LE MATEMATICHE
Vol. LXVIII (2013) – Fasc. I, pp. 137–164
doi: 10.4418/2013.68.1.12

ZERO-DIMENSIONAL FAMILIES OF POLYNOMIAL SYSTEMS

LORENZO ROBBIANO - MARIA-LAURA TORRENTE

If a *real world* problem is modelled with a system of polynomial equations, the coefficients are in general not exact. The consequence is that *small* perturbations of the coefficients may lead to *big* changes of the solutions. In this paper we address the following question: how do the zeros change when the coefficients of the polynomials are perturbed? In the first part we show how to construct semi-algebraic sets in the parameter space over which the family of all ideals shares the number of isolated real zeros. In the second part we show how to modify the equations and get new ones which generate the same ideal, but whose real zeros are more stable with respect to perturbations of the coefficients.

1. Introduction

Systems of polynomial equations with imprecise coefficients arise in mathematical models of practical problems. Due to their relevance in many applications, several methods have been considered for determining their solutions, in particular their real solutions. All of them face a difficulty, i.e. the potentially erratic behaviour of the solutions with respect to small perturbations in the coefficients of the polynomials involved. Tackling this problem entails a preliminary analysis of the following question of an algebraic nature. Given a zero-dimensional

Entrato in redazione: 30 maggio 2012

AMS 2010 Subject Classification: Primary 13C40, Secondary 14M10, 65F35, 65H04.
Keywords: Algebraic family, Real zeros, Condition number.

system of polynomials with smooth zeros, how far can we perturb the coefficients so that their zeros remain smooth and the number of real zeros does not change? It is clear that constancy of smoothness and the number of zeros are essential if we want to consider the perturbation a good one.

To address these issues, we concentrate on systems having as many equations as indeterminates and a finite set of smooth solutions. The fact that every zero-dimensional smooth scheme can be represented in this way follows, for instance, from the Shape Lemma (see [18], Theorem 3.7.25). In our case the focus is not on the ideal but on the given set of generators. Our method, explained in Section 2, prescribes the embedding of the given polynomials into an algebraic family which is manufactured by substituting some coefficients of the polynomials with parameters. The key remark is that the family is a family of ideals which depend on the given set of polynomials, not on the ideal they generate.

Once we have a family, we can describe a good subset of the parameter space over which the members of the family share the property that their zero sets have the same number of smooth real points. This is the content of Section 2 where we describe a free (see Proposition 2.7), and a smooth (see Theorem 2.13) locus in the parameter space. Then we provide a suitable algorithm to compute what we call an I-optimal subscheme of the parameter space (see Corollary 2.17): it is a subscheme over which the schemes are smooth and have the same number of complex points. The last important result of Section 2 is Theorem 2.21 which treats the *real case* and proves the existence of an open non-empty semialgebraic subscheme of the I-optimal subscheme over which the number of real zeros is constant.

The second part of the paper starts with Section 3 where we focus on a problem closely connected to the one addressed in Section 2. It is well-known (see [5]) that for a linear system with n equations and n unknowns, the most stable situation occurs when the coefficient matrix is orthonormal. Is there an analogue to orthonormality when we deal with polynomial systems?

In numerical analysis the condition number of a problem measures the sensitivity of the solution to small changes in the input data, and so it reveals how numerically well-conditioned the problem is. There exists a huge body of results about condition numbers for various numerical problems, for instance the solution of a linear system, the problem of matrix inversion, the least squares problem, and the computation of eigenvalues and eigenvectors.

Regarding the condition numbers of polynomial systems, much has been done during the last few decades by several authors, mainly following the paper [23] of Shub and Smale which treated the case of zero-dimensional homogeneous polynomial systems and became the source of inspiration of several other papers, such as [4], [8], [9], [10], [20], [21].

After some preparatory results which use the results of Section 2, we introduce a local condition number (see Definition 3.14) and with its help we prove Theorem 3.15 which has the merit of fully generalizing a classical result in numerical linear algebra (see Remark 3.16).

The subsequent short Section 4 illustrates how to manipulate the equations in order to lower, and sometimes to minimize, the local condition number (see Proposition 4.1). Then we concentrate on the case of the matrix 2-norm and show how to achieve the minimum when the polynomials involved have equal degree (see Proposition 4.3). Finally, Section 5 presents examples which indicate that our approach is good, in particular we see that when the local condition number is lowered, indeed the corresponding solution is more stable.

A natural question is how to compare our numbers with those of other authors. In general the comparison is not easy since condition numbers simply provide upper bounds, however some remarks can be made, particularly in connection to the above mentioned work [23] of Shub and Smale.

A minus of our method is that our number is not invariant under orthonormal transformations, and that it does not take into account the univariate case since we are looking for orthogonality of tangents while in the univariate case there is only one derivative. A plus is that our definition generalizes the classical definition of condition number in numerical linear algebra. Another plus is that we are able to use methods from commutative algebra to certify the admissibility of a given perturbation of the data.

2. Families of Zero-Dimensional Polynomial Systems

Given a zero-dimensional polynomial system which defines a smooth scheme $\mathbb{X}$, we want to embed it into a family of zero-dimensional schemes and study when and how it can move inside the family. In particular, we study the locus of the parameter-space over which the fibers are smooth with the same number of points as $\mathbb{X}$, and we give special emphasis to the case of real points.

We start the section by recalling some definitions. The notation is borrowed from [18] and [19], in particular we let $x_1, \dots, x_n$ be indeterminates and let $\mathbb{T}^n$ be the monoid of the power products in the symbols $x_1, \dots, x_n$. Most of the times, for simplicity we use the notation $\mathbf{x} = x_1, \dots, x_n$. If K is a field, the multivariate polynomial ring $K[\mathbf{x}] = K[x_1, \dots, x_n]$ is denoted by P, and if $f_1(\mathbf{x}), \dots, f_k(\mathbf{x})$ are polynomials in P, the set $\{f_1(\mathbf{x}), \dots, f_k(\mathbf{x})\}$ is denoted by $\mathbf{f}(\mathbf{x})$ (or simply by $\mathbf{f}$).

Definition 2.1. The **polynomial system** defined by $\mathbf{f}(\mathbf{x})$ is denoted by $\mathbf{f}(\mathbf{x}) = 0$ (or simply by $\mathbf{f} = 0$). We say that the system is **zero-dimensional** if the ideal generated by $\mathbf{f}(\mathbf{x})$ is zero-dimensional (see [18], Section 3.7).

Definition 2.2. We let m be a positive integer and let $\mathbf{a} = (a_1, \ldots, a_m)$ be an m-tuple of indeterminates which play the role of parameters. If we pick k polynomials $F_1(\mathbf{a},\mathbf{x}), \ldots, F_k(\mathbf{a},\mathbf{x}) \in K[\mathbf{a},\mathbf{x}]$, the set $\{F_1(\mathbf{a},\mathbf{x}), \ldots, F_k(\mathbf{a},\mathbf{x})\}$ is denoted by $F(\mathbf{a},\mathbf{x})$. We let $F(\mathbf{a},\mathbf{x}) = 0$ be the corresponding **family of systems of equations parametrized by a** and call $I(\mathbf{a},\mathbf{x})$ the ideal generated by $F(\mathbf{a},\mathbf{x})$ in $K[\mathbf{a},\mathbf{x}]$. If the scheme of the **a**-parameters is $\mathcal{S}$, then there is a K-algebra homomorphism $\varphi : K[\mathbf{a}] \longrightarrow K[\mathbf{a},\mathbf{x}]/I(\mathbf{a},\mathbf{x})$ or, equivalently, a **morphism of schemes** $\Phi : \mathrm{Spec}(K[\mathbf{a},\mathbf{x}]/I(\mathbf{a},\mathbf{x})) \longrightarrow \mathcal{S}$.

Although it is not strictly necessary for the theory, for our applications it suffices to consider independent parameters. Here is the formal definition.

Definition 2.3. If $\mathcal{S} = \mathbb{A}^m_K$ and $I(\mathbf{a},\mathbf{x}) \cap K[\mathbf{a}] = (0)$, then the parameters **a** are said to be **independent** with respect to $F(\mathbf{a},\mathbf{x})$, or simply independent if the context is clear.

The first important step is to embed the system $\mathbf{f}(\mathbf{x}) = 0$ into a family, but we must be careful and exclude families of the following type.

Example 2.4. Consider the family $F(a,\mathbf{x}) = \{x_1(ax_2+1), x_2(ax_2+1)\}$. It specializes to a zero dimensional scheme only for $a = 0$ while the generic member is positive-dimensional.

Definition 2.5. Let $\mathbf{f}(\mathbf{x})$ be a set of polynomials in P so that $\mathbf{f}(\mathbf{x})$ defines a zero-dimensional scheme and let $F(\mathbf{a},\mathbf{x})$ be a family parametrized by m independent parameters **a**. We say that $F(\mathbf{a},\mathbf{x})$ (and similarly $K[\mathbf{a},\mathbf{x}]/I(\mathbf{a},\mathbf{x})$ and $\mathrm{Spec}(K[\mathbf{a},\mathbf{x}]/I(\mathbf{a},\mathbf{x}))$) is a **generically zero-dimensional family** which contains $\mathbf{f}(\mathbf{x})$, if $\mathbf{f}(\mathbf{x}) = 0$ is a member of the family and the generic member of the family is zero-dimensional.

A theorem called *generic flatness* (see [11], Theorem 14.4) prescribes the existence of a non-empty Zariski-open subscheme $\mathcal{U}$ of $\mathcal{S}$ over which the morphism $\Phi^{-1}(\mathcal{U}) \longrightarrow \mathcal{U}$ is *flat*. In particular, it is possible to explicitly compute a subscheme over which the morphism is free. To do this, Gröbner bases reveal themselves as a fundamental tool.

Definition 2.6. Let $\mathbf{f}(\mathbf{x})$ be a set of polynomials in P such that $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme and let $F(\mathbf{a},\mathbf{x})$ be a generically zero-dimensional family containing $\mathbf{f}(\mathbf{x})$. Let $\mathcal{S} = \mathbb{A}^m_K$ be the scheme of the independent **a**-parameters and let $\Phi : \mathrm{Spec}(K[\mathbf{a},\mathbf{x}]/I(\mathbf{a},\mathbf{x})) \longrightarrow \mathcal{S}$ be the associated morphism of schemes. A dense Zariski-open subscheme $\mathcal{U}$ of $\mathcal{S}$ such that $\Phi^{-1}(\mathcal{U}) \longrightarrow \mathcal{U}$ is free (flat, faithfully flat), is said to be an ***I*-free (*I*-flat, *I*−faithfully flat)** subscheme of $\mathcal{S}$ or simply an I-free (I-flat, I-faithfully flat) scheme.

Proposition 2.7. *With the above assumptions and notation, let $I(\mathbf{a},\mathbf{x})$ be the ideal generated by $F(\mathbf{a},\mathbf{x})$ in $K[\mathbf{a},\mathbf{x}]$, let σ be a term ordering on $\mathbb{T}^n$, let $G(\mathbf{a},\mathbf{x})$ be the reduced σ-Gröbner basis of the ideal $I(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}]$, let $d(\mathbf{a})$ be the least common multiple of all the denominators of the coefficients of the polynomials in $G(\mathbf{a},\mathbf{x})$, and let $T = \mathbb{T}^n \setminus \mathrm{LT}_\sigma(I(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}])$.*

(a) The open subscheme $\mathcal{U}_\sigma$ of $\mathbb{A}^m_K$ defined by $d(\mathbf{a}) \neq 0$ is I-free.

(b) The multiplicity of each fiber over $\mathcal{U}$ coincides with the cardinality of T.

Proof. The assumption that $F(\mathbf{a},\mathbf{x})$ is a generically zero-dimensional family implies that $\mathrm{Spec}\big(K(\mathbf{a})[\mathbf{x}]/I(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}]\big) \longrightarrow \mathrm{Spec}(K(\mathbf{a}))$ is finite, in other words that $K(\mathbf{a})[\mathbf{x}]/I(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}]$ is a finite-dimensional $K(\mathbf{a})$-vector space. A standard result in Gröbner basis theory (see for instance [18], Theorem 1.5.7) shows that the residue classes of the elements in T form a $K(\mathbf{a})$-basis of this vector space. We denote by $\mathcal{U}_\sigma$ the open subscheme of $\mathbb{A}^m_K$ defined by $d(\mathbf{a}) \neq 0$. For every point in $\mathcal{U}$, the given reduced Gröbner basis evaluates to the reduced Gröbner basis of the corresponding ideal. Therefore the leading term ideal is the same for all these fibers, and so is its complement T. If we denote by $K[\mathbf{a}]_{d(\mathbf{a})}$ the localization of $K[\mathbf{a}]$ at the element $d(\mathbf{a})$ and by $I(\mathbf{a},\mathbf{x})^e$ the extension of the ideal $I(\mathbf{a},\mathbf{x})$ to the ring $K[\mathbf{a}]_{d(\mathbf{a})}$, then $K[\mathbf{a}]_{d(\mathbf{a})}[\mathbf{x}]/I(\mathbf{a},\mathbf{x})^e$ turns out to be a free $K[\mathbf{a}]_{d(\mathbf{a})}$-module. So claim (a) is proved. Claim (b) follows immediately from (a). □

Remark 2.8. We collect here a few remarks about this proposition. First of all we observe that the term ordering σ can be chosen arbitrarily, but clearly the open set $\mathcal{U}_\sigma$ may vary. Even if we fix the term ordering, but change the generators of the ideal, the open set $\mathcal{U}_\sigma$ may vary. The following example illustrates this remark.

