On the Complexity of Path-Following Newton Algorithms for Solving Systems of Polynomial Equations with Integer Coefficients.

by

Gregorio Malajovich-Muñoz

B.S. (Universidade Federal do Rio de Janeiro) 1989
M.S. (Universidade Federal do Rio de Janeiro) 1990

A thesis submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Mathematics
in the
GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Steve Smale, Chair
Professor Hendrik W. Lenstra
Professor John Canny

1993

The thesis of Gregorio Malajovich Muñoz is approved :

Chair Date

Date

Date

University of California, Berkeley
1993

On the Complexity of Path-Following Newton Algorithms
for Solving Systems of Polynomial Equations
with Integer Coefficients.

by

Gregorio Malajovich-Muñoz.

Contents

CHAPTER 1

Introduction

Complexity of solving systems of polynomial equations with integer coefficients can be bounded for a generic class of systems.

1. Systems of polynomials with Integer coefficients

Let $\mathcal{H}_d$ be the space of all systems of n homogeneous polynomial equations with degree $d = (d_1, \ldots, d_n)$ in variables $(x_0, \ldots, x_n)$ with complex coefficients. The space $\mathcal{H}_d$ is a complex vector space, with coordinates given by the coefficients of each polynomial.

A *zero* of a system $f \in \mathcal{H}_d$ is a point $\zeta \in \mathbb{C}^{n+1}$, $\zeta \neq 0$, such that $f(\zeta) = 0$. Alternatively, a zero of f is a ray $\{\lambda\zeta : \lambda \in \mathbb{C}\}$, for $f(\zeta) = 0$. This ray can be interpreted as a point in projective space $\mathbb{CP}^n$.

We will consider several versions of *Newton iteration.* One of them will be given by the operator $N^{\text{pseu}} : x \mapsto x - Df(x)^\dagger f(x)$, where $Df(x)^\dagger$ denotes the Moore-Penrose pseudo-inverse of $Df(x)$ (See Chapter 2).

Since the concept of a zero is not practical computationally, we will consider the following concept of *approximate zero* :

Definition 1. Let $f \in \mathcal{H}_d$, and let $z = z_0 \in \mathbb{C}^{n+1}$, $z \neq 0$. Let $z_{i+1} = N^{\text{pseu}}(z_i)$. The point z is said to be an *approximate zero* of f if and only if there is a zero ζ of f *associated* to z, such that :

$$\min_\lambda \frac{\|z_i - \lambda\zeta\|_2}{\|z_i\|_2} < 2^{-2^i - 1}$$

Another important notion is a version of the concept of *height.* If $a, b \in \mathbb{Z}$, we define $H(a) = |a|$ and $H(a + bi) = |a| + |b|$. (This is not the standard definition). If v is a vector, we define $H(v) = \max H(v_j)$. If f is a polynomial (or a system of polynomials), $H(f)$ is the maximum of the height of the coefficients of f.

Let $S(f)$ denote the number of non-zero coefficients of f. A reasonable measure of the *size* of an input $f \in \mathcal{H}_d$ would be the quantity $S(f) \log H(f)$.

2. Global complexity of polynomial-solving

For the purpose of our complexity analysis, a *system f of polynomials with integer coefficients* (resp. *with Gaussian integer coefficients*) will be the list of integers $n, (d_1, \dots, d_n), (S(f_1), f_1), \dots, (S(f_n), f_n)$, where each f_i is a list of monomials $f_{iJ}\, x^J$. Those monomials will be represented by integers $f_{iJ}, J_0, \dots, J_n$ (resp. $\mathrm{Re}(f_{iJ}), \mathrm{Im}(f_{iJ}), J_0, \dots, J_n$).

We define an *approximate solution* of a system f of polynomials as a list X_1, $\dots$, $X_{\prod d_i}$ of points of $\mathbb{C}^{n+1}$ with Gaussian integer coefficients, such that each X_i is an approximate zero of f associated to a different zero ray of f, and there is a X_i associated to every zero ray of f. Each X_i will be represented by the list of its coordinates (real and imaginary parts).

We will give an algorithm to obtain an approximate solution of a *generic* system of polynomial equations with Gaussian integer coefficients. By generic, we mean that it should not belong to a certain *real* algebraic variety of the realization of $\mathcal{H}_d$. The algorithm will require :

$$O\left((\max d_i)^{3/2} \mu^2 \left((n+1) S(f) \max d_i + (n+1)^3\right)\right)$$

floating point operations, with relative precision

$$O\left(\frac{1}{\mu^2 (\max d_i)^2 (n+1)^3 + \max S(f_i)}\right)$$

Here, μ is a condition number, to be defined later.

This result can be stated more precisely in the language of the theory of computation. Our main model of computation will be the Blum - Shub - Smale machine over $\mathbb{Z}$ (See [**3**]). The complexity results we will obtain will be equivalent, up to polynomial time, to Turing complexity. In this setting, we will prove the

Main Theorem . *There is (we can construct) a Blum-Shub-Smale machine M over $\mathbb{Z}$, that to the input $f \in \mathcal{H}_d$ (a system of polynomials with Gaussian integer coefficients), associates the list of integers $M(f)$ representing $\prod d_i$ points of $\mathbb{C}^{n+1}$ with Gaussian integer coefficients ;*

For each n, for each $d = (d_1, \dots, d_n)$ there is a Zariski open set U_d in $\mathcal{H}_d$ (considered as a real projective space) ;

If $f \in U_d$, then :

1 – $M(f)$ is an approximate solution to f.

2 – The output $M(f)$ is computed within time :

$$\left((\prod d_i)(\max d_i)^{3/2} \mu^2 + n(\prod d_i)^2 \sum d_i\right) \left((n+1) S(f) \max d_i + (n+1)^3\right) \times$$

$$\times P(\log \mu, \max d_i, \log n, \log S(f), \log\log H(f))$$

where P is a polynomial, and μ is a condition number, to be defined.

3 – The condition number μ can be bounded by :

$$\mu < \mu_0 H(f)^{d_0}$$

where the numbers μ_0 and d_0 will be defined later, and depend solely on $d = (d_1, \ldots, d_n)$.

3. The geometry of polynomial-solving

There are mainly two known classes of algorithms for solving systems of polynomial equations. The oldest class consists of algorithms derived from elimination theory. Elimination theory reduces polynomial solving to a linear problem in a high-dimensional space (e.g. dimension $\left(\begin{array}{c} \sum d_i \\ n \end{array}\right)$), the space of monomials of degree $\sum d_i - n$.

Some of those algorithms were developed at the start of this century (See Macaulay [**7**] or Van der Waerden [**17**]) . At that time, those algorithms were deemed untractable for more than 2 or 3 equations.

Modern algorithms using algebraic ideas include Gröbner Bases factorizations or numeric algorithms like those of Canny [**4**] or Renegar [**10**]. Some of those algorithms are particularly suitable to some kinds of sparse systems. When the system is sparse, it is possible to reduce the linear problem to a more modest dimension.

The other class of algorithms are the homotopy (path-following) methods. Several path-following procedures have been suggested (see Morgan, [**8**]). Basically, the algorithm follows a path $[f_0, f_1]$ in the space of polynomial systems $\mathcal{H}_d$. Here, f_1 is the system we want to solve, and f_0 is some system that we know beforehand how to solve.

If we complexify and projectivize both the spaces of systems and of solutions, then the solution rays $\lambda x_t : f_t(x_t) = 0$ depend smoothly on f, except in some degenerate locus Σ. It turns out that the degenerate locus is an algebraic variety, called the *discriminant variety.*

Path-following algorithms produce a finite sequence f_{t_i} in $[f_0, f_1]$; at each step, an approximate solution x_{t_i} for f_i is computed, using the previous approximate solution $x_{t_{i-1}}$ of $f_{t_{i-1}}$. In this thesis, this is done by a generalization of Newton iteration.

The complexity of that kind of path-following methods was studied by Shub and Smale in [**12**], assuming a model of real number computations. Among other important results, they bounded the number of homotopy steps necessary to follow a path in terms of a condition number $\mu([f_0, f_1]) = \max_{f \in [f_0, f_1]} \mu(f)$.

The condition number $\mu(f)$ has a simple geometric interpretation : it is the inverse of the minimum of $\rho(f, \zeta)$, the *distance* between f and the discriminant variety along the fiber $\{f(\zeta) = 0\}$.

Therefore, the number of homotopy steps depends on how far is the path $[f_0, f_1]$ from the discriminant variety.

4. Outline of this Thesis.

In Chapter 2, the results of [**12**] on the the quadratic convergence of Newton iteration and on the count of homotopy steps will be extended to the case of approximate Newton iteration.

Indeed, the iterates of good enough approximations x_0 of a zero ζ will verify a condition of the form :

$$\min_{\lambda} \frac{\|x_i - \lambda\zeta\|_2}{\|x_i\|_2} < \max\left(2^{-2^i-1}, k\delta\right)$$

where k is a small integer, provided the *error* of each iteration is bounded by δ.

The robustness theorem of [**12**] will also be extended to the case of approximate Newton iteration.

The worst case complexity of approximate Newton iteration is the subject of Chapter 3. For instance, given f, x and δ, we want an algorithm to approximate $x - Df(x)^\dagger f(x)$, with *error* at most δ, i.e., with precision $\delta\|x\|_2$.

Methods like exact rational or integer calculation, or like interval arithmetics, are not very efficient for computational or even theoretical complexity purposes, because of coefficient growth. Therefore, the construction of approximate Newton operators will be carried on using a machine with a fixed finite precision. Rigorous error bounds will be obtained, using standard numerical analysis techniques. The fixed precision machine can easily be constructed from a machine over the integers, equivalent (up to polynomial time classes) to a Turing machine.

The machine precision necessary to obtain the $\delta\|x\|_2$ approximation of $x - Df(x)^\dagger f(x)$ can be bounded in terms of n, d, the height $H(f)$ of f and the condition number $\mu(f, x)$.

The global complexity of solving systems of polynomial equations will be , roughly speaking , about $\mu^2 \log \mu$ times a polynomial in n and $\max d_i$, times the number of solutions $\prod d_i$, times the number of non-zero coefficients.

In order to give complexity bounds in terms of classical complexity theory, it makes sense to require f to have only integer (or Gaussian integer) coefficients. Although this precludes average-case complexity analysis using measure theory, this requirement will allow us to obtain some worst case bounds that would be impossible to obtain using computability over the reals. In Chapter 4, it will be proved that there are numbers $\mu(\Sigma)$ and $d(\Sigma)$ depending solely on n and d, such that whenever $\mu(f)$ is finite, $\mu(f) < \mu(\Sigma) H(f)^{d(\Sigma)}$. Here, H is the height of f.

A similar analysis carries on to the conditioning of a whole path in the space $\mathcal{H}_d$. This will provide us with a worst case bound for μ, hence for the complexity of solving a generic system f. By *generic*, we understand that f should not belong to a real algebraic variety in the realization of $\mathcal{H}_d$, that we will define.

Chapter 5 will provide the algorithms for solving f, and the algorithms for certifying a given approximate solution.

Although it is not practical to implement a computer program that will always succeed (provided enough memory is available, of course), it is cheap to implement a reasonably reliable program, with certification procedures that would guarantee the results obtained. Some implementation details will also be discussed.

5. Acknowledgements

I am thankful to Steve Smale, John Canny, Hendrik W. Lenstra for their suggestions and advice while writing this thesis.

Also, conversations with the following people were extremely useful : Lenore Blum, Felipe Cucker, Jim Demmel, Alan Edelman, Tom Emerson, Ioannis Emiris, Dima Grigoriev, Jan Hetting, Yili Huang, Myong-Hi Kim, Pascal Koiran T.Y. Li, Klaus Meer, Christian Michaux, Erich Novak, Victor Pan, Doug Priest, Maurice Rojas, Raymond Russell, Mike Shub, Rose St. John, Denise Wolf.

I was supported by a fellowship from the Conselho Nacional de Pesquisa (CNPq) of Brasil. Special thanks to Patricia Pinheiro, Marilene Farinasso, and all other CNPq staff that helped to mantain CNPq active while under very adverse circumstances.

Special thanks to all the people that helped to convince the government of Brasil to stop retaining the founds for tuition and fees of the fellowship recipients; This was at a time some of us were having our registration blocked.

During the years of my PhD, I was on a leave of absence from the Universidade Federal do Rio de Janeiro. I am also thankful to IBM-Brasil, IBM-Yorktown Heights, and to the Centre de Recerca Matemàtica of the Institut d'Estudis Catalans.

Special thanks to Luiz Carlos Guimaraes, without whom I probably would not be studying mathematics at this time.

Special thanks to my parents, for their encouragement and support.

I also would like to thank all the people that brought to existence all those electronic lists and newsgroups : Brasnet, Bras-Noticias, Sinopse, JCL noticias, CNPQ-l, pg-net, RUI.

This thesis was typeset in AMS-LaTeX, using a modification of ucthesis document style. Berkeley, July 28, 1993 and Bellaterra, November 1, 1993.

CHAPTER 2

On generalized Newton algorithms : Quadratic Convergence, Path-following and error analysis

Three versions of the Newton method for approximating zeros of a system of homogeneous polynomial equations are presented. Quadratic convergence theorems are obtained for exact and approximate iteration of the corresponding operators.

A bound on the number of approximate Newton steps necessary for path-following is estimated. This extends results in Shub - Smale [**12**]

1. Introduction

Let $\mathcal{H}_d$ be the space of all systems of n homogeneous polynomials in $n+1$ complex variables, of degree $d = (d_1, d_2, \ldots, d_n)$, with complex coefficients. Let $D = \max\ d_i$, and assume $D \geq 2$. The space $\mathcal{H}_d$ is endowed with the unitarily invariant *Kostlan norm* $\|.\|_{\mathrm{k}}$, defined by $\|f\|_{\mathrm{k}} = \sqrt{\sum \|f_i\|_{\mathrm{k}}{}^2}$, where :

$$\|f_i\|_{\mathrm{k}} = \sqrt{\sum_{|J|=d_i} \frac{|f_{iJ}|^2}{\begin{pmatrix} d_i \\ J \end{pmatrix}}} = \sqrt{\sum_{|J|=d_i} \frac{|f_{iJ}|^2}{\left(\frac{d_i!}{J_0!J_1!\ldots J_n!}\right)}}$$

The Kostlan norm $\|.\|_{\mathrm{k}}$ induces a metric $d_{\mathrm{proj}}(f,g) = \min_\lambda \frac{\|\lambda f - g\|_{\mathrm{k}}}{\|g\|_{\mathrm{k}}}$ in $\mathcal{H}_d$.

A zero of $f \in \mathcal{H}_d$ is a point $\zeta \in \mathbb{CP}^n$ such that $f(\zeta) = 0$. Alternatively, it is a line through the origin in $\mathbb{C}^{n+1}$ such that $f(\zeta) = 0$ for all ζ in that line.

Let $f \in \mathcal{H}_d$, and x range over $\mathbb{C}^{n+1}$. $Df(x)$ is a linear operator from $\mathbb{C}^{n+1}$ into $\mathbb{C}^n$. A generalized Newton operator is defined by the mapping :

$$N_V : x \mapsto x - Df(x)_{|V(x)}{}^{-1} f(x)$$

where V is a smooth family of hyperplanes in $\mathbb{C}^{n+1}$, $V(ax) = aV(x)$, and each $V(x)$ contains the point x. $V(x)$ will inherit the metric of $\mathbb{C}^{n+1}$. The notation $Df(x)_{|V(x)}{}^{-1} f(x)$ represents a point of $V(x)$, as contained in $\mathbb{C}^{n+1}$.

Different choices of V lead to different versions of Newton iteration, as we will see.

To each generalized Newton operator, we can associate a few *invariants*. Those are functions of $\mathcal{H}_d \times \mathbb{C}^{n+1}$, and are invariant under the group generated by the following transformations :

Unitary :

$$(f, x) \mapsto (f \circ U^{-1}, Ux) \ , \ U \in U(n+1)$$

Scaling :

$$(f_1, \dots, f_n, x) \mapsto (a_1 f_1, \dots, a_n f_n, bx) \ , \ a_i, b \in \mathbb{C}_\star$$

Scaling invariance implies that those invariants can be considered as functions of $\mathbb{P}(\mathcal{H}_d) \times \mathbb{CP}^n$. We define :

$$\beta(f, z) = \beta(z) = \frac{1}{\|z\|_2} . \left\| Df(z)_{|V(z)}{}^{-1} f(z) \right\|_2$$

$$\gamma(f, z) = \gamma(z) = \max \left\{ 1, \|z\|_2 \max_{k \geq 2} \left(\frac{1}{k!} \left\| Df(z)_{|V(z)}{}^{-1} D^k f(z) \right\|_2 \right)^{\frac{1}{k-1}} \right\}$$

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

Here, we have always $\gamma \geq 1$. Also, $D^k f(z)$ is a multilinear operator from $\left(\mathbb{C}^{n+1}\right)^k$ into $\mathbb{C}^n$. Therefore, those definitions are slightly different from the ones in [**12**], where $D^k f(z)$ is restricted to what we call $V(x)^k$.

Invariance of α, β and γ follows from the definitions.

The Newton operator in affine space :

If we set :

$$V(x) = x + \{0, y_1, \cdots, y_n\}$$

we obtain the (classical) Newton operator in affine space N^{aff}. If $\alpha(z_0)$ is small enough, successive iterates of z_0 will converge *quadratically* to a zero of f . The following theorem is essentially the Quadratic Convergence Theorem by Shub and Smale in [**12**] :

THEOREM 1. *Let $f \in \mathcal{H}_d$ Let $z_0 \in \mathbb{C}^{n+1}$ have its first coordinate non-zero. Let $\alpha^{\text{aff}}(z_0) < 1/8$. Let the sequence (z_i) be defined by $z_{i+1} = N^{\text{aff}}(z_i)$. Then there is a zero ζ of f such that $d_{\text{proj}}(z_i, \zeta) \leq 2^{-2^i - 1}$*

Above, distance in projective space is measured by :

$$d_{\text{proj}}(x,y) = \min_{\lambda \in \mathbb{C}} \left\{ \frac{\|x - \lambda y\|_2}{\|x\|_2} \right\}$$

There is a robust form of this theorem, that incorporates some error in each iteration. Since $N^{\text{aff}}(f, z_i)$ scales in $\|z_i\|_2$, it makes sense to measure the error at each iteration by :

$$\frac{\left\|z_{i+1} - N^{\text{aff}}(f, z_i)\right\|_2}{\|z_i\|_2}$$

Indeed, we will prove :

THEOREM 2. *Let $f \in \mathcal{H}_d$, let $z_0 \in \mathbb{C}^{n+1}$, let the first coordinate of z_0 be non-zero, and let $\delta \geq 0$ verify : $(\beta^{\text{aff}}(f, z_0) + \delta)\gamma^{\text{aff}}(f, z_0) < 1/16$, and $\gamma^{\text{aff}}(f, z_0)\delta < 1/384$. Let the sequence (z_i), where the first coordinates of z_i and z_0 are equal, verify :*

$$\frac{\left\|z_{i+1} - N^{\text{aff}}(f, z_i)\right\|_2}{\|z_i\|_2} \leq \delta$$

Then there is a zero ζ of f such that :

$$d_{\text{proj}}(z_i, \zeta) \leq \max \left(2^{-2^i - 1}, 6\delta \right)$$

Another version of the affine Newton operator was constructed by Morgan [**8**] by fixing a random vector y, and setting $V(x) = x + y^{\perp}$. This *random change of coordinates* allows him to use the classical Newton operator (in affine space) with systems that have zeros at infinity. There are more general Newton operators that allow to approximate zeros at infinity.

The Newton operator in projective space : If we define :

$$V(x) = x + x^{\perp}$$

we obtain the Newton operator in projective space, which may also be described by:

$$N^{\text{proj}}(x) = x - \begin{bmatrix} Df(x) \\ x^* \end{bmatrix}^{-1} \begin{bmatrix} f(x) \\ 0 \end{bmatrix}$$

where x^* means complex transpose of x. This operator was defined by Shub [**11**]. We will prove the Theorems :

THEOREM 3. *Let $f \in \mathcal{H}_d$ and let $z_0 \in \mathbb{C}^{n+1}$ be such that $\alpha^{\text{proj}}(z_0) < 1/32$. Let the sequence (z_i) be defined by $z_{i+1} = N^{\text{proj}}(z_i)$. Then there is a zero ζ of f such that $d_{\text{proj}}(z_i, \zeta) \leq 2^{-2^i-1}$*

THEOREM 4. *Let $f \in \mathcal{H}_d$, $z_0 \in \mathbb{C}^{n+1}$ and assume that $\delta \geq 0$ verifies :*

$$(\beta^{\text{proj}}(f, z_0) + \delta)\gamma^{\text{proj}}(f, z_0) < 1/32,$$

and $\gamma^{\text{proj}}(f, z_0)\delta < 1/640$. Let the sequence (z_i) verify :

$$\frac{\left\| z_{i+1} - N^{\text{proj}}(f, z_i) \right\|_2}{\left\| z_i \right\|_2} \leq \delta$$

Then there is a zero ζ of f such that :

$$d_{\text{proj}}(z_i, \zeta) \leq \max\left(2^{-2^i-1}, 10\delta\right)$$

Pseudo Newton operator If we make :

$$V(x) = x + \ker Df(x)^{\perp}$$

we obtain the pseudo-Newton operator :

$$N^{\text{pseu}}(x) = x - Df(x)^{\dagger} f(x)$$

where $A^{\dagger}$ is the Moore-Penrose pseudo-inverse of A, defined by :

$$A^{\dagger} = A_{|(kerA)^{\perp}}{}^{-1}$$

This notation refers to the case rank$Df(x) = n$. In the case rank$Df(x) < n$, the operator N^{pseu} is not defined.

An equivalent definition in our case is the following : In the particular case the matrix to invert is diagonal, we set :

$$\begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \cdots & & \\ & & & \lambda_n & 0 \end{bmatrix}^{\dagger} = \begin{bmatrix} \lambda_1{}^{-1} & & & & \\ & \lambda_2{}^{-1} & & & \\ & & \cdots & & \\ & & & \lambda_n{}^{-1} & 0 \end{bmatrix}$$

Then we extend this definition to all matrices of rank n by setting, for any U, V unitary:

$$(U\Lambda V)^{\dagger} = V^{*}\Lambda^{\dagger}U^{*}$$

A very important property of the pseudo-inverse of $A : \mathbb{C}^{n+1} \to \mathbb{C}^n$ is that $A^\dagger y$ is the vector of minimal norm in the linear space $A^{-1}y$. Hence,

$$\left\| A^\dagger \right\|_2 = \min \left\| A_{|V}{}^{-1} \right\|_2$$

when V ranges over all hyperplanes through the origin. This Newton operator was suggested by Allgower and Georg [**1**]. We will prove the Theorems :

THEOREM 5. *Let $f \in \mathcal{H}_d$ and let $z_0 \in \mathbb{C}^{n+1}$ be such that $\alpha^{\text{pseu}}(z_0) < 1/8$. Let the sequence (z_i) be defined by $z_{i+1} = N^{\text{pseu}}(z_i)$. Then there is a zero ζ of f such that $d_{\text{proj}}(z_i, \zeta) \leq 2^{-2^i-1}$*

THEOREM 6. *Let $f \in \mathcal{H}_d$, $z_0 \in \mathbb{C}^{n+1}$ and let $\delta \geq 0$ verify : $(\beta^{\text{pseu}}(f, z_0) + \delta)\gamma^{\text{pseu}}(f, z_0) < 1/16$, and $\gamma^{\text{pseu}}(f, z_0)\delta < 1/384$. Let the sequence (z_i) verify :*

$$\frac{\|z_{i+1} - N^{\text{pseu}}(f, z_i)\|_2}{\|z_i\|_2} \leq \delta$$

Then there is a zero ζ of f such that :

$$d_{\text{proj}}(z_i, \zeta) \leq \max\left(2^{-2^i-1}, 6\delta\right)$$

Path-following and conditioning :

The robustness results in [**12**] come out naturally in the generalized case. We can define some more invariants associated to a generalized Newton operator :

$$\mu(f, x) = \max\left\{1, \|f\|_{\text{k}} \left\| Df(x)_{|V(x)}{}^{-1} \text{diag}(\sqrt{d_i} \|x\|^{d_i - 1}) \right\|_2 \right\}$$

$$\eta(f, x) = \frac{\left\| \text{diag}(d_i{}^{-1} \|x\|_2{}^{-d_i}) f(x) \right\|_2}{\|f\|_{\text{k}}}$$

Invariants μ and η are invariants under unitary transformations and under scalings of the form $(f, x) \mapsto (af, bx)$, $a, b \in \mathbb{C}_\star$. The following estimates relate μ and η to β and γ :

$$\beta(f, x) \leq \mu(f, x)\eta(f, x)$$

$$\gamma(f, x) \leq \frac{\mu(f, x) D^{3/2}}{2}$$

The first estimate is obvious. The second one follows from the same proof as in Shub and Smale [**12**], III-1 and (in the case $\gamma = 1$) from the fact $1 \leq D^{3/2}/2$ when $D \geq 2$.

