Numerical Conditioning of Zero-dimensional Gröbner basis *

ANTONIO MONTES

Universitat Politècnica de Catalunya
Dept. of Appl. Math 2
e-mail: montes@ma2.upc.es

Abstract

Gröbner bases approach is useful to solve systems of algebraic equations. New algorithms given by D. Lazard [5] allows us to reduce a zero-dimensional system of algebraic equations to a family of triangular sets of polynomials. In this paper, we define and compute the condition-number of each root for a triangular set of polynomials. For the actual algorithms this is enough from a numerical point of view. If we used the original Buchberger algorithm [1], we also need to compute numerical g.c.d.'s. For completeness, we define it in this paper. We complete the study with illustrative examples.

1 Introduction.

Let $F = \{f_1, \ldots, f_n\}$ and $F = 0$ be a system of algebraic equations in the variables $\{x_1, \ldots, x_s\}$ and $G = GB(F) = \{g_1, \ldots, g_m\}$ be the Gröbner basis of the ideal generated by F with respect to the purely lexicographical order $(x_1 < x_2 < \ldots < x_s)$.

Then the affine variety defined by $F = 0$ has finitely many solutions iff for all i, $(1 \leq i \leq s)$ a power product of the form $x_i^{j_i}$ occurs among the leading power products of the polynomials g_i in G.

This leads to an automatic algorithm to detect if the system $F = 0$ has finitely many solutions, and in this case, to solve it numerically. Computing the Gröbner basis triangularizes the system. G will contain exactly one

*An earlier version [6] of this paper was published in *SIGSAM Bull.* **27**, Num. 1, (1993), 12-19.

polynomial of $K[x_1]$. The successive elimination can then be carried out by the following basic Buchberger algorithm [1], [2]

BASIC BUCHBERGER ALGORITHM

$p :=$ the polynomial in $G \cap K[x_1]$
$X_1 := \{(\alpha) \mid p(\alpha) = 0\}$
for $i := 1$ to $s - 1$ do
    $X_{i+1} := \phi$
    for all $(\alpha_1, \ldots, \alpha_i) \in X_i$ do
        $H := \{g(\alpha_1, \ldots, \alpha_i, x_{i+1}) \mid g \in G \cap K[x_1, \ldots, x_{i+1}] - K[x_1, \ldots, x_i]\}$
        $p :=$ g.c.d. of the polynomials in H
        $X_{i+1} := X_{i+1} \cup \{(\alpha_1, \ldots, \alpha_i, \alpha) \mid p(\alpha) = 0\}$
    od;
od;

Upon termination, X_s will contain all the solutions.

Actually there are known variants of this algorithm not requiring to compute g.c.d.'s., that are optimal from a theoretical point of view. Lazard [5] gives algorithms to compute the solution as the union of a family of triangular sets of polynomials. Given an order of the variables $x_1 < x_2 < \cdots < x_n$, a set of n polynomials is triangular if the leading monomial of the i-th polynomial is $x_i^{j_i}$ for $i = 1, \cdots, n$. This form of the solution is also optimal for numerical purposes when the solutions are to be computed in $\mathbb{C}^n$.

The efficiency of these algorithms depends in part on the complexity of the Gröbner basis in pure lexicographical order. It is known [7], that even if it is not always the worst order to compute Gröbner bases, it isn't in general the best. The complexity of the Gröbner basis in pure lexicographical order is extremely sensitive to a permutation of the variables [4]. Unfortunately we don't know a priori which is the best order of the variables to be used. For practical purposes, it is important to prefix an order of the variables if we have arguments for suspecting that it is a good order. We shall give arguments of this kind in the examples.

From a numerical point of view, in solving zero-dimensional systems of algebraic equations in $\mathbb{C}^n$, we will be interested in the study of the numerical-condition numbers of each triangular set of polynomials resulting from the factorized algorithms.

2 Condition-number of a root of an univariate polynomial.

Let α be a simple root of the equation $f(x) = 0$, $(f(\alpha) = 0, f'(\alpha) \neq 0)$, and let $\epsilon = 10^{-d}$ be the accuracy of the floating point arithmetic (denoted f.p.a.) we are working with. We mean that ϵ is a bound of the relative error of the representation of numbers and also of the elemental operations $+, -, *, /$. In our case, the equation is polynomial.