Example 2.9. We consider the polynomials $f_1, g \in K[x,y]$ where $f_1 = x^3 - y$, $g = x(x-1)(x+1)(x-2)(x+2)(x-3)(x+3)(x+13)(x^2+x+1)$. We let I be the ideal generated by $\{f_1, g\}$, and check that $I = (f_1, f_2)$ where $f_2 = xy^3 + 504x^2y - 183xy^2 + 14y^3 - 504x^2 + 650xy - 147y^2 - 468x + 133y$. It defines a zero-dimensional scheme and we embed it into the family $I(\mathbf{a},\mathbf{x}) = (ax^3 - y, g)$.

If we pick $\sigma = \mathtt{Lex}$ with $y > x$ and perform the computation as suggested by the proposition, we get the freeness of the family for all a. Instead, we get the freeness of the family $I(\mathbf{a},\mathbf{x}) = (ax^3 - y, f_2)$ for $a \neq 0$ (see a further discussion in Example 2.15).

If we pick $\sigma = \mathtt{Lex}$ with $x > y$ we get the freeness of the family for all $a \neq 0$.

Example 2.10. We let $P = \mathbb{C}[x]$, the univariate polynomial ring, and embed the ideal I generated by the following polynomial $x^2 - 3x + 2$ into the generically zero-dimensional family $F(\mathbf{a},x) = \{a_1x^2 - a_2x + a_3\}$. Such family is given by the canonical K-algebra homomorphism

$$\varphi : \mathbb{C}[\mathbf{a}] \longrightarrow \mathbb{C}[\mathbf{a},x]/(a_1x^2 - a_2x + a_3)$$

Let $\boldsymbol{\alpha} = (a_1,a_2,a_3) \in \mathbb{C}^3$. The family is zero dimensional for
$\{\boldsymbol{\alpha} \in \mathbb{C}^3 \mid a_1 \neq 0\} \cup \{\boldsymbol{\alpha} \in \mathbb{C}^3 \mid a_1 = 0,\ a_2 \neq 0\}$.
It represents two distinct smooth points for
$\{\boldsymbol{\alpha} \in \mathbb{C}^3 \mid a_1 \neq 0,\ a_2^2 - 4a_1a_3 \neq 0\}$.
It represents a smooth point for $\{\boldsymbol{\alpha} \in \mathbb{C}^3 \mid a_1 = 0,\ a_2 \neq 0\}$.
It is not zero-dimensional for $\{\boldsymbol{\alpha} \in \mathbb{C}^3 \mid a_1 = 0,\ a_2 = 0\}$.

From now on, we assume that K **has characteristic** 0. Moreover, Examples 2.9 and 2.10 motivate the following definition.

Definition 2.11. Let $\mathbf{f}(\mathbf{x})$ be a set of polynomials in P such that $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme and let $F(\mathbf{a},\mathbf{x})$ be a generically zero-dimensional family containing $\mathbf{f}(\mathbf{x})$. Let $\mathcal{S} = \mathbb{A}_K^m$ be the scheme of the independent $\mathbf{a}$-parameters and let $\Phi : \mathrm{Spec}(K[\mathbf{a},\mathbf{x}]/I(\mathbf{a},\mathbf{x})) \longrightarrow \mathcal{S}$ be the associated morphism of schemes. A dense Zariski-open subscheme $\mathcal{U}$ of $\mathcal{S}$ such that $\Phi^{-1}(\mathcal{U}) \longrightarrow \mathcal{U}$ is **smooth**, i.e. all the fibers of $\Phi^{-1}(\mathcal{U}) \longrightarrow \mathcal{U}$ are zero-dimensional and smooth, is said to be an ***I*-smooth** subscheme of $\mathcal{S}$ or simply an I-smooth scheme.

For instance in Example 2.10 we have the equality $\mathcal{S} = \mathbb{A}_{\mathbb{C}}^3$ and the following open set $\mathcal{U} = \{\boldsymbol{\alpha} \in \mathbb{C}^3 \mid a_1 \neq 0,\ a_2^2 - 4a_1a_3 \neq 0\}$ is I-smooth.

Remark 2.12. We observe that a dense I-smooth scheme may not exist. It suffices to consider the ideal $I = (x-1)^2$ embedded into the family $(x-a)^2$. In any event, a practical way to find one, if there is one, is via Jacobians, as we are going to show.

As anticipated in the introduction, at this point we restrict ourselves to the special and fundamental case where our zero-dimensional system $\mathbf{f}(\mathbf{x}) = 0$ is given by n equations, i.e. $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}),\dots,f_n(\mathbf{x})\}$.

Theorem 2.13. *Let $f_1(\mathbf{x}),\dots,f_n(\mathbf{x}) \in P$, let $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}),\dots,f_n(\mathbf{x})\}$ be such that $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme. Then let $F(\mathbf{a},\mathbf{x}) \in K[\mathbf{a},\mathbf{x}]$ be a generically zero-dimensional family containing $\mathbf{f}(\mathbf{x})$. Let $\mathcal{S} = \mathbb{A}_K^m$ be the scheme of the independent $\mathbf{a}$-parameters and let $I(\mathbf{a},\mathbf{x})$ be the ideal generated by $F(\mathbf{a},\mathbf{x})$ in $K[\mathbf{a},\mathbf{x}]$. Let $D(\mathbf{a},\mathbf{x}) = \det(\mathrm{Jac}_F(\mathbf{a},\mathbf{x}))$ be the determinant of the Jacobian matrix of $F(\mathbf{a},\mathbf{x})$ with respect to the indeterminates $\mathbf{x}$, let $J(\mathbf{a},\mathbf{x})$ be the ideal sum $I(\mathbf{a},\mathbf{x}) + (D(\mathbf{a},\mathbf{x}))$ in $K[\mathbf{a},\mathbf{x}]$, and let H be the ideal in $K[\mathbf{a}]$ defined by the equality $H = J(\mathbf{a},\mathbf{x}) \cap K[\mathbf{a}]$.*

(a) *There exists an I-smooth subscheme of $\mathcal{S}$ if and only if $H \neq (0)$.*

(b) *If $0 \neq h(\mathbf{a}) \in H$ then the open subscheme of $\mathcal{S}$ defined by the inequality $h(\mathbf{a}) \neq 0$ is I-smooth.*

Proof. To prove one implication of claim (a), and simultaneously claim (b), we assume that $H \neq (0)$ and let $0 \neq h(\mathbf{a}) \in H$. We have an equality of type $h(\mathbf{a}) = a(\mathbf{a},\mathbf{x})f(\mathbf{a},\mathbf{x}) + b(\mathbf{a},\mathbf{x})D(\mathbf{a},\mathbf{x})$ with $f(\mathbf{a},\mathbf{x}) \in I(\mathbf{a},\mathbf{x})$, and hence an equality $1 = \frac{a(\mathbf{a},\mathbf{x})}{h(\mathbf{a})}f(\mathbf{a},\mathbf{x}) + \frac{b(\mathbf{a},\mathbf{x})}{h(\mathbf{a})}D(\mathbf{a},\mathbf{x})$ in $J(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}]$. For every $\boldsymbol{\alpha} \in \mathcal{S}$ such that $h(\boldsymbol{\alpha}) \neq 0$ the equality implies that the corresponding scheme has no common zeros with the determinant of its Jacobian matrix, hence it is smooth. Conversely, assume that $H = (0)$. Then the canonical K-algebra homomorphism $K[\mathbf{a}] \longrightarrow K[\mathbf{a},\mathbf{x}]/J(\mathbf{a},\mathbf{x})$ is injective and hence it induces a morphism $\operatorname{Spec}\big(K[\mathbf{a},\mathbf{x}]/J(\mathbf{a},\mathbf{x})\big) \longrightarrow \mathbb{A}^m_K$ which is dominant. Hence, for a generic point of $\mathbb{A}^m_K$, the scheme $\operatorname{Spec}\big(K[\mathbf{a},\mathbf{x}]/J(\mathbf{a},\mathbf{x})\big)$ is not empty and so the affine scheme $\operatorname{Spec}\big(K[\mathbf{a},\mathbf{x}]/I(\mathbf{a},\mathbf{x})\big)$ is not smooth. □

The following example illustrates these results.

Example 2.14. Let us consider the polynomials $f_1 = x_1^2 + x_2^2 - 1$, $f_2 = x_2^2 + x_1$ in $\mathbb{C}[x_1,x_2]$ and the ideal $I = (f_1, f_2)$ generated by them. We embed it into $I(\mathbf{a},\mathbf{x}) = (x_1^2 + a_1x_2^2 - 1,\ x_2^2 + a_2x_1)$. It is a free family over $\mathbb{A}^2_{\mathbb{C}}$, and the multiplicity of each fiber is 4. We compute $D(\mathbf{a},\mathbf{x}) = \det(\mathrm{Jac}_F(\mathbf{a},\mathbf{x}))$ and get the equality $D(\mathbf{a},\mathbf{x}) = -2a_1a_2x_2 + 4x_1x_2$. We let

$$J(\mathbf{a},\mathbf{x}) = I(\mathbf{a},\mathbf{x}) + (D(\mathbf{a},\mathbf{x})) = (x_1^2 + a_1x_2^2 - 1,\ x_2^2 + a_2x_1,\ -2a_1a_2x_2 + 4x_1x_2)$$

A computation with CoCoA of `Elim`$([x_1,x_2],J)$ yields $(\frac{1}{2}a_1^2a_2^3 + 2a_2)$, and hence $J(\mathbf{a},\mathbf{x}) \cap K[\mathbf{a}] = (\frac{1}{2}a_1^2a_2^3 + 2a_2)$. According to the theorem, if $\mathcal{U}$ is the complement in $\mathbb{A}^2_{\mathbb{C}}$ of the curve defined by $\frac{1}{2}a_1^2a_2^3 + 2a_2 = 0$, then $\mathcal{U}$ is an I-smooth subscheme of $\mathbb{A}^2_{\mathbb{C}}$. On the other hand, the curve has three components, $a_2 = 0$, and $a_1a_2 \pm 2i = 0$. If $a_2 = 0$ then the corresponding ideal is $(x_1^2 - 1, x_2^2)$ which is not smooth. If we have $a_1a_2 \pm 2i = 0$, then the corresponding ideals are $(x_1^2 \mp \frac{2i}{a_2}x_2^2 - 1,\ x_2^2 + a_2x_1)$ which can be written as $((x_1 \pm i)^2,\ x_2^2 + a_2x_1)$ and hence are not smooth.

Let us now consider the set $\{f_1, f_2\}$ where $f_1 = x_1^2 + x_2^2$, $f_2 = x_2^2 + x_1$. We embed it into the family $I(\mathbf{a},\mathbf{x}) = (x_1^2 - a_1x_2^2,\ x_2^2 + a_2x_1)$. As before, we check that it is a free family over $\mathbb{A}^2_{\mathbb{C}}$, and the multiplicity of each fiber is 4. We compute $D(\mathbf{a},\mathbf{x}) = \det(\mathrm{Jac}_F(\mathbf{a},\mathbf{x}))$ and get $D(\mathbf{a},\mathbf{x}) = 2a_1a_2x_2 + 4x_1x_2$. The computation with CoCoA of `Elim`$([x,y],J)$ yields (0), and hence there is no subscheme of $\mathbb{A}^2_K$ which is I-smooth. Indeed, for $a_2 \neq 0$ we have $I(\mathbf{a},\mathbf{x}) = (x_1 + \frac{1}{a_2}x_2^2,\ \frac{1}{a_2^2}x_2^4 - a_1x_2^2)$ which is not smooth. Incidentally, we observe that also for $a_2 = 0$ the corresponding zero-dimensional scheme is not smooth.

The following example illustrates other subtleties related to the theorem.