Also, as in [**12**], the following estimates are true :

LEMMA 1.

$$\eta(g,\zeta) \le d_{\text{proj}}(f,g) + \eta(f,\zeta)$$
$$\mu(g,\zeta) \le \frac{\mu(f,\zeta)(1+d_{\text{proj}}(f,g))}{1-\sqrt{D}d_{\text{proj}}(f,g)\mu(f,\zeta)}$$

The number of steps and precision necessary for following a path (f_t,ζ_t) will depend on the following Theorems, that are modified versions of Theorem 3 in [**12**], I-3 :

THEOREM 7. *There are* $\bar{\alpha}=0.02$, $\bar{u}=0.05$ *such that, if* $\bar{\gamma}\ge 1$ *and :*

$$\eta^{\text{aff}}(f,\zeta)\mu^{\text{aff}}(f,\zeta) \le \frac{\bar{\alpha}}{\bar{\gamma}}$$
$$d_{\text{proj}}(x,\zeta) \le \frac{\bar{u}}{\bar{\gamma}}$$
$$\gamma^{\text{aff}}(f,\zeta) \le \bar{\gamma}$$

Then setting $x' = N^{\text{aff}}(f,x)$, *and* ζ' *the zero associated to* x', *we get :*

$$d_{\text{proj}}(x',\zeta') \le \frac{\bar{u}}{2\bar{\gamma}}$$

THEOREM 8. *There are* $\bar{\alpha}=0.01$, $\bar{u}=0.005$ *such that, if* $\bar{\gamma}\ge 1$ *and :*

$$\eta^{\text{proj}}(f,\zeta)\mu^{\text{proj}}(f,\zeta) \le \frac{\bar{\alpha}}{\bar{\gamma}}$$
$$d_{\text{proj}}(x,\zeta) \le \frac{\bar{u}}{\bar{\gamma}}$$
$$\gamma^{\text{proj}}(f,\zeta) \le \bar{\gamma}$$

Then setting $x' = N^{\text{proj}}(f,x)$, *and* ζ' *the zero associated to* x', *we get :*

$$d_{\text{proj}}(x',\zeta') \le \frac{\bar{u}}{2\bar{\gamma}}$$

THEOREM 9. *There are* $\bar{\alpha}=0.02$, $\bar{u}=0.05$ *such that, if* $\bar{\gamma}\ge 1$ *and :*

$$\eta^{\text{pseu}}(f,\zeta)\mu^{\text{pseu}}(f,\zeta) \le \frac{\bar{\alpha}}{\bar{\gamma}}$$
$$d_{\text{proj}}(x,\zeta) \le \frac{\bar{u}}{\bar{\gamma}}$$
$$\gamma^{\text{pseu}}(f,\zeta) \le \bar{\gamma}$$

Then setting $x' = N^{\text{pseu}}(f, x)$, and ζ' the zero associated to x', we get :

$$d_{\text{proj}}(x', \zeta') \leq \frac{\bar{u}}{2\bar{\gamma}}$$

It is immediate that :

COROLLARY 1. *In each of the three cases $N = N^{\text{aff}}, N^{\text{proj}}, N^{\text{pseu}}$, there are $\bar{\alpha}$, $\bar{u}$ such that, if $\bar{\gamma} \geq 1$ and :*

$$\eta(f,\zeta)\mu(f,\zeta) \leq \frac{\bar{\alpha}}{\bar{\gamma}}$$
$$d_{\text{proj}}(x,\zeta) \leq \frac{\bar{u}}{\bar{\gamma}}$$
$$\gamma(f,\zeta) \leq \bar{\gamma}$$
$$\delta \leq \frac{\bar{u}}{2\bar{\gamma}}$$

Then setting x' such that $\frac{\left\|x' - N(f,x)\right\|_2}{\|x\|_2} \leq \delta$, and if ζ' is the zero associated to x', we get :

$$d_{\text{proj}}(x', \zeta') \leq \frac{\bar{u}}{\bar{\gamma}}$$

A generalization of the Main Theorem of Shub and Smale in [**12**] for approximate Newton iteration follows :

THEOREM 10. *Assume that $N = N^{\text{aff}}$, N^{proj} or N^{pseu}. Let $\bar{\alpha}$ and $\bar{u}$ be given by Theorems 7, 8 or 9, respectively.*

Let (f_t, ζ_t) be a path in $\mathcal{H}_d \times \mathbb{C}^{n+1}$, so that $f_t(\zeta_t) = 0$. Let $\mu \geq \max(\mu(f_t, \zeta_t))$ be finite. Let $\bar{\gamma} \geq \frac{2}{3} D^{3/2} \mu$. Let z_0 verify $d_{\text{proj}}(z_0, \zeta_0) \leq \frac{\bar{u}}{\bar{\gamma}}$. Let (t_i) be a sequence such that $d_{\text{proj}}(f_{t_i}, f_{t_{i+1}}) \leq \Delta \leq \frac{3}{8}\frac{\bar{\alpha}}{\mu\bar{\gamma}}$. Let (z_i) verify :

$$\frac{\left\|z_{i+1} - N(f_{t_{i+1}}, z_i)\right\|_2}{\|z_i\|_2} \leq \delta \leq \frac{\bar{u}}{2\bar{\gamma}}$$

Then $d_{\text{proj}}(z_i, \zeta_{t_i}) \leq \frac{\bar{u}}{\bar{\gamma}}$, and hence $\alpha(f_{t_i}, z_i) \leq \bar{\alpha}$.

In particular, if the length of the path f_t is bounded by L, then $\frac{16}{9}\frac{L\mu^2 D^{3/2}}{\bar{\alpha}}$ steps of approximate Newton iteration with error less than $\frac{9}{4}\frac{\bar{u}}{D^{3/2}\bar{\mu}}$ suffices to *follow* the path (f_t, ζ_t) and obtain a zero of f_1.

2. Estimates on β

Let the generalized Newton operator N be one of N^{aff}, N^{proj} or N^{pseu}. Let z_i be a sequence of points satisfying :

$$\frac{\|z_{i+1} - N(f, z_i)\|_2}{\|z_i\|_2} \leq \delta$$

for some $\delta \geq 0$. In the affine case $N = N^{\text{aff}}$, assume furthermore that $V(z_i) = V(z_{i+1})$. This follows from the hypothesis of Theorem 2 according to which the first coordinates of all the z_i are equal.

The case $\delta = 0$ represents the exact iteration $z_{i+1} = N(f, z_i)$. For notational convenience, we will write :

$$\begin{aligned}
\beta_i &= \beta(f, z_i) \\
\gamma_i &= \gamma(f, z_i) \\
\alpha_i &= \alpha(f, z_i) = \beta_i \gamma_i \\
u_i &= \frac{\|z_{i+1} - z_i\|_2}{\|z_i\|_2} \gamma_i \\
\tilde{\alpha}_i &= (\beta_i + \delta)\gamma_i \\
\psi(u) &= 1 + 2u^2 - 4u
\end{aligned}$$

The following bounds are obvious, since $\gamma_i \geq 1$:

$$\frac{\|z_{i+1}\|_2}{\|z_i\|_2} \leq 1 + u_i \tag{1}$$

$$\frac{\|z_i\|_2}{\|z_{i+1}\|_2} \leq \frac{1}{1 - u_i} \tag{2}$$

Let $p(z)$ be the projection of $\mathbb{C}^{n+1}$ into the n-plane $V(z)$, in the direction of $\ker Df(z)$. (We assume that $Df(z)$ has rank n). Let $p(z', z)$ be the restriction of $p(z)$ to $V(z')$. Let κ be a constant, $\kappa \geq \|p(z_i, z_{i+1})\|_2$ for all i. In the cases $N = N^{\text{proj}}$ and $N = N^{\text{pseu}}$, we require the stronger condition $\kappa \geq \|p(z_{i+1})\|_2$.

If we are using $N = N^{\text{aff}}$, we have $V(z_i) = V(z_{i+1})$, hence can take $\kappa = 1$.

If we are using $N = N^{\text{pseu}}$, then by construction we have that $V(z_i) \perp \ker Df(z_i)$. It follows that we can also take $\kappa = 1$. Later on, we will bound κ in the case $N = N^{\text{proj}}$.

The idea of the proof of the quadratic convergence theorem will be to show that, under certain circumstances, $\tilde{\alpha}_{i+1} \leq 4{\tilde{\alpha}_i}^2$

We start with :

Lemma 2. *Under the conditions above,*

$$\beta_{i+1} - \kappa^2 \frac{\|z_i\|_2}{\|z_{i+1}\|_2} \frac{(1-u_i)^2}{\psi(u_i)} \delta \leq \kappa \frac{\|z_i\|_2}{\|z_{i+1}\|_2} \frac{1-u_i}{\psi(u_i)} \gamma_i (\beta_i + \delta)^2$$

Proof of Lemma 2

In order to prove Lemma 2, we break β_{i+1} as follows :

$$\begin{aligned}
\beta_{i+1} &= \frac{1}{\|z_{i+1}\|_2} \left\| {Df(z_{i+1})_{|V(z_{i+1})}}^{-1} f(z_{i+1}) \right\|_2 \\
&\leq \frac{1}{\|z_{i+1}\|_2} \left\| {Df(z_{i+1})_{|V(z_{i+1})}}^{-1} Df(z_{i+1})_{|V(z_i)} \right\|_2 \\
&\left\| {Df(z_{i+1})_{|V(z_i)}}^{-1} Df(z_i)_{|V(z_i)} \right\|_2 \left\| {Df(z_i)_{|V(z_i)}}^{-1} f(z_{i+1}) \right\|_2
\end{aligned} \tag{3}$$

Part 1 : ${Df(z_{i+1})_{|V(z_{i+1})}}^{-1} Df(z_{i+1})_{|V(z_i)}$ is the projection $p(z_i, z_{i+1})$ from $V(z_i)$ into $V(z_{i+1})$ in the direction $\ker Df(z_{i+1})$. It follows that its norm is bounded by κ :

$$\left\| {Df(z_{i+1})_{|V(z_{i+1})}}^{-1} Df(z_{i+1})_{|V(z_i)} \right\|_2 \leq \kappa \tag{4}$$

Part 2 : We first write :

$${Df(z_i)_{|V(z_i)}}^{-1} Df(z_{i+1})_{|V(z_i)} = I + \sum_{k \geq 2} k \frac{{Df(z_i)_{|V(z_i)}}^{-1} D^k f(z_i)}{k!} (z_{i+1} - z_i)^{k-1}$$

We obtain the inequality :

$$\left\| {Df(z_i)_{|V(z_i)}}^{-1} Df(z_{i+1})_{|V(z_i)} - I \right\|_2 \leq \sum_{k \geq 2} k {u_i}^{k-1} \leq \frac{1}{(1-u_i)^2} - 1$$

It follows that :

$$\left\| {Df(z_{i+1})_{|V(z_i)}}^{-1} Df(z_i)_{|V(z_i)} \right\|_2 \leq \frac{1}{2 - \frac{1}{(1-u_i)^2}} \leq \frac{(1-u_i)^2}{\psi(u_i)} \tag{5}$$

Part 3 : We expand :

$$\begin{aligned}
{Df(z_i)_{|V(z_i)}}^{-1} f(z_{i+1}) &= {Df(z_i)_{|V(z_i)}}^{-1} f(z_i) + {Df(z_i)_{|V(z_i)}}^{-1} Df(z_i)(z_{i+1} - z_i) \\
&+ \sum_{k \geq 2} \frac{{Df(z_i)_{|V(z_i)}}^{-1} D^k f(z_i)}{k!} (z_{i+1} - z_i)^k
\end{aligned}$$

Since (by hypothesis) z_{i+1} cannot be at distance more than $\|z_i\|_2\delta$ of $z_i - Df(z_i)_{|V(z_i)}{}^{-1}f(z_i)$, the projection of z_{i+1} into $V(z_i)$ cannot be at distance more than $\kappa\|z_i\|_2\delta$ of the projection of $z_i - Df(z_i)_{|V(z_i)}{}^{-1}f(z_i)$. Thus :

$$\left\| Df(z_i)_{|V(z_i)}{}^{-1}f(z_i) + Df(z_i)_{|V(z_i)}{}^{-1}Df(z_i)(z_{i+1}-z_i)\right\|_2 \leq \kappa\delta\|z_i\|_2$$

For the terms of order ≥ 2, we have :

$$\begin{aligned}\left\|\sum_{k\geq 2}\frac{Df(z_i)_{|V(z_i)}{}^{-1}D^kf(z_i)}{k!}(z_{i+1}-z_i)^k\right\|_2 &\leq \sum_{k\geq 2}\left(\frac{\|z_{i+1}-z_i\|_2}{\|z_i\|_2}\gamma_i\right)^{k-1}\|z_{i+1}-z_i\|_2\\ &\leq \frac{u_i}{1-u_i}\|z_{i+1}-z_i\|_2\\ &\leq \frac{u_i}{1-u_i}(\beta_i+\delta)\|z_i\|_2\end{aligned}$$

Hence, we obtain :

$$\left\|Df(z_i)_{|V(z_i)}{}^{-1}f(z_{i+1})\right\|_2 \leq \frac{u_i}{1-u_i}(\beta_i+\delta)\|z_i\|_2 + \kappa\delta\|z_i\|_2 \tag{6}$$

Putting all together :

Inserting bounds (4), (5) and (6) into inequality (3), we get :

$$\beta_{i+1} \leq \frac{\|z_i\|_2}{\|z_{i+1}\|_2}\kappa\frac{1-u_i}{\psi(u_i)}\gamma_i(\beta_i+\delta)^2 + \frac{\|z_i\|_2}{\|z_{i+1}\|_2}\kappa^2\frac{(1-u_i)^2}{\psi(u_i)}\delta$$

Hence,

$$\beta_{i+1} - \left(\frac{\|z_i\|_2}{\|z_{i+1}\|_2}\kappa^2\frac{(1-u_i)^2}{\psi(u_i)}\right)\delta \leq \frac{\|z_i\|_2}{\|z_{i+1}\|_2}\kappa\frac{1-u_i}{\psi(u_i)}\gamma_i(\beta_i+\delta)^2$$

This proves Lemma 2.

3. Estimates on γ

Lemma 3. *Let $z_i \in \mathbb{C}^{n+1}$ verify $\|p(z_i,z_{i+1})\|_2 \leq \kappa$, let $\gamma_i = \gamma(f,z_i)$ and $u_i = \frac{\|z_{i+1}-z_i\|_2}{\|z_i\|_2}\gamma_i$. Then we have :*

$$\gamma_{i+1} \leq \kappa\frac{\|z_{i+1}\|_2}{\|z_i\|_2}\frac{1}{\psi(u_i)(1-u_i)}\gamma_i$$

Note that in the statement above, we do not require $\frac{\|z_{i+1}-N(z_i)\|_2}{\|z_i\|_2} \leq \delta$.

Proof of Lemma 3 : We first estimate $\left\|\frac{Df(z_{i+1})_{|V(z_{i+1})}{}^{-1}D^kf(z_{i+1})}{k!}\right\|_2$.

According to the estimates (4) and (5), we have :

$$\left\|\frac{Df(z_{i+1})_{|V(z_{i+1})}{}^{-1}D^kf(z_{i+1})}{k!}\right\|_2 \leq \kappa\frac{(1-u_i)^2}{\psi(u_i)}\left\|\frac{Df(z_i)_{|V(z_i)}{}^{-1}D^kf(z_{i+1})}{k!}\right\|_2$$

Moreover,

$$\left\| \frac{Df(z_i)_{|V(z_i)}{}^{-1} D^k f(z_{i+1})}{k!} \right\|_2 \leq \sum_{l \geq 0} \left\| \frac{Df(z_i)_{|V(z_i)}{}^{-1} D^{k+l} f(z_i)}{k! l!} \right\|_2 \|z_{i+1} - z_i\|_2{}^l$$

$$\leq \sum_{l \geq 0} \frac{k + l! (\gamma_i)^{k-1} u_i{}^l}{k! l! \|z_i\|_2{}^{k-1}}$$

$$\leq \frac{\gamma_i{}^{k-1}}{\|z_i\|_2{}^{k-1}} \sum_{l \geq 0} \frac{k + l! u_i{}^l}{k! l!}$$

$$\leq \frac{\gamma_i{}^{k-1}}{(1 - u_i)^{k+1} \|z_i\|_2{}^{k-1}}$$

Thus,

$$\left\| \frac{Df(z_{i+1})_{|V(z_{i+1})}{}^{-1} D^k f(z_{i+1})}{k!} \right\|_2 \leq \kappa \frac{\gamma_i{}^{k-1}}{\psi(u_i)(1 - u_i)^{k-1} \|z_i\|_2{}^{k-1}}$$

Using $\psi(u_i) \leq 1$, $\kappa \geq 1$ and extracting the $(k-1)$-th root, we obtain :

$$\gamma_{i+1} \leq \kappa \frac{\|z_{i+1}\|_2}{\|z_i\|_2} \frac{1}{\psi(u_i)(1 - u_i)} \gamma_i$$

This proves Lemma 3.

4. Estimates on α

In this section, we prove Theorems 1, 2, 5 and 6.

Combining Lemma 2, equation (2) and Lemma 3, we obtain the following result :

Lemma 4. *Under the hypotheses and notations of Lemma 2,*

$$\left(\beta_{i+1} - \kappa^2 \frac{1 - u_i}{\psi(u_i)} \delta \right) \gamma_{i+1} \leq \kappa^2 \frac{1}{\psi(u_i)^2} (\beta_i + \delta)^2 \gamma_i{}^2$$

Proof of theorems 1 and 5 : If we make $\delta = 0$, Lemma 4 reads:

$$\alpha_{i+1} \leq \frac{\kappa^2}{\psi(u_i)^2} \alpha_i{}^2 \tag{7}$$

Assume that we are in the hypotheses of theorems 1 or 5. Then $\kappa = 1$. Also, $u_i = \alpha_i$. Assume by induction that $\alpha_i \leq 1/8$, we obtain $\psi(u_i) > 0.531 > 1/2$ and equation (7) implies :

$$\alpha_{i+1} \leq 4 \alpha_i{}^2 \leq 1/16 \leq 1/8$$

By induction, $\alpha_i \leq 2^{-2^i - 2}$ and hence :

$$\beta_i \leq \alpha_i \leq 2^{-2^i - 2}$$

$$d_{\mathrm{proj}}(z_i, \zeta) \leq \sum_{j \geq i} \beta_i \leq 2\alpha_i \leq 2^{-2^i - 1}$$

This proves theorems 1 and 5.

Lemma 4 allows us to prove the following statement :

LEMMA 5. *Assuming the hypotheses of Lemma 2, and using the same notation, let* $\frac{\|z_{i+1} - N(z_i)\|_2}{\|z_i\|_2} \leq \delta$, $\delta > 0$, $(\beta_0 + \delta)\gamma_0 \leq 1/8$, *and suppose that for* $\delta \neq 0$, $0 \leq i < j$ *we have :*

$$\beta_{i+1} \geq \frac{1 + \frac{\kappa^2(1-u_i)}{\psi(u_i)}}{\frac{4\psi(u_i)^2}{\kappa^2} - 1}\delta \tag{8}$$

where the denominator is positive.

Then $(\beta_{i+1} + \delta)\gamma_{i+1} \leq 4((\beta_i + \delta)\gamma_i)^2$, *and hence* $(\beta_j + \delta)\gamma_j \leq 2^{-2^j - 2}$.

This also implies $\beta_j \leq 2^{-2^j - 2}$, *and* $d_{\mathrm{proj}}(z_j, \zeta) \leq 2^{-2^j - 1}$, *where* ζ *is a zero of* f.

Proof of Lemma 5 : Equation (8) is the same as :

$$\beta_{i+1} + \delta \leq \frac{4\psi(u_i)^2}{\kappa^2} \left(\beta_{i+1} - \kappa^2 \frac{1 - u_i}{\psi(u_i)} \delta \right)$$

Plugging this formula in Lemma 4, we obtain :

$$(\beta_{i+1} + \delta)\gamma_{i+1} \leq 4((\beta_i + \delta)\gamma_i)^2$$

This proves lemma 5.

Lemma 5 means that in the conditions of Theorems 2 , 4 and 6, as long as δ is *small enough* relatively to β, we have quadratic convergence. We still have to prove that as soon as we are no more in the conditions of Lemma 5, the sequence z_i gets *trapped* in a disk of radius 6δ over ζ.

LEMMA 6. *If equation (8) is not true, and* $u_i \leq 1/16$, *then* $\beta_{i+1} \leq \frac{28}{15}\kappa^4\delta$

Proof of Lemma 6 :

$$\beta_{i+1} < \frac{1 + \frac{\kappa^2(1-u_i)}{\psi(u_i)}}{\frac{4\psi(u_i)^2}{\kappa^2} - 1}\delta \leq \frac{1 + \frac{4}{3}\kappa^2}{\frac{9}{4\kappa^2} - 1}\delta \leq \frac{28}{15}\kappa^4\delta$$

LEMMA 7. *Let* ζ *be a zero of* f. *Let the disk* D *of center* ζ *and radius* $2k\delta$ *,* $k \geq 3$, *verify for each* $z \in D$ *the condition* $\gamma(f, z) \leq \Gamma$, *with* $(k+1)\Gamma\delta < 0.1$. *Let* $\beta(z_i) < k\delta$, $z_i \in D$. *Then* $\beta(z_{i+1}) \leq (\frac{k\kappa}{6} + 2\kappa^2)\delta$. *In particular, if* $\kappa = 1$, $\beta(z_{i+1}) \leq k\delta$ *and* $z_{i+1} \in D$.

Proof of Lemma 7 : According to Lemma 2,

$$\beta_{i+1} \leq \kappa \frac{\|z_i\|_2}{\|z_{i+1}\|_2} \frac{1-u_i}{\psi(u_i)} \gamma_i (\beta_i + \delta)^2 + \kappa^2 \frac{\|z_i\|_2}{\|z_{i+1}\|_2} \frac{(1-u_i)^2}{\psi(u_i)} \delta$$

Using equation (2),

$$\begin{aligned} \beta_{i+1} &\leq \frac{\kappa}{\psi(u_i)} \Gamma((k+1)\delta)^2 + \frac{\kappa^2(1-u_i)}{\psi(u_i)}\delta \\ &\leq \left(\frac{\kappa(k+1)^2}{\psi((k+1)\delta\Gamma)} \Gamma\delta + \frac{\kappa^2}{\psi((k+1)\Gamma\delta)} \right) \delta \\ &\leq \frac{\kappa(k+1)^2\Gamma\delta + \kappa^2}{\psi((k+1)\Gamma\delta)} \delta \end{aligned}$$

Using $(k+1)\Gamma\delta < 0.1$, we get $\psi((k+1)\Gamma\delta) > 0.6$, hence :

$$\beta_{i+1} \leq \frac{\kappa(k+1)0.1 + \kappa^2}{0.6} \delta \leq \left(\frac{1}{6} \kappa k + \frac{11}{6} \kappa^2 \right) \delta \leq \left(\frac{k\kappa}{6} + 2\kappa^2 \right) \delta$$

This proves lemma 7.

Lemma 8. *Let* $u = \frac{\|z - z_i\|_2}{\|z_i\|_2} \gamma_i \leq \frac{1}{16}$. *Then* $\gamma(z) \leq 1.52 \kappa \gamma_i$.

For the proof, we use Lemma 3 and equation (1), according to which :

$$\gamma(z) \leq \frac{\kappa(1+u)}{\psi(u)(1-u)} \gamma_i$$

If $u \leq 1/16$, then $\psi(u) > 3/4$ and :

$$\gamma(z) \leq \frac{4 \times 17}{3 \times 15} \kappa \gamma_i \leq 1.52 \kappa \gamma_i$$

Proof of theorems 2 and 6 :

Assume the hypotheses of Theorem 2 (resp. of Theorem 6).

Let $k = 3$.

Let us fix j such that $\beta_i \geq k\delta$ for $i \leq j$ and $\beta_{j+1} \leq k\delta$. Let D be the disk of radius $2k\delta$ over ζ.

Assuming that $\alpha_i < \frac{1}{8}$, Lemma 2 implies that $d_{\text{proj}}(z_i, \zeta) \leq 2\beta_i$. Indeed, by applying Lemma 2 to the *exact* Newton iteration starting from z_i, one would obtain :

$$\beta_{i+1} \leq \frac{\kappa}{\psi(\alpha_i)} \alpha_i \beta_i < \frac{\beta_i}{2} \tag{9}$$

Therefore, ζ is at distance at most $2\beta_j$ of z_j, and all points in D are within distance $4\beta_j$ of z_j. We have to consider several cases :

General case : $j \geq 1$. In that case, $\alpha_j \leq 4{\alpha_0}^2 \leq 1/64$. If $z \in D$ is scaled properly, $\frac{\|z-z_j\|_2}{\|z_j\|_2}\gamma_j \leq 4\alpha_j \leq 1/16$. Therefore, we can apply Lemma 8 and obtain $\gamma(z) \leq 1.52\kappa\gamma_j$.

We fix $\Gamma \leq 2\gamma_j$.

We also have $(k+1)\Gamma\delta < 2\frac{k+1}{k}\alpha_j \leq 3\alpha_j < 0.1$.

Hence, we apply Lemma 7 by induction, and conclude that $z_{j+1}, z_{j+2}, \dots$ belong to D.

Special cases : $j = 0$ and j does not exist (this means that $\beta_0 < 3\delta$). The case $j = 0$ is the more difficult, so we prove only this case. The proof of the other case is similar.

We claim that $\Gamma = 4\gamma_0$ verifies $\max_D \gamma \leq 2\gamma_1 \leq 4\gamma_0 = \Gamma$. Indeed, $u_0 < 1/16$, hence by Lemma 8, $\gamma_1 \leq 2\gamma_0$. The distance from z_1 to any point of D is bounded by $4k\delta$, and $4k\delta\gamma_1 \leq 8k\delta\gamma_0 < 1/16$. We can use Lemma 8 again, and conclude that $\max_D \gamma(z) \leq 2\gamma_1 \leq 4\gamma_0$. Thus, we can set $\Gamma = 4\gamma_0$.

In order to use Lemma 7, we have to check that $(k+1)\Gamma\delta < 0.1$. This amounts to check that $4(k+1)\gamma_0\delta < 0.1$. This follows from the hypothesis on $\gamma_0\delta$. Thus, we use Lemma 7 by induction, and conclude that $z_{j+k} \in D$.

Theorems 2 and 6 are now proved. In order to prove Theorems 3 and 4 we still need to be able to bound κ.

5. Estimates on κ

We prove here Theorems 3 and 4. Let z_i be a sequence such that :

$$\frac{\left\|z_{i+1} - N^{\text{proj}}(z_i)\right\|_2}{\|z_i\|_2} \leq \delta \ , \delta \geq 0$$

Assume that $\tilde{\alpha}_i^{\text{proj}} = (\beta_i^{\text{proj}} + \delta)\gamma_i^{\text{proj}} \leq \bar{\alpha}_0$, $\bar{\alpha}_0$ a constant no more than $1/32$.

Since $\alpha^{\text{pseu}}(z_i) \leq \alpha^{\text{proj}}(z_i) \leq \tilde{\alpha}^{\text{proj}}(z_i) \leq \bar{\alpha}_0$, it follows from equation (9) that there is a zero ζ at distance of z_i no more than $2\beta^{\text{pseu}}(z_i) \leq 2\bar{\alpha}_0/\gamma^{\text{pseu}}(z_i) \leq 1/16\gamma^{\text{pseu}}(z_i)$.

