brouwer degree of $f_{|B(\zeta,\delta)}$ is a strictly positive integer. Since it is locally constant, there is $\epsilon > 0$ so that it is constant in $B(f, \epsilon)$. □

Chapter 4

Differential forms

Through this section, vectors are represented boldface such as $\mathbf{x}$ and coordinates are represented as x_j. Whenever we are speaking about a collection of vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$, x_{ij} is the j-th coordinate of the i-th vector.

4.1 Multilinear algebra over $\mathbb{R}$

Let $\mathcal{A}^k$ be the space of alternating k-forms in $\mathbb{R}^n$, that is the space of all k-linear forms $\alpha : (\mathbb{R}^n)^k \to \mathbb{R}$ such that, for all permutation $\sigma \in S_k$ (the permutation group of k elements), we have:

$$\alpha(\mathbf{u}_{\sigma_1}, \dots, \mathbf{u}_{\sigma_k}) = (-1)^{|\sigma|} \alpha(\mathbf{u}_1, \dots, \mathbf{u}_k).$$

Above, $|\sigma|$ is minimal so that σ is the composition of $|\sigma|$ *elementary permutations* (permutations fixing all elements but two).

The *canonical basis* of $\mathcal{A}^k$ is given by the forms $\mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_k}$,

Gregorio Malajovich, *Nonlinear equations*.
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

with $1 \le i_1 < i_2 < \cdots < i_k \le n$, defined by

$$\mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_k}(\mathbf{u}_1, \ldots, \mathbf{u}_k) = \sum_{\sigma \in S_k} (-1)^{|\sigma|} u_{\sigma(1)i_1} u_{\sigma(2)i_2} \cdots u_{\sigma(k)i_k}.$$

The *wedge product* $\wedge : \mathcal{A}^k \times \mathcal{A}^l \to \mathcal{A}^{k+l}$ is defined by

$$\alpha \wedge \beta \ (\mathbf{u}_1, \ldots, \mathbf{u}_{k+l}) = \\ = \frac{1}{k!l!} \sum_{\sigma \in S_{k+l}} (-1)^{|\sigma|} \alpha(\mathbf{u}_{\sigma(1)}, \ldots, \mathbf{u}_{\sigma(k)}) \beta(\mathbf{u}_{\sigma(k+1)}, \ldots, \mathbf{u}_{\sigma(k+l)})$$

The coefficient $\frac{1}{k!l!}$ above may be replaced by $\begin{pmatrix} k+l \\ k \end{pmatrix}$ if one replaces the sum by the anti-symmetric average over S_{k+l}. This convention makes the wedge product associative, in the sense that

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma). \tag{4.1}$$

so we just write $\alpha \wedge \beta \wedge \gamma$. This is also compatible with the notation $\mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_n}$.

Another important property of the wedge product is the following: if $\alpha \in \mathcal{A}^k$ and $\beta \in \mathcal{A}^l$, then

$$\alpha \wedge \beta = (-1)^{kl} \beta \wedge \alpha. \tag{4.2}$$

Let $U \subseteq \mathbb{R}^n$ be an open set (in the usual topology), and let $\mathcal{C}^\infty(U)$ denote the space of all smooth real valued functions defined on U. The fact that a linear k-form takes values in $\mathbb{R}$ is immaterial in all the definitions above.

Definition 4.1. The space of *differential k-forms in U*, denoted by $\mathcal{A}^k(U)$, is the space of linear k-forms defined in $\mathbb{R}^n$ with values in $\mathcal{C}^\infty(U)$.

This is equivalent to smoothly assigning to each point $\mathbf{x}$ on U, a linear k-form with values in $\mathbb{R}$. If $\alpha \in \mathcal{A}^k$, we can therefore write

$$\alpha_{\mathbf{x}} = \sum_{1 \le i_1 < \cdots < i_k \le n} \alpha_{i_1, \ldots, i_k}(\mathbf{x}) \ \mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_k}.$$

Properties (4.1) and (4.2) hold in this context. We introduce the *exterior derivative* operator $\mathrm{d} : \mathcal{A}^k \to \mathcal{A}^{k+1}$:

$$\mathrm{d}\alpha_{\mathbf{x}} = \sum_{\substack{1 \le i_1 < \cdots < i_k \le n \\ 1 \le j \le n \\ j \ne i_1, \dots, i_k}} \frac{\partial \alpha_{i_1,\dots,i_k}}{\partial x_j}(\mathbf{x}) \, \mathrm{d}x_j \wedge \mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_k}$$

Setting $\mathcal{A}^0(U) = \mathcal{C}^\infty(U)$, we see that d coincides with the ordinary derivative of functions. The exterior derivative is $\mathbb{R}$-linear and furthermore

$$\mathrm{d}^2 = \mathrm{d} \circ \mathrm{d} = 0 \tag{4.3}$$

and

$$\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^k \alpha \wedge \mathrm{d}\beta \tag{4.4}$$

Definition 4.2. Let $f : U \subseteq \mathbb{R}^m \to V \subseteq \mathbb{R}^n$ be of class $\mathcal{C}^\infty$. The *pull-back* of a differential form $\alpha \in \mathcal{A}^k(V)$ by f, denoted by $f^*\alpha$, is the element of $\mathcal{A}^k(U)$ given by

$$(f^*\alpha)_{\mathbf{x}}(\mathbf{u}_1, \dots, \mathbf{u}_k) = \alpha_{f(\mathbf{x})} \left(Df(\mathbf{x})\mathbf{u}_1, \dots, Df(\mathbf{x})\mathbf{u}_k \right).$$

The chain rule for functions can be written simply as

$$\mathrm{d}(f \circ g) = g^* \mathrm{d}f$$

Exercise 4.1. Check formulas (4.1), (4.2), (4.3), (4.4).

Exercise 4.2. Show that if A is an $n \times n$ matrix,

$$\begin{aligned} &\det(A)\ \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_n = \\ &\qquad = (A_{11}\mathrm{d}x_1 + \cdots + A_{1n}\mathrm{d}x_n) \wedge \cdots \wedge (A_{n1}\mathrm{d}x_1 + \cdots + A_{nn}\mathrm{d}x_n) \end{aligned}$$

4.2 Complex differential forms

An old tradition dictates that x means the 'thing', the unknown on one equation. While I try to comply in most of this text, here I will switch to another convention: if z is a complex number, x is its real part and y its imaginary part. This convention extends to vectors so

$$\mathbf{z} = \mathbf{x} + \sqrt{-1}\ \mathbf{y}.$$

The sets $\mathbb{C}^n$ and $\mathbb{R}^{2n}$ may be identified by

$$\mathbf{z} = \begin{bmatrix} x_1 \\ y_1 \\ x_2 \\ \vdots \\ y_n \end{bmatrix}.$$

It is possible to define alternating k-forms in $\mathbb{C}^n$ as complex-valued alternating k-forms in $\mathbb{R}^{2n}$. However, this approach misses some of the structure related to the linearity over $\mathbb{C}$ and holomorphic functions. Instead, it is usual to define $\mathcal{A}^{k0}$ as the space of complex valued alternating k-forms in $\mathbb{C}^n$. A basis for $\mathcal{A}^{k0}$ is given by the expressions

$$\mathrm{d}z_{i_1} \wedge \cdots \wedge \mathrm{d}z_{i_k}, \quad 1 \le i_1 < i_2 < \cdots < i_k \le n.$$

They are interpreted as

$$\mathrm{d}z_{i_1} \wedge \cdots \wedge \mathrm{d}z_{i_k}(\mathbf{u}_1, \dots, \mathbf{u}_k) = \sum_{\sigma \in S_k} (-1)^{|\sigma|} u_{\sigma(1)i_1} u_{\sigma(2)i_2} \cdots u_{\sigma(k)i_k}.$$

Notice that $\mathrm{d}z_i = \mathrm{d}x_i + \sqrt{-1}\ \mathrm{d}y_i$. We may also define $\mathrm{d}\bar{z}_i = \mathrm{d}x_i - \sqrt{-1}\ \mathrm{d}y_i$. Next we define $\mathcal{A}^{kl}$ as the complex vector space spanned by all the expressions

$$\mathrm{d}z_{i_1} \wedge \cdots \wedge \mathrm{d}z_{i_k} \wedge \mathrm{d}\bar{z}_{j_1} \wedge \cdots \wedge \mathrm{d}\bar{z}_{j_l}$$

for $1 \le i_1 < i_2 < \cdots < i_k \le n, 1 \le j_1 < j_2 < \cdots < j_l \le n$. Since

$$\mathrm{d}x_i \wedge \mathrm{d}y_i = -2\sqrt{-1}\ \mathrm{d}z_i \wedge \mathrm{d}\bar{z}_i,$$

the standard volume form in $\mathbb{C}^n$ is

$$\mathrm{d}V = \mathrm{d}x_1 \wedge \mathrm{d}y_1 \wedge \cdots \wedge \mathrm{d}y_n = \left(\frac{\sqrt{-1}}{2}\right)^n \mathrm{d}z_1 \wedge \mathrm{d}\bar{z}_1 \wedge \cdots \wedge \mathrm{d}\bar{z}_n.$$

The following fact is quite useful:

Lemma 4.3. *If A is an $n \times n$ matrix, then*

$$|\det(A)|^2 \;\mathrm{d}V = \bigwedge_{k=1}^{n} \sum_{i,j=1}^{n} \frac{\sqrt{-1}}{2} A_{ki} \bar{A}_{kj} \;\mathrm{d}z_i \wedge \mathrm{d}\bar{z}_j$$

Proof. As in exercise 4.2,

$$\det(A) \;\mathrm{d}z_1 \wedge \cdots \wedge \mathrm{d}z_n = \bigwedge_{k=1}^{n} \sum_{i=1}^{n} A_{ki} \;\mathrm{d}z_i$$

and

$$\overline{\det(A)} \;\mathrm{d}\bar{z}_1 \wedge \cdots \wedge \mathrm{d}\bar{z}_n = \bigwedge_{k=1}^{n} \sum_{j=1}^{n} \bar{A}_{kj} \;\mathrm{d}\bar{z}_j .$$

The Lemma is proved by wedging the two expressions above and multiplying by $(\sqrt{-1}/2)^n$. □

If U is an open subset of $\mathbb{C}^n$, then $\mathcal{C}^\infty(U, \mathbb{C})$ is the *complex* space of all smooth complex valued functions of U. Here, *smooth* means of class $\mathcal{C}^\infty$ and *real* derivatives are assumed. The holomorphic and anti-holomorphic derivatives are defined as

$$\frac{\partial f}{\partial z_i} = \frac{1}{2}\left(\frac{\partial f}{\partial x_i} - \sqrt{-1} \frac{\partial f}{\partial y_i} \right)$$

and

$$\frac{\partial f}{\partial \bar{z}_i} = \frac{1}{2}\left(\frac{\partial f}{\partial x_i} + \sqrt{-1} \frac{\partial f}{\partial y_i} \right)$$

The Cauchy-Riemann equations for a function f to be holomorphic are just

$$\frac{\partial f}{\partial \bar{z}_i} = 0.$$

We denote by $\partial : \mathcal{A}^{kl}(U) \to \mathcal{A}^{k+1,l}(U)$ the holomorphic differential, and by $\bar{\partial} : \mathcal{A}^{kl}(U) \to \mathcal{A}^{k,l+1}(U)$ the anti-holomorphic differential. If

$$\alpha(\mathbf{z}) = \sum_{\substack{1 \le i_1 < i_2 < \cdots < i_k \le n \\ 1 \le j_1 < j_2 < \cdots < j_l \le n}} \alpha_{i_1,\ldots,j_l}(\mathbf{z}) \;\mathrm{d}z_{i_1} \wedge \cdots \wedge \mathrm{d}\bar{z}_{j_l},$$

then

$$\partial\alpha(\mathbf{z}) = \sum_{\substack{1\le i_1<i_2<\dots<i_k\le n\\ 1\le j_1<j_2<\dots<j_l\le n\\ 1\le k\le n,\ k\ne i_r}} \frac{\partial\alpha_{i_1,\dots,j_l}}{\partial z_k}(\mathbf{z})\, \mathrm{d}z_k \wedge \mathrm{d}z_{i_1} \wedge \dots \wedge \mathrm{d}\bar{z}_{j_l},$$

and

$$\bar{\partial}\alpha(\mathbf{z}) = \sum_{\substack{1\le i_1<i_2<\dots<i_k\le n\\ 1\le j_1<j_2<\dots<j_l\le n\\ 1\le k\le n,\ k\ne j_r}} \frac{\partial\alpha_{i_1,\dots,j_l}}{\partial \bar{z}_k}(\mathbf{z})\, \mathrm{d}\bar{z}_k \wedge \mathrm{d}z_{i_1} \wedge \dots \wedge \mathrm{d}\bar{z}_{j_l},$$

The *total* differential is $\mathrm{d} = \partial + \bar{\partial}$. Another useful fact is that $\partial^2 = \bar{\partial}^2 = 0$.

4.3 Kähler geometry

Let $U \subseteq \mathbb{C}^n$ be an open set, and let $g_{ij} : U \to \mathbb{C}$ be such that each $g_{ij} \in \mathcal{C}^\infty(U, \mathbb{C}^n)$ and furthermore, the matrix $g(z) = [g_{ij}(z)]_{1\le i,j\le n}$ is Hermitian positive definite at each z.

This defines, at each $\mathbf{z} \in U$, the Hermitian inner product

$$\langle \mathbf{u}, \mathbf{v}\rangle_{\mathbf{z}} = \sum_{i,j=1}^{n} g_{ij}(\mathbf{z}) u_i \bar{v}_j .$$

The corresponding volume form is $\mathrm{d}V(\mathbf{z}) = |\det g_{ij}(\mathbf{z})|$ (compare with the Riemannian case).

Because $g(\mathbf{z})$ is Hermitian, its real part is symmetric and defines a Riemannian metric. Thus $\omega_{\mathbf{z}} = -\mathrm{im}\ g_{\mathbf{z}}(\cdot,\cdot)$ is skew-symmetric whence in $\mathcal{A}^{11}(U)$.

Definition 4.4. A *Kähler form* is a form $\omega_{\mathbf{z}} \in \mathcal{A}^{11}(U)$ that is:

1. *positive*:

$$\omega_{\mathbf{z}}(\mathbf{u}, \sqrt{-1}\ \mathbf{u}) \ge 0$$

with equality only if $\mathbf{u} = 0$, and

2. *closed:* $\mathrm{d}\omega_{\mathbf{z}} \equiv 0$.

The canonical Kähler form in $\mathbb{C}^n$ is

$$\omega = \frac{\sqrt{-1}}{2}\mathrm{d}z_1 \wedge \mathrm{d}\bar{z}_1 + \frac{\sqrt{-1}}{2}\mathrm{d}z_2 \wedge \mathrm{d}\bar{z}_2 + \cdots + \frac{\sqrt{-1}}{2}\mathrm{d}z_n \wedge \mathrm{d}\bar{z}_n.$$

Given a Kähler form, its volume form can be written as

$$\mathrm{d}V_{\mathbf{z}} = \frac{1}{n!}\underbrace{\omega_{\mathbf{z}} \wedge \omega_{\mathbf{z}} \wedge \cdots \wedge \omega_{\mathbf{z}}}_{n \text{ times}}.$$

The definition above is for a Kähler structure on a subset of $\mathbb{C}^n$. This definition can be extended to a complex manifold, or to a $2n$-manifold where a 'complex multiplication' $J : T_{\mathbf{z}}M \to T_{\mathbf{z}}M$, $J^2 = -I$, is defined.

An amazing fact about Kähler manifolds is the following.

Theorem 4.5 (Wirtinger)**.** *Wirtinger Let S be a d-dimensional complex submanifold of a Kähler manifold M. Then it inherits its Kähler form, and*

$$\mathrm{Vol}(S) = \frac{1}{d!}\int_S \underbrace{\omega_{\mathbf{z}} \wedge \cdots \wedge \omega_{\mathbf{z}}}_{d\ \textit{times}}.$$

Since ω is a closed form, $\omega\wedge\cdots\wedge\omega$ is also closed. When S happens to be a boundary, its volume is zero.

4.4 The co-area formula

Definition 4.6. A smooth (real, complex) fiber bundle is a tuple (E, B, π, F) such that

1. E is a smooth (real, complex) manifold (known as *total space*).
2. B is a smooth (real , complex) manifold (known as *base space*).
3. $\pi : E \mapsto B$ is a smooth surjection (the *projection*).
4. F is a (real, complex) smooth manifold (the *fiber*).

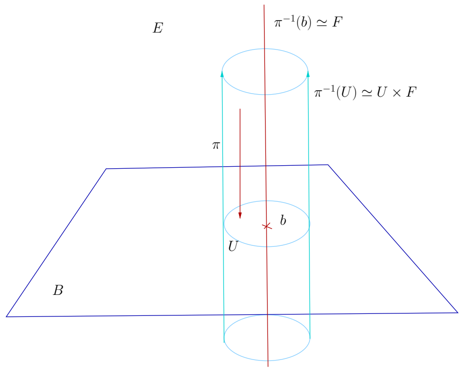

Figure 4.1: Fiber bundle.

5. The local triviality condition: for every $p \in E$, there is an open neighborhood $U \ni \pi(p)$ in B and a diffeomorphism $\Phi : \pi^{-1}(U) \to U \times F$. (the *local trivialization*).

6. Moreover, $\Phi_{|\pi^{-1}\circ\pi(p)} \to F$ is a diffeomorphism.

(See figure 4.1).

Familiar examples of fiber bundles are the tangent bundle of a manifold, the normal bundle of an embedded manifold, etc... In those case the fiber is a vector space, so we speak of a *vector bundle*. The fiber may be endowed of another structure (say a group) which is immaterial here.

Here is a less familiar example of a vector bundle. Recall that $\mathcal{P}_d$ is the space of complex univariate polynomials of degree $\leq d$. Let $\mathcal{V} = \{(f, x) \in \mathcal{P}_d \times \mathbb{C} : f(x) = 0\}$. This set is known as the *solution variety*. Let $\pi_2 : \mathcal{V} \to \mathbb{C}$ be the projection into the second set of coordinates, namely $\pi_2(f, x) = x$. Then $\pi_2 : \mathcal{V} \to \mathbb{C}$ is a vector bundle.

The co-area formula is a Fubini-type theorem for fiber bundles:

Theorem 4.7 (co-area formula). *Let (E, B, π, F) be a real smooth fiber bundle. Assume that B is finite dimensional. Let $f : E \to \mathbb{R}_{\geq 0}$*

be measurable. Then whenever the left integral exists,

$$\int_E f(p)\mathrm{d}E(p) = \int_B \mathrm{d}B(x) \int_{E_x} (\det D\pi(p)D\pi(p)^*)^{-1/2} f(p)\mathrm{d}E_x(p).$$

with $E_x = \pi^{-1}(x)$.

Lemma 4.8. *In the conditions of Theorem 4.7, there is a locally finite open covering* $\mathfrak{U} = \{U_\alpha\}$ *of* B, *and a family of smooth functions* $\psi_\alpha \geq 0$ *with domain* B *vanishing in* $B \setminus U_\alpha$ *such that*

1. *Each* $U_\alpha \in \mathfrak{U}$ *is such that there is a local trivialization* Φ *with domain* $\Phi^{-1}(U_\alpha)$.

2. $$\sum_\alpha \psi_\alpha(x) \equiv 1.$$

The family $\{\psi_\alpha\}$ is said to be a *partition of unity* for $\pi : E \to B$.

Proof of theorem 4.7. Let ψ_α be the partition of unity from Lemma 4.8. By replacing f by $f(\psi_\alpha \circ \pi)$ and then adding for all α, we can assume without loss of generality that f vanishes outside the domain $\pi^{-1}(U)$ of a local trivialization.

Now,

$$\begin{aligned}
\int_E f(p)\mathrm{d}E(p) &= \int_{\pi^{-1}(U)} f(p)\mathrm{d}E(p) \\
&= \int_{\Phi(\pi^{-1}(U))} \det D\Phi^{-1}(x,y) f(\Phi^{-1}(x,y))\mathrm{d}B(x)\mathrm{d}F(y) \\
&= \int_U \mathrm{d}B(x) \int_F \det D\Phi^{-1}(x,y) f(\Phi^{-1}(x,y))\mathrm{d}F(y)
\end{aligned}$$

using Fubini's theorem. Note that $\Phi_{|F_x} \to F$ is a diffeomorphism, so the inner integral can be replaced by

$$\int_{F_x} \det D\Phi_{|F_x} \det D\Phi^{-1}(p) f(p)\mathrm{d}F_x(p).$$

Moreover, by splitting $T_pE = \ker D\pi^\perp \oplus \ker D\pi$ and noticing that $F_x = \ker D\pi(p)$,

$$D\Phi = \begin{bmatrix} D\pi(p) & 0 \\ ? & D\Phi_{|F_x}(p) \end{bmatrix}.$$

Therefore

$$\det D\Phi_{|F_x} \det D\Phi^{-1} = \det\left(D\pi^{-1}_{|\ker D\pi^\perp}\right) = (\det D\pi D\pi^*)^{-1/2}.$$

□

When the fiber bundle is complex, we obtain a similar formula by assimilating $\mathbb{C}^n$ to $\mathbb{R}^{2n}$:

Theorem 4.9 (co-area formula)**.** *Let (E, B, π, F) be a complex smooth fiber bundle. Assume that B is finite dimensional. Let $f : E \to \mathbb{R}_{\geq 0}$ be measurable. Then whenever the left integral exists,*

$$\int_E f(p)\mathrm{d}E(p) = \int_B \mathrm{d}B(x) \int_{E_x} (\det D\pi(p)D\pi(p)^*)^{-1} f(p)\mathrm{d}E_x(p).$$

with $E_x = \pi^{-1}(x)$.

4.5 Projective space

Complex projective space $\mathbb{P}^n$ is the quotient of $\mathbb{C}^{n+1} \setminus \{0\}$ by the multiplicative group $\mathbb{C}_\times$. This means that the elements of $\mathbb{P}^n$ are complex 'lines' of the form

$$(x_0 : \cdots : x_n) = \{(\lambda x_0, \lambda x_1, \cdots, \lambda x_n) : 0 \neq \lambda \in \mathbb{C}\}.$$

It is possible to define local charts at $(p_0 : \cdots : p_n) : \mathbf{p}^\perp \subset \mathbb{C}^{n+1} \to \mathbb{P}^n$ by sending $\mathbf{x}$ into $(p_0 + x_0 : \cdots : p_n + x_n)$.

There is a canonical way to define a metric in $\mathbb{P}^n$, in such a way that for $\|\mathbf{p}\| = 1$, the chart $\mathbf{x} \mapsto \mathbf{p} + \mathbf{x}$ is a local isometry at $\mathbf{x} = 0$. Define the *Fubini-Study* differential form by

$$\omega_{\mathbf{z}} = \frac{\sqrt{-1}}{2}\partial\bar{\partial}\log\|\mathbf{z}\|^2. \tag{4.5}$$

Expanding the expression above, we get

$$\omega_{\mathbf{z}} = \frac{\sqrt{-1}}{2} \left(\frac{1}{\|\mathbf{z}\|^2} \sum_{j=0}^{n} \mathrm{d}z_j \wedge \mathrm{d}\bar{z}_j - \frac{1}{\|\mathbf{z}\|^4} \sum_{j,k=0}^{n} \bar{z}_j z_k \mathrm{d}z_j \wedge \mathrm{d}\bar{z}_k \right).$$

When (for instance) $\mathbf{z} = \mathrm{e}_0$,

$$\omega_{\mathrm{e}_0} = \frac{\sqrt{-1}}{2} \sum_{j=1}^{n} \mathrm{d}z_j \wedge \mathrm{d}\bar{z}_j.$$

Similarly, if E is any complex vector space, $\mathbb{P}(E)$ is the quotient of E by $\mathbb{C}_\times$. When E admits a norm, the Fubini-Study metric in $\mathbb{P}(E)$ can be introduced in a similar way.

Proposition 4.10.

$$\mathrm{Vol}(\mathbb{P}^n) = \frac{\pi^n}{n!}.$$

Before proving Proposition 4.10, we state and prove the formula for the volume of the sphere. The *Gamma function* is defined by

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} \, \mathrm{d}t.$$

Direct integration gives that $\Gamma(1) = 1$, and integration by parts shows that $\Gamma(r) = (r-1)\Gamma(r-1)$ so that if $n \in \mathbb{N}$, $\Gamma(n) = n - 1!$

Proposition 4.11.

$$\mathrm{Vol}(\mathbb{S}^k) = 2\frac{\pi^{(k+1)/2}}{\Gamma\left(\frac{k+1}{2}\right)}.$$

Proof. By using polar coordinates in $\mathbb{R}^{k+1}$, we can infer the following expression for the integral of the Gaussian normal:

$$\begin{aligned}
\int_{\mathbb{R}^{k+1}} \frac{1}{\sqrt{2\pi}^{k+1}} e^{-\|\mathbf{x}\|^2/2} \, \mathrm{d}V_{\mathbf{x}} &= \int_{S^k} \mathrm{d}S^k(\Theta) \int_0^\infty \frac{R^k}{\sqrt{2\pi}^{k+1}} e^{-R^2/2} \, \mathrm{d}R \\
&= \mathrm{Vol}(S^k) \int_0^\infty \frac{r^{(k-1)/2}}{2\sqrt{\pi}^{k+1}} e^{-r} \, \mathrm{d}r \\
&= \mathrm{Vol}(S^k) \frac{\Gamma\left(\frac{k+1}{2}\right)}{2\sqrt{\pi}^{k+1}}
\end{aligned}$$

The integral on the left is just

$$\left(\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x} \, \mathrm{d}x\right)^{k+1}$$

and from the case $k = 1$, we can infer that it is equal to 1. The proposition then follows for all k. $\square$

Proof of Proposition 4.10. Let $S^{2n+1} \subset \mathbb{C}^{n+1}$ be the unit sphere $|\mathbf{z}| = 1$. The *Hopf fibration* is the natural projection of S^{2n+1} onto $\mathbb{P}^n$. The preimage of any $(z_0 : \cdots : z_n)$ is always a great circle in S^{2n+1}.

We claim that

$$\mathrm{Vol}(\mathbb{P}^n) = \frac{1}{2\pi} \mathrm{Vol}(S^{2n+1}).$$

Since we know that the right-hand-term is $\pi^n/n!$, this will prove the Proposition.

The unitary group $U(n+1)$ acts on $\mathbb{C}^{n+1}_{\neq 0}$ by $Q, \mathbf{x} \mapsto Q\mathbf{x}$. This induces transitive actions in $\mathbb{P}^n$ and S^{2n+1}. Moreover, if $\|\mathbf{x}\| = 1$,

$$H(Q\mathbf{x}) = Q(x_0 : \cdots : x_n)$$

so $DH_{Q\mathbf{x}} = QDH_{\mathbf{x}}$. It follows that the Normal Jacobian $\det(DHDH^*)$ is invariant by $U(n+1)$-action, and we may compute it at a single point, say at e_0. Recall our convention $z_i = x_i + \sqrt{-1}\, y_i$. The tangent space $T_{\mathrm{e}_0} S^n$ has coordinates $y_0, x_1, y_1, \dots, y_n$ while the tangent space $T_{(1:0:\cdots:0)}\mathbb{P}^n$ has coordinates $x_1, y_1, \dots, y_n$. With those coordinates,

$$DH(\mathrm{e}_0) = \begin{bmatrix} 0 & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

(white spaces are zeros). Thus $DH(\mathrm{e}_0)\; DH(\mathrm{e}_0)^*$ is the identity.

The co-area formula (Theorem 4.7) now reads:

$$\begin{aligned} \mathrm{Vol} S^{2n+1} &= \int_{S^{2n+1}} \mathrm{d}S^{2n+1} \\ &= \int_{\mathbb{P}^n} d\mathbb{P}^n(\mathbf{x}) \int_{H^{-1}(\mathbf{x})} |\det(DH(\mathbf{y})\; DH^*(\mathbf{y}))|^{-1} \mathrm{d}S^1(\mathbf{y}) \\ &= 2\pi \mathrm{Vol}(\mathbb{P}^n) \end{aligned}$$

□

We come now to another consequence of Wirtinger's theorem. Let W be a variety (irreducible Zariski closed set) of complex dimension k in $\mathbb{P}^n$. By Lemma 2.31, the intersection of W with a generic plane Π of dimension $n-k$ is precisely d points.

We change coordinates so that Π is the plane $y_{k+1} = \cdots = y_n = 0$. Let $P = \{(y_0 : \cdots : y_k : 0 : \cdots 0)\}$ be a copy of $\mathbb{P}^k$. Then consider the formal sum (k-chain) $W - dP$. This is precisely the boundary of the $k+1$-chain

$$\mathcal{D} = \{(y_0 : \cdots : y_k : ty_{k+1} : \cdots : ty_n) : \mathbf{y} \in W, t \in [0,1]\}.$$

By Wirtinger's theorem (Th. 4.5), $W - dP$ has zero volume. We conclude that

Theorem 4.12. *Let $W \subset \mathbb{P}^n$ be a variety of dimension k and degree d. Then,*

$$\text{Vol } W = d\frac{\pi^k}{k!}.$$

Remark 4.13. Many authors such as [44] divide the Fubini-Study metric by π. This is a neat convention, because it makes the volume of $\mathbb{P}^n$ equal to $1/n!$. However, this conflicts with the notations used in the subject of polynomial equation solving (such as in [20]), so I opt here for maintaining the notational integrity of the subject.

Chapter 5

Reproducing kernel spaces and solution density

5.1 Fewspaces

Let M be an n-dimensional complex manifold. Our main object of study in this book are the systems of equations

$$f_1(\mathbf{x}) = f_2(\mathbf{x}) = \cdots = f_n(\mathbf{x}) = 0,$$

where $f_i \in \mathcal{F}_i$, and $\mathcal{F}_i$ is a suitable Hilbert space whose elements are functions from M to $\mathbb{C}$.

Main examples for M are $\mathbb{C}^n$, $(\mathbb{C}_{\neq 0})^n$, a 'quotient manifold' such as $\mathbb{C}^n/(2\pi\sqrt{-1}\ \mathbb{Z}^n)$, a polydisk $|z_1|, \dots, |z_n| < 1$, or a n-dimensional quasi-affine variety in $\mathbb{C}^n$. Examples of $\mathcal{F}_i$ are the space of polyno-

Gregorio Malajovich, *Nonlinear equations*.
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

mials of degree $\leq d_i$ for a certain d_i, or spaces spanned by a finite collection of arbitrary holomorphic functions.

It may be convenient to consider the f_i's as either given or random. By *random* we mean that the f_i are independently normally distributed random variables with unit variance.

Remark 5.1. The definition and main properties of holomorphic functions on several variables follow, in general lines, the main ideas from one complex variable. The unaware reader may want to read chapter 0 and maybe chapter 1 in [50] before proceeding. Regarding reproducing kernel spaces, a canonical reference is Aronszajn's paper [4]

The aim of this chapter is to define what sort of spaces are 'acceptable' for the problem above. Most of functional analysis deals with spaces that are made large enough to contain certain objects. In contrast, we need to avoid 'large' spaces if we want to *count* roots.

The general theory will include equations on quotient manifolds, such as homogeneous polynomials on projective space. We start with the simpler definition, where the equations are actual functions. (See definition 5.15 for general theory).

Definition 5.2. A *fewnomial space* (or *fewspace* for short) of functions over a complex manifold M is a Hilbert space of holomorphic functions from M to $\mathbb{C}$ such that the following holds. Let $V : M \to \mathcal{F}^*$ denote the *evaluation form* $V(\mathbf{x}) : f \mapsto f(\mathbf{x})$. For any $\mathbf{x} \in M$,

1. $V(\mathbf{x})$ is continuous as a linear form.

2. $V(\mathbf{x})$ is not the zero form.

In addition, we say that the fewspace is *non-degenerate* if and only if, for any $\mathbf{x} \in M$,

3. $P_{V(\mathbf{x})} DV(\mathbf{x})$ has full rank,

where P_W denotes the orthogonal projection onto $W^\perp$. (The derivative is with respect to $\mathbf{x}$). In particular, a non-degenerate fewspace has dimension $\geq n + 1$.

We say that a fewspace $\mathcal{F}$ is $\mathcal{L}^2$ if its elements have finite $\mathcal{L}^2$ norm. In this case the $\mathcal{L}^2$ inner product is assumed.

Example 5.3. Let M be an open connected subset of $\mathbb{C}^n$. *Bergman space* $\mathcal{A}(M)$ is the space of holomorphic functions defined in M with finite $\mathcal{L}^2$ norm. When M is bounded, it contains constant and linear functions, hence M is clearly a non-degenerate fewspace.

Remark 5.4. Condition 1 holds trivially for any finite dimensional fewnomial space, and less trivially for subspaces of Bergman space. (Exercise 5.1). Condition 2 may be obtained by removing points from M.

To each fewspace $\mathcal{F}$ we associate two objects: The *reproducing kernel* $K(\mathbf{x}, \mathbf{y})$ and a possibly degenerate Kähler form ω on M.

Item (1) in the definition makes $V(x)$ an element of the dual space $\mathcal{F}^*$ of $\mathcal{F}$ (more precisely, the 'continuous' dual space or space of continuous functionals). Here is a classical result about Hilbert spaces:

Theorem 5.5 (Riesz-Fréchet)**.** *Riesz Let $\mathbb{H}$ be a Hilbert space. If $\phi \in \mathbb{H}^*$, then there is a unique $\mathbf{f} \in \mathbb{H}$ such that*

$$\phi(\mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle_{\mathbb{H}} \qquad \forall \mathbf{v} \in \mathbb{H}.$$

Moreover, $\|\mathbf{f}\|_{\mathbb{H}} = \|\phi\|_{\mathbb{H}^*}$

For a proof, see [23] Th.V.5 p.81. Riesz-Fréchet representation Theorem allows to identify $\mathcal{F}$ and $\mathcal{F}^*$, whence the Kernel $K(\mathbf{x}, \mathbf{y}) = \overline{(V(\mathbf{x})^*)(\mathbf{y})}$. As a function of $\bar{\mathbf{y}}$, $K(\mathbf{x}, \mathbf{y}) \in \mathcal{F}$ for all $\mathbf{x}$.

By construction, for $f \in \mathcal{F}$,

$$f(\mathbf{y}) = \langle f(\cdot), K(\cdot, \mathbf{y}) \rangle.$$

There are two consequences. First of all,

$$K(\mathbf{y}, \mathbf{x}) = \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y}) \rangle = \overline{\langle K(\cdot, \mathbf{y}), K(\cdot, \mathbf{x}) \rangle} = \overline{K(\mathbf{x}, \mathbf{y})}$$

and in particular, for any fixed $\mathbf{y}$, $\mathbf{x} \mapsto K(\mathbf{x}, \mathbf{y})$ is an element of $\mathcal{F}$. Thus, $K(\mathbf{x}, \mathbf{y})$ is analytic in $\mathbf{x}$ and in $\bar{\mathbf{y}}$. Moreover, $\|K(\mathbf{x}, \cdot)\|^2 = K(\mathbf{x}, \mathbf{x})$.

Secondly, $Df(\mathbf{y})\dot{\mathbf{y}} = \langle f(\cdot), D_{\bar{\mathbf{y}}} K(\cdot, \mathbf{y}) \bar{\dot{\mathbf{y}}} \rangle$ and the same holds for higher derivatives.

Exercise 5.1. Show that V is continuous in Bergman space $\mathcal{A}(M)$. Hint: verify first that for u harmonic and r small enough,

$$\frac{1}{\mathrm{Vol}\ B(\mathbf{p},r)} \int_{B(\mathbf{p},r)} u(\mathbf{z})\ \mathrm{d}\mathbf{z} = u(\mathbf{p}).$$

5.2 Metric structure on root space

Because of Definition 5.2(2), $K(\cdot, y) \neq 0$. Thus, $y \mapsto K(\cdot, y)$ induces a map from M to $\mathbb{P}(\mathcal{F})$. The differential form ω is defined as the pull-back of the Fubini-Study form ω_f of $\mathbb{P}(\mathcal{F})$ by $y \mapsto K(\cdot, y)$.

Recall from (4.5) that The Fubini-Study differential 1-1 form in $\mathcal{F} \setminus \{0\}$ is defined by

$$\omega_f = \frac{\sqrt{-1}}{2} \partial\bar{\partial} \log \|f\|^2$$

and is equivariant by scaling. Its pull-back is

$$\omega_x = \frac{\sqrt{-1}}{2} \partial\bar{\partial} \log K(x,x).$$

When the form ω is non-degenerate for all $x \in M$, it induces a Hermitian structure on M. This happens if and only if the fewspace is a non-degenerate fewspace.

Remark 5.6. If $\mathcal{F}$ is the Bergman space, the kernel obtained above is known as the *Bergman Kernel* and the metric induced by ω as the *Bergman metric.*

Remark 5.7. If $\phi_i(x)$ denotes an orthonormal basis of $\mathcal{F}$ (finite or infinite), then the kernel can be written as

$$K(\mathbf{x},\mathbf{y}) = \sum \phi_i(\mathbf{x})\overline{\phi_i(\mathbf{y})}.$$

Remark 5.8. The form ω induces an element of the cohomology ring $H^*(M)$, namely the operator that takes a $2k$-chain C to $\int_C \omega \wedge \cdots \wedge \omega$.

If $\mathcal{F}$ is a fewspace and $\mathbf{x} \in M$, we denote by $\mathcal{F}_x$ the space $K(\cdot, \mathbf{x})^\perp$ of all $f \in \mathcal{F}$ vanishing at $\mathbf{x}$.

Proposition 5.9. *Let $\mathcal{F}$ be a fewspace. Let $\langle \mathbf{u}, \mathbf{w} \rangle_{\mathbf{x}} = \omega_{\mathbf{x}}(\mathbf{u}, J\mathbf{w})$ be the (possibly degenerate) Hermitian product associated to ω. Then,*

$$\langle \mathbf{u}, \mathbf{w} \rangle_{\mathbf{x}} = \frac{1}{2} \int_{\mathcal{F}_{\mathbf{x}}} \frac{(Df(\mathbf{x})\mathbf{u})\overline{Df(\mathbf{x})\mathbf{w}}}{K(\mathbf{x}, \mathbf{x})} \, \mathrm{d}\mathcal{F}_{\mathbf{x}} \tag{5.1}$$

where $\mathrm{d}\mathcal{F}_{\mathbf{x}} = \frac{1}{(2\pi)^{\dim \mathcal{F}_{\mathbf{x}}}} e^{-\|f\|^2} \mathrm{d}\lambda(f)$ is the zero-average, unit variance Gaussian probability distribution on $\mathcal{F}_{\mathbf{x}}$.