Let x_n be the best possible approximation of α with any numerical method. We have

$$f(x_n) = (x_n - \alpha) f'(\xi) \qquad \xi \in \mathrm{int}(x_n, \alpha) = J$$

and

$$\mid x_n - \alpha \mid \leq \frac{\mid f(x_n) \mid}{M}$$

where $\mid f'(x) \mid > M$ for $x \in J$.

Let $\bar{f}(x_n)$ be the floating point evaluation of $f(x_n)$ and $\triangle f(x_n)$ the error

$$\bar{f}(x_n) = f(x_n) + \triangle f(x_n)$$

Let $\mid \triangle f(x) \mid \leq \delta$ for $x \in J$. (Actually δ will depend on ϵ). The best possible computation of x_n will be obtained when $\bar{f}(x_n) = 0$, that implies $\mid f(x_n) \mid \leq \delta$. If $f'(x)$ doesn't change too much in $x \in J$ we have

$$\mid x_n - \alpha \mid \leq \frac{\delta}{\mid f'(\alpha) \mid}$$

and so the best floating point evaluation of α will have the relative error

$$\left| \frac{\triangle \alpha}{\alpha} \right| \leq \frac{\delta}{\mid \alpha f'(\alpha) \mid} \tag{1}$$

It's easy to generalize this result for a root of multiplicity p, obtaining

$$\left| \frac{\triangle \alpha}{\alpha} \right| \leq \frac{1}{\mid \alpha \mid} \left(\frac{\delta\, p!}{\mid f^{(p)}(\alpha) \mid} \right)^{1/p} \tag{2}$$

From now on, we shall consider only simple roots for simplicity. The trouble with multiple roots is that they turn the problem into an ill-conditioned one.

We have to evaluate now δ for a polynomial $p(x) = \sum_{i=0}^{n} a_i x^i$ in our floating point arithmetic characterized by the accuracy ϵ. We have to compute a term

$$|\,\overline{a_i x^i} - a_i x^i\,| = |\,a_i x^i((1+\delta_1)(1+\delta_2)^i(1+\mu_1)\dots(1+\mu_i) - 1\,|$$

where the factor $1+\delta_1$ stands for the representation of a_i, the factor $1+\delta_2$ for the representation of x, and the i factors $1+\mu_k$ for the i products in the term. Also $|\,\delta_i\,| < \epsilon$, $|\,\mu_i\,| < \epsilon$, for all i. As usual in this kind of calculation we use the inequality

$$|\,(1+u_1)\dots(1+u_n)-1\,| \le \frac{u}{1-\frac{u}{2}} \qquad \text{whenever} \qquad u = |\,u_1\,|+\dots+|\,u_n\,| < 2$$

The absolute error on the i-th term will be

$$|\triangle\, t_i| = |\,\overline{a_i x^i} - a_i x^i\,| \le |\,a_i x^i\,|\,(\prod_{j=1}^{j=2i+1}(1+\delta_j) - 1) \le |\,a_i x^i\,|\frac{(2i+1)\epsilon}{1-\frac{i\epsilon}{2}}$$

For reasonable f.p.a., for which $(2d+1)\epsilon < 0.01$, where d is the degree of p, the error $\triangle p(\alpha)$ is

$$|\ \triangle\, p(\alpha)\,| \le \sum_{i=0}^{n}\ |\triangle\, t_i|\ \le 1.006\,\epsilon \sum_{i=0}^{n}\ |\ a_i\alpha^i\ |\ (2i+1) \tag{3}$$

If $(2d+1)\epsilon < 0.01$, in order to be independent of ϵ, the condition-number C_α of the root α is defined [3]:

$$|\frac{\triangle\alpha}{\alpha}|\ \frac{1}{\epsilon} \le C_\alpha = \frac{\sum_{i=0}^{n}\ |\ a_i\alpha^i\ |\ (2i+1)}{|\ \sum_{i=1}^{n} i a_i \alpha^i\ |}\,1.006 \tag{4}$$

C_α defines a centered confidence-interval containing the exact root α.

$$\alpha = \bar{\alpha}(1+\mu C_\alpha), \qquad |\ \mu\ |\ < \epsilon \tag{5}$$

The major problem is to dispose of a set of numerical routines that are able to isolate and compute correctly all the complex roots of $p(x)$ with the best possible accuracy in the actual floating point system. This is not always possible to do, and the algorithms can fail. If the polynomial is not very ill-conditioned, this objective can be reached without difficulty, e.g. by the procedure `fsolve` in *Maple*. Then we shall be able to compute its condition-number and the confidence-interval.