Lemma 9. *In the conditions above,*

$$\|p(z_{i+1})\|_2 \leq \frac{1}{\sqrt{1 - \left(\frac{9.12\bar{\alpha}_0}{\psi(4.56\bar{\alpha}_0)}\right)^2}}$$

Proof of Lemma 9 : Let us scale ζ so that $\|\zeta\|_2 = 1$. Also, we can scale z_i so that $z_i \in \zeta + \zeta^\perp$. Let us choose $y \in \ker Df(z_{i+1})$. By similarity of triangles,

$$\|p(z_{i+1})\|_2 = \sqrt{1 - d_{\text{proj}}(y, z_{i+1})^2}^{\,-1}$$

We now estimate $d_{\text{proj}}(y, z_{i+1})$. We set :

$$v = \frac{\|z_{i+1} - \zeta\|_2}{\|\zeta\|_2} \gamma^{\text{pseu}}(\zeta)$$

We apply Lemma 8 to obtain :

$$\gamma^{\text{pseu}}(\zeta) \leq 1.52 \gamma^{\text{pseu}}(z_i) \leq 1.52 \gamma^{\text{proj}}(z_i)$$

Also,

$$\frac{\|z_{i+1} - \zeta\|_2}{\|\zeta\|_2} \leq \frac{\|z_{i+1} - z_i\|_2}{\|\zeta\|_2} + \frac{\|z_i - \zeta\|_2}{\|\zeta\|_2} \leq \frac{\|z_i\|_2}{\|\zeta\|_2}(\beta_i + \delta + 2\beta_i) \leq 3(\beta_i + \delta)$$

Hence,

$$v = \frac{\|z_{i+1} - \zeta\|_2}{\|\zeta\|_2} \gamma^{\text{pseu}}(\zeta) \leq 4.56 \bar{\alpha}_0$$

We can scale y so that we can write $y = \zeta + y^\perp$, $y^\perp \perp \zeta$, then $d_{\text{proj}}(y, \zeta) \leq \|y^\perp\|_2$. By the choice of y,

$$Df(z_{i+1}) y = Df(z_{i+1})(\zeta + y^\perp) = 0$$

Expanding Df around ζ, we obtain :

$$Df(\zeta)(\zeta + y^\perp) + \sum_{k \geq 2} k \left(\frac{D^k f(\zeta)}{k!} \right) (z_{i+1} - \zeta)^{k-1} \left(\zeta + y^\perp \right) = 0$$

Obviously, $Df\zeta = 0$; we apply $Df(z_{i+1})^\dagger$ to the equation, and obtain :

$$y^\perp + \sum_{k \geq 2} k \left(\frac{Df(\zeta)^\dagger D^k f(\zeta)}{k!} \right) (z_{i+1} - \zeta)^{k-1} \left(\zeta + y^\perp \right) = 0$$

Now, we have :

$$\begin{aligned} \left\| \sum_{k \geq 2} k \left\{ \frac{Df(\zeta)^\dagger D^k f(\zeta)}{k!} \right\} (z_{i+1} - \zeta)^{k-1} \right\|_2 &\leq \sum k \gamma(\zeta)^{k-1} (z_{i+1} - \zeta)^{k-1} \\ &\leq \sum k v^{k-1} \\ &\leq \frac{1}{(1-v)^2} - 1 \end{aligned}$$

Thus,

$$d_{\text{proj}}(y, \zeta) \leq \|y^\perp\|_2 \leq \frac{\frac{1}{(1-v)^2} - 1}{2 - \frac{1}{(1-v)^2}} \leq \frac{2 - v}{\psi(v)} v \leq \frac{2v}{\psi(v)}$$

Putting all together,

$$\|p(z_{i+1})\|_2 \leq \sqrt{1 - \left(\frac{2 \times 4.56\bar{\alpha}_0}{\psi(4.56\bar{\alpha}_0)}\right)^2}^{-1}$$

and Lemma 9 is proved.

Proof of Theorem 3 We check numerically (using Lemma 9) that for $\bar{\alpha}_0 = 1/32$, we have $\kappa \leq 1.26$.

Also, $\frac{\kappa^2}{\psi(\bar{\alpha}_0)^2} \leq 2.06 \leq 4$, so equation 7 gives :

$$\alpha_{i+1} \leq 4{\alpha_i}^2 \leq 1/32$$

And this proves theorem 3.

Proof of Theorem 4 : Using $\kappa \leq 1.2567$, Lemmas 6, 7 and 8 become :

LEMMA 10. *If equation (8) is not true, and $u_i \leq 1/16$, then $\beta_{i+1} \leq 4.66\delta < 5\delta$*

LEMMA 11. *Let ζ be a zero of f. Let the disk D of center ζ and radius $2k\delta$, $k \geq 4$, verify for each $z \in D$ the condition $\gamma(f,z) \leq \Gamma$, with $(k+1)\Gamma\delta < 0.1$. Let $\beta(z_i) < k\delta$, $z_i \in D$. Then $\beta(z_{i+1}) \leq (\frac{k\kappa}{6} + 2\kappa^2)\delta$. In particular, if $\kappa \leq 1.2567$, then $\beta(z_{i+1}) \leq k\delta$ and $z_{i+1} \in D$.*

LEMMA 12. *Let $u = \frac{\|z-z_i\|_2}{\|z_i\|_2}\gamma_i \leq \frac{1}{16}$. Then $\gamma(z) \leq 1.52\kappa\gamma_i \leq 1.92 < 2$.*

At this time, we set $k = 5$. The same proof of Theorems 2 and 6 applies word by word to prove Theorem 4.

6. Proof of the Robustness results

Proof of Lemma 1 :

The first estimate is easy. The second follows from :

$$\begin{aligned}
\mu(g,\zeta) &= \mu(\lambda g, \zeta) \\
&= \|\lambda g\|_{\mathrm{k}} \left\| \left(\operatorname{diag}({d_i}^{-1/2} {\|\zeta\|_2}^{1-d_i}) D(\lambda g(\zeta))_{|V_g(\zeta)} \right)^{-1} \right\|_2 \\
&\leq \|\lambda g\|_{\mathrm{k}} \left\| \left(\operatorname{diag}({d_i}^{-1/2} {\|\zeta\|_2}^{1-d_i}) D(\lambda g(\zeta))_{|V_f(\zeta)} \right)^{-1} \right\|_2
\end{aligned}$$

Then we proceed as in [**12**], using Lemma 5 of III-1.

Bounds on $\alpha(f,x)$: Let us put ourselves in the conditions of Theorems 7 , 8 or 9.

By hypothesis, $\frac{\|x-\zeta\|_2}{\|\zeta\|_2} \leq \bar{u}$. Also, we assume that $\bar{u} < 1/16$.

The following estimate is very similar to Lemma 2 :

Lemma 13.

$$\beta(f,x) \leq \kappa(1-\bar{u})\frac{(1-\bar{u})\beta(f,\zeta)+\kappa d_{\mathrm{proj}}(x,\zeta)}{\psi(\bar{u})}\frac{\|\zeta\|_2}{\|x\|_2}$$

Proof of Lemma 13 : Using the fact that $\frac{\|x-\zeta\|_2}{\|\zeta\|_2}\gamma(\zeta) \leq \bar{u}$, we can write, using Parts 1 and 2 of the proof of Lemma 2 :

$$\beta(f,x) \leq \kappa\frac{(1-\bar{u})^2}{\psi(\bar{u})}\frac{1}{\|x\|_2}\left\|Df(\zeta)_{|V(\zeta)}{}^{-1}f(x)\right\|_2$$

Expanding the last term into its Taylor series, we obtain :

$$\begin{aligned}\left\|Df(\zeta)_{|V(\zeta)}{}^{-1}f(x)\right\|_2 &\leq \left\|Df(\zeta)_{|V(\zeta)}{}^{-1}f(\zeta)+Df(\zeta)_{|V(\zeta)}{}^{-1}Df(\zeta)(x-\zeta)\right\|_2\\ &+\left\|\sum_{k\geq 2}\frac{Df(\zeta)_{|V(\zeta)}{}^{-1}D^kf(\zeta)}{k!}(x-\zeta)^{k-1}\right\|_2\\ &\leq \|\zeta\|_2(\beta(f,\zeta)+\kappa d_{\mathrm{proj}}(x,\zeta))+\|\zeta\|_2\frac{\bar{u}}{1-\bar{u}}d_{\mathrm{proj}}(x,\zeta)\end{aligned}$$

Therefore,

$$\begin{aligned}\beta(f,x) &\leq \kappa\frac{\|\zeta\|_2}{\|x\|_2}(1-\bar{u})\frac{(1-\bar{u})\beta(f,\zeta)+((1-\bar{u})\kappa+\bar{u})d_{\mathrm{proj}}(x,\zeta)}{\psi(\bar{u})}\\ &\leq \kappa\frac{\|\zeta\|_2}{\|x\|_2}(1-\bar{u})\frac{(1-\bar{u})\beta(f,\zeta)+\kappa d_{\mathrm{proj}}(x,\zeta)}{\psi(\bar{u})}\end{aligned}$$

This proves Lemma 13

Lemma 3 gives :

$$\gamma(f,x) \leq \kappa\frac{\gamma(f,\zeta)}{\psi(\bar{u})(1-\bar{u})}\frac{\|x\|_2}{\|\zeta\|_2}$$

Using Lemma 13 together with the previous estimate, we obtain :

Lemma 14.

$$\alpha(f,x) \leq \kappa^2\frac{(1-\bar{u})\alpha(f,\zeta)+\kappa\bar{u}}{\psi(\bar{u})^2}$$

Now we use Lemma 2 and obtain :

Lemma 15.

$$\beta(f,x') \leq \kappa\frac{\|x\|_2}{\|x'\|_2}\frac{1-\alpha(f,x)}{\psi(\alpha(f,x))}\alpha(f,x)\beta(f,x)$$

Proof of Theorems 7 and 9 :

We first set $\kappa = 1$. Let us assume for a while that :

$$\frac{(1-\bar{u})\bar{\alpha}+\bar{u}}{\psi(\bar{u})^2} < 1/32 \tag{10}$$

It follows from Lemma 14 that $\alpha(f,x) < 1/8$, and Lemma 15 implies :

$$\beta(f,x') \leq \frac{1}{8\psi(1/8)}\beta(f,x)$$

Hence :

$$\beta(f,x') \leq \frac{1}{8\psi(1/8)}(1-\bar{u})\frac{(1-\bar{u})\bar{\alpha}+\bar{u}}{\psi(\bar{u})}\frac{1}{\bar{\gamma}}$$

Hence, in order to obtain $\beta(f,x') \leq \bar{u}/2\bar{\gamma}$, we need :

$$\frac{1}{8\psi(1/8)}(1-\bar{u})\frac{(1-\bar{u})\bar{\alpha}+\bar{u}}{\psi(\bar{u})} < \frac{\bar{u}}{2} \tag{11}$$

Numerically, we can verify that : $\bar{u} = 0.05$ and $\bar{\alpha} = 0.02$ make conditions (10) and (11) true, proving Theorems 7 and 9.

Proof of Theorem 8 :

Let us assume now that :

$$\kappa^2\frac{(1-\bar{u})\bar{\alpha}+\kappa\bar{u}}{\psi(\bar{u})^2} < 1/32 \tag{12}$$

It follows from Lemma 14 that $\alpha(f,x) < 1/32$, and from Lemma 15 we obtain :

$$\beta(f,x') \leq \kappa\frac{1}{32\psi(1/32)}\beta(f,x)$$

Hence :

$$\beta(f,x') \leq \kappa\frac{1}{32\psi(1/32)}(1-\bar{u})\frac{(1-\bar{u})\bar{\alpha}+\bar{u}}{\psi(\bar{u})}\frac{1}{\bar{\gamma}}$$

Hence, in order to obtain $\beta(f,x') \leq \bar{u}/2\bar{\gamma}$, we need :

$$\kappa\frac{1}{32\psi(1/32)}(1-\bar{u})\frac{(1-\bar{u})\bar{\alpha}+\bar{u}}{\psi(\bar{u})} < \frac{\bar{u}}{2} \tag{13}$$

If we further assume $\bar{\alpha} < 1/32$, we have always $\kappa < 1.2567$. Numerically, we can verify that : $\bar{u} = 0.005$ and $\bar{\alpha} = 0.01$ make conditions (12) and (13) true, proving Theorem 8.

Proof of Theorem 10 :

We first set :

$$\bar{\gamma} \geq \frac{2}{3}D^{3/2}\mu$$

and :

$$\Delta \leq \frac{3}{8}\frac{\bar{\alpha}}{\mu\bar{\gamma}} \leq \frac{9}{16}\frac{\bar{\alpha}}{\mu^2 D^{3/2}}$$

We assume by induction that $d_{\text{proj}}(z_{i-1}, \zeta_{t_{i-1}}) \leq \frac{\bar{u}}{\bar{\gamma}}$. We want to verify that we are in the conditions of Corollary 1.

Using $\bar{\alpha} \leq 0.02$, we obtain the estimates : $\Delta \leq 0.04$ and $\sqrt{D}\Delta\mu \leq 0.04$.

We use Lemma 1 :

$$\begin{aligned}
\gamma(f_{t_i}, \zeta_{t_{i-1}}) &\leq \frac{D^{3/2}}{2}\mu(f_{t_i}, \zeta_{t_{i-1}}) \\
&\leq \frac{D^{3/2}}{2}\mu(f_{t_{i-1}}, \zeta_{t_{i-1}})\frac{1+\Delta}{1-\sqrt{D}\Delta\mu} \\
&\leq \frac{1.04}{1.92}D^{3/2}\mu \\
&\leq \frac{2}{3}D^{3/2}\mu \leq \bar{\gamma}
\end{aligned}$$

$$\begin{aligned}
\eta(f_{t_i}, \zeta_{t_{i-1}})\mu(f_{t_i}, \zeta_{t_{i-1}}) &\leq \Delta\mu\frac{1+\Delta}{1-\sqrt{D}\Delta\mu} \\
&\leq \frac{3}{8}\frac{1.04}{0.96}\frac{\bar{\alpha}}{\bar{\gamma}} \\
&\leq \frac{\bar{\alpha}}{\bar{\gamma}}
\end{aligned}$$

Now we can apply Corollary 1, and conclude that $d_{\text{proj}}(z_i, \zeta_{t_i}) \leq \frac{\bar{u}}{\bar{\gamma}}$. This proves Theorem 10.

CHAPTER 3

Construction of the Approximate Newton Operator

The approximate Newton Operators are constructed, and their complexity is estimated.

1. Introduction

Let $f = (f_1, \ldots, f_n)$ be a system of polynomials of degree $d = (d_1, \ldots, d_n)$ in variables x_i, with Gaussian integer coefficients. Let $D = \max\ d_i$. Let $S(f)$ be the number of non-zero coefficients of f. We will consider together the case in which the f_i are polynomials in variables $x_1, \ldots, x_n$ representing points in affine space, and the case in which the f_i are homogeneous in variables $x_0, \ldots, x_n$.

In chapter 2, we defined three versions of the Newton operator. In the case the f_i are non-homogeneous, the (classical) Newton method in affine space is defined by :

$$N^{\mathrm{aff}}(f, x) = x - Df(x)^{-1} f(x)$$

In the homogeneous case, we can define methods :

$$N^{\mathrm{proj}}(f, x) = x - \begin{bmatrix} Df(x) \\ x^* \end{bmatrix}^{-1} \begin{bmatrix} f(x) \\ 0 \end{bmatrix}$$

$$N^{\mathrm{pseu}}(f, x) = x - Df(x)^{\dagger} f(x)$$

where x^* is the complex transpose of vector x and $A^{\dagger}$ is the Moore-Penrose pseudo-inverse of matrix A.

The two methods above can be considered as mappings from the space $\mathbb{C}^{n+1}$ into itself. We can also consider the affine method as a mapping from a subset of $\mathbb{C}^{n+1}$ into itself :

Indeed, if f is non-homogeneous, we define :

$$N^{\mathrm{aff}}(x_0, \ldots, x_n) = {x_0}^d N^{\mathrm{aff}}(x_1/x_0, \ldots, x_n/x_0)$$

so that we have mappings of all spaces $x_0 = c$ into themselves, for $c \neq 0$. We unified the three methods by the notation $N(f, x) = x - Df(x)_{|V(x)}{}^{-1}$, where $V(x)$ associates to each x an hyperplane passing through x, and $V(\lambda x) = \lambda V(x)$.

Since $N(f, \lambda x) = \lambda N(f, x), \lambda \in \mathbb{C}$, it follows that $N(f, .)$ sends lines through the origin into lines through the origin. Therefore, it makes sense to consider $N(f, .)$ as a mapping of the projective space.

The system f will be given as a list of polynomials $f_1, \dots, f_n$. Each polynomial f_i is a list of monomials $\dots f_{iJ} x^J \dots$. A monomial is a list of $n + 3$ integers : $\text{Re}(f_{iJ}), \text{Im}(f_{iJ}), J_0, J_1, \dots, J_n$, where J_0 can be omitted in the non-homogeneous case.

We define the height $H(f) = \max(|\text{Re}(f_{iJ})| + |\text{Im}(f_{iJ})|)$. This is not the standard definition of Height. The point x will be given as a list $(x_0, \dots, x_n)$ of Gaussian integers. Of course, if the input is given as a list of complex floating point numbers, the same complexity analysis below will be true. We define the *height* $H(x)$ of x as $H(x) = \max(|\text{Re}(x_i)| + |\text{Im}(x_i)|)$.

The objective of this chapter is to construct algorithms to compute *approximately* the operators N^{aff}, N^{proj} and N^{pseu}. This means that if N is one of those operators, and if N' is the mapping computed by some algorithm, we want to guarantee that the *error* $\sup_x \frac{\|N'(f,x) - N(f,x)\|_2}{\|x\|_2}$ is small. Also, we want a bound on the *complexity* of those algorithms.

The concepts of *algorithm* and *complexity* require the definition of a computation and complexity model. Here, we will consider two different settings : numerical analysis and theoretical computer science.

From the numerical analysis point of view, the complexity will be the number of floating point operations performed for a given input. For the model of computation, we will explain in section 2 what we mean by an ϵ_m*-machine*, or a *machine* performing *correctly rounded* finite precision floating point arithmetics *with monotone subtraction.* The later are standard properties verified by the arithmetic of most computers today.

One consequence of the correctly rounded arithmetic is that the computed result $\text{fl}(a \diamond b)$ of an arithmetic operation $a \diamond b$ verifies : $\text{fl}(a \diamond b) = (a \diamond b)(1 + \epsilon)$ with $|\epsilon|$ not greater than a fixed constant ϵ_m, called the *machine epsilon.*

THEOREM 11. *Let* $16 < k \in \mathbb{N}$ *be fixed. Then there is a machine* N' *performing correctly rounded arithmetic with monotone subtraction, with machine epsilon*

$\epsilon_m = 2^{-k-1}$, *and requiring* $O(nDS(f)+n^3)$ *floating point operations, such that, if* $H(f), H(x) \leq \frac{1}{2\epsilon_m}$ *:*

$$\begin{aligned}\frac{\|N'(f,x)-N(f,x)\|_2}{\|x\|_2} \leq & 4\mu(f,x)\sqrt{D}\left(p(n)+2D+2+2\max(S(f_i))\right)\epsilon_m \\ & + O({\epsilon_m}^2)\end{aligned}$$

where $O({\epsilon_m}^2)$ *means a function bounded by some* $B(n,D,S(f),\mu){\epsilon_m}^2$*.*

Above, $p(n)$ is a function of n, defined as $5n^2(g_{cp}(n)+1)$ for the affine method, as $5(n+1)^2(g_{cp}(n+1)+1)$ for the projective method, and $(3.25(n+1)^3+11(n+1)^2)$ for the pseudo-Newton method. $g_{cp}(n)$ is the *maximum complete pivot growth* for a matrix of size n , see Golub and Van Loan [**5**] or Wilkinson [**19**]. The function $g_{cp}(n)$ can be bounded by the formula $g_{cp}(n) \leq \sqrt{n}\sqrt{2 \times 3^{\frac{1}{2}} \times \cdots \times n^{\frac{1}{n-1}}}$.

The *condition number* $\mu(f,x)$ is defined by :

$$\mu(f,x) = \|f\|_{\mathrm{k}} \left\| {Df(x)_{|V(x)}}^{-1} \mathrm{diag}(\sqrt{d_i}{\|x\|_2}^{d_i-1}) \right\|_2$$

The *Kostlan norm* $\|.\|_{\mathrm{k}}$ is the unitarily invariant norm in the space of polynomials , and is defined by $\|f\|_{\mathrm{k}} = \sqrt{\sum {\|f_i\|_{\mathrm{k}}}^2}$, where :

$$\|f_i\|_{\mathrm{k}} = \sqrt{\sum_{|J|=d_i} \frac{|f_{iJ}|^2}{\begin{pmatrix} d_i \\ J \end{pmatrix}}}$$

As it often happens in numerical analysis, the linear factor in theorem 11 may be extremely pessimistic. The actual error may be much smaller. On the other hand, the result above says little about the behavior of the high order terms in ${\epsilon_m}^2$.

From the point of view of theoretical computer science, we will use the Blum - Shub - Smale model over $\mathbb{Z}$ (See [**3**]) (equivalent to the Turing model), and give a bound for the height of numbers used in calculations. The cost of each operation with numbers of height H will be $O(\log H)$ or $O(\log H \log\log H)$. The cost of accessing position n in the memory is n. In this setting, we obtain :

THEOREM 12. *There is a machine over* $\mathbb{Z}$ *such that, given a system* f *of* n *polynomials of degree at most* D*, with S(f) non-zero Gaussian integer coefficients, a vector* x *of floating-point complex numbers,* $\bar{\mu} \geq \mu(f,x)$*, and* $\delta > 0$*, returns* $N_\delta(f,x)$ *such that :*

$$\frac{\|N_\delta(f,x) - N^{\mathrm{pseu}}(f,x)\|_2}{\|x\|_2} \leq \delta$$

within polynomial time in n, D, $\log H(f)$, $\log H(x)$, $S(f)$, $\log \bar{\mu}$, $-\log \delta$.

We will prove this theorem only for the pseudo-Newton method. The proof for the affine and projective Newton methods is similar.

In all this chapter, we assume without loss of generality that x is scaled so that $\frac{1}{2} < \|x\|_2 \leq 1$ (just by performing a division by a power of 2).

The results above were obtained after a few arbitrary choices. In many senses, they are not the best possible algorithms for that problem.

Simple precision is used wherever it is possible. While the usual numerical analysis literature uses double precision, for instance, to add a list of single precision numbers, this procedure requires an additional hypothesis on the size of the list. See, for instance, Golub and Van Loan [**5**] sections 2.4.6 and 2.4.7, Wilkinson [**18**] equation (5.2) of chapter 3, page 113, or ibid., start of section 37, chapter 3, page 152.

In our setting, computing the dot product $< v, w >= v^* w$, where v^* is the complex transpose of v, and v and w are vectors of dimension n, will introduce a total error in the result that is bounded by :

$$((1+\epsilon_m)^n - 1)\, \|v\|_2 \|w\|_2 \leq n\epsilon_m \|v\|_2 \|w\|_2$$

Using double precision we would obtain error bounded by $\epsilon_m \|v\|_2 \|w\|_2$, but only provided n is small enough.

Also, simple and vectorizable algorithms are given preference over more precise and expensive ones. For instance, there are single precision algorithms for adding a list of numbers up to precision ϵ_m, or up to arbitrary precision (See Priest, [**9**]). But precision is not a severe issue here, while speed is crucial.

2. Basic definitions

First of all, we have to define (for the numerical analysis purpose) a model of computation. We want to model the floating point arithmetic of modern computers. Let us fix a number $k \geq 16$, the number of bits of mantissa.

DEFINITION 2. The set $\mathbb{R}_{fp}$ of *floating point numbers* is the class of all reals of the form $A2^B$, with $A, B \in \mathbb{Z}$ and $|A| \leq 2^k$.

This definition deliberately ignores the possibility of overflow or underflow (exponents too small or to large). Now we can define a model of computer :

DEFINITION 3. An ϵ_m*-machine* means a Blum-Shub-Smale machine over the reals [**3**], with coefficients in $\mathbb{R}_{fp}$, with the following modifications :

- The even positions of memory contain always integer values.
- The odd positions of memory contain always floating point numbers in $\mathbb{R}_{fp}$.
- Computation nodes contain one arithmetic operation each, from $+$, $-$, $/$, $*$, and square root.
- Decision nodes are in the form $x_i \geq 0$ or $x_i > 0$ or $x_i = 0$.
- If a division by zero or a square root of a negative number occur, execution stops and the machine outputs nothing.
- There is a function $\text{fl} : \mathbb{R} \to \mathbb{R}_{fp}$ such that, when a real number x is to be stored in an odd position of memory, it gets rounded-off to $\text{fl}(x)$
- *Correct arithmetics :* $\text{fl}(x)$ is one of the elements of $\mathbb{R}_{fp}$ nearest to x.
- *Monotone subtraction :* For all $a, b \in \mathbb{R}_{fp}$, we have : $\text{fl}(a-b) = -\text{fl}(b-a)$

Now we introduce some definitions regarding rounding-off :

A consequence of Definition 3 is the $1+\epsilon$ property. If representable numbers have t bits of mantissa, let $\epsilon_m = 2^{-t-1}$. Assuming that $t \geq 16$, we will have always $\epsilon_m \leq 2^{-17} = \frac{1}{131072}$. Then the machine verifies the $1+\epsilon$ property, defined below :

DEFINITION 4. A machine verifies the $1+\epsilon$ property if it is true, for all $a, b \in \mathbb{R}_{fp}$, that :

$$\text{fl}(a \diamond b) = (a \diamond b)(1+\epsilon) \ , \ |\epsilon| \leq \epsilon_m$$

where $\diamond$ is one of $+$, $-$, $\times$, $/$. Also, $\text{fl}(\sqrt{x}) = \sqrt{x}(1+\epsilon), |\epsilon| \leq \epsilon_m$.

We say that the relative error of $a \diamond b$ is less than ϵ_m.

This property is known to hold in most computer systems available at this time, except CRAY systems.