Example 2.15. (Example 2.9 continued)
We consider the family $I(\mathbf{a},\mathbf{x}) = (ax^3 - y, f_2)$ for $a \neq 0$ of Example 2.9, compute $D(\mathbf{a},\mathbf{x}) = \det(\mathrm{Jac}_F(\mathbf{a},\mathbf{x}))$ and get $D(\mathbf{a},\mathbf{x}) = 9ax^3y^2 + 1512ax^4 - 1098ax^3y + 126ax^2y^2 + 1950ax^3 - 882ax^2y + y^3 + 399ax^2 + 1008xy - 183y^2 - 1008x + 650y - 468$. We let $J(\mathbf{a},\mathbf{x}) = I(\mathbf{a},\mathbf{x}) + (D(\mathbf{a},\mathbf{x}))$ and get $J(\mathbf{a},\mathbf{x}) \cap K[\mathbf{a}] = (h(\mathbf{a}))$ where

$$h(\mathbf{a}) = a^9 - \frac{738170716516748}{7749152384519}a^8 + \frac{21803946383594456350074\,6}{91409877182005574647}a^7$$

$$-\frac{1665570115630099814740616 68}{3135358787342791210392 1}a^6 - \frac{2761692608914197508465522 07}{31353587873427912103921}a^5$$

$$+\frac{98680911599871901908167889 6}{31353587873427912103921}a^4 - \frac{6324760741392623787151795 2}{31353587873427912103921}a^3$$

$$-\frac{1316764479863922379654192128}{31353587873427912103921}a^2 + \frac{31787255080429647770419 2}{13058553883143653521}a - \frac{974975584016793600000}{266501099655992929}$$

Therefore, if $\mathcal{U}$ denotes the complement in $\mathbb{A}^1_K$ of the zeros of $h(a)$, the theorem says that it is a Zariski-open I-smooth subscheme. However, we have already seen in Example 2.9 that $a = 0$ (the origin is in $\mathcal{U}$) is not in the free locus: we observe that the corresponding scheme is smooth, but it has only two points. The other subtlety is that the Bézout number of the family is $3 \times 4 = 12$, but if we substitute $y = ax^3$ into f_2 we get a univariate polynomial of degree 10. The two *missing* points are at infinity. No member of the family represents twelve points. The final remark is that if we move the parameter a inside the locus described by $a \cdot h(a) \neq 0$ we always get a smooth scheme of 10 points. If $K = \mathbb{C}$ the ten points have complex coordinates, some of them are real, but there are no values of a for which all the 10 points are real. The reason is that if $r_1 = \frac{-1+\sqrt{3}i}{2}$, $r_2 = \frac{-1-\sqrt{3}i}{2}$ are the two complex roots of $x^2 + x + 1 = 0$, then two of the ten points are (r_1, r_1^3), (r_2, r_2^3) which are not real points (see Theorem 2.21 and Example 2.23).

An alternative method to construct the smooth locus of the family is obtained by using the discriminant, as explained for instance in the book [14].

Combining Theorem 2.13 and Proposition 2.7 we get a method to select a Zariski-open subscheme of the parameter space over which all the fibers are smooth schemes of constant multiplicity (see [24] for similar results). Before describing the algorithm, we need a definition which captures this concept.

Definition 2.16. With the above notation, a dense Zariski-open subscheme $\mathcal{U}$ of $\mathcal{S}$ such that $\Phi^{-1}(\mathcal{U}) \longrightarrow \mathcal{U}$ is smooth and free is said to be an ***I*-optimal** subscheme of $\mathcal{S}$.

Corollary 2.17. *Let $\mathcal{S} = \mathbb{A}^m_K$ and consider the following sequence of instructions.*

(1) Compute $D(\mathbf{a},\mathbf{x}) = \det(\mathrm{Jac}_F(\mathbf{a},\mathbf{x}))$.

(2) Let $J(\mathbf{a},\mathbf{x}) = I(\mathbf{a},\mathbf{x}) + (D(\mathbf{a},\mathbf{x}))$ *and compute* $H = J(\mathbf{a},\mathbf{x}) \cap K[\mathbf{a}]$.

(3) If $H = (0)$ *return* "There is no I-smooth subscheme of $\mathbb{A}^m_K$" *and stop.*

(4) Choose $h(\mathbf{a}) \in H \setminus 0$ *and let* $\mathcal{U}_1 = \mathbb{A}^m_K \setminus \{\boldsymbol{\alpha} \in \mathbb{A}^m_K \mid h(\boldsymbol{\alpha}) = 0\}$.

(5) Choose a term ordering σ on $\mathbb{T}^n$ and compute the reduced σ-Gröbner basis $G(\mathbf{a},\mathbf{x})$ *of* $I(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}]$

(6) Let $T = \mathbb{T}^n \setminus \mathrm{LT}_\sigma(I(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}])$, *compute the cardinality of T and call it μ; then compute the least common multiple of all the denominators of the coefficients of the polynomials in $G(\mathbf{a},\mathbf{x})$, and call it $d(\mathbf{a})$; finally, let* $\mathcal{U}_2 = \mathbb{A}^m_K \setminus \{\boldsymbol{\alpha} \in \mathbb{A}^m_K \mid d(\boldsymbol{\alpha}) \neq 0\}$ *and let* $\mathcal{U} = \mathcal{U}_1 \cap \mathcal{U}_2$.

(7) Return $\mathcal{U}_1$, $\mathcal{U}_2$, $\mathcal{U}$, T, μ.

This is an algorithm which returns $\mathcal{U}_1$ which is I-smooth, $\mathcal{U}_2$ which is I-free, $\mathcal{U}$ which is I-optimal, T which provides a basis as K-vector spaces of all the fibers over $\mathcal{U}_2$, and μ which is the multiplicity of all the fibers over $\mathcal{U}_2$.

Proof. It suffices to combine Theorem 2.13 and Proposition 2.7. □

Example 2.18. We consider the ideal $I = (f_1, f_2)$ of $K[x,y]$ where $f_1 = xy - 6$, $f_2 = x^2 + y^2 - 13$. We embed it into the family $I(\mathbf{a},\mathbf{x}) = (a_1xy + a_2,\ a_3x^2 + a_4y^2 + a_5)$. We compute the reduced `DegRevLex`-Gröbner basis of the ideal $I(\mathbf{a},\mathbf{x})K(\mathbf{a})[\mathbf{x}]$ and get

$$\{x^2 + \tfrac{a_4}{a_3}y^2 + \tfrac{a_5}{a_3},\ \ xy + \tfrac{a_2}{a_1},\ \ y^3 - \tfrac{a_2a_3}{a_1a_4}x + \tfrac{a_1a_5}{a_1a_4}y\}$$

according to the above results, a free locus is given by $a_1a_3a_4 \neq 0$. Now we compute $D(\mathbf{a},\mathbf{x}) = \det(\mathrm{Jac}_F(\mathbf{a},\mathbf{x}))$ and get $D(\mathbf{a},\mathbf{x}) = -2a_1a_3x^2 + 2a_1a_4y^2$.

We let $J(\mathbf{a},\mathbf{x}) = I(\mathbf{a},\mathbf{x}) + (D(\mathbf{a},\mathbf{x}))$ and compute $J(\mathbf{a},\mathbf{x}) \cap K[\mathbf{a}]$. We get the principal ideal generated by $a_2^2a_3a_4 - \frac{1}{4}a_1^2a_5^2$. In conclusion, an I-optimal subscheme is $\mathcal{U} = \mathbb{A}^5_K \setminus F$ where F is the closed subscheme defined by the equation $a_1a_3a_4(a_2^2a_3a_4 - \frac{1}{4}a_1^2a_5^2) = 0$, and $\mu = 4$.

Definition 2.19. We say that **a point is complex** if its coordinates are complex numbers, and we say that **a point is real** if its coordinates are real numbers.

The following example illustrates the fact that even if we start with a set of real points, a zero-dimensional complete intersection which contains them may also contain complex non-real points.

Example 2.20. Let $\mathbb{X}$ be the set of the 10 real points $\{(-1,-1),(2,8),(-2,-8),$ $(3,27)$, $(-3,-27)$, $(4,64)$, $(5,125)$, $(-5,-125)$, $(6,216)$, $(-6,-216)\}$.

A zero-dimensional scheme containing $\mathbb{X}$ is defined by $\mathbf{f} = \{f_1, f_2\}$ where $f_1 = y - x^3$ and $f_2 = x^2y^2 - 1/4095y^4 + 1729/15x^2y - 74/15xy^2 + 1/15y^3 - 8832/5x^2 + 5852/15xy - 10754/315y^2 + 2160x - 4632/5y + 250560/91$. Let I denote the vanishing ideal of the 10 points and let J denote the ideal generated by $\mathbf{f}$. The colon ideal $J : I$ defines the residual intersection. Since J is the intersection of a cubic and a quartic curve, the residual intersection is a zero-dimensional scheme of multiplicity 2. Indeed, a computation (performed with CoCoA) shows that $J : I$ is generated by $(x + 1/78y - 87/26, y^2 - 756y + 658503)$. Since $756^2 - 4 \cdot 658503 = -2062476 < 0$, the two extra points on the zero-dimensional complete intersection are complex, non real points.

Theorem 2.21. *Let $f_1(\mathbf{x}),\dots,f_n(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$, let $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}),\dots,f_n(\mathbf{x})\}$ such that $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme. Then let $F(\mathbf{a},\mathbf{x}) \in \mathbb{R}[\mathbf{a},\mathbf{x}]$ be a generically zero-dimensional family containing* $\mathbf{f}(\mathbf{x})$. *Assume that there exists an I-optimal subscheme $\mathcal{U}$ of $\mathbb{A}^m_{\mathbb{R}}$, and let $\boldsymbol{\alpha}_I \in \mathcal{U}$ be the point in the parameter space which corresponds to I. If $\mu_{\mathbb{R},I}$ is the number of distinct real points in the fiber over $\boldsymbol{\alpha}_I$ (i.e. zeroes of I), then there exists an open semi-algebraic subscheme $\mathcal{V}$ of $\mathcal{U}$ such that for every $\boldsymbol{\alpha} \in \mathcal{V}$ the number of real points in the fiber over $\boldsymbol{\alpha}$ is $\mu_{\mathbb{R},I}$.*

Proof. We consider the ideal $\mathcal{I} = I(\mathbf{a},\mathbf{x})\mathbb{R}(\mathbf{a})[\mathbf{x}]$. It is zero-dimensional and the field $\mathbb{R}(\mathbf{a})$ is infinite. Since a linear change of coordinates does not change the problem, we may assume that $\mathcal{I}$ is in x_n-normal position (see [18], Section 3.7). Moreover, we have already observed (see Remark 2.8) that in Proposition 2.7 the choice of σ is arbitrary. We choose $\sigma = \texttt{Lex}$ and hence the reduced Lex-Gröbner basis of $\mathcal{I}$ has the shape prescribed by the Shape Lemma (see [18] Theorem 3.7.25). Therefore there exists a univariate polynomial $h_{\mathbf{a}} \in \mathbb{R}(\mathbf{a})[x_n]$ whose degree is the multiplicity of both the generic fiber and the fiber over $\boldsymbol{\alpha}_I$, which is the number of complex zeros of I. Due to the shape of the reduced Gröbner basis, a point is real if and only if its x_n-coordinate is real. Therefore it suffices to prove the following statement: given a univariate square-free polynomial $h_{\mathbf{a}} \in \mathbb{R}(\mathbf{a})[x_n]$ such that $h_{\boldsymbol{\alpha}_I}$ has exactly $\mu_{\mathbb{R},I}$ real roots, there exists an open semi-algebraic subset of $\mathbb{A}^m_{\mathbb{R}}$ such that for every point $\boldsymbol{\alpha}$ in it, the polynomial $h_{\boldsymbol{\alpha}}$

has exactly $\mu_{\mathbb{R},I}$ real roots. The correctedness of this statement follows from [3], Theorem 5.12 where it is shown that for every root there exists an open semi-algebraic set in $\mathbb{A}^m_{\mathbb{R}}$ which isolates the root. Since complex non-real roots have to occur in conjugate pairs, this implies that real roots stay real. □

Let us see some examples.

Example 2.22. We consider the ideal $I = (xy-2y^2+2y,\ x^2-y^2-2x)$ in $\mathbb{R}[x,y]$, and we embed it into the family $I(\mathbf{a},\mathbf{x}) = (xy-ay^2+ay,\ x^2-y^2-2x)$. We compute the reduced Lex-Gröbner basis of $I(\mathbf{a},\mathbf{x})\mathbb{R}(\mathbf{a})[\mathbf{x}]$ and get

$$\{x^2-2x-y^2,\ xy-ay^2+ay,\ y^3-\tfrac{2a}{a-1}y^2+\tfrac{a^2+2a}{a^2-1}y\}$$

Applying the algorithm illustrated in Corollary 2.17 we get an I-smooth subscheme of $\mathbb{A}^1_{\mathbb{R}}$ for $a(a+2)\neq 0$, and an I-free subscheme for $(a-1)(a+1)\neq 0$. For a different from $0,-2,\ 1,-1$ we have an I-optimal subscheme and the multiplicity is 4.

Our ideal I is obtained for $a=2$, and hence it lies over the optimal subscheme. It has multiplicity 4 and the four zeros are real.

The computed Lex-Gröbner basis does not have the shape prescribed by the Shape Lemma, so we perform a linear change of coordinates by setting $x = x+y,\ y = x-y$. We compute the reduced Lex-Gröbner basis and get

$$\{x+4\tfrac{a+1}{a-1}y^3-2\tfrac{a+1}{a-1}y^2-\tfrac{3a+1}{a-1}y,\ \ y^4-y^3-\tfrac{1}{2}\tfrac{a}{a+1}y^2+\tfrac{1}{2}\tfrac{a}{a+1}y\}$$

It has the good shape, so we can use the polynomial

$$h_{\mathbf{a}} = y^4-y^3-\tfrac{1}{2}\tfrac{a}{a+1}y^2+\tfrac{1}{2}\tfrac{a}{a+1}y = y(y-1)(y^2-\tfrac{1}{2}\tfrac{a}{a+1})$$

We get the following result.

