J. Symbolic Computation (1992) **13**, 133–138

A Gröbner Basis Technique for Padé Approximation

PATRICK FITZPATRICK AND JOHN FLYNN

Department of Mathematics, University College, Cork, Ireland

(*Received* 27 *August* 1990)

We consider solving for a and b the congruence $a \equiv bh \bmod \mathbf{I}$, where a, b and h are (multivariable) polynomials and $\mathbf{I}$ is a polynomial ideal. This is a generalization of the well-known problem of Padé approximation of which decoding Hensel codes is a special case. We show how Gröbner bases of modules may be used to generalize the Euclidean algorithm method of solution of the 1-variable problem.

1. Introduction

Let $\mathbf{R} = k[x_1, \ldots, x_n]$ where k is a field. Interpreting a polynomial $h \in \mathbf{R}$ of total degree m as a truncation of a formal power series we may ask for relatively prime polynomials a, b where $b(0) \neq 0$ such that the expansion of a/b as far as terms of degree m is equal to h. This is the well-known problem of Padé approximation and there is an extensive literature on both the 1-variable and the multivariable cases. Usually, restrictions are placed on the degrees of a and b in order that the required solution (a, b) be unique, and it is clear that the solution may be determined by solving the appropriate system of linear equations and cancelling any common factors. This problem may be regarded as a special case of that considered in this note, namely, solving (for a and b) the congruence

$$a \equiv bh \bmod \mathbf{I}, \qquad (*)$$

where $\mathbf{I}$ is an ideal in $\mathbf{R}$ and h is a given polynomial in $\mathbf{R}$. In general we shall require that a and b be relatively prime but drop the condition that $b(0) \neq 0$.

For Padé approximation, $\mathbf{I}$ is the ideal generated by the monomials of total degree $m+1$, and it is well known that in the 1-variable case the solution may be determined (assuming that arithmetic in k is exact) using the extended Euclidean algorithm or the Berlekamp–Massey algorithm. Neither of these algorithms is valid for $n > 1$ and one may ask for a generalization of these *algebraic* (as opposed to *linear*) techniques of solution. Sakata (1990) has given an extension to n variables of the Berlekamp–Massey algorithm and it is our purpose to give a corresponding generalization of the method based on the extended Euclidean algorithm.

The natural context for such a generalization is that of Gröbner bases of polynomial modules. Indeed, our initial approach to the problem was motivated by Buchberger *et al.* (1985) where attention is focused on multivariable Hensel codes. These codes were developed extensively by Gregory & Krishnamurthy (1984) and Krishnamurthy (1986) as a means of performing exact arithmetic with rationals a/b where a, b are 2integers or polynomials in one or more variables. The decoding of a Hensel code can be regarded as a special case of Padé approximation.

0747-7171/92/020133+06 $03.00/0

In section 2 we show by an elementary argument that Gröbner bases of modules can be used to solve (*) for *any* ideal I. When certain restrictions are imposed on the leading terms of a and b (for example, degree bounds), the solution is the *unique reduced element of least leading term* in the solution module $\mathbf{M} = \{(a, b) \in \mathbf{R}^2: a \equiv bh \bmod \mathbf{I}\}$ relative to an order $<_w$ on terms in $\mathbf{R}^2$ which is induced from a weight vector w obtained from these restrictions. This element must appear in any Gröbner basis of $\mathbf{M}$ relative to $<_w$. This is sufficient to provide the algebraic solution required for Padé approximation, and indeed to enable *all* the Padé approximants for given h, $\mathbf{I}$ to be determined. Section 3 contains some examples.

In terms of complexity, the currently known algorithms for calculating Gröbner bases (of syzygy modules—see Bayer & Stillman (1988), for example) do not appear to provide any practical advantage over linear methods for solving (*), although the particular form of the ideal I used may make the calculation very much better than the worst case. Consequently, our contribution must be seen primarily as setting a theoretical context rather than providing a practical algorithm for Padé approximation.