We will also need to perform complex arithmetics. A complex number can be represented by its real and imaginary parts, where elementary operations between real numbers are performed with correct arithmetics. Therefore, complex arithmetics can be performed by an ϵ-machine. However, it is convenient to set $\epsilon_m = 2^{-t-1} \times 6$, so the $1+\epsilon$ property will hold for complex sum, subtraction, multiplication and division. Let us prove this only for divisions :

$$\frac{z}{w} = \frac{z\bar{w}}{w\bar{w}}$$

Let $\epsilon = 2^{-t-1}$. Then $\mathrm{fl}(w\bar{w}) = w\bar{w}(1+\epsilon_1)^2$ and $\mathrm{Re}(\mathrm{fl}(z\bar{w})) = \mathrm{Re}(z\bar{w})(1+\epsilon_2)^2$, where $|\epsilon_1|, |\epsilon_2| \leq \epsilon$. It follows that :

$$\mathrm{Re}\left(\mathrm{fl}\left(\frac{z}{w}\right)\right) = \frac{\mathrm{Re}(z\bar{w})}{w\bar{w}} \frac{(1+\epsilon_3)^3}{(1-\epsilon_3)^2}$$

Since $\epsilon < 2^{-17}$, we get :

$$\mathrm{Re}\left(\mathrm{fl}\left(\frac{z}{w}\right)\right) = \frac{\mathrm{Re}(z\bar{w})}{w\bar{w}}(1+6\epsilon_4)$$

Proceeding in the same way for the imaginary part, we obtain :

$$\begin{aligned} \mathrm{fl}(z/w) - z/w &= 6\epsilon_4 \mathrm{Re}(z/w) + 6i\epsilon_5 \mathrm{Im}(z/w) \\ |\frac{\mathrm{fl}(z/w)}{z/w} - 1| &\leq 6\epsilon \\ \mathrm{fl}(z/w) &= \frac{z}{w}(1+6\epsilon_6) \end{aligned}$$

where $|\epsilon_3|, |\epsilon_4|, |\epsilon_5|, |\epsilon_6| \leq \epsilon$.

The following notation will be very useful. If ϵ_m is the *machine epsilon*, the relative error of each calculation, we define the function :

$$\epsilon_m(k) = \frac{1}{(1-\epsilon_m)^k} - 1$$

Then we have the following properties, for $k, l, n \geq 1$:

$$\begin{aligned} \epsilon_m &< \epsilon_m(1) \\ (1+\epsilon_m(k))(1+\epsilon_m(l)) &= 1+\epsilon_m(k+l) \\ \epsilon_m(k) + \epsilon_m(l) &< \epsilon_m(k+l) \\ n\epsilon_m(k) &\leq \epsilon_m(nk) \\ \frac{1+\epsilon_m(k)}{1-\epsilon_m(l)} &< 1+\epsilon_m(k+l) \\ \frac{1-\epsilon_m(k)}{1+\epsilon_m(l)} &> 1-\epsilon_m(k+l) \end{aligned}$$

However, under the assumption that $k\epsilon_m$ is small, $\epsilon_m(k) \simeq k\epsilon_m$.

We will often find inequalities of the form $p\epsilon_m(q) < 1$; those can be solved by using the

LEMMA 16. *Let $p, q > 1$, and let $\epsilon_m < \frac{1}{2pq}$. Then $p\epsilon_m(q) < 1$.*

Proof : $pq\epsilon_m < \frac{1}{2}$ implies that $1 - pq\epsilon_m > \frac{1}{2}$, hence $(1-\epsilon_m)^{pq} > \frac{1}{2}$, so $\frac{1}{(1-\epsilon_m)^{pq}} < 2$, and therefore $p\epsilon_m(q) \leq \epsilon_m(pq) < 1$.

This shows that there is a machine using k bits of precision such that $p\epsilon_m(q) < 1$, where $k < 1 + \log_2 p + \log_2 q$.

3. Algorithms

Suppose we are given x and f as in the hypotheses of Theorem 11. $S(f)$ is the number of non-zero coefficients of f. If f is dense, $S(f) = \sum \binom{d_i + n}{n}$.

If $1 \leq i \leq n$ and if K is a multi-index of degree d_i, we denote f_{iK} the coefficient of f_i associated to multi-index K.

We define the following procedure in order to compute the approximate Newton method in affine space :

```
ALGORITHM   $z'' \leftarrow$ Affine ( $f$, $x$ )
    1 Compute $y_I = x^I$, where multi-index $I$ appears in $f$ or in $Df$.
    2 Compute $b_i = \sum f_{iK} \ \ y_K$
    3 Compute $A_{ij} = \sum \deg_j(K) \ \ f_{iK} \ \ y_{K-e_j}$
    4 Compute $L$, $U$, $P_1$, $P_2$ by gaussian elimination
      with complete pivoting, where $P_1 A P_2 = LU$,
      $P_1$ and $P_2$ are permutations,
      $L$ is lower triangular with entries $\leq 1$,
      $U$ is upper triangular.  For notational convenience,
      we will forget about $P_1$ and $P_2$ and write $A = LU$.
    5 Compute $z = L^{-1} b$ by forward-substitution.
    6 Compute $z' = U^{-1} z$ by back-substitution.
    7 Compute $z'' = x - z'$.
```

The following are upper bounds of the floating point operation count for each line.

Line 1 costs at most $(1+n)(D-1)S(f)$ floating point operations, since there are $S(f)$ non-zero coefficients, each of which appears at most once in f and n times in Df.

Line 2 requires at most $2S(f)$ operations, while Line 3 requires at most $3nS(f)$ operations.

For Line 4, we need to perform Gaussian elimination, and that takes at most $\frac{2}{3}n^3$ floating point operations. See Golub and Van Loan [**5**], Algorithm 3.2.3, page 97.

Forward and backward substitution require $\frac{n(n-1)}{2}$ multiplications, $\frac{n(n-1)}{2}$ sums or subtractions and n divisions, hence each of them takes at most n^2 operations. In the case of Line 5, we know that the diagonal of L contains only ones, so we do not need to perform the n divisions. Hence, the cost of Lines 5 and 6 is bounded by $n(n-1)$ and n^2, respectively.

Line 7 requires only n operations.

Hence, the total operation count is :

$$(nD + 2n + D + 1)S(f) + \frac{2}{3}n^3 + 2n^2 \tag{14}$$

For the Newton method in Projective space, we will have $n+1$ variables, and we use the formula :

$$N^{\text{proj}}(f,x) = x - \begin{bmatrix} Df(x) \\ x^* \end{bmatrix}^{-1} \begin{bmatrix} f(x) \\ 0 \end{bmatrix}$$

Using the same algorithm, the operation count becomes :

$$(nD + 2n + 2D + 3)S(f) + \frac{2}{3}(n+1)^3 + 2(n+1)^2 \tag{15}$$

For the pseudo-Newton case, we can modify the algorithm as follows :

```
ALGORITHM   $z'' \leftarrow$ Pseudo ( $f$, $x$ )
    1 Compute $y_I = x^I$, where multi-index $I$ appears in $f$ or in $Df$.
    2 Compute $b_i = \sum f_{iK} \ y_K$
    3 Compute $A_{ij} = \sum \deg_j(K) \ f_{iK} \ y_{K-e_j}$
    4 Compute $Q$, $R$ such that $A^t = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$
    5 Compute $z = {R^t}^{-1} b$ by forward-substitution.
    6 Compute $z' = Q_{:,1:n} z$.
    7 Compute $z'' = x - z'$.
```

Lines 1 to 3 are as above.

For Line 4, we use Householder QR factorization, as in Algorithm 5.2.1, page 212 in Golub and Van Loan [**5**]. This algorithm takes $2n^2(n+1-n/3) = \frac{4}{3}n^3 + 2n^2$ floating point operations. This includes n square root operations on real numbers.

Line 5 will take at most $(n+1)^2$ operations.

Q is returned not as a matrix, but as a list of Householder vectors. Thus, computing step 6 will require $3n^2$ operations.

Operation count (summary)

Line	Affine	Projective	Pseudo-Newton
1	$(1+n)(D-1)S(f)$	$(2+n)(D-1)S(f)$	$(2+n)(D-1)S(f)$
2	$2S(f)$	$2S(f)$	$2S(f)$
3	$3nS(f)$	$3(n+1)S(f)$	$3(n+1)S(f)$
4	$\frac{2}{3}n^3$	$\frac{2}{3}(n+1)^3$	$\frac{4}{3}n^3+2n^2$
5	$n(n-1)$	$(n+1)n$	$(n+1)^2$
6	n^2	$(n+1)^2$	$3n^2$
7	n	$n+1$	$n+1$
Total	$(nD+2n+D+1)S(f)$ $+\frac{2}{3}n^3+2n^2$	$(nD+2n+2D+3)S(f)$ $+\frac{2}{3}n^3+6n^2+6n+\frac{8}{3}$	$(nD+2n+2D+3)S(f)$ $+\frac{4}{3}n^3+4n^2+3n+2$

4. Sketch of the proof of Theorems 11 and 12

Let us consider for a while the operators :

$$M^{\text{aff}} : A, b \mapsto A^{-1}b$$

$$M^{\text{proj}} : x', A, b \mapsto \begin{bmatrix} A \\ {x'}^* \end{bmatrix}^{-1} \begin{pmatrix} b \\ 0 \end{pmatrix}$$

$$M^{\text{pseu}} : A, b \mapsto A^{\dagger}b$$

We divide the algorithm in three phases : Phase 1 is lines 1 to 3, Phase 2 is lines 4 to 6, and Phase 3 is line 7.

Given x, f, Phase 1 computes A and b, with a certain error $\delta_1 A$ and $\delta_1 b$. Those are called *forward error* of Phase 1. It will be easy to find a bound for them. $x = x'$ is passed exactly to Phase 2.

The forward error of Phase 2 is harder to bound. It is much easier to prove that given A, x', b, Phase 2 returns $M(x'+\delta_2 x, A+\delta_2 A, b+\delta_2 b)$, where $\delta_2 x$, $\delta_2 A$ and $\delta_2 b$ are called the *backward error* of Phase 2. We will give bounds for those quantities.

If we set $\delta'_x = \delta_2 x$, $\delta A = \delta_1 A + \delta_2 A$, $\delta b = \delta_1 b + \delta_2 b$, then we are indeed computing $M(x+\delta x', Df(x)_{|V(x)} + \delta A, f(x) + \delta b)$, M one of M^{aff}, M^{proj}, M^{pseu}.

We will bound the derivatives of M, and then obtain bounds on the total error of Phases 1 and 2.

Phase 3 is just a subtraction, and hence, the total error of the algorithm is easy to bound.

5. Forward error analysis of Phase 1

If A is a matrix, we define $\|A\|_{\max} = \max |A_{ij}|$. This is a *vector* norm, not a *matrix* norm : it is not true in general that $\|AB\|_{\max} \leq \|A\|_{\max} \|B\|_{\max}$. In this section, we bound $\|\delta_1 A\|_{\max}$ and $\|\delta_1 b\|_2$. We recall that $\|x\| \leq 1$, so :

$$\|f(x)\| \leq \|f\|_{\mathrm{k}}$$

where $\|.\|_{\mathrm{k}}$ is the Kostlan norm.

Since each of the $f_{iK}\ x^K$ is computed with relative error $\epsilon_m(d_i)$, the sum $f_i(x)$ is computed up to a total error $\delta_1 b$ such that :

$$|\delta_1 b_i| \leq \epsilon_m (d_i + S(f_i)) \sum |f_{iK}|\ |x^K|$$

Adding squares for each i and extracting the square root, we obtain :

(16) $$\|\delta_1 b\|_2 \leq \|f\|_{\mathrm{k}} \epsilon_m (D + \max S(f_i))$$

In order to bound $\|\delta_1 A\|_{\max}$, we claim that

$$\left\| \frac{\partial}{\partial x_j} f_i \right\|_{\mathrm{k}} \leq \sqrt{d_i} \|f_i\|_{\mathrm{k}}$$

Indeed,

$$\begin{aligned} \left\| \frac{\partial}{\partial x_j} f_i \right\|_{\mathrm{k}} &\leq \sqrt{ \sum \frac{{f_{iJ}}^2}{\begin{pmatrix} d_i - 1 \\ J - e_j \end{pmatrix}} } \max J_j \\ &\leq \sqrt{d_i} \sqrt{ \sum \frac{{f_{iJ}}^2}{\begin{pmatrix} d_i \\ J \end{pmatrix}} } \end{aligned}$$

From this claim we deduce that :

(17) $$\|\delta_1 A\|_{\max} \leq \sqrt{d_i} \|f\|_{\mathrm{k}} \epsilon_m (D + S(f_i) + 1)$$

6. Some backward error analysis identities

Products : From the $(1+\epsilon)$ property, we deduce :

(18) $$\mathrm{fl}(ab) = (a + \delta a) b$$

where $|\delta a| \leq |a| \epsilon_m$.

Divisions :

$$\mathrm{fl}(b/a) = \frac{b}{a+\delta a} \tag{19}$$

with $|\delta a| \leq |a|\epsilon_m(1)$. Indeed, $\frac{b}{a}(1+\epsilon_m) = \frac{b}{a+\delta a}$ with $|\delta a| \leq |a|(\frac{1}{1+\epsilon} - 1)$, where $|\epsilon| \leq \epsilon_m$, so $(\frac{1}{1+\epsilon} - 1) \leq \epsilon_m(1)$.

Sums of products : For sums of products, we have :

$$\mathrm{fl}(\sum_{1\leq i\leq n} a_i b_i) = \sum_{1\leq i\leq n} (a_i + \delta a_i) b_i \tag{20}$$

with $|\delta a_i| \leq |a_i|\epsilon_m(n)$.

Backward and forward substitution : Now, let U be a $n\times n$ upper-triangular matrix. We want to compute $z = U^{-1}y$ by backward substitution :

$$\text{For } \ i = n \text{ down to } \ 1, \ z_i = \frac{y_i - \sum_{j>i} U_{ij} z_j}{U_{ii}}$$

Let $|U|$ denote the matrix of the absolute values of the coefficients of U. Then we have :

$$z = (U + \delta U)^{-1} y$$

where $|\delta U_{ij}| \leq \epsilon_m(n)|U_{ij}|$ when $i \neq j$, and $|\delta U_{ii}| \leq \epsilon_m(2)|U_{ii}|$. Thus, we have always, for $n \geq 2$:

$$|\delta U_{ij}| \leq \epsilon_m(n)|U_{ij}| \tag{21}$$

This is still true when $n = 1$, trivially. The same bound is true for forward substitution.

Householder transforms : We will now bound the backward error of a Householder transform. We compute the Householder transform in the following order :

$$H_v(y) = y + \left(\frac{-2(v^t y)}{(v^t v)}\right) v$$

According to equation (20), $v^t v$ and $v^t y$ are computed with total error bounded by $\|v\|_2{}^2\epsilon_m(n)$ and $\|v\|_2\|y\|_2\epsilon_m(n)$, respectively. Multiplication by -2 is just adding one to the exponent, and changing sign, so it does not introduce new rounding-off error.

Let $w = -2\frac{v^t y}{v^t v} v$. Each w_i is computed up to error $\epsilon_m(2n+2)\|y\|_2$. Therefore, and since $\|H_v(y)\|_2 = \|y\|_2$, the total error is less than $\epsilon_m(2n+3)\|y\|_2$. Since H_v is unitary, this is also the backward error :

$$\|\delta y\|_2 \leq \|y\|_2 \epsilon_m(2n+3) \tag{22}$$

Products of Householder transforms : Let $Q = H_1 H_2 \ldots H_n$ be a product of Householder transforms. Then :

$$\mathrm{fl}(Qy) = (Q + \delta Q)y = (H_1 + \delta H_1)(H_2 + \delta H_2)\ldots(H_n + \delta H_n)y$$

That means that :

$$\begin{aligned}\|\delta Q\|_2 &\leq \sum \|\delta H_i\|_2 + \sum \|\delta H_i\|_2 \|\delta H_j\|_2 + \ldots \\ &\leq \epsilon_m (2n^2 + 3n)\end{aligned}$$

If, however, the i-th Householder vector contains at most i non-zero coordinates, we obtain the bound :

$$\epsilon_m (\sum_i^n 2i + 3) = \epsilon_m (n^2 + 4n)$$

LU decomposition : We also need to bound the backward error $\delta_g A$ of the LU decomposition : $A + \delta_g A = LU$. Recall that we use complete pivoting, but we do not write the permutation matrices, since permutation introduces no floating point operation. Let $|L|$ be the matrix of the absolute values of elements of L. Since we perform $n-1$ pivot steps, we have the formula below (see also Theorem 3.3.1 page 105 in Golub and Van Loan [**5**]) :

$$|\delta_g A| \leq |L||U|\epsilon_m (3n - 3)$$

By construction $|L_{ij}| \leq 1$, but $|U_{ij}|$ is harder to bound ; indeed, it is usual to define the *pivot growth factor* g_{cp} such that, *in exact arithmetic* :

$$|U_{ij}| \leq g_{cp} \max |A_{ij}| \qquad (23)$$

It is known that g_{cp}, in the case of complete pivoting, can be bounded by the formula

$$g_{cp}(n) \leq \sqrt{n}\sqrt{2 \times 3^{\frac{1}{2}} \times \cdots \times n^{\frac{1}{n-1}}}$$

We obtain the formula :

$$|\delta A_{ij}| \leq n g_{cp}(n) \max |A_{ij}| \epsilon_m (3n - 3) \qquad (24)$$

QR decomposition : Again, we set :

$$A + \delta_{qr} A = QR$$

At step i, in order to obtain the Householder vector v, setting $x = A_{i:n,i}$, we first compute $\beta = x_1 + \mathrm{sgn}(x_1)\|x\|_2$. This is done with relative error bounded by

$\epsilon_m(\frac{n-i+1}{2}+1)$. Then we do $v_1 = x_1$ and for $j = 2, \ldots, n-i+1$, we compute $v_j = x_j/\beta$. We obtain v with relative error (in each coordinate) bounded by $\epsilon_m(\frac{n-i+1}{2}+2)$. This also bounds the error $\left\|H_v - H_{\mathrm{fl}(v)}\right\|_2$.

Applying $H_{\mathrm{fl}(v)}$ introduces further error, bounded by $\epsilon_m(2(n-i+1)+3)$. It follows that per step, we introduce error :

$$\|\delta H\|_2 \leq \epsilon_m \left(\frac{5n-5i+10}{2}\right)$$

It follows that at the end, we obtain Q and R such that :

$$(Q+\delta Q)R = A$$

with error :

$$\|\delta Q\|_2 \leq \epsilon_m(2.25n^2 + 5.25n)$$

We easily see that $\delta_{qr}A = -\delta Q\ R$. We know that $\|R\|_2 = \|A\|_2$, so we obtain :

$$\|\delta_{qr}A\|_2 \leq \|A\|_2 \epsilon_m(2.25n^2+5.25n) \leq n\max(|A_{ij}|)\epsilon_m(2.25n^2+5.25n)$$

7. Backward error analysis of Phase 2

The Newton operator in affine space produces matrices L and U such that $A - LU = \delta_g A$, where $\delta_g A$ verifies equation (24). Because of rounding-off in lines 5 and 6, the algorithm computes $\mathrm{fl}(M(A,b)) = (U+\delta U)^{-1}(L+\delta L)^{-1}b$, so we write :

$$(A+\delta_2 A)^{-1} = (U+\delta U)^{-1}(L+\delta L)^{-1}$$

Or :

$$\begin{aligned} A+\delta_2 A &= (L+\delta L)(U+\delta U) \\ &= A + \delta_g A + \delta L U + L\delta U + \delta L \delta U \end{aligned}$$

Subtracting A from both sides, and passing to the matrix of absolute values, we obtain :

$$|\delta_2 A| \leq |\delta_g A| + |\delta L||U| + |L||\delta U| + |\delta L||\delta U|$$

Inserting formulas (21) , (23) and (24), we obtain :

$$\|\delta_2 A\|_{\max} = \max(|A_{ij}|) \leq \epsilon_m(5n^2(g_{cp}(n)+\epsilon_m(n)))(\|A\|_{\max} + \|\delta_2 A\|_{\max}) \tag{25}$$

For the Newton operator in projective space, we obtain :

$$\|\delta_2 A\|_{\max} \leq \epsilon_m(5(n+1)^2(g_{cp}(n+1)+\epsilon_m(n+1)))(\|A\|_{\max} + \|\delta_2 A\|_{\max}) \tag{26}$$

For the pseudo-Newton method, we write :

$$\delta_2 A_{ij} = \delta_{qr}A_{ij} + \delta R\ Q^t + R\ \delta Q^t + \delta R\ \delta Q^t$$

$$\|\delta_2 A_{ij}\|_2 \leq \|\delta_{qr} A_{ij}\|_2 + \|\delta R\|_2 \left\|Q^t\right\|_2 + \|R\|_2 \left\|\delta Q^t\right\|_2 + \|\delta R\|_2 \left\|\delta Q^t\right\|_2$$

Hence,

$$\begin{aligned} \|\delta_2 A\|_{\max} &\leq \|\delta_2 A\|_2 \\ &\leq \epsilon_m (3.25(n+1)^2 + 10.25(n+1) + \epsilon_m((n+1)^3 + 4(n+1)^2)) \\ &(n+1)(\|A\|_{\max} + \|\delta_2 A\|_{\max}) \end{aligned} \tag{27}$$

8. Conditioning of M

Conditioning of matrix inversion : Let $A(t)$ be a non-singular matrix, depending on a parameter t. From the formula $A(t)^{-1}A(t) = I$ we obtain :

$$\frac{\partial}{\partial t}(A(t)^{-1}) = A(t)^{-1} \frac{\partial}{\partial t} A(t) A(t)^{-1} \tag{28}$$

Let us use that formula for A as depending of entry A_{ij} :

$$\frac{\partial}{\partial A_{ij}} A^{-1} = A^{-1} \left[0 \ e_i \ 0\right] A^{-1}$$

Recall that $M^{\text{aff}}(A,b) = A^{-1}b$, so we write :

$$\frac{\partial}{\partial A_{ij}} M(A,b) = \left[0 A^{-1} e_i 0\right] A^{-1} b$$

We now consider the operator $\frac{\partial}{\partial A} M(A,b)$. This operator associates to each $n \times n$ matrix X the vector $Y A^{-1} b$, where $Y_{ij} = (A^{-1})_{ij} X_{ij}$.

Let $\|.\|_{\max,2}$ be defined as the norm of operators from the vector space of $n \times n$ matrices endowed with the norm $\|.\|_{\max}$ into the space of n vectors endowed with the 2-norm $\|.\|_2$. From the inequality : $\|Y\|_2 \leq \|X\|_{\max} \left\|A^{-1}\right\|_2$, we deduce the formula :

$$\left\|\frac{\partial}{\partial A} M(A,b)\right\|_{\max,2} \leq \left\|A^{-1}\right\|_2 \left\|A^{-1}b\right\|_2 \tag{29}$$

Also, considering $\frac{\partial}{\partial b} M(A,b)$ as a linear operator and using the matrix 2-norm, we have :

$$\left\|\frac{\partial}{\partial b} M(A,b)\right\|_2 \leq \left\|A^{-1}\right\|_2$$

Conditioning for Newton in projective space : Recall the definition of M^{proj} :

$$M^{\text{proj}}(x,A,b) = \begin{bmatrix} A \\ x^* \end{bmatrix}^{-1} \begin{bmatrix} b \\ 0 \end{bmatrix}$$

As in the case of Newton in affine space,

$$\frac{\partial}{\partial A_{ij}} M^{\text{proj}}(x, A, b) = \left[0 \begin{bmatrix} A \\ x^* \end{bmatrix}^{-1} e_i 0 \right] \begin{bmatrix} A \\ x^* \end{bmatrix}^{-1} b$$

Hence :

$$\left\| \frac{\partial}{\partial A} M^{\text{proj}}(x, A, b) \right\|_{\max,2} \leq \left\| \begin{bmatrix} A \\ x^* \end{bmatrix}^{-1} \right\|_2 \left\| \begin{bmatrix} A \\ x^* \end{bmatrix}^{-1} b \right\|_2$$

The derivative on x is the same as the derivative in A. As in the affine case, also,

$$\left\| \frac{\partial}{\partial b} M(x, A, b) \right\|_2 \leq \left\| \begin{bmatrix} A \\ x^* \end{bmatrix}^{-1} \right\|_2$$

The pseudo-Newton case : Now, $M^{\text{pseu}}(A, b) = A^\dagger b$. It is easy to check that :

$$\left\| \frac{\partial}{\partial b} N^{\text{pseu}}(A, b) \right\|_2 = \|A^\dagger\|_2$$

For the derivative in A_{ij}, we parametrize :

$$A(t) = A + e_i {e_j}^t t$$

Let $V(t)$ be the space $\ker A(t)^\perp$. Then we have :

$$\|A(t)^\dagger\|_2 = \left\| A(t)_{|V(t)}{}^{-1} \right\|_2 \leq \left\| A(t)_{|V(0)}{}^{-1} \right\|_2 = \|A(0)^\dagger\|_2$$

by construction of the pseudo-inverse. Thus :

$$\left\| \frac{\partial}{\partial A} M^{\text{pseu}}(A, b) \right\|_{\max,2} \leq \|A^\dagger\|_2 \|A^\dagger b\|_2$$

What does this mean : In a first order setting, we can assume that $Df(x) - A = \delta A$ with $\|\delta A\|_{\max} \in O(\epsilon_m)$, so that it makes sense to discard the terms in ${\epsilon_m}^2$, and write :

$$\begin{aligned} \left\| \text{fl}(M^{\text{aff}}) - M^{\text{aff}} \right\|_2 < & \left\| Df(x)^{-1} \right\|_2 \left\| Df(x)^{-1} f(x) \right\|_2 \|\delta A\|_{\max} \\ & + \left\| Df(x)^{-1} \right\|_2 \|\delta b\|_2 + O({\epsilon_m}^2) \end{aligned}$$

Inserting inequalities $\left\| Df(x)^{-1} \right\|_2 \leq \frac{2\mu(f,x)}{\|f\|_{\text{k}}}$ and $\left\| Df(X)^{-1} f(x) \right\|_2 \leq \beta(f, x)$, we get the formula :

$$\left\| \text{fl}(M^{\text{aff}}) - M^{\text{aff}} \right\|_2 < \frac{2\mu(f, x)}{\|f\|_{\text{k}}} \left(\beta(f, x) \|\delta A\|_{\max} + \|\delta b\|_2 \right) \tag{30}$$

In the same way, we obtain formulas :

(31)

$$\left\|\mathrm{fl}(M^{\mathrm{proj}}) - M^{\mathrm{proj}}\right\|_2 < \frac{2\mu(f,x)}{\|f\|_{\mathrm{k}}}\left(\beta(f,x)\|\delta A\|_{\max} + \beta(f,x)\|\delta x\|_{\max} + \|\delta b\|_2\right)$$

$$\|\mathrm{fl}(M^{\mathrm{pseu}}) - M^{\mathrm{pseu}}\|_2 < \frac{2\mu(f,x)}{\|f\|_{\mathrm{k}}}\left(\beta(f,x)\|\delta A\|_{\infty} + \|\delta b\|_2\right) \tag{32}$$

It is possible to make the formulas above more rigorous, by bounding $\left\|A^{-1}\right\|_2$ and $\left\|A^{-1}b\right\|_2$ for A, b in neighborhoods of $Df(x)_{|V(x)}{}^{-1}$ and $f(x)$, respectively. This will be done to establish the polynomial time bound for the algorithm.