Proof. Let

$$P_{\mathbf{x}} = I - \frac{K(\cdot, \mathbf{x})K(\cdot, \mathbf{x})^*}{K(\mathbf{x}, \mathbf{x})}$$

be the orthogonal projection $\mathcal{F} \to \mathcal{F}_{\mathbf{x}}$. We can write the left-hand-side as:

$$\langle \mathbf{u}, \mathbf{w} \rangle_{\mathbf{x}} \quad = \quad \frac{\langle P_{\mathbf{x}} DK(\cdot, \mathbf{x})\mathbf{u}, P_{\mathbf{x}} DK(\cdot, \mathbf{x})\mathbf{w} \rangle}{K(\mathbf{x}, \mathbf{x})}$$

For the right-hand-side, note that

$$Df(\mathbf{x})\mathbf{u} = \langle f(\cdot), DK(\cdot, \mathbf{x})\mathbf{u} \rangle = \langle f(\cdot), P_{\mathbf{x}} DK(\cdot, \mathbf{x})\mathbf{u} \rangle.$$

Let $\mathbf{U} = \frac{1}{\|K(\cdot,\mathbf{x})\|} P_{\mathbf{x}} DK(\cdot, \mathbf{x})\mathbf{u}$ and $\mathbf{W} = \frac{1}{\|K(\cdot,\mathbf{x})\|} P_{\mathbf{x}} DK(\cdot, \mathbf{x})\mathbf{w}$. Both $\mathbf{U}$ and $\mathbf{W}$ belong to $\mathcal{F}_{\mathbf{x}}$. The right-hand-side is

$$\begin{aligned} \frac{1}{2} \int_{\mathcal{F}_{\mathbf{x}}} \frac{(Df(\mathbf{x})\mathbf{u})\overline{Df(\mathbf{x})\mathbf{w}}}{\|K(\mathbf{x}, \mathbf{x})\|^2} \, \mathrm{d}\mathcal{F}_{\mathbf{x}} &= \frac{1}{2} \int_{\mathcal{F}_{\mathbf{x}}} \langle \mathbf{f}, \mathbf{U} \rangle \overline{\langle f, \mathbf{W} \rangle} \, \mathrm{d}\mathcal{F}_{\mathbf{x}} \\ &= \frac{1}{2} \langle \mathbf{U}, \mathbf{W} \rangle \int_{\mathbb{C}} \frac{1}{2\pi} |z|^2 e^{-|z|^2/2} \, \mathrm{d}z \\ &= \langle \mathbf{U}, \mathbf{W} \rangle \end{aligned}$$

which is equal to the left-hand-side. $\square$

For further reference, we state that:

Lemma 5.10. *The metric coefficients g_{ij} associated to the (possibly degenerate) inner product above are*

$$g_{ij}(\mathbf{x}) = \frac{1}{K(\mathbf{x}, \mathbf{x})} \left(K_{ij}(\mathbf{x}, \mathbf{x}) - \frac{K_{i\cdot}(\mathbf{x}, \mathbf{x}) K_{\cdot j}(\mathbf{x}, \mathbf{x})}{K(\mathbf{x}, \mathbf{x})} \right)$$

with the notation $K_{i\cdot}(\mathbf{x},\mathbf{y}) = \frac{\partial}{\partial x_i} K(\mathbf{x},\mathbf{y})$, $K_{\cdot j}(\mathbf{x},\mathbf{y}) = \frac{\partial}{\partial \bar{y}_j} K(\mathbf{x},\mathbf{y})$ *and* $K_{ij}(\mathbf{x},\mathbf{y}) = \frac{\partial}{\partial x_i} \frac{\partial}{\partial \bar{y}_j} K(\mathbf{x},\mathbf{y})$.

The Fubini 1-1 form is then:

$$\omega = \frac{\sqrt{-1}}{2} \sum_{ij} g_{ij} dz_i \wedge d\bar{z}_j$$

and the volume element is $\frac{1}{n!} \bigwedge_{i=1}^{n} \omega$.

Exercise 5.2. Prove Lemma 5.10.

5.3 Root density

We will deduce the famous theorems by Bézout, Kushnirenko and Bernstein from the statement below. Recall that $n_{\mathcal{K}}(f)$ is the number of isolated zeros of f that belong to $\mathcal{K}$.

Theorem 5.11 (Root density)**.** *root density Let* $\mathcal{K}$ *be a locally measurable set of an* n*-dimensional manifold* M*. Let* $\mathcal{F}_1, \dots, \mathcal{F}_n$ *be fewspaces. Let* $\omega_1, \dots, \omega_n$ *be the induced symplectic forms on* M*. Assume that* $\mathbf{f} = f_1, \dots, f_n$ *is a zero average, unit variance variable in* $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$*. Then,*

$$\mathbb{E}(n_{\mathcal{K}}(\mathbf{f})) = \frac{1}{\pi^n} \int_{\mathcal{K}} \omega_1 \wedge \cdots \wedge \omega_n.$$

Proof of Theorem 5.11. Let $\mathcal{V} \subset \mathcal{F} \times M$, where $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_n$ be the *incidence locus*, $\mathcal{V} \stackrel{\text{def}}{=} \{(\mathbf{f}, x) : \mathbf{f}(x) = 0\}$. (It is a variety when M is a variety). Let $\pi_1 : \mathcal{V} \to \mathcal{F}$ and $\pi_2 : \mathcal{V} \to M$ be the canonical projections.

For each $\mathbf{x} \in M$, denote by $\mathcal{F}_{\mathbf{x}} = \{\mathbf{f} \in \mathcal{F} : \mathbf{f}(\mathbf{x}) = 0\}$. Then $\mathcal{F}_{\mathbf{x}}$ is a linear space of codimension n in $\mathcal{F}$. More explicitly,

$$F_{\mathbf{x}} = K_1(\cdot, \mathbf{x})^{\perp} \times \cdots \times K_n(\cdot, \mathbf{x})^{\perp} \subset \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$$

using the notation K_i for the reproducing kernel associated to $\mathcal{F}_i$.

Let $O \in M$ be an arbitrary particular point, and let $F = \mathcal{F}_O$.

We claim that $(\mathcal{V}, M, \pi_2, F)$ is a vector bundle.

First, we should check that $\mathcal{V}$ is a manifold. Indeed, $\mathcal{V}$ is defined implicitly as $\mathrm{ev}^{-1}(0)$, where $\mathrm{ev}(\mathbf{f}, \mathbf{x}) = \mathbf{f}(\mathbf{x})$ is the evaluation function.

Let $p = (\mathbf{f}, \mathbf{x}) \in \mathcal{V}$ be given. The differential of the evaluation function at p is

$$D\mathrm{ev}(p) : \dot{\mathbf{f}}, \dot{\mathbf{x}} \mapsto D\mathbf{f}(\mathbf{x})\dot{\mathbf{x}} + \dot{\mathbf{f}}(\mathbf{x}).$$

Let us prove that $D\mathrm{ev}(p)$ has rank n.

$$D\mathrm{ev}(p)(\dot{\mathbf{f}}, 0) = \begin{bmatrix} \langle \dot{f}_1(\cdot), K_1(\cdot, \mathbf{x})\rangle_{\mathcal{F}_1} \\ \vdots \\ \langle \dot{f}_n(\cdot), K_n(\cdot, \mathbf{x})\rangle_{\mathcal{F}_n} \end{bmatrix}$$

and in particular, $D\mathrm{ev}(p)(\mathrm{e}_i K_i(\mathbf{x}, \cdot)/K_i(\mathbf{x}, \mathbf{x}), 0) = \mathrm{e}_i$. Therefore 0 is a regular value of ev and hence (Proposition 3.2) $\mathcal{V}$ is an embedded manifold.

Now, we should produce a local trivialization. Let U be a neighborhood of $\mathbf{x}$. Let $i_O : \mathcal{F}_\mathbf{x} \to F$ be a linear isomorphism. For $\mathbf{y} \in U$, we define $i_\mathbf{y} : \mathcal{F}_\mathbf{y} \to \mathcal{F}_\mathbf{x}$ by othogonal projection in each component. The neighborhood U should be chosen so that $i_\mathbf{y}$ is always a linear isomorphism. Explicitly,

$$\begin{aligned} i_\mathbf{y} = I_{\mathcal{F}_1} - &\frac{1}{K_1(\mathbf{x}, \mathbf{x})} K_1(\mathbf{x}, \cdot) K_1(\mathbf{x}, \cdot)^* \oplus \cdots \\ &\oplus I_{\mathcal{F}_n} - \frac{1}{K_n(\mathbf{x}, \mathbf{x})} K_n(\mathbf{x}, \cdot) K_n(\mathbf{x}, \cdot)^* \end{aligned}$$

so $U = \{\mathbf{y} : K_j(\mathbf{y}, \mathbf{x}) \neq 0 \ \forall j\}$.

For $q = (\mathbf{g}, \mathbf{y}) \in \pi_2^{-1}(\mathbf{x})$, set

$$\Phi(q) = (\pi_2(q), i_O \circ i_\mathbf{y} \circ \pi_1(q)).$$

This is clearly a diffeomorphism.

The expected number of roots of $\mathcal{F}$ is

$$\mathbb{E}(n_\mathcal{K}(f)) = \int_\mathcal{V} \chi_{\pi_2^{-1}(\mathcal{K})}(p)(\pi_1^* \mathrm{d}\mathcal{F})(p).$$

Denote by $\mathrm{d}\mathcal{F}$, $\mathrm{d}\mathcal{F}_{\mathbf{x}}$ the zero-average, unit variance Gaussian probability distributions. Note that in $\mathcal{F}_{\mathbf{x}}$, $\pi_1^* \mathrm{d}F = \frac{1}{(2\pi)^n}\mathrm{d}F_{\mathbf{x}}$. The coarea formula for $(\mathcal{V}, M, \pi_2, F)$ (Theorem 4.9) is

$$\mathbb{E}(\#(Z(\mathbf{f}) \cap \mathcal{K})) = \frac{1}{(2\pi)^n} \int_{\mathcal{K}} \mathrm{d}M(\mathbf{x}) \int_{F_{\mathbf{x}}} NJ(\mathbf{f}, i\mathbf{x})^{-2} \mathrm{d}F_{\mathbf{x}}$$

with Normal Jacobian $NJ(\mathbf{f}, \mathbf{x}) = \det(D\pi_2(\mathbf{f}, \mathbf{x}) D\pi_2(\mathbf{f}, \mathbf{x})^*)^{1/2}$.

The Normal Jacobian can be computed by

$$\begin{aligned} NJ(\mathbf{f}, \mathbf{x})^2 &= \det \left(D\mathbf{f}(\mathbf{x})^{-*} \begin{bmatrix} K_1(\mathbf{x}, \mathbf{x}) & & \\ & \ddots & \\ & & K_n(\mathbf{x}, \mathbf{x}) \end{bmatrix} D\mathbf{f}(\mathbf{x})^{-1} \right) \\ &= \frac{\prod K_i(\mathbf{x}, \mathbf{x})}{|\det D\mathbf{f}(\mathbf{x})|^2} \end{aligned}$$

We pick an arbitrary system of coordinates around $\mathbf{x}$. Using Lemma 4.3,

$$|\det D\mathbf{f}(\mathbf{x})|^2 \mathrm{d}M = \bigwedge_{i=1}^{n} \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j} f_i(\mathbf{x}) \overline{\frac{\partial}{\partial x_k} f_i(\mathbf{x})}\ \frac{\sqrt{-1}}{2} \mathrm{d}x_j \wedge d\bar{x}_k$$

Thus,

$$\begin{aligned} &\mathbb{E}(\#(Z(\mathbf{f}) \cap \mathcal{K})) = \\ &\quad = \frac{1}{(2\pi)^n} \int_{\mathcal{K}} \bigwedge_{i=1}^{n} \sum_{jk} \int_{\mathcal{F}_{ix}} \frac{\langle Df(x)\frac{\partial}{\partial x_j}, Df(\mathbf{x})\frac{\partial}{\partial x_k}\rangle}{K_i(\mathbf{x}, \mathbf{x})} \\ &\qquad\qquad \frac{\sqrt{-1}}{2} \mathrm{d}x_j \wedge \mathrm{d}\bar{x}_k\ \mathrm{d}\mathcal{F}_{i\mathbf{x}}(f_i) \\ &\quad = \frac{1}{\pi^n} \int_{\mathcal{K}} \bigwedge_{i=1}^{n} \sum_{jk} \omega_i \left(\frac{\partial}{\partial x_j}, J \frac{\partial}{\partial x_k} \right) \frac{\sqrt{-1}}{2} \mathrm{d}x_j \wedge \mathrm{d}\bar{x}_k \\ &\quad = \frac{1}{\pi^n} \int_{\mathcal{K}} \bigwedge_{i=1}^{n} \omega_i(\mathbf{x}) \end{aligned}$$

using Proposition 5.9. □

5.4 Affine and multi-homogeneous setting

We start by particularizing Theorem 5.11 for the Bézout Theorem setting.

The space $\mathcal{P}_{d_i}$ of all polynomials of degree $\le d_i$ is endowed with the Weyl inner product [85] given by

$$\langle \mathbf{x}^{\mathbf{a}}, \mathbf{x}^{\mathbf{b}} \rangle = \begin{cases} \binom{d_i}{\mathbf{a}}^{-1} & \text{if } \mathbf{a} = \mathbf{b} \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

With this choice, $\mathcal{P}_{d_i}$ is a non-degenerate fewspace with Kernel

$$K(\mathbf{x}, \mathbf{y}) = \sum_{|\mathbf{a}| \le d_i} \binom{d_i}{\mathbf{a}} \mathbf{x}^{\mathbf{a}} \bar{\mathbf{y}}^{\mathbf{a}} = (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^{d_i}$$

The geometric reason behind Weyl's inner product will be explained in the next section. A consequence of this choice is that the metric depends linearly in d_i.

We compute $K_{j\cdot}(\mathbf{x}, \mathbf{x}) = d_j \bar{x}_j K(\mathbf{x}, \mathbf{x})/R^2$ and

$$K_{jk}(\mathbf{x}, \mathbf{x}) = \delta_{jk} d_i K(\mathbf{x}, \mathbf{x})/R^2 + d_i(d_{i-1}) \bar{x}_j x_k / R^4,$$

with $R^2 = 1 + \|\mathbf{x}\|^2$. Lemma 5.10 implies

$$g_{jk} = d_i \left(\frac{1}{R^2} \left(\delta_{jk} - \frac{\bar{x}_j x_k}{R^2} \right) \right),$$

with $R^2 = 1 + \|\mathbf{x}\|^2$. Thus, if ω_i is the metric form of $\mathcal{P}_{d_i}$ and ω_0 the metric form of $\mathcal{P}_1$,

$$\bigwedge_{i=1}^{n} \omega_1 = (\prod_{i=1}^{n} d_i) \bigwedge_{i=1}^{n} \omega_0.$$

Comparing the bounds in Theorem 5.11 for the linear case (degree 1 for all equations) and for $\mathbf{d}$, we obtain:

Corollary 5.12. *Let $\mathbf{f} \in \mathcal{P}_{\mathbf{d}} = \mathcal{P}_{d_1} \times \cdots \times \mathcal{P}_{d_n}$ be a zero average, unit variance variable. Then,*

$$\mathbb{E}(n_{\mathbb{C}^n}(\mathbf{f})) = \prod d_i$$

Remark 5.13. Mario Wschebor pointed out that if one could give a similar expression for the variance (which is zero) it would be possible to deduce and 'almost everywhere' Bézout's theorem from a purely probabilistic argument.

Now, let $\mathcal{F}_i$ is the space of polynomials with degree d_{ij} in the j-th set of variables. We write $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_s)$ for $\mathbf{x}_i \in \mathbb{C}^{n_i}$, and the same convention holds for multi-indices.

The inner product will be defined by:

$$\langle \mathbf{x}_1^{\mathbf{a}_1} \dots \mathbf{x}_s^{\mathbf{a}_n}, \mathbf{x}_1^{\mathbf{b}_1} \dots \mathbf{x}_s^{\mathbf{b}_n} \rangle = \frac{\delta_{\mathbf{a}_1\mathbf{b}_1} \cdots \delta_{\mathbf{a}_s\mathbf{b}_s}}{\binom{d_{i1}}{\mathbf{a}_1} \cdots \binom{d_{is}}{\mathbf{a}_s}} \tag{5.3}$$

The integral kernel is now

$$K(\mathbf{x}, \mathbf{y}) = (1 + \langle \mathbf{x}_1, \mathbf{y}_1 \rangle)^{d_{i1}} \cdots (1 + \langle \mathbf{x}_s, \mathbf{y}_s \rangle)^{d_{is}}$$

We need more notations: the j-th variable belongs to the $l(j)$-th group, and $R_l^2 = 1 + \|\mathbf{x}_l\|^2$.

With this notations,

$$\begin{aligned}
K_{j\cdot}(\mathbf{x}, \mathbf{x}) &= d_{l(j)} \frac{\bar{\mathbf{x}}_j K(\mathbf{x}, \mathbf{x})}{R_{l(j)}^2} \\
K_{jk}(\mathbf{x}, \mathbf{x}) &= \delta_{jk} d_{l(j)} \frac{K(\mathbf{x}, \mathbf{x})}{R_{l(j)}^2} + d_{l(j)}(d_{l(k)} - \delta_{l(j)l(k)}) \frac{\bar{\mathbf{x}}_j \mathbf{x}_k}{R_{l(j)}^2 R_{l(k)}^2} \\
g_{jk} &= d_{l(j)} \left(\frac{\delta_{jk}}{R_{l(j)}^2} - \delta_{l(j)l(k)} \frac{\bar{\mathbf{x}}_j \mathbf{x}_k}{R_{l(j)}^2 R_{l(k)}^2} \right)
\end{aligned}$$

Recall that ω_i is the symplectic form associated to $\mathcal{F}_i$. We denote by ω_{jd} the form associated to the polynomials that have degree $\leq d$ in the j-th group of variables, and are independent of the other variables. From the calculations above,

$$\omega_i = \omega_{1d_1} + \cdots + \omega_{sd_s} = d_{i1}\omega_{11} + \cdots + d_{is}\omega_{s1}$$

Hence,

$$\bigwedge \omega_i = \bigwedge d_{i1}\omega_{11} + \cdots + d_{is}\omega_{s1}.$$

This is a polynomial in variables $Z_1 = \omega_{11}, \dots, Z_s = \omega_{ss}$. Notice that $Z_1 \wedge Z_2 = Z_2 \wedge Z_1$ so we may drop the wedge notation. Moreover, $Z_i^{n_i+1} = 0$. Hence, only the monomial in $Z_1^{n_1} Z_2^{n_2} \cdots Z_s^{n_s}$ may be nonzero.

Corollary 5.14. *Let B be the coefficient of $Z_1^{n_1} Z_2^{n_2} \cdots Z_s^{n_s}$ in*

$$\prod (d_{i1} Z_1 + \cdots + d_{is} Z_s).$$

Let $\mathbf{f} \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$ be a zero average, unit variance variable. Then,

$$\mathbb{E}(n_{\mathbb{C}^n}(\mathbf{f})) = B$$

Proof. By Theorem 5.11,

$$\begin{aligned} \mathbb{E}(n_{\mathbb{C}^n}(\mathbf{f})) &= \frac{1}{\pi^n} \int_{\mathbb{C}^n} \bigwedge \omega_i \\ &= \frac{B}{\pi^n} \int_{\mathcal{K}} \underbrace{\omega_{11} \wedge \cdots \wedge \omega_{11}}_{n_1 \text{times}} \wedge \cdots \wedge \underbrace{\omega_{s1} \wedge \cdots \wedge \omega_{s1}}_{n_s \text{times}} \end{aligned}$$

In order to evaluate the right-hand-term, let $\mathcal{G}_j$ be the space of affine polynomials on the j-th set of variables. Its associated symplectic form is ω_{i1}.

A generic polynomial system in

$$\mathcal{G} = \underbrace{\mathcal{G}_1 \times \cdots \mathcal{G}_1}_{n_1 \text{times}} \times \cdots \times \underbrace{\mathcal{G}_s \times \cdots \mathcal{G}_s}_{n_s \text{times}}$$

is just a set of decoupled linear systems, hence has one root. Hence,

$$1 = \frac{1}{\pi^n} \int_{\mathbb{C}^n} \underbrace{\omega_{11} \wedge \cdots \wedge \omega_{11}}_{n_1 \text{times}} \wedge \cdots \wedge \underbrace{\omega_{s1} \wedge \cdots \wedge \omega_{s1}}_{n_s \text{times}}$$

and the expected number of roots of a multi-homogeneous system is B. □

5.5 Compactifications

The Corollaries in the section above allow to prove Bézout and Multi-Homogeneous Bézout theorems, if one argues as in Chapter 1 that the set of systems with root 'at infinity' is contained in a non-trivial Zariski closed set. It is more geometric to compactify $\mathbb{C}^n$ and to homogenize all polynomials.

In the homogeneous setting, the manifold of roots is projective space $\mathbb{P}^n$. In the multi-homogeneous setting, the manifold of roots is $\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_s}$. Both of them are connected and compact. Note that

- Polynomials are **not** ordinary functions of $\mathbb{P}^n$ or multi-projective spaces, and
- The only global holomorphic functions from a compact connected manifold are constant.

Let $\mathcal{H}_d$ denote the space of homogeneous $n+1$-variate polynomials. It is a fewspace associated to the manifold $\mathbb{C}^{n+1} \setminus 0$. The complex multiplicative group $\mathbb{C}_\times$ acts on the manifold $\mathbb{C}^{n+1}$ as

$$\mathbf{x} \overset{\lambda}{\rightsquigarrow} \lambda \mathbf{x}$$

A property of this action is that f vanishes at $\mathbf{x}$ if and only if it vanishes at all the orbit of $\mathbf{x}$.

Definition 5.15. Let M be an m-dimensional complex manifold, and let a group H act on M so that M/H is an n-dimensional complex manifold.

A *fewnomial space* (or *fewspace* for short) of equations over the quotient M/H is a Hilbert space of holomorphic functions from M to $\mathbb{C}$ such that the following holds. Let $V : M \to \mathcal{F}^*$ denote the *evaluation form* $V(x) : f \mapsto f(x)$. For any $x \in M$,

1. $V(x)$ is continuous as a linear form.
2. $V(x)$ is not the zero form.
3. There is a multiplicative character of H, denoted χ, such that for every $x \in M$, for every $h \in H$ and for every $f \in \mathcal{F}$,

$$f(hx) = \chi(h) f(x).$$

In addition, the fewspace is said to be non-degenerate if and only if, for each $x \in M$,

4. the kernel of $P_{V(x)}DV(x)$ is tangent to the group action,

where P_W denotes the orthogonal projection onto $W^{\perp}$. (The derivative is with respect to x).

Example 5.16. $\mathcal{H}_d$ is a non-degenerate fewspace of equations for $\mathbb{P}^n = \mathbb{C}^{n+1}/\mathbb{C}_{\times}$, with $\chi(h) = h^d$.

Example 5.17. Let $n = n-1+\cdots+n_s-s$ and $\Omega = \{\mathbf{x} \in \mathbb{C}^{n+s} : \mathbf{x}_i = 0 \text{ for some } i\}$. In the multi-homogeneous setting, the homogenization group $(\mathbb{C}_{\times})^s$ acts on $M = \mathbb{C}^{n+s} \setminus \Omega$ by

$$(\mathbf{x}_1, \dots, \mathbf{x}_s) \overset{\mathbf{h}}{\rightsquigarrow} (h_1\mathbf{x}_1, \dots, h_s\mathbf{x}_s)$$

and the multiplicative character for $\mathcal{F}_i$ is

$$\chi_i(\mathbf{h}) = h_1^{d_{i1}} h_2^{d_{i2}} \cdots h_s^{d_{is}}$$

By tracing through the definitions, we obtain:

Lemma 5.18. *Let $\mathcal{F}$ be a fewspace of equations on M/H with character χ. Then,*

$$\begin{aligned} V(h\mathbf{x}) &= \chi(h)V(\mathbf{x}) \\ K(h\mathbf{x}, h\mathbf{y}) &= |\chi(h)|^2 K(\mathbf{x}, \mathbf{y}) \\ h^*\omega &= \omega. \end{aligned}$$

In particular, ω induces a form on M/H.

All this may be summarized as a principal bundle morphism:

$$\begin{array}{ccccc}
H & & \xrightarrow{\chi} & & \mathbb{C}_{\times} \\
 & \circlearrowleft & & & & \circlearrowleft \\
 & M & & \xrightarrow{V} & & \mathcal{F}^* \setminus \{0\} \\
 & \downarrow & & & & \downarrow \\
 & M/H & & \xrightarrow{v} & & \mathbb{P}(\mathcal{F}^*)
\end{array}$$

This diagram should be understood as a commutative diagram. The down-arrows are just the canonical projections.

The quotient M/H is endowed with the possibly degenerate Hermitian metric given by $\omega_{\mathcal{F}}$.

Remark 5.19. Given f in a fewspace $\mathcal{F}$ of equations, define $E_f = \{(\mathbf{x}, f(\mathbf{x})) : \mathbf{x} \in M\}$. Then E_f is invariant by $H \times \mathbb{C}_\times$-action. Therefore

$$(E_f/(H \times \mathbb{C}_\times, M/H, \pi, \mathbb{C})$$

is a line bundle. In this sense, solving a system of polynomial equations is the same as finding simultaneous zeros of n line bundles.

Theorem 5.20 (Homogeneous root density)**.** *Let $\mathcal{K}$ be a locally measurable set of M/H. Let $\mathcal{F}_1, \dots, \mathcal{F}_n$ be fewspaces on the quotient M/H, with $\omega_1, \dots, \omega_n$ be the induced (possibly degenerate) symplectic forms. Assume that $\mathbf{f} = f_1, \dots, f_n$ is a zero average, unit variance variable in $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. Then,*

$$\mathbb{E}(n_{\mathcal{K}}(\mathbf{f})) = \frac{1}{\pi^n} \int_{\mathcal{K}} \omega_1 \wedge \cdots \wedge \omega_n.$$

Proof. There is a covering U_α of M/H such that each U_α may be diffeomorphically embedded in M Now, the $\mathcal{F}_i$ are fewspaces of functions in U_α.

Write $\mathcal{K}$ as a disjoint union of sets K_α where each K_α is measurable and contained in U_α. By Theorem 5.11,

$$\mathbb{E}(n_{\mathcal{K}_\alpha}(\mathbf{f})) = \frac{1}{\pi^n} \int_{\mathcal{K}_\alpha} \omega_1 \wedge \cdots \wedge \omega_n.$$

Then we add over all the α's. □

It is time to explain the choice of the inner product (5.2) and (5.3). Suppose that we want to write $f \in \mathcal{H}_d$ as a symmetric tensor. Then,

$$f(\mathbf{x}) = \sum_{1 \le x_{j_1}, \dots, x_{j_d} \le n} T_{j_1, \dots, j_d} x_{j_1} x_{j_2} \cdots x_{j_{i_d}}$$

with

$$T_{j_1, \dots, j_d} = \frac{1}{\binom{d}{\mathrm{e}_{j_1} + \cdots + \mathrm{e}_{j_d}}} f_{\mathrm{e}_{j_1} + \cdots + \mathrm{e}_{j_d}}.$$

The Frobenius norm of T is precisely $\|T\|_F = \|f\|$. The reader shall check (Exercise 5.3) that $\|T\|_F$ is invariant for the $U(n+1)$-action on $\mathbb{C}^{n+1}$.

As a result, the Weyl inner product is invariant under unitary action $f \rightsquigarrow f \circ U^*$ and moreover,

$$K(U\mathbf{x}, U\mathbf{y}) = K(\mathbf{x}, \mathbf{y}).$$

Hence ω is 'equivariant' by $U(n+1)$. This action therefore generates an action in quotient space $\mathbb{P}^n$. Moreover, $U(n+1)$ acts *transitively* on $\mathbb{P}^n$, meaning that for all $\mathbf{x}, \mathbf{y} \in \mathbb{P}^n$ there is $U \in U(n+1)$ with $\mathbf{y} = U\mathbf{x}$.

In this sense, $\mathbb{P}^n$ is said to be 'homogeneous'. The formal definition states that a homogeneous manifold is a manifold that is quotient of two Lie groups, and $\mathbb{P}^n = U(n+1)/(U(1) \times U(n))$.

We can now mimic the argument given for Theorem 1.3

Theorem 5.21. *Let $\mathcal{F}_1, \dots, \mathcal{F}_n$ be fewspaces of equations on M/H. Suppose that*

1. *M/H is compact.*
2. *A group G acts transitively on M/H, in such a way that the induced forms ω_i on M/H are G-equivariant.*
3. *Assume furthermore that the set of regular values of $\pi_1 : \mathcal{V} \to \mathcal{F}$ is path-connected.*

Let $\mathbf{f} = f_1, \dots, f_n \in \mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_n$. Then,

$$n_{M/H}(\mathbf{f}) \leq \frac{1}{\pi^n} \int_{M/H} \omega_1 \wedge \dots \wedge \omega_n$$

with equality almost everywhere.

Proof. Let Σ be the set of critical values of $\mathcal{F}$. From Sard's Theorem it has zero measure.

For all $\mathbf{f}, \mathbf{g} \in \mathcal{F} \setminus \Sigma$, we claim that $n_M(\mathbf{f}) \geq n_M(\mathbf{g})$. Indeed, there is a path $(\mathbf{f}_t)_{t \in [0,1]}$ in $\mathcal{F} \setminus \Sigma$. By the inverse function theorem and because M/H is compact, each root of $\mathbf{f}$ can be continued to a root of $\mathbf{g}$.

It follows that $n_M(\mathbf{f})$ is independent of $\mathbf{f} \in \mathcal{F} \setminus \Sigma$. Thus with probability one,

$$n_{\mathcal{M}}(\mathbf{f}) = \frac{1}{\pi^n} \int_{\mathcal{M}} \omega_1 \wedge \dots \wedge \omega_n.$$

Corollary 3.9 completes the proof. □

We can prove Bézout's Theorem by combining Theorem 5.21 with Corollary 5.12. The multi-homogeneous Bézout theorem is more intricate and implies Bézout's theorem, so we write down a formal proof of it instead.

Proof of Theorem 1.5. Let $H = (\mathbb{C}_{\times})^s$ act in $\mathbb{C}^{n+s} \setminus V^{-1}(0)$ as explained above. Then $\mathcal{H}_{d_1}, \dots, \mathcal{H}_{d_n}$ are fewspaces of equations on

$$\mathbb{C}^{n+s}/H = \mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_s}$$

which is compact. The group $U(n_1+1) \times \cdots \times U(n_s+1)$ acts transitively and preserves the symplectic forms.

It remains to prove that the set of critical points of π_1 is contained in a Zariski closed set. We proceed by induction in s.

The case $s = 1$ (Bézout's theorem setting) follows directly from the Main Theorem of Elimination Theory (Th.2.33) applied to the systems $f_1(\mathbf{x}) = 0, \cdots, f_n(\mathbf{x}) = 0, g_j(\mathbf{x}) = 0$ where $g(\mathbf{x})$ is the determinant of $Df(\mathbf{x})_{\mathrm{e}_j^\perp}$. According to that theorem, $\Sigma_j = \{\mathbf{f} : \exists \mathbf{x} \in \mathbb{P}^n : f_1(\mathbf{x}) = \cdots = f_n(\mathbf{x}) = g_j(\mathbf{x}) = 0\}$ is Zariski closed. Hence $\Sigma = \cap \Sigma_j$ is Zariski closed.

For the induction step, we assume that the induction hypothesis above was established up to stage $s-1$. As before,

$$\Sigma_j = \{(\mathbf{f}, \mathbf{x}_1, \dots, \mathbf{x}_{s-1}) : \exists \mathbf{x} \in \mathbb{P}^{n_s} : f_1(\mathbf{x}) = \cdots = f_n(\mathbf{x}) = g_J(\mathbf{x}) = 0\}$$

with $g_J(\mathbf{x}_s) = \det Df(\mathbf{x})_J$ and J is a coordinate space of $\mathbb{C}^{n+s}$ of dimension n. By Theorem 2.33 $\Sigma' = \cap \Sigma'_J$ is a Zariski closed subset of $\mathcal{F} \times \mathbb{C}^{n_1 + \cdots + n_{s-1} + s - 1}$. Its defining polynomial(s) are homogeneous in $\mathbf{x}_1, \dots, \mathbf{x}_s$. Then by induction, we know that the set Σ of all $\mathbf{f}$ such that those defining polynomials vanish for some $\mathbf{x}_1, \dots, \mathbf{x}_{s-1}$ is Zariski closed.

As it is a zero-measure set, $\Sigma \subsetneq \mathcal{F}$. Thus, the set $\mathcal{F} \setminus \Sigma$ of regular values of π_1 is path-connected. Theorem 1.5 is now a consequence of Theorem 5.21 together with Corollary 5.14. □

Exercise 5.3. The Frobenius norm for tensors $T^{i_1\cdots i_p}_{j_1\cdots j_q}$ is

$$\|T\|_F = \sqrt{\sum_{i_1,\cdots,j_q=1}^{n} |T^{i_1\cdots i_p}_{j_1\cdots j_q}|^2}$$

The unitary group acts on the variable j_1 by composition:

$$T^{i_1\cdots i_p}_{j_1\cdots j_q} \overset{U}{\rightsquigarrow} \sum_{k=1}^{N} T^{i_1\cdots i_p}_{k\cdots j_q} U^k_{j_1}.$$

Show that the Frobenius norm is invariant for the $U(n)$-action. Deduce that it is invariant when $U(n)$ acts simultaneously on all lower (or upper) indices. Deduce that Weyl's norm is invariant by unitary action $f \rightsquigarrow f \circ U$.

Exercise 5.4. This is another proof that the inner product defined in (5.2) is $U(n+1)$-invariant. Show that for all $f \in \mathcal{H}_d$,

$$\|f\|^2 = \frac{1}{2^d d!} \int_{\mathbb{C}^{n+1}} \|f(\mathbf{x})\|^2 \ \frac{1}{(2\pi)^{n+1}} e^{-\|\mathbf{x}\|^2/2} \ \mathrm{d}V(\mathbf{x}).$$

The integral is the $\mathcal{L}^2$ norm of $\mathbf{f}$ with respect to zero average, unit variance probability measure. Conclude that $\|f\|$ is invariant.

Exercise 5.5. Show that if $\mathcal{F} = \mathcal{H}_d$, then the induced norm defined in Lemma 5.10 is d times the Fubini-Study metric. Hint: assume without loss of generality that $\mathbf{x} = \mathrm{e}_0$.

Chapter 6

Exponential sums and sparse polynomial systems

The objective of this chapter is to prove Kushnirenko's and Bernstein's theorems. We will need a few preliminaries of convex geometry.

6.1 Legendre's transform

Through this section, let $\mathbb{E}$ be a Hilbert space.

Definition 6.1. Recall that a subset U of $\mathbb{E}$ is *convex* if and only if, for all $\mathbf{v}_0, \mathbf{v}_1 \in U$ and for all $t \in [0, 1]$, $(1-t)\mathbf{v}_0 + t\mathbf{v}_1 \in U$.

Gregorio Malajovich, *Nonlinear equations*.
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

Lemma 6.2. *A set U is convex if and only if U is an intersection of closed half-spaces.*

In order to prove this Lemma we need a classical fact about Hilbert spaces:

Lemma 6.3. *Let U be a convex subset in a Hilbert space, and let $\mathbf{p} \notin U$. Then there is a hyperplane separating U and $\mathbf{p}$, namely*

$$\mathbf{x} \in U \Rightarrow \alpha(\mathbf{x}) < \alpha(\mathbf{p})$$

where $\alpha \in \mathbb{E}^*$.

This is a consequence of the Hahn-Banach theorem, see [23] Lemma I.3 p.6.

Proof of Lemma 6.2. Assume that U is convex. Then, let $\mathcal{S}$ be the collection of all half-spaces $H_{\alpha,\alpha_0} = \{\alpha(\mathbf{x}) - \alpha_0 \geq 0\}$, $\alpha \in \mathbb{E}^*, \alpha_0 \in \mathbb{R}$, such that $U \subseteq H_{\alpha,\alpha_0}$.

Clearly

$$U \subseteq \bigcap_{\alpha,\alpha_0 \in \mathcal{S}} H_{\alpha,\alpha_0}.$$

Equality follows from Lemma 6.3.

The reciprocal is easy and left to the reader. □

Definition 6.4. A function $f : U \subseteq \mathbb{E} \to \mathbb{R}$ is *convex* if and only if its *epigraph*

$$\mathrm{Epi}_f = \{(\mathbf{x}, y) : f(\mathbf{x}) \leq y\}$$

is convex.

Note that from this definition, the domain of a convex function is always convex. In this book we shall convention that a convex function has non-empty domain.

Definition 6.5. The Legendre-Fenchel transform of a function $f : U \subseteq \mathbb{E} \to \mathbb{R}$ is the function $f^* : U^* \subseteq \mathbb{E}^* \to \mathbb{R}$ given by

$$f^*(\alpha) = \sup_{\mathbf{x} \in U} \alpha(\mathbf{x}) - f(\mathbf{x}).$$

Proposition 6.6. *Let $f : \mathbb{E} \to \mathbb{R}$ be given. Then,*

1. *f^* is convex. In part. U^* is convex.*
2. *For all $\mathbf{x} \in U, \alpha \in U^*$,*
$$f^*(\alpha) + f(\mathbf{x}) \geq \alpha(\mathbf{x}).$$
3. *If furthermore f is convex then $f^{**}_{|U} \equiv f$.*

Proof. Let $(\alpha_0, \beta_0), (\alpha_1, \beta_1) \in \mathrm{Epi}_{f^*}$ This means that $\beta_i \geq f^*(\alpha_i)$, $i = 1, 2$ so
$$\beta_i \geq \alpha_i(\mathbf{x}) - f(\mathbf{x}) \ \forall \mathbf{x} \in U.$$

Hence, if $t \in [0, 1]$,
$$(1-t)\beta_0 + t\beta_1 \geq ((1-t)\alpha_0 + t\alpha_1)(\mathbf{x}) - f(\mathbf{x}) \ \forall \mathbf{x} \in U$$
and $((1-t)\alpha_0 + t\alpha_1, (1-t)\beta_0 + t\beta_1) \in \mathrm{Epi}_{f^*}$.

