Exact Solutions of Linear Equations with Rational Coefficients by Congruence Techniques*

By I. Borosh and A. S. Fraenkel

1. Introduction. Sometimes one is interested in exact solutions of linear equations and cannot tolerate any errors at all, be it round-off errors, truncation errors or otherwise. In such situations division cannot be used. Eliminating variables by cross-multiplying leads to a tremendous growth in the size of intermediate results. This is well illustrated in an example of J. Barkley Rosser [5], in which a 6 × 6 matrix is given whose elements are integers in the interval $[-192, 195]$. Intermediate results of the order of 10^{29} are obtained after cross-multiplying. Therefore Rosser proposes to accomplish the elimination by means of a long sequence of additions and cross-multiplications by small numbers only. However, as can be seen from his example, numbers still grow rather fast, and for large matrices most of the computation would still have to be done in multiple-precision.

Luther and Guseman [3] developed an iterative method for obtaining the exact adjoint of a matrix. Multiple-precision computation is again required for most of the process.

The method described in this paper consists of converting the given system of equations to a system of congruences modulo a number of primes. At the end of the kth step, the solution of the system of congruences mod p_k is combined with the previous solution by the Chinese Remainder Theorem. The process is continued until the solutions produced in the kth and $(k + 1)$st steps are identical. At this point a substitution check is made, and usually the process terminates there, although the check may reveal that more primes are required.

All computations are performed in single-precision arithmetic, except for the following three computations:

(i) combining the previous solution with the solution mod p_k;

(ii) the substitution check;

(iii) a final reduction step designed to produce a solution in lowest terms.

All these three computations are rather short.

A series of standard FORTRAN subroutines was written to implement the congruence method on the CDC 1604-A Computer (32K memory, 48 bit word). For example, it took 7 (60) minutes computing time to obtain the 6 (9) independent solutions of a system of 54 (111) homogeneous equations in 60 (120) unknowns of rank 54 (111) with integral coefficients in the range $[-1008, 856]$ ($[-2180, 2568]$). These times could have been about halved, had we available primes of the order of 10^{14}, rather than only primes of the order of 10^7. The present program can handle matrices of order up to about 150 × 150. However, the method is by no means limited to this size. By using magnetic tape, and extending the precision of the multiple-precision part of the program, larger matrices can be handled.

In fact, it appears that the gain of the congruence method over conventional

Received April 6, 1965.

* Presented at the IFIP Congress 65 at the Symposium on "Application of Computers to Number Theory and Discrete Problems," May 1965, New York.

multiple-precision methods is larger for large matrices. The computation time for working in t-precision is roughly proportional to t^2. In the congruence method, where most of the work is single-precision, computation time is roughly proportional to t. But each single-precision operation in the congruence method is more time-consuming than in conventional methods, since in the congruence method one works with residues mod p_k which should always lie in a fixed interval $(K, K + p_i]$, and division by p_k is necessary to keep them there.

Another property of the congruence method is that it preserves the maximal nonzero determinant of the original coefficient matrix. Using cross-multiplication elimination, this determinant is usually multiplied by a very large factor, which leads to very large intermediate results, and consequently higher precision is required.[1]

There are many other areas besides number theory in which exact solutions are desirable. For example, in group theoretical investigations in atomic and nuclear spectroscopy, where one often encounters symmetric matrices whose elements and eigenvalues are of the form $a\sqrt{b}/c$, a, b, c integers [1], [4]. It is required to find the eigenvectors, whose elements are of the same form. Sometimes it is much more enlightening to examine the exact form of the elements than to observe numerical results only [1]. By treating the square roots as parameters, this can be done by the congruence method.

The congruence method is not confined to solving linear diophantine equations. It can be used, e.g., for evaluating rational functions with rational coefficients exactly. The need for this arises again in atomic spectroscopy [6], and of course in numerical investigations in number theory. Since all numbers represented in a computer are rational, it appears feasible to use the method in some situations where high-precision is required for intermediate results, although no exact final solution is required. However, in the present paper we confine attention to the exact solution of linear equations only.

2. The Congruence Method. By multiplying through by a suitable integer, we may assume that the rational system of equations is integral. We may further assume without loss of generality that the system is homogeneous. Because consider the nonhomogeneous system $AX = Y$, where A is of order $m \times (n - 1)$, and X, Y are column vectors of orders $n - 1$ and m respectively. Replacing A by the augmented matrix B and $X^T = (x_1, \cdots, x_{n-1})$ by $Z^T = (z_1, \cdots, z_n)$, leads to the homogeneous system $BZ = 0$. Considering this as a linear dependence relation on the columns of B, we see that if this system has no solution with $z_n \neq 0$, the nonhomogeneous system is not solvable. Otherwise, a particular solution of $AX = Y$ is given by $(-z_1/z_n, \cdots, -z_{n-1}/z_n)$, to which can be added any solution of $BZ = 0$ for which $z_n = 0$. Therefore we shall henceforth only deal with the problem of obtaining an integral solution to an integral homogeneous system $AX = 0$, where A is of order $m \times n$ and rank r, with $m \leqq n$. It has a nontrivial solution and hence a nontrivial integral solution if and only if $r < n$.