9. First order analysis

At this time, we can compute the error in $N(f,x)$, up to $O(\epsilon_m{}^2)$ and prove Theorem 11. Recall we defined the function $p(n)$ in three different cases :

$$p^{\mathrm{aff}}(n) = 5n^2(g_{cp}(n)+1)$$
$$p^{\mathrm{proj}}(n) = 5(n+1)^2(g_{cp}(n+1)+1)$$
$$p^{\mathrm{pseu}}(n) = (3.25(n+1)^3 + 11(n+1)^2)$$

We summarize the results obtained in equations (17), (16), (25), (26) , (27) , (30), (31) and (32) in the following table. Depending on the algorithm, $p = p^{\mathrm{aff}}$, p^{proj} or p^{pseu}. In the projective case, x' is considered as part of matrix A :

First order analysis

Quantity	Bound
$\|\delta_1 A\|_{\max}$	$(D+1+\max S(f_i))\sqrt{D}\|f\|_{\mathrm{k}}\epsilon_m$
$\|\delta_2 A\|_{\max}$	$p(n)\sqrt{D}\|f\|_{\mathrm{k}}\epsilon_m$
$\left\|\frac{\partial N}{\partial A}\right\|_{\max,2}$	$\frac{2\mu(f,x)}{\|f\|_{\mathrm{k}}}$
$\|\delta_1 b\|_2$	$(D+\max S(f_i))\|f\|_{\mathrm{k}}\epsilon_m$
$\left\|\frac{\partial N}{\partial b}\right\|_2$	$\frac{2\mu(f,x)\beta(f,x)}{\|f\|_{\mathrm{k}}}$
Step 7	$(1+\beta(f,x))\epsilon_m$
Total	$(2\mu(f,x)\sqrt{D}(p(n)+D+1+\max(S(f_i)))$ $+\beta(f,x)(D+\max(S(f_i)))) + 1 + \beta(f,x))\epsilon_m$

In any of the three cases, taking into account that $\beta(f,x) < 1$ (Indeed, β is typically much smaller, near an approximate zero), and dividing by $\|x\|_2 \geq \frac{1}{2}$, we obtain the bound :

$$\frac{\|N'(f,x)-N(f,x)\|}{\|x\|} \leq 4\mu(f,x)\sqrt{D}\,(p(n)+2D+2+2\max(S(f_i)))\,\epsilon_m + O({\epsilon_m}^2)$$

This proves Theorem 11 .

10. Construction of the finite precision machine

In this section, we show how to construct a Blum-Shub-Smale machine over $\mathbb{Z}$ simulating a complex ϵ-machine, and therefore performing correct arithmetics with the $1+\epsilon$ property. Let H be an integer not less than $2/\epsilon_m$, where ϵ_m is fixed. We will prove that

LEMMA 17. *An ϵ-machine can be simulated by a machine over $\mathbb{Z}$, such that sum, subtraction, multiplication, division, and square rooting can be performed in time $O(-\log(\epsilon)(\log(-\log\epsilon))^2)$, provided no overflow or underflow occur.*

In particular, if $c = \max(-\log(\epsilon), r+e_0)$ where r is the number of floating point operations and e_0 is the maximum number of exponent bits in the input and in the constants in the machine, then no overflow or underflow occur and complex sum, subtraction, multiplication, division and square rooting can be performed in time $O(c(\log c)^2)$

Proof : The second part of the Lemma follows from the fact that if we start with e_0 bits of exponent, we can only multiply the exponent by 2 at each floating point operation, so the final exponent will be at most $2^{e_0+r} \leq 2^c$, so that c bits will accommodate all the possible values of the exponent during the computation.

Floating point numbers can be represented by two integers : the mantissa and the exponent. Computation of sum, subtraction, multiplication and division are easy. We will show how to compute square roots within the time bound $O(-\log(\epsilon)(\log(-\log\epsilon))^2)$.

Let z be a real floating-point number, and assume without loss of generality that $1/2 < z \leq 2$. We want to find a root of :

$$f(x) = x^2 - z$$

We compute :

$$f'(x) = 2x$$

$$f''(x) = 2$$

We will use the affine Newton algorithm in one variable. The invariants are :

$$\beta(x) = \frac{x^2 - z}{2x}$$
$$\gamma(x) = \frac{1}{2x}$$
$$\alpha(x) = \frac{x^2 - z}{4x^2}$$

Let $x_0 = \frac{z+1}{2}$, then :

$$\alpha(x_0) = \frac{1}{4}\frac{(z-1)^2}{(z+1)^2}$$

It is easy to verify that the function $\alpha(x_0)$ in function of z is bounded above by $\frac{1}{36}$ when z is in $[1/2, 2]$. Also, $\gamma(z) \leq 4$.

Assume we have a machine with precision $\epsilon_m \leq \epsilon/48$, $\epsilon_m < 5768$.

The operator $N^{\text{aff}} : x \mapsto x - \frac{f(x)}{f'(x)} = \frac{x + \frac{z}{x}}{2}$ can be computed in two floating point operations.

$$|\text{fl}(\frac{z}{x}) - (\frac{z}{x})| < |x|\epsilon_m|\frac{z}{x^2}| \leq 2|x|\epsilon_m$$
$$\frac{|\text{fl}(N^{\text{aff}}(x)) - N^{\text{aff}}(x)|}{|x|} < (1 + 2\epsilon_m)(1 + \epsilon_m) - 1 < 4\epsilon_m$$

Hence

$$\frac{|\text{fl}(N^{\text{aff}}(x)) - N^{\text{aff}}(x)|}{\sqrt{1 + x^2}} < 4\epsilon_m$$

Thus, we can compute the operator $N^{\text{aff}} : x \mapsto x - \frac{f(x)}{f'(x)}$ in such a way that :

$$\frac{\left\|\text{fl}(N^{\text{aff}}(x)) - N^{\text{aff}}(x)\right\|_2}{\|x\|_2} \leq \delta$$

with $\delta \leq \epsilon/12$, with $\delta < 1442$. We are in the conditions of Theorem 2, Chapter 2 :

$$\alpha(x_0) + \delta\gamma(x_0) < 1/16, \epsilon\gamma(x_0) < 1/384$$

According to that Theorem, we are able obtain an approximation of precision $6\delta < \epsilon/2$ in $\log_2 - \log_2 \delta$ iterations, each of these consisting of 2 floating point operations with precision $\epsilon/48$.

This does not give us always the correctly rounded result. This algorithm may give us a value x at distance $\epsilon/2$ of two representable numbers x_1 and x_2. Those numbers are the candidates to be the correctly rounded result. In order to choose the right one, we still have to compute $\text{sgn}(x^2 - z)$. But this introduces a very small additional cost.

Hence, the cost of square-rooting is still $O(\log H(\log\log H)^2)$, and Lemma 17 is proved.

11. Polynomial time analysis

We are proving Theorem 12 for the pseudo-Newton method only. A very similar proof could be given for the other two methods, by using QR factorization instead of LU factorization.

We will first give rigorous bounds to $\|A^{-1}\|$ and $\|A^{-1}b\|$. This bounds will be provided by the Lemma :

Lemma 18. *Assume that :*

$$\|\delta A\|_2 < \frac{\|f\|_{\mathrm{k}}}{4\mu(f,x)}$$
$$\|\delta b\|_2 < \frac{\|f\|_{\mathrm{k}}}{2}$$

Then :

$$\left\|(Df(x)_{|V(x)} + \delta A)^{-1}\right\|_2 < \frac{4\mu(f,x)}{\|f\|_{\mathrm{k}}}$$
$$\left\|(Df(x)_{|V(x)} + \delta A)^{-1}(f(x) + \delta b)\right\|_2 < \beta(f,x) + 4\mu(f,x)$$

Let $A = Df(x)_{|V(x)} + \delta A$, and $b = f(x) + \delta b$.

Under the hypotheses of the Lemma, we have :

$$\|\delta A\|_2 < \frac{1}{2\left\|Df(x)_{|V(x)}{}^{-1}\right\|_2}$$

Hence

$$\left\|Df(x)_{|V(x)}{}^{-1}A - I\right\|_2 < \frac{1}{2}$$

Inverting, we get :

$$\left\|A^{-1}Df(x)_{|V(x)} - I\right\|_2 < \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots < 1$$

Multiplying by $Df(x)_{|V(x)}{}^{-1}$:

$$\left\|\delta A^{-1}\right\|_2 = \left\|A^{-1} - Df(x)_{|V(x)}{}^{-1}\right\|_2 < \left\|Df(x)_{|V(x)}{}^{-1}\right\|_2$$

It follows that :

$$\left\|A^{-1}\right\|_2 < 2\left\|Df(x)_{|V(x)}{}^{-1}\right\|_2 < \frac{4\mu(f,x)}{\|f\|_{\mathrm{k}}}$$

For $\left\|A^{-1}b\right\|_2$, we observe that we have :

$$\|\delta b\|_2 < \frac{\|f\|_{\mathrm{k}}}{2}$$

Now we write :

$$\begin{aligned}\left\|A^{-1}b\right\|_2 \leq & \left\|{Df(x)_{|V(x)}}^{-1}f(x)\right\|_2 + \left\|{Df(x)_{|V(x)}}^{-1}\right\|_2\|\delta b\|_2 \\ & + \|\delta A\|_2\|f(x)\|_2 + \|\delta A\|_2\|\delta b\|_2 \\ & \leq \beta(f,x) + \mu(f,x) + 2\mu(f,x) + \mu(f,x) \\ & \leq \beta(f,x) + 4\mu(f,x)\end{aligned}$$

This proves lemma 18

It follows from Lemma 18 that for A in a neighborhood of radius δA of $Df(x)_{|V(x)}$ and for b in a neighborhood of radius δb of $f(x)$, such that $\|\delta A\|_2 \leq \|f\|_{\mathrm{k}}/4\mu(f,x)$ and $\|\delta b\|_2 \leq \|f\|_{\mathrm{k}}/2$, the following bounds hold :

$$\begin{aligned}\left\|\frac{\partial}{\partial A}M^{\mathrm{pseu}}(A,b)\right\|_{\max,2} & \leq \left\|A^\dagger\right\|_2\left\|A^\dagger b\right\|_2 \\ & \leq \frac{4\mu(f,x)}{\|f\|_{\mathrm{k}}}(\beta(f,x)+4\mu(f,x)) \\ \left\|\frac{\partial}{\partial b}M^{\mathrm{pseu}}(A,b)\right\|_2 & \leq \left\|A^\dagger\right\|_2 \leq \frac{4\mu(f,x)}{\|f\|_{\mathrm{k}}}\end{aligned}$$

We can use the mean value inequality and, for any A, b in the same neighborhood, we obtain :

$$\begin{aligned}&\left\|M^{\mathrm{pseu}}(A,b) - Df(x)^\dagger f(x)\right\|_2 \leq \\ &\leq \frac{4\mu(f,x)}{\|f\|_{\mathrm{k}}}(\beta(f,x)+4\mu(f,x))\sqrt{n+1}\|\delta A\|_{\max} + \frac{4\mu(f,x)}{\|f\|_{\mathrm{k}}}\|\delta b\|_2\end{aligned} \tag{33}$$

Now, assume the hypotheses of theorem 12. We choose ϵ_m so that the following conditions are true :

$$16\bar{\mu}(1+4\bar{\mu})\sqrt{D}\epsilon_m(2D+2+2\max S(f_i)+p(n)) \quad < \delta \tag{34}$$

$$8\bar{\mu}\sqrt{D}\epsilon_m(2D+1+2\max S(f_i)+p(n)) \quad < 1 \tag{35}$$

$$\epsilon_m(p(n)) \quad < 1/2 \tag{36}$$

$$\epsilon_m H(f) \quad < 1/2 \tag{37}$$

$$\epsilon_m H(x) \quad < 1/2 \tag{38}$$

It is easy to see that condition (34) implies conditions (35) and (36). By using Lemma 16, we choose :

$$\epsilon_m = \frac{1}{\frac{32}{\delta}\bar{\mu}(1+4\bar{\mu})\sqrt{D}(2D+1+2\max S(f_i)+p(n))+2H(f)+2H(x)} \tag{39}$$

This choice of ϵ_m limits the computing time for each elementary floating point operation to a polynomial in $-\log\delta$, $\log\bar{\mu}$, $\log D$, $\log n$, $\log\max S(f_i)$, $\log H(f)$, $\log H(x)$, and logarithms of the above. The total number of floating point operations is already bounded by a polynomial in n, D, and $S(f)$. Lemma 17 gives a worst case time bound that is polynomial in n, D, $\log H(f)$, $\log H(x)$, $S(f)$, $\log\bar{\mu}$ and $-\log\delta$, as stated in the theorem.

It remains to prove that the result obtained is within error δ of the correct pseudo-Newton iteration.

Equation (27), together with condition (36) becomes :

$$\|\delta_2 A\|_{\max} \le \epsilon_m(p^{\text{pseu}}(n))(\|A\|_{\max}+\|\delta A\|_{\max})$$

Adding equation (17), and knowing that $\|A\|_{\max} \le \|f\|_{\text{k}}\sqrt{D}$, we obtain :

$$\|\delta A\|_{\max} \le \|f\|_{\text{k}}\sqrt{D}\epsilon_m(D+1+\max S(f_i)) + \epsilon_m(p^{\text{pseu}}(n))(\|A\|_{\max}+\|\delta A\|_{\max})$$

Rearranging,

$$\|\delta A\|_{\max} \le \frac{\|f\|_{\text{k}}\sqrt{D}\epsilon_m(D+1+\max S(f_i)+p^{\text{pseu}}(n))}{1-\epsilon_m(p(n))}$$

Introducing again condition (36), we get :

$$\|\delta A\|_{\max} \le 2\|f\|_{\text{k}}\sqrt{D}\epsilon_m(D+1+\max S(f_i)+p^{\text{pseu}}(n)) \le \frac{\|f\|_{\text{k}}}{4\bar{\mu}}$$

The last inequality following from condition (35). For $\|\delta b\|_2$, we had the bound (16) :

$$\|\delta b\|_2 \le \|f\|_{\text{k}}\epsilon_m(D+\max S(f_i)) < \|f\|_{\text{k}}/2$$

Hence, we can use bound (33). The total error is less than :

$$8\bar{\mu}(\beta+4\bar{\mu})\sqrt{D}\epsilon_m(D+1+\max S(f_i)+p^{\text{pseu}}(n))+$$

$$+4\bar{\mu}\epsilon_m(D+\max S(f_i))+(1+\beta)\epsilon_m$$

Since $\beta \le 1$, the total error of the algorithm is less than :

$$8\bar{\mu}(1+4\bar{\mu})\sqrt{D}\epsilon_m(2D+2+\max S(f_i)+p^{\text{pseu}}(n))$$

According to condition (34), this is less than $\delta/2$, so that the error over $\|x\|_2$ is less than δ, qed.

CHAPTER 4

Gap theory and estimate of the condition number

> In this chapter, gap theorems are introduced. The condition number of systems and paths is estimated in terms of heights. Newton iteration is used to decide if a root is rational. A result on root separation is also stated.

1. Introduction

Let $\mathcal{H}_d$ be the space of all systems of n homogeneous polynomial equations of degree $d = (d_1, \ldots, d_n)$ in $n+1$ variables $(x_0, \ldots, x_n)$, with complex coefficients. Let Σ be the locus of all systems in $\mathcal{H}_d$ such that there is a solution $x \neq 0$, $f(x) = 0$ such that $Df(x)$ does not have full rank. Σ is an algebraic variety, called the *discriminant variety*. Systems belonging to Σ are called *degenerate*.

Let $f = (f_1, \ldots, f_n) \in \mathcal{H}_d$ be non-degenerate, $n \geq 2$, $D = \max d_i \geq 2$. If we require f to have only integer (or rational, or gaussian integer, or gaussian rational) coefficients, then we will be able to obtain a certain number of conclusions about the roots of the system.

In this chapter, we are interested about two kind of bounds : Bounds on the minimal distance between two different roots, and bounds on some *height* of rational roots. By *height*, we understand a measure of the number of bits necessary to represent a point in projective space :

If $a \in \mathbb{Z}$, we define the height $H(a) = |a|$. For Gaussian integers, we define : $H(a + bi) = H(a) + H(b)$. Then we have $H(x \pm y) \leq H(x) + H(y)$ and $H(xy) \leq H(x)H(y)$. This is not the standard definition of height.

Let z be a point in $\mathbb{C}^{n+1}$, considered as a point in projective space. We say that z is rational if there is $\lambda \in \mathbb{C}$ such that $\lambda z \in \mathbb{Q}^{n+1}$. By multiplying by the smallest common multiple, we can assume without loss of generality that z has integer coordinates, without common factor. Then we set the height $H(z) = \max |z_j|$.

If f_j is a polynomial with rational coefficients, we define $H(f_j)$ as the height of the vector of its coefficients. The height of the system f is the maximum of $H(f_j)$.

This definition of height is not adequate to all the situations we will have to face, and we will see definitions more suitable to several particular cases.

In Shub and Smale [**12**], and in chapter 2, the complexity of path-following was bounded in terms of a condition number μ. It was proved by Shub and Smale in [**12**] that the condition number

$$\mu(f) = \max_{f(\zeta)=0} \|f\|_{\mathrm{k}} \left\| Df(x)^{-1} \mathrm{diag}({\|x\|_2}^{d_i-1}\sqrt{d_i}) \right\|_2$$

verifies $\mu(f) = 1/\rho(f)$, where $\rho(f)$ is the Maximum distance in Kostlan's metric between f and the discriminant variety Σ, along a fiber $\{f : f(\zeta_i) = 0\}$. This implies the inequality :

$$\mu(f) \leq \frac{1}{d_{\mathrm{proj}}(f, \Sigma)}$$

A consequence of that is the existence of a finite *complexity exponent* $d(\Sigma)$ and of a *geometric condition number* $\mu(\Sigma)$ depending solely on d and n, such that :

THEOREM 13. *Let $f \in \mathcal{H}_d$ have gaussian integer coefficients. Then either $f \in \Sigma$, either*

$$\mu(f) \leq \mu(\Sigma) H(f)^{d(\Sigma)}$$

where we can set :

$$d(\Sigma) = n \prod d_j \sum d_j \geq \sum r_i$$
$$\mu(\Sigma) = \sqrt{D!} d(\Sigma) \left(3n!n(d(\Sigma) + \max S(f_i))(2 \sum d_j)^n \prod (d_j - 1) \right)^{d(\Sigma)}$$

It is immediate from Theorem 13 that :

COROLLARY 2. *There is a universal polynomial M_μ such that if $\mu(f)$ is finite, then :*

$$\mu(f) < 2^{M_\mu(n, D, \prod d_i, \log H(f), \log \dim \mathcal{H}_d)}$$

Moreover, it is possible to obtain a result about conditioning of a whole path in projective space. Indeed, it will be necessary to extend the path to a *real* line, embedded in projective space. Generically, this line does not cut the discriminant variety. If we denote by $\mu([f_0, f_1])$ the condition number of the path $[f_0, f_1]$, and write $f_t = (1-t)f_0 + tf_1$, then we will prove :

THEOREM 14. *There is a proper variety Σ_0 in $\mathcal{H}_d \times \mathcal{H}_d$ considered as a real projective space, so that for $f_t \not\in \Sigma_0$ the following estimate is true :*

$$\mu(f_t) \leq \mu(\Sigma_0) H(f_t)^{d(\Sigma_0)}$$

where we can set :

$$\begin{aligned} d(\Sigma_0) &= 2n \prod (d_j+1)^2 (1 + \sum d_j)^2 \\ \mu(\Sigma_0) &= \sqrt{D!} d(\Sigma_0) \\ &\quad \times \left(32^D n! n (d(\Sigma_0) + \max S(f_i) - 1) \prod (d_j - 1)(-1 + 4 \sum d_j)^{2n+1} \right)^{d(\Sigma_0)} \end{aligned}$$

Theorems 13 and 14 are obtained by estimating the height and degree of polynomials defining Σ and Σ_0.

Besides the estimate on the number of approximate Newton steps defined in chapter 2, several other useful results can be obtained in terms of μ or $\log \mu$.

We will need the notations and conclusions of chapters 2 and 3. Since we will be using mainly the pseudo-Newton operator, we recall some definitions :

$$\begin{aligned} \beta(f,x) &= \frac{1}{\|x\|_2} \left\| Df(x)^\dagger f(x) \right\|_2 \\ \gamma(f,x) &= \|x\|_2 \max_k \left\{ 1, \left(\frac{\left\| Df(x)^\dagger D^k f(x) \right\|_2}{k!} \right)^{\frac{1}{k-1}} \right\} \\ \alpha(f,x) &= \beta(f,x)\gamma(f,x) \end{aligned}$$

THEOREM 15. *There is a machine over $\mathbb{Z}$ such that, given input ($f \in \mathcal{H}_d$, $x_0 \in \mathbb{C}^{n+1}$), such that f is non-degenerate, $\alpha(f,x_0) \leq 1/16$, and ζ is the zero associate to x_0, returns ζ if ζ is rational, and fails otherwise. Moreover, the execution time of that program is bounded by a polynomial in $\prod d_i$, n, D, $\log H(f)$, and* $\dim \mathcal{H}_d$.

The condition on α is the condition of the Theorem 6, Chapter 2. Together with the construction of the approximate Newton operator (Theorem 12, Chapter 3), theorem 6, Chapter 2 can be used to generate a sequence x_i quadratically convergent to the nearest zero ζ :

$$d(x_i, \zeta) \leq \max\{2^{-2^i - 1}, 6\epsilon\}$$

where ϵ is the *error* of the approximate Newton operator. The cost of obtaining x_i was found to be a polynomial in $\log H(f)$, $\dim \mathcal{H}_d$, $\log \mu(f,\zeta)$ and $-\log \epsilon$. We will use this analysis to prove the theorem.

Also, the following result on separation of roots comes naturally in terms of γ, hence of μ :

THEOREM 16. *Let $f \in \mathcal{H}_d$, and assume that x_1, x_2 are different roots of f. Then $d_{\text{proj}}(x_1, x_2) \geq \sin \tan^{-1} 1/2\gamma(f, x_1)$*

It follows that $d_{\text{proj}}(x_1, x_2) \geq \frac{1}{\sqrt{2} D^{3/2} \mu(f)} \geq \frac{1}{\sqrt{2} D^{3/2} \mu(\Sigma) H(f)^{d(\Sigma)}}$.

2. The Macaulay resultant

THEOREM 17 (Macaulay [**7**]). *Let $\mathcal{H}$ be the space of all the systems $f = (f_0, \dots, f_n)$ of homogeneous polynomials of degree $d = (d_0, \dots, d_n)$ in variables $x = (x_0, \dots, x_n)$. Let Σ be the locus in $\mathcal{H}$ of systems f having a non-trivial solution x, $x \neq 0$, $f(x) = 0$. Then :*

(1) *The locus Σ is an algebraic variety, the zero set of a polynomial R of degree $\sum_i \prod_{j \neq i} d_i$ in the coefficients of f, given by :*

$$R(f) = \gcd \det A_j$$

where A_j are matrices to be defined, of size $\left(\begin{array}{c} \sum d_j \\ n \end{array} \right)$, and with entries either coefficients of f, either zero.

(2) *There is a submatrix B_j of A_j, such that :*

$$R(f) = \frac{\det A_j}{\det B_j}$$

(3) *Poisson formula : If we fix some $f_1, \dots, f_n$ non-singular, then R is given, as a polynomial in the coefficients of f_0 and up to a multiplicative constant, by :*

$$R(f_0) = \prod_{f_1(\zeta)=0, \dots, f_n(\zeta)=0} f_0(\zeta)$$

The polynomial R is called the *resultant* of the (undetermined) system f. It follows from Bezout theorem that the resultant has degree $d_1 \dots d_n$ in the coefficients of f_0.

An important particular case is when f_0 is linear, of the form : $f_0(x) = ux$. In that case, we write :

$$R_u(f_1, \dots, f_n) = R(f_0, \dots, f_n)$$

When $f_1, \dots f_n$ are fixed, R_u is called the u-resultant of $f_1, \dots, f_n$ and is a polynomial in $u_0, \dots, u_n$.

In this section, we follow some of the ideas in Macaulay's paper [**7**] in order to construct matrices A_j and to sketch a proof of Theorem 17, but with a different notation.

Idea of the Proof of Theorem 17 : For reasons that will become apparent later, we set $\mathcal{D} = (\sum d_j) - n$

Consider the condition :

$$\exists x \neq 0 : f(x) = 0 \tag{40}$$

It is equivalent to :

$$\exists x \neq 0 : \forall g = (g_0, \dots, g_n),\ \deg g_j = \mathcal{D} - d_j,\ \sum f_j(x) g_j(x) = 0 \tag{41}$$

Proof: (40) $\Rightarrow$ (41) is trivial. Assume x is given by (41). There is i with $x_i \neq 0$. Set $g_j = \lambda_j {x_i}^{\mathcal{D}-d_j}$. In other words :

$$\forall \lambda,\ \sum \lambda_j {x_i}^{\mathcal{D}-d_j} f_j(x) = 0$$

Changing to coordinates $\lambda'_j = \lambda_j {x_i}^{\mathcal{D}-d_j}$, we obtain :

$$\forall \lambda',\ \sum \lambda'_j f_j(x) = 0$$

It follows that $f_j(x) = 0$.