2. Gröbner Bases of Polynomial Modules and Padé Approximation

Our reference for this section is Möller & Mora (1986) where the necessary background material may be found as well as references to original sources where appropriate. Well-known properties of Gröbner bases are used freely without further comment. We denote by T the set of terms $\alpha = x_1^{a_1} \cdots x_n^{a_n}$ and choose a term order $<_T$ on T. Throughout the paper $<_T$ will denote graduated lexicographic order with $x_1 < \cdots < x_n$. The leading term of a polynomial p is denoted $\mathrm{Lt}(p)$. The *total degree* $\tau(\alpha)$ of α is the sum $a_1 + a_2 + \cdots + a_n$ and the total degree $\tau(p)$ of p is the maximum of the total degrees of its terms. If $\mathbf{R}^r$ denotes the free module of rank r over $\mathbf{R}$, the set T_r of terms in $\mathbf{R}^r$ is defined as $\{(0, \ldots, \varphi, 0, \ldots, 0): \varphi \in T\}$ or, equivalently, as the set $\{(\varphi, k): \varphi \in T, 1 \le k \le r\} = T \times \{1, 2, \ldots, r\}$. We shall use the latter representation exclusively. (Ordered pairs are used in several different contexts, but in each case the interpretation will be clear from the context.) All our modules are submodules of $\mathbf{R}^r$ for some r, and have the appropriate graded structure induced from that of $\mathbf{R}^r$. If $w = (\psi_1, \ldots, \psi_r)$ where $\psi_j \in T$, is any *weight vector*, then the term order $<_w$ on T_r *induced by* $<_T$ *and* w is defined by $(\sigma, i) <_w (\tau, j)$ if $\psi_i \sigma <_T \psi_j \tau$ or $\psi_i \sigma = \psi_j \tau$ and $i < j$. We also say that a Gröbner basis of a module relative to the term order $<_w$ is *adapted to* w (when $<_T$ is understood). For any submodule $\mathbf{U}$ of $\mathbf{R}^r$ we write $\mathrm{Lt}(\mathbf{U})$ for the submodule generated by $\{\mathrm{Lt}(u): u \in \mathbf{U}\}$.

If $\{g_1, \ldots, g_r\}$ is a set of polynomials in $\mathbf{R}$ which generate an ideal $\mathbf{G}$ and if $<_T$ is a term order on T, then Algorithm 4.2 of Möller & Mora (1986) produces further elements $g_{r+1}, \ldots, g_s$ such that $G = \{g_1, \ldots, g_s\}$ is a Gröbner basis of $\mathbf{G}$. The algorithm also gives a Gröbner basis for the module of syzygies

$$\mathbf{syz}(G) \subseteq \mathbf{R}^r, \ \mathbf{syz}(G) = \left\{(h_1, \ldots, h_r): \sum_{j=1}^{r} h_j g_j = 0\right\},$$

with respect to the term order on T_r induced by $<_T$ and $w = (\mathrm{Lt}(g_1), \ldots, \mathrm{Lt}(g_r))$. A term φ is in *normal form mod* $\mathbf{G}$ if $\varphi \notin \mathrm{Lt}(\mathbf{G})$ and a polynomial p is in *normal form mod* $\mathbf{G}$ if each of its terms is in normal form mod $\mathbf{G}$.