[1] Dr. Morris Newman informed us that he has a machine program which transforms an integral matrix to Hermite Normal Form. These transformations also preserve the size of all subdeterminants.

Definitions.

(i) A *minor* $M = M(i_1, \cdots, i_t; j_1, \cdots, j_t)$ of order t of a matrix C is the determinant of the submatrix composed of rows $i_1, \cdots, i_t$ and columns $j_1, \cdots, j_t$ of C.

(ii) If $G = (g_1, \cdots, g_n)$, $H = (h_1, \cdots, h_n)$ are two vectors satisfying $g_j \equiv h_j \pmod p$ for all $1 \leqq j \leqq n$, we write $G \equiv H \pmod p$.

(iii) Let $\{F^{(k)}\}$ be a sequence of vectors and G a fixed vector. We say that $\{F^{(k)}\}$ *converges* to G if $F^{(k)} = G$ for all sufficiently large k.

Let $p_1, \cdots, p_k$ be distinct primes, and suppose that $H^{(1)}, \cdots, H^{(k)}$ are vectors satisfying

$$G \equiv H^{(s)} \pmod{p_s}, \qquad s = 1, \cdots, k'$$

where G is a fixed vector. By the Chinese Remainder Theorem, there exists a unique vector $F^{(k)} = (f_1^{(k)}, \cdots, f_n^{(k)})$ satisfying

$$(1) \qquad F^{(k)} \equiv H^{(s)} \pmod{p_s}, \qquad s = 1, \cdots, k$$

and

$$(2) \qquad [-p_1 \cdots p_k/2] < f_j^{(k)} \leqq [p_1 \cdots p_k/2], \qquad j = 1, \cdots, n.$$

For all sufficiently large k we clearly have also

$$[-p_1 \cdots p_k/2] < g_j \leqq [p_1 \cdots p_k/2], \qquad j = 1, \cdots, n.$$

Hence $F^{(k)}$ converges to G. We state this formally in the following form:

THEOREM I. *Suppose that*

$$(3) \qquad G \equiv H^{(s)} \pmod{p_s}, \qquad s = 1, \cdots, k'$$

where G is a fixed vector Let $F^{(k)}$ be the unique vector satisfying (1) *and* (2). *Then $\{F^{(k)}\}$ converges to G.*

We shall apply this result to the case where the $H^{(s)}$ are solutions of $AX \equiv 0 \pmod{p_s}$, and G is an unknown but fixed solution of $AX = 0$. A way to find such vectors $H^{(s)}$ is suggested by the following simple fact:

THEOREM II. *Let G be a solution of $AX = 0$, where $A = (a_{ij})$ is an integral $m \times n$ matrix of rank r. Let H be a solution of $AX \equiv 0 \pmod p$, p prime and $M = M(i_1, \cdots, i_r; j_1, \cdots, j_r)$ a minor of A not divisible by p. If $g_j \equiv h_j \pmod p$ for $j \neq j_1, \cdots, j_r$, then $G \equiv H \pmod p$.*

Proof. G and H and hence $G - H$ are solutions of

$$\sum_{j=1}^{n} a_{ij}x_j \equiv 0 \pmod p, \qquad i = i_1, \cdots, i_r.$$

Since $g_j - h_j \equiv O(\text{mod}\, p)$ for $j \neq j_1, \cdots, j_r$, $g_{j_l} - h_{j_l}$ $(l = 1, \cdots, r)$ is also a solution of

$$\sum_{l=1}^{r} a_{i_k, j_l}x_{j_l} \equiv 0 \pmod p, \qquad k = 1, \cdots, r.$$

But the coefficient matrix of this system is nonsingular. Hence it has only the trivial solution, and $G \equiv H(\text{mod}\, p)$ as asserted.

The system of equations considered in Theorem II has an integral solution space of $n - r$ dimensions. A particular solution, to be denoted by G, is determined by putting $g_{j_{r+1}} = M$, $g_{j_{r+i}} = 0$ for $1 < i \leqq n - r$. In order to find solutions $H^{(s)}$ of $AX \equiv 0 \pmod{p_s}$ all of which satisfy (3) for the same G, it suffices therefore to do the following:

(I) put $h_{j_{r+1}} \equiv M \pmod{p_s}$;

(II) put $h_{j_{r+i}} \equiv 0 \pmod{p_s}$, $1 < i \leqq n - r$, $p_s \nmid M$.