Now we introduce some notation. $M_d, d \in \mathbb{N}$ is the linear space generated by all monomials of degree d in $n+1$ variables. ρ_d is the d-uple embedding :

$$\begin{array}{rccc} \rho_d : & \mathbb{C}^{n+1} & \to & M_d \\ & x & \mapsto & (x^{\alpha_0}, \dots, x^{\alpha_N}) \end{array}$$

Where N is the dimension of M_d, and α_i ranges through all possible monomials of degree d.

$$N = \dim M_d = \begin{pmatrix} n+d \\ n \end{pmatrix}$$

There is a natural isomorphism between the dual of M_d and the linear space of homogeneous polynomials of degree d :

$${\rho_d}^* : \tilde{f} \mapsto f, f(x) = \tilde{f} \rho_d(x)$$

When $d = (d_0, \dots, d_n)$, those maps can be extended as follows :

$$\begin{array}{rccc} \rho_d : & \mathbb{C}^{n+1} & \to & M_{d_0} \times \cdots \times M_{d_n} \\ & x & \mapsto & \big(\rho_{d_0}(x), \dots, \rho_{d_{n}(x)}\big) \end{array}$$

$$\begin{array}{rrcl} {\rho_d}^* : & (M_{d_0} \times \cdots \times M_{d_n})^* & \to & \mathcal{H}_d \\ & \left(\tilde{f}_0, \ldots, \tilde{f}_n\right) & \mapsto & \left(\rho^*_{d_0}(f_0), \ldots, \rho^*_{d_n}(f_n)\right) \end{array}$$

The latest one is still an isomorphism of linear spaces.

Multiplication of polynomial g_0 of degree d'_0 by a fixed polynomial f_0 of degree d_0 corresponds to a linear mapping :

$$\begin{array}{rrcl} \hat{f}_0 : & {M_{d'_0}}^* & \to & {M_{d_0+d'_0}}^* \\ & \tilde{g}_0 & \mapsto & \tilde{g}_0 \hat{f}_0 \end{array}$$

Where

$$(\tilde{g}_0 \hat{f}_0)\rho_{d_0+d'_0}(x) = g_0(x) f_0(x)$$

We write $\tilde{g}_0 \hat{f}_0$ in that order because $\tilde{g}_0$ is a row vector (a covector).

Also, the sum of two polynomials of same degree $\mathcal{D}$ is associated to the sum in ${M_\mathcal{D}}^*$. So we can define the operator corresponding to :

$$g_0, \ldots, g_n \mapsto f_0 g_0 + \cdots + f_n g_n$$

whenever the $f_i g_i$ have the same degree ; in our case we will want degree $\mathcal{D}$ (we will see why shortly). So we get a linear operator $\hat{f}$:

$$\begin{array}{rrcl} \hat{f} : & {M_{\mathcal{D}-d_0}}^* \times \cdots \times {M_{\mathcal{D}-d_n}}^* & \to & {M_\mathcal{D}}^* \\ & \tilde{g}_0, \ldots, \tilde{g}_n & \mapsto & \tilde{g}_0 \hat{f}_0 + \cdots + \tilde{g}_n \hat{f}_n \\ & & & = (\tilde{g}_0, \ldots, \tilde{g}_n)\hat{f} \\ & & & = \tilde{g}\hat{f} \end{array}$$

Here, $\hat{f}$ can be seen as a matrix with more rows than columns. It defines a bilinear mapping

$$\tilde{g}, \tilde{x} \mapsto \tilde{g}\hat{f}\tilde{x}$$

Indeed, we are speaking about a tri-linear map, if we consider the coefficients of f as input.

We get the following conditions, clearly equivalent to conditions (40) and (41) :

$$\exists x : \forall \tilde{g}, \tilde{g}\hat{f}\rho_\mathcal{D}(x) = 0 \tag{42}$$

$$\exists x : \hat{f}\rho_\mathcal{D}(x) = 0 \tag{43}$$

There is a very convenient condition for (43) to be true.

We define the subspaces $M^{(i,j)}$ of ${M_{\mathcal{D}}}^*$ as follows : let let $L(i,j)$ be the set $\{j+1,\dots,i-1\}$ or eventually the set $\{j+1,\dots,n,0,\dots,i-1\}$. In particular, $L(i,0)=\{1,\dots,i-1\}$.

We define $M^{(i,j)}$ as the subspace of $M_{\mathcal{D}}$ of monomials of degree $< d_k$ in x_k for k in $L(i,j)$, and of degree $\geq d_i$ in x_i.

Because we set $\mathcal{D}=(\sum d_j)-n$, We can write :

$$ {M_{\mathcal{D}}}^* \simeq M^{(0,j)} \times \cdots \times M^{(n,j)} $$

There is a natural embedding of $M^{(i,j)}$ into ${M_{\mathcal{D}-d_I}}^*$ given by the division by ${x_i}^{d_i}$.

This extends to embeddings I_j of ${M_{\mathcal{D}}}^*$ into ${M_{\mathcal{D}-d_0}}^* \times \dots {M_{\mathcal{D}-d_n}}^*$:

$$\begin{tikzcd}
{M_{\mathcal{D}}}^* \arrow[r] \arrow[ddr, "I_j"'] & M^{(0,j)} \arrow[dd] & \times & \dots & \times & M^{(n,j)} \arrow[dd] \\
 & & & \dots & & \\
 & {M_{\mathcal{D}-d_0}}^* & \times & \dots & \times & {M_{\mathcal{D}-d_n}}^*
\end{tikzcd}$$

We denote by $A_j = I_j\hat{f}$ the linear operator $\tilde{g} \mapsto \tilde{g}I_j\hat{f}$ (notice again that $\tilde{g}$ is a row vector, so we write $\tilde{g}I_j$ for I_j applied to $\tilde{g}$). This definition gives us a square matrix A_j for every j. $\det A_j$ is a polynomial in the coefficients of f.

We can finally define the resultant of f as:

$$ R = \gcd(\det A_j) $$

where $\det A_j$ are considered as polynomials in variables $\tilde{f}$. The condition :

$$ R(\tilde{f}) = 0 \tag{44} $$

is equivalent to conditions (40) to (43). It follows from the construction that condition (43) implies condition (44).

We will use the Lemma :

LEMMA 19. *Let $f_1,\dots,f_n$ be non-degenerate. Then the numerator $R(\tilde{f})$ is a polynomial of degree $d_1d_2\dots d_n$ in the coefficients of f_0.*

Lemma 19 is proved in [**7**], pages 10 and 11. The idea of the proof is to show that what we call $\det A_j {\det A_0}^{-1}$ contains no coefficient of f_0 in the denominator. (We will not perform that computation here). It follows that the product of all the irreducible factors of $\det A_0$ containing coefficients of f_0 divides $R(f)$. Therefore, the degree of $R(f)$ in the coefficients of f_0 is not less than the degree of $\det A_0$

in those coefficients. But since $R(f)$ divides $\det A_0$, those degrees are equal ; by construction of A_0, they are equal to $\dim M^{(0,0)} = d_1 d_2 \dots d_n$.

Assuming Lemma 19, when we specialize the coefficients of $f_1, \dots, f_n$, generically, we have the following expression, up to a multiplicative constant :

$$R(f) = \prod f_0(\zeta_i)$$

where ζ_i ranges over all the roots of $f_1, \dots, f_n$. Indeed, it is clear that $f_0(\zeta_i)$ divides $R(f_0, f_1, \dots, f_n)$ as a polynomial in $\tilde{f}_0$ (ζ_i is fixed, so $f_0(\zeta_i)$ is a polynomial of degree 1 in $\tilde{f}_0$). Just note that $f_0(\zeta) = 0$ for ζ solution of $f_1, \dots, f_n$ implies that $R(f) = 0$. Hence according to Hilbert's Nullstellensatz, $\prod f_0(\zeta_i)$ divides $R(f)^k$, as polynomials in $\tilde{f}_0$, for some k. But since by hypothesis the roots ζ_i are disjoint, $\prod f_0(\zeta_i)$ cannot have a multiple factor, so it has to divide $R(f)$.

Moreover, by Bezout theorem, there are $d_1 d_2 \dots d_n$ different roots ζ_i. Using Lemma 19, the two polynomials have the same degree, so they are equal up to a multiplicative constant. This implies item (3) of Theorem 17.

Using Item 3, the condition $R(\tilde{f}) = 0$ implies that there is ζ such that $f(\zeta) = 0$.

If we add the degree in the coefficient of each f_i, we conclude that the total degree of each $R(f)$ is $\sum_i \prod_{j \neq i} d_j$. This will concludes the proof of Item 1 of Theorem 17.

We still have to sketch the proof of item 2 of Theorem 17.

Proof of Item (2) of Theorem 17 : Let $V^* = \cap_{j,k} \ker I_k - I_j \subset M^*_{\mathcal{D}}$. Let W^* be its orthogonal complement. Spaces V^* and W^* are subspaces of $M^*_{\mathcal{D}}$, and V^* is spanned by all monomials x^α with only one $\alpha_i \geq d_i$. Therefore, it has dimension $\sum_i \prod_{j \neq i} d_j$.

Let $A_k = I_k \hat{f}$ as above, and let B_k be the sub-matrix of A_k corresponding to rows and columns in W^* and W, respectively. We want to prove that :

$$R(f) = \frac{\det A_k}{\det B_k}$$

It is enough to show that $\frac{\det A_k}{\det B_k}$ is a polynomial. For in that case, $R(f)$ and $\frac{\det A_k}{\det B_k}$ have the same degree ; by definition of R, there is a polynomial $b(f)$ such that $R(f)b(f) = \det A_k$, and that polynomial is clearly of the same degree than $\det B_k$. Therefore, $b(f) = \det B_k$, up to a multiplicative constant. This implies Item 2.

We write :

$$A_k = \begin{bmatrix} M & P \\ N & B_k \end{bmatrix}$$

where $M \in L(V,V)$, $N \in L(W,V)$, $P \in L(V,W)$ and of course $B_k \in L(W,W)$. We also define :

$$\tilde{B}_k = \begin{bmatrix} I_V & 0 \\ 0 & B_k \end{bmatrix}$$

We consider the operator $p = A_k \tilde{B}_k^{-1}$ in $M_{\mathcal{D}}$. Let e_j be a basis (column) vector. Then either $e_j \in V$, either $e_j \in W$.

If $e_j \in V$, then $\tilde{B}_k^{-1} e_j = e_j$, and hence $A_k \tilde{B}_k^{-1} e_j$ is a vector whose coordinates are coefficients of f.

If $e_j \in W$, then $\tilde{B}_k^{-1} e_j \in W$, and :

$$A_k \tilde{B}_k^{-1} e_j = e_j + \begin{bmatrix} P \\ 0 \end{bmatrix} {B_k}^{-1} e_j$$

Therefore, p has the form :

$$p = \begin{bmatrix} M & P' \\ 0 & I_W \end{bmatrix}$$

The matrix M is non-singular, since the coefficients in ${x_i}^{d_i}$ appear in the main diagonal of M, and only there. Therefore, $\det M$ is a non-zero polynomial.

It follows that $\det p = \det M \det I_W = \det M$, and therefore $\det p$ is a polynomial in the coefficients of f. Moreover,

$$\det p = \frac{\det A_k}{\det \tilde{B}_k} = \frac{\det A_k}{\det B_k}$$

This concludes the sketch of the proof of Theorem 17.

3. Gap Theorems

Let $f = (f_1, \ldots, f_n) \in \mathcal{H}_d$ be non-degenerate, with integer coefficients of absolute value less than some integer $H(f)$. Let $D = \max d_j$. Let x be a solution of f. Assume $|x_i|$ and $|x_j|$ are non-zero. We will show that $\frac{|x_i|}{|x_j|}$ is bounded away from zero. The following is a version of Canny's gap theorem [**4**].

Theorem 18 (Canny). *In the conditions above,*

$$\frac{|x_i|}{|x_j|} \geq \left(3H(f) \binom{\sum d_j}{n} \right)^{-(n+1) \prod d_j}$$

This is also true when the coefficients of f are gaussian integers. We also want to be able to decide when x is a rational solution (i.e., all the $\frac{|x_i|}{|x_j|}$ are rational). Almost the same proof of Canny's theorem will lead to :

THEOREM 19. *Let $f(x) = 0$, and $|x_i|, x_j \in \mathbb{N}_*$ have no common factor. Then :*

$$|x_i|, x_j \leq \left(3H(f) \begin{pmatrix} \sum d_j \\ n \end{pmatrix}\right)^{(n+1)\prod d_j}$$

In order to prove the theorems, we will need some alternative definitions of Height. First of all, assume that H is defined in a ring R. The main examples are the integers and the Gaussian integers. Assume furthermore that :

$$\begin{aligned} H(1) = H(-1) = H(i) = H(-i) = 1 \\ H(a+b) \leq H(a) + H(b) \\ H(ab) \leq H(a)H(b) \end{aligned}$$

Let K be the field of fractions of R, and let $L = K[f_1, \dots, f_N]$ where the f_i are indeterminates (transcendental) over K. If a is an integer in L, a can be written in the form $a = \sum a_I f^I$, where $a_I \in R$ and I are multi-indices. We define a height B on the ring of integers of L :

$$B(\sum a_I f^I) = \sum H(a_I)$$

The following properties of B are obvious :

$$\begin{aligned} B(1) = B(-1) = B(i) = B(-i) = 1 \\ B(g+h) \leq B(g) + B(h) \\ B(gh) \leq B(g)B(h) \end{aligned}$$

We also want to extend this definition to integral polynomials in $L[t]$, but in a different way. We define :

$$C(t^d + p_{d-1}t^{d-1} + \cdots + p_0) = \max\left(B(p_i)^{\frac{1}{d-i}}\right)$$

The following facts were proved in [**4**] for the particular case $L = K = \mathbb{Q}$.

LEMMA 20. *Let p, q integral polynomials in L and let M be a $n \times n$ matrix with integral entries in L.*

$$C(pq) \leq C(p) + C(q)$$

$$C(p/q) \leq C(p) + 2C(q)$$

$$C(detM - tI) \leq n \max B(M_{ij})$$

Proof of lemma 20 :

Part 1 : We write :

$$p(t) = t^m + p_{m-1}t^{m-1} + \cdots + p_0$$

$$q(t) = t^n + q_{n-1}t^{n-1} + \cdots + q_0$$

$$r(t) = p(t)q(t) = t^{m+n} + r_{m+n-1}t^{m+n-1} + \cdots + t_0$$

where :

$$r_i = \sum_{0 \leq j \leq m, 0 \leq i-j \leq n} p_j q_{i-j}$$

By definition,

$$C(pq) = \max B\left(\sum p_j q_{i-j}\right)^{\frac{1}{n+m-i}}$$

$$C(pq) \leq \max \left(\sum B(p_j)B(q_{i-j})\right)^{\frac{1}{n+m-i}}$$

So there is i such that :

$$C(pq)^{n+m-i} \leq \sum C(p)^{m-j}C(q)^{n-i+j}$$

On the other hand,

$$(C(p) + C(q))^{n+m-i} = \sum_{i-n \leq j \leq m} \begin{pmatrix} n+m-i \\ m-j \end{pmatrix} C(p)^{m-j}C(q)^{n-i+j}$$

Comparing term by term,

$$C(pq)^{n+m-i} \leq (C(p) + C(q))^{n+m-i}$$

Hence

$$C(pq) \leq C(p) + C(q)$$

Part 2 :

$$\begin{aligned}
p(t) &= t^m + p_{m-1}t^{m-1} + \cdots + p_0\\
q(t) &= t^n + q_{n-1}t^{n-1} + \cdots + q_0\\
r(t) &= p(t)/q(t) = t^{m-n} + r_{m-n-1}t^{m-n-1} + \cdots + t_0
\end{aligned}$$

Since p and q are monic, the quotient $r(t)$ can be computed by the following recurrence :

$$\begin{aligned}
r_{m-n} &= 1\\
r_{m-n-j-1} &= p_{m-j-1} - \sum_{0\leq i\leq j} r_{m-n-i}q_{n+i-j-1}
\end{aligned}$$

We have :

$$\begin{aligned}
B(r_{m-n-j-1}) &\leq B(p_{m-j-1}) + \sum_{0\leq i\leq j} B(r_{m-n-i})B(q_{n+i-j-1})\\
B(r_{m-n-j-1}) &\leq C(p)^{j+1} + \sum_{0\leq i\leq j} B(r_{m-n-i})C(q)^{j+1-i}
\end{aligned}$$

We proceed by induction on i. Assume that $B(r_{m-n-j}) \leq (C(p)+2C(q))^j$ for all $j \leq i$. This is trivially true for $i = 0$. By induction,

$$\begin{aligned}
B(r_{m-n-j-1}) &\leq C(p)^{j+1} + \sum_{0\leq i\leq j} (C(p)+2C(q))^i C(q)^{j+1-i}\\
&\leq C(p)^{j+1} + C(q)\sum_{0\leq i\leq j} (C(p)+2C(q))^i C(q)^{j-i}\\
&\leq C(p)^{j+1} + C(q)\sum_{0\leq i\leq j} 2^{i-j}(C(p)+2C(q))^j\\
&\leq C(p)^{j+1} + 2C(q)(C(p)+2C(q))^j\\
&\leq C(p)(C(p)+2C(q))^j + 2C(q)(C(p)+2C(q))^j\\
&\leq (C(p)+2C(q))^{j+1}
\end{aligned}$$

Thus,

$$C(r) \leq C(p) + 2C(q)$$

Part 3 :

$$\det(M - tI) = t^n + a_{n-1}t^{n-1} + \cdots + a_0$$

Coefficient a_i is the sum of n^{n-i} products of $n-i$ entries of B-height less than $B(M)$, so we have : $|a_i| \leq (nB(M))^{n-i}$.

$$C(\det(M - tI)) \leq nB(M)$$

LEMMA 21. *Let $f = (f_0, \dots, f_n)$ be an indeterminate system of homogeneous equations of degree $d_0, \dots, d_n$ in $n+1$ variables, with integer (resp. Gaussian integer) coefficients. Let $\tilde{f}$ denote the coefficients of f, and let $L = \mathbb{Q}[\tilde{f}]$ (resp. $L = \mathbb{Q}[i][\tilde{f}]$).*

Then :

$$B(R(\tilde{f})) \leq \left(3 \begin{pmatrix} \sum d_j \\ n \end{pmatrix} \right)^{\deg R}$$

If $f_1, \dots, f_n$ are determinate, then :

$$B(R(\tilde{f_0})) \leq \left(3 \begin{pmatrix} \sum d_j \\ n \end{pmatrix} H(f_1, \dots, f_n) \right)^{\deg R}$$

As it was proved in Theorem 17, $\deg R = \sum_i \prod_{j \neq i} d_j$. Theorems 18 and 19 follow directly from Lemma 21 :

Set $\tilde{f_0} = ux_i + x_j$. Then $R(\tilde{f_0})$ is a polynomial in u, given by :

$$R(u) = r_{\prod d_i} u^{\prod d_i} + \dots + r_0$$

In particular, if x_i and x_j are rational, they divide $r_{\prod d_i}$ and r_0, respectively. This implies theorem 19.

In any case, $|u| = | - \frac{x_j}{x_i}|$ is either less than one, either

$$\begin{aligned} |u| &\leq \frac{|r_0| + \dots + |r_{\prod d_i - 1}|}{|r_{\prod d_i}|} \\ &\leq H(r_0) + \dots + H(r_{\prod d_i - 1}) \\ &\leq B(R(u)) \end{aligned}$$

Hence,

$$\frac{|x_i|}{|x_j|} \geq \left(3H(f) \begin{pmatrix} \sum d_j \\ n \end{pmatrix} \right)^{-\sum_i \prod_{j \neq i} d_j}$$

proving Theorem 18

Proof of Lemma 21 :

Consider the system g_v, depending on indeterminate v, defined by :

$$g_i(x) = f_i(x) - v{x_i}^{d_i}$$

Assume that some or all of the coefficients of f are determinate (the same proof works if they are all undeterminate, by setting $H(f) = 1$). As in the section before, set :

$$A_k = I_k \hat{g}$$

and let B_k be the sub-matrix of A_k from Theorem 17.

The variable v appears in the main diagonal of A_k and B_k. Using Part 3 of Lemma 20, we conclude that $\det A_k$ and $\det B_k$ verify, as polynomials in v :

$$C(\det A_k) \leq H(f) \begin{pmatrix} \sum d_j \\ n \end{pmatrix}$$

$$C(\det B_k) \leq H(f) \begin{pmatrix} \sum d_j \\ n \end{pmatrix}$$

Therefore, if we consider $R(f)$ as a polynomial in v, Part 2 of Lemma 20 implies :

$$C(R(f)(v)) \leq 3H(f) \begin{pmatrix} \sum d_j \\ n \end{pmatrix}$$

The degree in v of $R(f)$ is $\sum_i \prod_{j \neq i} d_j$. Therefore the resultant $R(f)(0)$ of f verifies :

$$B(R(f)(0)) \leq \left(3H(f) \begin{pmatrix} \sum d_j \\ n \end{pmatrix} \right)^{\deg R}$$

This proves Lemma 21.

4. Worst possible conditioning

Let $f = (f_1, \dots f_n) \in \mathcal{H}_d$. If $v \in \mathbb{C}^{n+1}$, $v \neq 0$, we define the following polynomial from $\mathcal{H}_d$ into $\mathbb{C}$:

$$\operatorname{discr}_v(f) = R(f_1, \dots, f_n, \det \begin{bmatrix} Df \\ v^* \end{bmatrix})$$

In the particular case $v = e_1 = (1, 0, \dots 0)$, $\operatorname{discr}_v(f) = 0$ is a necessary and sufficient condition for the existence of $\zeta \neq 0$, $f(\zeta) = 0$, such that $Df(\zeta)_{|V(\zeta)}$ does not have full rank. Therefore, we can write :

$$\Sigma^{\text{aff}} = Z(\operatorname{discr}_{e_1}(f))$$

In the projective and pseudo-Newton case, f is degenerate if and only if there is $\zeta \neq 0$, $f(\zeta) = 0$, such that $Df(\zeta)$ has rank $< n$. Indeed, $Df(\zeta)\zeta = 0$, so $Df(\zeta)$

has rank $< n$ if and only if $\begin{bmatrix} Df(\zeta) \\ \zeta^* \end{bmatrix}$ has rank $< n+1$. This is equivalent, in other notation, to say that $Df(\zeta)_{|V(\zeta)}$ has rank $< n$.

Thus,

$$\Sigma = \Sigma^{\text{proj}} = \Sigma^{\text{pseu}} = \{f : \exists \zeta \neq 0, f(\zeta) = 0, \text{rank}(Df(\zeta)) < n\}$$

Lemma 22. *Let $N > \prod d_i$, and let $V = \{v_1, v_2, \dots v_n\}$ be a family of non-zero vectors of $\mathbb{C}^{n+1}$, not two of them colinear. Let* discr *be the ideal generated by the polynomials discr$_v$ while $v \in V$. Then $\Sigma = Z(\text{discr})$.*

Proof of Lemma 22 : If $f \in \Sigma$, then by definition there is $\zeta \neq 0$ such that $f(\zeta) = 0$ and $\text{rank}Df(\zeta) < n$. It follows that for any $v \in V$, $\begin{bmatrix} Df(\zeta) \\ v^* \end{bmatrix}$ has rank $< n+1$, hence $\text{discr}_v(f) = 0$ for all $v \in V$.

Reciprocally, let $\text{discr}_v(f) = 0$ for all $v \in V$. Then for all $v \in V$, there is $\zeta_v \neq 0$ such that $f(\zeta_v) = 0$ and rank $\begin{bmatrix} Df(\zeta_v) \\ v^* \end{bmatrix} < n+1$.

If there are infinitely many zeros of f, then $f \in \Sigma$ and we are done. Assume there are only finitely many zeros. By Bezout's theorem, there are at most $\prod d_i$ different zeros of f.

This implies that at least two of the ζ_v should be the same. Call this point ζ. There are $v, w \in V$ such that : $Df(\zeta)v = Df(\zeta)w = 0$, v and w linearly independent. Therefore, $\text{rank}Df(\zeta) < n$, hence $f \in \Sigma$. This proves Lemma 22.

Let us fix $V = \{(1, v_1, \dots, v_n), v_i \in \mathbb{N}, 0 \leq v_i \leq n\}$. Then V contains more than $\prod d_i$ non-colinear elements.

Let r_i be the degree of discr_v in the coefficients of f_i. Then :

Lemma 23.

$$r_i = \deg_{f_i}(\text{discr}_v) = \prod d_j + (-n + \sum d_j) \prod_{j \neq i} d_j$$

$$B(\text{discr}_v) \leq \left(3n!n \begin{pmatrix} -n + 2\sum d_j \\ n \end{pmatrix} \prod (d_j - 1) \right)^{\sum r_i}$$

Indeed, the degree in ζ of det $\begin{bmatrix} Df \\ v^* \end{bmatrix}$ is $-n + \sum d_i$. Each coefficient of this polynomial is a product of a coefficient of each f_i and a numeric value not bigger than $n!n \prod (d_j - 1)$. It follows from Theorem 17 that $r_i = \prod d_j + (-n + \sum d_j) \prod_{j \neq i} d_j$.