Referring now to congruence (*), let $\{p_1, \ldots, p_m\}$ be a basis for the ideal $\mathbf{I} \subseteq \mathbf{R}$ and suppose that h is in normal form mod I. For simplicity we may assume that the given basis is a Gröbner basis of I. It is clear that the set of all polynomial pairs (a, b) satisfying

(*) forms a module **M** each of whose elements is associated with a solution $(a, b, c_1, \ldots, c_m)$ of the equation

$$a(-1)+bh+c_1p_1+\cdots+c_mp_m=0,$$

and thus with an element of the syzygy module $\mathbf{syz}(F)\subseteq \mathbf{R}^{m+2}$ where $F=\{-1, h, p_1, \ldots, p_m\}$. In view of the special form of the set F, it is easy to determine the basis produced by the algorithm mentioned above and thence derive a basis for **M** by projection on the first two components. We say that $(a, b)\in \mathbf{M}$ is *reduced mod* **I** if both a and b are in normal form mod **I**. (Thus the reduced solutions are just representatives of the factor module of **M** by the submodule generated by elements of the forms $(p, 0)$ and $(0, p)$ for $p\in \mathbf{I}$.)

THEOREM 2.1. *The set* $\{(h, 1, 0, \ldots, 0), (p_j, 0, \ldots, 0, 1, 0, \ldots, 0), 1\leq j\leq m\}$ *(where the second vector has a* 1 *in the* $j+2$ *position) is a Gröbner basis for* $\mathbf{syz}(F)$ *with respect to the term order on* T_{m+2} *induced by* $<_T$ *and* $(1, \mathrm{Lt}(h), \mathrm{Lt}(p_1), \ldots, \mathrm{Lt}(p_m))$. *The set* $M=\{(h, 1), (p_j, 0), 1\leq j\leq m\}$ *is a Gröbner basis for* **M** *with respect to the term order* $<_w$ *on* T_2 *induced by* $<_T$ *and* $w=(1, \mathrm{Lt}(h))$. *Moreover,* $(h, 1)$ *is (up to a constant multiple) the unique reduced element of* **M** *of least leading term relative to* $<_w$.

PROOF. The first part follows from Möller & Mora (1986, Theorem 7.8). For the second, observe that M is a basis by the projection and $\mathrm{Lt}(\mathbf{M})$ is generated by $\{(1, 2), (\mathrm{Lt}(p_j), 1), 1\leq j\leq m\}\subseteq T\times\{1, 2\}$ relative to $<_w$. It is clear that $(h, 1)$ is reduced mod **I** and that it has leading term $(1, 2)$. If $(a, b)\in \mathbf{M}$ has leading term less than this then $\mathrm{Lt}(a, b)$ has the form $(\varphi, 1)$ where $\varphi <_T \mathrm{Lt}(h)$. Since M is a Gröbner basis $(\varphi, 1)$ is a multiple of $(\mathrm{Lt}(p_j), 1)$ for some j and hence (a, b) is not reduced. Uniqueness is immediate.

Since $(h, 1)$ is a reduced solution of (*) it follows that for any term order $<$ on T_2 there exists a reduced element in **M** of least leading term relative to $<$ which is unique up to a constant multiple. We call such an element the *minimal reduced solution of* (*) *relative to* $<$. Our aim is to show that under certain conditions on the leading terms of a, b and on the ideal **I**, *every* reduced solution of (*) arises as the minimal reduced solution relative to an order $<_w$ induced from some weight vector w. We give three versions of the condition which are appropriate in different circumstances (illustrated in the examples in the next section) as follows. Let (a, b) be a reduced solution of (*). Define the

total degree condition: $\tau(a)\leq k$, $\tau(b)\leq m$, where k, m are non-negative integers, and $\tau(\varphi)>k+m$ for all $\varphi\in \mathrm{Lt}(\mathbf{I})$. We say that $(a, b, \mathbf{I})$ satisfies $tdc(k, m)$

strong term order condition: $\mathrm{Lt}(a)\leq_T \varphi$, $\mathrm{Lt}(b)\leq_T \psi$, where $\varphi, \psi\in T$ and φ, ψ, **I** satisfy: for all $\rho, \sigma\in T$ with $\rho\leq_T \varphi$, $\sigma\leq_T \psi$ the product $\rho\sigma$ does not lie in $\mathrm{Lt}(\mathbf{I})$. We say $(a, b, \mathbf{I})$ satisfies $stoc(\varphi, \psi)$