Item 2 follows directly from the definition.

Let $\mathbf{x} \in U$. By Lemma 6.3, there is a separating hyperplane between $(\mathbf{x}, f(\mathbf{x}))$ and the interior of Epi_f. Namely, there are α, β so that for all $\mathbf{y} \in U$, for all z with $z > f(\mathbf{y})$,
$$\alpha(\mathbf{y}) + \beta z < \alpha(\mathbf{x}) + \beta f(x).$$
Since $\mathbf{x} \in U$, $\beta < 0$ and we may scale coefficients so that $\beta = -1$. Under this convention,
$$\alpha(\mathbf{x} - \mathbf{y}) - f(\mathbf{x}) + f(\mathbf{y}) \geq 0$$
with equality when $\mathbf{x} = \mathbf{y}$. Thus,
$$\begin{aligned} f^{**}(\mathbf{x}) &= \sup_{\alpha} \alpha(\mathbf{x}) - f^*(\alpha) \\ &= \sup_{\alpha} \inf_{\mathbf{y}} \alpha(\mathbf{x} - \mathbf{y}) + f(\mathbf{y}) \\ &= \sup_{\alpha} f(\mathbf{x}) \\ &= f(\mathbf{x}) \end{aligned}$$

□

6.2 The momentum map

Let $M = \mathbb{C}^n/(2\pi\sqrt{-1}\ \mathbb{Z}^n)$. Let $A \subset \mathbb{Z}^n_{\geq 0} \subset (\mathbb{R}^n)^*$ be finite, and let $\mathcal{F}_A = \{f : \mathbf{x} \mapsto f(\mathbf{x}) = \sum_{\mathbf{a}\in A} f_{\mathbf{a}} e^{\mathbf{ax}}\}$.

If we set $z_i = e^{x_i}$, then elements of $\mathcal{F}_A$ are actually polynomials in $\mathbf{z}$. (The roots that have a real negative coordinate z_i are irrelevant for this section).

We assume an inner product on $\mathcal{F}_A$ of the form.

$$\langle e^{\mathbf{ax}}, e^{\mathbf{bx}} \rangle = \begin{cases} c_{\mathbf{a}} & \text{if } \mathbf{a} = \mathbf{b} \\ 0 & \text{otherwise} \end{cases}$$

where the *variances* $c_{\mathbf{a}}$ are arbitrary.

In this context,

$$K(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{a}\in A} c_{\mathbf{a}}^{-1} e^{\mathbf{a}(\mathbf{x}+\bar{\mathbf{y}})}.$$

Notice the property that for any purely imaginary vector g, $K(\mathbf{x}+g, \mathbf{y}+g) = K(\mathbf{x}, \mathbf{y})$. In particular, $K_{i\cdot}(\mathbf{x}, \mathbf{x})$ is always real. This is a particular case of *toric action* which arises in a more general context. Properly speaking, the n-torus $(\mathbb{R}^n/2\pi\mathbb{R}^n, +)$ acts on M by $\mathbf{x} \overset{\theta}{\mapsto} \mathbf{x} + i\theta)$.

The *momentum map* $m : M \to (\mathbb{R}^n)^*$ for this action is defined by

$$m_{\mathbf{x}} = \frac{1}{2} \mathrm{d} \log K(\mathbf{x}, \mathbf{x}) \tag{6.1}$$

The terminology *momentum* arises because it corresponds to the angular momentum of the Hamiltonian system

$$\dot{q}_i = \frac{\partial}{\partial p_i} H(\mathbf{x}) \qquad\qquad \dot{p}_i = -\frac{\partial}{\partial q_i} H(\mathbf{x})$$

where $x_i = p_i + \sqrt{-1} q_i$ and $H(\mathbf{x}) = m_{\mathbf{x}} \cdot \xi$. The definition for an arbitrary action is more elaborate, see [75].

Proposition 6.7. *1. The image $\{m_{\mathbf{x}} : \mathbf{x} \in M\}$ of m is the the interior $\mathring{\mathcal{A}}$ of the convex hull $\mathcal{A}$ of A.*

2. *The map $m : M \rightarrow \mathcal{A} \subset (\mathbb{R}^n)^*$ is volume preserving, in the sense that for any measurable $U \subseteq \mathcal{A}$,*

$$\mathrm{Vol}(m^{-1}(U)) = \pi^n \mathrm{Vol}(U)$$

Proof. We compute explicitly

$$m(\mathbf{x}) = \frac{\sum_{\mathbf{a} \in A} \mathbf{a} c_{\mathbf{a}} e^{2\mathbf{a} \ \mathrm{re}(\mathbf{x})}}{\sum_{\mathbf{a} \in A} c_{\mathbf{a}} e^{2\mathbf{a} \ \mathrm{re}(\mathbf{x})}}$$

where we assimilate $\mathbf{a}$ to $a_1 \mathrm{d}q_1 + \cdots + a_n \mathrm{d}q_n$.

Every vertex of $\mathcal{A}$ is in the closure of the image of m. Indeed, let $\mathbf{a} \in (\mathbb{R}^n)^*$ be a vertex of $\mathcal{A}$ and let $p \in \mathbb{R}^n$ be a vector such that $\mathbf{a}p \geq \mathbf{a}'p$ for all $\mathbf{a}' \neq \mathbf{a}$. In that case, $m(e^{tp}) \rightarrow \mathbf{a}$ when $t \rightarrow \infty$.

Also, it is clear from the formula above that the image of m is a subset of $\mathcal{A}$.

The will prove that the image of m is a convex set as follows: $f(\mathbf{x}) = -m(\mathbf{x}) = -\frac{1}{2} \log K(\mathbf{x}, \mathbf{x})$ is a convex function. Its Legendre transform is

$$f^*(\alpha) = \alpha \mathbf{x} + m(\mathbf{x})$$

Therefore, the domain of f^* is $\{-m(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$ which is convex (Proposition 6.6).

Now, we consider the map $\hat{m}$ from M to $\mathcal{A} \times \mathbb{R}^n \subset \mathbb{C}^n / 2\pi\mathbb{Z}^n$, given by

$$\hat{m}(\mathbf{x} + \sqrt{-1}\mathbf{y}) = m(\mathbf{x}) + \sqrt{-1}\mathbf{y}.$$

The canonical symplectic form in $\mathbb{C}^n$ is $\eta = \mathrm{d}x_1 \wedge \mathrm{d}y_1 + \cdots + \mathrm{d}x_n \wedge \mathrm{d}y_n$. We compute its pull-back $\hat{m}^*\eta$:

$$\hat{m}^*\eta = \eta(D\hat{m}\mathbf{u}, D\hat{m}\mathbf{v})$$

Differentiating,

$$D\hat{m}(\mathbf{x} + \sqrt{-1}\mathbf{y}) : \dot{\mathbf{x}} + \sqrt{-1}\dot{\mathbf{y}} \mapsto D^2(\frac{1}{2} \log K(\mathbf{x}, \mathbf{x}))\dot{\mathbf{x}} + \sqrt{-1}\dot{\mathbf{y}}$$

Thus,

$$
\begin{aligned}
\hat{m}^*\eta(\mathbf{u},\mathbf{v}) &= D^2(\frac{1}{2}\log K(\mathbf{x},\mathbf{x}))(\mathrm{re}(\mathbf{u}),\mathrm{im}(\mathbf{v})) \\
&\quad -D^2(\frac{1}{2}\log K(\mathbf{x},\mathbf{x}))(\mathrm{im}(\mathbf{u}),\mathrm{re}(\mathbf{v})) \\
&= 2^n\langle \mathbf{u}, J\mathbf{v}\rangle_{\mathbf{x}+\sqrt{-1}\mathbf{y}} \\
&= 2^n\omega_{\mathbf{x}+\sqrt{-1}\mathbf{y}}(\mathbf{u},\mathbf{v})
\end{aligned}
$$

using Lemma 5.10. □

As a consequence toward the proof of Kushnirenko's theorem, we note that

Proposition 6.8.

$$\mathbb{E}(n_M(\mathbf{f})) = n!\mathrm{Vol}(\mathcal{A})$$

Proof. The preimage $M = m^{-1}(\mathcal{A})$ has volume $\pi^n\mathrm{Vol}(\mathcal{A})$. Theorem 5.11) implies then that expected number of roots is

$$\mathbb{E}(n_M(\mathbf{f})) = \frac{1}{\pi^n}\int_M \bigwedge_{i=1}^{n} \omega = \frac{n!}{\pi^n}\mathrm{Vol}(M) = n!\mathrm{Vol}(\mathcal{A}).$$

□

6.3 Geometric considerations

To achieve the proof of the Kushnirenko theorem, we still need to prove that the number of roots is generically constant. The following step in the proof of that fact was used implicitly in other occasions:

Lemma 6.9. *Let M be a holomorphic manifold, and $\mathcal{F} = \mathcal{F}_1\times\cdots\times\mathcal{F}_n$ be a product of fewspaces. Let $\mathcal{V} \subset \mathcal{F}\times M$ and let $\pi_1 : \mathcal{V}\to\mathcal{F}$ and $\pi_2 : \mathcal{V}\to M$ be the canonical projections.*

Assume that $(\mathbf{f}_t)_{t\in[0,1]}$ is a smooth path in $\mathcal{F}$ and that for all t, $\mathbf{f}_t$ is a regular value for f_t. Let $v_0 \in \pi_1^{-1}(f_0)$.

Then, the path $\mathbf{f}_t$ can be lifted to a path v_t with $\pi_1(v_t) = \mathbf{f}_t$ in an interval I such that either $I = [0,1]$ or $I = [0,\tau)$, $\tau < 1$ and $\pi_2(v_t)$ diverges for $t\to\tau$.

Proof. The implicit function theorem guarantees that (v_t) is defined for some interval $(0, \tau)$. Take τ maximal with that property. If $\tau < 1$ and v_t converges for $t \to \tau$, then we could apply the implicit function theorem at $t = \tau$ and increase τ. Therefore v_t diverges, and since the first projection is smooth $\pi_2(v_t)$ diverges. □

It would be convenient to have a compact M. Recall that in the Kushnirenko setting, M can be thought as a subset of $\mathbb{P}(\mathcal{F}_A)$ (while $\mathcal{F} = \mathcal{F}_A^n$). More precisely,

$$\begin{array}{rccc} K: & M & \to & \mathcal{F}, \\ & \mathbf{x} & \mapsto & K(\cdot, \bar{\mathbf{x}}) \end{array}$$

is an embedding and an isometry into $\mathbb{P}(\mathcal{F}_A)$. Let $\bar{M}$ be the ordinary closure of $\mathbf{K}(M)$. In this setting, it is the same as the Zariski closure. The set $\bar{M}$ is an example of a *toric variety.*

Can we then replace M by $\bar{M}$ in the theory? The answer is **not always**.

Example 6.10. Let

$$A = \{0, \mathrm{e}_1, \mathrm{e}_2, \mathrm{e}_3, \mathrm{e}_1 + \mathrm{e}_2\} \subset \mathbb{Z}^3$$

Then $\bar{M}$ has a singularity at $(0 : 0 : 0 : 1 : 0)$ and hence is not a manifold.

This phenomenon can be averted if the polytope A satisfies a geometric-combinatorial condition [34]. Here, however, we need to proceed in a more general setting to prove theorems 1.6 and 1.9.

Let $\mathcal{B}$ be a facet of $\mathcal{A}$, that is the set of maxima of linear functional $0 \neq \omega_B : \mathbb{R}^n \to \mathbb{R}$ while restricted to $\mathcal{A}$. Let $B = A \cap \mathcal{B}$ be the set of corresponding exponents.

We say that $P \in \bar{M}$ is a *zero at B-infinity* for f if and only if, $P \perp f$ in $\mathcal{F}_A$ and moreover, $P = \lim K(\cdot, x_j$ with $m_{x_j} \to \mathcal{B}$. A zero *at toric infinity* is a zero at B-infinity for some facet B.

Toric varieties are manifolds if and only if they satisfy a certain condition on their vertices [34]. In view of this example, we will not assume this condition. Instead,

Lemma 6.11. *The set of* $\mathbf{f} \in \mathcal{F}_A^n$ *with a zero at toric infinity is contained in a non-trivial Zariski-closed set of* $\mathcal{F}_A$*.*

Proof. Let B be a facet of A. f_B is the coordinate projection of f onto $\mathcal{F}_B \subset \mathcal{F}_A$, and $\{_B = (f_{1B}, \dots, f_{nB})$ is a holomorphic function of M. However, $\mathcal{B}$ is s-dimensional for some $s < n$. Then (after eventually changing variables), $\mathbf{f}_B$ is a system of n equations in $s < n$ variables. The set of $\mathbf{f}_B$ with a common root is therefore contained in a Zariski closed set (Theorem 2.33).

There are finitely many facets, so the set of $\mathbf{f} \in \mathcal{F}_A^n$ with a root at infinity is contained inside a Zariski closed set. $\square$

Proof of Kushnirenko's Theorem. Any point of M is smooth, so non-smooth points of $\bar{M}$ are necessarily contained at toric infinity. By Lemma 6.11, those are contained in a strict Zariski closed subset of $\mathcal{F}_A$. The same is true for critical values of π_1. Hence, given $\mathbf{f}_0, \mathbf{f}_1$ on a Zariski open set, there is a path $\mathbf{f}_t$ between them that contains only regular values of π_1 and no $\mathbf{f}_t$ has a zero at toric infinity. Therefore, there is a compact set $C \subset M$ containing all the roots $(\pi_2 \circ \pi_1^{-1}(\mathbf{f}_t)$. Lemma 6.9 then assures that $\mathbf{f}_0$ and $\mathbf{f}_1$ have the same number of roots. Proposition 6.8 finishes the proof. $\square$

6.4 Calculus of polytopes and kernels

We will use the same technique to give a proof of Bernstein's Theorem. Rather than repeating verbatim, we will stress the differences.

First the setting. Now, $\mathcal{F} = \mathcal{F}_{A_1} \times \cdots \times \mathcal{F}_{A_n}$. Each space $\mathcal{F}_{A_i}$ corresponds to one reproducing kernel K_{A_i}, one possibly degenerate symplectic form ω_{A_i}, and so on. In order to make $M = \mathbb{C}^n$ mod $2\pi\sqrt{-1}\mathbb{Z}^n$ into a Kähler manifold, we endow it with the following form:

$$\omega = \lambda_1 \omega_{A_1} + \cdots + \lambda_n \omega_{A_n}.$$

where the λ_1 strictly positive real numbers. This form can actually be degenerate.

Theorem 5.11 will give us the root expectancy,

$$\mathbb{E}(n_M(\mathbf{f})) = \frac{1}{\pi^n} \int_M \omega_{A_1} \wedge \cdots \wedge \omega_{A_n}$$

This is $1/n!$ times the coefficient in $\lambda_1^2 \lambda_2^2 \cdots \lambda_n^2$ of

$$\frac{1}{\pi^n} \int_M \omega^n$$

Note that if ω is degenerate, then the expected number of roots is zero.

It is time for the calculus of reproducing kernels. If $K(x,y) = \overline{K(y,x)}$ is smooth, and $K(x,x)$ is non-zero, then we define ω_K as the form given by the formulas of Lemma 5.10:

$$\omega = \frac{\sqrt{-1}}{2} \sum_{ij} g_{ij} dz_i \wedge d\bar{z}_j$$

with

$$g_{ij}(x) = \frac{1}{K(x,x)} \left(K_{ij}(x,x) - \frac{K_{i\cdot}(x,x) K_{\cdot j}(x,x)}{K(x,x)} \right).$$

Proposition 6.12. *Let $A = \lambda_1 A_1 + \cdots + \lambda_n A_n$. Let*

$$K_A(x,y) = \prod K_{A_i}(\lambda x, \lambda y)$$

with K_{A_i} as above. Then, K_A is a reproducing kernel corresponding to exponential sums with support in A, and

$$\int_M \omega_{K_A}^{\wedge n} = \lambda_1 \int_M \omega_{K_{A_1}}^{\wedge n} + \cdots + \lambda_n \int_M \omega_{K_{A_n}}^{\wedge n}$$

In particular, the integral of the root density is precisely $\pi^n/n!$ times the mixed volume of $\mathcal{A}_1, \ldots, \mathcal{A}_n$. Since the proof of Proposition 6.12 is left to the exercises.

Now we come to the points at toric infinity.

Definition 6.13. Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be polytopes in $\mathbb{R}^n$. A *facet* of $(\mathcal{A}_1, \ldots, \mathcal{A}_n)$ is a n-tuple $(\mathcal{B}_1, \ldots, \mathcal{B}_n)$ such that there is one linear form η in $\mathbb{R}^n$ and the points of each $\mathcal{B}_i$ are precisely the maxima of η in $\mathcal{A}_i$.

Let $B_1, \ldots, B_n$ be the lattice points in facet $(\mathcal{B}_1, \ldots, \mathcal{B}_n)$. A system $\mathbf{f}$ has a root at $(B_1, \ldots, B_n)$ infinity if and only if $(f_{1,B_1}, \ldots, f_{n,B_n})$

has a common root. Since facets have dimension $< n$, one variable may be eliminated. Hence, systems with such a common root are confined to a certain non-trivial Zariski closed set.

Since the number of facets is finite, the systems with a root at toric infinity are contained in a Zariski closed set.

The proof of Bernstein's theorem follows now exactly as for Kushnirenko's theorem.

Remark 6.14. We omitted many interesting mathematical developments related to the contents of this chapter, such as isoperimetric inequalities. A good reference is [45].

Exercise 6.1. Assume that ω is degenerate. Show that the polytopes are all orthogonal to some direction. Show that the set of $\mathbf{f}$ with common roots is a non-trivial closed Zariski set.

Exercise 6.2. Let $K(x, y), L(x, y)$ be complex symmetric functions on M and are linear in x, and $\lambda, \mu > 0$, then

$$\omega_{KL} = \omega_K + \omega_L$$

Exercise 6.3. Let

$$K(x, y) = \sum_{a \in A} c_a e^{a(x+\bar{y})}$$

and $L(x, y) = \sum_{a \in A} c_a e^{\lambda a(x+\bar{y})}$. Then $(\omega_L)_x = \lambda^2 (\omega_K)_{\lambda x}$.

Exercise 6.4. Complete the proof of Proposition 6.12

Chapter 7

Newton Iteration and Alpha theory

Let $\mathbf{f}$ be a mapping between Banach spaces. *Newton Iteration* is defined by

$$N(\mathbf{f}, \mathbf{x}) = \mathbf{x} - D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

wherever $D\mathbf{f}(\mathbf{x})$ exists and is bounded. Its only possible fixed points are those satisfying $\mathbf{f}(\mathbf{x}) = 0$. When $\mathbf{f}(\mathbf{x}) = 0$ and $D\mathbf{f}(\mathbf{x})$ is invertible, we say that $\mathbf{x}$ is a *nondegenerate zero* of $\mathbf{f}$.

It is well-known that Newton iteration is quadratically convergent in a neighborhood of a nondegenerate zero ζ. Indeed, $N(\mathbf{f}, \mathbf{x}) - \zeta = D^2\mathbf{f}(\zeta)(\mathbf{x} - \zeta)^2 + \cdots$.

There are two main approaches to quantify how fast is quadratic convergence. One of them, pioneered by Kantorovich [48] assumes that the mapping $\mathbf{f}$ has a bounded second derivative, and that this bound is known.

Gregorio Malajovich, *Nonlinear equations.*
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

The other approach, developed by Smale [76, 77] and described here, assumes that the mapping $\mathbf{f}$ is analytic. Then we will be able to estimate a neighborhood of quadratic convergence around a given zero (Theorem 7.5) or to certify an 'approximate root' (Theorem 7.15) from data that depends only on the value and derivatives of $\mathbf{f}$ at one point.

A more general exposition on this subject may be found in [29], covering also overdetermined and undetermined polynomial systems.

7.1 The gamma invariant

Through this chapter, $\mathbb{E}$ and $\mathbb{F}$ are Banach spaces, $\mathcal{D} \subseteq \mathbb{E}$ is open and $\mathbf{f} : \mathbb{E} \to \mathbb{F}$ is analytic.

This means that if $\mathbf{x}_0 \in \mathbb{E}$ is in the domain of $\mathbb{E}$, then there is $\rho > 0$ with the property that the series

$$\mathbf{f}(\mathbf{x}_0) + \mathbf{D}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \mathbf{D}^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0) + \cdots \tag{7.1}$$

converges uniformly for $\|\mathbf{x} - \mathbf{x}_0\| < \rho$, and its limit is equal to $\mathbf{f}(\mathbf{x})$ (For more details about analytic functions between Banach spaces, see [65, 66]).

In order to abbreviate notations, we will write (7.1) as

$$\mathbf{f}(\mathbf{x}_0) + \mathbf{D}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \sum_{k \geq 2} \frac{1}{k!} \mathbf{D}^k f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^k$$

where the exponent k means that $\mathbf{x} - \mathbf{x}_0$ appears k times as an argument to the preceding multi-linear operator.

The maximum of such ρ will be called the *radius of convergence.* (It is ∞ when the series (7.1) is globally convergent). This terminology comes from one complex variable analysis. When $\mathbf{E} = \mathbb{C}$, the series will converge for all $\mathbf{x} \in B(\mathbf{x}_0, \rho)$ and diverge for all $\mathbf{x} \notin \overline{B(\mathbf{x}_0, \rho)}$. This is no more true in several complex variables, or Banach spaces (Exercise 7.3).

The norm of a k-linear operator in Banach Spaces (such as the k-th derivative) is the *operator norm*, for instance

$$\|D^k \mathbf{f}(\mathbf{x}_0)\|_{\mathbb{E} \to \mathbb{F}} = \sup_{\|\mathbf{u}_1\|_{\mathbb{E}} = \cdots = \|\mathbf{u}_k\|_{\mathbb{E}} = 1} \|D^k \mathbf{f}(\mathbf{x}_0)(\mathbf{u}_1, \ldots, \mathbf{u}_k)\|_{\mathbb{F}}.$$

As long as there is no ambiguity, we drop the subscripts of the norm.

Definition 7.1 (Smale's γ invariant)**.** Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic mapping between Banach spaces, and $\mathbf{x} \in \mathbb{E}$. When $D\mathbf{f}(\mathbf{x})$ is invertible, define

$$\gamma(\mathbf{f}, \mathbf{x}_0) = \sup_{k \geq 2} \left(\frac{\|D\mathbf{f}(\mathbf{x}_0)^{-1} D^k \mathbf{f}(\mathbf{x}_0)\|}{k!} \right)^{\frac{1}{k-1}}.$$

Otherwise, set $\gamma(\mathbf{f}, \mathbf{x}_0) = \infty$.

In the one variable setting, this can be compared to the radius of convergence ρ of $\mathbf{f}'(\mathbf{x})/\mathbf{f}'(\mathbf{x}_0)$, that satisfies

$$\rho^{-1} = \limsup_{k \geq 2} \left(\frac{\|\mathbf{f}'(\mathbf{x}_0)^{-1} \mathbf{f}^{(k)}(\mathbf{x}_0)\|}{k!} \right)^{\frac{1}{k-1}}.$$

More generally,

Proposition 7.2. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be a C^∞ map between Banach spaces, and $\mathbf{x}_0 \in \mathcal{D}$ such that $\gamma(\mathbf{f}, \mathbf{x}_0) < \infty$. Then f is analytic in x_0 if and only if, $\gamma(f, x_0)$ is finite. The series*

$$\mathbf{f}(\mathbf{x}_0) + \mathbf{D}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \sum_{k \geq 2} \frac{1}{k!} \mathbf{D}^k f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^k \tag{7.2}$$

is uniformly convergent for $\mathbf{x} \in B(\mathbf{x}_0, \rho)$ for any $\rho < 1/\gamma(\mathbf{f}, \mathbf{x}_0))$.

Proposition 7.2, if. The series

$$\mathbf{D}f(\mathbf{x}_0)^{-1}\mathbf{f}(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) + \sum_{k \geq 2} \frac{1}{k!} \mathbf{D}f(\mathbf{x}_0)^{-1} \mathbf{D}^k f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^k$$

is uniformly convergent in $B(\mathbf{x}_0, \rho)$ where

$$\begin{aligned}
\rho^{-1} &< \limsup_{k \geq 2} \left(\frac{\|D\mathbf{f}(\mathbf{x}_0)^{-1} D^k \mathbf{f}(\mathbf{x}_0)\|}{k!} \right)^{\frac{1}{k}} \\
&\leq \limsup_{k \geq 2} \gamma(\mathbf{f}, \mathbf{x}_0)^{\frac{k-1}{k}} \\
&= \lim_{k \to \infty} \gamma(\mathbf{f}, \mathbf{x}_0)^{\frac{k-1}{k}} \\
&= \gamma(\mathbf{f}, \mathbf{x}_0)
\end{aligned}$$

□

Before proving the *only if* part of Proposition 7.2, we need to relate the norm of a multi-linear map to the norm of the corresponding polynomial.

Lemma 7.3. *Let $k \geq 2$. Let $\mathbf{T} : \mathbb{E}^k \to \mathbb{F}$ be k-linear and symmetric. Let $\mathbf{S} : \mathbb{E} \to \mathbb{F}$, $\mathbf{S}(\mathbf{x}) = T(\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x})$ be the corresponding polynomial. Then,*

$$\|\mathbf{T}\| \leq e^{k-1} \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{S}(\mathbf{x})\|$$

Proof. The polarization formula for (real or complex) tensors is

$$\mathbf{T}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \frac{1}{2^k k!} \sum_{\substack{\epsilon_j = \pm 1 \\ j=1,\ldots,k}} \epsilon_1 \cdots \epsilon_k \mathbf{S}\left(\sum_{l=1}^{k} \epsilon_l \mathbf{x}_l \right)$$

It is easily derived by expanding the expression inside parentheses. There will be $2^k k!$ terms of the form

$$\epsilon_1 \cdots \epsilon_k T(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k)$$

or its permutations. All other terms miss at least one variable (say $\mathbf{x}_j$). They cancel by summing for $\epsilon_j = \pm 1$.

It follows that when $\|\mathbf{x}\| \leq 1$,

$$\begin{aligned} \mathbf{T}(\mathbf{x}_1, \cdots, \mathbf{x}_k) &\leq \frac{1}{k!} \max_{\substack{\epsilon_j = \pm 1 \\ j=1,\ldots,k}} \|\mathbf{S}\left(\sum_{l=1}^{k} \epsilon_l \mathbf{x}_l \right)\| \\ &\leq \frac{k^k}{k!} \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{S}(\mathbf{x})\| \end{aligned}$$

The Lemma follows from using Stirling's formula,

$$k! \geq \sqrt{2\pi k} k^k e^{-k} e^{1/(12k+1)}.$$

We obtain:

$$\|\mathbf{T}\| \leq \left(\frac{1}{\sqrt{2\pi k}} e^{12k+1} \right) e^k \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{S}(\mathbf{x})\|.$$

Then we use the fact that $k \geq 2$, hence $\sqrt{2\pi k} \geq e$. □

Proposition 7.2, only if. Assume that the series (7.2) converges uniformly for $\|\mathbf{x} - \mathbf{x}_0\| < \rho$. Without loss of generality assume that $\mathbb{E} = \mathbb{F}$ and $D\mathbf{f}(\mathbf{x}_0) = I$.

We claim that

$$\limsup_{k \geq 2} \sup_{\|\mathbf{u}\|=1} \|\frac{1}{k!} D^k \mathbf{f}(\mathbf{x}_0)\mathbf{u}^k\|^{1/k} \leq \rho^{-1}.$$

Indeed, assume that there is $\delta > 0$ and infinitely many pairs $(k_i, \mathbf{u}_i)$ with $\|\mathbf{u}_i\| = 1$ and

$$\|\frac{1}{k!} D^k \mathbf{f}(\mathbf{x}_0)\mathbf{u}^k\|^{1/k} > \rho^{-1}(1+\delta).$$

In that case,

$$\|\frac{1}{k!} D^k \mathbf{f}(\mathbf{x}_0) \left(\frac{\rho}{\sqrt{1+\delta}}\mathbf{u}\right)^k \| > \sqrt{1+\delta}^k$$

infinitely many times, and hence (7.2) does not converge uniformly on $B(\mathbf{x}_0, \rho)$.

Now, we can apply Lemma 7.3 to obtain:

$$\begin{aligned} \limsup_{k \geq 2} \|\frac{1}{k!} D^k \mathbf{f}(\mathbf{x}_0)\|^{1/(k-1)} &\leq e \limsup_{k \geq 2} \sup_{\|\mathbf{u}\|=1} \|\frac{1}{k!} D^k \mathbf{f}(\mathbf{x}_0)\mathbf{u}^k\|^{\frac{1}{k-1}} \\ &\leq e \lim_{k \to \infty} \rho^{-(1+1/(k-1))} \\ &= e\rho^{-1} \end{aligned}$$

and therefore $\|\frac{1}{k!} D^k f(x_0)\|^{1/(k-1)}$ is bounded. □

Exercise 7.1. Show the polarization formula for Hermitian product:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{4} \sum_{\epsilon^4 = 1} \epsilon \|\mathbf{u} + \epsilon \mathbf{v}\|^2$$

Explain why this is different from the one in Lemma 7.3.

Exercise 7.2. If one drops the uniform convergence hypothesis in the definition of analytic functions, what happens to Proposition 7.2?

7.2 The γ-Theorems

The following concept provides a good abstraction of quadratic convergence.

Definition 7.4 (Approximate zero of the first kind)**.** Let $\mathbf{f} : \mathcal{D} \subseteq \mathbf{E} \to \mathbf{F}$ be as above, with $\mathbf{f}(\zeta) = 0$. An *approximate zero of the first kind* associated to ζ is a point $\mathbf{x}_0 \in \mathcal{D}$, such that

1. The sequence $(\mathbf{x})_i$ defined inductively by $\mathbf{x}_{i+1} = N(\mathbf{f}, \mathbf{x}_i)$ is well-defined (each $\mathbf{x}_i$ belongs to the domain of $\mathbf{f}$ and $D\mathbf{f}(\mathbf{x}_i)$ is invertible and bounded).

2. $$\|\mathbf{x}_i - \zeta\| \leq 2^{-2^i+1} \|\mathbf{x}_0 - \zeta\|.$$

The existence of approximate zeros of the first kind is not obvious, and requires a theorem.

Theorem 7.5 (Smale)**.** *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let ζ be a non-degenerate zero of $\mathbf{f}$. Assume that*

$$B = B\left(\zeta, \frac{3-\sqrt{7}}{2\gamma(\mathbf{f},\zeta)}\right) \subseteq \mathcal{D}.$$

Every $\mathbf{x}_0 \in B$ is an approximate zero of the first kind associated to ζ. The constant $(3-\sqrt{7})/2$ is the smallest with that property.

Before going further, we remind the reader of the following fact.

Lemma 7.6. *Let $d \geq 1$ be integer, and let $|t| < 1$. Then,*

$$\frac{1}{(1-t)^d} = \sum_{k \geq 0} \binom{k+d-1}{d-1} t^k.$$

Proof. Differentiate $d-1$ times the two sides of the expression $1/(1-t) = 1 + t + t^2 + \cdots$, and then divide both sides by $d-1$! $\square$

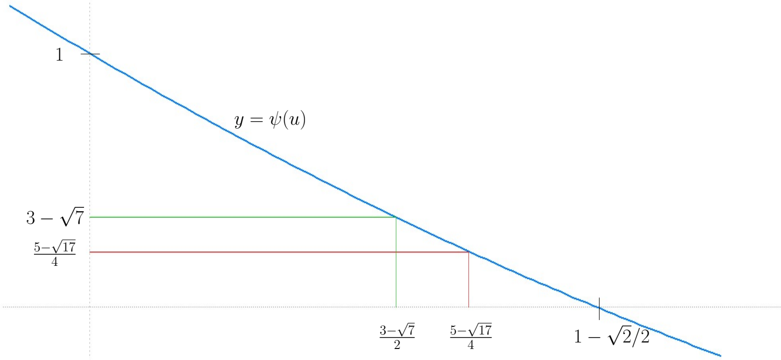

Figure 7.1: $y = \psi(u)$

Lemma 7.7. *The function $\psi(u) = 1 - 4u + 2u^2$ is decreasing and non-negative in $[0, 1 - \sqrt{2}/2]$, and satisfies:*

$$\frac{u}{\psi(u)} < 1 \qquad \text{for } u \in [0, (5-\sqrt{17})/4) \tag{7.3}$$

$$\frac{u}{\psi(u)} \le \frac{1}{2} \qquad \text{for } u \in [0, (3-\sqrt{7})/2]\ . \tag{7.4}$$

The proof of Lemma 7.7 is left to the reader (but see Figure 7.1).

Another useful result is:

Lemma 7.8. *Let A be a $n \times n$ matrix. Assume $\|A - I\|_2 < 1$. Then A has full rank and, for all y,*

$$\frac{\|y\|}{1 + \|A - I\|_2} \le \|A^{-1}y\|_2 \le \frac{\|y\|}{1 - \|A - I\|_2}.$$

Proof. By hypothesis, $\|Ax\| > 0$ for all $x \neq 0$ so that A has full rank. Let $y = Ax$. By triangular inequality,

$$\|Ax\| \ge \|x\| - \|(A - I)x\| \ge (1 - \|(A - I)\|_2)\|x\|.$$

Also by triangular inequality,

$$\|Ax\| \le \|x\| + \|(A - I)x\| \le (1 + \|(A - I)\|_2)\|x\|.$$

$\square$

The following Lemma will be needed:

Lemma 7.9. *Assume that* $u = \|\mathbf{x} - \mathbf{y}\|\gamma(\mathbf{f}, \mathbf{x}) < 1 - \sqrt{2}/2$. *Then,*

$$\|D\mathbf{f}(\mathbf{y})^{-1} D\mathbf{f}(\mathbf{x})\| \leq \frac{(1-u)^2}{\psi(u)}.$$

Proof. Expanding $\mathbf{y} \mapsto D\mathbf{f}(\mathbf{x})^{-1} D\mathbf{f}(\mathbf{y})$ around $\mathbf{x}$, we obtain:

$$D\mathbf{f}(\mathbf{x})^{-1} D\mathbf{f}(\mathbf{y}) = I + \sum_{k \geq 2} \frac{1}{k-1!} D\mathbf{f}(\mathbf{x})^{-1} D^k \mathbf{f}(\mathbf{x})(\mathbf{y} - \mathbf{x})^{k-1}.$$

Rearranging terms and taking norms, Lemma 7.6 yields

$$\|D\mathbf{f}(\mathbf{x})^{-1} D\mathbf{f}(\mathbf{y}) - I\| \leq \frac{1}{(1 - \gamma\|\mathbf{y} - \mathbf{x}\|)^2} - 1.$$

By Lemma 7.8 we deduce that $D\mathbf{f}(\mathbf{x})^{-1} D\mathbf{f}(\mathbf{y})$ is invertible, and

$$\|D\mathbf{f}(\mathbf{y})^{-1} D\mathbf{f}(\mathbf{x})\| \leq \frac{1}{1 - \|D\mathbf{f}(\mathbf{x})^{-1} D\mathbf{f}(\mathbf{y}) - I\|} = \frac{(1-u)^2}{\psi(u)}. \tag{7.5}$$

$\square$

Here is the method for proving Theorem 7.5 and similar ones: first we study the convergence of Newton iteration applied to a 'universal' function. In this case, set

$$h_\gamma(t) = t - \gamma t^2 - \gamma^2 t^3 - \cdots = t - \frac{\gamma t^2}{1 - \gamma t}.$$

(See figure 7.2).

The function h_γ has a zero at $t = 0$, and $\gamma(h_\gamma, 0) = \gamma$. Then, we compare the convergence of Newton iteration applied to an arbitrary function to the convergence when applied to the universal function.

Lemma 7.10. *Assume that* $0 \leq u_0 = \gamma t_0 < \frac{5-\sqrt{17}}{4}$. *Then the sequences*

$$t_{i+1} = N(h_\gamma, t_i) \text{ and } u_{i+1} = \frac{u_i^2}{\psi(u_i)}$$

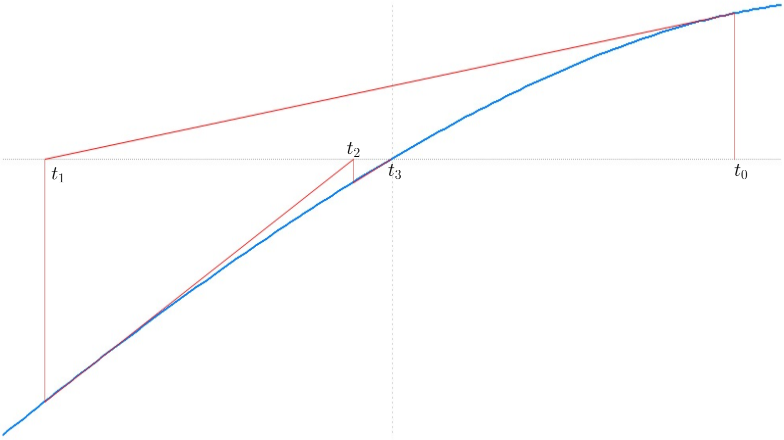

Figure 7.2: $y = h_\gamma(t)$

are well-defined for all i, $\lim_{i\to\infty} t_i = 0$, *and*

$$\frac{|t_i|}{|t_0|} = \frac{u_i}{u_0} \leq \left(\frac{u_0}{\psi(u_0)}\right)^{2^i-1}.$$

Moreover,

$$\frac{|t_i|}{|t_0|} \leq 2^{-2^i+1}$$

for all i *if and only if* $u_0 \leq \frac{3-\sqrt{7}}{2}$.