Consider now the polynomial $R\left(f_1, \dots, f_n, \det\begin{bmatrix} Dg \\ v^* \end{bmatrix}\right)$. By Lemma 21,

$$B(R(f_1, \dots, f_n, \det\begin{bmatrix} Dg \\ v^* \end{bmatrix})) \leq \left(3\begin{pmatrix} -n+2\sum d_j \\ n \end{pmatrix} n!n\prod(d_j-1)\right)^{\sum r_i}$$

Since specialization equating variables do not increase B, we obtain :

$$B(R(f_1, \dots, f_n, \det\begin{bmatrix} Df \\ v^* \end{bmatrix})) \leq \left(3\begin{pmatrix} -n+2\sum d_j \\ n \end{pmatrix} n!n\prod(d_j-1)\right)^{\sum r_i}$$

Proof of Theorem 13 :

Consider the mapping :

$$\begin{array}{rccl} \sigma : & \mathcal{H}_d & \to & \mathbb{C}^N \\ & f = f_1 f_2 \dots f_n & \mapsto & \dots (f^I = f_1^{I_1} f_2^{I_2} \dots f_n^{I_n}) \dots \quad , |I_j| = r_j \end{array}$$

where $N = \prod \begin{pmatrix} r_i + S(f_i) - 1 \\ r_i \end{pmatrix}$. $N \leq (\max r_i + \max S(f_i) - 1)^{\sum r_i}$

This mapping associates to each system f, a vector whose coordinates are the values of all the monomials of degree r_i in f_i. Since according to Lemma 23 the discriminant discr_v has degree r_i in f_i, we obtain :

$$\mathrm{discr}_v(f) = \sigma_*(\mathrm{discr}_v)\sigma(f)$$

Above, the *pull-forward* σ_* associates to each polynomial of degree r_i in each f_i, the (row) vector of its coefficients. Using that $|a| \leq H(a)$, $\|\sigma_*(\mathrm{discr}_v)\|_1 \leq B(\mathrm{discr}_v)$.

LEMMA 24. *Let* $p, q \in \mathbb{C}^N$, $p^t q \neq 0$. *Then* $d_{\mathrm{proj}}(p^\perp, q) \geq \frac{1}{\|p\|_2\|q\|_2} \geq \frac{1}{\|p\|_1\|q\|_2}$

Lemma 24 follows from the fact that the nearest point to q in $p^\perp$ is its orthogonal projection $q - \frac{pp^*q}{\|p\|_2{}^2}$. Therefore, the distance $d_{\mathrm{proj}}(q, p^\perp)$ is not less than $\frac{p^*q}{\|p\|_2\|q\|_2}$

Lemma 24 gives us a bound on the distance between $\sigma(f)$ and $\sigma(\mathrm{discr}_v)^\perp$:

$$d_{\mathrm{proj}}(\sigma(f), \sigma(\mathrm{discr}_v)^\perp) \geq$$

$$\geq \left(3n!n\begin{pmatrix} -n+2\sum d_j \\ n \end{pmatrix}\prod(d_j-1)\right)^{-\sum r_i} \left(\sqrt{N}\prod(H(f_i)^{r_i})\right)^{-1}$$

Now we consider a path $\gamma : [0,1] \to \mathcal{H}_d$, with $\|\gamma(t)\|_{\mathrm{k}} = c$ and $\|\gamma'(t)\|_{\mathrm{k}} = cv$. Then $d_{\mathrm{proj}}(\gamma(0), \gamma(1)) \leq v$. We compute :

$$\begin{aligned}
\|(\sigma \circ \gamma)(t)\|_2 &\geq \sqrt{\sum_I (\sigma_I \circ \gamma)(t)^2} \\
&\geq \prod_i (\max_j |\gamma_i(t)_j|^{r_i - 1}) \|\gamma(t)\|_2 \\
&\geq \prod_i (\max_j |\gamma_i(t)_j|^{r_i - 1}) \|\gamma(t)\|_{\mathrm{k}} \\
\|\gamma'(t)\|_2 &\leq \|\gamma'(t)\|_{\mathrm{k}} \sqrt{D!} \\
\|(\sigma \circ \gamma)'(t)\|_2 &\leq cv\sqrt{D!}\sqrt{N} \max_i r_i \prod_i \max_j |\gamma_i(t)_J|^{r_i - 1}
\end{aligned}$$

Thus,

$$\frac{\|(\sigma \circ \gamma)'(t)\|_2}{\|(\sigma \circ \gamma)(t)\|_2} \leq v\sqrt{D!}\sqrt{N} \max_i r_i$$

And hence :

$$\frac{\|(\sigma \circ \gamma)(1) - (\sigma \circ \gamma)(0)\|_2}{\|(\sigma \circ \gamma)(0)\|_2} \leq v\sqrt{D!}\sqrt{N} \max_i r_i$$

Assume furthermore that $\gamma(0) = f$ and that $\mathrm{discr}_v(\gamma(1)) = 0$. Then :

$$\frac{\|(\sigma \circ \gamma)(1) - (\sigma \circ \gamma)(0)\|_2}{\|(\sigma \circ \gamma)(0)\|_2} \leq \sqrt{N}\sqrt{D!} \max r_i d_{\mathrm{proj}}(f, Z(\mathrm{discr}_v))$$

Introducing Lemma 24, we obtain :

$$\begin{aligned}
&d_{\mathrm{proj}}(f, Z(discr_v)) \geq \\
&\quad \geq \left(\sqrt{D!}N \max r_i\right)^{-1} \left(3 \begin{pmatrix} -n + 2\sum d_i \\ n \end{pmatrix} n!n \prod (d_j - 1) H(f)\right)^{-\sum r_i}
\end{aligned}$$

Where $N \leq (\max r_i + \max S(f_i) - 1)^{\sum r_i}$.

Let $f \not\in \Sigma$. Then there is v in V such that $\mathrm{discr}_v(f) \neq 0$. The equation above bounds the distance of f to the variety $Z(\mathrm{discr}_v)$, that contains Σ.

Therefore, we can set the bound :

$$d(\Sigma) = n \prod d_j \sum d_j \geq \sum r_i$$

This exponent is clearly less than $\sum r_i$. We also can set :

$$\mu(\Sigma) = \sqrt{D!}d(\Sigma) \left(3n!n(d(\Sigma) + \max S(f_i))(2\sum d_j)^n \prod (d_j - 1)\right)^d (\Sigma)$$

and Theorem 13 is proved.

Proof of Theorem 14 :

Let us suppose that there are reals $\tilde{t}_1, \tilde{t}_2$ such that $\tilde{t}_1 f + \tilde{t}_2 g \in \Sigma$. Then in particular, there is ζ such that :

$$\tilde{t}_1 f(\zeta) + \tilde{t}_2 g(\zeta) = 0$$

and

$$\det \begin{bmatrix} \frac{\partial}{\partial \zeta} \tilde{t}_1 f(\zeta) + \tilde{t}_2 g(\zeta) \\ (1 \ , \ 0 \ldots 0) \end{bmatrix} = 0$$

Moreover, we can equate one of the $\tilde{t}_i$ to one of the ζ_i, say $t_1 = \zeta_{n+1}$. Then, we obtain a system of the form :

$$\begin{array}{lcl} \operatorname{Re}(t_1 f(\zeta) + t_2 g(\zeta)) & = & 0 \\ \operatorname{Im}(t_1 f(\zeta) + t_2 g(\zeta)) & = & 0 \\ \operatorname{Re}\left(\det \begin{bmatrix} \frac{\partial}{\partial \zeta}(t_1 f(\zeta) + t_2 g(\zeta)) \\ (1 \ , \ 0 \ldots 0) \end{bmatrix}\right) & = & 0 \\ \operatorname{Im}\left(\det \begin{bmatrix} \frac{\partial}{\partial \zeta}(t_1 f(\zeta) + t_2 g(\zeta)) \\ (1 \ , \ 0 \ldots 0) \end{bmatrix}\right) & = & 0 \end{array} \tag{45}$$

If we have in mind that t_1 and t_2 are *real* variables, it is clear that system (45) is a system of $2n + 2$ real homogeneous equations in $2n + 2$ real variables $(t_1, t_2, \operatorname{Re}(\zeta_1), \operatorname{Im}(\zeta_1), \ldots, \operatorname{Im}(\zeta_n))$.

If there are $\tilde{t}_1$ and $\tilde{t}_2$ such that $\tilde{t}_1 f + \tilde{t}_2 g \in \Sigma$, then system (45) has a solution, and therefore its resultant vanishes. The converse is not true, since there may exist a complex solution that is not a real solution.

Let us call $R(f, g)$ the resultant of system (45). Then, in the *realization* of $\mathcal{H}_d \times \mathcal{H}_d$, we define the variety :

$$\Sigma_0 = \{f, g : R(f, g) = 0\}$$

This variety is proper : if we set $f = g$, there is f such that there is no real solution for system (45), since any real solution would imply the existence of a complex solution for $f(\zeta) = 0, \det \begin{bmatrix} Df(\zeta) \\ {e_1}^* \end{bmatrix} = 0$, and not all f are degenerate.

Let r_i be the degree of R in the coefficients of f_i and g_i. Then :

$$r_i = 2 \prod (d_j + 1) \prod_{j \neq i} (d_j + 1) (\sum d_j)^2 \ + \ 2 \prod (d_j + 1)^2 \sum d_j$$

Therefore,

$$\sum r_i \leq 2n \prod (d_j + 1)^2 (1 + \sum d_j)^2$$

So we set :

$$d(\Sigma_0) = 2n \prod (d_j+1)^2 (1+\sum d_j)^2$$

When we *realize* a complex polynomial of degree d, we obtain two polynomials of height at most 2^d. Therefore, the height of system (45) is bounded by :

$$2^D n! \prod (d_j - 1)$$

Therefore, Lemma 21 implies :

$$B(R) \le \left(3 \begin{pmatrix} -1+4\sum d_j \\ n \end{pmatrix} 2^D n! n \prod (d_j-1) \right)^{d(\Sigma_0)}$$

Thus, as in the proof of Theorem 13, we can bound $\mu(f_t)$ by $\sqrt{D!} N B(R) d(\Sigma_0)$, and set :

$$\mu(\Sigma_0) = \sqrt{D!} d(\Sigma_0) \times$$
$$\times \left(32^D n! n (d(\Sigma_0) + \max S(f_i) - 1) \prod (d_j - 1)(-1+4\sum d_j)^{2n+1} \right)^{d(\Sigma_0)}$$

And theorem 14 is proved.

5. Diophantine decision problem

In this section, we are concerned about the following problem : Let $0 \le x \le 1$ be a real number, such that $2^{-k}x$ is an integer. Let H be an integer, and let $\epsilon < 1/2H^2$. We want to decide if there is a rational $\frac{p}{q}$ in $[x-\epsilon, x+\epsilon]$, such that $q \le H$. In the case it exists, we want to be able to find it. Consider that $\epsilon = 2^{-k'}$ for some $k' \ge k$.

If a solution exists, it is unique. Indeed, let $\frac{p}{q}$ and $\frac{p'}{q'}$ be different solutions. Then $|\frac{p}{q} - \frac{p'}{q'}| = |\frac{pq'-p'q}{qq'}| \ge \frac{1}{H^2} \ge 2\epsilon$, a contradiction.

Lemma 25. *There is a program that solves this problem within polynomial time on* $\log k'$

The cases $x - \epsilon \le 0$ and $x + \epsilon \ge 1$ are trivial, and can be discarded. Thus, we assume that $x \in [\epsilon, 1-\epsilon]$.

Geometrically, we consider the (q,p) plane. Let Q be the square $0 < q, p \le H$. The set of rationals $1 \le q, p \le 1$ is represented by the lattice of points of Q with integer coordinates.

Let Δ be the line of equation $p = qx$. Let Δ_1 and Δ_2 be the lines of equation $p = (q+\epsilon)x$ and $p = (q-\epsilon)x$, respectively. Δ corresponds to the real number x, and the condition $x - \epsilon \leq \frac{p}{q} \leq x + \epsilon$ is true if and only if (p, q) belongs to the cone C limited by Δ_1 and Δ_2.

Thus, we want to find a lattice point (p, q) in $Q \cap C$. In order to do that, we use the continued fraction expansion of x.

Consider the recurrence :

$$r_0 = x$$

$$r_{i+1} = \frac{1}{r_i} - \lfloor \frac{1}{r_i} \rfloor$$

We also write :

$$\sigma_{i+1} = \lfloor \frac{1}{r_i} \rfloor$$

Notice that when r_0 is rational, $H(r_{i+1}) \leq H(r_0)$, so the r_i can be computed with rational arithmetics in a bounded number of bits.

We can also define the approximants p_i, q_i of x by the recurrence :

$$p_0 = 0, q_0 = 1, p_1 = 1, q_1 = 1$$

$$p_{i+1} = p_{i-1} + \sigma_i p_i$$

$$q_{i+1} = q_{i-1} + \sigma_i q_i$$

It is a fact that :

$$\frac{p_i}{q_i} = \frac{1}{\sigma_1 + \frac{1}{\cdots \frac{1}{\sigma_{i-1}}}}$$

The pairs p_i, q_i have no common denominator, and in the case x is rational (our case) x is attained in a finite number of steps. Indeed, from the recurrence, we easily obtain that $p_{i+2} \geq 2p_i$ and $q_{i+2} \geq 2q_i$, so $p_i/q_i = x$ for some $i \leq 2k$.

p_i/q_i are known to be the best possible approximations of x with a given height :

THEOREM 20 (181 in [**6**]). *If $n > 1$, $0 < q < q_n$, and $p/q \neq p_n/q_n$, then*

$$\left| \frac{p_n}{q_n} - x \right| < \left| \frac{p}{q} - x \right|$$

That implies that if (q_i, p_i) does not belong to C, then there is no lattice point in C at the left of q_i.

For odd i's, q_i, p_i is below Δ. For even i's, q_i, p_i is above Δ.

Now we come to the algorithm : we compute q_i, p_i, until q_i, p_i is in C or q_{i-2}, p_{i-2} is not in Q.

If q_i, p_i is in Q, then we are done (output p_i, q_i).

If q_{i-2}, p_{i-2} is not in Q, then there is no rational solution of bounded height, and we are also done (output NO SOLUTION).

There is a third case. There might be lattice points in $C \cap Q$ along segments $((q_{i-2}, p_{i-2})(q_i, p_i))$ or $((q_{i-1}, p_{i-1})(q_{i+1}, p_{i+1}))$.

Assume for instance that i is even, and we want to check for lattice points in $Q \cap C$ along $((q_{i-2}, p_{i-2})(q_i, p_i))$. Then we write the parametric equation :

$$(q, p) = (q_{i-2}, p_{i-2}) + \lambda(q_{i-1}, p_{i-1})$$

We combine it with the equation of Δ_2, and obtain :

$$\lambda = \frac{q_{i-2}(x-\epsilon) - p_{i-2}}{p_{i-1} - q_{i-1}(x-\epsilon)}$$

The natural candidate to a solution on this side is the point :

$$(q, p) = (q_{i-2}, p_{i-2}) + \lceil \lambda \rceil (q_{i-1}, p_{i-1})$$

If this fail, we have to look for the point :

$$(q, p) = (q_{i-1}, p_{i-1}) + \lceil \lambda' \rceil (q_i, p_i)$$

Where

$$\lambda' = \frac{q_{i-1}(x+\epsilon) - p_{i-1}}{p_i - q_i(x+\epsilon)}$$

If this one also fails to be in $Q \cap C$, there is no rational solution to our problem. The other case (i odd) is analogous.

Since all the computations can be performed by rational arithmetics in precision k', and there are at most $O(\log k')$ continued fraction iterations, the total time is polynomial in $\log k'$.

6. Proof of theorem 15

In order to deal with points in the projective space, the usual concept of height may be unadequate; instead we define the height of ratios:

$$H_r(x_0 : \ldots x_n) = \max_{x_j \neq 0} H(\frac{x_i}{x_j})$$

where the height of a rational is given by $H(p/q) = \max(|p|, q)$, whenever p and q have no common divisor.

Theorem 19 can be restated more conveniently in terms of H_r :

Theorem 21. *Let $f \in \mathcal{H}_d$ be non-degenerate, $f(\zeta) = 0$, and let ζ be rational. Then $H_r(x) \leq \prod {d_i}^{-1} (3nH(f) \dim \mathcal{H}_d)^{\prod d_i}$.*

Also, there is a "gap" between different rational points of a bounded height :

Lemma 26. *Let x,x' be rational points of $\mathbb{C}^{n+1}$ (this means that $x, x' \in \mathbb{Q}^{n+1}$), such that $d_{\mathrm{proj}}(x, x') \neq 0$. Let $H = \max(H_r(x), H_r(x'))$. Then $d_{\mathrm{proj}}(x, x') \geq \frac{1}{H^2(n+1)}$*

Proof : assume without loss of generality that $x_0 = 1$, $|x_j| \leq 1$ for all j, $x'_i = 1$ for some i, and $|x'_j| \leq 1$.

$$d(\frac{x}{\|x\|}, \frac{x'}{\|x'\|}) \geq \frac{1}{\sqrt{n+1}} d(x, x') \geq \frac{1}{H^2\sqrt{n+1}}$$

since we are projecting from an hypercube into its inscribed hypersphere. Now we project into the hyperplane passing by $x/\|x\|$ perpendicular to x'. In the worst case :

$$d_{\mathrm{proj}}(x, x') \geq \frac{1}{\sqrt{n+1}} d(\frac{x}{\|x\|}, \frac{x'}{\|x'\|}) \geq \frac{1}{H^2(n+1)}$$

Now, the Quadratic Convergence Theorem provides us a point $x \in \mathbb{C}^{n+1}$ and a neighborhood of radius $\delta < \frac{1}{2H^2(n+1)}$ over x, containing a root ζ. Here, we set $H = \prod d_i (3nH(f) \dim \mathcal{H}_d)^{\prod d_i}$, so that if ζ is rational, it belongs to a ball of radius δ over x.

Thus, we need about $1 + \log(-\log \delta)$ iterations of N_ϵ, $\epsilon = \delta^2$. We need also that $\epsilon\gamma < 1/40$, but γ can be bounded in terms of μ by using theorem 13. Thus we obtain a polynomial time bound in $\prod d_i$, n, D, $\log H(f)$, $\dim H_d$ and $\log \mu(\zeta)$. However, since $\log \mu(\zeta)$ is bounded by the polynomial in Corollary 2, we obtain a polynomial time bound in $\prod d_i$, n, D, $\log H(f)$ and $\dim \mathcal{H}_d$ alone.

Without loss of generality, we assume $x_0 = 1$ and $0 \leq x_j \leq 1$. We have $|x_i - \zeta_i| \leq \frac{1}{2H^2}$, so we can use the diophantine approximation algorithm to find the ζ_i's , if they exist. This can be done in time n times a polynomial in $\log H$. This proves theorem 15.

7. Separation of roots

We prove theorem 16 as follows : Let $h = x_2 - x_1$. By scaling x_2, assume first that $x_2 \in x_1 + ker Df(x_1)^\perp = x_1 + x_1^\perp$. We have :

$$f(x_1) = 0$$

$$f(x_2) = 0 = f(x_1) + Df(x_1)h + \frac{D^2 f(x_1)h^2}{2} + \cdots + \frac{D^k f(x_1)h^k}{k!} + \ldots$$

So :

$$-Df(x_1)h = \sum_{k \geq 2} \frac{1}{k!} D^k f(x_1) h^k$$

Multiplying both terms by $Df(x_1)^\dagger$, and using that $h \in kerDf(x_1)^\perp$, we obtain :

$$h = \sum_{k \geq 2} \frac{1}{k!} Df(x_1)^\dagger D^k f(x_1) h^k$$

Passing to the norms, and dividing by $\|h\|$:

$$1 \leq \sum_{k \geq 2} \frac{1}{k!} \|Df(x_1)^\dagger D^k f(x_1)\| \|h\|^{k-1} \leq \sum_{k \geq 2} \frac{1}{k!} \frac{\gamma(f, x_1)^{k-1} \|h\|^{k-1}}{\|x_1\|^{k-1}}$$

This implies :

$$1 \leq \frac{1}{1 - \gamma(f, x_1) \frac{\|x_2 - x_1\|}{\|x_1\|}} - 1$$

We get :

$$\gamma(f, x_1) \frac{\|x_2 - x_1\|}{\|x_1\|} \geq \frac{1}{2}$$

$$\frac{\|x_2 - x_1\|}{\|x_1\|} \geq \frac{1}{2\gamma(f, x_1)}$$

Now, assuming again $x_2 \in x_1 + x_1^\perp$, we have :

$$d_p(x_1, x_2) = \sin \tan^{-1} \frac{\|x_2 - x_1\|}{\|x_1\|} \quad \square$$

However, we can say nothing about the height of a root x solely in terms of γ, since γ is rotation-invariant : any root x can be mapped anywhere in projective space, without changing γ.

CHAPTER 5

Global complexity of solving systems of polynomials

A global algorithm for solving systems of polynomial equations is constructed. Main Theorem is proved.

1. Introduction

A Theorem on the global complexity of solving systems of polynomial equations with integer coefficients was stated in Chapter 1. Here, we will construct an algorithm corresponding to the machine in Main Theorem , and prove the Theorem.

We will use the analysis of the Pseudo-Newton iteration from Chapter 2, together with the construction of a Pseudo-Newton operator $\texttt{Pseudo}(f, x)$ in Chapter 3. Chapter 4 will provide us with a worst case estimate of the condition number of f.

Basically, we have to answer here to the following questions : How to get a starting system, how to certify an approximate solution, and how to design a global algorithm that will succeed generically.

Since the worst case bounds on μ are too pessimistic, the algorithm will rather guess an upper bound for μ, and attempt to solve the system. If this fails, the algorithm will increase the upper bound, and try again. This approach leads to algorithms that are also tractable in practice.

2. How to obtain starting points.

Obtaining starting points for a homotopy path is easy. Obtaining good starting points is, at this time, an open problem (see [**14**]). We give only the easy and less efficient way :

Let $g_k(x) = x^k - 1 = (x - \zeta^0) \dots (x - \zeta^{k-1})$, where ζ is a k-th root of unity. Then the system f, $f_i(x) = g_{d_i}(x_0, x_i)$ is clearly a non-degenerate starting point.

LEMMA 27. *There is an algorithm* `STARTPOINT` *, returning a system of polynomials $f =$* `STARTPOINT` *$(n, d) \in \mathcal{H}_d$, with height 1, together with an approximate solution X of f.*

We will indicate the proof of the Lemma. We have to bound the cost of obtaining a d-th root of unity.

Consider the polynomial in one variable : $g(x) = x^d - 1$. Let ζ be a d-th root of unity, $g(\zeta) = 0$. $g'(\zeta) = d\zeta^{d-1} = \frac{d}{\zeta}$.

$$\mu^{\mathrm{aff}}(g, \zeta) \le \|g\|_{\mathrm{k}} \left\| \frac{\zeta}{d} \sqrt{d} |\zeta|^{d-1} \right\|_2 = \sqrt{\frac{2}{d}}$$

We use Lemma 14 of chapter 2. According to that Lemma, if :

$$\frac{\|(1,x) - (1,\zeta)\|_2}{\|(1,\zeta)\|_2} \gamma(\zeta) \le \bar{u}$$

then :

$$\alpha(f, x) \le \frac{\bar{u}}{\psi(\bar{u})^2}$$

Using that $\gamma(f, \zeta) \le \frac{d^{3/2}}{2} \mu(f, \zeta)$, we obtain that :

LEMMA 28. *If $|x - \zeta| \le \frac{2\bar{u}}{d}$, then $\alpha(f, x) \le \frac{\bar{u}}{\psi(\bar{u})^2}$*

That means that if we are able to obtain an approximation good enough of ζ, we obtain an approximate zero associated to ζ. We can compute ζ to a higher precision by Affine Newton iteration, and then compute all the ζ^i to obtain the approximate solution of f.

In order to obtain a good approximation of ζ, we can write :

$$\zeta = 1 + ai + \frac{1}{2}(ai)^2 + \cdots + \frac{1}{k!}(ai)^k + E_k$$

We will approximate ζ by truncating the series. Suppose we want to do that with error δ, we need to choose k such that $|E_k| < \delta$.

Since $a = \frac{2\pi}{d}$ and the cases $d = 1$ and $d = 2$ are trivial, we can assume that $a < 3$. Then for $k > 6$, $a^k/k! \le 2^{6-k}$. This allows us to compute the approximation of ζ using a number of terms logarithmic in δ^{-1}.

Therefore, it is easy to see that a 2δ approximation of ζ can be obtained within time polynomial in $-\log \delta$. Setting $\delta < \frac{1}{16d}$, we get $\alpha(f, x) < 1/8$. According to Theorem 1, x is an approximate zero associated to ζ.

3. How to bound μ

In this section, we give procedures to certify that a given list of points is an approximate zero of a system of polynomial equations.

LEMMA 29. *There is a procedure to compute a bound M (finite or not) such that :*

$$\mu(f,x) \le M \le 4\sqrt{n}\mu(f,x)$$

That procedure is polynomial time in n, D, $\prod d_i$, $S(f)$, $\log H(f)$.

The following algorithm returns the bound $\text{mu}(f,x)$ for $\mu(f,x)$, such that, ignoring the numeric error, $2\mu(f,x) \le \text{mu}(f,x) \le 2\sqrt{n}\mu(f,x)$. Of course, numerical error will occur, and that is the reason Lemma 29 is weaker than the bound above. Let us first state the algorithm, and then proceed with the error analysis.

```
ALGORITHM   $z'' \leftarrow$ mu ( $f$, $x$ )
      1 Compute $A = \left[ \frac{\partial f_i}{\partial x_j} {d_i}^{-1/2} {\|x\|_2}^{-d_i+1} \right]$
      2 Make $A^t = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$
      3 Compute $c = {R^t}^{-1} I$
      4 Set mu $= 2\|f\|_k \|c\|_f$
```

Proof of correctness : Assume there is no numerical error.

$$\begin{aligned} \mu(f,x) &= \|f\|_{\mathrm{k}} \left\| Df^{\dagger} \operatorname{diag}(\sqrt{d_i} {\|x\|_2}^{d_i-1}) \right\|_2 \\ &= \|f\|_{\mathrm{k}} \left\| Q \begin{bmatrix} R^{\mathrm{t}} \\ 0 \end{bmatrix}^{-1} \right\|_2 \\ &= \|f\|_{\mathrm{k}} \left\| {R^{\mathrm{t}}}^{-1} \right\|_2 \end{aligned}$$

We use the fact :

$$\left\| {R^{\mathrm{t}}}^{-1} \right\|_2 \le \left\| {R^{\mathrm{t}}}^{-1} \right\|_{\mathrm{f}} \le \sqrt{n} \left\| {R^{\mathrm{t}}}^{-1} \right\|_2$$

It follows that :

$$2\mu(f,x) \le 2\text{mu}(f,x) = 2\|f\|_{\mathrm{k}} \left\| {R^{\mathrm{t}}}^{-1} \right\|_{\mathrm{f}} \le 2\sqrt{n}\mu(f,x)$$

First order error analysis :

Let $q = 4D + S(f) + \max S(f_i) + 2.25n^3 + 6.25n^2 + 2n + 1$. Let $M_i(f,x) = {R^{\mathrm{t}}}^{-1} e_i$, then $\|M_i(f,x)\|_2 = \left\| A^{\dagger} b_i \right\|_2$.

