Numer. Math. 40, 137 – 141 (1982)

Numerische
Mathematik


# Exact Solution of Linear Equations Using *P*-Adic Expansions*

John D. Dixon

Department of Mathematics and Statistics, Carleton University, Ottawa K1S 5B6, Canada

**Summary.** A method is described for computing the exact rational solution to a regular system $Ax=b$ of linear equations with integer coefficients. The method involves: (i) computing the inverse (mod $p$) of $A$ for some prime $p$; (ii) using successive refinements to compute an integer vector $\bar{x}$ such that $A\bar{x} \equiv b \pmod{p^m}$ for a suitably large integer $m$; and (iii) deducing the rational solution $x$ from the $p$-adic approximation $\bar{x}$. For matrices $A$ and $b$ with entries of bounded size and dimensions $n \times n$ and $n \times 1$, this method can be implemented in time $O(n^3(\log n)^2)$ which is better than methods previously used.

*Subject classifications:* (MR 1980) AMS(MOS); 65F05, 15A06, 10M10, 10A30 CR: 5.14.

## Introduction

In some situations it is desirable to obtain the exact rational solution to a nonsingular system $Ax=b$ of linear equations with integer coefficients; for example, in some number theory problems and in cases where the system is so ill-conditioned that the usual floating point calculations are inadequate. There are presently two approaches used to find such exact solutions: direct computation using multiprecision arithmetic and congruence techniques. To date the only form of the congruence technique which has appeared feasible is: (i) computation of $\det A$ and $(\operatorname{adj} A)b$ modulo $p$ for a number of different moduli $p$ (usually chosen prime); (ii) use of the Chinese Remainder Theorem to combine these results for a composite modulus; and (iii) application of the Hadamard inequality to $\det A$ and the entries of $(\operatorname{adj} A)b$ to deduce the required rational solution from the congruential values. The congruence method has the advantage that for suitable choice of the moduli most calculations are carried out in

* This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (Grant No. A 7171)

0029-599X/82/0040/0137/$01.00

single precision, but on the other hand involves considerable redundancy. Which of the two methods is superior is not entirely clear (see [1] and the literature quoted there); but in either case for an $n \times n$ system in which all coefficients lie in a fixed range $[-\beta, \beta]$ the number of arithmetic operations between numbers the size of $\beta$ grows at least as fast as $n^4$.

In the present paper we describe another method, which is also congruential but based on a single prime modulus $p$ and with time complexity $O(n^3(\log n)^2)$. Although $p$-adic methods have been proposed before (see [4]), they have not appeared practical. Our method is both theoretically attractive and efficient in practice.

## Outline of the Method

Consider the system $Ax=b$ where $A$ is an $n \times n$ nonsingular integer matrix and $b$ is an $n \times 1$ integer matrix. Let $\lambda_i$ denote the (Euclidean) length of the ith column of $A$ and $\lambda_0$ be the length of $b$. Then Hadamard's determinant inequality shows that $|\det A| \leqq \prod_{i=1}^{n} \lambda_i$. If we define $\delta = \prod' \lambda_i$ where the product is over the $n$ largest of $\lambda_0, \ldots, \lambda_n$ then Cramer's rule shows that the rational entries in the solution $x$ to $Ax=b$ have their numerators and denominators all bounded in absolute value by $\delta$. Now let $\beta$ be a bound on the absolute values of the entries of $A$ and $b$. Then certainly $\delta \leqq \beta^n n^{n/2}$. We shall suppose that $p$ is a prime such that $p \leqq \beta$ and $p \nmid \det A$. In what follows we shall measure the time taken by our algorithm in terms of the number of arithmetic operations between integers of size up to $\beta$.

The first step in the solution of $Ax=b$ is to compute an inverse (mod $p$) to $A$. In other words we find an $n \times n$ integer matrix $C$ whose entries lie in the range $[0, p-1]$ and such that $AC \equiv I (\bmod p)$. Using classical methods $C$ can be computed in $O(n^3)$ operations (where an operation is defined to mean an integer arithmetic operation between numbers of magnitude $\beta$).