weak term order condition: $\mathrm{Lt}(a)\leq_T \varphi$, $\mathrm{Lt}(b)\leq_T \psi$, where $\varphi, \psi\in T$ and φ, ψ, **I** satisfy: for all $\rho, \sigma\in T$ with $\rho\leq_T \varphi$, $\sigma\leq_T \psi$ *and* $\rho, \sigma\notin \mathrm{Lt}(\mathbf{I})$ the product $\rho\sigma$ does not lie in $\mathrm{Lt}(\mathbf{I})$. We say $(a, b, \mathbf{I})$ satisfies $wtoc(\varphi, \psi)$

Notice that the total degree condition implies a strong term order condition (take $\varphi=x_n^k$, $\psi=x_n^m$). Also, the strong term order condition implies the weak term order condition since if $\rho\sigma$ does not lie in $\mathrm{Lt}(\mathbf{I})$ for *any* ρ, σ then certainly $\rho=\rho 1$ and $\sigma=1\sigma$ do not lie in $\mathrm{Lt}(\mathbf{I})$: the essential difference is that under the weak condition some of the terms ρ, σ are allowed to lie in $\mathrm{Lt}(\mathbf{I})$, whereas this is not the case under the strong condition.

LEMMA 2.2. *Let (a, b) be a reduced solution of congruence (*) where a and b are relatively prime and $(a, b, \mathbf{I})$ satisfies wtoc(φ, ψ). Then every reduced solution (a_1, b_1) such that $(a_1, b_1, \mathbf{I})$ satisfies wtoc(φ, ψ) is a multiple $p(a, b) = (pa, pb)$ for some $p \in \mathbf{R}$.*

PROOF. If (a_1, b_1) is a reduced solution such that $(a_1, b_1, \mathbf{I})$ satisfies wtoc(φ, ψ) then $a_1 b - a b_1$ lies in $\mathbf{I}$. But $\mathrm{Lt}(a_1 b - a b_1) = \rho\sigma$ where $\rho \leq_T \varphi$, $\sigma \leq_T \psi$ and since both solutions are reduced neither ρ nor σ lies in $\mathrm{Lt}(\mathbf{I})$. Thus $\rho\sigma \notin \mathrm{Lt}(\mathbf{I})$ and consequently $a_1 b - a b_1 = 0$ which implies the result.

The next lemma shows that by adapting the term order to the weight vector (ψ, φ) where φ and ψ are the terms appearing in the term order conditions, we make all the *relevant* terms lowest in order.

LEMMA 2.3. *Let $w = (\psi, \varphi)$ where $\psi, \varphi \in T$. Let $S = \{(\rho, 1), (\sigma, 2)\} \subseteq \mathbf{R}^2$ where $\rho \leq_T \varphi$ and $\sigma \leq_T \psi$. Then in the term order induced by $<_T$ and w every element of S is less than any term $(\gamma, 1)$ with $\gamma >_T \varphi$ and any term $(\delta, 2)$ with $\delta >_T \psi$.*

PROOF. We have $(\rho, 1) \leq_w (\varphi, 1) <_w (\gamma, 1)$. Also, $(\sigma, 2) \leq_w (\psi, 2) <_w (\gamma, 1)$ since $\varphi\psi <_T \psi\gamma$. A similar argument for $(\delta, 2)$ completes the proof.

Finally, we consider solving (*) as finding a minimal element in a Gröbner basis of $\mathbf{M}$ under a term order adapted to the conditions on a, b and $\mathbf{I}$.