Line 1 : A is computed with forward error bounded by :

$$\|\delta_1 A\|_{\max} \le \|f\|_{\mathrm{k}} \epsilon_m (2D + \max S(f_i) + 2n + 1)$$

Line 2 : This line introduces some backward error in A :

$$\|\delta_{qr} A\|_2 \leq \|A\|_2 \epsilon_m (2.25n^2 + 5.25n)$$

$$\|\delta_{qr} A\|_{\max} \leq n\|A\|_{\max} \epsilon_m (2.25n^2 + 5.25n)$$

Since $\|A\|_{\max} \leq \|f\|_{\mathrm{k}}$,

$$\|\delta_{qr} A\|_{\max} \leq \|f\|_{\mathrm{k}} \epsilon_m (2.25n^2 + 5.25n)$$

Line 3 : This line introduces backward error in R, hence in A :

$$\|\delta_R A\|_{\max} \leq \|\delta_2 R\|_{\max} \leq \|R\|_{\max} \epsilon_m (n) \leq n\|A\|_{\max} \epsilon_m (n)$$

We can write :

$$\|\delta_2 A\|_{\max} \leq \|f\|_{\mathrm{k}} \epsilon_m (n^2)$$

Therefore :

$$\begin{aligned} \|\delta A\|_{\max} &\leq \|f\|_{\mathrm{k}} \epsilon_m (2D + \max S(f_i) + 2.25n^3 + 6.25n^2 + 2n + 1) \\ &\leq \|f\|_{\mathrm{k}} \epsilon_m (q - S(f) - 2D) \end{aligned}$$

Now,

$$\|DM_i(f,x)\|_{\max,2} \leq \left\|A^\dagger e_i\right\|_2 \left\|A^\dagger\right\|_2$$

Line 4 introduces forward error bounded by $\|f\|_{\mathrm{k}} \epsilon_m (S(f) + 2D)$, therefore, in first order analysis,

$$\|\delta \mathrm{mu}\|_2 \leq \mu^2 \epsilon_m (q)$$

A preliminary Lemma : In order to give some rigorous bounds for μ, we shall need the following Lemma :

LEMMA 30. *Let C, δC be matrices, so that $\|\delta C\|_{\max} \leq \|f\|_{\mathrm{k}} \epsilon_m (q)$. Let $2n\|f\|_{\mathrm{k}} \left\|C^\dagger\right\|_2 \epsilon_m (q) < 1$. Then :*

$$\frac{1}{2}\left\|C^\dagger\right\|_2 \leq \left\|(C + \delta C)^\dagger\right\|_2 \leq 2\left\|C^\dagger\right\|_2$$

Proof :

$$\|\delta C\|_2 \leq n\|\delta C\|_{\max} \leq n\|f\|_{\mathrm{k}} \epsilon_m (q) \leq \frac{1}{2\|C^\dagger\|_2}$$

Therefore, $\left\|(C + \delta C)^\dagger\right\|_2$ verifies :

$$\frac{1}{2}\left\|C^\dagger\right\|_2 \leq \left\|(C + \delta C)^\dagger\right\|_2 \leq 2\left\|C^\dagger\right\|_2$$

Proof of Lemma 29 :

Step 1 : Choose ϵ_m so that :

$$4n^{3/2}\mu(\Sigma)H(f)^{d(\Sigma)}\epsilon_m(q) < 1 \tag{46}$$

where $\mu(\Sigma)$ and $d(\Sigma)$ are defined in Theorem 13 of Chapter 4. It is clear from the corollary of the same theorem that the number of bits of precision we are requiring is bounded by a polynomial in n, D, $\prod d_i$, $\dim \mathcal{H}_d$ and also $\log H(f)$.

Step 2 : Compute $\mathtt{mu}(f,x)$. In case a division by zero occurs, or in case $\frac{\mathrm{mu}(f,x)}{2\sqrt{n}} > \mu(\Sigma)H(f)^{d(\Sigma)}$, then return ∞.

Indeed, assume that μ is finite. $\|\delta A\|_{\max} \leq \|f\|_{\mathrm{k}}\epsilon_m(q)$, so that :

$$\|f\|_{\mathrm{k}}\|A+\delta A\|_{\mathrm{f}} \leq \|f\|_{\mathrm{k}}\|A+\delta A\|_2\sqrt{n} \leq 2\|f\|_{\mathrm{k}}\|A\|_2\sqrt{n} \leq 2\mu\sqrt{n}$$

That implies that :

$$\mu \geq \mu(\Sigma)H(f)^{d(\Sigma)}$$

contradicting Theorem 13 of Chapter 4.

Step 3 : Since $\frac{\mathrm{mu}(f,x)}{2\sqrt{n}} \leq \mu(\Sigma)H(f)^{d(\Sigma)}$, we can use equation (46) to obtain :

$$2n\|f\|_{\mathrm{k}}\left\|(A+\delta A)^{\dagger}\right\|_{\mathrm{f}}\epsilon_m(q) < 1$$

Hence,

$$2n\|f\|_{\mathrm{k}}\left\|(A+\delta A)^{\dagger}\right\|_2\epsilon_m(q) < 1$$

Therefore, Lemma 30 implies :

$$\frac{1}{2}\left\|A^{\dagger}\right\|_2 \leq \left\|(A+\delta A)^{\dagger}\right\|_2 \leq 2\left\|A^{\dagger}\right\|_2$$

$$\mu \leq 2\|f\|_{\mathrm{k}}\left\|(A+\delta A)^{\dagger}\right\|_2 \leq 4\mu$$

$$\mu \leq 2\|f\|_{\mathrm{k}}\left\|(A+\delta A)^{\dagger}\right\|_{\mathrm{f}} \leq 4\sqrt{n}\mu$$

Thus,

$$\mu \leq \mathrm{mu} \leq 4\sqrt{n}\mu$$

4. More estimates on μ

LEMMA 31. *Let $f \in \mathcal{H}_d$, and assume that $x, \zeta \in \mathbb{C}^{n+1}$ are non-zero. Let $u = d_{\mathrm{proj}}(x,\zeta)\gamma^{\mathrm{pseu}}(f,\zeta)$. Then :*

$$\mu^{\mathrm{pseu}}(f,x) \leq \frac{(1-u)^2}{\psi(u)}\mu^{\mathrm{pseu}}(f,\zeta)$$

Proof of Lemma 31 : Let ζ be scaled in such a way that $d_{\text{proj}}(x, \zeta) = \frac{\|x-\zeta\|_2}{\|\zeta\|_2}$. We proceed as in Lemma 2 of Chapter 2, and break $\mu(f, x)$ in :

$$\begin{aligned}\mu(f,x) \leq & \|f\|_{\text{k}} \left\| Df(x)_{|V(x)}{}^{-1} Df(x)_{|V(\zeta)} \right\|_2 \\ & \left\| Df(x)_{|V(\zeta)}{}^{-1} Df(\zeta)_{|V(\zeta)} \right\|_2 \\ & \left\| Df(\zeta)_{|V(\zeta)}{}^{-1} \text{diag} \|\zeta\|_2{}^{d_i} \sqrt{d_i} \right\|_2 \\ & \left\| \text{diag} \left(\frac{\|x\|_2}{\|\zeta\|_2} \right)^{d_i} \right\|_2\end{aligned}$$

Using the estimates of Part 1 and Part 2 of the proof of Lemma 2 Chapter 2, and using also the fact that $\|x\|_2 \leq \|\zeta\|_2$, we obtain :

$$\mu(f, x) \leq \frac{(1-u)^2}{\psi(u)} \mu(f, x)$$

This proves the Lemma 31.

In particular, if $\mu(f, \zeta) \leq \bar{\mu}$, and if $4 d_{\text{proj}}(x, \zeta) \gamma^{\text{pseu}}(f, \zeta) < \frac{1}{2}$, it is possible to conclude that $\mu(f, x) \leq 2\bar{\mu}$. This kind of estimate allows to bound μ in a neighborhood of the path we are following. It allows us to bound the conditioning of approximate zeros, and therefore to set the machine precision necessary for Newton iteration, as it is required in Chapter 3.

5. How to bound η

Let $q = 4D + S(f) + \max S(f_i) + 2.25n^3 + 6.25n^2 + 2n + 1$. Computing a bound for η is straight-forward :

```
ALGORITHM    eta ← eta ( $f$, $x$ )
      1 Compute $n_1 = \|f\|_k$
      2 Compute $n_2 = \|x\|_2$
      3 Set $b_i = \frac{f_i(x)}{(\sqrt{d_i} n_1 n_2)}$
      4 Set eta= $(1 + \frac{1}{8}) \|b\|_2$
```

Forward error analysis :

Line 1 : The forward error is $|\delta n_1| \leq \|f\|_{\text{k}} \epsilon_m (S(f) + 2D)$

Line 2 : We have $|\delta n_2| \leq \|x\|_2 \epsilon_m (2n + 2)$

Line 3 : The error is bounded by :

$$|\delta b_i| \leq \frac{1}{\sqrt{d_i}} \epsilon_m (S(f) + \max S(f_i) + 3D + 2n + 2)$$

Line 4 : We get :

$$|\delta \text{eta}| \leq \epsilon_m(S(f) + \max S(f_i) + 3D + 4n + 1) \leq \epsilon_m(q)$$

And therefore,

LEMMA 32. *In the conditions above, $\eta \leq \texttt{eta} \leq 2\eta$.*

6. How to certify an approximate solution.

We already proved the following facts :

LEMMA 33. *Assume the following condition is verified :*

$$4n^{3/2}\mu(\Sigma)H(f)^{d(\Sigma)}\epsilon_m(q) < 1. \tag{47}$$

Then :

$$\begin{aligned}
\mu(f,x) &< \texttt{mu}(f,x) < 4\sqrt{n}\mu(f,x) \\
\eta(f,x) &< \texttt{eta}(f,x) < 2\eta(f,x) \\
\beta(f,x) &< \texttt{mu}(f,x)\texttt{eta}(f,x) \\
\gamma(f,x) &< D^{3/2}\texttt{mu}(f,x) \\
\alpha(f,x) &< \frac{D^{3/2}}{2}\texttt{mu}(f,x)^2\texttt{eta}(f,x)
\end{aligned}$$

Let f and x be given. Let ζ be the zero associated to x. The following algorithm will return an approximate zero of f, or fail. If it fails, one of the following conditions will be false :

$$d_{\text{proj}}(x,\zeta) \leq \frac{\bar{u}}{\bar{\gamma}} \tag{48}$$

$$\mu(f,\zeta) \leq \bar{\mu} \tag{49}$$

Constant $\bar{\mu}$ was defined in Theorem 9 of Chapter 2. Constant $\bar{\gamma}$ is defined in Corollary 1, Chapter 2.

If conditions (48) and (49) are true, then by Lemma 31, $\mu(f,x) \leq 2\bar{\mu}$. We will prove that in that case, the output of the algorithm is an approximate zero of f, associated to ζ.

```
ALGORITHM    $y \leftarrow$ Zero ($f$, $x$, $\bar{\mu}$)
```

```
Let $k, \delta \leftarrow \lceil 2 + \log\log n \rceil, \lceil \frac{1}{200n} \rceil$
Let $\epsilon_m$ be such that Pseudo has precision $\delta$
  and condition (47) is true.
Let $y \leftarrow$Pseudo $^k(f, x)$
Let $a \leftarrow \frac{D^{3/2}}{2}$mu$(f, x)$ eta$(f, x)$
If $a \geq \frac{1}{8}$ Fail.
Return $y$.
```

It is clear that $\alpha(f, y) < a < 32n\alpha(f, y)$. So if the algorithm succeeds, y is certainly an approximate zero of f.

Now assume conditions (48) and (49). Then $\alpha(f, y) < \min(2^{-2^k-1}, \delta)\alpha(f, x) \leq \frac{1}{32n}\alpha(f, x)$.

Therefore, $a \leq \alpha(f, x) \leq 2\bar{u} < 1/8$, and the algorithm shall succeed. Moreover, the result y is obtained from x by an approximate Newton iteration, so it is associated to the same zero ray ζ.

Let we note $Y =$Zero(X) for the application of Zero to a list of points X. The next algorithm if it succeeds, returns an approximate solution of f.

If it is given an approximate solution of f as input, and conditions (48) and (49) are true for every point in X, it succeeds.

ALGORITHM $Z \leftarrow$ Solution (f, X, $\bar{\mu}$)

```
1 If Length($X$) $\neq \prod d_i$ then FAIL.
2 Let $Y \leftarrow$Zero$(f, X, \bar{\mu})$.
3 Let $Z =$Pseudo$^k(f, Y)$, where $k$, $\delta$ and $\epsilon$ are chosen
  so that each point of $Z$ is at distance less than $\frac{1}{4\sqrt{2}D^{3/2}\bar{\mu}}$
  of its associated zero.
4 If two of the $Z_i$ are at distance less than $\frac{1}{2\sqrt{2}D^{3/2}\bar{\mu}}$, then FAIL.
5 Return $Z$.
```

Lines 3 and 4 check that there are not two elements of X with the same associated zero. This follows from Theorem 6 of Chapter 2 and Theorem 16 of Chapter 4. Line 1 checks that there is one X_i for each zero of f.

7. How to solve a generic system.

Given a path $[f_0, f_1]$, it is easy to produce equally spaced systems f_{t_i} in $\mathcal{H}_d$. We just set :

$$t_i = \frac{1}{2} + \frac{\tan\theta_i}{\tan\theta_{\max}}$$

where :

$$2 \sin \theta_{\max} = \left\| \frac{f_0}{\|f_0\|_{\mathrm{k}}} - \frac{f_1}{\|f_1\|_{\mathrm{k}}} \right\|_{\mathrm{k}}$$

and $\theta_i = \left(2\frac{i}{N} - 1\right)\theta_{\max}$, where i ranges between 0 and N. $\theta_{\max}$ is half of the angular distance between f_0 and f_1. It follows from trigonometry that the points :

$$f_{t_i} = (1 - t_i)\frac{f_0}{\|f_0\|_{\mathrm{k}}} + t_i \frac{f_1}{\|f_1\|_{\mathrm{k}}}$$

are equally spaced. Now, path-following can be codified by the following algorithm :

```
ALGORITHM    $Y_1 \leftarrow$ Pathfollow $(f_0, f_1, Y_0)$
      1 $\bar{\alpha}, \bar{u}, \bar{\gamma}, \delta, Z, i \leftarrow 0.02, 0.05, \frac{2}{3}D^{3/2}\bar{m}u, \frac{\bar{m}u}{2\bar{\gamma}}, Y_0, 0$.
      2 Let $t_i$ be s.t.  $d_{proj}(f_{t_i}, f_{t_{i+1}}) \leq \frac{3\bar{\alpha}}{8\bar{\mu}\bar{\gamma}}$
      3 REPEAT $i, Z \leftarrow i + 1,$Pseudo$(f_{t_i}, Z)$
      4 $Y_1 \leftarrow Z$
```

Line 2 can be done as described above, using that $d_{\mathrm{proj}}(f_0, f_1) \leq 1$. Line 3 should be executed with the precision given by Corollary 1 of Chapter 2. The actual machine precision necessary for that is bounded in Theorem 12 of Chapter 3. For a bound on the condition number, we should take $2\bar{\mu}$, to use the result of Lemma 31.

The following algorithm is corresponds to the machine of the Main Theorem.

```
ALGORITHM    $X_1 \leftarrow$ SOLVE $(f_1)$
      $f_0, X_0 \leftarrow$ STARTPOINT$(n, d)$ ; $\bar{\mu} = 2$ ;
      REPEAT
        $\bar{\mu} \leftarrow 2\bar{\mu}$
        $X_1 \leftarrow$Pathfollow$(f_0, f_1, X_0, \bar{\mu})$
        $Z \leftarrow$Solve$(f, X, \bar{\mu})$.
        UNTIL $Z \neq$ FAIL
      Output $Z$
```

A better version of this algorithm retraces only the paths where the final result was not proven to be an approximate zero, and the paths where the final result was near the final result of another path.

Proof of the Main Theorem :

The set U_d are the complements of a section of the real algebraic variety Σ_0 defined in Theorem 14 Chapter 4. This section is just the section of all the (f_0, f)

where $f_0 =$`STARTPOINT`(n,d) . Since this section is proper in $f_0 \times \mathcal{H}_d$, the set U_d is a non-trivial open set.

We set $\mu_0 = \mu(\Sigma_0)$ and $d_0 = d(\Sigma_0)$. The bound on μ comes from Theorem 14.

The complexity analysis of `Solve` follows from the complexity analysis of all the called procedures.

It follows from Corollary 1 in Chapter 2 that the number of homotopy steps is bounded by $O(\mu^2 D^{3/2})$. There are $\prod d_i$ paths to follow. The cost of each approximate Newton iteration is $O(nDS(f)+n^3)$ floating point operations (Theorem 11) with precision ϵ_m, where ϵ_m can be bound, as in the proof of Theorem 12 of Chapter 3, by equation (39) :

$$\epsilon_m = \frac{1}{\frac{32}{\delta}\bar{\mu}(1+4\bar{\mu})\sqrt{D}(2D+1+2\max S(f_i)+p(n))+2H(f)+2H(x)}$$

The cost of performing each arithmetic operation is about $\log \epsilon_m \log\log \epsilon_m$. Therefore, The cost of Path-following can be bounded by :

$$(\prod d_i)(\max d_i)^{3/2}\mu^2\left((n+1)S(f)\max d_i + (n+1)^3\right) \times \\ \times P(\log\mu, \max d_i, \log n, \log S(f), \log\log H(f)) \tag{50}$$

where P is a polynomial.

In Lemma 33, we require ϵ_m to verify condition (47) :

$$4n^{3/2}\mu(\Sigma)H(f)^{d(\Sigma)}\epsilon_m(q) < 1.$$

This is the reason we cannot give a more convenient bound for the time spent checking the results of path-following. We can only state that the cost of certifying is bounded by :

$$n(\prod d_i)^2 \sum d_i \left((n+1)S(f)\max d_i + (n+1)^3\right) \times \\ \times P(\log\mu, \max d_i, \log n, \log S(f), \log\log H(f)) \tag{51}$$

Adding equations (50) and (51), we obtain the time bound of the Main theorem. This concludes the proof.

8. Computational matters.

A simplified version of the algorithms in this Thesis was implemented. It is available through anonymous ftp at *math.berkeley.edu* until Dec 31, 1993. Starting Jan 1, 1994, it will be available in another site, hopefuly *labma.ufrj.br*.

Some of the algorithms mentioned earlier in this text can be adapted to benefit from the architecture of modern computers. Three issues are extremely important :

- Overhead. Each time an evaluation of f or Df is performed, it is necessary to go through the data structure representing f. In some machines, this can be relatively time-consuming for big systems. Therefore, it is a good idea to follow a large number of paths at once, and to go through the data structure only once per evaluation, for all this group of paths.
- Pipelining. In order to take advantage of vector facilities, it is necessary to reduce the most time-consuming routines to certain vector operations. One possibility is, when following a large number of paths, to perform the arithmetic operations with vectors, each coordinate corresponding to one path.
- Parallelism. If several processors are available, there is a natural parallelization of the algorithm, that is to assign a group of paths to every processor. The algorithm is suitable to SIMD machines (Single Instruction Multiple Data) as well as to MIMD machines (Multiple Instruction Multiple Data). If there are also vector facilities available, there will be a trade-off between parallelism and pipe-lining.

APPENDIX A

Glossary of notations

$\|\cdot\|_2$	Vector and matrix 2-norm	
$\|\cdot\|_{\mathrm{f}}$	Frobenius norm of a matrix	
$\|\cdot\|_1$	Matrix 1 norm	
$\|\cdot\|_\infty$	Matrix sup norm	
M^{t}	Real transpose of Matrix M	
M^*	Complex transpose of Matrix M	
$\mathrm{diag}(a_i)$	Diagonal matrix A such that $A_{i,i} = a_i$	
$\mathrm{rank}M$	Rank of matrix M	
$\mathrm{Re}(z)$	Real part of z	
$\mathrm{Im}(z)$	Imaginary part of z	
$\mathcal{H}_d$	A space of systems of polynomial equations	1
n	In most places, number of equations.	1
$d = (d_1, \dots, d_n)$	Degree of the systems of polynomials	1
$H(x)$	Height of a rational point or polynomial	2
$S(f)$	the number of non-zero coefficients of f	2
D	Abreviation for $\max d_i$	8
$\|\cdot\|_{\mathrm{k}}$	Kostlan's unitary invariant norm	8
$d_{\mathrm{proj}}(f, g)$	Projective distance between f and g by $\|\cdot\|_{\mathrm{k}}$	8
N_V	Generalized Newton operator	9
V	Certain family of hyperplanes	9
β	Newton step length	9
γ	A kind of measure of the non-linearity of a mapping	9
α	Product of β and γ	9
N^{aff}	Newton operator in affine space	10
$d_{\mathrm{proj}}(x, y)$	Distance between points x and y in projective space	10
N^{proj}	Newton operator in projective space	11
N^{pseu}	Pseudo-Newton operator	12
$A^\dagger$	Moore-Penrose pseudo-inverse of matrix A	12

$\mu(f,x)$	Condition number of f at x	13
$\eta(f,x)$	A corrected norm of $f(x)$	13
$\mathrm{fl}(a \diamond b)$	Computed result of $a \diamond b$	33
ϵ_m	Machine epsilon	36
$\lVert\cdot\rVert_{\max}$	sup norm of a matrix, considered as a vector	41
$\lVert\cdot\rVert_{\max,2}$	A norm of the space of mappings of matrices into vectors	47
$R_u(f)$	u-resultant of a system f of n polynomials	61
$R(f)$	Resultant of a system f of $n+1$ polynomials	64
B	A different definition of height	68
C	A different definition of height	68
$\mathrm{discr}_v(f)$	v-discriminant of f	73
Σ^{aff}	Affine discriminant variety	73
Σ	Discriminant variety	73
$d(\Sigma)$	Complexity exponent of the discriminant variety Σ	77
$\mu(\Sigma)$	Condition number of the discriminant variety Σ	77
Σ_0	Real variety of paths with a degenerate system	78
$d(\Sigma_0)$	Complexity exponent of Σ_0	78
$\mu(\Sigma_0)$	Condition number of Σ_0	79
q	A polynomial in D, $S(f)$ and n.	88
U_d	Non-degenerate locus for global solving	95
μ_0	Conditioning of solving polynomial systems	95
d_0	Complexity exponent of solving polynomial systems	95

APPENDIX B

The Author's Address and e-mail

The author's permanent mailing address is :

Gregorio Malajovich

Departamento de Matematica Aplicada
Instituto de Matematica
Universidade Federal do Rio de Janeiro
Caixa Postal 68530
Rio de Janeiro, RJ
CEP 21945
BRASIL

gregorio@@labma.ufrj.br

Bibliography

[1] Eugene L. Allgower and Kurt Georg, Continuation and Path-Following. Preprint, Colorado State University, may 1992. *Acta Numerica* (1992) 1-64

[2] E. Anderson et alli, *LAPACK users' guide.* SIAM, Philadelphia, 1992.

[3] 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. *Bulletin of the AMS,* **21**, 1, July 1989.

[4] John Canny, *The complexity of robot motion planning*, MIT Press, Cambridge, Mass., 1988.

[5] Gene H. Golub and Charles F. Van Loan, *Matrix Computations.* The John Hopkins University Press, Baltimore and London, 1989, second printing, 1990.

[6] G. H. Hardy and E. M. Wright, *Theory of Numbers*, 4th edition, Oxford at the Clarendon Press, 1965.

[7] F. S. Macaulay, Some formulæ in elimination, *Proceedings of the London Mathematical Society*, Vol XXXV, 1903.

[8] Alexander Morgan *Solving polynomial systems using continuation for engineering and scientific purposes.* Prentice Hall Inc, Englewood Cliffs, New Jersey, 1987.

[9] Douglas M. Priest, *On properties of floating point arithmetics : numerical stability and cost of accurate computations.* Draft of PhD Thesis, Berkeley, 9 Nov 1992.

[10] James Renegar, On the worst-case arithmetic complexity of approximating zeros of systems of polynomials. *Siam J. on Computing,* **18**, 2 , 350-370, Apr 1989.

[11] Michael Shub, Some remarks on Bezout's Theorem and complexity theory. *in* M. Hirsch, J. Marsden and M. Shub (ed), *From Topology to Computation*, Proceedings of the Smalefest, 1993.

[12] Michael Shub and Steve Smale, On the Complexity of Bezout's Theorem I - Geometric aspects. *Journal of the AMS*, **6**, 2, Apr 1993.

[13] Michael Shub and Steve Smale, On the complexity of Bezout's Theorem II - Volumes and Probabilities. in: F. Eysette and A. Galligo, eds : *Computational Algebraic geometry.* Progress in Mathematics **109**, Birkhauser, 267-285, 1993.

[14] Michael Shub and Steve Smale, Complexity of Bezout's Theorem III ; Condition number and packing. *Journal of Complexity* **9**, 4-14, 1993.

[15] Michael Shub and Steve Smale, Complexity of Bezout's Theorem IV ; Probability of success ; Extensions Preprint, Berkeley, 1993.

[16] Steve Smale, Newton method estimates from data at one point, in R. Erwing, K. Gross and C. Martin (editors). *The merging of disciplines : New directions in Pure, Applied and Computational Mathematics.* Springer, New York, 1986

[17] B. L. Van der Waerden, *Modern Algebra, Vol 2.* F. Ungar publishing Co, New York, 1950

[18] J. H. Wilkinson, *The algebraic eigenvalue problem*, Clarendon Press, Oxford, 1965, Reprinted 1969.

[19] J. H. Wilkinson, Error analysis of direct methods of matrix inversion, *Journal of the ACM*, **8**, 281-330, 1961.