- For $a<-1,\ a\neq -2$ there are 4 real points.
- For $-1<a<0$ there are 2 real points.
- For $a>0,\ a\neq 1$ there are 4 real points.

To complete our analysis, let us see what happens at the *bad* points $0,-2,\ 1,-1$.

At 0 the primary decomposition of the ideal I_0 is $(x-2,y)\cap(y^2+2x,xy,x^2)$, hence the fiber consists in the simple point $(2,0)$ and a triple point at $(0,0)$.

At -2 we see that $(x+\frac{2}{3},\ y-\frac{4}{3})\cap(x,y)\cap(x-2,y^2)$ is the primary decomposition of the ideal I_{-2}, and hence the fiber consists in the simple point $(-\frac{2}{3},\frac{4}{3})$, the simple point $(0,0)$ and a double point at $(2,0)$.

At -1 the primary decomposition of the ideal I_{-1} is $(x,y)\cap(x-2,y)$, hence the fiber consists of the two simple real points $(0,0)$ and $(2,0)$.

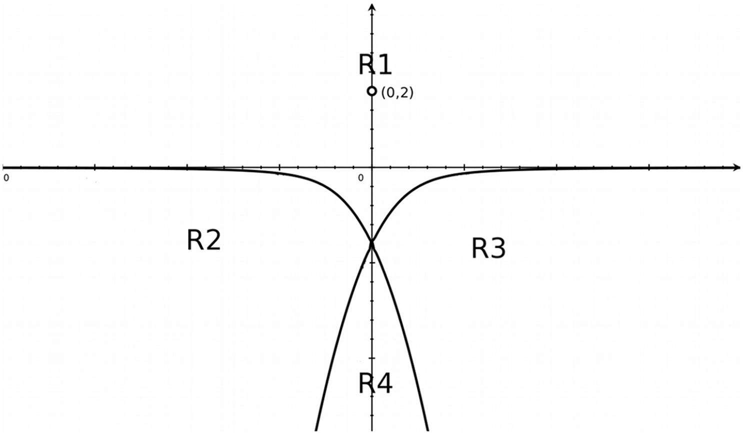
At 1 we see that $(x,y)\cap(x-2,y)\cap(x+\frac{1}{4},y-\frac{3}{4})$ is the primary decomposition of the ideal I_1, hence the fiber consists of the three simple real points $(0,0)$, $(2,0)$, $(-\frac{1}{4},\frac{3}{4})$.

Example 2.23. We consider the ideal $I=(xy+1,\ x^2+y^2-5)$ in $\mathbb{R}[x,y]$, and we embed it into the family $I(\mathbf{a},x,y)=(xy+a_1x+1,\ x^2+y^2+a_2)$. We compute the reduced Lex-Gröbner basis of $I(\mathbf{a},\mathbf{x})K(\mathbf{a})[x,y]$ and get $G(\mathbf{a},x,y)=\{g_1,g_2\}$ where

$$\begin{aligned} g_1 &= x-y^3-a_1y^2-a_2y-a_1a_2, \\ g_2 &= y^4+2a_1y^3+(a_1^2+a_2)y^2+2a_1a_2y+(a_1^2a_2+1) \end{aligned}$$

which has the shape prescribed by the Shape Lemma (see [18] Theorem 3.7.25). There is no condition for the free locus, and $D(\mathbf{a},x,y)=\det(\mathrm{Jac}_F(\mathbf{a},x,y))=-2x^2+2y^2+2a_1y$. We let $J(\mathbf{a},x,y)=I(\mathbf{a},x,y)+(D(\mathbf{a},x,y))$ and compute the intersection $J(\mathbf{a},x,y)\cap K[\mathbf{a}]$. We get the principal ideal generated by the following polynomial $h(\mathbf{a})=a_1^6a_2+3a_1^4a_2^2+a_1^4+3a_1^2a_2^3+20a_1^2a_2+a_2^4-8a_2^2+16$. An I-optimal subscheme is $\mathcal{U}=\mathbb{A}^4_{\mathbb{R}}\setminus F$ where F is the closed subscheme defined by the equation $h(\mathbf{a})=0$, and we observe that $\mu=4$.

At this point we know that for $h(\mathbf{a})\neq 0$ each fiber is smooth and has multiplicity 4, hence it consists of 4 distinct complex points. What about real points?



The real curve defined by $h(\mathbf{a})=0$ is shown in the above picture. It is the union of two branches and the isolated point $(0,2)$. The upper region R1 (with the exception of the point $(0,2)$) corresponds to the ideals in the family whose zeros are four complex non-real points. The regions R2 and R3 correspond to the ideals whose zeros are two complex non-real points and two real points. The region R4 corresponds to the ideals whose zeros are four real points. To describe the four regions algebraically, we use the Sturm-Habicht sequence (see [15]) of $g_2 \in \mathbb{R}(\mathbf{a})[y]$. The leading monomials are y^4, $4y^3$, $4r(\mathbf{a})y^2$, $-8\ell(\mathbf{a})y$, $16h(\mathbf{a})$ where we have $r(\mathbf{a}) = a_1^2 - 2a_2$, $\ell(\mathbf{a}) = a_1^4 a_2 + 2a_1^2 a_2^2 + 2a_1^2 + a_2^3 - 4a_2$. To get the total number of real roots we count the sign changes in the sequence at $-\infty$ and $+\infty$; in particular, we observe that in the parameter space the ideal I corresponds to the point $(0,-5)$ which belongs to the region R4. We get

$$\mathrm{R4} = \{\boldsymbol{\alpha} \in \mathbb{R}^2 \mid r(\boldsymbol{\alpha}) > 0,\ \ell(\boldsymbol{\alpha}) < 0,\ h(\boldsymbol{\alpha}) > 0\}$$

which is semi-algebraic open, not Zariski-open.

3. Condition Numbers

In this section we introduce a notion of *condition number* for zero-dimensional polynomial systems of $\mathbb{R}[\mathbf{x}]$ which define a smooth scheme; the aim is to give a measure of the sensitivity of its real roots with respect to small perturbations of the input data, that is small changes of the coefficients of the involved polynomials.

The section starts with the recall of well-known facts about numerical linear algebra. We let m,n be positive integers and let $\mathrm{Mat_{m\times n}}(\mathbb{R})$ be the set of $m \times n$ matrices with entries in $\mathbb{R}$; if $m = n$ we simply write $\mathrm{Mat_n}(\mathbb{R})$.

Definition 3.1. Let $M = (m_{ij})$ be a matrix in $\mathrm{Mat_{m\times n}}(\mathbb{R})$, $v = (v_1,\dots,v_n)$ a vector in $\mathbb{R}^n$ and $\|\cdot\|$ a vector norm.

(a) Let $r \geq 1$ be a real number; the r**-norm** on the vector space $\mathbb{R}^n$ is defined by the formula $\|v\|_r = (\sum_{i=1}^n |v_i|^r)^{\frac{1}{r}}$ for every $v \in \mathbb{R}^n$.

(b) The **infinity norm** on $\mathbb{R}^n$ is defined by the formula $\|v\|_\infty = \max_i |v_i|$.

(c) The **spectral radius** $\rho(M)$ of the matrix M is defined by the formula $\rho(M) = \max_i |\lambda_i|$, where the λ_i are the *complex* eigenvalues of M.

(d) The real function defined on $\mathrm{Mat_{m\times n}}(\mathbb{R})$ by $M \mapsto \max_{\|v\|=1} \|Mv\|$ is a matrix norm called the **matrix norm induced** by $\|\cdot\|$. A matrix norm induced by a vector norm is called an **induced matrix norm**.

(e) The matrix norm induced by $\|\cdot\|_1$ is given by the following formula $\|M\|_1 = \max_j(\sum_i |m_{ij}|)$. The matrix norm induced by $\|\cdot\|_\infty$ is given by the formula $\|M\|_\infty = \max_i(\sum_j |m_{ij}|)$. Finally, the matrix norm induced by $\|\cdot\|_2$ is given by the formula $\|M\|_2 = \max_i(\sigma_i)$ where the σ_i are the singular values of M.

If no confusion arises, from now on we will use the symbol $\|\cdot\|$ to denote both a vector norm and a matrix norm. We recall some facts about matrix norms.

Proposition 3.2. *Let M be a matrix in $\mathrm{Mat_n}(\mathbb{R})$, let I be the identity matrix of size n and let $\|\cdot\|$ be an induced matrix norm on $\mathrm{Mat_n}(\mathbb{R})$. If the matrix $I+M$ is invertible then $(1-\|M\|)\,\|(I+M)^{-1}\| \leq 1$.*

Proof. See [5], Theorem 3.13. □

Proposition 3.3. *Let $M \in \mathrm{Mat_{m\times n}}(\mathbb{R})$ and denote by M_i the i-th row of M. Let $r_1 \geq 1, r_2 \geq 1$ be real numbers such that $\frac{1}{r_1}+\frac{1}{r_2}=1$; then*

$$\max_i \|M_i\|_{r_2} \leq \|M\|_{r_1} \leq m^{1/r_1} \max_i \|M_i\|_{r_2}$$

In particular, for $r_1 = r_2 = 2$

$$\max_i \|M_i\|_2 \leq \|M\|_2 \leq \sqrt{m}\max_i \|M_i\|_2$$

Proof. See [16], inequality (6.13). □

This introductory part ends with the recollection of some facts about the polynomial ring $K[\mathbf{x}]$. In particular, given $\eta = (\eta_1,\dots,\eta_n) \in \mathbb{N}^n$ we denote by $|\eta|$ the number $\eta_1+\dots+\eta_n$, by $\eta!$ the number $\eta_1!\dots\eta_n!$, and by $\mathbf{x}^\eta$ the power product $x_1^{\eta_1}\dots x_n^{\eta_n}$.

Definition 3.4. Let p be a point of K^n; the K-linear map on $K[\mathbf{x}]$ defined by $f \mapsto f(p)$ is called the **evaluation map** associated to p.

Definition 3.5. Let d be a nonnegative integer, let $r \geq 1$ be a real number, let p be a point of $\mathbb{R}^n$ and let $g(\mathbf{x})$ be a polynomial in $\mathbb{R}[\mathbf{x}]$.

(a) The formal Taylor expansion of $g(\mathbf{x})$ at p is given by the following expression: $g(\mathbf{x}) = \sum_{|\eta|\geq 0} \frac{1}{\eta!}\frac{\partial^\eta g}{\partial \mathbf{x}^\eta}(p)(\mathbf{x}-p)^\eta$.

(b) The polynomial $\sum_{|\eta|\geq d} \frac{1}{\eta!}\frac{\partial^\eta g}{\partial \mathbf{x}^\eta}(p)(\mathbf{x}-p)^\eta$ is denoted by $g^{\geq d}(\mathbf{x},p)$.

(c) The ***r*-norm of** $g(\mathbf{x})$ **at** p is defined as the r-norm of the vector $\frac{\partial g}{\partial \mathbf{x}}(p)$. If $\|\frac{\partial g}{\partial \mathbf{x}}(p)\|_r = 1$ then $g(\mathbf{x})$ is called **unitary at** p.

We use the following formulation of Taylor's theorem.

Proposition 3.6. *Let p be a point of $\mathbb{R}^n$ and let $g(\mathbf{x})$ be a polynomial in $\mathbb{R}[\mathbf{x}]$. For every point $q \in \mathbb{R}^n$ we have*

$$g(q) = g(p) + \mathrm{Jac}_g(p)(q-p) + \frac{1}{2}(q-p)^t H_g(\xi)(q-p)$$

where ξ is a point of the line connecting p to q and $H_g(\xi)$ is the Hessian matrix of g at ξ.

Let $f_1(\mathbf{x}), \dots, f_n(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$ and let $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$ so that the ideal $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme; we introduce a notion of admissible perturbation of $\mathbf{f}(\mathbf{x})$. Roughly speaking, the polynomial set $\boldsymbol{\varepsilon}(\mathbf{x}) = \{\varepsilon_1(\mathbf{x}), \dots, \varepsilon_n(\mathbf{x})\} \subset \mathbb{R}[\mathbf{x}]$ is considered to be an admissible perturbation of $\mathbf{f}(\mathbf{x})$ if the real solutions of $(\mathbf{f}+\boldsymbol{\varepsilon})(\mathbf{x}) = 0$ are nonsingular and derive from perturbations of the real solutions of $\mathbf{f}(\mathbf{x}) = 0$. Using the results of Section 2 we formalize this concept as follows.