3 Condition-number of a triangular set of polynomials.

We consider now a triangular set of two polynomials in two variables.

$$\left\{ \begin{array}{l} \sum_i a_i x^i = 0 \\ \sum_{ij} b_{ij} x^i y^j = 0 \end{array} \right. \tag{6}$$

Let α_k be the roots of the first equation and $\bar{\alpha}_k$ the approximate numerical solutions computed with accuracy ϵ.

Upon substitution of $x = \bar{\alpha}_k$ in the second equation we have to solve the roots β_{kl} in y, and this is also an univariate problem, but no more with integer coefficients. From now on, we shall suppress the subindex of the roots. We have to evaluate the following term

$$\begin{array}{l} \mid \overline{b_{ij}\bar{\alpha}^i\beta^j} - b_{ij}\alpha^i\beta^j \mid = \\ = \mid b_{ij}\bar{\alpha}^i\beta^j(1+\delta_1)(1+\delta_2)^i(1+\delta_3)^j(1+\mu_1)\ldots(1+\mu_{i+j}) - b_{ij}\alpha^i\beta^j \mid = \\ = \mid b_{ij}\alpha^i\beta^j[(1+\mu_0 C_\alpha)^i(1+\mu_1)\ldots(1+\mu_{2i+2j+1}-1] \mid) \end{array}$$

where $\mid \mu_i \mid \leq \epsilon$. We obtain

$$\mid \triangle\, t_{ij} | = | \overline{b_{ij}\bar{\alpha}^i\beta^j} - b_{ij}\alpha^i\beta^j \mid = \mid b_{ij}\alpha^i\beta^j \left[(1+\mu_0 C_\alpha)^i (\prod_{k=1}^{2i+2j+1} (1+\mu_k)) - 1 \right] \mid$$

that for reasonable f.p.a, i.e., $\epsilon d_1 C_\alpha < 0.01$ and $(1+2d)\epsilon < 0.01$, where d_1 y d are respectively the degree with respect to x and the total degree of the second polynomial, gives

$$\mid \triangle\, t_{ij} | \leq \epsilon \mid b_{ij}\alpha^i\beta^j \mid \big(iC_\alpha + 2i + 2j + 1 \big)\, 1.011$$

Finally, the condition-number of β will be

$$\frac{1}{\epsilon} | \frac{\Delta\beta}{\beta} | \leq \frac{1}{\epsilon} \frac{\mid \triangle\, f(\beta) \mid}{\mid \beta f'(\beta) \mid} \leq C_\beta = \frac{\sum_{ij} \mid b_{ij}\alpha^i\beta^j \mid \big(iC_\alpha + 2i + 2j + 1 \big)}{\mid \sum_{ij} j b_{ij}\alpha^i\beta^j \mid} 1.011 \tag{7}$$

This process can be easy generalized to a triangular set of n polynomials. The result is

$$\frac{1}{\epsilon} | \frac{\triangle x_n}{x_n} | \leq C_n = \tag{8}$$

$$= \frac{\sum\limits_{i_1 \ldots i_n} \mid a^{(n)}_{i_1 \ldots i_n} {x_1}^{i_1} \ldots {x_n}^{i_n} \mid \big(2i_n + 1 + \sum\limits_{k=1}^{n-1} i_k (C_k + 2) \big)}{\mid \sum\limits_{i_1 \ldots i_n} i_n a^{(n)}_{i_1 \ldots i_n} {x_1}^{i_1} \ldots {x_n}^{i_n} \mid} f(n)$$

where

$$x_k = \bar{x}_k(1+\mu_k C_k) \qquad |\, \mu_k \,| \le \epsilon \qquad \text{for} \quad 1 \le k \le n$$

and

$$f(n) = \frac{1}{1 - 0.005n}$$

is the second order corrector coefficient, which can in practice be neglected.

The conditions on ϵ for this formulas to be valid are

$$(2d+1)\epsilon < 0.01, \qquad \epsilon d_k C_k < 0.01 \quad \text{for} \quad k = 1 \ldots n-1$$

where d is the biggest total degree and d_k the biggest x_k degree of all the polynomials in the set. With these simple conditions, the C_k will not depend on ϵ.

If the coefficients of the polynomials in the given system are not complex, then if $(\alpha_1, \ldots, \alpha_s)$ is a solution, its complex conjugate $(\alpha_1^*, \ldots, \alpha_s^*)$ is obviously also a solution and has the same condition-numbers. This reduces to the half the calculations for the complex roots.