Proof. We just compute

$$\begin{aligned} h'_\gamma(t) &= \frac{\psi(\gamma t)}{(1-\gamma t)^2} \\ t h'_\gamma(t) - h_\gamma(t) &= -\frac{\gamma t^2}{(1-\gamma t)^2} \\ N(h_\gamma, t) &= -\frac{\gamma t^2}{\psi(\gamma t)}. \end{aligned}$$

When $u_0 < \frac{5-\sqrt{17}}{4}$, (7.3) implies that the sequence u_i is decreasing, and by induction

$$u_i = \gamma |t_i|.$$

Moreover,

$$\frac{u_{i+1}}{u_0} = \left(\frac{u_i}{u_0}\right)^2 \frac{u_0}{\psi(u_i)} \leq \left(\frac{u_i}{u_0}\right)^2 \frac{u_0}{\psi(u_0)} < \left(\frac{u_i}{u_0}\right)^2 .$$

By induction,

$$\frac{u_i}{u_0} \leq \left(\frac{u_0}{\psi(u_0)}\right)^{2^i-1} .$$

This also implies that $\lim t_i = 0$.

When furthermore $u_0 \leq (3-\sqrt{7})/2$, $u_0/\psi(u_0) \leq 1/2$ by (7.4) hence $u_i/u_0 \leq 2^{-2^i+1}$. For the converse, if $u_0 > (3-\sqrt{7})/2$, then

$$\frac{|t_1|}{|t_0|} = \frac{u_0}{\psi(u_0)} > \frac{1}{2}.$$

□

Before proceeding to the proof of Theorem 7.5, a remark is in order.

Both Newton iteration and γ are invariant with respect to translation and to linear changes of coordinates: let $\mathbf{g}(\mathbf{x}) = A\mathbf{f}(\mathbf{x}-\zeta)$, where A is a continuous and invertible linear operator from $\mathbb{F}$ to $\mathbb{E}$. Then

$$N(\mathbf{g}, \mathbf{x}+\zeta) = N(\mathbf{f}, \mathbf{x}) + \zeta \text{ and } \gamma(\mathbf{g}, \mathbf{x}+\zeta) = \gamma(\mathbf{f}, \mathbf{x}).$$

Also, distances in $\mathbb{E}$ are invariant under translation.

Proof of Theorem 7.5. Assume without loss of generality that $\zeta = 0$ and $D\mathbf{f}(\zeta) = I$. Set $\gamma = \gamma(\mathbf{f}, \mathbf{x})$, $u_0 = \|\mathbf{x}_0\|\gamma$, and let h_γ and the sequence (u_i) be as in Lemma 7.10.

We will bound

$$\|N(\mathbf{f},\mathbf{x})\| = \left\|\mathbf{x} - D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})\right\| \leq \|D\mathbf{f}(\mathbf{x})^{-1}\| \|\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{x}\|. \tag{7.6}$$

The Taylor expansions of $\mathbf{f}$ and $D\mathbf{f}$ around 0 are respectively:

$$\mathbf{f}(\mathbf{x}) = \mathbf{x} + \sum_{k \geq 2} \frac{1}{k!} D^k \mathbf{f}(0)\mathbf{x}^k$$

and

$$D\mathbf{f}(\mathbf{x}) = I + \sum_{k \geq 2} \frac{1}{k-1!} D^k \mathbf{f}(0)\mathbf{x}^{k-1}. \tag{7.7}$$

Combining the two equations, above, we obtain:

$$\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{x} = \sum_{k \geq 2} \frac{k-1}{k!} D^k \mathbf{f}(0)\mathbf{x}^k.$$

Using Lemma 7.6 with $d = 2$, the rightmost term in (7.6) is bounded above by

$$\|\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{x}\| \leq \sum_{k \geq 2} (k-1)\gamma^{k-1}\|\mathbf{x}\|^k = \frac{\gamma\|\mathbf{x}\|^2}{(1-\gamma\|\mathbf{x}\|)^2}. \tag{7.8}$$

Combining Lemma 7.9 and (7.8) in (7.6), we deduce that

$$\|N(\mathbf{f}, \mathbf{x})\| \leq \frac{\gamma\|\mathbf{x}\|^2}{\psi(\gamma\|\mathbf{x}\|)}.$$

By induction, $u_i \leq \gamma\|\mathbf{x}_i\|$. When $u_0 \leq (3-\sqrt{7})/2$, we obtain as in Lemma 7.10 that

$$\frac{\|\mathbf{x}_i\|}{\|\mathbf{x}_0\|} \leq \frac{u_i}{u_0} \leq 2^{-2^i+1}.$$

We have seen in Lemma 7.10 that the bound above fails for $i = 1$ when $u_0 > (3-\sqrt{7})/2$. □

Notice that in the proof above,

$$\lim_{i \to \infty} \frac{u_0}{\psi(u_i)} = u_0.$$

Therefore, convergence is actually faster than predicted by the definition of approximate zero. We proved actually a sharper result:

	1/32	1/16	1/10	1/8	$\frac{3-\sqrt{7}}{2}$
1	4.810	3.599	2.632	2.870	1.000
2	14.614	11.169	8.491	6.997	3.900
3	34.229	26.339	20.302	16.988	10.229
4	73.458	56.679	43.926	36.977	22.954
5	151.917	117.358	91.175	76.954	48.406

Table 7.1: Values of $-log_2(u_i/u_0)$ in function of u_0 and i.

Theorem 7.11. *Let* $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ *be an analytic map between Banach spaces. Let* ζ *be a non-degenerate zero of* $\mathbf{f}$. *Let* $u_0 < (5 - \sqrt{17})/4$.

Assume that

$$B = B\left(\zeta, \frac{u_0}{\gamma(\mathbf{f}, \zeta)}\right) \subseteq \mathcal{D}.$$

If $\mathbf{x}_0 \in B$*, then the sequences*

$$\mathbf{x}_{i+1} = N(\mathbf{f}, \mathbf{x}_i) \text{ and } u_{i+1} = \frac{u_i^2}{\psi(u_i)}$$

are well-defined for all i*, and*

$$\frac{\|\mathbf{x}_i - \zeta\|}{\|\mathbf{x}_0 - \zeta\|} \le \frac{u_i}{u_0} \le \left(\frac{u_0}{\psi(u_0)}\right)^{-2^i+1}.$$

Table 7.1 and Figure 7.3 show how fast u_i/u_0 decreases in terms of u_0 and i.

To conclude this section, we need to address an important issue for numerical computations. Whenever dealing with digital computers, it is convenient to perform calculations in floating point format. This means that each real number is stored as a *mantissa* (an integer, typically no more than 2^{24} or 2^{53}) times an exponent. (The IEEE-754 standard for computer arithmetics [47] is taught at elementary numerical analysis courses, see for instance [46, Ch.2]).

By using floating point numbers, a huge gain of speed is obtained with regard to exact representation of, say, algebraic numbers. However, computations are inexact (by a typical factor of 2^{-24} or 2^{-53}).

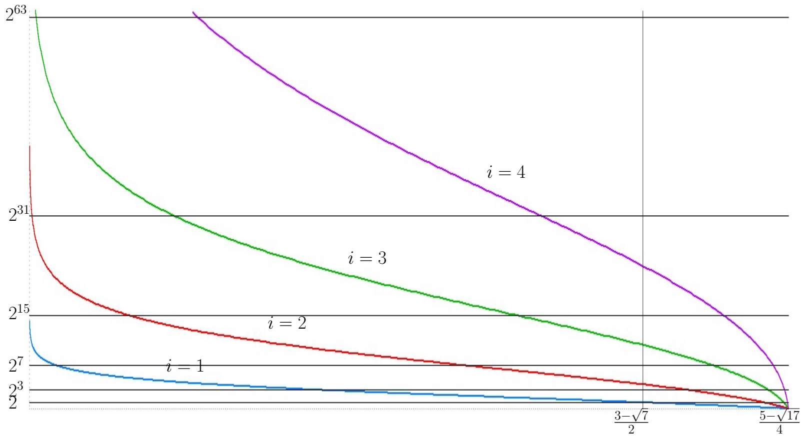

Figure 7.3: Values of $log_2(u_i/u_0)$ in function of u_0 for $i = 1, \dots, 4$.

Therefore, we need to consider *inexact* Newton iteration. An obvious modification of the proof of Theorem 7.5 gives us the following statement:

Theorem 7.12. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \rightarrow \mathbb{F}$ be an analytic map between Banach spaces. Let ζ be a non-degenerate zero of $\mathbf{f}$. Let*

$$0 \leq 2\delta \leq u_0 \leq 2 - \frac{\sqrt{14}}{2} \simeq 0.129 \cdots$$

Assume that

1. $$B = B\left(\zeta, \frac{u_0}{\gamma(\mathbf{f}, \zeta)}\right) \subseteq \mathcal{D}.$$

2. *$\mathbf{x}_0 \in B$, and the sequence $\mathbf{x}_i$ satisfies*

$$\|\mathbf{x}_{i+1} - N(\mathbf{f}, \mathbf{x}_i)\| \gamma(\mathbf{f}, \zeta) \leq \delta$$

3. *The sequence u_i is defined inductively by*

$$u_{i+1} = \frac{u_i^2}{\psi(u_i)} + \delta.$$

Then the sequences u_i and $\mathbf{x}_i$ are well-defined for all i, $\mathbf{x}_i \in \mathcal{D}$, and

$$\frac{\|\mathbf{x}_i - \zeta\|}{\|\mathbf{x}_0 - \zeta\|} \le \frac{u_i}{u_0} \le \max\left(2^{-2^i+1}, 2\frac{\delta}{u_0}\right).$$

Proof. By hypothesis,

$$\frac{u_0}{\psi(u_0)} + \frac{\delta}{u_0} < 1$$

so the sequence u_i is decreasing and positive. For short, let $q = \frac{u_0}{\psi(u_0)} \le 1/4$. By induction,

$$\frac{u_{i+1}}{u_0} \le \frac{u_0}{\psi(u_i)}\left(\frac{u_i}{u_0}\right)^2 + \frac{\delta}{u_0} \le \frac{1}{4}\left(\frac{u_i}{u_0}\right)^2 + \frac{\delta}{u_0}.$$

Assume that $u_i/u_0 \le 2^{-2^i+1}$. In that case,

$$\frac{u_{i+1}}{u_0} \le 2^{-2^{i+1}} + \frac{\delta}{u_0} \le \max\left(2^{-2^{i+1}+1}, 2\frac{\delta}{u_0}\right).$$

Assume now that $2^{-2^i+1}, u_i/u_0 \le 2\delta/u_0$. In that case,

$$\frac{u_{i+1}}{u_0} \le \frac{\delta}{u_0}\left(\frac{\delta}{4u_0} + 1\right) \le \frac{2\delta}{u_0} = \max\left(2^{-2^{i+1}+1}, 2\frac{\delta}{u_0}\right).$$

From now on we use the assumptions, notations and estimates of the proof of Theorem 7.5. Combining (7.5) and (7.8) in (7.6), we obtain again that

$$\|N(\mathbf{f}, \mathbf{x})\| \le \frac{\gamma\|\mathbf{x}\|^2}{\psi(\gamma\|\mathbf{x}\|)}.$$

This time, this means that

$$\|\mathbf{x}_{i+1}\|\gamma \le \delta + \|N(\mathbf{f}, \mathbf{x})\|\gamma \le \delta + \frac{\gamma^2\|\mathbf{x}\|^2}{\psi(\gamma\|\mathbf{x}\|)}.$$

By induction that $\|\mathbf{x}_i - \zeta\|\gamma(\mathbf{f}, \zeta) < u_i$ and we are done. $\square$

Exercise 7.3. Consider the following series, defined in $\mathbb{C}^2$:

$$g(x) = \sum_{i=0}^{\infty} x_1^i x_2^i.$$

Compute its radius of convergence. What is its domain of absolute convergence ?

Exercise 7.4. The objective of this exercise is to produce a non-optimal algorithm to approximate $\sqrt{y}$. In order to do that, consider the mapping $f(x) = x^2 - y$.

1. Compute $\gamma(f, x)$.

2. Show that for $1 \leq y \leq 4$, $x_0 = 1/2 + y/2$ is an approximate zero of the first kind for x, associated to y.

3. Write down an algorithm to approximate $\sqrt{y}$ up to relative accuracy 2^{-63}.

Exercise 7.5. Let $\mathbf{f}$ be an analytic map between Banach spaces, and assume that ζ is a non-degenerate zero of $\mathbf{f}$.

1. Write down the Taylor series of $D\mathbf{f}(\zeta)^{-1}\left(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\zeta)\right)$.

2. Show that if $\mathbf{f}(\mathbf{x}) = 0$, then

$$\gamma(\mathbf{f}, \zeta)\|\mathbf{x} - \zeta\| \geq 1/2.$$

This shows that two non-degenerate zeros cannot be at a distance less than $1/2\gamma(\mathbf{f}, \zeta)$. (Results of this type appeared in [28], but some of them were known before [55, Th.16]).

7.3 Estimates from data at a point

Theorem 7.5 guarantees quadratic convergence in a neighborhood of a known zero ζ. In practical situations, ζ is not known. A major result in alpha-theory is the criterion to detect an approximate zero with just local information. We need to slightly modify the definition.

Definition 7.13 (Approximate zero of the second kind)**.** Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be as above. An *approximate zero of the first kind* associated to $\zeta \in \mathcal{D}$, $\mathbf{f}(\zeta) = 0$, is a point $\mathbf{x}_0 \in \mathcal{D}$, such that

1. The sequence $(\mathbf{x})_i$ defined inductively by $\mathbf{x}_{i+1} = N(\mathbf{f}, \mathbf{x}_i)$ is well-defined (each $\mathbf{x}_i$ belongs to the domain of $\mathbf{f}$ and $D\mathbf{f}(\mathbf{x}_i)$ is invertible and bounded).
2. $$\|\mathbf{x}_{i+1} - \mathbf{x}_i\| \leq 2^{-2^i+1} \|\mathbf{x}_1 - \mathbf{x}_0\|.$$
3. $\lim_{i\to\infty} \mathbf{x}_i = \zeta$.

For detecting approximate zeros of the second kind, we need:

Definition 7.14 (Smale's β and α invariants)**.**

$$\beta(\mathbf{f}, \mathbf{x}) = \|D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})\| \text{ and } \alpha(\mathbf{f}, \mathbf{x}) = \beta(\mathbf{f}, \mathbf{x})\gamma(\mathbf{f}, \mathbf{x}).$$

The β invariant can be interpreted as the size of the Newton step $N(\mathbf{f}, \mathbf{x}) - \mathbf{x}$.

Theorem 7.15 (Smale)**.** *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let*

$$\alpha \leq \alpha_0 = \frac{13 - 3\sqrt{17}}{4}.$$

Define

$$r_0 = \frac{1 + \alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha} \text{ and } r_1 = \frac{1 - 3\alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha}.$$

Let $\mathbf{x}_0 \in \mathcal{D}$ be such that $\alpha(\mathbf{f}, \mathbf{x}_0) \leq \alpha$ and assume furthermore that $B(\mathbf{x}_0, r_0\beta(\mathbf{f}, \mathbf{x}_0)) \subseteq \mathcal{D}$. Then,

1. *$\mathbf{x}_0$ is an approximate zero of the second kind, associated to some zero $\zeta \in \mathcal{D}$ of $\mathbf{f}$.*
2. *Moreover, $\|\mathbf{x}_0 - \zeta\| \leq r_0\beta(\mathbf{f}, \mathbf{x}_0)$.*
3. *Let $\mathbf{x}_1 = N(\mathbf{f}, \mathbf{x}_0)$. Then $\|\mathbf{x}_1 - \zeta\| \leq r_1\beta(\mathbf{f}, \mathbf{x}_0)$.*

The constant α_0 is the largest possible with those properties.

This theorem appeared in [77]. The value for α_0 was found by Wang Xinghua [84]. Numerically,

$$\alpha_0 = 0.157,670,780,786,754,587,633,942,608,019\cdots$$

Other useful numerical bounds, under the hypotheses of the theorem, are:

$$r_0 \leq 1.390,388,203\cdots \text{ and } r_1 \leq 0.390,388,203\cdots .$$

The proof of Theorem 7.15 follows from the same method as the one for Theorem 7.5. We first define the 'worst' real function with respect to Newton iteration. Let us fix $\beta, \gamma > 0$. Define

$$h_{\beta\gamma}(t) = \beta - t + \frac{\gamma t^2}{1-\gamma t} = \beta - t + \gamma t^2 + \gamma^2 t^3 + \cdots .$$

We assume for the time being that $\alpha = \beta\gamma < 3-2\sqrt{2} = 0.1715\cdots$. This guarantees that $h_{\beta\gamma}$ has two distinct zeros $\zeta_1 = \frac{1+\alpha-\sqrt{\Delta}}{4\gamma}$ and $\zeta_2 = \frac{1+\alpha+\sqrt{\Delta}}{4\gamma}$ with of course $\Delta = (1+\alpha)^2 - 8\alpha$. An useful expression is the product formula

$$h_{\beta\gamma}(x) = 2\frac{(x-\zeta_1)(x-\zeta_2)}{\gamma^{-1}-x}. \tag{7.9}$$

From (7.9), $h_{\beta\gamma}$ has also a pole at γ^{-1}. We have always $0 < \zeta_1 < \zeta_2 < \gamma^{-1}$.

The function $h_{\beta\gamma}$ is, among the functions with $h'(0) = -1$ and $\beta(h,0) \leq \beta$ and $\gamma(h,0) \leq \gamma$, the one that has the first zero ζ_1 furthest away from the origin.

Proposition 7.16. *Let $\beta, \gamma > 0$, with $\alpha = \beta\gamma \leq 3-2\sqrt{2}$. let $h_{\beta\gamma}$ be as above. Define recursively $t_0 = 0$ and $t_{i+1} = N(h_{\beta\gamma}, t_i)$. then*

$$t_i = \zeta_1 \frac{1-q^{2^i-1}}{1-\eta q^{2^i-1}}, \tag{7.10}$$

with

$$\eta = \frac{\zeta_1}{\zeta_2} = \frac{1+\alpha-\sqrt{\Delta}}{1+\alpha+\sqrt{\Delta}} \text{ and } q = \frac{\zeta_1-\gamma\zeta_1\zeta_2}{\zeta_2-\gamma\zeta_1\zeta_2} = \frac{1-\alpha-\sqrt{\Delta}}{1-\alpha+\sqrt{\Delta}}.$$

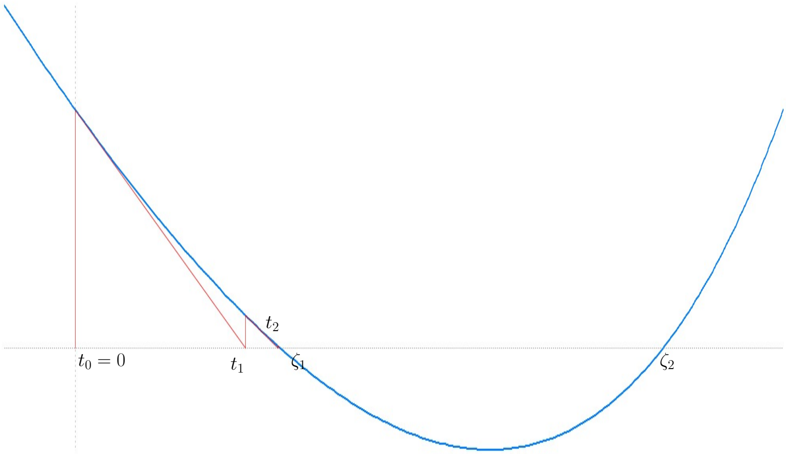

Figure 7.4: $y = h_{\beta\gamma}(t)$.

Proof. By differentiating (7.9), one obtains

$$h'_{\beta\gamma}(t) = h_{\beta\gamma}(t)\left(\frac{1}{t-\zeta_1} + \frac{1}{t-\zeta_2} + \frac{1}{\gamma^{-1}-t}\right)$$

and hence the Newton operator is

$$N(h_{\beta\gamma}, t) = t - \frac{1}{\frac{1}{t-\zeta_1} + \frac{1}{t-\zeta_2} + \frac{1}{\gamma^{-1}-t}}.$$

A tedious calculation shows that $N(h_{\beta\gamma}, t)$ is a rational function of degree 2. Hence, it is defined by 5 coefficients, or by 5 values.

In order to solve the recurrence for t_i, we change coordinates using a fractional linear transformation. As the Newton operator will have two attracting fixed points (ζ_1 and ζ_2), we will map those points to 0 and ∞ respectively. For convenience, we will map $t_0 = 0$ into $y_0 = 1$. Therefore, we set

$$S(t) = \frac{\zeta_2 t - \zeta_1\zeta_2}{\zeta_1 t - \zeta_1\zeta_2} \quad \text{and} \quad S^{-1}(y) = \frac{-\zeta_1\zeta_2 y + \zeta_1\zeta_2}{-\zeta_1 y + \zeta_2}$$

Let us look at the sequence $y_i = S(t_i)$. By construction $y_0 = 1$, and subsequent values are given by the recurrence

$$y_{i+1} = S(N(h_{\beta\gamma}, S^{-1}(y_i))).$$

It is an exercise to check that

$$y_{i+1} = q y_i^2, \tag{7.11}$$

Therefore we have $y_i = q^{2^i - 1}$, and equation (7.10) holds. □

Proposition 7.17. *Under the conditions of Proposition 7.16, 0 is an approximate zero of the second kind for $h_{\beta\gamma}$ if and only if*

$$\alpha = \beta\gamma \le \frac{13 - 3\sqrt{17}}{4}.$$

Proof. Using the closed form for t_i, we get:

$$\begin{aligned} t_{i+1} - t_i &= \frac{1 - q^{2^{i+1}-1}}{1 - \eta q^{2^{i+1}-1}} - \frac{1 - q^{2^i - 1}}{1 - \eta q^{2^i - 1}} \\ &= q^{2^i - 1} \frac{(1-\eta)(1-q^{2^i})}{(1 - \eta q^{2^{i+1}-1})(1 - \eta q^{2^i - 1})} \end{aligned}$$

In the particular case $i = 0$,

$$t_1 - t_0 = \frac{1-q}{1-\eta q} = \beta$$

Hence

$$\frac{t_{i+1} - t_i}{\beta} = C_i q^{2^i - 1}$$

with

$$C_i = \frac{(1-\eta)(1-\eta q)(1-q^{2^i})}{(1-q)(1 - \eta q^{2^{i+1}-1})(1 - \eta q^{2^i - 1})}.$$

Thus, $C_0 = 1$. The reader shall verify in Exercise 7.6 that C_i is a non-increasing sequence. Its limit is non-zero.

From the above, it is clear that 0 is an approximate zero of the second kind if and only if $q \le 1/2$. Now, if we clear denominators and rearrange terms in $(1+\alpha-\sqrt{\Delta})/(1+\alpha+\sqrt{\Delta}) = 1/2$, we obtain the second degree polynomial

$$2\alpha^2 - 13\alpha + 2 = 0.$$

This has solutions $(13 \pm \sqrt{17})/2$. When $0 \leq \alpha \leq \alpha_0 = (13 - \sqrt{17})/2$, the polynomial values are positive and hence $q \leq 1/2$. $\square$

Proof of Theorem 7.15. Let $\beta = \beta(\mathbf{f}, \mathbf{x}_0)$ and $\gamma = \gamma(\mathbf{f}, \mathbf{x}_0)$. Let $h_{\beta\gamma}$ and the sequence t_i be as in Proposition 7.16. By construction, $\|\mathbf{x}_1 - \mathbf{x}_0\| = \beta = t_1 - t_0$. We use the following notations:

$$\beta_i = \beta(\mathbf{f}, \mathbf{x}_i) \text{ and } \gamma_i = \gamma(\mathbf{f}, \mathbf{x}_i).$$

Those will be compared to

$$\hat{\beta}_i = \beta(h_{\beta\gamma}, t_i)) \text{ and } \hat{\gamma}_i = \gamma(h_{\beta\gamma}, t_i)).$$

Induction hypothesis: $\beta_i \leq \hat{\beta}_i$ and for all $l \geq 2$,

$$\|D\mathbf{f}(\mathbf{x}_i)^{-1} D^l \mathbf{f}(\mathbf{x}_i)\| \leq -\frac{h_{\beta\gamma}^{(l)}(t_i)}{h'_{\beta\gamma}(t_i)}.$$

The initial case when $i = 0$ holds by construction. So let us assume that the hypothesis holds for i. We will estimate

$$\beta_{i+1} \leq \|D\mathbf{f}(\mathbf{x}_{i+1})^{-1} D\mathbf{f}(\mathbf{x}_i)\| \|D\mathbf{f}(\mathbf{x}_i)^{-1} \mathbf{f}(\mathbf{x}_{i+1})\| \tag{7.12}$$

and

$$\gamma_{i+1} \leq \|D\mathbf{f}(\mathbf{x}_{i+1})^{-1} D\mathbf{f}(\mathbf{x}_i)\| \frac{\|D\mathbf{f}(\mathbf{x}_i)^{-1} D^k \mathbf{f}(\mathbf{x}_{i+1})\|}{k!}. \tag{7.13}$$

By construction, $\mathbf{f}(\mathbf{x}_i) + D\mathbf{f}(\mathbf{x}_i)(\mathbf{x}_{i+1} - \mathbf{x}_i) = 0$. The Taylor expansion of $\mathbf{f}$ at $\mathbf{x}_i$ is therefore

$$D\mathbf{f}(\mathbf{x}_i)^{-1}\mathbf{f}(\mathbf{x}_{i+1}) = \sum_{k\geq 2} \frac{D\mathbf{f}(\mathbf{x}_i)^{-1} D^k \mathbf{f}(\mathbf{x}_i)(\mathbf{x}_{i+1} - \mathbf{x}_i)^k}{k!}$$

Passing to norms,

$$\|D\mathbf{f}(\mathbf{x}_i)^{-1}\mathbf{f}(\mathbf{x}_{i+1})\| \le \frac{\beta_i^2\gamma_i}{1-\gamma_i}$$

The same argument shows that

$$-\frac{h_{\beta\gamma}(t_{i+1})}{h'_{\beta\gamma}(t_i)} = \frac{\beta(h_{\beta\gamma},t_i)^2\gamma(h_{\beta\gamma},t_i)}{1-\gamma(h_{\beta\gamma},t_i)}$$

From Lemma 7.9,

$$\|D\mathbf{f}(\mathbf{x}_{i+1})^{-1}D\mathbf{f}(\mathbf{x}_i)\| \le \frac{(1-\beta_i\gamma_i)^2}{\psi(\beta_i\gamma_i)}.$$

Also, computing directly,

$$\frac{h'_{\beta\gamma}(t_{i+1})}{h'_{\beta\gamma}(t_i)} = \frac{(1-\hat{\beta}\hat{\gamma})^2}{\psi(\hat{\beta}\hat{\gamma})}. \tag{7.14}$$

We established that

$$\beta_{i+1} \le \frac{\beta_i^2\gamma_i(1-\beta_i\gamma_i)}{\psi(\beta_i\gamma_i)} \le \frac{\hat{\beta}_i^2\hat{\gamma}_i(1-\hat{\beta}_i\hat{\gamma}_i)}{\psi(\hat{\beta}_i\hat{\gamma}_i)} = \hat{\beta}_{i+1}.$$

Now the second part of the induction hypothesis:

$$D\mathbf{f}(\mathbf{x}_i)^{-1}D^l\mathbf{f}(\mathbf{x}_{i+1}) = \sum_{k\ge 0}\frac{1}{k!}\frac{D\mathbf{f}(\mathbf{x}_i)^{-1}D^{k+l}\mathbf{f}(\mathbf{x}_i)(\mathbf{x}_{i+1}-\mathbf{x}_i)^k}{k+l}$$

Passing to norms and invoking the induction hypothesis,

$$\|D\mathbf{f}(\mathbf{x}_i)^{-1}D^l\mathbf{f}(\mathbf{x}_{i+1})\| \le \sum_{k\ge 0} -\frac{h_{\beta\gamma}^{(k+l)}(t_i)\hat{\beta}_i^k}{k!h'_{\beta\gamma}(t_i)}$$

and then using Lemma 7.9 and (7.14),

$$\|D\mathbf{f}(\mathbf{x}_{i+1})^{-1}D^l\mathbf{f}(\mathbf{x}_{i+1})\| \le \frac{(1-\hat{\beta}_i\hat{\gamma}_i)^2}{\psi(\hat{\beta}_i\hat{\gamma}_i)}\sum_{k\ge 0} -\frac{h_{\beta\gamma}^{(k+l)}(t_i)\hat{\beta}_i^k}{k!h'_{\beta\gamma}(t_i)}.$$

A direct computation similar to (7.14) shows that

$$-\frac{h_{\beta\gamma}^{(k+l)}(t_{i+1})}{k!h'_{\beta\gamma}(t_{i+1})} = \frac{(1-\hat{\beta}_i\hat{\gamma}_i)^2}{\psi(\hat{\beta}_i\hat{\gamma}_i)} \sum_{k\geq 0} -\frac{h_{\beta\gamma}^{(k+l)}(t_i)\hat{\beta}_i^k}{k!h'_{\beta\gamma}(t_i)}.$$

and since the right-hand-terms of the last two equations are equal, the second part of the induction hypothesis proceeds. Dividing by $l!$, taking $l-1$-th roots and maximizing over all l, we deduce that $\gamma_i \leq \hat{\gamma}_i$.

Proposition 7.17 then implies that $\mathbf{x}_0$ is an approximate zero.

The second and third statement follow respectively from

$$\|\mathbf{x}_0 - \zeta\| \leq \beta_0 + \beta_1 + \cdots = \zeta_1$$

and

$$\|\mathbf{x}_1 - \zeta\| \leq \beta_1 + \beta_2 + \cdots = \zeta_1 - \beta.$$

□

The same issues as in Theorem 7.5 arise. First of all, we actually proved a sharper statement. Namely,

Theorem 7.18. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let*

$$\alpha \leq 3 - 2\sqrt{2}.$$

Define

$$r = \frac{1 + \alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha}.$$

Let $\mathbf{x}_0 \in \mathcal{D}$ be such that $\alpha(\mathbf{f}, \mathbf{x}_0) \leq \alpha$ and assume furthermore that $B(\mathbf{x}_0, r\beta(\mathbf{f}, \mathbf{x}_0)) \subseteq \mathcal{D}$. Then, the sequence $\mathbf{x}_{i+1} = N(\mathbf{f}, \mathbf{x}_i)$ is well defined, and there is a zero $\zeta \in \mathcal{D}$ of $\mathbf{f}$ such that

$$\|\mathbf{x}_i - \zeta\| \leq q^{2^i-1} \frac{1-\eta}{1-\eta q^{2^i-1}} r\beta(\mathbf{f}, \mathbf{x}_0).$$

for η and q as in Proposition 7.16.

	1/32	1/16	1/10	1/8	$\frac{13-3\sqrt{17}}{4}$
1	4.854	3.683	2.744	2.189	1.357
2	14.472	10.865	7.945	6.227	3.767
3	33.700	25.195	18.220	14.41	7.874
4	72.157	53.854	38.767	29.648	15.881
5	149.71	111.173	79.861	60.864	31.881
6	302.899	225.811	162.49	123.295	63.881

Table 7.2: Values of $-log_2(\|\mathbf{x}_i - \zeta\|/\beta)$ in function of α and i.

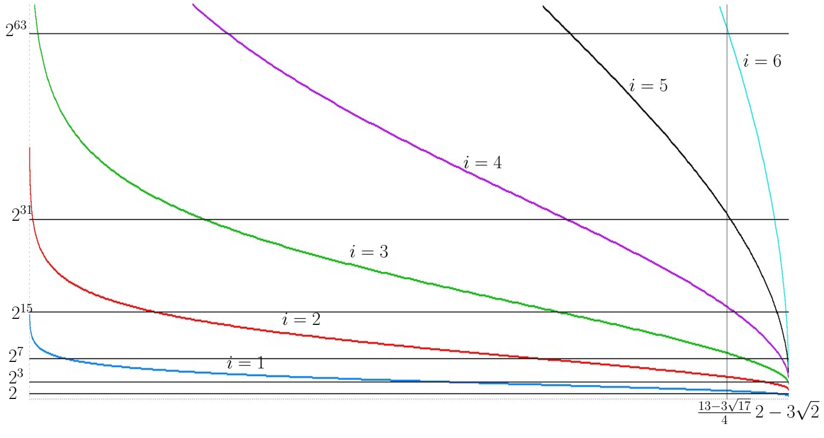

Figure 7.5: Values of $-log_2(\|\mathbf{x}_i - \zeta\|/\beta)$ in function of α for $i = 1$ to 6.

Table 7.2 and Figure 7.5 show how fast $\|\mathbf{x}_i - \zeta\|/\beta$ decreases in terms of α and i.

The final issue is robustness. There is no obvious modification of the proof of Theorem 7.15 to provide a nice statement, so we will rely on Theorem 7.12 indeed.

Theorem 7.19. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \rightarrow \mathbb{F}$ be an analytic map between Banach spaces. Let δ, α and $u0$ satisfy*

$$0 \leq 2\delta < u_0 = \frac{r\alpha}{(1-r\alpha)\psi(r\alpha)} < 2 - \frac{\sqrt{14}}{2}$$

with $r = \frac{1+\alpha-\sqrt{1-6\alpha+\alpha^2}}{4\alpha}$. Assume that

1.
$$B = B\left(\mathbf{x}_0, 2r\beta(\mathbf{f},\mathbf{x}_0)\right) \subseteq \mathcal{D}.$$

2. $\mathbf{x}_0 \in B$, and the sequence $\mathbf{x}_i$ satisfies

$$\|\mathbf{x}_{i+1} - N(\mathbf{f},\mathbf{x}_i)\| \frac{r\beta(f,x_0)}{(1-r\alpha)\psi(r\alpha)} \leq \delta$$

3. The sequence u_i is defined inductively by

$$u_{i+1} = \frac{u_i^2}{\psi(u_i)} + \delta.$$

Then the sequences u_i and $\mathbf{x}_i$ are well-defined for all i, $\mathbf{x}_i \in \mathcal{D}$, and

$$\frac{\|\mathbf{x}_i - \zeta\|}{\|\mathbf{x}_1 - \mathbf{x}_0\|} \leq \frac{ru_i}{u_0} \leq r\max\left(2^{-2^i+1}, 2\frac{\delta}{u_0}\right).$$

Numerically, $\alpha_0 = 0.074, 290\cdots$ satisfies the hypothesis of the Theorem. A version of this theorem (not as sharp, and another metric) appeared as Theorem 2 in [56].

The following Lemma will be useful:

Lemma 7.20. *Assume that $u = \gamma(\mathbf{f},\mathbf{x})\|\mathbf{x}-\mathbf{y}\| \leq 1 - \sqrt{2}/2$. Then,*

$$\gamma(\mathbf{f},\mathbf{y}) \leq \frac{\gamma(\mathbf{f},\mathbf{x})}{(1-u)\psi(u)}.$$

Proof. In order to estimate the higher derivatives, we expand:

$$\frac{1}{l!} D\mathbf{f}(\mathbf{x})^{-1} D^l \mathbf{f}(\mathbf{y}) = \sum_{k \ge 0} \binom{k+l}{l} \frac{D\mathbf{f}(\mathbf{x})^{-1} D^{k+l}\mathbf{f}(\mathbf{x})(\mathbf{y}-\mathbf{x})^k}{k+l}$$

and by Lemma 7.6 for $d = l+1$,

$$\frac{1}{l!} \| D\mathbf{f}(\mathbf{x})^{-1} D^l \mathbf{f}(\mathbf{y}) \| \le \frac{\gamma(\mathbf{f},\mathbf{x})^{l-1}}{(1-u)^{l+1}}.$$

Combining with Lemma 7.9,

$$\frac{1}{l!} \| D\mathbf{f}(\mathbf{y})^{-1} D^l \mathbf{f}(\mathbf{y}) \| \le \frac{\gamma(\mathbf{f},\mathbf{x})^{l-1}}{(1-u)^{l-1}\psi(u)}.$$

Taking the $l-1$-th power,

$$\gamma(\mathbf{f},\mathbf{y}) \le \frac{\gamma(\mathbf{f},\mathbf{x})}{(1-u)\psi(u)}.$$

□

Proof of Theorem 7.19. We have necessarily $\alpha < 3 - 2\sqrt{2}$ or r is undefined. Then (Theorem 7.18) there is a zero ζ of $\mathbf{f}$ with $\|\mathbf{x}_0 - \zeta\| \le r\beta(f, x_0)$. Then, Lemma 7.20 implies that $\|\mathbf{x}_0 - \zeta\|\gamma(\mathbf{f},\zeta) \le u_0$. Now apply Theorem 7.12.

□

Exercise 7.6. The objective of this exercise is to show that C_i is non-increasing.

1. Show the following trivial lemma: *If* $0 \le s < a \le b$, *then* $\frac{a-s}{b-s} \le \frac{a}{b}$.
2. Deduce that $q \le \eta$.
3. Prove that $C_{i+1}/C_i \le 1$.

Exercise 7.7. Show that

$$\zeta_1 \gamma(\zeta_1) = \frac{1+\alpha-\sqrt{\Delta}}{3-\alpha+\sqrt{\Delta}} \frac{1}{\psi\left(\frac{1+\alpha-\sqrt{\Delta}}{4}\right)}.$$

Chapter 8

Condition number theory

8.1 Linear equations

The following classical theorem in linear algebra is known as the *singular value decomposition* (*svd* for short).

Theorem 8.1. *Let* $A : \mathbb{R}^n \mapsto \mathbb{R}^m$ *(resp.* $\mathbb{C}^n \to \mathbb{C}^m$*) be linear. Then, there are* $\sigma_1 \geq \cdots \geq \sigma_r > 0$, $r \leq m, n$, *such that*

$$A = U \Sigma V^*$$

with $U \in O(m)$ *(resp.* $U(m)$*),* $V \in O(n)$ *(resp.* $U(n)$*) and* $\Sigma_{ij} = \sigma_i$ *for* $i = j \leq r$ *and* 0 *otherwise.*

It is due to Sylvester (real $n \times n$ matrices) and to Eckart and Young [37] in the general case, now exercise 8.1 below.

Gregorio Malajovich, *Nonlinear equations*.
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

Σ is a $m \times n$ matrix. It is possible to rewrite this in an 'economical' formulation with Σ an $r \times r$ matrix, U and V orthogonal (resp. unitary) $m \times r$ and $n \times r$ matrices. The numbers $\sigma_1, \dots, \sigma_r$ are called *singular values* of A. They may be computed by extracting the positive square root of the non-zero eigenvalues of A^*A or AA^*, whatever matrix is smaller. The operator and Frobenius norm of A may be written in terms of the σ_i's:

$$\|A\|_2 = \sigma_1 \qquad \|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}.$$

The discussion and the results above hold when A is a linear operator between finite dimensional inner product spaces. It suffices to choose an orthonormal basis, and apply Theorem 8.1 to the corresponding matrix.