Definition 3.7. Let $f_1(\mathbf{x}), \dots, f_n(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$ and let $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$ such that $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme; let $\mu_{\mathbb{R},I}$ be the number of real solutions of $\mathbf{f}(\mathbf{x}) = 0$, and let $\boldsymbol{\varepsilon}(\mathbf{x}) = \{\varepsilon_1(\mathbf{x}), \dots, \varepsilon_n(\mathbf{x})\}$ be a set of polynomials in $\mathbb{R}[\mathbf{x}]$. Suppose that the assumptions of Theorem 2.21 are satisfied, let $\mathcal{V} \subset \mathbb{A}^m_{\mathbb{R}}$ be an open semi-algebraic subset of $\mathcal{U}$ such that $\boldsymbol{\alpha}_I \in \mathcal{V}$, and for every $\boldsymbol{\alpha} \in \mathcal{V}$ the number of real roots of $F(\boldsymbol{\alpha}, \mathbf{x}) = 0$ is equal to $\mu_{\mathbb{R},I}$. If there exists $\boldsymbol{\alpha} \in \mathcal{V}$ such that $(\mathbf{f}+\boldsymbol{\varepsilon})(\mathbf{x}) = F(\boldsymbol{\alpha}, \mathbf{x})$, then $\boldsymbol{\varepsilon}(\mathbf{x})$ is called an **admissible perturbation** of $\mathbf{f}(\mathbf{x})$.

Henceforth we let $\boldsymbol{\varepsilon}(\mathbf{x}) = \{\varepsilon_1(\mathbf{x}), \dots, \varepsilon_n(\mathbf{x})\}$ be an admissible perturbation of $\mathbf{f}(\mathbf{x})$, and let $\mathcal{Z}_{\mathbb{R}}(\mathbf{f}) = \{p_1, \dots, p_{\mu_{\mathbb{R},I}}\}$, $\mathcal{Z}_{\mathbb{R}}(\mathbf{f}+\boldsymbol{\varepsilon}) = \{r_1, \dots, r_{\mu_{\mathbb{R},I}}\}$ be the sets of real solutions of $\mathbf{f}(\mathbf{x}) = 0$ and $(\mathbf{f}+\boldsymbol{\varepsilon})(\mathbf{x}) = 0$ respectively. We consider each r_i as a perturbation of the root p_i, hence we write $r_i = p_i + \Delta p_i$ for $i = 1, \dots, \mu_{\mathbb{R},I}$.

Now we concentrate on a single element p of $\mathcal{Z}_{\mathbb{R}}(\mathbf{f})$.

Corollary 3.8. *Let p be one of the real solutions of $\mathbf{f} = 0$, and $p + \Delta p$ the corresponding real solution of $\mathbf{f}+\boldsymbol{\varepsilon} = 0$. Then we have*

$$0 = (\mathbf{f}+\boldsymbol{\varepsilon})(p+\Delta p) = \boldsymbol{\varepsilon}(p) + \mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)\Delta p + (v_1(\xi_1), \dots, v_n(\xi_n))^t \qquad (1)$$

where $\xi_1, \dots, \xi_n$ are points on the line which connects the points p and $p+\Delta p$, and $v_j(\xi_j) = \frac{1}{2}\Delta p^t H_{f_j+\varepsilon_j}(\xi_j)\Delta p$ for each $j = 1, \dots, n$.

Proof. It suffices to put $q = p + \Delta p$, apply the formula of Proposition 3.6 to the polynomial system $(\mathbf{f}+\boldsymbol{\varepsilon})(\mathbf{x})$, and use the fact that $\mathbf{f}(p) = 0$. □

Example 3.9. We consider $\mathbf{f} = \{f_1, f_2\}$ where $f_1 = xy - 6$, $f_2 = x^2 + y^2 - 13$ and observe that $\mathcal{Z}_{\mathbb{R}}(\mathbf{f}) = \{(-3,-2),(3,2),(-2,-3),(2,3)\}$. The set $\mathbf{f}(\mathbf{x})$ is embedded into the following family $F(\mathbf{a},\mathbf{x}) = \{xy + a_1, x^2 + a_2y^2 + a_3\}$.

Let $\boldsymbol{\alpha} = (a_1, a_2, a_3) \in \mathbb{R}^3$; the semi-algebraic open set

$$\mathcal{V} = \{\boldsymbol{\alpha} \in \mathbb{R}^3 \mid a_3^2 - 4a_1^2 a_2 > 0, a_2 > 0, a_3 < 0\}$$

is a subset of the I-optimal scheme $\mathcal{U} = \{\boldsymbol{\alpha} \in \mathbb{A}^3_{\mathbb{R}} \mid a_2(a_3^2 - 4a_1^2 a_2) \neq 0\}$. Moreover, it contains the point $\boldsymbol{\alpha}_I = (-6, 1, -13)$, and the fiber over each $\boldsymbol{\alpha} \in \mathcal{V}$ consists of 4 real points. The set $\boldsymbol{\varepsilon}(\mathbf{x}) = \{\delta_1, \delta_2 y^2 + \delta_3\}$ is an admissible perturbation of $\mathbf{f}(\mathbf{x})$ if and only if the conditions $(\delta_3 - 13)^2 - 4(\delta_1 - 6)^2(\delta_2 + 1) > 0$, $\delta_2 > -1$, and $\delta_3 < 13$ are satisfied. Since the values $\delta_1 = 2$, $\delta_2 = \frac{5}{4}$, and $\delta_3 = 0$ satisfy the previous conditions, the polynomial set $\boldsymbol{\varepsilon}(\mathbf{x}) = \{2, \frac{5}{4}y^2\}$ is an admissible perturbation of $\mathbf{f}(\mathbf{x})$. The real roots of $(\mathbf{f} + \boldsymbol{\varepsilon})(\mathbf{x}) = 0$ are

$$\mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \boldsymbol{\varepsilon}) = \left\{\left(-3, -\tfrac{4}{3}\right), \left(3, \tfrac{4}{3}\right), (-2,-2), (2,2)\right\}$$

For each $r_i \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \boldsymbol{\varepsilon})$ the matrix $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(r_i)$ is invertible, as predicted by the theory. On the contrary, by evaluating $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(\mathbf{x})$ at the third and the fourth point of $\mathcal{Z}_{\mathbb{R}}(\mathbf{f})$ we obtain a singular matrix. This is an obstruction to the development of the theory which suggests further restrictions (see the following discussion).

Our idea is to evaluate Δp using equation (1) of Corollary 3.8. However, while the assumption that $\boldsymbol{\varepsilon}(\mathbf{x})$ is an admissible perturbation of $\mathbf{f}(\mathbf{x})$ combined with the Jacobian criterion guarantee the non singularity of $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p + \Delta p)$, they do not imply the non singularity of the matrix $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)$, as we have just seen in Example 3.9. The next step is to find a criterion which guarantees the non singularity of $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)$.

Lemma 3.10. *Let* $\|\cdot\|$ *be an induced matrix norm on* $\mathrm{Mat}_n(\mathbb{R})$ *and assume that* $\|\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)\| < 1$. *Then* $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)$ *is invertible.*

Proof. By assumption p is a nonsingular root of $\mathbf{f}(\mathbf{x}) = 0$, hence $\mathrm{Jac}_{\mathbf{f}}(p)$ is invertible and so $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)$ can be rewritten as $\mathrm{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p) = \mathrm{Jac}_{\mathbf{f}}(p) + \mathrm{Jac}_{\boldsymbol{\varepsilon}}(p) = \mathrm{Jac}_{\mathbf{f}}(p)\left(I + \mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)\right)$. Consequently, it suffices to show that the matrix $I + \mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)$ is invertible. And we achieve it by proving that the spectral radius $\rho(\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p))$ is smaller than 1.

We have $\rho(\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)) \leq \|\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)\| < 1$, and the proof is now complete. □

Note that the requirement $\|\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)\| < 1$ gives a restriction on the admissible choices of $\boldsymbol{\varepsilon}(\mathbf{x})$, as we see in the following example.

Example 3.11. (Example 3.9 continued)
Let $\boldsymbol{\varepsilon}(\mathbf{x}) = \{\delta_1, \delta_2 y^2 + \delta_3\}$, with $\delta_i \in \mathbb{R}$, be an admissible perturbation of $\mathbf{f}(\mathbf{x})$ of Example 3.9. We consider the real solution $p_4 = (2,3)$ of $\mathbf{f} = 0$ and compute $\|\operatorname{Jac}_{\mathbf{f}}(p_4)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p_4)\|_2^2 = \frac{117}{25}\delta_2^2$. Using Lemma 3.10 we see that the condition $|\delta_2| < \frac{5}{39}\sqrt{13}$ is sufficient to have $\operatorname{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p_4)$ invertible.

From now on we assume that the hypothesis of Lemma 3.10 is satisfied. In order to deduce an upper bound for $\|\Delta p\|$ we consider an approximation of it.

Definition 3.12. If $\|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\|$ is different from 1, we denote the number $1/(1-\|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\|)$ by $\Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)$. Moreover, by Δp^1 we denote the vector $-\operatorname{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)^{-1}\boldsymbol{\varepsilon}(p)$, which is the solution of equation (1) of Corollary 3.8 truncated at the first order.

Proposition 3.13. *Let* $\|\cdot\|$ *be an induced matrix norm on* $\operatorname{Mat}_n(\mathbb{R})$ *and assume that* $\|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\| < 1$. *Then we have*

$$\|\Delta p^1\| \leq \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\ \|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\ \|\boldsymbol{\varepsilon}(p)\| \tag{2}$$

Proof. Lemma 3.10 guarantees that the matrix $\operatorname{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)$ is invertible, so

$$\begin{aligned}\Delta p^1 &= -\operatorname{Jac}_{\mathbf{f}+\boldsymbol{\varepsilon}}(p)^{-1}\boldsymbol{\varepsilon}(p) = -(\operatorname{Jac}_{\mathbf{f}}(p)+\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p))^{-1}\boldsymbol{\varepsilon}(p)\\ &= -\left(I+\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\right)^{-1}\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\boldsymbol{\varepsilon}(p)\end{aligned}$$

We apply the inequality of Proposition 3.2 to $\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)$, and get

$$\begin{aligned}\|\Delta p^1\| &\leq \|(I+\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)^{-1})\|\ \|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\ \|\boldsymbol{\varepsilon}(p)\|\\ &\leq \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\ \|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\ \|\boldsymbol{\varepsilon}(p)\|\end{aligned}$$

which concludes the proof. □

We introduce the local condition number of the polynomial system $\mathbf{f}(\mathbf{x}) = 0$.

Definition 3.14. Let $f_1(\mathbf{x}),\dots,f_n(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$, let $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}),\dots,f_n(\mathbf{x})\}$ such that the ideal generated by $\mathbf{f}(x)$ defines a zero-dimensional scheme, and let p be a nonsingular real solution of $\mathbf{f}(\mathbf{x}) = 0$. Let $\|\cdot\|$ be a norm.

(a) The number $\kappa(\mathbf{f},p) = \|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\|\operatorname{Jac}_{\mathbf{f}}(p)\|$ is called the **local condition number** of $\mathbf{f}(\mathbf{x})$ at p.

(b) If the norm is an r-norm, the local condition number is denoted by $\kappa_r(\mathbf{f},p)$.

The following theorem illustrates the importance of the local condition number. It depends on $\mathbf{f}$ and p, not on $\boldsymbol{\varepsilon}$, and is a key ingredient to provide an upper bound for the relative error $\frac{\|\Delta p^1\|}{\|p\|}$.

Theorem 3.15. (Local Condition Number)
Let $f_1(\mathbf{x}),\dots,f_n(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$ *and let* $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}),\dots,f_n(\mathbf{x})\}$ *such that the ideal generated by* $\mathbf{f}(x)$ *defines a zero-dimensional scheme; let* $\boldsymbol{\varepsilon}(\mathbf{x})$ *be an admissible perturbation of* $\mathbf{f}(\mathbf{x})$*, let p be a nonsingular real solution of* $\mathbf{f}(\mathbf{x}) = 0$*, let* $\|\cdot\|$ *be an induced matrix norm, and assume that* $\|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\| < 1$*. Then we have*

$$\frac{\|\Delta p^1\|}{\|p\|} \le \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\ \kappa(\mathbf{f},p)\left(\frac{\|\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\|}{\|\operatorname{Jac}_{\mathbf{f}}(p)\|} + \frac{\|\boldsymbol{\varepsilon}(0)-\boldsymbol{\varepsilon}^{\ge 2}(0,p)\|}{\|\mathbf{f}(0)-\mathbf{f}^{\ge 2}(0,p)\|}\right) \tag{3}$$

Proof. By Definition 3.5 the evaluation of $\boldsymbol{\varepsilon}$ at 0 can be expressed in this way $\boldsymbol{\varepsilon}(0)=\boldsymbol{\varepsilon}(p)-\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)p+\boldsymbol{\varepsilon}^{\ge 2}(0,p)$. We get $\boldsymbol{\varepsilon}(p)=\boldsymbol{\varepsilon}(0)+\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)p-\boldsymbol{\varepsilon}^{\ge 2}(0,p)$. Dividing (2) of Proposition 3.13 by $\|p\|$ we obtain

$$\begin{aligned}
\frac{\|\Delta p^1\|}{\|p\|} &\le \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\ \|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\ \frac{\|\boldsymbol{\varepsilon}(p)\|}{\|p\|} \\
&\le \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\ \|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\ \frac{\|\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\|\|p\| + \|\boldsymbol{\varepsilon}(0)-\boldsymbol{\varepsilon}^{\ge 2}(0,p)\|}{\|p\|} \\
&= \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\left(\|\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\| + \frac{\|\boldsymbol{\varepsilon}(0)-\boldsymbol{\varepsilon}^{\ge 2}(0,p)\|}{\|p\|}\right)
\end{aligned}$$