The second step is to compute a $p$-adic approximation $\bar{x}$ to $x$. This is done as follows. Consider the two sequences of integer column vectors $\{x_i\}$, $\{b_i\}$ (where all entries of $x_i$ lie in $[0, p-1]$) defined by: $b_0 = b$; and $x_i \equiv C b_i (\bmod p)$, $b_{i+1} = p^{-1}(b_i - Ax_i)$ for $i \geqq 0$. Note that $b_i - Ax_i \equiv A(Cb_i - x_i) \equiv 0 (\bmod p)$, so the entries of $b_{i+1}$ are integers. Also, since all entries of $x_i$ lie in $[0, p-1]$, simple induction shows all entries of $b_i$ lie in $[-n\beta, n\beta]$. The entries of $b_i$ are integer of length $O(\log n + \log \beta) = O(\log n)$, so $x_i$, $b_{i+1}$ can be computed from $b_i$ in $O(n^2 \log n)$ operations. We shall terminate the construction of these sequences after computing $x_{m-1}$, $b_m$ where $m$ is an integer to be defined below; this construction takes a total of $O(mn^2 \log n)$ operations. Putting $\bar{x} = \sum_{i=0}^{m-1} x_i p^i$ we have

$$A\bar{x} = \sum_{i=0}^{m-1} p^i A x_i = \sum_{i=0}^{m-1} p^i (b_i - p b_{i+1}) = b_0 - p^m b_m.$$

Hence $A\bar{x} \equiv b (\bmod p^m)$ as required.

The final step is to recover the rational solution $x$ from the $p$-adic approximation $\bar{x}$. In the next section we shall show that there is a choice of $m$ such that $m=O(n\log n)$, and that each component of $x$ can be recovered from the corresponding component of $\bar{x}$ in $O(m^2)$ operations; thus $x$ can be recovered from $\bar{x}$ in $O(nm^2)=O(n^3(\log n)^2)$ operations. Hence the total computation of the exact solution to $Ax=b$ takes $O(n^3(\log n)^2)$ operations.

## Decoding to Rational Form

To perform the final step in the above solution $Ax=b$ we must solve the following problem. Given that there is an unknown fraction $f/g$ with $|f|, |g|\leqq\delta$ and that $gs\equiv f(\bmod h)$ for known integers $s,h$ ($s=\bar{x}$ and $h=p^m$ in the case above), give an efficient algorithm to compute $f/g$. It is clear that $h$ cannot be too small if the solution is to be unique. The following theorem gives appropriate conditions.

**Theorem.** *Let $s,h>1$ be integers and suppose that there exist integers $f,g$ such that*

$$gs\equiv f(\bmod h) \quad \text{and} \quad |f|,|g|\leqq\lambda h^{\frac{1}{2}}$$

*where $\lambda=0.618\ldots$ is a root of $\lambda^2+\lambda-1=0$. Let $w_i/v_i$ $(i=1,2,\ldots)$ be the convergents to the continued fraction for $s/h$ and put $u_i=v_is-w_ih$. If $k$ is the least integer such that $|u_k|<h^{\frac{1}{2}}$, then $f/g=u_k/v_k$.*

*Proof.* It is well known that the sequences $\{w_i\}$, $\{v_j\}$ are increasing while $\{u_i\}$ is alternating in sign and decreasing in absolute value (see, for example, [2] for properties of continued fractions).

Now put $f=gs-th$. Then

$$\left|\frac{s}{h}-\frac{t}{g}\right|=\left|\frac{fg}{hg^2}\right|<\frac{1}{2g^2}$$

and so $t/g$ equals one of the convergents, say $w_j/v_j$, of $s/h$ ([2], Theorem 19). Since $w_j$ and $v_j$ are relatively prime, $|u_j|\leqq|f|\leqq\lambda h^{\frac{1}{2}}$, and so the definition of $k$ shows that $j\geqq k$. On the other hand, $u_j=v_js-w_jh$ and $u_k=v_ks-w_kh$, and so $u_jv_k-u_kv_j\equiv 0(\bmod h)$. Since $j\geqq k$, $|u_jv_k-u_kv_j|\leqq(|u_j|+|u_k|)|v_j|<(\lambda+1)\lambda h=h$; hence $u_jv_k=u_kv_j$ and so $j=k$. Thus $f/g=u_k/v_k$ as asserted and the theorem is proved.