When $m = n = r$, $\|A^{-1}\|_2 = \sigma_n$. In this case, the *condition number* of A for linear solving is defined as

$$\kappa(A) = \|A\|_* \|A^{-1}\|_{**}.$$

The choice of norms is arbitrary, as long as operator and vector norms are consistent. Two canonical choices are

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 \text{ and } \kappa_D(A) = \|A\|_F \|A^{-1}\|_2.$$

The second choice was suggested by Demmel [35]. Using that definition he obtained bounds on the probability that a matrix is poorly conditioned. The exact probability distribution for the most usual probability measures in matrix space was computed in [38].

Assume that $A(t)\mathbf{x}(t) \equiv \mathbf{b}(t)$ is a family of problems and solutions depending smoothly on a parameter t. Differentiating implicitly,

$$\dot{A}\mathbf{x} + A\dot{\mathbf{x}} = \dot{\mathbf{b}}$$

which amounts to

$$\dot{\mathbf{x}} = A^{-1}\dot{\mathbf{b}} - A^{-1}\dot{A}\mathbf{x}.$$

Passing to norms and to relative errors, we quickly obtain

$$\frac{\|\dot{\mathbf{x}}\|}{\|\dot{\mathbf{x}}\|} \le \kappa_D(A) \left(\frac{\|\dot{A}\|_F}{\|A\|_F} + \frac{\|\dot{\mathbf{b}}\|}{\|\mathbf{b}\|} \right).$$

This bounds the relative error in the solution $\mathbf{x}$ in terms of the relative error in the coefficients. The usual paradigm in numerical linear algebra dates from [81] and [86]. After the rounding-off during computation, we obtain the exact solution of a perturbed system. Bounds for the perturbation or *backward error* are found through line by line analysis of the algorithm. The output error or *forward error* is bounded by the backward error, times the condition number.

Condition numbers provide therefore an important metric invariant for numerical analysis problems. A geometric interpretation in the case of linear equation solving is:

Theorem 8.2. *Let A be a non-degenerate square matrix.*

$$\|A^{-1}\|_2 = \min_{\det(A+B)=0} \|B\|_F$$

In particular, this implies that

$$\kappa_D(A)^{-1} = \min_{\det(A+B)=0} \frac{\|B\|_F}{\|A\|_F}$$

A pervading principle in the subject is: *the inverse of the condition number is related to the distance to the ill-posed problems.*

It is possible to define the condition number for a full-rank non-square matrix by

$$\kappa_D(A) = \|A\|_F \ \sigma_{\min(m,n)}(A)^{-1}.$$

Theorem 8.3. *[Eckart and Young,* [36]*] Let A be an $m \times n$ matrix of rank r. Then,*

$$\sigma_r(A)^{-1} = \min_{\sigma_r(A+B)=0} \|B\|_F.$$

In particular, if $r = \min(m, n)$,

$$\kappa_D(A)^{-1} = \min_{\sigma_r(A+B)=0} \frac{\|B\|_F}{\|A\|_F}.$$

Exercise 8.1. Prove Theorem 8.1. Hint: let u, v, σ such that $Av = \sigma u$ with σ maximal, $\|u\| = 1$, $\|v\| = 1$. What can you say about $A_{|v^\perp}$?

Exercise 8.2. Prove Theorem 8.3.

Exercise 8.3. Assume furthermore that $m < n$. Show that the same interpretation for the condition number still holds, namely the norm of the perturbation of *some* solution is bounded by the condition number, times the perturbation of the input.

8.2 The linear term

As in Chapter 5, let M be an analytic manifold and let $\mathcal{F}$ be a non-degenerate fewspace of holomorphic functions from M to C. A possibly trivial homogenization group H acts on M, and $f(h\mathbf{x}) = \chi(h)f(\mathbf{x})$ for all $f \in \mathcal{F}$, $\mathbf{x} \in M$, where $\chi(h)$ is a multiplicative character. Furthermore, we assume that M/H is an n-dimensional manifold.

Given $\mathbf{x} \in M$, $\mathcal{F}_{\mathbf{x}}$ denotes the space of functions $f \in \mathcal{F}$ vanishing at $\mathbf{x}$. Using the kernel notation, $\mathcal{F}_{\mathbf{x}} = K(\cdot, \mathbf{x})^{\perp}$. The later is non-zero by Definition 5.2(2).

Let $\mathbf{x} \in M$ and $f \in \mathcal{F}_{\mathbf{x}}$. The derivative of f at $\mathbf{x}$ is

$$Df(\mathbf{x})\mathbf{u} \mapsto \langle f(\cdot), D_{\bar{\mathbf{x}}}K(\cdot, \mathbf{x})\mathbf{u}\rangle_{\mathcal{F}} = \langle f(\cdot), P_x D_{\bar{\mathbf{x}}}K(\cdot, \mathbf{x})\mathbf{u}\rangle_{\mathcal{F}_{\mathbf{x}}}$$

where $P_x : \mathcal{F} \to \mathcal{F}_x$ is the orthogonal projection operator (Lemma 5.10). Note that since $\mathcal{F}$ is a linear space, $D_{\bar{\mathbf{x}}}K(\cdot, \mathbf{x})$ and $P_{\mathbf{x}}D_{\bar{\mathbf{x}}}K(\cdot, \mathbf{x})$ are also elements of $\mathcal{F}$.

Lemma 8.4. *Let $L = L_{\mathbf{x}} : \mathcal{F} \to T_x M^*$ be defined by*

$$L_{\mathbf{x}}(f) : u \mapsto \left\langle f(\cdot), \frac{1}{\sqrt{K(\mathbf{x}, \mathbf{x})}} P_{\mathbf{x}} D_{\bar{\mathbf{x}}} K(\cdot, \mathbf{x})\bar{u} \right\rangle_{\mathcal{F}} .$$

Then L is onto, and $L_{|\ker L^{\perp}}$ is an isometry.

Proof. Recall that the metric in M is the pull-back of the Fubini-Study metric in $\mathcal{F}$ by $\mathbf{x} \mapsto K(\cdot, \mathbf{x})$. The adjoint of $L = L_{\mathbf{x}}$ is

$$\begin{array}{rccl} L^* : & T_{\mathbf{x}}M & \to & \mathcal{F}^*, \\ & u & \mapsto & \left(f \mapsto \left\langle f(\cdot), \frac{1}{\sqrt{K(\mathbf{x},\mathbf{x})}} P_{\mathbf{x}} D_{\bar{\mathbf{x}}} K(\cdot, \mathbf{x})\bar{u} \right\rangle_{\mathcal{F}} \right). \end{array}$$

Thus, for all $f, g \in \mathcal{F}$,

$$\langle L^* f(\cdot), L^* g(\cdot) \rangle_{\mathcal{F}^*} = \langle L^* f(\cdot)^*, L^* g(\cdot)^* \rangle_{\mathcal{F}} = \langle f(\cdot), g(\cdot) \rangle_{\mathbf{x}}.$$

This says that L^* is unitary, hence it has zero kernel and is an isometry onto its image. Thus (Theorem 8.1) $L_{| \ker L^\perp}$ is an isometry. □

8.3 The condition number for unmixed systems

Let $\mathbf{f} = (f_1, \dots, f_s) \in \mathcal{F}^s$. Let $K(\cdot, \cdot)$ and $L = L_{\mathbf{x}}$ be as above. We define now

$$\begin{array}{rccc} \mathbf{L} = \mathbf{L}_{\mathbf{x}} : & \mathcal{F}^s & \to & L(T_{\mathbf{x}} M, \mathbb{C}^s), \\ & (f_1, \dots, f_s) & \mapsto & \begin{bmatrix} L_{\mathbf{x}}(f_1) \\ \vdots \\ L_{\mathbf{x}}(f_s) \end{bmatrix}. \end{array}$$

The space $L(T_{\mathbf{x}} M, \mathbb{C}^s)$ is endowed with 'Frobenius norm',

$$\left\| \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_s \end{bmatrix} \right\|_F^2 = \sum_{i=1}^{s} \|\theta_i\|_{\mathbf{x}}^2$$

each θ_i interpreted as a 1-form, that is an element of $T_{\mathbf{x}} M^*$. An immediate consequence of Lemma 8.4 is

Lemma 8.5. $\mathbf{L}_{\mathbf{x}}$ *is onto, and* $\mathbf{L}_{| \ker \mathbf{L}^\perp}$ *is an isometry.*

The condition number of $\mathbf{f}$ at $\mathbf{x}$ is defined by

$$\mu(\mathbf{f}, \mathbf{x}) = \|\mathbf{f}\| \, \sigma_{\min(n,s)}(\mathbf{L}_{\mathbf{x}}(\mathbf{f}))^{-1}.$$

We will see in the next section that when $\mathcal{F} = \mathcal{H}_{d,d,\cdots,d}$ and $n = s$, this is exactly the Shub-Smale condition number of [70], known as the normalized condition number μ_{norm} in [20].

Theorem 8.6 (Condition number theorem, unmixed)**.** *Let* $\mathbf{f} \in \mathcal{F}^s$. *Let* $r = \min(n, s)$. *Then*

$$\mu(\mathbf{f}, \mathbf{x})^{-1} = \min_{\substack{\mathbf{g} \in \mathcal{F}_{\mathbf{x}} \\ \mathrm{rank}(D(\mathbf{f}+\mathbf{g})(\mathbf{x}))<r}} \frac{\|\mathbf{g}\|}{\|\mathbf{f}\|}.$$

This theorem by Malajovich and Rojas [58, 59] generalizes a theorem by Shub and Smale (see Theorem 8.10 below and comments) to the exponential sum setting.

Proof.

$$\begin{aligned} \mu(\mathbf{f}, x)^{-1} &= \frac{1}{\|\mathbf{f}\|} \sigma_{\min(n,s)}(\mathbf{L}_x(\mathbf{f})) \\ &= \frac{1}{\|\mathbf{f}\|} \sigma_{\min(n,s)}(A) \end{aligned}$$

where $A = \mathbf{L}_x(\mathbf{f})_{| \ker \mathbf{L}_x^\perp}$. By Theorem 8.3,

$$\mu(\mathbf{f}, x)^{-1} = \min_{det(A+B)=0} \frac{\|B\|_F}{\|A\|_F}.$$

Let B be such that the minimum is attained. By Lemma 8.5 there is $\mathbf{h} \perp \ker \mathbf{L}_x$ with $L_x(\mathbf{h}) = B$ and $\|\mathbf{h}\|_{\mathcal{F}_x} = \|B\|_F$. Hence,

$$\mu(\mathbf{f}, \mathbf{x})^{-1} = \min_{\mathrm{rank}(D(\mathbf{f}+\mathbf{h})(\mathbf{x}))<r} \frac{\|\mathbf{h}\|}{\|\mathbf{f}\|}.$$

□

Here is a consequence of Theorem 8.6. Recall that $\mathcal{F}_A$ is the space of linear combinations of $e^{\mathbf{ax}}$, that we assimilate to sparse polynomials in $\mathbf{y} = \mathbf{e}^{\mathbf{x}}$.

Theorem 8.7 (Malajovich and Rojas)**.** *Assume that* $\mathbf{f} \in \mathcal{F}_A^n$ *is a normally distributed, zero average and unit variance random variable. Then,*

$$\begin{aligned} &\mathrm{Prob}\left[\mu(f, z) \geq \epsilon^{-1} \textit{for some } z \in (\mathbb{C} \setminus 0)^n \textit{ with } \mathbf{f}(e^{\mathbf{x}}) = 0\right] \leq \\ &\qquad\qquad \leq Bn^3(n+1)(\#A-1)(\#A-2)\epsilon^4 \end{aligned}$$

where $B = n!\mathrm{Vol}(\mathcal{A})$ *is Kushnirenko's bound.*

(See [58]).

8.4 Condition numbers for homogeneous systems

We consider now a possibly unmixed situation. Let $\mathbf{f} \in \mathcal{H}_{d_1} \times \cdots \times \mathcal{H}_{d_n}$, where each f_i is homogeneous in $n+1$ variables. Let $M = \mathbb{C}^{n+1} \setminus \{0\}$, $H = \mathbb{C}_\times$ and thus $M/H = \mathbb{P}^n$.

Projective space is endowed with the Fubini-Study metric $\langle \cdot, \cdot \rangle$. Each of the $\mathcal{H}_{d_i}$ has reproducing kernel $K_i(\mathbf{x}, \mathbf{y}) = (x_0\bar{y}_0 + \cdots + x_n\bar{y}_n)^{d_i}$ and therefore (Exercise 5.5) induces a metric $\langle \cdot, \cdot \rangle_{\mathbb{P}^n, i} = d_i \langle \cdot, \cdot \rangle$.

Lemma 8.8. *Let $L = L_{i\mathbf{x}} : \mathcal{H}_{d_i} \to T^*_{\mathbf{x}}(\mathbb{P}^n)$ be defined by*

$$L_{i\mathbf{x}}(f) : u \mapsto \frac{1}{\sqrt{d_i}} \left\langle f(\cdot), \frac{1}{\sqrt{K(\mathbf{x}, \mathbf{x})}} P_{\mathbf{x}} D_{\bar{\mathbf{x}}} K(\cdot, \mathbf{x}) \bar{u} \right\rangle_{\mathcal{H}_{d_i}} .$$

Then L is onto, and $L_{|\ker L^\perp}$ is an isometry.

Proof. If we assume the $\langle \cdot, \cdot \rangle_{\mathbb{P}^n, i}$ norm on $T^*_{\mathbf{x}}(\mathbb{P}^n)$, Lemma 8.4 implies that the operator above is onto and $L_{|\ker L^\perp}$ is $d_i^{-1/2}$ times an isometry.

For vectors, the relation between Fubini-Study and $\mathcal{H}_{d_i}$-induced norm is

$$\|\mathbf{u}\| = \frac{1}{\sqrt{d_i}} \|\mathbf{u}\|_i .$$

For covectors, it is therefore

$$\|\omega\| = \sqrt{d_i} \|\omega\|_i .$$

Hence, we deduce that $L_{|\ker L^\perp}$ is an isometry, when Fubini-Study metric is assumed on $\mathbb{P}^n$. □

Now we define

$$\begin{array}{cccc} \mathbf{L}_{\mathbf{x}} : & \mathcal{F}^s & \to & L(T_{\mathbf{x}} M, \mathbb{C}^s), \\ & (f_1, \dots, f_s) & \mapsto & \begin{bmatrix} L_{1\mathbf{x}}(f_1) \\ \vdots \\ L_{s\mathbf{x}}(f_s) \end{bmatrix} . \end{array}$$

As before,

Lemma 8.9. $\mathbf{L_x}$ *is onto, and* $\mathbf{L}_{|\ker \mathbf{L}^\perp}$ *is an isometry.*

The condition number of $\mathbf{f}$ at $\mathbf{x}$ is defined by

$$\mu(\mathbf{f},\mathbf{x}) = \|\mathbf{f}\| \ \sigma_{\min(n,s)}(\mathbf{L_x}(f))^{-1}.$$

When $n = s$, this is precisely the Shub-Smale condition number:

$$\mu(\mathbf{f},\mathbf{x}) = \|\mathbf{f}\|_{\mathcal{H}_\mathbf{d}} \left\| (D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp})^{-1} \begin{bmatrix} \sqrt{d_1}\|\mathbf{x}\|_2^{d_1-1} & & \\ & \ddots & \\ & & \sqrt{d_n}\|\mathbf{x}\|_2^{d_n-1} \end{bmatrix} \right\|_2 . \tag{8.1}$$

Theorem 8.10 (Condition number theorem, homogeneous)**.** *Let* $\mathbf{f} \in \mathcal{F}_\mathbf{x} = (\mathcal{H}_{d_1} \times \cdots \times \mathcal{H}_{d_s})_\mathbf{x}$*. Let* $r = \min(n, s)$*. Then*

$$\mu(\mathbf{f},\mathbf{x})^{-1} = \min_{\substack{\mathbf{g}\in\mathcal{F}_\mathbf{x} \\ \operatorname{rank}(D(\mathbf{f}+\mathbf{g})(\mathbf{x}))<r}} \frac{\|\mathbf{g}\|}{\|\mathbf{f}\|}.$$

The proof is the same as in the unmixed case, and will be omitted.

8.5 Condition numbers in general

We consider now the general case. M is a holomorphic manifold, and $\mathcal{F}_1, \dots, \mathcal{F}_s$ are possibly different non-degenerate fewspaces of holomorphic functions on M . A possibly trivial group H of homogenizations acts on M, in such a way that $f_i(h\mathbf{x}) = \chi_i(h) f_i(\mathbf{x})$ for $f_i \in \mathcal{F}_i$, $\mathbf{x} \in M$. The quotient M/H is assumed to be a n-dimensional holomorphic manifold.

We denote by $K_i(\cdot,\cdot)$, ω_i and $\langle\cdot,\cdot\rangle_{\mathbb{P}^n,i}$ the corresponding invariants. This is a highly unfamiliar situation. We have several metrics for just one manifold. We will *choose* an arbitrary metric and assume that there are real numbers $0 \le e_i \le d_i$ such that for all $\mathbf{x} \in M/H$, for all $\mathbf{u} \in T_\mathbf{x} M$,

$$e_i \|\mathbf{u}\|^2 \le \|\mathbf{u}\|_i^2 \le d_i \|\mathbf{u}\|^2.$$

For covectors $\theta \in T^*_\mathbf{x} M$, we will have

$$\frac{1}{d_i}\|\theta\|^2 \le \|\theta\|_i^2 \le \frac{1}{e_i}\|\theta\|^2$$

Example 8.11. As in the previous section, let $M = \mathbb{C}^{n+1} \setminus \{0\}$, $H = \mathbb{C}_\times$ and $\mathcal{F}_i = \mathcal{H}_{D_i}$. In that case, $M \setminus H = \mathbb{P}^n$, and we set $\langle \cdot, \cdot \rangle_{\mathbb{P}^n}$ equal to the Fubini-Study metric. In that case, $e_i = d_i = D_i$.

Example 8.12. Assume that $\mathcal{F}_1, \dots \mathcal{F}_s$ are non-degenerate fewspaces and that M/H is compact. Let

$$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_1 + \cdots + \langle \cdot, \cdot \rangle_s.$$

There we can take $d_i = 1$. Because $\mathcal{F}_i$ is a non-degenerate fewspace we know that $\langle \cdot, \cdot \rangle_i$ is non-degenerate. By compactness, $e_i > 0$.

In [58], we introduced this mysterious local invariant:

Definition 8.13. Let $\langle \cdot, \cdot \rangle$ be Hermitian inner products in an n-dimensional complex vector space $\mathbb{E}$. Their *mixed dilation* is

$$\Delta = \min_{T \in L(\mathbb{E}, \mathbb{C}^n)} \max_i \frac{\max_{\|T\mathbf{u}\|=1} \langle \mathbf{u}, \mathbf{u} \rangle_i}{\min_{\|T\mathbf{u}\|=1} \langle \mathbf{u}, \mathbf{u} \rangle_i}.$$

Finiteness of Δ follows from the fact that the fraction in its expression is always ≥ 1 and finite. The reader can check that the minimum is attained for some T.

The quotient manifold M/H or a compact subset therein may be endowed with a 'minimal dilation metric', namely

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{x}} = \mathbf{v}^* T^* T \mathbf{u}$$

where T is a point of minimum of the dilation at that point $\mathbf{x}$. This metric is arbitrary up to a multiple, so we may scale the metric so that, for instance,

$$\operatorname{tr} \langle \cdot, \cdot \rangle = \sum \langle \cdot, \cdot \rangle_i$$

Open Problem 8.14. Under what conditions this local metric extends to a Hermitian metric on all of M/H? It would be nice to find a uniform bound for the dilation that is polynomially bounded in the input size.

From now on, we fix a Hermitian metric $\langle \cdot, \cdot \rangle$ on M/H for reference.

Lemma 8.15. *Let* $L = L_{i\mathbf{x}} : \mathcal{F}_i \rightarrow T^*_{\mathbf{x}}(M/H)$ *be defined by*

$$L_{i\mathbf{x}}(f) : u \mapsto \frac{1}{\sqrt{d_i}} \left\langle f(\cdot), \frac{1}{\sqrt{K(\mathbf{x},\mathbf{x})}} P_{\mathbf{x}} D_{\bar{\mathbf{x}}} K(\cdot, \mathbf{x}) \bar{u} \right\rangle_{\mathcal{F}_i} .$$

Then L *is onto, and* $L_{|\ker L^\perp}$ *satisfies:*

$$\sqrt{\frac{e_i}{d_i}} \|f\| \leq \|L_{|\ker L^\perp} f\|_{T^*_{\mathbf{x}}(M/H)} \leq \|f\|$$

Again,

$$\begin{array}{rccc} \mathbf{L}_{\mathbf{x}} : & \mathcal{F}_1 \times \cdots \times \mathcal{F}_s & \rightarrow & L(T_x M, \mathbb{C}^s), \\ & (f_1, \dots, f_s) & \mapsto & \begin{bmatrix} L_{1\mathbf{x}}(f_1) \\ \vdots \\ L_{s\mathbf{x}}(f_s) \end{bmatrix} . \end{array}$$

As before,

Lemma 8.16. $\mathbf{L}_x$ *is onto, and*

$$\left(\min \sqrt{e_i/d_i} \right) \|\mathbf{h}\| \leq \|\mathbf{L}_{|\ker \mathbf{L}^\perp} \mathbf{h}\| \leq \|\mathbf{h}\|$$

The condition number of $\mathbf{f}$ at $\mathbf{x}$ is defined by

$$\mu(\mathbf{f}, \mathbf{x}) = \|\mathbf{f}\| \ \left(\sigma_{\min(n,s)}(\mathbf{L}_{\mathbf{x}}(\mathbf{f})) \right)^{-1} .$$

By construction and the implicit (inverse) function theorem,

Proposition 8.17. *Let* $\mathbf{f}_t \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_s$ *a one-parameter family, with* $\mathbf{f}_0(\mathbf{x}_0) = 0$*. If* $s \leq n$*, then there is locally a solution* $\mathbf{x}_t$*,* $\mathbf{f}_t(\mathbf{x}_t)$ *with*

$$\|\dot{\mathbf{x}}_t\| \leq \frac{1}{\min \sqrt{d_i}} \mu(\mathbf{f}_0, \mathbf{x}_t) \|\dot{\mathbf{f}}_t\|$$

Moreover, we have:

Theorem 8.18 (Condition number theorem)**.** *Let*

$$\mathbf{f} \in \mathcal{F}_{\mathbf{x}} = (\mathcal{F}_1 \times \cdots \times \mathcal{F}_s)_{\mathbf{x}} .$$

Let $r = \min(n, s)$. *Then*

$$\left(\min \sqrt{\frac{e_i}{d_i}}\right) \min_{\substack{g \in \mathcal{F}_{\mathbf{x}} \\ \operatorname{rank}(D(\mathbf{f}+\mathbf{h})(\mathbf{x}))<r}} \frac{\|\mathbf{h}\|}{\|\mathbf{f}\|} \leq \mu(\mathbf{f},\mathbf{x})^{-1} \leq \min_{\substack{\mathbf{h} \in \mathcal{F}_{\mathbf{x}} \\ \operatorname{rank}(D(\mathbf{f}+\mathbf{h})(\mathbf{x}))<r}} \frac{\|\mathbf{h}\|}{\|\mathbf{f}\|}.$$

Proof.

$$\begin{aligned} \mu(\mathbf{f},\mathbf{x})^{-1} &= \frac{1}{\|\mathbf{f}\|}\sigma_{\min(n,s)}(\mathbf{L}_{\mathbf{x}}(\mathbf{f})) \\ &= \frac{1}{\|\mathbf{f}\|}\sigma_{\min(n,s)}(A) \end{aligned}$$

where $A = \mathbf{L}_{\mathbf{x}}(\mathbf{f})_{|\ker \mathbf{L}_{\mathbf{x}}^{\perp}}$. By Theorem 8.3,

$$\mu(\mathbf{f},\mathbf{x})^{-1} = \frac{1}{\|\mathbf{f}\|} \min_{det(A+B)=0} \|B\|_F.$$

Let B be such that the minimum is attained. By Lemma 8.16 there is $\mathbf{g} \perp \ker \mathbf{L}_x$ with $\mathbf{L}_x(\mathbf{h}) = B$ and

$$\min \sqrt{\frac{e_i}{d_i}} \|\mathbf{h}\|_{\mathcal{F}_{\mathbf{x}}} \leq \|B\|_F \leq \|\mathbf{h}\|_{\mathcal{F}_{\mathbf{x}}}$$

Hence,

$$\left(\min \sqrt{\frac{e_i}{d_i}}\right) \min_{\operatorname{rank}(D(\mathbf{f}+\mathbf{h})(\mathbf{x}))<r} \frac{\|\mathbf{h}\|}{\|\mathbf{f}\|} \leq \mu(\mathbf{f},\mathbf{x})^{-1} \leq \min_{\operatorname{rank}(D(\mathbf{f}+\mathbf{h})(\mathbf{x}))<r} \frac{\|\mathbf{h}\|}{\|\mathbf{f}\|}.$$

□

The definition of the condition number and the sharpness of the above theorem depend upon an arbitrary choice of the metric $\langle \cdot, \cdot \rangle$. This motivates the introduction of an *invariant* condition number, namely

$$\underline{\mu}(\mathbf{f},\mathbf{x}) = \left(\min_{\substack{\mathbf{h} \in \mathcal{F}_{\mathbf{x}} \\ \operatorname{rank}(D(\mathbf{f}+\mathbf{h})(\mathbf{x}))<r}} \frac{\|\mathbf{h}\|}{\|\mathbf{f}\|} \right)^{-1}.$$

We always have

$$\underline{\mu}(\mathbf{f},\mathbf{x}) \leq \mu(\mathbf{f},\mathbf{x}) \leq \max \sqrt{\frac{d_i}{e_i}} \underline{\mu}(\mathbf{f},\mathbf{x}).$$

8.6 Inequalities about the condition number

The following is easy:

Lemma 8.19. *Assume that* $\|\mathbf{f}\| = \|\mathbf{g}\| = 1$. *Then*

$$\underline{\mu}(\mathbf{f}, \mathbf{x})^{-1} - \|\mathbf{f} - \mathbf{g}\| \le \underline{\mu}(\mathbf{g}, \mathbf{x})^{-1} \le \underline{\mu}(\mathbf{f}, \mathbf{x})^{-1} + \|\mathbf{f} - \mathbf{g}\|$$

Definition 8.20. A *symmetry group* G is a Lie group acting on M/H and leaving ω, $\omega_1, \dots, \omega_n$ invariant. It acts *transitively* iff for all $\mathbf{x}, \mathbf{y} \in M/H$ there is $Q \in G$ such that $G\mathbf{x} = \mathbf{y}$. The action is *smooth* if $Q, \mathbf{x} \mapsto Q\mathbf{x}$ is smooth.

The action of G in M/H induces an action on each $\mathcal{F}_i$, by

$$f_i \overset{Q}{\rightsquigarrow} f_i \circ Q^{-1}.$$

When each $f \mapsto f \circ Q$ is an isometry, we say that G acts on $\mathcal{F}_i$ by isometries. In this later case, μ and $\bar{\mu}$ are G-invariants.

Example 8.21. The group $U(n+1)$ is a symmetry group acting smoothly and transitively on $\mathbb{P}^n$. It acts on each $\mathcal{H}_{d_i}$ by isometries.

Proposition 8.22. *Let G be a compact, connected symmetry group acting smoothly and transitively on M/H, such that the induced action into the $\mathcal{F}_i$ is by isometries.*

Then, there is D such that for all $\mathbf{f} \in \mathcal{F}$ and $Q \in G$, $\|\mathbf{f}\| = 1$,

$$\|\mathbf{f} - \mathbf{f} \circ Q^{-1}\| \le D d(\mathbf{x}, Q\mathbf{x})$$

where d denotes Riemannian distance. In the particular case $\mathcal{F} = \mathcal{H}_{\mathbf{d}}$ and $G = U(n+1)$, $D = \max d_i$.

Proof. The existence of D is easy: take $Q(t)$ so that $Q(t)\mathbf{x}$ is a minimizing geodesic between $\mathbf{x}$ and $Q\mathbf{x}$. Since the action is smooth,

$$\mathbf{f}_i \circ Q_t^* : \mathbf{x} \mapsto \langle \mathbf{f}_i(\cdot), K_i(\cdot, Q_t^* \mathbf{x}) \rangle$$

is also smooth. Hence

$$D = \sup_{i, \dot{Q} \in T_I G} \|DK_i(\cdot, \dot{Q}\mathbf{x})\|$$

For the particular case of homogeneous systems, we consider $f_i \circ U(t)^*(\cdot) \in \mathcal{H}_{d_i}$ in function of t. We will compute its derivative at $t = 0$. We write down $f_i(\mathbf{x})$ as a tensor, using the notation of Exercise 5.3:

$$f_i(\mathbf{x}) = \sum_{0 \leq j_k \leq n} T_{j_1 \cdots j_{d_i}} x_{j_1} x_{j_2} \cdots x_{j_{d_i}}$$

We can pick coordinates so that

$$U(t) = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} \oplus I_{n-k}$$

Its derivative at $t = 0$ is

$$\dot{U} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \oplus 0_{n-k}.$$

So the derivative of f_i at zero is

$$\dot{f}_i(\mathbf{x}) = \sum_{0 \leq j_k \leq n} \sum_{k=1}^{D} \begin{cases} -T_{j_1 \cdots j_{d_i}} \frac{x_1}{x_0} x_{j_1} x_{j_2} \cdots x_{j_{d_i}} & \text{if } j_k = 0 \\ T_{j_1 \cdots j_{d_i}} \frac{x_0}{x_1} x_{j_1} x_{j_2} \cdots x_{j_{d_i}} & \text{if } j_k = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Rearranging terms and writing $\mathbf{J} = [j_0, \dots, j_{d_i}]$,

$$\dot{f}_i(\mathbf{x}) = \sum_{0 \leq j_k \leq n} x_{j_1} x_{j_2} \cdots x_{j_{d_i}} \sum_{k=1}^{d_i} \begin{cases} -T_{\mathbf{J}+\mathrm{e}_k} & \text{if } j_k = 0 \\ T_{\mathbf{i}-\mathrm{e}_k} & \text{if } j_k = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Comparing the two sides,

$$\|\dot{f}_i\| \leq d_i \|f_i\|.$$

so

$$\|\dot{\mathbf{f}}\| \leq D \|\mathbf{f}\|.$$

□

Theorem 8.23. *Under the assumptions of Proposition 8.22, Let G be a compact, connected symmetry group acting smoothly and transitively on M/H, such that the induced action into the $\mathcal{F}_i$ is by isometries. Let D be the number of 8.22. Let $\mathbf{f}, \mathbf{g} \in \mathcal{F}$, $\|\mathbf{f}\| = \|\mathbf{g}\| = 1$ and $\mathbf{x}, \mathbf{y} \in M/H$. Then,*

$$\frac{1}{1+u+v}\ \underline{\mu}(\mathbf{f}, \mathbf{x}) \leq \underline{\mu}(\mathbf{g}, \mathbf{y}) \leq \frac{1}{1-u-v}\ \underline{\mu}(\mathbf{f}, \mathbf{x})$$

for $u = \underline{\mu}(\mathbf{f}, \mathbf{x}) D d(\mathbf{x}, \mathbf{y})$ *and* $v = \underline{\mu}(\mathbf{f}, \mathbf{x}) \|\mathbf{f} - \mathbf{g}\|$.

In particular, if $\mathcal{F} = \mathcal{H}_{\mathbf{d}}$*, then* $D = \max d_i$ *and* $\underline{\mu} = \mu$.

This theorem appeared in the context of the Shub-Smale condition number (8.1) in several recent papers [25, 31, 69], with larger constants.

Proof. Let $Q(t)\mathbf{x}$ be a geodesic, such as in Proposition 8.22 with $Q(0)\mathbf{x} = \mathbf{x}$ and $Q(1)\mathbf{x} = \mathbf{y}$. Then,

$$
\begin{aligned}
\underline{\mu}(\mathbf{f}, \mathbf{x})^{-1} &\leq \underline{\mu}(\mathbf{g}, \mathbf{x})^{-1} + \|\mathbf{g} - \mathbf{f}\| \\
&\leq \underline{\mu}(\mathbf{g} \circ Q(1), \mathbf{y})^{-1} + \|\mathbf{g} - \mathbf{f}\| \\
&\leq \underline{\mu}(\mathbf{g}, \mathbf{y})^{-1} + \|g - g \circ Q(1)\| + \|\mathbf{g} - \mathbf{f}\| \\
&\leq \underline{\mu}(\mathbf{g}, \mathbf{y})^{-1} + D d(x, y) + \|\mathbf{g} - \mathbf{f}\|
\end{aligned}
$$

Similarly,

$$
\underline{\mu}(\mathbf{f}, \mathbf{x})^{-1} \geq \underline{\mu}(\mathbf{g}, \mathbf{y})^{-1} - D d(x, y) - \|\mathbf{g} - \mathbf{f}\|
$$

Now we just have to multiply both inequalities by $\underline{\mu}(\mathbf{f}, \mathbf{x}) \underline{\mu}(\mathbf{g}, \mathbf{y})$ and a trivial manipulation finishes the proof. $\square$

Chapter 9

The pseudo-Newton operator

Newton iteration was originally defined on linear spaces, where it makes sense to add a vector to a point. Manifolds in general lack this operation. A standard procedure in geometry is to replace the sum by the exponential map

$$\begin{array}{rccl} \exp : & TM & \rightarrow & M, \\ & (\mathbf{x}, \dot{\mathbf{x}}) & \mapsto & \exp_{\mathbf{x}}(\dot{\mathbf{x}}), \end{array}$$

that is the map such that $\exp_{\mathbf{x}}(t\dot{\mathbf{x}}/\|\dot{\mathbf{x}}\|)$ is a geodesic with speed $\dot{\mathbf{x}}$ at zero. This approach was developed by many authors, such as [82] or [40]. The alpha-theory for the Riemannian Newton operator

$$N^{\mathrm{Riem}}(\mathbf{f}, \mathbf{x}) = \exp_{\mathbf{x}} -D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

appeared in [32]. This approach can be algorithmically cumbersome, as it requires the computation of the exponential map, which in turn

Gregorio Malajovich, *Nonlinear equations.*
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

depends on the connection.

Luckily, it turns out that of the two conditions defining the geodesic, only one is actually relevant for the purpose of Newton iteration: the condition at $t = 0$ should be $\dot{\mathbf{x}}$.

A more general procedure is to replace the exponential map by a *retraction* map $R : TM \to M$ with

$$\frac{\partial}{\partial t}_{|t=0} R(\mathbf{x}, t\dot{\mathbf{x}})\dot{\mathbf{x}}.$$

This is discussed in [1]. A previous example, studied in the literature, is *projective Newton* [20, 68, 70].

Through this chapter and the next, we adopt the following notations. Given a point $\mathbf{x} \in \mathbb{P}^n$ or in a quotient manifold M/H, $\mathbf{X}$ denotes a representative of it in $\mathbb{C}^{n+1}$ (or in M). The class of equivalence of $\mathbf{X}$ may be denoted by $\mathbf{x}$ or by $[\mathbf{X}]$. With this convention, projective Newton is

$$N^{\text{proj}}(\mathbf{f}, \mathbf{x}) = [\mathbf{X} - D\mathbf{f}(\mathbf{X})^{-1}_{\mathbf{X}^\perp}\mathbf{f}(\mathbf{X})].$$

This iteration has advantages and disadvantages. The main disadvantage is that its alpha-theory is much harder than the usual Newton iteration.

In this book, we will follow a different approach. The following operator was suggested by [2]:

$$N^{\text{pseu}}(\mathbf{f}, \mathbf{X}) = \mathbf{X} - D\mathbf{f}(\mathbf{X})^{-1}_{|\ker D\mathbf{f}(\mathbf{X})^\perp}\mathbf{f}(\mathbf{X}).$$

This holds in general for manifolds that are quotient of a linear space (or an adequate subset of it) by a group. For instance, $\mathbb{P}^n$ as quotient of $\mathbb{C}^{n+1} \setminus 0$ by $\mathbb{C}_\times$. In this case, results of convergence and robustness are not harder than in the classical setting [56].

This whole approach was extended to the multi-projective setting in [33]. More precisely, let $n = n_1 + \cdots + n_s - s$ and consider multi-homogeneous polynomials in $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_s)$. Let Ω be the set of $\mathbf{X} \in \mathbb{C}^{n+s}$ such that at least one of the $\mathbf{X}_i$ vanishes. Then we set $M = \mathbb{C}^{n+s} \setminus \Omega$ and $H = (\mathbb{C}_\times)^s$, acting on M by $\mathbf{h}\mathbf{X} = (h_1\mathbf{X}_1, \dots, h_s\mathbf{X}_s)$.

Through this chapter, $\mathcal{F}_1, \dots, \mathcal{F}_n$ will denote spaces of multi-homogeneous polynomials, such that elements of $\mathcal{F}_i$ have degree d_{ij}

in X_j. An alternative definition of Ω is: the set of points $\mathbf{X}$ at $\mathbb{C}^{n+s}$ where axiom 5.2.2 fails, namely the evaluation map at X is the zero map for some $\mathcal{F}_i$.