Using again Definition 3.5 we express $\mathbf{f}(0) = \mathbf{f}(p) - \operatorname{Jac}_{\mathbf{f}}(p)p + \mathbf{f}^{\ge 2}(0,p)$; since $\mathbf{f}(p) = 0$ we have $\|\mathbf{f}(0) - \mathbf{f}^{\ge 2}(0,p)\| = \|\operatorname{Jac}_{\mathbf{f}}(p)p\| \le \|\operatorname{Jac}_{\mathbf{f}}(p)\|\|p\|$ from which

$$\frac{1}{\|p\|} \le \frac{\|\operatorname{Jac}_{\mathbf{f}}(p)\|}{\|\mathbf{f}(0)-\mathbf{f}^{\ge 2}(0,p)\|}$$

We combine the inequalities to obtain

$$\begin{aligned}
\frac{\|\Delta p^1\|}{\|p\|} &\le \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\left(\|\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\| + \|\operatorname{Jac}_{\mathbf{f}}(p)\|\frac{\|\boldsymbol{\varepsilon}(0)-\boldsymbol{\varepsilon}^{\ge 2}(0,p)\|}{\|\mathbf{f}(0)-\mathbf{f}^{\ge 2}(0,p)\|}\right) \\
&\le \Lambda(\mathbf{f},\boldsymbol{\varepsilon},p)\|\operatorname{Jac}_{\mathbf{f}}(p)^{-1}\|\|\operatorname{Jac}_{\mathbf{f}}(p)\|\left(\frac{\|\operatorname{Jac}_{\boldsymbol{\varepsilon}}(p)\|}{\|\operatorname{Jac}_{\mathbf{f}}(p)\|} + \frac{\|\boldsymbol{\varepsilon}(0)-\boldsymbol{\varepsilon}^{\ge 2}(0,p)\|}{\|\mathbf{f}(0)-\mathbf{f}^{\ge 2}(0,p)\|}\right)
\end{aligned}$$

and the proof is concluded. □

The next remark contains observations about the local condition number.

Remark 3.16. We call attention to the following observations.

(a) The notion of local condition number given in Definition 3.14 is a generalization of the classical notion of condition number of linear systems (see [5]). In fact, if $\mathbf{f}(\mathbf{x})$ is linear, that is $\mathbf{f}(\mathbf{x}) = A\mathbf{x} - b$ with $A \in \mathrm{Mat_n}(\mathbb{R})$

invertible, and $\mathcal{Z}_{\mathbb{R}}(\mathbf{f}) = \{p\} = \{A^{-1}b\}$, then $\kappa(\mathbf{f},p)$ is the classical condition number of the matrix A. In fact $\mathrm{Jac}_{\mathbf{f}}(\mathbf{x}) = A$, and so $\kappa(\mathbf{f},p) = \|\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\| \|\mathrm{Jac}_{\mathbf{f}}(p)\| = \|A^{-1}\| \|A\|$. Further, if we consider the perturbation $\boldsymbol{\varepsilon}(\mathbf{x}) = \Delta A\mathbf{x} - \Delta b$, relation (3) becomes

$$\frac{\|\Delta p\|}{\|p\|} \leq \frac{1}{1 - \|A^{-1}\| \, \|\Delta A\|} \|A^{-1}\| \, \|A\| \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right) \tag{4}$$

which is the relation that quantifies the sensitivity of the $Ax = b$ problem (see [5], Theorem 4.1).

(b) Using any induced matrix norm, the condition number $\kappa(\mathbf{f},p)$ turns out to be greater than or equal to 1. In particular, using the 2-norm we have $\kappa_2(\mathbf{f},p) = \frac{\sigma_{\max}(\mathrm{Jac}_{\mathbf{f}}(p))}{\sigma_{\min}(\mathrm{Jac}_{\mathbf{f}}(p))}$; in this case the local condition number attains its minimum, that is $\kappa_2(\mathbf{f},p) = 1$, when $\mathrm{Jac}_{\mathbf{f}}(p)$ is orthonormal.

(c) The condition number $\kappa(\mathbf{f},p)$ is invariant under a scalar multiplication of the polynomial system $\mathbf{f}(\mathbf{x})$ by a unique nonzero real number γ. On the contrary, $\kappa(\mathbf{f},p)$ is not invariant under a generic scalar multiplication of each polynomial $f_j(\mathbf{x})$ of $\mathbf{f}(\mathbf{x})$. The reason is that if we multiply each $f_j(\mathbf{x})$ by a nonzero real number γ_j we obtain the new polynomial set $\mathbf{g}(\mathbf{x}) = \{\gamma_1 f_1(\mathbf{x}), \ldots, \gamma_n f_n(\mathbf{x})\}$ whose condition number at p is

$$\kappa(\mathbf{g},p) = \|\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\Gamma^{-1}\| \, \|\Gamma \mathrm{Jac}_{\mathbf{f}}(p)\| \neq \kappa(\mathbf{f},p)$$

where $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_n) \in \mathrm{Mat_n}(\mathbb{R})$ is the diagonal matrix with entries $\gamma_1, \ldots, \gamma_n$.

(d) It is interesting to observe that if p is the origin then Formula (3) of the theorem is not applicable. However, one can translate p away from the origin, and the nice thing is that the local condition number does not change.

4. Optimization of the local condition number

In this section we introduce a strategy to improve the numerical stability of zero-dimensional polynomial systems of $\mathbb{R}[\mathbf{x}]$ which define a smooth scheme. Let $f_1(\mathbf{x}), \ldots, f_n(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$ and let $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$ such that $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme; our aim is to find an alternative representation of I with minimal local condition number.

Motivated by Remark 3.16, item (b) and (c), we consider the strategy of resizing each polynomial of $\mathbf{f}(\mathbf{x})$, and study its effects on the condition number.

The following proposition shows that rescaling each $f_j(\mathbf{x})$ so that $\frac{\partial f_j}{\partial \mathbf{x}}(p)$ has unitary norm is a nearly optimal, in some cases optimal, strategy. The result is obtained by adapting the method of Van der Sluis (see [16], Section 7.3) to the polynomial case.

Proposition 4.1. *Let $f_1(\mathbf{x}),\dots,f_n(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]$, let $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}),\dots,f_n(\mathbf{x})\}$ such that $I = (\mathbf{f}(x))$ defines a zero-dimensional scheme, and let p be a nonsingular real solution of $\mathbf{f}(\mathbf{x}) = 0$. Let $r_1 \geq 1, r_2 \geq 1$ be real numbers such that $\frac{1}{r_1} + \frac{1}{r_2} = 1$, including the pairs $(1,\infty)$ and $(\infty,1)$, let $\gamma = (\gamma_1,\dots,\gamma_n)$ be an n-tuple of nonzero real numbers, and let $\mathbf{g}_\gamma(\mathbf{x})$, $\boldsymbol{u}(\mathbf{x})$ be the polynomial systems defined by the polynomials $\mathbf{g}_\gamma(\mathbf{x}) = \{\gamma_1 f_1(\mathbf{x}),\dots,\gamma_n f_n(\mathbf{x})\}$ and also the polynomials $\boldsymbol{u}(\mathbf{x}) = \{\|\frac{\partial f_1}{\partial \mathbf{x}}(p)\|_{r_2}^{-1} f_1(\mathbf{x}),\dots,\|\frac{\partial f_n}{\partial \mathbf{x}}(p)\|_{r_2}^{-1} f_n(\mathbf{x})\}$.*

(a) We have the inequality $\kappa_{r_1}(\boldsymbol{u},p) \leq n^{1/r_1}\kappa_{r_1}(\mathbf{g}_\gamma,p)$.

(b) In particular, if $(r_1,r_2) = (\infty,1)$ we have the equality

$$\kappa_\infty(\boldsymbol{u},p) = \min\nolimits_\gamma \kappa_\infty(\mathbf{g}_\gamma,p)$$

where $\boldsymbol{u}(\mathbf{x}) = \{\|\frac{\partial f_1}{\partial \mathbf{x}}(p)\|_1^{-1} f_1(\mathbf{x}),\dots,\|\frac{\partial f_n}{\partial \mathbf{x}}(p)\|_1^{-1} f_n(\mathbf{x})\}$.

Proof. Let $\Gamma = \operatorname{diag}(\gamma_1,\dots,\gamma_\mathrm{n})$ and $D = \operatorname{diag}(\|\frac{\partial \mathrm{f}_1}{\partial \mathbf{x}}(\mathrm{p})\|_{\mathrm{r}_2}^{-1},\dots,\|\frac{\partial \mathrm{f}_\mathrm{n}}{\partial \mathbf{x}}(\mathrm{p})\|_{\mathrm{r}_2}^{-1})$; then $\mathrm{Jac}_{\mathbf{g}_\gamma}(\mathbf{x}) = \Gamma\,\mathrm{Jac}_\mathbf{f}(\mathbf{x})$ and $\mathrm{Jac}_{\boldsymbol{u}}(\mathbf{x}) = D\,\mathrm{Jac}_\mathbf{f}(\mathbf{x})$. The condition numbers of $\mathbf{g}_\gamma(\mathbf{x})$ and $\boldsymbol{u}(\mathbf{x})$ at p are given by

$$\begin{aligned}\kappa_{r_1}(\mathbf{g}_\gamma,p) &= \|(\Gamma\,\mathrm{Jac}_\mathbf{f}(p))^{-1}\|_{r_1}\|\Gamma\,\mathrm{Jac}_\mathbf{f}(p)\|_{r_1}\\ \kappa_{r_1}(\boldsymbol{u},p) &= \|(D\,\mathrm{Jac}_\mathbf{f}(p))^{-1}\|_{r_1}\|D\,\mathrm{Jac}_\mathbf{f}(p)\|_{r_1}\end{aligned}$$

From Proposition 3.3 we have

$$\begin{aligned}\|D\,\mathrm{Jac}_\mathbf{f}(p)\|_{r_1} &\leq n^{1/r_1}\max_i \|(D\,\mathrm{Jac}_\mathbf{f}(p))_i\|_{r_2} = n^{1/r_1}\\ \|(D\,\mathrm{Jac}_\mathbf{f}(p))^{-1}\|_{r_1} &= \|\mathrm{Jac}_\mathbf{f}^{-1}(p)D^{-1}\|_{r_1} = \|\mathrm{Jac}_\mathbf{f}^{-1}(p)\Gamma^{-1}\Gamma D^{-1}\|_{r_1}\\ &\leq \|\mathrm{Jac}_\mathbf{f}^{-1}(p)\Gamma^{-1}\|_{r_1}\max_i\left(|\gamma_i|\left\|\frac{\partial f_i}{\partial \mathbf{x}}(p)\right\|_{r_2}\right)\\ &\leq \|\mathrm{Jac}_\mathbf{f}^{-1}(p)\Gamma^{-1}\|_{r_1}\|\Gamma\,\mathrm{Jac}_\mathbf{f}(p)\|_{r_1} = \kappa_{r_1}(\mathbf{g}_\gamma,p)\end{aligned}$$

therefore $\kappa_{r_1}(\boldsymbol{u},p) \leq n^{1/r_1}\kappa_{r_1}(\mathbf{g}_\gamma,p)$ and (a) is proved. To prove (b) it suffices to use (a) and observe that $n^{1/\infty} = 1$ □

Remark 4.2. The above proposition implies that the strategy of rescaling each polynomial $f_j(\mathbf{x})$ to make it unitary at p (see Definition 3.5) is beneficial for lowering the local condition number of $\mathbf{f}(\mathbf{x})$ at p. This number is minimum when $r=\infty$, it is within factor $\sqrt{n}$ of the minimum when $r=2$. However, for $r=2$ we can do better, at least when all the polynomials $f_1(\mathbf{x}),\dots,f_n(\mathbf{x})$ have equal degree. The idea is to use Remark 3.16, item (b) which says that when using the matrix 2-norm, the local condition number attains its minimum when the Jacobian matrix is orthonormal.

Proposition 4.3. *With the same assumptions of Proposition 4.1 we assume that* $\deg(f_1)=\cdots=\deg(f_n)$*; moreover, let* $C=(c_{ij})\in \mathrm{Mat}_n(\mathbb{R})$ *be an invertible matrix, and let* $\mathbf{g}$ *be defined by* $\mathbf{g}^{\mathrm{tr}}=C\cdot\mathbf{f}^{\mathrm{tr}}$. *Then the following conditions are equivalent:*

(a) $\kappa_2(\mathbf{g},p)=1$*, the minimum possible;*

(b) $C^tC=(\mathrm{Jac}_{\mathbf{f}}(p)\,\mathrm{Jac}_{\mathbf{f}}(p)^t)^{-1}$

Proof. We know that $\kappa_2(\mathbf{g},p)=1$ if and only if the matrix $\mathrm{Jac}_{\mathbf{g}}(p)$ is orthonormal. This condition can be expressed by the equality $\mathrm{Jac}_{\mathbf{g}}(p)\,\mathrm{Jac}_{\mathbf{g}}(p)^t=I_n$, that is $C\,\mathrm{Jac}_{\mathbf{f}}(p)\,\mathrm{Jac}_{\mathbf{f}}(p)^t\,C^t=I_n$ and the conclusion follows. □

We observe that condition (b) of Proposition 4.3 requires that the entries of C satisfy an underdetermined system of $(n^2+n)/2$ independent quadratic equations in n^2 unknowns.