THEOREM 2.4. *Suppose that (*) has a reduced solution (a, b) with a and b relatively prime such that $(a, b, \mathbf{I})$ satisfies wtoc(φ, ψ) and let $w = (\psi, \varphi)$. Then (a, b) is the minimal reduced solution relative to $<_w$. A constant multiple of (a, b) appears in any Gröbner basis of $\mathbf{M}$ under this order.*

PROOF. The previous lemma shows that any reduced solution with leading term less than $\mathrm{Lt}(a, b)$ also satisfies wtoc(φ, ψ). Now Lemma 2.2 provides a contradiction. If M is a Gröbner basis of $\mathbf{M}$ relative to $<_w$ then M contains some *reduced* element with leading term at most $\mathrm{Lt}(a, b)$ and by minimality this must be (a, b) up to a constant multiple.

3. Examples

As noted in the Introduction we now have an algebraic technique for Padé approximation. It is instructive to see how this applies even in the 1-variable situation. The following example is derived from the calculations indicated in Knuth (1981, Exercise 13, p. 515).

EXAMPLE 3.1. Let $h = 7z^3 + 3z^2 + z + 1$ in $\mathbf{Q}[z]$. Then there are "essentially" four Padé approximants (a, b) to h modulo $\mathbf{I} = (z^4)$, namely,

$$(h, 1), (-2z^2 + 4z - 3, 7z - 3), (z - 1, z^2 + 2z - 1), (1, -2z^3 - 2z^2 - z + 1).$$

These may be determined according to Proposition 2.4 by finding minimal elements in Gröbner bases of the module $\mathbf{M}$ generated by $\{(h, 1), (z^4, 0)\}$ relative to the term orders adapted to the weight vectors $(1, z^3)$, (z, z^2) (equivalently $(1, z)$), (z^2, z) (equivalently $(z, 1)$) and $(z^3, 1)$, respectively.

The next two examples may be regarded as illustrating Padé approximation in two variables, or alternatively, as decoding Hensel codes (over $\mathbf{F}_2$). In the first we have a restriction on the leading terms of a and b in the form of a strong term order condition, while in the second we have a total degree condition.

EXAMPLE 3.2. Let $h = y^3+x^3+y^2+xy+x^2+y+x+1$ in $\mathbf{F}_2[x, y]$. Suppose that (a, b) exists satisfying congruence (*) with $\mathrm{Lt}(a) \leq_T xy$, $\mathrm{Lt}(b) \leq_T y$ and $\mathbf{I}$ generated by the monomials of total degree 4. Thus $(a, b, \mathbf{I})$ satisfies $stoc(xy, y)$. The term order $<_w$ adapted to $w = (y, xy)$ (equivalently $(1, x)$) is

$$(1,1) <_w (x,1) <_w (1,2) <_w (y,1) <_w (x^2,1) <_w (x,2) <_w (xy,1) <_w (y,2) <_w \ldots$$

which illustrates Lemma 2.3. The solution module $\mathbf{M}$ has as basis given by Theorem 2.1 the set $M = \{(h, 1), (x^4, 0), (x^3y, 0), \ldots, (y^4, 0)\}$. Converting this to a basis relative to $<_w$ we obtain

$$\{(h,1), (x^4,0), (x^3,x^3), (xy^2+x^2y+x^3+x^2+x+1, y+1),$$
$$(xy+1, y+x+1), (x^2y+x^3+x^2, x^2)\},$$

in which the leading terms of the elements are $(y^3, 1)$, $(x^4, 1)$, $(x^3, 2)$, $(xy^2, 1)$, $(y, 2)$, $(x^2y, 1)$, respectively. Clearly the fifth element in the list is the desired solution.