In order to define the Newton iteration on multiprojective space $\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_s}$, Dedieu and Shub [33] endow $M = \mathbb{C}^{n+s} \setminus \Omega$ with a metric that is H-invariant. Their construction amounts to scaling $\mathbf{X}$ by $\mathbf{h}$ such that $\|h_1\mathbf{X}_1\| = \cdots = \|h_s\mathbf{X}_s\| = 1$ and then

$$N^{\mathrm{pseu}}(\mathbf{f}, \mathbf{x}) = [\mathbf{hX} - D\mathbf{f}(\mathbf{hX})^{-1}_{\ker D\mathbf{f}(\mathbf{hX})^{\perp}}\mathbf{f}(\mathbf{hX})].$$

In this book, we are following a different philosophy. While condition numbers are geometric invariants that live in quotient space (or on manifolds), Newton iteration operates only on linear spaces. Hence we will define

$$N(\mathbf{f}, \mathbf{X}) = \mathbf{X} - D\mathbf{f}(\mathbf{X})^{-1}_{\ker D\mathbf{f}(\mathbf{X})^{\perp}}\mathbf{f}(\mathbf{X})$$

as a mapping from M into itself. It may be undefined for certain values of $\mathbf{X}$. While it coincides with N^{pseu} for values of $\mathbf{X}$ scaled such that $\|\mathbf{X}_1\| = \cdots = \|\mathbf{X}_s\|$, it is not in general a mapping in quotient space. This will allow for iteration of N, **without** rescaling. In chapter 10 we will take care of rescaling the vector $\mathbf{X}$ when convenient, and will say that explicitly.

9.1 The pseudo-inverse

The iteration N^{pseu} is usually expressed in terms of a generalization of the inverse of a matrix:

Definition 9.1. Let A be a matrix, with svd decomposition $A = U\Sigma V^*$ (see Th. 8.1). Its *pseudo-inverse* $A^\dagger$ is

$$A^\dagger = V\Sigma^\dagger U^*$$

where $(\Sigma^\dagger)_{ii} = \Sigma_{ii}^{-1}$ when $\Sigma_{ii} \neq 0$, or zero otherwise.

Note that if A is a rank m, $m \times n$ matrix with $m \leq n$, then $AA^\dagger = I_m$ and $A^\dagger A$ is the orthogonal projection onto $\ker A^\perp$. Moreover,

$$A^\dagger = (AA^*)^{-1}A.$$

Another convenient interpretation is the following: $\mathbf{x} = A^\dagger \mathbf{y}$ is the solution of the least-squares problem:

$$\text{Minimize}\|A\mathbf{x} - \mathbf{y}\|^2 \text{ with } \|\mathbf{x}\|^2 \text{ minimal.}$$

If A is $m \times n$ of full rank, $m \leq n$, then $\mathbf{x}$ is the vector with minimal norm such that $A\mathbf{x} = \mathbf{y}$.

Lemma 9.2 (Minimality property)**.** *Let A be a $m \times n$ matrix of rank m, $m \leq n$. Let Π be a m-dimensional space such that $A_{|\Pi}$ is invertible. Then,*

$$\|A^\dagger\| \leq \|(A_{|\Pi})^{-1}\|.$$

The same definition and results hold for linear operators between inner product spaces.

In particular, when Let $\mathbf{f} \in \mathcal{H}_{\mathbf{d}}$ and $\mathbf{X} \in \mathbb{C}^{n+1}$. Then,

$$D\mathbf{f}(\mathbf{X})^\dagger = \left(D\mathbf{f}(\mathbf{X})_{|\ker D\mathbf{f}(\mathbf{X})^\perp}\right)^{-1}$$

whenever this derivative is invertible. In particular,

$$\|D\mathbf{f}(\mathbf{X})^\dagger\| \leq \| \left(D\mathbf{f}(\mathbf{X})_{|\Pi}\right)^{-1} \|$$

for any hyperplane Π.

While the minimality property is extremely convenient, we will need later the following lower bound:

Lemma 9.3. *Let A be a full rank, $n \times (n+1)$ real or complex matrix. Assume that $w = \|A^\dagger\| \|A - B\| < 1$. Let $\Pi : \ker A^\perp \to \ker B^\perp$ denote orthogonal projection. Then for all $\mathbf{x} \in (\ker A)^\perp$,*

$$\|\Pi \mathbf{x}\| \geq \|\mathbf{x}\| \sqrt{1 - w^2}.$$

In particular, for all $\mathbf{y}$,

$$\|B^\dagger A \mathbf{y}\| \geq \|\mathbf{y}\| \frac{\sqrt{1 - w^2}}{1 + w}.$$

Proof. First of all, pick $\mathbf{b}$ with norm one in $\ker B$. If $\mathbf{b} \in \ker A$ then Π is the identity and we are done. Therefore, assume that $\mathbf{b} \not\in \ker A$.

The kernel of A is then spanned by $\mathbf{b}+\mathbf{c}$, where

$$\mathbf{c} = A^{\dagger}(B-A)\mathbf{b}.$$

From this expression, $\|\mathbf{c}\| \leq w$.

Now, assume without loss of generality that $\mathbf{x} \in \ker A^{\perp}$ has norm one. Since

$$\Pi\mathbf{x} = \mathbf{x} - \mathbf{b}\langle \mathbf{x}, \mathbf{b}\rangle,$$

we bound

$$\|\Pi\mathbf{x}\|^2 = \|\mathbf{x}^2\| - 2|\langle \mathbf{x}, \mathbf{b}\rangle|^2 + \|\mathbf{b}\|^2 |\langle \mathbf{x}, \mathbf{b}\rangle|^2 = 1 - |\langle \mathbf{x}, \mathbf{b}\rangle|^2.$$

Note that $\mathbf{x} \perp \mathbf{b}+\mathbf{c}$ so the latest bound is $1 - |\langle \mathbf{x}, \mathbf{c}\rangle|^2 \geq 1 - w^2$.

In order to prove the lower bound on $\|B^{\dagger}A\mathbf{y}\|$, we write

$$B^{\dagger}A = \Pi B^{-1}_{|\ker A^{\perp}} A.$$

Since $\|A^{\dagger}B_{|\ker A^{\perp}} - I_{\ker A^{\perp}}\| \leq \|A^{\dagger}\|\|B-A\| \leq w$, Lemma 7.8 implies that

$$\|B^{-1}_{|\ker A^{\perp}} A\mathbf{y}\| \geq \|\mathbf{y}\| \frac{1}{1+w}.$$

□

9.2 Alpha theory

We define Smale's invariants in $M = \mathbb{C}^{n+s} \setminus \Omega$ in the obvious way:

$$\beta(\mathbf{f}, \mathbf{X}) = \|D\mathbf{f}(\mathbf{X})^{\dagger}\mathbf{f}(\mathbf{X})\|_2$$

and

$$\gamma(\mathbf{f}, \mathbf{X}) = \sup_{k \geq 2} \left(\frac{\|D\mathbf{f}(\mathbf{X})^{\dagger} D^k \mathbf{f}(\mathbf{X})\|_2}{k!} \right)^{1/(k-1)}.$$

and of course

$$\alpha(\mathbf{f}, \mathbf{X}) = \beta(\mathbf{f}, \mathbf{X})\gamma(\mathbf{f}, \mathbf{X})$$

In the projective case $s = 1$, β scales as $\|\mathbf{X}\|$ while γ scales as $\|\mathbf{X}\|^{-1}$. α is invariant. This is no more true when $s \geq 2$.

We can extend those definitions to projective or multiprojective space by setting $\beta(\mathbf{f}, \mathbf{x}) = \beta(\mathbf{f}, \mathbf{X})$ where $\mathbf{X}$ is scaled such that $\|\mathbf{X}_1\| = \cdots = \|\mathbf{X}_s\| = 1$. (The same for γ and α).

Lemma 7.9 that was crucial for alpha theory. Now it becomes:

Lemma 9.4. *Let $\mathbf{X}, \mathbf{Y} \in M$ and $f \in \mathcal{F}$. Assume that $u = \|\mathbf{X} - \mathbf{Y}\|\gamma(\mathbf{f}, \mathbf{X}) < 1 - \sqrt{2}/2$. Then,*

$$\|D\mathbf{f}(\mathbf{Y})^\dagger D\mathbf{f}(\mathbf{X})\| \leq \frac{(1-u)^2}{\psi(u)}.$$

Proof. Expanding $\mathbf{Y} \mapsto D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{Y})$ around $\mathbf{X}$, we obtain:

$$\begin{aligned} D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{Y}) =& D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{X}) + \\ &+ \sum_{k \geq 2} \frac{1}{k-1!} D\mathbf{f}(\mathbf{X})^\dagger D^k\mathbf{f}(\mathbf{X})(\mathbf{Y}-\mathbf{X})^{k-1}. \end{aligned}$$

Rearranging terms and taking norms, Lemma 7.6 yields

$$\|D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{Y}) - D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{X})\| \leq \frac{1}{(1-\gamma\|\mathbf{Y}-\mathbf{X}\|)^2} - 1.$$

In particular,

$$\begin{aligned} &\|D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{Y})_{|\ker D\mathbf{f}(\mathbf{X})^\perp} - D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{X})_{|\ker D\mathbf{f}(\mathbf{X})^\perp}\| \leq \\ &\qquad \leq \frac{1}{(1-\gamma(\mathbf{f},\mathbf{X})\|\mathbf{Y}-\mathbf{X}\|)^2} - 1. \end{aligned}$$

Now we have full rank endomorphisms of $\ker D\mathbf{f}(\mathbf{X})^\perp$ on the left, so we can apply Lemma 7.8 to get:

$$\|D\mathbf{f}(\mathbf{Y})^{-1}_{|\ker D\mathbf{f}(\mathbf{X})^\perp} D\mathbf{f}(\mathbf{X})\| \leq \frac{(1-u)^2}{\psi(u)}. \tag{9.1}$$

Because of the minimality property of the pseudo-inverse (see Lemma 9.2),

$$\|D\mathbf{f}(\mathbf{Y})^\dagger D\mathbf{f}(\mathbf{X})\| \leq \|D\mathbf{f}(\mathbf{Y})^{-1}_{|\ker D\mathbf{f}(\mathbf{X})^\perp} D\mathbf{f}(\mathbf{X})\|$$

so (9.1) proves the Lemma. □

Here is another useful estimate, that we state for homogeneous systems only:

Lemma 9.5. *Let* $\mathbf{X} \in \mathbb{C}^{n+1}$ *and* $\mathbf{f}, \mathbf{g} \in \mathcal{H}_{\mathbf{d}}$. *Assume that* $v = \frac{\|\mathbf{f}-\mathbf{g}\|}{\|\mathbf{f}\|}\mu(\mathbf{f}, \mathbf{X}) < 1$. *Then, for all* $\mathbf{Y} \perp \ker Df(\mathbf{X})$,

$$\|\mathbf{Y}\| \frac{\sqrt{1-v^2}}{1+v} \leq \|D\mathbf{g}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{X})\mathbf{Y}\| \leq \frac{\|\mathbf{Y}\|}{1-v}.$$

The rightmost inequality holds unconditionally.

Proof. By Lemma 8.9,

$$\left\|D\mathbf{f}(\mathbf{X})^\dagger\right\| \|D\mathbf{g}(\mathbf{X}) - D\mathbf{f}(\mathbf{X})\| \leq \mu(\mathbf{f}, \mathbf{X}) \left\| L_{\mathbf{x}}\left(\frac{\mathbf{g}-\mathbf{f}}{\|\mathbf{f}\|}\right)\right\| \leq v$$

In particular

$$\left\|D\mathbf{f}(\mathbf{X})^\dagger D\mathbf{g}(\mathbf{X})_{\ker D\mathbf{f}(\mathbf{X})^\perp} - I_{\ker D\mathbf{f}(\mathbf{X})^\perp}\right\| \leq v.$$

By Lemmas 9.2 and 7.8,

$$\left\|D\mathbf{g}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{X})Y\right\| \leq \left\|D\mathbf{g}(\mathbf{X})^{-1}_{\ker D\mathbf{f}(\mathbf{X})^\perp} D\mathbf{f}(\mathbf{X})Y\right\| \leq \frac{\|Y\|}{1-v}$$

The lower bound follows from Lemma 9.3:

$$\left\|D\mathbf{g}(\mathbf{X})^\dagger D\mathbf{f}(\mathbf{X})Y\right\| \geq \frac{\|Y\|\sqrt{1-v^2}}{1+v}$$

□

9.3 Approximate zeros

The projective distance is defined in $\mathbb{C}^{n+1}$ by

$$d_{\text{proj}}(\mathbf{X}, \mathbf{Y}) = \inf_{\lambda \in \mathbb{C}_\times} \frac{\|\mathbf{X} - \lambda\mathbf{Y}\|}{\|\mathbf{X}\|}.$$

Since it is scaling invariant, is defines a metric in projective space that is related to the Riemannian distance by

$$d_{\text{proj}}(\mathbf{x}, \mathbf{y}) = \sin(d_{\text{Riem}}(\mathbf{x}, \mathbf{y})) \leq d_{\text{Riem}}(\mathbf{x}, \mathbf{y})$$

In the multi-projective setting, we define

$$d_{\text{proj}}(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{s} d_{\text{proj}}(\mathbf{X}_i, \mathbf{Y}_i)^2}.$$

Again, this is scaling invariant and we have

$$d_{\text{proj}}(\mathbf{x}, \mathbf{y}) \le d\text{Riem}(\mathbf{x}, \mathbf{y})$$

Definition 9.6 (Approximate zero of the first kind)**.** Let $\mathbf{f} \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$, and $\mathbf{z} \in M/H$ with $\mathbf{f}(z) = 0$. An *approximate zero of the first kind* associated to $\mathbf{z}$ is a point $\mathbf{X}_0 \in M$, such that

1. The sequence $(\mathbf{X})_i$ defined inductively by $\mathbf{X}_{i+1} = N_{\text{pseu}}(\mathbf{f}, \mathbf{X}_i)$ is well-defined.

2. $$d_{\text{proj}}(\mathbf{X}_i, \mathbf{Z}) \le 2^{-2^i+1} d_{\text{proj}}(\mathbf{X}_0, \mathbf{Z}).$$

Theorem 9.7 (Smale)**.** *Let $\mathbf{f} \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_s$ and let $\mathbf{Z}$ be a non-degenerate zero of $\mathbf{f}$, scaled such that $\|\mathbf{Z}_1\| = \cdots = \|\mathbf{Z}_s\| = 1$. Let $\mathbf{X}_0$ be scaled such that $d_{\text{proj}}(\mathbf{X}_0, \mathbf{Z}) = \|\mathbf{X}_0 - \mathbf{Z}\|$. If*

$$\|\mathbf{X}_0 - \mathbf{Z}\| \le \frac{3 - \sqrt{7}}{2\gamma(\mathbf{f}, \mathbf{Z})},$$

then $\mathbf{X}_0$ is an approximate zero of the first kind associated to $\mathbf{Z}$.

This is an improvement of Corollary 1 in [33]. The improvement is made possible because we do not rescale $\mathbf{X}_1, \mathbf{X}_2, \ldots$.

Proof of Theorem 7.5. Set $\gamma = \gamma(\mathbf{f}, \mathbf{Z})$, $u_0 = \|\mathbf{X}_0 - \mathbf{Z}\|\gamma$, and let h_γ, (u_i) be as in Lemma 7.10.

We bound

$$\begin{aligned} \|N(\mathbf{f}, \mathbf{X}) - Z\| &= \left\| \mathbf{X} - Z - D\mathbf{f}(\mathbf{X})^\dagger \mathbf{f}(\mathbf{X}) \right\| \\ &\le \|D\mathbf{f}(\mathbf{X})^\dagger\| \|\mathbf{f}(\mathbf{X}) - D\mathbf{f}(\mathbf{X})(\mathbf{X} - \mathbf{Z})\|. \end{aligned} \tag{9.2}$$

The Taylor expansions of $\mathbf{f}$ and $D\mathbf{f}$ around Z are respectively:

$$\mathbf{f}(\mathbf{X}) = D\mathbf{f}(\mathbf{Z})(\mathbf{X} - \mathbf{Z}) + \sum_{k \ge 2} \frac{1}{k!} D^k \mathbf{f}(\mathbf{Z})(\mathbf{X} - \mathbf{Z})^k$$

and

$$D\mathbf{f}(\mathbf{X}) = D\mathbf{f}(\mathbf{Z}) + \sum_{k\geq 2} \frac{1}{k-1!} D^k\mathbf{f}(\mathbf{Z})(\mathbf{X}-\mathbf{Z})^{k-1}.$$

Combining the two equations, above, we obtain:

$$\mathbf{f}(\mathbf{X}) - D\mathbf{f}(\mathbf{X})(\mathbf{X}-\mathbf{Z}) = \sum_{k\geq 2} \frac{k-1}{k!} D^k\mathbf{f}(Z)(\mathbf{X}-\mathbf{Z})^k.$$

Using Lemma 7.6 with $d = 2$, the rightmost term in (9.2) is bounded above by

$$\begin{aligned} \|\mathbf{f}(\mathbf{X}) - D\mathbf{f}(\mathbf{X})(\mathbf{X}-\mathbf{Z})\| &\leq \sum_{k\geq 2} (k-1)\gamma^{k-1}\|\mathbf{X}-\mathbf{Z}\|^k \\ &= \frac{\gamma\|\mathbf{X}-\mathbf{Z}\|^2}{(1-\gamma\|\mathbf{X}-\mathbf{Z}\|)^2}. \end{aligned} \tag{9.3}$$

Combining Lemma 9.4 and (9.3) in (9.2), we deduce that

$$\|N(\mathbf{f},\mathbf{X}) - \mathbf{Z}\| \leq \frac{\gamma\|\mathbf{X}-\mathbf{Z}\|^2}{\psi(\gamma\|\mathbf{X}-\mathbf{Z}\|)}.$$

By induction, $u_i \leq \gamma\|\mathbf{X}_i - \mathbf{Z}_i\|$. When $u_0 \leq (3-\sqrt{7})/2$, we obtain as in Lemma 7.10 that

$$\frac{d_{\text{proj}}(\mathbf{X}_i,\mathbf{Z})}{d_{\text{proj}}(\mathbf{X}_0,\mathbf{Z})} \leq \frac{\|\mathbf{X}_i-\mathbf{Z}\|}{\|\mathbf{X}_0-\mathbf{Z}\|} \leq \frac{u_i}{u_0} \leq 2^{-2^i+1}.$$

We have seen in Lemma 7.10 that the bound above fails for $i = 1$ when $u_0 > (3-\sqrt{7})/2$. □

The same comments as the ones for theorem 7.5 are in order. We actually proved stronger theorems, see exercises.

Exercise 9.1. Show that the projective distance in $\mathbb{P}^n$ satisfies the triangle inequality. Same question in the multi-projective case.

Exercise 9.2. Restate and prove Theorem 7.11 in the context of pseudo-Newton iteration.

Exercise 9.3. Restate and prove Theorem 7.12 in the context of pseudo-Newton iteration.

9.4 The alpha theorem

Definition 9.8 (Approximate zero of the second kind)**.** Let $\mathbf{f} \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. An *approximate zero of the second kind* associated to $\mathbf{z} \in M/H$, $\mathbf{f}(\mathbf{z}) = 0$, is a point $\mathbf{X}_0 \in M$, scaled s.t. $\|(\mathbf{X}_0)_1\| = \cdots = \|(\mathbf{X}_0)_s\| = 1$, and satisfying the following conditions:

1. The sequence $(\mathbf{X})_i$ defined inductively by $\mathbf{X}_{i+1} = N(\mathbf{f}, \mathbf{X}_i)$ is well-defined (each $\mathbf{X}_i$ belongs to the domain of $\mathbf{f}$ and $D\mathbf{f}(\mathbf{X}_i)$ is invertible and bounded).

2. $$d_{\text{proj}}(\mathbf{X}_{i+1}, \mathbf{X}_i) \le 2^{-2^i+1} d_{\text{proj}}(\mathbf{X}_1, \mathbf{X}_0).$$

3. $\lim_{i\to\infty} \mathbf{X}_i = \mathbf{Z}$.

Theorem 9.9. *Let* $\mathbf{f} \in \mathcal{H}_{\mathbf{d}}$. *Let*

$$\alpha \le \alpha_0 = \frac{13 - 3\sqrt{17}}{4}.$$

Define

$$r_0 = \frac{1 + \alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha} \text{ and } r_1 = \frac{1 - 3\alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha}.$$

Let $\mathbf{X}_0 \in \mathbb{C}^{n+s}$, $\|(\mathbf{X}_0)_1\| = \cdots = \|(\mathbf{X}_0)_s\| = 1$, *be such that* $\alpha(\mathbf{f}, \mathbf{X}_0) \le \alpha$. *Then,*

1. $\mathbf{X}_0$ *is an approximate zero of the second kind, associated to some zero* $z \in \mathbb{P}^n$ *of* $\mathbf{f}$.

2. *Moreover,* $d_{\text{proj}}(\mathbf{X}_0, \mathbf{z}) \le r_0 \beta(\mathbf{f}, \mathbf{X}_0)$.

3. *Let* $\mathbf{X}_1 = N(\mathbf{f}, \mathbf{x}_0)$. *Then* $d_{\text{proj}}(\mathbf{X}_1, \mathbf{z}) \le r_1 \frac{\beta(\mathbf{f}, \mathbf{X}_0)}{1 - \beta(\mathbf{f}, \mathbf{X}_0))}$.

Proof of Theorem 9.9. Let $\beta = \beta(\mathbf{f}, \mathbf{X}_0)$ and $\gamma = \gamma(\mathbf{f}, \mathbf{X}_0)$. Let $h_{\beta\gamma}$ and the sequence t_i be as in Proposition 7.16. By construction of the pseudo-Newton operator, $d_{\text{proj}}(\mathbf{X}_1, \mathbf{X}_0) = \beta = t_1 - t_0$. We use the following notations:

$$\beta_i = \beta(\mathbf{f}, \mathbf{X}_i) \text{ and } \gamma_i = \gamma(\mathbf{f}, \mathbf{X}_i).$$

Those will be compared to

$$\hat{\beta}_i = \beta(h_{\beta\gamma}, t_i)) \text{ and } \hat{\gamma}_i = \gamma(h_{\beta\gamma}, t_i)).$$

Induction hypothesis: $\beta_i \leq \hat{\beta}_i$ and for all $l \geq 2$,

$$\|D\mathbf{f}(\mathbf{X}_i)^\dagger D^l \mathbf{f}(\mathbf{X}_i)\| \leq -\frac{h^{(l)}_{\beta\gamma}(t_i)}{h'_{\beta\gamma}(t_i)}.$$

The initial case when $i = 0$ holds by construction. So let us assume that the hypothesis holds for i. We will estimate

$$\beta_{i+1} \leq \|D\mathbf{f}(\mathbf{X}_{i+1})^\dagger D\mathbf{f}(\mathbf{X}_i)\| \|D\mathbf{f}(\mathbf{X}_i)^\dagger \mathbf{f}(\mathbf{X}_{i+1})\| \tag{9.4}$$

and

$$\gamma_{i+1} \leq \|D\mathbf{f}(\mathbf{X}_{i+1})^\dagger D\mathbf{f}(\mathbf{X}_i)\| \frac{\|D\mathbf{f}(\mathbf{X}_i)^\dagger D^k \mathbf{f}(\mathbf{X}_{i+1})\|}{k!}. \tag{9.5}$$

By construction, $\mathbf{f}(\mathbf{X}_i) + D\mathbf{f}(\mathbf{X}_i)(\mathbf{X}_{i+1} - \mathbf{X}_i) = 0$. The Taylor expansion of $\mathbf{f}$ at $\mathbf{X}_i$ is therefore

$$D\mathbf{f}(\mathbf{X}_i)^\dagger \mathbf{f}(\mathbf{X}_{i+1}) = \sum_{k \geq 2} \frac{D\mathbf{f}(\mathbf{X}_i)^\dagger D^k \mathbf{f}(\mathbf{X}_i)(\mathbf{X}_{i+1} - \mathbf{X}_i)^k}{k!}$$

Passing to norms,

$$\|D\mathbf{f}(\mathbf{X}_i)^\dagger \mathbf{f}(\mathbf{X}_{i+1})\| \leq \frac{\beta_i^2 \gamma_i}{1 - \gamma_i}$$

while we know from (7.14) that

$$\hat{\beta}_{i+1} = -\frac{h_{\beta\gamma}(t_{i+1})}{h'_{\beta\gamma}(t_i)} = \frac{\beta(h_{\beta\gamma}, t_i)^2 \gamma(h_{\beta\gamma}, t_i)}{1 - \gamma(h_{\beta\gamma}, t_i)} = \frac{\hat{\beta}_i^2 \hat{\gamma}_i}{1 - \hat{\gamma}_i}$$

From Lemma 9.4,

$$\|D\mathbf{f}(\mathbf{X}_{i+1})^\dagger D\mathbf{f}(\mathbf{X}_i)\| \leq \frac{(1 - \beta_i \gamma_i)^2}{\psi(\beta_i \gamma_i)}.$$

Thus,

$$\beta_{i+1} \leq \frac{\beta_i^2 \gamma_i (1 - \beta_i \gamma_i)}{\psi(\beta_i \gamma_i)} \tag{9.6}$$

By (7.14) and induction,

$$\beta_{i+1} \leq \frac{\hat{\beta}_i^2 \hat{\gamma}_i (1 - \hat{\beta}_i \hat{\gamma}_i)}{\psi(\hat{\beta}_i \hat{\gamma}_i)} = \hat{\beta}_{i+1}.$$

Now the second part of the induction hypothesis:

$$D\mathbf{f}(\mathbf{X}_i)^\dagger D^l \mathbf{f}(\mathbf{X}_{i+1}) = \sum_{k \geq 0} \frac{1}{k!} \frac{D\mathbf{f}(\mathbf{X}_i)^\dagger D^{k+l} \mathbf{f}(\mathbf{X}_i)(\mathbf{X}_{i+1} - \mathbf{X}_i)^k}{k+l}$$

Passing to norms and invoking the induction hypothesis,

$$\|D\mathbf{f}(\mathbf{X}_i)^\dagger D^l \mathbf{f}(\mathbf{X}_{i+1})\| \leq \sum_{k \geq 0} -\frac{h_{\beta\gamma}^{(k+l)}(t_i) \hat{\beta}_i^k}{k! h'_{\beta\gamma}(t_i)}$$

and then using Lemma 9.4 and (7.14),

$$\|D\mathbf{f}(\mathbf{X}_{i+1})^\dagger D^l \mathbf{f}(\mathbf{X}_{i+1})\| \leq \frac{(1 - \hat{\beta}_i \hat{\gamma}_i)^2}{\psi(\hat{\beta}_i \hat{\gamma}_i)} \sum_{k \geq 0} -\frac{h_{\beta\gamma}^{(k+l)}(t_i) \hat{\beta}_i^k}{k! h'_{\beta\gamma}(t_i)}.$$

A direct computation similar to (7.14) shows that

$$-\frac{h_{\beta\gamma}^{(k+l)}(t_{i+1})}{k! h'_{\beta\gamma}(t_{i+1})} = \frac{(1 - \hat{\beta}_i \hat{\gamma}_i)^2}{\psi(\hat{\beta}_i \hat{\gamma}_i)} \sum_{k \geq 0} -\frac{h_{\beta\gamma}^{(k+l)}(t_i) \hat{\beta}_i^k}{k! h'_{\beta\gamma}(t_i)}.$$

and since the right-hand-terms of the last two equations are equal, the second part of the induction hypothesis proceeds. Dividing by $l!$, taking $l-1$-th roots and maximizing over all l, we deduce that $\gamma_i \leq \hat{\gamma}_i$.

Proposition 7.17 then implies that $\mathbf{X}_0$ is an approximate zero.

Let $Z = \lim_{k \to \infty} N^k(\mathbf{f}, \mathbf{Z})$. The second statement follows from

$$d_{\text{proj}}(\mathbf{X}_0, \mathbf{Z}) \leq \|\mathbf{X}_0 - \mathbf{Z}\| \leq \beta_0 + \beta_1 + \cdots = r_0 \beta.$$

For the third statement, note that $\|X_1\| \geq (1-\beta)$. Then

$$d_{\text{proj}}(\mathbf{X}_1, \mathbf{Z}) \leq \frac{\|\mathbf{X}_1 - \mathbf{Z}\|}{\|\mathbf{X}_1\|} \leq \frac{\beta_1 + \beta_2 + \cdots}{1-\beta} \leq \frac{r_1 \beta}{1-\beta}.$$

□

9.5 Alpha-theory and conditioning

The reproducing kernel $K_i(\mathbf{X}, \mathbf{Y})$ associated to a fewspace $\mathcal{F}$ is analytic in $\mathbf{X}$. This implies that $\bar{\mathbf{X}} \mapsto K_i(\cdot, \mathbf{X})$ is also an analytic map from M to $\mathcal{F}_i$. Let ρ_i denote its radius of convergence, with respect to a scaling invariant metric. Then, the value of ρ_i at one point $\mathbf{X}$ determines the value for all $\mathbf{X}$.

In general, if

$$\rho_i^{-1} = \limsup_{k \geq 2} \left(\frac{\|D^k K_i(\cdot, \mathbf{X})\|}{k!} \right)^{1/(k-1)}$$

is finite, then

$$R_i^{-1} = \sup_{k \geq 2} \left(\frac{\|D^k K_i(\cdot, \mathbf{X})\|}{k!} \right)^{1/(k-1)}$$

is also finite. This will provide bounds for the higher derivatives of K.

Through this section, we assume for convenience that $M/H = \mathbb{P}^n$ and $\mathcal{F}_i = \mathcal{H}_{d_i}$. The unitary group $U(n+1)$ acts transitively on $\mathbb{P}_n$. Since $K_i = (\sum \mathbf{X}_i \bar{\mathbf{Y}}_i)^{d_i}$, $\rho_i = \infty$ for polynomials are globally analytic.

Taking $\mathbf{X} = \mathrm{e}_0$ and then scaling, we obtain

$$\begin{aligned} \left(\frac{\|D^k K_i(\cdot, \mathbf{X})\|}{k!} \right)^{\frac{1}{k-1}} &= \|\mathbf{X}\| \left(\frac{d_i(d_i-1)\cdots(d_i-k+1)}{k!} \right)^{\frac{1}{k-1}} \\ &\leq \frac{d_i}{2} \|\mathbf{X}\| \end{aligned}$$

with equality for $k=2$.

Proposition 9.10. *Assume that* $\mathbf{f} \in \mathcal{H}_{\mathbf{d}}$*, Let* $R_1, \dots, R_s$ *be as above, and assume the canonical norm in* $\mathbb{C}^{n+1}$*. Then, for* $\|\mathbf{X}\| = 1$*,*

$$\left(\frac{\|D^k \mathbf{f}(\mathbf{X})\|}{k!} \right)^{1/(k-1)} \leq \|\mathbf{f}\|^{1/(k-1)} \frac{D}{2}$$

with $D = \max d_i$*.*

Proof.

$$D^k f_i(\mathbf{X}) = \langle f_i(\cdot), D^k K_i(\cdot, \bar{\mathbf{X}}) \rangle.$$

□

Thus,

Theorem 9.11 (Higher derivative estimate)**.** *Let* $\mathbf{f} \in \mathcal{H}_{\mathbf{d}}$ *and* $\mathbf{X} \in \mathbb{C}^{n+1} \setminus \{0\}$*. Then,*

$$\gamma(\mathbf{f}, \mathbf{X}) \leq \|\mathbf{X}\| \frac{(\max d_i)^{3/2}}{2} \mu(\mathbf{f}, \mathbf{x})$$

Proof. Without loss of generality, scale $\mathbf{X}$ so that $\|\mathbf{X}\| = 1$. For each $k \geq 2$,

$$\begin{aligned} \left(\frac{\|D\mathbf{f}(\mathbf{X})^\dagger D^k \mathbf{f}(\mathbf{X})\|}{k!} \right)^{\frac{1}{k-1}} &\leq \|D\mathbf{f}(\mathbf{X})^{-1}_{|\mathbf{X}^\perp}\|^{1/(k-1)} \|\mathbf{f}\|^{1/(k-1)} \frac{D}{2} \\ &\leq \|\mathbf{L}_{\mathbf{x}}(\mathbf{f})^{-1}\|^{1/(k-1)} \|\mathbf{f}\|^{1/(k-1)} \frac{D^{1+\frac{1}{k-1}}}{2}. \\ &\leq \frac{D^{3/2}}{2} \mu(\mathbf{f}, \mathbf{x})^{1/(k-1)} \\ &\leq \frac{D^{3/2}}{2} \mu(\mathbf{f}, \mathbf{x}) \end{aligned}$$

using that $\mu(\mathbf{f}, \mathbf{x}) \geq \sqrt{n} \geq 1$. □

Exercise 9.4. Show that Proposition 9.10 holds for multi-homogeneous polynomials, with

$$D = \max d_{ij}.$$

Exercise 9.5. Let $\mathbf{f}$ denote a system of multi-homogeneous equations. Let $\mathbf{X} \in \mathbb{C}^{n+s} \setminus \Omega$, scaled such that $\|\mathbf{X}_i\| = 1$. Show that,

$$\gamma(\mathbf{f}, \mathbf{X}) \leq \|\mathbf{X}\| \frac{(\max d_{ij})^{3/2}}{2} \underline{\mu}(\mathbf{f}, \mathbf{x}).$$

Chapter 10

Homotopy

Several recent breakthroughs made Smale's 17^{th} problem an active, fast-moving subject. The first part of the *Bézout saga* [70–74] culminated in the existential proof of a non-uniform, average polynomial time algorithm to solve Problem 1.11. Namely,

Theorem 10.1 (Shub and Smale)**.** *Let $\mathcal{H}_{\mathbf{d}}$ be endowed with the normal (Gaussian) probability distribution $\mathrm{d}\mathcal{H}_{\mathbf{d}}$ with mean zero and variance 1.*

There is a constant c such that, for every n, for every $\mathbf{d} = (d_1, \dots, d_n)$, **there is** *an algorithm to find an approximated root of a random $f \in (\mathcal{H}_{\mathbf{d}}, \mathrm{d}\mathcal{H}_{\mathbf{d}})$ within expected time cN^4, where $N = \dim \mathcal{H}_{\mathbf{d}}$ is the input size.*

This theorem was published in 1994, and motivated the statement of Smale's 17^{th} problem. It was obtained through the painful complexity analysis of a linear homotopy method. Given $\mathbf{F}_0, \mathbf{F}_1 \in \mathcal{H}_{\mathbf{d}}$ and $\mathbf{x}_0$ and approximate zero of $\mathbf{F}_0$, the homotopy method was of the

Gregorio Malajovich, *Nonlinear equations.*
28^o Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2011.

form

$$\mathbf{x}_{i+1} = N^{\text{proj}}(\mathbf{F}_{t_i}, \mathbf{x}_i),$$

for

$$\mathbf{F}_t = (1-t)\mathbf{F}_0 + t\mathbf{F}_1, \quad 0 = t_0 \le t_i \le t_\tau = 1.$$

The major difficulty was finding an adequate starting pair $(\mathbf{F}_0, \mathbf{x}_0)$. Only the existence of such a pair was known, without any clue on how to find one in polynomial time.

A minor difficulty was the choice of the t_i. This can be done by trial and error. By doing so, there is no guarantee that one is approximating an actual continuous solution path $\mathbf{F}_t(\mathbf{x}_t) \equiv 0$. This is trouble when attempting to find all the roots of a polynomial system, or when investigating the corresponding Galois group.

In 2006, Carlos Beltrán and Luis Miguel Pardo demonstrated in his doctoral thesis [6, 11] the existence of a good 'questor set' from which an adequate random pair $(\mathbf{F}_0, \mathbf{x}_0)$ could be drawn with a good probability.

A randomized algorithm is said to be of *Vegas* type if it returns an answer with probability $1 - \epsilon$ for some ϵ, and the answer it returns is always correct. This is by opposition to *Monte-Carlo* type algorithms, that would return a correct answer with probability $1 - \epsilon$.

Theorem 10.2 (Beltrán and Pardo)**.** *Let $\epsilon > 0$. Then there is a Vegas type algorithm such that, given $n, \mathbf{d} = d_1, \dots d_n$ and a random $\mathbf{F}_1 \in (\mathcal{H}_{\mathbf{d}}, \mathrm{d}H_{\mathbf{d}})$, finds with probability $1 - \epsilon$ an approximate zero $\mathbf{X}$ for $\mathbf{F}_1$, within expected time $O(N^5 \epsilon^{-2})$, where $N = \dim \mathcal{H}_{\mathbf{d}}$ is the input size.*

This result and its proof was greatly improved in subsequent papers by Beltrán and Pardo such as [13]. The running time was reduced to

$$\mathbb{E}(\tau) = C(\max d_i)^{3/2} nN$$

homotopy steps.

In another development, Peter Bürgisser and Felipe Cucker gave a deterministic algorithm for solving random systems within expected

$$\mathbb{E}(\tau) = N^{O(\log \log N)}$$

homotopy steps. They pointed out that this solves Smale's 17^{th} problem for the 'case' $\max d_i \le n^{\frac{1}{1+\epsilon}}$ while the 'case' $\max d_i \ge n^{1+\epsilon}$ follows from resultant based algorithms such as [67]. When

$$n^{\frac{1}{1+\epsilon}} \le \max d_i \le n^{1+\epsilon},$$

Smale's 17^{th} problem is still open.

Another recent advance are 'condition-length' based algorithms. While previous algorithm have a complexity bound in terms of the line integral of $\mu(\mathbf{F}_t, \mathbf{z}_t)^2$ in $\mathbb{P}(\mathcal{H}_\mathbf{d})$, condition-length algorithms (suggested in [14, 69] and developed in [7, 31] have a complexity bound in terms of a geometric invariant, the *condition length*. This allows to reduce Smale's 17^{th} problem (Open Problem 1.11) to a 'variational' problem.

In the rest of this chapter, I will give a simplified version of the algorithm in [31], together with its complexity analysis. Then, I will discuss how to use this algorithm to obtain results analogous to those of [13] and [25]. In the last section, I will review some recent results on the geometry of the condition metric.

10.1 Homotopy algorithm

Let $\mathbf{d} = (d_1, \dots, d_n)$ be fixed, and set $D = \max d_i$. Recall that $\mathcal{H}_\mathbf{d}$ is the space of homogeneous polynomial systems in n variables of degree $d_1, \dots, d_n$. We want to find solutions $\mathbf{z} \in \mathbb{P}^n$, and those will be represented by elements of $\mathbb{C}^{n+1} \setminus \{0\}$. We keep the convention of the previous chapter, where we set $\mathbf{Z}$ for a representative of $\mathbf{z}$. However, we will prefer representatives with norm one whenever possible.