In solving interesting applied problems of this kind, we often have to compute Gröbner basis with parameters. In this case, the coefficients of the polynomials will be functions of the parameters, and for a concrete case we will have to specialize the resulting basis by evaluating the coefficients numerically. This will comport also a rounding error in the coefficients. If the coefficients are polynomial functions of the parameters we can use formulas like (3) to evaluate the rounding error for a coefficient and then divide by it and ϵ to obtain its condition number. We will have to compute the rounding error of polynomials in several variables someones having condition numbers (they are results of previous calculations) and someones having only floating point representation errors ($C_k = 0$). Denoting $P = \sum_i p_i$, the condition number in isolating x_n in P will be

$$\frac{1}{\epsilon}\,|\,\frac{\triangle x_n}{x_n}\,| \le C_n \le \frac{\sum_i \left\{|\,p_i\,| \left[1 + 2\deg(p_i) + \sum_{k=1}^{n-1} \deg(p_i, x_k)\, C_k\right]\right\} f(n)}{|\sum_i \deg(p_i, x_n) p_i|}$$

4 Computation of numerical g.c.d.

If we use the basic Buchberger algorithm without separating in triangular sets, we have another difficulty in the numerical solution, namely, the evaluation of g.c.d.'s. It is obvious that we cannot use nor Euclid's algorithm nor any other exact algorithm, because computing with approximations, we generally would obtain 1 as result. The computation of numerical g.c.d. can introduce new avoidable errors in the solution, and is not recommended. But if we do so, we have to define how to compute the numerical g.c.d..

For each approximate root $\bar{\alpha}$ of a polynomial $g \in H$, we obtained in §3 a confidence-interval I_α containing the exact root α. We can define the *numerical solution* of a polynomial $g \in H$ as the union of the confidence-intervals of all its roots.

$$r(g(\alpha_1, \ldots, \alpha_i, x)) = \bigcup_{g(x=\alpha)=0} I_\alpha$$

and the *numerical* g.c.d. as

$$\text{n.g.c.d.}\,(H) = \bigcap_{g \in H} r(g)$$

These definitions depend of the f.p.a. used. The computer algebra systems, like *Maple*, allow us to change the numerical accuracy. So we can verify if the computed n.g.c.d. remain stable with increasing accuracy of the f.p.a. By stable we mean that the n.g.c.d. conserves the same number of disjoint intervals and that each one has a common intersection with the corresponding one's obtained with fewer accuracy.

5 Examples.

We want to consider two characteristic examples. All calculations are done with *Maple*.

Example 1. Consider the following system of algebraic equations

$$\left\{ \begin{array}{l} x^3 - 5\,x^2 + 4\,x - 2 = 0 \\ y^2 - 2\,yx^2 + 4\,yx - 3\,x^5 + 4\,x^3 + 8 = 0 \\ 10\,z^3 + 5\,z^2y^2x^3 + 8\,z^2y - 7\,z^2x + 1 = 0 \end{array} \right.$$

which, obviously has finitely many solutions. For each of the three values of x we shall have two values of y, and for each pair (x, y), three values of z. In this case, even not being necessary, it's not difficult to compute Gröbner basis for the pure lexicographical order with $x < y < z$, namely

$$
\begin{aligned}
G \;=\; & [10\,z^3+450\,yx^2z^2-380\,yxz^2+228\,z^2y+85955\,z^2x^2-72837\,z^2x \\
& +41410\,z^2+1, \\
& y^2-2\,yx^2+4\,yx-241\,x^2-110+206\,x, \\
& x^3-5\,x^2+4\,x-2]
\end{aligned}
$$

Using f.p.a. with `Digits:=10` and Buchberger basic algorithm, combined with the numerical method given in §3, we obtain all the roots and its corresponding condition-numbers. We only give the significant digits of the roots according to its condition-numbers and the f.p.a. used ($\epsilon = 10^{-\mathrm{Digits}}$). Remember that condition-numbers are ϵ-independent if the f.p.a. is the same in all computations.

x (C_x)	y (C_y)	z	(C_z)
.423621199+	$-.9237683$	$.18488 \pm .16221i$	(81171)
.549684246i	$\pm 1.876131i$	$-.21958 \mp .14024i$	(95782)
(9)	(434)	$.43461 \mp 1.4987i$	(172503)
	-1.0161122	$-.11034 \mp .10367i$	(32799)
	$\pm 3.1434360i$	$.10533 \pm .10404i$	(30672)
	(282)	$.16657 \mp 4.4581i$	(63432)
4.15275760	-50.141367	-89983.836	(134)
(17)	(47)	$.61750574 \cdot 10^{-11} \pm$	
		$\pm .10541872 \cdot 10^{-2} i$	(64)
	68.021128	-165730.49	(83)
	(40)	$.18203931 \cdot 10^{-11} \pm$	
		$\pm .77678137 \cdot 10^{-3} i$	(38)