5. Experiments

In numerical linear algebra it is known (see for instance [5], Ch. 4, Section 1) that the upper bound given by the classical formula (4) of Remark 3.16 (a) is not necessarily sharp. Since our upper bound (3) generalizes the classical one, as shown in Remark 3.16, we provide some experimental evidence that lowering the condition number not only sharpens the upper bound, but indeed stabilizes the solution point. To do that, we exhibit two explicit examples where we show that a suitable change of generators of our ideals leads to a substantial lowering of our condition number. As a side remark, we mention that we computed the corresponding condition number of Shub and Smale as described in the paper [23], and observed that also in their case the number obtained with our new polynomials is better (smaller) than the number obtained with the original ones. But in our case we can benefit from the claim of Proposition 4.3 since, by changing the generators of our ideals, we achieve the optimal condition number 1.

Example 5.1. We consider $f_1, f_2 \in \mathbb{R}[x,y]$, where

$$\begin{array}{rcl} f_1 & = & \frac{1}{4}x^2y+xy^2+\frac{1}{4}y^3+\frac{1}{5}x^2-\frac{5}{8}xy+\frac{13}{40}y^2+\frac{9}{40}x-\frac{3}{5}y+\frac{1}{40} \\ f_2 & = & x^3+\frac{14}{13}xy^2+\frac{57}{52}x^2-\frac{25}{52}xy+\frac{8}{13}y^2-\frac{11}{52}x-\frac{4}{13}y-\frac{4}{13} \end{array}$$

Let $\mathbf{f} = \{f_1, f_2\}$ and let I be the ideal generated by $\mathbf{f}$. The set $\mathcal{Z}_{\mathbb{R}}(\mathbf{f})$ has 7 real roots; we consider the point $p = (0,1) \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f})$. The polynomial system $\mathbf{f}$ is unitary at p and its condition number is $\kappa_2(\mathbf{f}, p) = 8$. Using Proposition 4.3 we construct a new polynomial system $\mathbf{g}$ with minimal local condition number at p. The new pair of generators $\mathbf{g}$ is defined (see Proposition 4.3) by the following the formula $\mathbf{g}^{\mathrm{tr}} = C \cdot \mathbf{f}^{\mathrm{tr}}$, where $C = (c_{ij}) \in \mathrm{Mat}_2(\mathbb{R})$ is an invertible matrix whose entries satisfy the following system

$$\left\{ \begin{array}{lcl} c_{11}^2 + c_{21}^2 & = & \frac{25}{16} \\ c_{11}c_{12} + c_{21}c_{22} & = & -\frac{15}{16} \\ c_{12}^2 + c_{22}^2 & = & \frac{25}{16} \end{array} \right.$$

A solution is given by $c_{11} = 1$, $c_{12} = 0$, $c_{21} = \frac{63}{16}$, $c_{22} = -\frac{65}{16}$, and we observe that the associated unitary polynomial system $\mathbf{g} = \{f_1, \frac{63}{16}f_1 - \frac{65}{16}f_2\}$ provides an alternative representation of I with minimal local condition number $\kappa_2(\mathbf{g}, p) = 1$ at the point p.

Now we embed the system $\mathbf{f}(x,y)$ into the family $F(a,x,y) = \{F_1, F_2\}$ where

$$\begin{array}{rcl} F_1(a,x,y) & = & \frac{1}{4}x^2y+xy^2+\frac{1}{4}y^3+\frac{1}{5}x^2-\frac{5}{8}xy+\left(\frac{13}{40}-a\right)y^2 \\ & & +\left(\frac{9}{40}+a\right)x+\left(-\frac{3}{5}+a\right)y+\frac{1}{40}-2a \\ F_2(a,x,y) & = & x^3+\frac{14}{13}xy^2+\frac{57}{52}x^2-\frac{25}{52}xy+\left(\frac{8}{13}+a\right)y^2 \\ & & +\left(-\frac{11}{52}+a\right)x-\left(\frac{4}{13}+a\right)y-\frac{4}{13}+a^2 \end{array}$$

We denote by $I_F(a,x,y)$ the ideal generated by $F(a,x,y)$ in $\mathbb{R}[a,x,y]$, compute the reduced `Lex`-Gröbner basis of $I_F(a,x,y)\mathbb{R}(a)[x,y]$, and get

$$\{x + \tfrac{l_1(a,y)}{d_F(a)},\ \ y^9 + l_2(a,y)\}$$

where $l_1(a,y), l_2(a,y) \in \mathbb{R}[a,y]$ have degree 8 in y and $d_F(a) \in \mathbb{R}[a]$ has degree 12. This basis has the shape prescribed by the Shape Lemma and a flat locus is given by $\{\alpha \in \mathbb{R} \mid d_F(\alpha) \neq 0\}$. We let $D_F(a,x,y) = \det(\mathrm{Jac}_F(a,x,y))$, $J_F(a,x,y) = I_F(a,x,y) + (D_F(a,x,y))$, compute $J_F(a,x,y) \cap \mathbb{R}[a]$, and we get the principal ideal generated by a univariate polynomial $h_F(a)$ of degree 28. An I-optimal subscheme is $\mathcal{U}_F = \{\alpha \in \mathbb{R} \mid d_F(\alpha)h_F(\alpha) \neq 0\}$. An open semi-algebraic subset $\mathcal{V}_F$ of $\mathcal{U}_F$ which contains the point $\alpha_I = 0$ and such that the fiber over each $\alpha \in \mathcal{V}_F$ consists of 7 real points, is given by the open interval (α_1, α_2), where $\alpha_1 < 0$ and $\alpha_2 > 0$ are the real roots of $d_F(a)h_F(a) = 0$ closest to the origin. Their approximate values are $\alpha_1 = -0.00006$ and $\alpha_2 = 0.01136$.

To produce similar perturbations, we embed the system $\mathbf{g}(x,y)$ into the family $G(a,x,y) = \{G_1, G_2\}$ where

$$\begin{aligned}
G_1(a,x,y) &= \tfrac{1}{4}x^2y + xy^2 + \tfrac{1}{4}y^3 + \tfrac{1}{5}x^2 - \tfrac{5}{8}xy + \left(\tfrac{13}{40} - a\right)y^2 \\
&\quad + \left(\tfrac{9}{40} + a\right)x + \left(-\tfrac{3}{5} + a\right)y + \tfrac{1}{40} - 2a \\
G_2(a,x,y) &= -\tfrac{65}{16}x^3 + \tfrac{63}{64}x^2y - \tfrac{7}{16}xy^2 + \tfrac{63}{64}y^3 - \tfrac{1173}{320}x^2 - \tfrac{65}{128}xy \\
&\quad + \left(-\tfrac{781}{640} + a\right)y^2 + \left(\tfrac{1117}{640} + a\right)x + \left(-\tfrac{89}{80} - a\right)y + \tfrac{863}{640} + a^2
\end{aligned}$$

We denote by $I_G(a,x,y)$ the ideal generated by $G(a,x,y)$ in $\mathbb{R}[a,x,y]$, compute the reduced Lex-Gröbner basis of $I_G(a,x,y)\mathbb{R}(a)[x,y]$, and get

$$\{x + \tfrac{l_3(a,y)}{d_G(a)},\ y^9 + l_4(a,y)\}$$

where $l_3(a,y), l_4(a,y) \in \mathbb{R}[a,y]$ have degree 8 in y and $d_G(a) \in \mathbb{R}[a]$ has degree 12, therefore the basis has the shape prescribed by the Shape Lemma. A flat locus is given by $\{\alpha \in \mathbb{R} \mid d_G(\alpha) \neq 0\}$. We let $D_G(a,x,y) = \det(\mathrm{Jac}_G(a,x,y))$, $J_G(a,x,y) = I_G(a,x,y) + (D_G(a,x,y))$ and compute $J_G(a,x,y) \cap \mathbb{R}[a]$. We get the principal ideal generated by a univariate polynomial $h_G(a)$ of degree 28. An I-optimal subscheme is $\mathcal{U}_G = \{\alpha \in \mathbb{R} \mid d_G(\alpha)h_G(\alpha) \neq 0\}$. An open semialgebraic subset $\mathcal{V}_G$ of $\mathcal{U}_G$ containing the point $\alpha_I = 0$ and such that the fiber over each $\alpha \in \mathcal{V}_G$ consists of 7 real points is given by the open interval (α_3, α_4), where $\alpha_3 < 0$ and $\alpha_4 > 0$ are the real roots of $d_G(a)h_G(a) = 0$ closest to the origin. Their approximate values are $\alpha_3 = -0.00009$ and $\alpha_4 = 0.00914$.

Let $\alpha \in (\alpha_1, \alpha_4)$. According to Definition 3.7 the polynomial set $\boldsymbol{\varepsilon}(x,y) = \{-\alpha y^2 + \alpha x + \alpha y - 2\alpha,\ \alpha y^2 + \alpha x - \alpha y + \alpha^2\}$ is an admissible perturbation of $\mathbf{f}(x,y)$ and $\mathbf{g}(x,y)$. Further, since $\|\mathrm{Jac}_{\mathbf{f}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)\|_2 = \sqrt{65}|\alpha| < 1$ and $\|\mathrm{Jac}_{\mathbf{g}}(p)^{-1}\,\mathrm{Jac}_{\boldsymbol{\varepsilon}}(p)\|_2 = \sqrt{2}|\alpha| < 1$ Theorem 3.15 can be applied.

We let $q \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \boldsymbol{\varepsilon})$ and $r \in \mathcal{Z}_{\mathbb{R}}(\mathbf{g} + \boldsymbol{\varepsilon})$ be the two perturbations of the point p. In order to compare the numerical behaviour of $\mathbf{f}$ and $\mathbf{g}$ at the real root p we compare the relative errors $\frac{\|q-p\|_2}{\|p\|_2}$ and $\frac{\|r-p\|_2}{\|p\|_2}$ for different values of α.

$\kappa_2(\mathbf{f},p)$	$\mathrm{UB}(\mathbf{f},p)$	$\frac{\|q-p\|_2}{\|p\|_2}$
8	0.1729	0.000097
$\kappa_2(\mathbf{g},p)$	$\mathrm{UB}(\mathbf{g},p)$	$\frac{\|r-p\|_2}{\|p\|_2}$
1	0.0275	0.000023

The first column of the above table contains the values of the local condition numbers of $\mathbf{f}$ and $\mathbf{g}$ at p. The second column contains the mean values of the upper bounds $\mathrm{UB}(\mathbf{f},p)$ and $\mathrm{UB}(\mathbf{g},p)$ given by Theorem 3.15, computed for 100 random values of $\alpha \in (\alpha_1,\alpha_4)$. The third column contains the mean values of $\frac{\|q-p\|_2}{\|p\|_2}$ and $\frac{\|r-p\|_2}{\|p\|_2}$ for the same values of α. The mean values of $\frac{\|q-p\|_2}{\|p\|_2}$ are smaller than the mean values of $\frac{\|r-p\|_2}{\|p\|_2}$. This fact suggests that p is more stable when it is considered as a root of $\mathbf{g}$ instead of as a root of $\mathbf{f}$.

Example 5.2. We consider $f_1, f_2, f_3 \in \mathbb{R}[x,y,z]$, where

$$\begin{array}{rcl} f_1 & = & \frac{6}{17}x^2 + xy - \frac{24}{85}x - \frac{8}{85}y - \frac{6}{85} \\ f_2 & = & \frac{39}{89}x^2 + \frac{70}{89}xy + yz - \frac{39}{89}x + \frac{10}{89}y \\ f_3 & = & y^2 + 2xz + z^2 - z \end{array}$$

Let $\mathbf{f} = \{f_1, f_2, f_3\}$ and let I be the ideal generated by $\mathbf{f}$. The set $\mathcal{Z}_{\mathbb{R}}(\mathbf{f})$ has 6 real roots; we consider the point $p = (1,0,0) \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f})$. The polynomial system $\mathbf{f}$ is unitary at p and its condition number is $\kappa_2(\mathbf{f},p) = 123$. Using Proposition 4.3 we construct a new polynomial system $\mathbf{g}$ with minimal local condition number at p. The new set $\mathbf{g}$ is defined by $\mathbf{g}^{\mathrm{tr}} = C \cdot \mathbf{f}^{\mathrm{tr}}$, where $C = (c_{ij}) \in \mathrm{Mat}_3(\mathbb{R})$ is an invertible matrix whose entries satisfy the following system

$$\left\{ \begin{array}{lcc} c_{11}^2 + c_{21}^2 + c_{31}^2 & = & \frac{57229225}{15129} \\ c_{11}c_{12} + c_{21}c_{22} + c_{31}c_{32} & = & -\frac{57221660}{15129} \\ c_{11}c_{13} + c_{21}c_{23} + c_{31}c_{33} & = & 0 \\ c_{12}^2 + c_{22}^2 + c_{32}^2 & = & \frac{57229225}{15129} \\ c_{12}c_{13} + c_{22}c_{23} + c_{32}c_{33} & = & 0 \\ c_{13}^2 + c_{23}^2 + c_{33}^2 & = & 1 \end{array} \right.$$

A solution is given by $c_{11} = c_{33} = 1$, $c_{12} = c_{13} = c_{23} = c_{32} = 0$, $c_{21} = \frac{7564}{123}$, $c_{22} = -\frac{7565}{123}$. Therefore the associated unitary polynomial system is the following $\mathbf{g} = \{f_1, \frac{7564}{123}f_1 - \frac{7565}{123}f_2, f_3\}$. It provides an alternative representation of I with minimal local condition number $\kappa_2(\mathbf{g},p) = 1$ at the point p.