We will consider an affine path in $\mathcal{H}_\mathbf{d}$ given by

$$\mathbf{F}_t = (1-t)\mathbf{F}_0 + t\mathbf{F}_1$$

where $\mathbf{F}_0$ and $\mathbf{F}_1$ are scaled such that

$$\|\mathbf{F}_0\| = 1 \qquad \mathbf{F}_0 \perp \mathbf{F}_1 - \mathbf{F}_0 \tag{10.1}$$

with an extra bound,

$$\|\mathbf{F}_1 - \mathbf{F}_0\| \le 1. \tag{10.2}$$

Again, $\mathbf{f}_t$ is the equivalence class of $\mathbf{F}_t$ in $\mathbb{P}(\mathcal{H}_\mathbf{d})$. Given representatives for $\mathbf{f}_0$ and $\mathbf{f}_1$, two cases arise: either we can find $\mathbf{F}_0$ and $\mathbf{F}_1$ satisfying (10.1) and (10.2), or we may find $\mathbf{f}_{1/2}$ half-way in projective space such that $(\mathbf{f}_0, \mathbf{f}_{1/2})$ and $(\mathbf{f}_{1/2}, \mathbf{f}_1)$ fall into the previous case. Therefore, (10.2) is not a big limitation.

Let $0 < a < \alpha_0$, where α_0 is the constant of Theorem 9.9. We will say that $\mathbf{X}$ is a (β, μ, a)-*certified* approximate zero of $\mathbf{f}$ if and only if

$$\frac{D^{3/2}}{2} \|\mathbf{X}\|^{-1} \beta(\mathbf{F}, \mathbf{X}) \mu(\mathbf{f}, \mathbf{x}) \le a.$$

This condition implies, in particular (Theorems 9.9 and 9.11) that $\mathbf{X}$ is an approximate zero of the second kind for $\mathbf{f}$.

We address the following computational task:

Problem 10.3 (true lifting)**.** Given $0 \ne \mathbf{F}_0$ and $0 \ne \mathbf{F}_1 \in \mathcal{H}_\mathbf{d}$ satisfying (10.1) and (10.2), and given also a (β, μ, a_0)-certified approximate zero $\mathbf{X}_0$ of $\mathbf{F}_0$, associated to a root $\mathbf{z}_0$, find a (β, μ, a_0)-certified approximate zero of $\mathbf{f}_1$, associated to the zero $\mathbf{z}_1$ where $\mathbf{z}_t$ is continuous and $\mathbf{F}_t(\mathbf{z}_t) \equiv 0$ for $t \in [0, 1]$.

A *true lifting* is not always possible. Moreover, the cost of the algorithm will depends on certain invariant of the path $(\mathbf{f}_t, \mathbf{z}_t)$ that can be infinite. However, we may understand this invariant geometrically.

The set $\mathcal{V} = \{(\mathbf{f}, \mathbf{z}) \in \mathbb{P}(\mathcal{H}_\mathbf{d}) \times \mathbb{P}^n : \mathbf{f}(\mathbf{z}) = 0\}$ is known as the *solution variety* of the problem. The solution variety inherits a metric from the product of the Fubini-Study metrics in $\mathbb{P}(\mathcal{H}_\mathbf{d})$ and $\mathbb{P}^{n+1}$.

The discriminant variety Σ' in $\mathcal{V}$ is the set of critical points for the projection $\pi_1 : V \to \mathcal{H}_\mathbf{d}$. This is a Zariski closed set, hence its complement is path-connected. For a probability one choice of $\mathbf{F}_0, \mathbf{F}_1$, the corresponding path $(\mathbf{f}_t, \mathbf{z}_t)$ exists and keeps a certain distance to this discriminant variety. We will see that in that case, the algorithm succeeds. Before we define the invariant:

Definition 10.4. The *condition length* of the path $(\mathbf{f}_t, \mathbf{z}_t)_{t \in [a,b]} \in \mathcal{V}$ is

$$\mathcal{L}(\mathbf{f}_t; a, b) = \int_a^b \mu(\mathbf{f}_s, \mathbf{z}_s) \|(\dot{\mathbf{f}}_s, \dot{\mathbf{z}}_s)\|_{(\mathbf{f}_s, \mathbf{z}_s)} \, \mathrm{d}s$$

As this is expository material, we will make suppositions about intermediate quantities that need to be computed. Namely, the following operations are assumed to be performed exactly and at unit cost: Sum, subtraction, multiplication, division, deciding $x > 0$, and square root.

In particular, Newton iteration $N(\mathbf{F}, \mathbf{X}) = \mathbf{X} - D\mathbf{F}(\mathbf{X})^{\dagger}\mathbf{F}(\mathbf{X})$ can be computed in $O(n \dim(\mathcal{H}_{\mathbf{d}}))$ operations.

It would be less realistic to assume that we can compute condition numbers (that have an operator norm). Operator norms can be approximated (up to a factor of $\sqrt{n}$) by the Frobenius norm, which is easy to compute. Therefore, let

$$\mu_F(\mathbf{F}, \mathbf{X}) =$$
$$= \|\mathbf{F}\| \left\| D\mathbf{F}(\mathbf{X})^{-1}_{|\mathbf{X}^{\perp}} \begin{bmatrix} \|\mathbf{X}\|^{d_1-1}\sqrt{d_1} & & \\ & \ddots & \\ & & \|\mathbf{X}\|^{d_n-1}\sqrt{d_n} \end{bmatrix} \right\|_F$$

be the 'Frobenius' condition number. It is invariant by scaling, and

$$\mu(\mathbf{f}, \mathbf{x}) \le \mu_F(\mathbf{f}, \mathbf{x}) \le \sqrt{n}\ \mu(\mathbf{f}, \mathbf{x}).$$

Also, we need to define the following quantity:

$$\Phi_{t,\sigma}(\mathbf{X}) = \left\| D\mathbf{F}_t(\mathbf{X})^{\dagger}(\mathbf{F}_\sigma(\mathbf{X}) - \mathbf{F}_t(\mathbf{X})) \right\|.$$

The algorithm will depend on constants $a_0, \alpha, \epsilon_1, \epsilon_2$. The constant a_0 is fixed so that

$$\frac{a_0 + \epsilon_2}{(1-\epsilon_1)^2} = \alpha. \tag{10.3}$$

The value of the other constants was computed numerically (see remark 10.14 below). The constant C will appear as a complexity bound, and depends on the other constants. There is no claim of optimality in the values below:

Constant	Value
α	7.110×10^{-2}
ϵ_1	5.596×10^{-2}
ϵ_2	5.656×10^{-2}
a_0	$6.805, 139, 185, 76 \times 10^{-3}$
C	16.26 (upper bound).

We will need routines to compute the following quantities:

- $S_1(\mathbf{X}, t)$ is the minimal value of $s > t$ with

$$\|\mathbf{F}_s - \mathbf{F}_t\| = \frac{\epsilon_1}{\mu_F(\mathbf{F}_t, \mathbf{X})}.$$

  This can be computed by computing easily with elementary operations and exactly one square root.

- $S_2(\mathbf{X}, t)$ is the maximal value of $s > t$ such that, for all $t < \sigma < s$,

$$\Phi_{t,\sigma}(\mathbf{X}) \le \frac{2\epsilon_2}{D^{3/2}\mu_F(\mathbf{F}_t, \mathbf{X})}$$

  In particular, when $S_2(t)$ is finite,

$$\Phi_{t,S_2(t)}(\mathbf{X}) = \frac{2\epsilon_2}{D^{3/2}\mu_F(\mathbf{F}_t, \mathbf{X})}$$

  Again, S_2 may be computed by elementary operations, and then solving one degree two polynomial (that is, one square root).

> **Algorithm** Homotopy.
> **Input:** $\mathbf{F}_0, \mathbf{F}_1 \in \mathcal{H}_{\mathbf{d}} \setminus \{0\}$, $\mathbf{X}_0 \in \mathbb{C}^{n+1} \setminus \{0\}$.
>
> $i \leftarrow 0$, $t_0 \leftarrow 0$, $\mathbf{X}_0 \leftarrow \frac{1}{\|\mathbf{X}_0\|}\mathbf{X}_0$.
> **Repeat**
> $\quad t_{i+1} \leftarrow \min\left(S_1(\mathbf{X}_i, t_i), S_2(\mathbf{X}_i, t_i), 1\right)$.
> $\quad \mathbf{X}_{i+1} \leftarrow \frac{1}{\|N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)\|} N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)$.
> $\quad i \leftarrow i + 1$.
> **Until** $t_i = 1$.
> **Return** $\mathbf{X} \leftarrow \mathbf{X}_i$

Theorem 10.5 (Dedieu-Malajovich-Shub)**.** *Let $n, D = \max d_i \ge 2$. Assume that $\mathbf{F}_0$ and $\mathbf{F}_1$ satisfy* (10.1), (10.2) *and moreover $\mathbf{X}_0$ is a (β, μ, a_0) certified approximate zero for $\mathbf{F}_0$.*

1. *If the algorithm terminates, then* $\mathbf{X}$ *is a* (β, μ, a_0) *certified approximate zero for* $\mathbf{F}_1$.

2. *If the algorithm terminates, and* $\mathbf{z}_0$ *denotes the zero of* $\mathbf{F}_0$ *associated to* $\mathbf{X}_0$, *then* $\mathbf{z}_1$ *is the zero of* $\mathbf{F}_1$ *associated to* $\mathbf{X}$ *where* $\mathbf{f}_t(\mathbf{z}_t) \equiv 0$ *is a continuous path.*

3. *There is a constant* $C < 16.26$ *such that if the condition length* $\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; 0, 1)$ *is finite, then the algorithm always terminates after at most*
$$1 + Cn^{1/2}D^{3/2}\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; 0, 1) \tag{10.4}$$
steps.

The actual theorem in [31] is stronger, because the algorithm thereby allows for approximations instead of exact calculations. It is more general, as the path does not need to be linear. Also, it is worded in terms of the projective Newton operator N^{proj}. This is why the constants are different. But the important feature of the theorem is an explicit step bound in terms of the condition length, and this is reproduced here.

Remark 10.6. We can easily bound

$$\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; 0, 1) \le \int_0^1 \|\dot{\mathbf{f}}_t\|_{\mathbf{f}_t} \mu(\mathbf{f}_t, \mathbf{z}_t)^2 \, \mathrm{d}t$$

and recover the complexity analysis of previously known algorithms.

Remark 10.7. The factor on $\sqrt{n}$ in the complexity bound comes from the approximation of μ by μ_F. It can be removed at some cost. The price to pay is a more complicated subroutine for norm estimation, and a harder complexity analysis.

10.2 Proof of Theorem 10.5

Towards the proof of Theorem 10.5, we need five technical Lemmas. For the geometric insight, see figure 10.1.

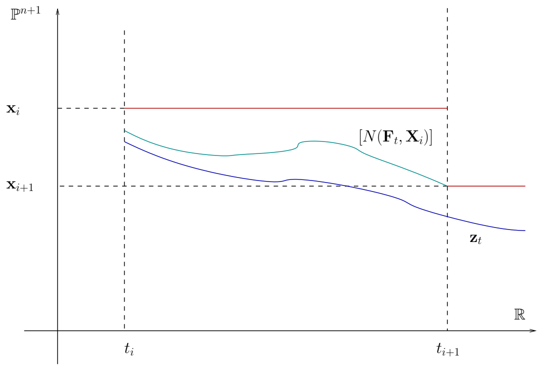

Figure 10.1: The homotopy step. This picture is in projective space. For the picture in linear space, the reader can imagine that he stands at the origin. The points X_{i+1}, $N(F_{t_{i+1}}, X_i)$ and the origin are in the same complex line.

Lemma 10.8. *Assume the conditions of Theorem 10.5. For short, write* $\beta = \beta(\mathbf{F}_{t_i}, \mathbf{X}_i)$ *and* $\mu = \mu(\mathbf{F}_{t_i}, \mathbf{X}_i)$. *If*

$$\frac{D^{3/2}}{2}\beta\mu \leq a_0, \tag{10.5}$$

$$\|\mathbf{F}_t - \mathbf{F}_s\| \leq \frac{\epsilon_1}{\mu}, \text{ and} \tag{10.6}$$

$$\Phi_{t,s}(\mathbf{X}) \leq \frac{2\epsilon_2}{D^{3/2}\mu} \qquad \forall s \in [t_i, t_{i+1}], \tag{10.7}$$

then the following estimates hold for all $s \in [t_i, t_{i+1}]$*:*

$$\mu(\mathbf{f}_s, \mathbf{x}_i) \leq \frac{\mu}{1-\epsilon_1} \tag{10.8}$$

$$\beta(\mathbf{F}_s, \mathbf{X}_i) \leq \frac{2}{D^{3/2}} \frac{(1-\epsilon_1)\alpha}{\mu} \tag{10.9}$$

$$\beta(\mathbf{F}_s, \mathbf{X}_i) \geq \frac{2}{D^{3/2}} \frac{(\epsilon_2 - a_0)\sqrt{1-\epsilon_1^2}}{(1+\epsilon_1)\mu} \tag{10.10}$$

$$\frac{D^{3/2}}{2}\beta(\mathbf{F}_s, \mathbf{X}_i)\mu(\mathbf{f}_s, \mathbf{x}_i) \leq \alpha \tag{10.11}$$

Proof. Because of (10.1), $\|\mathbf{F}_{t_i}\|, \|\mathbf{F}_s\| \geq 1$ and

$$\left\| \frac{\mathbf{F}_{t_i}}{\|\mathbf{F}_{t_i}\|} - \frac{\mathbf{F}_s}{\|\mathbf{F}_s\|} \right\| \leq \|\mathbf{F}_{t_i} - \mathbf{F}_s\| \leq \frac{\epsilon_1}{\mu}$$

Then Lemma 8.22 with $u = 0$, $v = \epsilon_1$ implies (10.8).

For (10.9) and (10.10), we write

$$\begin{aligned} \beta(\mathbf{F}_s, \mathbf{X}_i) = D\mathbf{F}_s(\mathbf{X}_i)^\dagger D\mathbf{F}_{t_i}(\mathbf{X}_i) \big(& D\mathbf{F}_{t_i}(\mathbf{X}_i)^\dagger \mathbf{F}_{t_i}(\mathbf{X}_i) + \\ & + D\mathbf{F}_{t_i}(\mathbf{X}_i)^\dagger (\mathbf{F}_s(\mathbf{X}_i) - \mathbf{F}_{t_i}(\mathbf{X}_i)) \big) . \end{aligned}$$

Let $v = \frac{\|\mathbf{F}_s - \mathbf{F}_{t_i}\|}{\|\mathbf{F}_{t_i}\|}\mu$. By (10.2) $\|\mathbf{F}_{t_i}\| > 1$ so that $v \leq \epsilon_1$. From Lemma 9.5, we deduce that

$$\frac{\sqrt{1-\epsilon_1^2}}{1+\epsilon_1} \left(\frac{2\epsilon_2}{D^{3/2}\mu} - \beta \right) \leq \beta(\mathbf{F}_s, \mathbf{X}_i) \leq \frac{\beta + \frac{2\epsilon_2}{D^{3/2}\mu}}{1-\epsilon_1}$$

Now equation (10.3) implies (10.9) and (10.10). (10.11) is obtained by multiplying (10.8) and (10.9). □

Lemma 10.9. *Under the conditions of Lemma 10.8,*

$$\begin{aligned} \mu(\mathbf{f}_s, [N(\mathbf{F}_s, \mathbf{X}_i)]) &\leq \frac{\mu}{1 - \epsilon_1 - \pi a_0/\sqrt{D}} && (10.12) \\ \beta(\mathbf{F}_s, N(\mathbf{F}_s, \mathbf{X}_i)) &\leq \frac{2}{D^{3/2}} \frac{1-\epsilon_1}{\mu} \frac{1-\alpha}{\psi(\alpha)} \alpha^2 && (10.13) \end{aligned}$$

and

$$\frac{D^{3/2}}{2} \beta(\mathbf{F}_s, N(\mathbf{F}_s, \mathbf{X}_i)) \mu(\mathbf{f}_s, [N(\mathbf{F}_s, \mathbf{X}_i)]) \leq (1 - (1-\epsilon_1)\alpha/2)\, a_0 \tag{10.14}$$

Proof. The proof of (10.12) is similar to the one of (10.8). We need to keep in mind that $\mathbf{X}_{t_i}$ is scaled but $N(\mathbf{F}_s, \mathbf{X}_{t_i})$ is not assumed scaled. Anyway, we know that

$$\|\mathbf{X}_{t_i} - \mathbf{N}(\mathbf{F}_s, \mathbf{X}_{t_i})\| = \beta.$$

Let d_{Riem} denote the Riemannian distance between $\mathbf{x}_{t_i}$ and Newton iteration $[\mathbf{N}(\mathbf{F}_s, \mathbf{X}_{t_i})]$.

$$\sin(d_{\text{Riem}}) = d_{\text{proj}}(\mathbf{X}_{t_i}, N(\mathbf{F}_s, \mathbf{X}_{t_i})) \leq \beta.$$

Because projective space has radius $\pi/2$, we may always bound

$$d_{\text{Riem}}(\mathbf{x}, \mathbf{y}) \leq \frac{\pi}{2} d_{\text{proj}}(\mathbf{x}, \mathbf{y})$$

so that we should set $u = \frac{D\pi}{2}\mu\beta$ in order to apply Theorem 8.23. We obtain

$$\mu(\mathbf{f}_s, [N(\mathbf{F}_s, \mathbf{X}_i)]) \leq \frac{\mu}{1 - \epsilon_1 - \pi a_0/\sqrt{D}}$$

The estimate on (10.13) follows from (9.6). Using (10.11),

$$\beta(\mathbf{F}_s, N(\mathbf{F}_s, \mathbf{X}_i)) \leq \frac{\alpha(1-\alpha)}{\psi(\alpha)} \beta(\mathbf{F}_s, \mathbf{X}_i)$$

The estimate

$$\frac{(1-\epsilon_1)(1-\alpha)}{(1-(1-\epsilon_1)\alpha/2)(1-\epsilon_1 - \pi a_0/\sqrt{2})} \frac{\alpha^2}{\psi(\alpha)} \leq a_0 \tag{10.15}$$

was obtained numerically. It implies (10.14) □

Remark 10.10. (10.15) seems to be the main 'active' constraint for the choice of $\alpha, \epsilon_1, \epsilon_2$.

Lemma 10.11. *Under the conditions of Lemma 10.8,*

$$\mu(\mathbf{f}_s, \mathbf{z}_s) \geq \frac{\mu}{1 + \epsilon_1 + \pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{D}} \tag{10.16}$$

where $r_0 = r_0(\alpha)$ is defined in Theorem 9.9.

Proof. From Theorem 9.9 applied to $\mathbf{F}_s$ and $\mathbf{X}_i$, the projective distance from $\mathbf{X}_i$ to $\mathbf{z}_s$ is bounded above by $r_0(\alpha)\beta(\mathbf{F}_s, \mathbf{X}_i)$. Therefore, we set

$$u = \pi(1-\epsilon_1) r_0(\alpha)\alpha/\sqrt{D} \quad v = \epsilon_1$$

and apply Theorem 8.23. □

Lemma 10.12. *Assume the conditions of Lemma 10.8, and assume furthermore that* $\|\mathbf{F}_{t_i} - \mathbf{F}_{t_{i+1}}\| = \epsilon_1/\mu_F(\mathbf{f}_{t_i}, \mathbf{x}_i)$. *Then,*

$$\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; t_i, t_{i+1}) \geq \frac{1}{CD^{3/2}\sqrt{n}}$$

Proof.

$$\begin{aligned}\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; t_i, t_{i+1}) &= \int_{t_i}^{t_{i+1}} \mu(\mathbf{f}_s, \mathbf{z}_s)\|(\dot{\mathbf{f}}_s, \dot{\mathbf{z}_s})\|_{\mathbf{f}_s, \mathbf{z}_s} \; \mathrm{d}s \\ &\geq \int_{t_i}^{t_{i+1}} \mu(\mathbf{f}_s, \mathbf{z}_s)\|\dot{\mathbf{f}}_s\|_{\mathbf{f}_s} \; \mathrm{d}s \\ &\geq \frac{\mu}{1+\epsilon_1+\pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{D}} \int_{t_i}^{t_{i+1}} \|\dot{\mathbf{f}}_s\|_{\mathbf{f}_s} \; \mathrm{d}s\end{aligned}$$

The rightmost integral evaluates to $d_{\mathrm{Riem}}(f_{t_i}, f_{t_{i+1}})$. Assume that

$$\tan\theta_1 = \|\mathbf{F}_{t_i} - \mathbf{F}_0\| \qquad \text{and} \qquad \tan\theta_2 = \|\mathbf{F}_{t_{i+1}} - \mathbf{F}_0\|$$

We know from elementary calculus that

$$\frac{\tan\theta_2 - \tan\theta_1}{\theta_2 - \theta_1} \leq \frac{1}{\cos^2\theta_2} = 1 + \tan^2\theta_2$$

Therefore, using $\tan\theta_2 \leq \|\mathbf{F}_1 - \mathbf{F}_0\|$, we obtain that

$$\theta_2 - \theta_1 \geq \frac{1}{2}\|\mathbf{F}_{t_{i+1}} - \mathbf{F}_{t_i}\|$$

Using that bound,

$$\begin{aligned}\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; t_i, t_{i+1}) &\geq \frac{1}{2}\frac{\mu}{1+\epsilon_1+\pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{D}}\|\mathbf{F}_{t_i} - \mathbf{F}_{t_{i+1}}\| \\ &\geq \frac{\sqrt{2}}{D^{3/2}\sqrt{n}}\frac{\epsilon_1}{1+\epsilon_1+\pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{D}}\end{aligned}$$

Numerically, we obtain

$$\sqrt{2}\frac{\epsilon_1}{1+\epsilon_1+\pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{2}} \geq C^{-1}. \tag{10.17}$$

□

Lemma 10.13. *Assume the conditions of Lemma 10.8, and suppose furthermore that*

$$\min_{t_i \le \sigma \le t_{i+1}} \Phi_{t_i,\sigma}(\mathbf{X}_i) \le \frac{2\epsilon_2}{D^{3/2}\mu_F(\mathbf{F}_{t_i}, \mathbf{X}_i)}$$

with equality for $\sigma = t_{i+1}$. Then,

$$\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; t_i, t_{i+1}) \ge \frac{1}{CD^{3/2}\sqrt{n}}$$

Proof.

$$\begin{aligned}
\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; t_i, t_{i+1}) &= \int_{t_i}^{t_{i+1}} \mu(\mathbf{f}_s, \mathbf{z}_s)\|(\dot{\mathbf{f}_s}, \dot{\mathbf{z}_s})\|_{\mathbf{f}_s,\mathbf{z}_s}\, \mathrm{d}s \\
&\ge \int_{t_i}^{t_{i+1}} \mu(\mathbf{f}_s, \mathbf{z}_s)\|\dot{\mathbf{z}_s}\|_{\mathbf{z}_s}\, \mathrm{d}s \\
&\ge \frac{\mu}{1+\epsilon_1+\pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{D}} \int_{t_i}^{t_{i+1}} \|\dot{\mathbf{z}_s}\|_{\mathbf{z}_s}\, \mathrm{d}s \\
&\ge \frac{\mu}{1+\epsilon_1+\pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{D}} d_{\mathrm{proj}}(\mathbf{z}_{t_{i+1}}, \mathbf{z}_{t_i}).
\end{aligned}$$

At this point we use triangular inequality:

$$\begin{aligned}
d_{\mathrm{proj}}(\mathbf{z}_{t_{i+1}}, \mathbf{z}_{t_i}) \ge & d_{\mathrm{proj}}(N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i), \mathbf{X}_i) - d_{\mathrm{proj}}(\mathbf{X}_i, \mathbf{z}_{t_i}) \\
& - d_{\mathrm{proj}}(N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i), \mathbf{z}_{t_{i+1}})
\end{aligned}$$

The first norm is precisely $\beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)$. From (10.10),

$$d_{\mathrm{proj}}(N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i), \mathbf{X}_i) \ge \frac{2}{D^{3/2}} \frac{(\epsilon_2 - a_0)\sqrt{1-\epsilon_1^2}}{(1+\epsilon_1)\mu}.$$

The second and third norms are distances to a zero. From Theorem 9.9 applied to $\mathbf{F}_{t_i}, \mathbf{X}_i$,

$$d_{\mathrm{proj}}(\mathbf{X}_i, \mathbf{z}_{t_i}) \le r_0(a_0)\beta \le \frac{2}{D^{3/2}\mu} a_0 r_0(a_0).$$

Applying the same theorem to $\mathbf{F}_{t_{i+1}}, \mathbf{X}_i$ with $\alpha(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i) < \alpha$ by (10.11), and estimating $\|N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)\| \ge 1 - \beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)$,

$$d_{\mathrm{proj}}(N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i), \mathbf{z}_{t_{i+1}}) \le r_1(\alpha) \frac{\beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)}{1 - \beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)}$$

By (10.9) and taking $\mu \geq \sqrt{2}, D \geq 2$, $\beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i) \leq (1-\epsilon_1)\alpha/2$. Therefore,

$$d_{\text{proj}}(N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i), \mathbf{z}_{t_{i+1}}) \leq \frac{2}{D^{3/2}} \frac{1-\epsilon_1}{\mu} \frac{1-\alpha}{\psi(\alpha)} \alpha^2 r_1(\alpha) \frac{1}{1-(1-\epsilon_1)\alpha/2}$$

using (10.13).

Putting all together,

$$\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; t_i, t_{i+1}) \geq \frac{2}{D^{3/2}\sqrt{n}} \times$$
$$\times \frac{\frac{(\epsilon_2 - a_0)\sqrt{1-\epsilon_1^2}}{(1+\epsilon_1)} - a_0 r_0(a_0) - (1-\epsilon_1)\frac{1-\alpha}{\psi(\alpha)(1-(1-\epsilon_1)\alpha/2)} \alpha^2 r_1(\alpha)}{1 + \epsilon_1 + \pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{D}}$$

The final bound was obtained numerically, assuming $D \geq 2$. We check computationally that

$$2\frac{\frac{(\epsilon_2 - a_0)\sqrt{1-\epsilon_1^2}}{(1+\epsilon_1)} - a_0 r_0(a_0) - (1-\epsilon_1)\frac{1-\alpha}{\psi(\alpha)(1-(1-\epsilon_1)\alpha/2)} \alpha^2 r_1(\alpha)}{1 + \epsilon_1 + \pi(1-\epsilon_1)\alpha r_0(\alpha)/\sqrt{2}} \geq C^{-1} \tag{10.18}$$

□

Proof of Theorem 10.5. Suppose the algorithm terminates. We claim that for each t_i, X_i is a (β, μ, a_0)-certified approximate zero of F_{t_i}, and that its associated zero is z_{t_i}. This is true by hypothesis when $i = 0$. Therefore, assume this is true up to a certain i.

Recall that $\beta(\mathbf{F}, \mathbf{X})$ scales as $\|\mathbf{X}\|$. In particular,

$$\beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_{i+1}) = \frac{\beta(\mathbf{F}_{t_{i+1}}, N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i))}{\|N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)\|} \leq \frac{\beta(\mathbf{F}_{t_{i+1}}, N(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i))}{1 - \beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i)}.$$

By (10.9) again, $\beta(\mathbf{F}_{t_{i+1}}, \mathbf{X}_i) \leq (1-\epsilon_1)\alpha/2$.

We apply (10.14) to obtain that

$$\frac{D^{3/2}}{2} \beta(\mathbf{F}_s, \mathbf{X}_{i+1}) \mu(\mathbf{f}_s, [N(\mathbf{F}_s, \mathbf{X}_i)]) \leq a_0.$$

From (10.11), X_i is an approximate zero of the second kind for F_s, $s \in [t_i, t_{i+1}]$. Since both $\alpha(\mathbf{F}_s, \mathbf{X}_i)$ and $\beta(\mathbf{F}_s, \mathbf{X}_i)$ are bounded above, the sequence of continuous functions $h_k(s) = N^k(\mathbf{F}_s, \mathbf{X}_i)$ is uniformly convergent to $\mathbf{Z}_s = \lim_{k\to\infty} N^k(\mathbf{F}_s, \mathbf{X}_i)$. Hence, $\mathbf{Z}_s$ is continuous and is a representative of $\mathbf{z}_s$. Since $[\lim N^k(\mathbf{F}_s, \mathbf{X}_i)] = [\lim N^k(\mathbf{F}_s, \mathbf{X}_{i+1})]$, item 2 of the Theorem follows.

Now to item 3: except for the final step, every step in the algorithm falls within two possibilities: either $s = S_1$, or $s = S_2$. Then Lemma 10.12 and 10.13 say that

$$\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; t_i, t_{i+1}) \geq \frac{1}{CD^{3/2}\sqrt{n}}$$

$\square$

Remark 10.14. The constants were computed using the free computer algebra package *Maxima* [60] with 40 digits of precision, and checked with 100 digits. The first thing to do is to guess a viable point $(\alpha, \epsilon_1, \epsilon_2)$ satisfying (10.3), (10.15), (10.17) and (10.18), for instance $(0.05, 0.02, 0.04)$.

Then, those values are optimized for $\min(\epsilon_1, \epsilon_2)$ by adding a small Gaussian perturbation, and discarding moves that do not improve the objective function or leave the viable set. Slowly, the variance of the Gaussian is reduced and the point converges to a local optimum. This optimization method is called *simulated annealing.*

10.3 Average complexity of randomized algorithms

In the sections above, we constructed and analyzed a linear homotopy algorithm. Now it is time to explain how to obtain a proper starting pair $(\mathbf{F}_0, \mathbf{x}_0)$.

Here is a simplified version of the Beltrán-Pardo construction of a randomized starting system. It is assumed that our randomized computer can sample points of $N(0, 1)$. The procedure is as follows.

Let M be a random (Gaussian) complex matrix of size $n \times n+1$. Then find a nonzero $Z_0 \in \ker M$. Next, draw $\mathbf{F}_0$ at random in the subspace R_M of $\mathcal{H}_\mathbf{d}$ defined by $\mathbf{L}_{\mathbf{Z}_0}(\mathbf{F}_0) = M$, $\mathbf{F}_0(\mathbf{Z}_0) = 0$. This can be done by picking F_0 at random, and then projecting.

Thus we obtain a pair (f_0, z_0) in the solution variety $\mathcal{V} \subset \mathbb{P}(\mathcal{H}_{\mathbf{d}}) \times \mathbb{P}^n$. This pair is a random variable, and hence has a certain probability distribution.

Proposition 10.15 (Beltrán-Pardo)**.** *The procedure described above provides a random pair* $(\mathbf{f}_0, \mathbf{z}_0)$ *in* $\mathcal{V}$*, with probability distribution*

$$\frac{1}{B}\ \pi_1^*\ \mathrm{d}\mathcal{H}_{\mathbf{d}},$$

where $B = \prod d_i$ *is the Bézout bound and* $\mathrm{d}\mathcal{H}_{\mathbf{d}}$ *is the Gaussian probability volume in* $\mathcal{H}_{\mathbf{d}}$*. Thus* $\pi_1^*\ \mathrm{d}\mathcal{H}_{\mathbf{d}}$ *denotes its pull-back through the canonical projection* π_1 *onto the first coordinate.*

Proof. For any integrable function $h : \mathcal{V} \to \mathbb{R}$,

$$\begin{aligned}
\frac{1}{B}\int_{\mathcal{V}} h(v)\pi_1^*\mathrm{d}\mathcal{H}_{\mathbf{d}}(v) = & \\
&= \frac{1}{B}\int_{\mathbb{P}^n} \mathrm{d}V(z) \int_{(\mathcal{H}_{\mathbf{d}})_z} h(F,z)\frac{\det|Df(z)Df(z)^*|}{\prod K_i(z,z)} \mathrm{d}\mathcal{H}_{\mathbf{d}})_z \\
&= \int_{\mathbb{P}^n} \mathrm{d}V(z) \int_{(\mathcal{H}_{\mathbf{d}})_z} h(F,z)\frac{\det|L_z(f)L_z(f)^*|}{(1+\|z\|^2)^n} \mathrm{d}\mathcal{H}_{\mathbf{d}})_z \\
&= \int_{\mathcal{H}_{\mathbf{1}}} \int_{\mathcal{R}_M} h(M+F,z)\mathrm{d}\mathcal{H}_{\mathbf{1}}
\end{aligned}$$

□

We need to quote from their paper [13, Theorem 20] the following estimate:

Theorem 10.16. *Let* $\mathbf{M}$ *be a random complex matrix of dimension* $(n+1) \times n$ *picked with Gaussian probability distribution of mean 0 and variance 1. Then,*

$$\mathbb{E}\left(\|M^\dagger\|^2\right) \le \frac{n}{2}\left(1+\frac{1}{n}\right)^{n+1} - n - \frac{1}{2}$$

Assuming $n \ge 2$, the right-hand-side is immediately bounded above by $(\frac{e^{3/2}}{2} - 1)n < 1.241n$. In exercise 10.1, the reader will show that when the variance is σ^2, then

$$\mathbb{E}\left(\|M^\dagger\|^2\right) \le \left(\frac{e^{3/2}}{2} - 1\right) n\sigma^{-2}. \tag{10.19}$$

Corollary 10.17. *Let $(\mathbf{f}, \mathbf{z}) \in \mathcal{V}$ be random in the following sense: $\mathbf{f}$ is normal with mean zero and variance σ^2, and $\mathbf{z}$ is a random zero of $\mathbf{f}$ (each one has same probability). Then,*

$$\mathbb{E}\left(\frac{\mu(\mathbf{f}, \mathbf{z})^2}{\|\mathbf{f}\|^2}\right)\left(\frac{e^{3/2}}{2} - 1\right) n\sigma^{-2}.$$

Bürgisser and Cucker introduced the following invariant:

Definition 10.18.

$$\begin{array}{rccl} \mu_2^2 : & \mathbb{P}(\mathcal{H}_{\mathbf{d}}) & \rightarrow & \mathbb{R}, \\ & \mathbf{f} & \mapsto & \frac{1}{B} \sum_{\mathbf{z} \in Z(\mathbf{f})} \mu(\mathbf{f}, \mathbf{z})^2 \end{array}$$

where $B = \prod d_i$ is the Bézout number.

Define the line integral

$$\mathcal{M}(\mathbf{f}_t; 0, 1) = \int_0^1 \|\dot{f}_t\|_{f_t} \mu_2^2(f_t) \mathrm{d}t = \int_{(\mathbf{f}_t)_{t \in [0,1]}} \mu_2^2(f_t) \mathrm{d}t.$$

When $\mathbf{F}_1$ is Gaussian random and $\mathbf{F}_0, \mathbf{z}_0$ are random as above, each zero $\mathbf{z}_0$ of $\mathbf{F}_0$ is equiprobable and

$$\mathbb{E}\left(\int_0^1 \|\dot{f}_t\|_{f_t} \mu(f_t, z_t)^2 \mathrm{d}t\right) = \mathbb{E}\left(\mathcal{M}(\mathbf{f}_t; 0, 1)\right)$$

Also, $\mathcal{M}(\mathbf{f}_t; 0, 1)$ is a line integral in $\mathbb{P}(\mathcal{H}_{\mathbf{d}})$, and depends upon $\mathbf{F}_0$ and $\mathbf{F}_1$. The curve $(\mathbf{f}_t)_{t \in [0,1]}$ is invariant under **real** rescaling of $\mathbf{F}_0$ and $\mathbf{F}_1$.

Bürgisser and Cucker suggested to sample $\mathbf{F}_0$ and $\mathbf{F}_1$ in the probability space

$$\left(B(0, \sqrt{2N}), \kappa^{-1} \mathrm{d}\mathcal{H}_{\mathbf{d}}\right)$$

instead of $(\mathcal{H}_{\mathbf{d}}, \mathrm{d}\mathcal{H}_{\mathbf{d}})$. Here, N is the complex dimension of sampling space ($\mathcal{H}_{\mathbf{d}}$ and κ is the constant that makes the new sampling space into a probability space. It is known that $\kappa \geq 1/2$.

Therefore, when $\mathbf{F}_0, \mathbf{Z}_0$ and $\mathbf{F}_1$ are random in the sense of Proposition 10.15, the expected value of $\mathcal{M}$ will be computed as if $\mathcal{F}_0$, $\mathcal{F}_1$ were sampled in the new probability space. We will need a geometric lemma before proceeding.

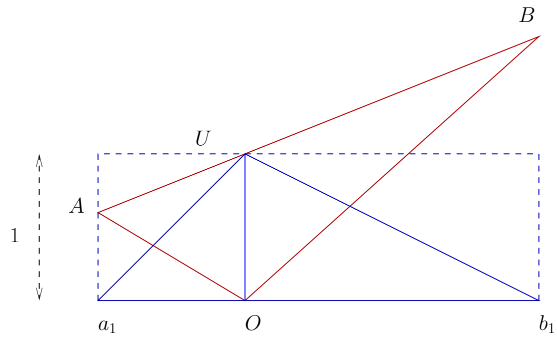

Figure 10.2: Geometric Lemma.

Lemma 10.19. *Let $A = (a_1, a_2), B = (b_1, b_2) \in \mathbb{R}^2$ be two points in the plane, such that $U = (0, 1) \in [A, B]$. Then,*

$$|b_1 - a_1| \le \|A\|\|B\|.$$

Proof. (See figure 10.2) We interpret $|b_1 - a_1|$ as the area of the rectangle of corners $(a_1, 0), (b_1, 0), (b_1, 1), (a_1, 1)$.