The system $AX \equiv 0 \pmod{p_s}$ is solved by a triangularizing process T, which transforms A into a triangular matrix B_{p_s}. If the rank of $A \pmod{p_s}$ is ρ_s, the minor consisting of the first ρ_s rows and columns of B_{p_s} will be denoted by

$$M_{p_s} = M_{p_s}(i_1^{(s)}, \cdots, i_{\rho_s}^{(s)}; j_1^{(s)}, \cdots, j_{\rho_s}^{(s)}),$$

where $i_1^{(s)}, \cdots, i_{\rho_s}^{(s)}; j_1^{(s)}, \cdots, j_{\rho_s}^{(s)}$ are the ρ_s rows and columns of A appearing as the first ρ_s rows and columns of B_{p_s}. (Here and in the following we use the phrase "rows and columns of A" in a slightly extended sense: a row (column) of A to which is added any linear combination of other rows (columns), is still considered the same row (column) of A.) Since B_{p_s} contains diagonal elements $\not\equiv 0 \pmod{p_s}$ in its first ρ_s rows and columns and since p_s is prime, $M_{p_s} \not\equiv 0 \pmod{p_s}$.

The triangularizing process $T = T_{p_s}$ consists of performing elementary row and column operations on $A \pmod{p_s}$, which leave all minors of A invariant $\pmod{p_s}$. In searching for an element $\not\equiv 0 \pmod{p_s}$, rows are first scanned sequentially, and only then columns. More precisely, suppose that by performing elementary row and column operations we already obtained a matrix $C = (c_{ij})$ such that $c_{ii} \not\equiv 0 \pmod{p_s}$ for $1 \leqq i < k$, and $c_{ij} \equiv 0 \pmod{p_k}$ for $j < i \leqq m$, $1 \leqq j < k$. Then the element in the kth row and kth column of C is the unique element $c_{\sigma\tau} \not\equiv 0 \pmod{p_s}$ of C, such that $c_{ij} \equiv 0 \pmod{p_s}$ for $k \leqq j < \tau$, $k \leqq i \leqq m$ and $c_{i\tau} \equiv 0 \pmod{p_s}$ for $k \leqq i < \sigma$.

As usual, all elements $\not\equiv 0 \pmod{p_s}$ of B_{p_s} and of all partially triangularized matrices leading to B_{p_s} are congruent to the product of a minor of A by the inverse of another $\pmod{p_s}$.

Since neither r nor any minor M of A of order r is known à priori, the remaining problem is to choose primes p_s so that:

(i) the same rank $\rho_s \pmod{p_s}$ of A is obtained for all s;

(ii) B_{p_s} shall contain the same rows and columns of A for all s;

(iii) ρ_s shall be identical with the rank r of A.

Suppose that (i) and (ii) were already solved for all primes $p_1, \cdots, p_{k-1}$. Denote by ρ the common rank of the B_{p_i}, and let

$$V = (i_1, \cdots, i_\rho; j_1, \cdots, j_\rho) = (v_1, \cdots, v_{2\rho}),$$

where $i_1, \cdots, i_\rho; j_1, \cdots, j_\rho$ are the rows and columns of A appearing as the first ρ rows and columns of B_{p_i}, $i = 1, \cdots, k - 1$. Let

$$V^{(l)} = (i_1^{(l)}, \cdots, i_\rho^{(l)}; j_1^{(l)}, \cdots, j_\rho^{(l)}) = (v_1^{(l)}, \cdots, v_{2\rho}^{(l)})$$

be another 2ρ-tuple, and introduce a lexicographic ordering in the set of all 2ρ-tuples as follows: $V^{(l)} > V$ if and only if $v_i^{(l)} > v_i$ for the smallest integer i for which $v_i^{(l)} \neq v_i$. For selecting a prime p_k, we consider the following three cases:

1. $\rho_k < \rho$. In this case p_k is replaced by another prime. This procedure is repeated until a prime p_k is found for which $\rho_k \geqq \rho$. It is clear that such a prime exists

2. $\rho_k > \rho$. In this case we drop the result obtained by the primes $p_1, \cdots, p_{k-1}$ and put $\rho = \rho_k$, $V = V^{(k)}$.

3. $\rho_k = \rho$. Here we have three subcases:

(i) $V^{(k)} > V$. From the definition of the process T, it is clear that this happens only if p_k divides one of the minors of A, and hence there is only a finite number of such primes. By repeatedly replacing p_k, we will eventually find a prime p_k for which $V^{(k)} \leqq V$.

(ii) $V^{(k)} < V$. In this case we drop $p_1, \cdots, p_{k-1}$ and put $V = V^{(k)}$. This situation can happen only for a finite number of primes, since the set of all V contains a lexicographically minimal element.