EXAMPLE 3.3. Let $h = xy^3+x^4+y^3+xy^2+x^3+x^2+y+x$ in $\mathbf{F}_2[x, y]$. Suppose that h is the Hensel code (Padé approximant) for (a, b) relative to the ideal $\mathbf{I}$ generated by the monomials of total degree 5, where a and b are restricted to have total degree (at most) 2. Then $(a, b, \mathbf{I})$ satisfies $tdc(y^2, y^2)$ and we seek a Gröbner basis for $\mathbf{M}$ adapted to the weight vector $w = (y^2, y^2)$ (equivalently $(1, 1)$). This is just the well-known *term-order–position order* (TO–POS)

$$(1,1) <_w (1,2) <_w (x,1) <_w (x,2) <_w \ldots$$

Converting the basis given in Proposition 2.1 to a basis relative to this order we obtain

$$\{(x^3y+x^2y+x^3, x^3+x^2), (0, x^4), (x^4+x^2y+x^3, x^2), (xy+y+x, y^2+x+1),$$
$$(xy^2+x^3, x^3+xy+x^2), (0, x^3y), (y^3+x^2y+xy+y+x, y^3+x^2y+xy+x+1)\},$$

in which it is clear that the fourth element is the desired solution.

We end with a 3-variable example which illustrates the use of an ideal $\mathbf{I}$ other than that generated by monomials of a given total degree and also provides a situation in which it is appropriate to use the weak version of the term order condition. In the foregoing examples the minimal reduced solution is in fact the solution with least leading term reduced or not. This is not the case in the next example.

EXAMPLE 3.4. Let $h = x^2yz^2+xyz^2+x^2z^2+x^2yz+xz^2+xyz+x^2z+x^2y+xz+xy+x^2+x$ and suppose that h is invertible in the algebra $\mathbf{F}_2[x, y, z]/\mathbf{I}$ where $\mathbf{I}$ is generated by $\{x^3+x+1, y^3+y+1, z^3+z+1\}$. Finding the inverse of h amounts to solving (*) for (a, b) where $(a, b, \mathbf{I})$ satisfies $wtoc(1, x^2y^2z^2)$. Here $(a, b, \mathbf{I})$ *does not* satisfy $stoc(1, x^2y^2z^2)$. A Gröbner basis of $\mathbf{M}$ relative to $<_w$ where $w = (x^2y^2z^2, 1)$ is as follows:

$$\{(1, xy^2z^2+y^2z^2+xyz^2+yz^2), (0, x^3+x+1), (0, y^3+y+1), (0, z^3+z+1)\}$$

and so the inverse of h is $xy^2z^2+y^2z^2+xyz^2+yz^2$. We observe that the leading terms of the last three basis elements are all *less* than the leading term $(xy^2z^2, 2)$ of the first element which is the minimal *reduced* solution.

We wish to thank Graham H. Norton for helpful correspondence relating to this work and for his careful comments on an earlier draft. We are also indebted to an anonymous referee for helpful comments.

Note added in proof:

REMARK. The 1-variable technique can be applied to the decoding problem for BCH and Goppa error correcting codes, which is essentially the solution of (*) relative to a total degree condition. This provides a new decoding algorithm for these codes (in practice equivalent to that based on the extended Euclidean algorithm).

References

Bayer, D., Stillman, M. (1988). On the complexity of computing syzygies. *J. Symbolic Computation* **6**, 135–147.

Buchberger, B., Krishnamurthy, E. V., Winkler, F. (1985). Gröbner bases, polynomial remainder sequences and multivariable Hensel codes. In: Bose, N. K. (ed.) *Multidimensional Systems Theory*, 252–256. Dordrecht: Reidel.

Gregory, R. T., Krishnamurthy, E. V. (1984). *Methods and Applications of Error-free Computation.* New York: Springer-Verlag.

Knuth, D. E. (1981). *The Art of Computer Programming: Vol. 2—Seminumerical Algorithms,* 2nd edn. Reading, Mass.: Addison-Wesley.

Krishnamurthy, E. V. (1986). *Error-free polynomial matrix computation.* New York: Springer-Verlag.

Möller, H. M., Mora, F. (1986). New constructive methods in classical ideal theory. *J. Algebra* **100**, 138–178.

Sakata, S. (1990). Extension of the Berlekamp-Massey algorithm to n dimensions. *Information and Computation* **84**, 207–239.