We claim that this is twice the area of the triangle (O, A, B). Indeed,

$$\begin{aligned} \text{Area}(O, A, B) &= \text{Area}(O, U, A) + \text{Area}(O, U, B) \\ &= \text{Area}(O, U, (a_1, 0)) + \text{Area}(O, U, (b_1, 0)) \\ &= \frac{1}{2}|b_1 - a_1| \end{aligned}$$

Therefore,

$$|b_1 - a_1| = 2\text{Area}(O, A, B) = \|A\|\|B\| \sin(\widehat{AOB}) \le \|A\|\|B\|$$

□

$$\begin{aligned} \mathcal{M}(\mathbf{f}_t; 0, 1) &\le \int_0^1 \| \left(I - \frac{1}{\|\mathbf{F}_t\|^2} \mathbf{F}_t \mathbf{F}_t^* \right) \dot{\mathbf{F}}_t \| \|\mathbf{F}_t\| \frac{\mu_2^2(\mathbf{F}_t)}{\|\mathbf{F}_t\|^2} \mathrm{d}t \\ &\le \int_0^1 \|\mathbf{F}_0\|\|\mathbf{F}_1\| \frac{\mu_2^2(\mathbf{F}_t)}{\|\mathbf{F}_t\|^2} \mathrm{d}t \end{aligned}$$

by the geometric Lemma, setting $U = \mathbf{F}_t$, $A = \mathbf{F}_0$, $B = \mathbf{F}_1$ and scaling. Replacing $\|\mathbf{F}_0\|$, $\mathbf{F}_1$ by $\sqrt{2N}$ and passing to expectations,

$$\begin{aligned} \mathbb{E}\left(\mathcal{M}(\mathbf{f}_t;0,1)\right) &\leq 2N\mathbb{E}\left(\int_0^1 \frac{\mu_2^2(\mathbf{F}_t)}{\|\mathbf{F}_t\|^2}\mathrm{d}t\right) \\ &\leq 2N\int_0^1 \mathbb{E}\left(\frac{\mu_2^2(\mathbf{F}_t)}{\|\mathbf{F}_t\|^2}\mathrm{d}t\right). \end{aligned}$$

Now, in the rightmost integral, $\mathbf{F}_0$ and $\mathbf{F}_1$ are sampled from the probability space

$$\left(B(0,\sqrt{2N}), \kappa^{-1}\mathrm{d}\mathcal{H}_{\mathbf{d}}\right).$$

The integrand is positive, so we can bound the integral by

$$\mathbb{E}\left(\mathcal{M}(\mathbf{f}_t;0,1)\right) \leq \kappa^{-2}\int_0^1 \mathbb{E}\left(\frac{\mu_2^2(F_t)}{\|F_t\|^2}\mathrm{d}t\right)$$

where now $\mathbf{F}_0$ and $\mathbf{F}_1$ are Gaussian random variables. Using that $\kappa \geq 1/2$,

$$\mathbb{E}\left(\mathcal{M}(\mathbf{f}_t;0,1)\right) \leq 8N\int_0^1 \mathbb{E}\left(\frac{\mu_2^2(\mathbf{F}_t)}{\|\mathbf{F}_t\|^2}\mathrm{d}t\right).$$

Let $N(\bar{\mathbf{F}}, \sigma^2 I)$ denote the Gaussian normal distribution with mean $\bar{\mathbf{F}}$ and covariance $\sigma^2 I$ (a rescaling of what we called $d\mathcal{H}_{\mathbf{d}}$).

From Corollary 10.17,

$$\mathbb{E}\left(\mathcal{M}(\mathbf{f}_t;0,1)\right) \leq 8\left(\frac{e^{3/2}}{2}-1\right)N\int_0^1 \frac{n}{t^2+(1-t)^2}\mathrm{d}t = 4(\frac{e^{3/2}}{2}-1)\pi Nn.$$

This establishes:

Proposition 10.20. *The expected number of homotopy steps of the algorithm of Theorem 10.5 with F_0, z_0 sampled by the Beltrán-Pardo method, is bounded above by*

$$1+4\left(\frac{e^{3/2}}{2}-1\right)\pi C N n^{3/2} D^{3/2}$$

The deterministic algorithm by Bürgisser and Cucker is similar, with starting system

$$\hat{F}_0(X) = \begin{bmatrix} X_1^{d_1} - X_0^{d_1} \\ \vdots \\ X_n^{d_n} - X_0^{d_1} \end{bmatrix}$$

Therefore it is possible to average over all paths, because the starting system is 'symmetric'. The condition integral was bounded in two parts. When t is small, the condition $\mu_2(\mathbf{f}_t)$ can be bounded in terms of the condition of $\mathbf{f}_0$, which unfortunately grows exponentially in n. The rest of the analysis relies on the following 'smoothed analysis' theorem:

Theorem 10.21. *Let $\mathbf{d} = (d_1, \dots, d_n)$, let $\bar{\mathbf{F}} \in \mathcal{H}_{\mathbf{d}}$ and let $\mathbf{F}$ be random with probability density $N(\bar{\mathbf{F}}, \sigma^2 I)$. Then,*

$$\mathbb{E}\left(\frac{\mu_2^2(\mathbf{F})}{\|\mathbf{F}\|^2}\right) \le \frac{n^{3/2}}{\sigma^2}$$

I refer to the paper, but the reader may look at exercises 10.2 and 10.3 before.

Exercise 10.1. In Theorem 10.16, replace the variance by σ^2. Show (10.19).

Exercise 10.2. Show that the average over the complex ball $B(0,\epsilon) \subset \mathbb{C}^2$ of the function $1/(|z_1|^2 + |z_2|^2)$ is finite.

Exercise 10.3. Let $n = 1$ and $d = 1$. Then $\mathcal{H}_d$ is the set of linear forms in variables x_0 and x_1. Compute the expected value of $\mu_2^2(f)/\|f\|$. Conclude that its expected value is finite, for $F \in N(\mathrm{e}_1, \sigma)$.

10.4 The geometric version of Smale's 17th problem

In view of Theorem 10.5, one would like to be able to produce given $\mathbf{F}_1 \in \mathcal{H}_{\mathbf{d}}$, a path $(\mathbf{f}_t, \mathbf{z}_t)$ in the solution variety such that

1. An approximate zero $\mathbf{X}_0$ is known for $\mathbf{f}_0$.

2. The condition length $\mathcal{L}(\mathbf{f}_t, \mathbf{z}_t; 0, 1)$ is bounded by a uniform polynomial in $n, D, \dim \mathcal{H}_{\mathbf{d}}$.

It is unknown how to do that in general. A deterministic algorithm producing such paths within expected polynomial time would provide an affirmative answer for Smale's 17^{th} problem. Here is a possibility: pick a fixed initial zero (say $\mathbf{X}_0 = \mathbf{Z}_0 = \mathrm{e}_0$), a fixed initial polynomial having $\mathbf{Z}_0$ as a root, and follow a linear path. For instance,

$$\mathbf{F}_0(\mathbf{X}) = \begin{bmatrix} \sqrt{d_1}\left(X_0^{d_1-1}X_1 - X_0^{d_1}\right) \\ \vdots \\ \sqrt{d_n}\left(X_0^{d_n-1}X_n - X_0^{d_n}\right) \end{bmatrix} \tag{10.20}$$

or

$$\tilde{\mathbf{F}}_0(\mathbf{X}) = \mathbf{F}_1(\mathbf{X}) - \mathbf{F}_1(\mathrm{e}_0) \begin{bmatrix} X_0^{d_1} \\ \vdots \\ X_0^{d_n} \end{bmatrix}.$$

Then, one has to integrate the expected length of the path. None of those linear paths is known to be polynomially bounded length in average.

Another possibility is to look for more insight. The *condition metric* on $\mathcal{V} \setminus \Sigma'$ is

$$\langle \cdot, \cdot \rangle'_{\mathbf{f},\mathbf{x}} = \mu^2(\mathbf{f}, \mathbf{x}) \langle \cdot, \cdot \rangle_{\mathbf{f},\mathbf{x}}$$

This reduces complexity to lengths. This new Riemannian metric is akin to the hyperbolic metric in Poincaré plane $y > 2$,

$$\langle \cdot, \cdot \rangle_{x,y}^{\text{Poincaré}} = y^{-2} \langle \cdot, \cdot \rangle_2.$$

A new difficulty arises. All geometry books seem to be written under differentiability assumptions for the metric. Here, μ is not differentiable at all points. (See fig. 10.3) The differential equation defining geodesics has to be replaced by a differential inequality [21].

In [8, 9] it was proved in the **linear** case that the condition number is *self-convex*. This means that $\log \mu$ is a convex function along geodesics in the condition metric.

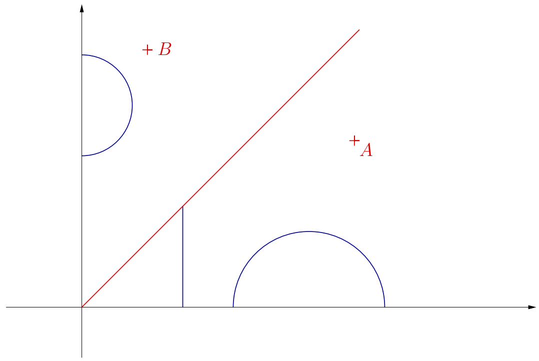

Figure 10.3: The condition metric for diagonal, real matrices is $\min(|x|,|y|)^{-2}\langle\cdot,\cdot\rangle$. Geodesics in the smooth part are easy to construct. But what is the shortest path from A to B?

In particular, the maximum of μ along a geodesic arc is attained at the extremities. The non-linear case is still open.

Starting the homotopy at a global minimum of μ (such as (10.20)), one would have a guarantee that the condition number along the path is bounded above by the condition number of the target $\mathbf{F}_1$. Moreover, a 'short' geodesic between $\mathbf{F}_1$ and a global minimum is known to exist [14].

There is nothing very particular about geodesics, except that they minimize distance. One can settle for a *short path*, that is a piecewise linear path with condition length bounded by a uniform polynomial in the input size.

This book finishes with a question.

Question 10.22. Given a random $\mathbf{f}_1$, is it possible to deterministically find a starting pair $(\mathbf{f}_0, \mathbf{z}_0)$ and a short path to $(\mathbf{f}_1, \mathbf{z}_1)$ in polynomial time?

Appendix A

Open Problems, by Carlos Beltrán, Jean-Pierre Dedieu, Luis Miguel Pardo and Mike Shub

A.1 Stability and complexity of numerical computations

Let us cite the first lines of the book [20]:

"The classical theory of computation had its origin in work of logicians (...) in the 1930's. The model of computation that developed in the following decades, the Turing machine has been extraordinarily successful in giving the foundations and framework for theoretical computer science. The point of view of this book is that the Turing model (we call it "classical") with its dependence on 0's and 1's is fundamentally inadequate for giving such a foundation to the theory

of modern scientific computation, where most of the algorithms ... are real number algorithms."

Then the authors develop a model of computation on the real numbers known today as the BSS model following the lines of a seminal paper [19]. This model is well adapted to study the complexity of numerical algorithms.

However this ideal picture suffers from an important defect. Numerical analysts do not use the exact arithmetic of real numbers but floating-point numbers and a finite precision arithmetic. The cited authors remark on the ultimate need to take input and round-off error into account in their theory. But now about twenty years later there is scant progress in this direction. For this reason we feel important to develop a model of computation based on floating-point arithmetic and to study, in this model, the concepts of stability and complexity of numerical computations.

A.2 A deterministic solution to Smale's 17th problem

Smale's 17th problem asks

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

The foundations to the study of this problem where set in the so–called "Bezout Series", that is [70–74]. The reader may see [79] for a description of this problem.

After the publication of [79] there has been much progress in the understanding of systems of polynomial equations. An Average Las Vegas algorithm (i.e. an algorithm which starts by choosing some points at random, with average polynomial running time) to solve this problem was described in [11, 12]. This algorithm is based on the idea of homotopy methods, as in the Bezout Series. Next, [69] showed that the complexity of following a homotopy path could actually be done in a much faster way than this proved in the Bezout Series (see (A.1) below). With this new method, the Average Las Vegas algorithm was improved to have running time which is almost quadratic in the input size, see [13]. Not only the expected value of the running time is known to be polynomial in the size of the input, also the variance and other higher moments, see [16].

The existence of a deterministic polynomial time algorithm for Smale's 17th problem is still an open problem. In [25] a deterministic algorithm is shown that has running time $O(N^{\log\log N})$, and indeed polynomial time for certain choices of the number of variables and degree of the polynomials. There exists a conjecture open since the nineties [74]: the number of steps will be polynomial time on the average if the starting point is the homogeneization of the identity map, that is

$$f_0(z) = \begin{cases} z_0^{d_1-1} z_1 & = 0 \\ \vdots & \\ z_0^{d_n-1} z_n & = 0 \end{cases}, \qquad \zeta_0 = (1, 0, \ldots, 0).$$

Another approach to the question is the one suggested by a conjecture in [15] on the averaging function for polynomial system solving.

A.3 Equidistribution of roots under unitary transformations

In the series of articles mentioned in the Smale's 17th problem section, all the algorithms cited use linear homotopy methods for solving polynomial equations. That is, let f_1 be a (homogeneous) system to be solved and let f_0 be another (homogeneous) system which has a known (projective) root ζ_0. Let f_t be the segment from f_0 to f_1 (sometimes we take the projection of the segment onto the set of systems of norm equal to 1). Then, try to (closely) follow the homotopy path, that is the path ζ_t such that ζ_t is a zero of f_t for $0 \leq t \leq 1$. If this path does not have a singular root, then it is well–defined. A natural question is the following: Fix f_1 and consider the orbit of f_0 under the action $f_0 \mapsto f_0 \circ U^*$ where U is a unitary matrix. The root

ζ_1 of f_1 which is reached by the homotopy starting at $f_0 \circ U^*$ will be different for different choices of U. The question is then, assuming that all the roots of f_1 are non–singular, what is the probability (of the set of unitary matrices with Haar measure) of finding each root? Some experiments [10] seem to show that all roots are equally probable, at least in the case of quadratic systems. But, there is no theoretical proof of this fact yet.

A.4 Log–Convexity

Let $\mathcal{H}_{\mathbf{d}}$ be the projective space of systems of n homogeneous polynomials of fixed degrees $(d) = (d_1, \ldots, d_n)$ and $n+1$ unknowns. In [69], it is proved that following a homotopy path (f_t, ζ_t) (where f_t is any C^1 curve in $\mathbb{P}(\mathcal{H}_{\mathbf{d}})$, and ζ_t is defined by continuation) requires at most

$$L_\kappa(f_t, \zeta_t) = CD^{3/2} \int_0^1 \mu(f_t, \zeta_t) \|(\dot{f}_t, \dot{\zeta}_t)\| \, dt \tag{A.1}$$

homotopy steps (see [7, 10, 25, 31] for practical algorithms and implementation, and see [55, 56] for different approaches to practical implementation of Newton's method). Here, C is a universal constant, D is the max of the d_i and μ is the normalized contition number, sometimes denoted μ_{norm}, and defined by

$$\mu(f, z) = \|f\| \left\| (Df(z) \mid_{z^\perp})^{-1} \operatorname{Diag}\left(\|z\|^{d_i - 1} d_i^{1/2} \right) \right\|,$$

$$\forall\, f \in \mathbb{P}(\mathcal{H}_{\mathbf{d}}),\ z \in \mathbb{P}(\mathbb{C}^{n+1}).$$

Note that $\mu(f, z)$ is essentially the operator norm of the inverse of the matrix $Df(z)$ restricted to the orthogonal complement of z. Then, (A.1) is the length of the curve (f_t, ζ_t) in the so–called *condition metric*, that is the metric in

$$W = \{(f, z) \in \mathbb{P}(\mathcal{H}_{\mathbf{d}}) \times \mathbb{P}^n : \mu(f, z) < +\infty\}$$

defined by pointwise multiplying the usual product structure by the condition number.

Thus, paths (f_t, ζ_t) which are, in some sense, *optimal* for the homotopy method, are those defined as *shortest geodesics in the condition metric.* They are known to exist and to have length which is logarithmic in the condition number of the extremes, see [14]. Their computation is however a difficult task. A simple question that one may ask is the following: let (f_t, ζ_t), $0 \le t \le 1$ be a geodesic for the condition metric. Is it true that $\max\{\mu(f_t, \zeta_t : 0 \le t \le 1\}$ is reached at the extremes $t = 0, 1$? More generally, one can ask for convexity of μ along these geodesics, or even convexity of $\log \mu$ (which implies convexity of μ).

Following [8,9,21], let us put the question in a general setting. Let $\mathcal{M}$ be a Riemannian manifold and let $\kappa : \mathcal{M} \to (0, \infty)$ be a Lipschitz function. We call that conformal metric in $\mathcal{M}$ obtained by pointwise multiplying the original one by κ the *condition metric*. We say that a curve $\gamma(t)$ in $\mathcal{M}$ is a minimizing geodesic (in the condition metric) if it has minimal (condition) length among all curves with the same extremes. A *geodesic* in the condition metric is then by definition any curve that is locally a minimizing geodesic. Then, we say that κ is self–convex if the function

$$t \to \log(\kappa(\gamma(t)))$$

is convex for any geodesic $\gamma(t)$ in $\mathcal{M}$. The question is then: Is μ self–convex in W?

It is interesting to point out that the usual unnormalized condition number of linear algebra (that is, $\kappa(A) = \|A^{-1}\|$) is a self–convex function in the set of maximal rank matrices, see [8,9] In [8] it is also proved that functions given by the inverse of the distance to a (sufficiently regular) submanifold of $\mathbb{R}^n$ is log–convex when restricted to an open set. Another interesting question is if that result can be extended to arbitrary submanifolds of arbitrary Riemannian manifolds.

A.5 Extension of the algorithms for Smale's 17th problem to other subspaces

The algorithms described above are all designed to solve polynomial systems which are assumed to be in dense representation. In particular, the "average" running time is for dense polynomial systems. As any affine subspace of $\mathcal{H}_{\mathbf{d}}$ has zero–measure in $\mathcal{H}_{\mathbf{d}}$, one cannot conclude that the average running time of any of these algorithms is polynomial for, say, sparse polynomial systems. Same question is open for real polynomial systems (i.e. for polynomial systems in $\mathcal{H}_{\mathbf{d}}$ with real coefficients). Some progress in this last problem has been done in [22]. Another interesting question is if some of these methods can be made to work for polynomial systems given by straight line programs.

A.6 Numerics for decision problems

Most of the algorithms nowadays used for polynomial system solving are based on numerics, for example all the homotopy methods discussed above. However, many problems in computation are decissional problems. The model problem is Hilbert's Nullstellensatz, that is given $f_1, \ldots, f_k$ polynomials with unknowns $z_1, \ldots, z_n$, does there exist a common zero $\zeta \in \mathbb{C}^n$? This problem asks if numeric algorithms can be designed to answer this kind of questions. Note that Hilbert's Nullstellensatz is a NP–hard problem, so one cannot expect worse case polynomial running time, but maybe average polynomial running time can be reached. Some progress in this direction may be available using the algorithms and theorems in [13, 25].

A.7 Integer zeros of a polynomial of one variable

A nice problem to include in this list is the so–called Tau Conjecture: is the number of integer zeros of a univariate polynomial, polynomially bounded by the length of the straight line program that generates it? This is Smale's 4th problem and we refer the reader to [79].

Another problem is the following: given $f_1, \ldots, f_k$ integer polynomials of one variable, find a bound for the maximum number of distinct integer roots of the composition $f_1 \circ \cdots \circ f_k$. In particular, can it happen that this number of zeros is equal to the product of the degrees?

This problem has been studied by Carlos Di Fiori, and he found an example of 4 polynomials of degree 2 such that their composition has 16 integer roots. An example of 5 degree 2 polynomials whose composition has 32 integer roots seems to be unknown to the date.

Bibliography

[1] P.A. Absil, J. Trumpf, R. Mahony, and B. Andrews, *All roads lead to Newton: Feasible second-order methods for equality-constrained optimization*. Tech Report UCL-INMA-2009.024.

[2] Eugene L. Allgower and Kurt Georg, *Continuation and path following*, Acta numerica, 1993, Acta Numer., Cambridge Univ. Press, Cambridge, 1993, pp. 1–64.

[3] Carlos d'Andrea, Teresa Krick, and Martín Sombra, *Heights of Varieties in Multiprojective spaces and arithmetic Nullstellensätze*, available at `http://front.math.ucdavis.edu/1103.4561`. Preprint, ArXiV, march 2011.

[4] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc. **68** (1950), 337–404.

[5] Jean-Marc Azaïs and Mario Wschebor, *Level sets and extrema of random processes and fields*, John Wiley & Sons Inc., Hoboken, NJ, 2009.

[6] Carlos Beltrán, *Sobre el problema 17 de Smale: Teoría de la Intersección y Geometría Integral*, PhD Thesis, Universidad de Cantábria, 2006, `http://sites.google.com/site/beltranc/publications`.

[7] ———, *A continuation method to solve polynomial systems and its complexity*, Numer. Math. **117** (2011), no. 1, 89–113, DOI 10.1007/s00211-010-0334-3.

[8] Carlos Beltrán, Jean-Pierre Dedieu, Gregorio Malajovich, and Mike Shub, *Convexity properties of the condition number*, SIAM Journal on Matrix Analysis and Applications **31** (2010), no. 3, 1491-1506, DOI 10.1137/080718681.

[9] ———, *Convexity properties of the condition number*. Preprint, ArXiV, 30 oct 2009, `http://arxiv.org/abs/0910.5936`.

[10] Carlos Beltrán and Anton Leykin, *Certified numerical homotopy tracking* (30 oct 2009). Preprint, ArXiV, `http://arxiv.org/abs/0912.0920`.

[11] Carlos Beltrán and Luis Miguel Pardo, *On Smale's 17th problem: a probabilistic positive solution*, Found. Comput. Math. **8** (2008), no. 1, 1–43, DOI 10.1007/s10208-005-0211-0.

[12] ______, *Smale's 17th problem: average polynomial time to compute affine and projective solutions*, J. Amer. Math. Soc. **22** (2009), no. 2, 363–385, DOI 10.1090/S0894-0347-08-00630-9.

[13] Carlos Beltrán and Luis Miguel Pardo, *Fast linear homotopy to find approximate zeros of polynomial systems*, Foundations of Computational Mathematics **11** (2011), 95–129.

[14] Carlos Beltrán and Michael Shub, *Complexity of Bezout's theorem. VII. Distance estimates in the condition metric*, Found. Comput. Math. **9** (2009), no. 2, 179–195, DOI 10.1007/s10208-007-9018-5.

[15] ______, *On the geometry and topology of the solution variety for polynomial system solving*. to appear.

[16] ______, *A note on the finite variance of the averaging function for polynomial system solving*, Found. Comput. Math. **10** (2010), no. 1, 115–125, DOI 10.1007/s10208-009-9054-4.

[17] D. N. Bernstein, *The number of roots of a system of equations*, Funkcional. Anal. i Priložen. **9** (1975), no. 3, 1–4 (Russian).

[18] D. N. Bernstein, A. G. Kušnirenko, and A. G. Hovanskiĭ, *Newton polyhedra*, Uspehi Mat. Nauk **31** (1976), no. 3(189), 201–202 (Russian).

[19] Lenore Blum, Mike Shub, and Steve Smale, *On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines*, Bull. Amer. Math. Soc. (N.S.) **21** (1989), no. 1, 1–46, DOI 10.1090/S0273-0979-1989-15750-9.

[20] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale, *Complexity and real computation*, Springer-Verlag, New York, 1998. With a foreword by Richard M. Karp.

[21] Paola Boito and Jean-Pierre Dedieu, *The condition metric in the space of rectangular full rank matrices*, SIAM J. Matrix Anal. Appl. **31** (2010), no. 5, 2580–2602, DOI 10.1137/08073874X.

[22] Cruz E. Borges and Luis M. Pardo, *On the probability distribution of data at points in real complete intersection varieties*, J. Complexity **24** (2008), no. 4, 492–523, DOI 10.1016/j.jco.2008.01.001.

[23] Haïm Brezis, *Analyse fonctionnelle*, Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree], Masson, Paris, 1983 (French). Théorie et applications. [Theory and applications].

[24] W. Dale Brownawell, *Bounds for the degrees in the Nullstellensatz*, Ann. of Math. (2) **126** (1987), no. 3, 577–591, DOI 10.2307/1971361.

[25] Peter Bürgisser and Felipe Cucker, *On a problem posed by Steve Smale*, Annals of Mathematics (to appear). Preprint, ArXiV, `arxiv.org/abs/0909.2114v1`.

[26] ______, *Conditionning*. In preparation.

[27] David Cox, John Little, and Donal O'Shea, *Ideals, varieties, and algorithms*, 3rd ed., Undergraduate Texts in Mathematics, Springer, New York, 2007. An introduction to computational algebraic geometry and commutative algebra.

[28] Jean-Pierre Dedieu, *Estimations for the separation number of a polynomial system*, J. Symbolic Comput. **24** (1997), no. 6, 683–693, DOI 10.1006/jsco.1997.0161.

[29] ______, *Estimations for the separation number of a polynomial system*, J. Symbolic Comput. **24** (1997), no. 6, 683–693, DOI 10.1006/jsco.1997.0161.

[30] ______, *Points fixes, zéros et la méthode de Newton*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 54, Springer, Berlin, 2006 (French). With a preface by Steve Smale.

[31] Jean-Pierre Dedieu, Gregorio Malajovich, and Michael Shub, *Adaptative Step Size Selection for Homotopy Methods to Solve Polynomial Equations.* Preprint, ArXiV, 11 apr 2011, `http://arxiv.org/abs/1104.2084`.

[32] Jean-Pierre Dedieu, Pierre Priouret, and Gregorio Malajovich, *Newton's method on Riemannian manifolds: convariant alpha theory*, IMA J. Numer. Anal. **23** (2003), no. 3, 395–419, DOI 10.1093/imanum/23.3.395.

[33] Jean-Pierre Dedieu and Mike Shub, *Multihomogeneous Newton methods*, Math. Comp. **69** (2000), no. 231, 1071–1098 (electronic), DOI 10.1090/S0025-5718-99-01114-X.

[34] Thomas Delzant, *Hamiltoniens périodiques et images convexes de l'application moment*, Bull. Soc. Math. France **116** (1988), no. 3, 315–339 (French, with English summary).

[35] James W. Demmel, *The probability that a numerical analysis problem is difficult*, Math. Comp. **50** (1988), no. 182, 449–480, DOI 10.2307/2008617.

[36] Carl Eckart and Gale Young, *The approximation of a matrix by another of lower rank*, Psychometrika **1** (1936), no. 3, 211–218, DOI 10.1007/BF02288367.

[37] ______, *A principal axis transformation for non-hermitian matrices*, Bull. Amer. Math. Soc. **45** (1939), no. 2, 118–121, DOI 10.1090/S0002-9904-1939-06910-3.

[38] Alan Edelman, *On the distribution of a scaled condition number*, Math. Comp. **58** (1992), no. 197, 185–190, DOI 10.2307/2153027.

[39] Ioannis Z. Emiris and Victor Y. Pan, *Improved algorithms for computing determinants and resultants*, J. Complexity **21** (2005), no. 1, 43–71, DOI 10.1016/j.jco.2004.03.003.

[40] O. P. Ferreira and B. F. Svaiter, *Kantorovich's theorem on Newton's method in Riemannian manifolds*, J. Complexity **18** (2002), no. 1, 304–329, DOI 10.1006/jcom.2001.0582.

[41] Noaï Fitchas, Marc Giusti, and Frédéric Smietanski, *Sur la complexité du théorème des zéros*, Approximation and optimization in the Caribbean, II (Havana, 1993), Approx. Optim., vol. 8, Lang, Frankfurt am Main, 1995, pp. 274–329 (French, with English and French summaries). With the collaboration of Joos Heintz, Luis Miguel Pardo, Juan Sabia and Pablo Solernó.

[42] Michael R. Garey and David S. Johnson, *Computers and intractability*, W. H. Freeman and Co., San Francisco, Calif., 1979. A guide to the theory of NP-completeness; A Series of Books in the Mathematical Sciences.

[43] Marc Giusti and Joos Heintz, *La détermination des points isolés et de la dimension d'une variété algébrique peut se faire en temps polynomial*, Computational algebraic geometry and commutative algebra (Cortona, 1991), Sympos. Math., XXXIV, Cambridge Univ. Press, Cambridge, 1993, pp. 216–256 (French, with English and French summaries).

[44] Phillip Griffiths and Joseph Harris, *Principles of algebraic geometry*, Wiley Classics Library, John Wiley & Sons Inc., New York, 1994. Reprint of the 1978 original.

[45] M. Gromov, *Convex sets and Kähler manifolds*, Advances in differential geometry and topology, World Sci. Publ., Teaneck, NJ, 1990, pp. 1–38.

[46] Nicholas J. Higham, *Accuracy and stability of numerical algorithms*, 2nd ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[47] The Institute of Electrical and Electronics Engineers Inc, *IEEE Standard for Floating Point Arithmetic IEEE Std 754-2008*, 3 Park Avenue, New York, NY 10016-5997, USA, 2008, `http://ieeexplore.ieee.org/xpl/standards.jsp`.

[48] L. V. Kantorovich, *On the Newton method*, in: L.V. Kantorovich, *Selected works. Part II, Applied functional analysis. Approximation methods and computers;*, Classics of Soviet Mathematics, vol. 3, Gordon and Breach Publishers, Amsterdam, 1996. Translated from the Russian by A. B. Sossinskii; Edited by S. S. Kutateladze and J. V. Romanovsky. Article originally published in *Trudy MIAN SSSR* **28** 104-144(1949).

[49] A. G. Khovanskiĭ, *Fewnomials*, Translations of Mathematical Monographs, vol. 88, American Mathematical Society, Providence, RI, 1991. Translated from the Russian by Smilka Zdravkovska.

[50] Steven G. Krantz, *Function theory of several complex variables*, 2nd ed., The Wadsworth & Brooks/Cole Mathematics Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1992.

[51] Teresa Krick, Luis Miguel Pardo, and Martín Sombra, *Sharp estimates for the arithmetic Nullstellensatz*, Duke Math. J. **109** (2001), no. 3, 521–598, DOI 10.1215/S0012-7094-01-10934-4.

[52] A. G. Kušnirenko, *Newton polyhedra and Bezout's theorem*, Funkcional. Anal. i Priložen. **10** (1976), no. 3, 82–83. (Russian).

[53] T. L. Lee, T. Y. Li, and C. H. Tsai, *HOM4PS-2.0: a software package for solving polynomial systems by the polyhedral homotopy continuation method*, Computing **83** (2008), no. 2-3, 109–133, DOI 10.1007/s00607-008-0015-6.

[54] Tien-Yien Li and Chih-Hsiung Tsai, *HOM4PS-2.Opara: parallelization of HOM4PS-2.O for solving polynomial systems*, Parallel Comput. **35** (2009), no. 4, 226–238, DOI 10.1016/j.parco.2008.12.003.

[55] Gregorio Malajovich, *On the complexity of path-following Newton algorithms for solving systems of polynomial equations with integer coefficients*, PhD Thesis, Department of Mathematics, University of California at Berkeley, 1993, `http://www.labma.ufrj.br/~gregorio/papers/thesis.pdf`.

[56] ______, *On generalized Newton algorithms: quadratic convergence, path-following and error analysis*, Theoret. Comput. Sci. **133** (1994), no. 1, 65–84, DOI 10.1016/0304-3975(94)00065-4. Selected papers of the Workshop on Continuous Algorithms and Complexity (Barcelona, 1993).

[57] Gregorio Malajovich and Klaus Meer, *Computing minimal multi-homogeneous Bézout numbers is hard*, Theory Comput. Syst. **40** (2007), no. 4, 553–570, DOI 10.1007/s00224-006-1322-y.

[58] Gregorio Malajovich and J. Maurice Rojas, *High probability analysis of the condition number of sparse polynomial systems*, Theoret. Comput. Sci. **315** (2004), no. 2-3, 524–555, DOI 10.1016/j.tcs.2004.01.006.

[59] ______, *Polynomial systems and the momentum map*, Foundations of computational mathematics (Hong Kong, 2000), World Sci. Publ., River Edge, NJ, 2002, pp. 251–266.

[60] Maxima.sourceforge.net, *Maxima, a Computer Algebra System, Version 5.18.1*, 2009.

[61] John W. Milnor, *Topology from the differentiable viewpoint*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1997. Based on notes by David W. Weaver; Revised reprint of the 1965 original.

[62] Ferdinand Minding, *On the determination of the degree of an equation obtained by elimination*, Topics in algebraic geometry and geometric modeling, Contemp. Math., vol. 334, Amer. Math. Soc., Providence, RI, 2003, pp. 351–362. Translated from the German (Crelle, 1841)and with a commentary by D. Cox and J. M. Rojas.

[63] Ketan D. Mulmuley and Milind Sohoni, *Geometric complexity theory: introduction*, Technical Report TR-2007-16, Department of Computer Science, University of Chicago, September 4, 2007, `http://www.cs.uchicago.edu/research/publications/techreports/TR-2007-16`.

[64] Kazuo Muroi, *Reexamination of the Susa mathematical text no. 12: a system of quartic equations*, SCIAMVS **2** (2001), 3–8.

[65] Leopoldo Nachbin, *Lectures on the Theory of Distributions*, Textos de Matemática, Instituto de Física e Matemática, Universidade do Recife, 1964.

[66] ______, *Topology on spaces of holomorphic mappings*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 47, Springer-Verlag New York Inc., New York, 1969.

[67] James Renegar, *On the worst-case arithmetic complexity of approximating zeros of systems of polynomials*, SIAM J. Comput. **18** (1989), no. 2, 350–370, DOI 10.1137/0218024.

[68] Michael Shub, *Some remarks on Bezout's theorem and complexity theory*, From Topology to Computation: Proceedings of the Smalefest (Berkeley, CA, 1990), Springer, New York, 1993, pp. 443–455.

[69] ———, *Complexity of Bezout's theorem. VI. Geodesics in the condition (number) metric*, Found. Comput. Math. **9** (2009), no. 2, 171–178, DOI 10.1007/s10208-007-9017-6.

[70] Michael Shub and Steve Smale, *Complexity of Bézout's theorem. I. Geometric aspects*, J. Amer. Math. Soc. **6** (1993), no. 2, 459–501, DOI 10.2307/2152805.

[71] M. Shub and S. Smale, *Complexity of Bezout's theorem. II. Volumes and probabilities*, Computational algebraic geometry (Nice, 1992), Progr. Math., vol. 109, Birkhäuser Boston, Boston, MA, 1993, pp. 267–285.

[72] Michael Shub and Steve Smale, *Complexity of Bezout's theorem. III. Condition number and packing*, J. Complexity **9** (1993), no. 1, 4–14, DOI 10.1006/jcom.1993.1002. Festschrift for Joseph F. Traub, Part I.

[73] ———, *Complexity of Bezout's theorem. IV. Probability of success; extensions*, SIAM J. Numer. Anal. **33** (1996), no. 1, 128–148, DOI 10.1137/0733008.

[74] M. Shub and S. Smale, *Complexity of Bezout's theorem. V. Polynomial time*, Theoret. Comput. Sci. **133** (1994), no. 1, 141–164, DOI 10.1016/0304-3975(94)90122-8. Selected papers of the Workshop on Continuous Algorithms and Complexity (Barcelona, 1993).

[75] S. Smale, *Topology and mechanics. I*, Invent. Math. **10** (1970), 305–331.

[76] Steve Smale, *On the efficiency of algorithms of analysis*, Bull. Amer. Math. Soc. (N.S.) **13** (1985), no. 2, 87–121, DOI 10.1090/S0273-0979-1985-15391-1.

[77] ———, *Newton's method estimates from data at one point*, computational mathematics (Laramie, Wyo., 1985), Springer, New York, 1986, pp. 185–196.

[78] ———, *Mathematical problems for the next century*, Math. Intelligencer **20** (1998), no. 2, 7–15, DOI 10.1007/BF03025291.

[79] ———, *Mathematical problems for the next century*, Mathematics: frontiers and perspectives, Amer. Math. Soc., Providence, RI, 2000, pp. 271–294.

[80] Andrew J. Sommese and Charles W. Wampler II, *The numerical solution of systems of polynomials*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005. Arising in engineering and science.

[81] A. M. Turing, *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math. **1** (1948), 287–308.

[82] Constantin Udrişte, *Convex functions and optimization methods on Riemannian manifolds*, Mathematics and its Applications, vol. 297, Kluwer Academic Publishers Group, Dordrecht, 1994.

[83] Jan Verschelde, *Polyhedral methods in numerical algebraic geometry*, Interactions of classical and numerical algebraic geometry, Contemp. Math., vol. 496, Amer. Math. Soc., Providence, RI, 2009, pp. 243–263.

[84] Wang Xinghua, *Some result relevant to Smale's reports*, in: M.Hirsch, J. Marsden and S. Shub(eds): *From Topolgy to Computation: Proceedings of Smalefest*, Springer, new-york, 1993, pp. 456-465.

[85] Hermann Weyl, *The theory of groups and quantum mechanics*, Dover Publications, New York, 1949. XVII+422 pp.

[86] J. H. Wilkinson, *Rounding errors in algebraic processes*, Dover Publications Inc., New York, 1994. Reprint of the 1963 original [Prentice-Hall, Englewood Cliffs, NJ; MR0161456 (28 #4661)].

Glossary of notations

As a general typographical convention, a stands for a scalar quantity, $\mathbf{a}$ for a vectorial quantity, A for a matrix or operator or geometrical entity, $\mathcal{A}$ for a space, $\mathbf{A}$ for a ring or algebra, $\mathfrak{a}$ for an ideal.

$I(X)$	–	Ideal of polynomials vanishing at X	17
$x \overset{L}{\rightsquigarrow} y$	–	Group action: $y = a(L, x)$.	19
$Z(f)$	–	Zero set.	21
$\mathcal{F}$	–	Fewspace (Def. 5.2 or 5.15) or a product of.	56
V	–	Evaluation function associated to a fewspace.	56
$K(\mathbf{x}, \mathbf{y})$	–	Reproducing kernel associated to a fewspace.	57
ω	–	Kähler form associated to a fewspace.	57
$\mathcal{F}_{\mathbf{x}}x$	–	Fiber of $\mathbf{f} \in \mathcal{F}$ with $f(\mathbf{x}) = 0$.	58
$\mathrm{d}\mathcal{F}$	–	Zero average, unit variance normal probab. distrib.	62
$\mathcal{P}_d$	–	Space of polynomials of degree $\leq d$ in n variables.	63
$\mathcal{P}_{\mathbf{d}}$	–	$\mathcal{P}_{d_1} \times \cdots \times \mathcal{P}_{d_n}$.	63
$\mathcal{H}_d$	–	Space of homog. polynomials of deg. d in $n+1$ vars.	66
$N(\mathbf{f}, \mathbf{x})$	–	Newton operator.	82
$\gamma(\mathbf{f}, \mathbf{x})$	–	Invariant related to Newton iteration.	84
$\psi(u)$	–	The function $1 - 4u + 2u^2$.	88
$\beta(\mathbf{f}, \mathbf{x})$	–	Invariant related to Newton iteration.	97
$\alpha(\mathbf{f}, \mathbf{x})$	–	Invariant related to Newton iteration.	97

Index