(iii) $V^{(k)} = V$. Then a solution $H^{(k)}$ is selected according to (I) and (II) above. This solution is combined with the previous one, whereby a vector $F^{(k)}$ satisfying (1) and (2) is produced.

By construction,

$$\begin{aligned} M_{p_k}(i_1^{(k)}, \cdots, i_{\rho_k}^{(k)}; j_1^{(k)}, \cdots, j_{\rho_k}^{(k)}) &\equiv M_{p_k}(i_1, \cdots, i_\rho; j_1, \cdots, j_\rho) \\ &\equiv M\,(i_1, \cdots, i_\rho; j_1, \cdots, j_\rho) \pmod{p_k}, \end{aligned}$$

where $i_1, \cdots, i_\rho; j_1, \cdots, j_\rho$ are the same for all k. Hence by (I), (II) and Theorem II, congruence (3) is satisfied, and $F^{(k)}$ converges to G by Theorem I.

Using a sufficiently large number of primes guarantees that $\rho = r$ by Hadamard's inequality. In practice we work with large primes, because then the probability that a prime divides a minor of A is very small, and practically every prime leads to subcase (iii) of case 3. Thus convergence is fast and the probability that $\rho < r$ for a number of successive primes is extremely small. Even if this should happen, it would be detected by the substitution check, and more primes would then be selected until $\rho = r$.

3. The Computer Program. Primes of the order of 10^7 which are just less than half a computer word are used for the computer program. They were taken from among the last entries of Lehmer's table [2]. Using primes of the order of a whole computer word would usually lead to less iteration steps, but such large primes were not readily available.

A master routine and seven subroutines were written in FORTRAN 63 for the CDC 1604-A computer. The master program entitled SOLVE reads in the matrix A from magnetic tape. The size of A is limited to about 150 × 150. The master program also contains the substitution check which is written in 8-precision. This means that the absolute value of all minors of A is limited to 2^{383}. Also the last three subroutines below are written in 8-precision. The last two of these carry out the final reduction step. The first four subroutines below are written in single-precision.

1. TRIANGLE. Carries out the triangularizing process and computes the principal minor (mod p). It also registers the order of the rows and columns of A in the triangularized matrix B_p.

2. SOLUTION. Generates the solution space (mod p) of dimension $n - \rho_p$ from B_p.

3. MODULO. Finds the residue of a number in the range $[-p/2] < a \leqq [p/2]$ by means of a division by p. This is carried out after each multiplication.

4. INVERSE. Computes the inverse of a (mod p), p prime. For any $a \neq 0$ satisfying $[-p/2] < a \leqq [p/2]$, we have $(a, p) = 1$. Hence $ka + lp = 1$ and $k \equiv a^{-1} \pmod{p}$. The integer k is computed by means of the Euclidean Algorithm.

5. CHIN. Computes the solution $x = a$ of

$$x \equiv A \pmod{P} \tag{4}$$

$$x \equiv B \pmod{Q} \tag{5}$$

lying in the range $[-PQ/2] < a \leqq [PQ/2]$, where A, P are single-precision numbers, B, Q multiple-precision numbers, P prime, $(P, Q) = 1$. From (4), $a = A + KP$. Hence from (5), $K \equiv (B - A)P^{-1} \pmod{Q}$.

6. GCD. Computes the greatest common divisor of two multiple precision numbers by the Euclidean Algorithm.

7. SIMPLIFY. Divides the n components of a multiple-precision vector by their g.c.d., computed by means of the GCD subroutine.

Department of Applied Mathematics
The Weizmann Institute of Science
Rehovot, Israel

1. B. Kaufman & C. Noack, "Unitary symmetry of oscillators and the Talmi transformation," *J. Mathematical Phys.*, v. 6, 1965, pp. 142–152.

2. D. N. Lehmer, *List of Prime Numbers From* 1 *to* 10,006,721, Hafner, New York, 1956.

3. H. A. Luther & L. F. Guseman, Jr., "A finite sequentially compact process for the adjoints of matrices over arbitrary integral domains," *Comm. ACM*, v. 5, 1962, pp. 447–448. MR **27** #2093.

4. G. Racah, "Use of the Weizac in theoretical spectroscopy," *Bull. Res. Council Israel Sect. F*, v. 8 1959, pp. 1–14.

5. J. B. Rosser, "A method of computing exact inverses of matrices with integer coefficients," *J. Res. Nat. Bureau Standards*, v. 49, 1952, pp. 349–358. MR **14**, 1128.

6. M. Rotenberg et al., *The* 3-*j and* 6-*j Symbols*, The Technology Press, Massachusetts Institute of Technology, Cambridge, Mass., 1959.