We embed the system $\mathbf{f}(x,y,z)$ into the family $F(a,x,y,z) = \{F_1, F_2, F_3\}$ where

$$\begin{array}{rcl} F_1(a,x,y,z) & = & \frac{6}{17}x^2 + (1-a^2)xy + (-\frac{24}{85}+a)x + (-\frac{8}{85}-a)y + (-\frac{6}{85}+a^2) \\ F_2(a,x,y,z) & = & \frac{39}{89}x^2 + (\frac{70}{89}+a)xy + yz + (\frac{39}{89}+a)x + (\frac{10}{89}+a)y \\ F_3(a,x,y,z) & = & y^2 + 2xz + (1-2a)z^2 + (-1+a)z \end{array}$$

We denote by $I_F(a,x,y,z)$ the ideal generated by $F(a,x,y,z)$ in $\mathbb{R}[a,x,y,z]$, compute the reduced $\mathtt{Lex}$-Gröbner basis of $I_F(a,x,y,z)\mathbb{R}(a)[x,y,z]$, and get

$$\{x + \tfrac{l_1(a,z)}{d_F(a)},\ y + \tfrac{l_2(a,z)}{d_F(a)},\ z^9 + \tfrac{l_3(a,z)}{e_F(a)}\}$$

where $l_1(a,z), l_2(a,z), l_3(a,z) \in \mathbb{R}[a,z]$ have degrees $\deg_z(l_1) = \deg_z(l_2) = 7$ and $\deg_z(l_3) = 8$ while $d_F(a) \in \mathbb{R}[a]$ has degree 54, and $e_F(a) \in \mathbb{R}[a]$ has degree 11. The basis has the shape prescribed by the Shape Lemma. A flat locus is given by $\{\alpha \in \mathbb{R} \mid d_F(\alpha)e_F(\alpha) \neq 0\}$. We let $D_F(a,x,y,z) = \det(\mathrm{Jac}_F(a,x,y,z))$, $J_F(a,x,y,z) = I_F(a,x,y,z) + (D_F(a,x,y,z))$ and compute $J_F(a,x,y,z) \cap \mathbb{R}[a]$. We get the principal ideal generated by a univariate polynomial $h_F(a)$ of degree 59. An I-optimal subscheme is $\mathcal{U}_F = \{\alpha \in \mathbb{R} \mid d_F(\alpha)e_F(\alpha)h_F(\alpha) \neq 0\}$. An open semi-algebraic subset $\mathcal{V}_F$ of $\mathcal{U}_F$ containing the point $\alpha_I = 0$ and such that the fiber over each $\alpha \in \mathcal{V}_F$ consists of 6 real points is given by the interval (α_1, α_2), where $\alpha_1 < 0$ and $\alpha_2 > 0$ are the real roots of $d_F(a)e_F(a)h_F(a) = 0$ closest to the origin. Their approximate values are $\alpha_1 = -0.17082$ and $\alpha_2 = 0.20711$.

To produce similar perturbations, we embed the system $\mathbf{g}(x,y,z)$ into the family $G(a,x,y,z) = \{G_1, G_2, G_3\}$ where

$$\begin{array}{rcl} G_1(a,x,y) & = & \frac{6}{17}x^2 + (1-a^2)xy + (-\frac{24}{85}+a)x + (-\frac{8}{85}-a)y + (-\frac{6}{85}+a^2) \\ G_2(a,x,y) & = & -\frac{3657}{697}x^2 + (\frac{538}{41}+a)xy - \frac{7565}{123}yz + (\frac{33413}{3485}+a)x \\ & & +(-\frac{44254}{3485}+a)y - \frac{15128}{3485} \\ G_3(a,x,y) & = & y^2 + 2xz + (1-2a)z^2 + (-1+a)z \end{array}$$

We denote by $I_G(a,x,y,z)$ the ideal generated by $G(a,x,y,z)$ in $\mathbb{R}[a,x,y,z]$, compute the reduced Lex-Gröbner basis of $I_G(a,x,y,z)\mathbb{R}(a)[x,y,z]$, and get

$$\{x + \tfrac{l_4(a,z)}{d_G(a)},\ y + \tfrac{l_5(a,z)}{d_G(a)},\ z^9 + \tfrac{l_6(a,z)}{e_G(a)}\}$$

where $l_4(a,z), l_5(a,z), l_6(a,z) \in \mathbb{R}[a,z]$ have degrees $\deg_z(l_4) = \deg_z(l_5) = 7$ and $\deg_z(l_6) = 8$ while $d_G(a) \in \mathbb{R}[a]$ has degree 54, and $e_G(a) \in \mathbb{R}[a]$ has degree 11. The basis has the shape prescribed by the Shape Lemma. A flat locus is given by $\{\alpha \in \mathbb{R} \mid d_{G1}(\alpha)d_{G2}(\alpha) \neq 0\}$. We let $D_G(a,x,y,z) = \det(\mathrm{Jac}_G(a,x,y,z))$, $J_G(a,x,y,z) = I_G(a,x,y,z) + (D_G(a,x,y,z))$ and compute $J_G(a,x,y,z) \cap \mathbb{R}[a]$. We get the principal ideal generated by a univariate polynomial $h_G(a)$ of degree 59. An I-optimal subscheme is $\mathcal{U}_G = \{\alpha \in \mathbb{R} \mid d_G(\alpha)e_G(\alpha)h_G(\alpha) \neq 0\}$. An open semi-algebraic subset $\mathcal{V}_G$ of $\mathcal{U}_G$ containing the point $\alpha_I = 0$ and such that the fiber over each $\alpha \in \mathcal{V}_G$ consists of 6 real points is given by the interval (α_3, α_4), where $\alpha_3 < 0$ and $\alpha_4 > 0$ are the real roots of $d_G(a)e_G(a)h_G(a) = 0$ closest to the origin. Their approximate values are $\alpha_3 = -0.02942$ and $\alpha_4 = 0.03312$.

Let $\alpha \in (\alpha_3, \alpha_4)$. According to Definition 3.7 the polynomial set $\boldsymbol{\varepsilon}(x,y) = \{-\alpha^2xy + \alpha x - \alpha y + \alpha^2,\ \alpha xy + \alpha x + \alpha y,\ -2\alpha z^2 + \alpha z\}$ is an admissible perturbation of $\mathbf{f}(x,y,z)$ and $\mathbf{g}(x,y,z)$.

We let $q \in \mathcal{Z}_{\mathbb{R}}(\mathbf{f} + \boldsymbol{\varepsilon})$ and $r \in \mathcal{Z}_{\mathbb{R}}(\mathbf{g} + \boldsymbol{\varepsilon})$ be the two perturbations of the point p. In order to compare the numerical behaviour of $\mathbf{f}$ and $\mathbf{g}$ at the real root p we compare the relative errors $\frac{\|q-p\|_2}{\|p\|_2}$ and $\frac{\|r-p\|_2}{\|p\|_2}$ for different values of α. The first column of the following table contains the values of the local condition numbers of $\mathbf{f}$ and $\mathbf{g}$ at p. The second column contains the mean values of $\frac{\|q-p\|_2}{\|p\|_2}$ and $\frac{\|r-p\|_2}{\|p\|_2}$ for 100 random values of $\alpha \in (\alpha_1, \alpha_4)$.

$\kappa_2(\mathbf{f}, p)$	$\frac{\|q-p\|_2}{\|p\|_2}$
123	0.0436
$\kappa_2(\mathbf{g}, p)$	$\frac{\|r-p\|_2}{\|p\|_2}$
1	0.0221

As in the example before, the fact that the mean values of $\frac{\|q-p\|_2}{\|p\|_2}$ are smaller than the mean values of $\frac{\|r-p\|_2}{\|p\|_2}$ suggests that p is more stable when it is considered as a root of $\mathbf{g}$ instead of as a root of $\mathbf{f}$.

Acknowledgements

All the supporting computations were performed with CoCoA (see [7]). We thank Marie-Françoise Roy and Saugata Basu for some help in the proof of Theorem 2.21. Finally, we like to spend a few words to thank the referee. His/her constructive criticism allowed us to improve the paper considerably.

REFERENCES

[1] J. Abbott - A. Bigatti - M. Kreuzer - L. Robbiano, *Computing Ideals of Points*, J. Symb. Comput. 30 (2000), 341–356.

[2] J. Abbott - M. Kreuzer - L. Robbiano, *Computing zero-dimensional Schemes*, J. Symb. Comput. 39 (2005), 31–49.

[3] S. Basu - R. Pollack - M. F. Coste-Roy, *Algorithms in Real Algebraic Geometry*, Algorithms and Computation in Mathematics Vol. 10, Springer-Verlag, 2006.

[4] C. Beltrán - L. M. Pardo, *On the probability distribution of condition numbers of complete intersection varieties and the average radius of convergence of Newton's method in the underdetermined case*, Math. Comp. 76 (259) (2007), 1393–1424.

[5] D. Bini - M. Capovani - O. Menchi, *Metodi numerici per l'algebra lineare*, Zanichelli, 1988.

[6] B. Buchberger - M. Möller, *The construction of multivariate polynomials with preassigned zeros*, In J. Calmet Editor, Proceedings of the European Computer Algebra Conference (EUROCAM '82, Lecture Notes in Comp. Sci. 144 (1982), Springer, 24–31, .

[7] CoCoATeam, CoCoA: a system for doing Computations in Commutative Algebra. Available at http://cocoa.dima.unige.it.

[8] F. Cucker - G. Malajovich - T. Krick - M. Wschebor, *A numerical algorithm for zero-counting. I: complexity and accuracy*, J. Complexity 24 (2008), 582–605.

[9] J. Dégot, *A Condition Number Theorem for Underdetermined Polynomial Systems*, Math. Comp. 70 (233) (2001), 329–335.

[10] J. W. Demmel, *The probability that a numerical analysis problem is difficult*, Math. Comp. 50 (182) (1988), 449–480.

[11] D. Eisenbud, *Commutative algebra with a view toward algebraic geometry*, Graduate Texts in Mathematics, Springer, 1995.

[12] C. Fassino, *Almost Vanishing Polynomials for Sets of Limited Precision Points*, J. Symb. Comput. 45 (2010), 19–37.

[13] C. Fassino - M. Torrente, *Simple Varieties for Limited Precision Points*, to appear in Theoret. Comput. Sci. (2012).

[14] I. M. Gelfand, M. M. Kapranov - A. V. Zelevinsky, *Discriminants, Resultants, and Multidimensional Determinants*, Birkhäuser, 2008.

[15] L. Gonzalez - H. Lombardi - T. Recio - M.-F. Roy, *Sturm-Habicht sequence*, In Proceedings of ISSAC'1989, ACM New York, USA, 136–146.

[16] N. J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, 1996.

[17] M. Kreuzer - H. Poulisse - L. Robbiano, *From Oil Fields to Hilbert Schemes*, in: L. Robbiano and J. Abbott (eds.), *Approximate Commutative Algebra*, Text and Monographs in Symbolic Computation, Springer-Verlag Wien, (2009), 1–54.

[18] M. Kreuzer - L. Robbiano, *Computational Commutative Algebra 1*, Springer, Heidelberg, 2000.

[19] M. Kreuzer - L. Robbiano, *Computational Commutative Algebra 2*, Springer, Heidelberg, 2005.

[20] G. Malajovich - J. M. Rojas, *High probability analysis of the condition number of sparse polynomial systems*, Theoret. Comput. Sci. 315 (2-3) (2004), 525–555.

[21] B. Mourrain - J. P. Pavone, *Subdivision methods for solving polynomial equations*, J. Symb. Comput. 44 (2009), 292–306.

[22] L. Robbiano - J. Abbott (eds.), *Approximate Commutative Algebra*, Text and Monographs in Symbolic Computation, Springer-Verlag Wien, 2009.

[23] M. Shub - S. Smale, *Complexity of Bezout's Theorem I: Geometric Aspects*, Journal of the American Mathematical Society, 6 (2) (1993), 459–501.

[24] A. J. Sommese - C. W. Wampler, *The numerical solution of systems of polynomials arising in engineering and science*, World Scientific, 2005.

LORENZO ROBBIANO
Dipartimento di Matematica
Università di Genova
Via Dodecaneso, 35
16146 Genova - ITALIA
e-mail: `robbiano@dima.unige.it`

MARIA-LAURA TORRENTE
Dipartimento di Matematica
Università di Genova
Via Dodecaneso, 35
16146 Genova - ITALIA
e-mail: `torrente@dima.unige.it`