Roots and condition-numbers of the system

To compute the lexicographical Gröbner basis for the order $z < y < x$ of the variables is a difficult task. The reason being that our system has an intrinsic order of the variables: for each x then for each y and then for each z. If we reverse the order, the Gröbner basis must obviously contain a polynomial in z of order $3 \cdot 2 \cdot 3 = 18$, and probably more than one polynomial in z, y in order to restraint the 18 z-solutions to 6 values of y and so on for x. This structure is much more complex than the previous one for the intrinsic natural order.

We can formulate an heuristic rule for choosing the order of variables for computing Gröbner basis for lexicographical order of a zero-dimensional set of algebraic equations. First of all we shall respect the order in which the given system is nearest to triangular form. Second, if we have some *physical arguments* to expect some priority of the variables in the sense that x_1 takes less different values than x_2 and so on, or that x_2 depends in some sense of x_1, x_3 of (x_1, x_2) and so on, then we must take the order $x_1 < x_2 < \cdots < x_n$, taking as first variable the more principal one in the actual physical problem.

Example 2. Consider now the following system of algebraic equations,

$$\left\{ \begin{array}{l} x^5 - 6\,x^4 - 11\,x^3 + 26\,x^2 + 105\,x - 150 = 0 \\ x^3y^2 + x^3y - 8\,x^3 + 3\,x^2y^2 + 3\,x^2y - 24\,x^2 + xy^2 + xy - 8\,x \\ -10\,y^2 - 10\,y + 80 = 0 \\ x^2y^3 + 7\,x^2y - 4\,x^2 - 9\,xy^3 - 63\,xy + 36\,x + 15\,y^3 + 105\,y - 60 = 0 \\ x^2z^2 + 5\,x^2z - 3\,x^2 - 9\,xz^2 - 45\,xz + 27\,x + 15\,z^2 + 75\,z - 45 = 0 \\ y^3z^2 - 3\,y^3z + 5\,y^3 + 7\,yz^2 - 21\,yz + 35\,y - 4\,z^2 + 12\,z - 20 = 0 \end{array} \right.$$

This is an illustrative example in which the basic algorithm would comport n.g.c.d.'s evaluations. Nevertheless, the factorized algorithm `gsolve` of *Maple*, gives the solution as a family of two separated triangular sets that can be easy solved numerically, and for which formulas given in §3 are directly applicable:

$$\begin{array}{rcl} G & = & [[x^2 - 9\,x + 15, y^2 + y - 8, z^2 - 3\,z + 5], \\ & & [z^2 + 5\,z - 3, y^3 + 7\,y - 4, x^3 + 3\,x^2 + x - 10]] \end{array}$$

There are $8 + 18 = 26$ solutions very simple to compute for which we only need the condition-number defined in §2.

References

[1] B.Buchberger, A Survey on the Method of Gröbner Bases for Solving Problems in Connection with Systems of Multivariate Polynomials, "Symbolic and Algebraic Computation by Computers", ed. N. Inada and T. Soma, World Scientific Publishing Co., (1985), 69-83.

[2] D. Cox, J. Little, D. O'Shea, "Ideals, Varieties, and Algorithms", Springer-Verlag, (1991).

[3] G. Dahlquist, A. Björck, "Numerical Methods", Prentice Hall Series in Automatic Computation, (1974), 51-55, 238-242, 246-248.

[4] R. Gebauer, H. Kredel, "The Buchberger Algorithm Package in SAC-2", Universität Heidelberg, Institut für Mathematik, Im Neuheimer Feld 294, D-6900 Heidelberg, West Germany (c/o Prof. W. Böge). (1983).

[5] D. Lazard, Solving Zero-dimensional Algebraic Systems, *J. Symb. Comp.* **13**, (1992), 117-131.

[6] A. Montes, Numerical conditioning of a system of algebraic equations with a finite number of solutions using Gröbner bases, *SIGSAM Bull.* **27**, Num. 1, (1993), 12-19.

[7] Alyson A. Reeves, The worst order is not always the lexicographic order, *SIGSAM Bull.*, **25**, Num. 4, (1991), 18-19, + erratas Num 5, p.13.