In practice $f/g$ is best calculated by using a modification of the Euclidean algorithm. We may assume $0\leqq s\leqq h$, and make a slight change in notation by replacing $u_i$ by $|u_i|$ (recall that the original $\{u_i\}$ is alternating). We generate the integer sequences $\{u_i\}$, $\{v_i\}$ and $\{q_i\}$ defined by:

$$u_{-1}=h, \qquad u_0=s, \qquad v_{-1}=0, \qquad v_0=1$$

and $q_i=[u_{i-1}/u_i]$, $u_{i+1}=u_{i-1}-q_iu_i$, $v_{i+1}=v_{i-1}+q_iv_i$ for $i=0,1,\ldots$, until the first index $k$ where $u_k<h^{\frac{1}{2}}$. Then (under the hypotheses of the theorem) we have $f/g=(-1)^ku_k/v_k$. (The replacement of $u_i$ by $|u_i|$ means that all sequences consist of positive integers and this simplifies the necessary multiprecision arithmetic.) To estimate the number of operators (involving numbers of size $\beta$) required to

compute $f/g$ we proceed as follows. Simple induction shows that $v_i \geqq \prod_{j=0}^{i-1} q_j$ and $v_i \geqq 2^{i/2}$ for $i \geqq 0$. Since $(-1)^k u_k/v_k$ is equal to $f/g$ in its lowest terms, $|v_k| \leqq |g| < h^{\frac{1}{2}}$. Thus $\sum_{j=0}^{k-1} \log q_j$ and $k$ are both $O(\log h)$. The integers $u_i$, $v_i$ are all bounded by $h$ and so the multiprecision computation of $f/g$ will take at most $\Sigma O(\log q_j + 1)\; O(\log h) = O(\log h)^2$ operations. (In fact, almost all steps will involve simple precision values of $q_i$; see [3], p. 305.)

Returning to the solution $Ax=b$ we have $h=p^m$ and can ensure that the hypotheses of the theorem hold for the components of $x$ provided $\delta \leqq \lambda p^{m/2}$. Thus take $m = 2[\log(\delta\lambda^{-1})/\log p]$. As we saw above $\delta \leqq \beta^n n^{n/2}$, and so, assuming $p$ has been fixed, we have $m = O(n \log n)$ as required.

*Remark.* Examining the above analysis of the time complexity with a little more care it will be seen that the $\log n$ factors all enter in the form $(\log n + \log \beta)$. In practice, for the values of $n$ and $\beta$ which are likely to occur, $\log n$ will be smaller than $\log \beta$ and this factor will be roughly constant. Thus, in practical cases, the time taken to solve $Ax=b$ may be expected to be proportional to $n^3$ rather than $n^3(\log n)^2$.

## Practical Considerations

About half of the calculations are carried out modulo $p$, so to simplify multiplication it is convenient to choose $p$ so that $p^2$ is no larger than the maximum integer which the computer used can handle. In rare cases, $\det A \neq 0$ but is divisible by $p$; this will be recognized in the computation of $C$ and then a new choice of $p$ must be made. There is a modest saving possible by using the $LU$-factorization of $A$ rather than computing $C$, but the major time is spent on the last two steps. Computing $\bar{x}$ is straight forward, and the fact that its components are expressed as integers to the base $p$ should be used in the computations of the final step. It is worthwhile to code the last step carefully using, for example, Lehmer's trick (see [3], p. 306) to speed up the calculation.

We noted in an earlier remark that in practice the time taken by this method to solve $Ax=b$ exactly is roughly proportional to $n^3$ and so comparable to the time taken by a singleprecision floating-point solution by direct methods. Some very limited experiments (with $\beta = p$ and $p^2$ approximately the maximum size of an integer for the computer used) suggest that the exact solution takes approximately 10–20 times longer. This compares very favourably with the figures given in [1] for the methods considered there. It should be noted however that the alternative methods mentioned in the introduction have two properties which our method does not: (i) they permit simple computation of $\det A$; and (ii) they can be used to compute $A^{-1}$ with only a little more work (while our method seems to require the equivalent of solving $Ax=b$ $n$ times). However, for the problem of finding the exact solution of a single system $Ax=b$, the method described here should be superior to the other methods for all but the smaller values of $n$.

## References

1. Cabay, S., Lam, T.P.L.: Congruence techniques for the exact solution of integer systems of linear equations. ACM Trans. Math. Software **3**, 386–397 (1977)
2. Khinchin, A.Ya.: Continued Fractions, 3rd ed. Chicago: Univ. Chicago Press, 1961
3. Knuth, D.: The Art of Computer Programming, Volume 2. Reading, MA: Addison-Wesley, 1969
4. Krishnamurthy, E.V., Rao, T.M., Subramanian, K.: *P*-adic arithmetic procedures for exact matrix computations. Proc. Indian Acad. Sci. **82**A, 165–175 (1975)

Received January 29, 1982/August 6, 1982