Contribution 1/2

Iterative Refinement of the Solution of a Positive Definite System of Equations*

by R. S. Martin, G. Peters, and J. H. Wilkinson

1. Theoretical Background

In an earlier paper in this series [1] the solution of a system of equations $Ax = b$ with a positive definite matrix of coefficients was described; this was based on the Cholesky factorization of $A$. If $A$ is ill-conditioned the computed solution may not be sufficiently accurate, but (provided $A$ is not almost singular to working accuracy) it may be improved by an iterative procedure in which the Cholesky decomposition is used repeatedly.

This procedure may be defined by the following relations

$$
A = LL^T, \quad x^{(0)} = 0,
$$

$$
r^{(s)} = b - A x^{(s)}, \quad (LL^T) d^{(s)} = r^{(s)}, \quad x^{(s+1)} = x^{(s)} + d^{(s)}.
$$

In each iteration the residual $r^{(s)}$ corresponding to the current $x^{(s)}$ is computed and the correction $d^{(s)}$ to be added to $x^{(s)}$ is determined using the computed $LL^T$ factorization.

The only reason that the first solution is not exact is because of the intervention of rounding errors. Since rounding errors are also made in the iterative refinement it is by no means obvious that the successive $x^{(s)}$ will be improved solutions. It has been shown by Wilkinson [2] that if $A$ is not “too ill-conditioned” the $x^{(s)}$ will converge to the correct solution “to working accuracy” provided $r^{(s)}$ is computed to double-precision. Since the elements of $A$ and of $x^{(s)}$ are single-precision numbers this may be achieved merely by accumulating inner-products. The remaining steps may be performed in single-precision.

Two terms used above call for some explanation. We say that $x^{(0)}$ is the correct solution “to working accuracy” if $\|x^{(0)} - x^{\infty}\|_\infty \leq 2^{-t}$ where the mantissae of our floating-point numbers have $t$ binary digits. For the meaning of the term “too ill-conditioned” we must refer to the error analysis of the Cholesky method. It has been shown ([2, 3]) that the computed solution of $Ax = b$ is the exact solution of a system $(A + \delta A)x = b$ where $\delta A$ is dependent on $b$ but is uniformly bounded. The upper bound $m$ of $\|\delta A\|_2$ is dependent on the details of the arithmetic used, particularly upon whether or not inner-products are accumulated without intermediate rounding. We shall say that $A$ is “too ill-conditioned” for the precision of computation that is being used if $m\|A^{-1}\|_2 \geq 1$. In this case $A + \delta A$ could be singular and in any case $(A + \delta A)^{-1}b$ may not

agree with $A^{-1}b$ in any of its figures. If $m\|A^{-1}\|_2 = 2^{-p}$ ($p > 0$) then successive iterates certainly satisfy the relation

$$\|x - x^{(n+1)}\|_2 \leq \|x - x^{(n)}\|_2 2^{-p}/(1 - 2^{-p}).$$

(2)

Usually the statistical distribution of rounding errors will ensure that the $x^{(n)}$ will improve more rapidly than this.

A similar process may be used to refine an inverse derived from the Cholesky factorization of $A$ as described in [1]. If we denote the first computed inverse by $X^{(n)}$ then we have the iterative refinement procedure defined by

$$X^{(n+1)} = X^{(n)} + X^{(n)}(I - AX^{(n)}),$$

$$= X^{(n)} + X^{(n)}B^{(n)},$$

(3)

$$= X^{(n)} + Z^{(n)}.$$

In order to be effective it is essential that the right-hand side of (3) should not be expressed in the form $2X^{(n)} - X^{(n)}AX^{(n)}$. The "residual matrix" $I - AX^{(n)}$ must be computed using double-precision or accumulation of inner-products, each component being rounded only on completion. The conditions for convergence are the same as those for the iterative refinement of the solution of $Ax = b$, but if the condition is satisfied the convergence is effectively quadratic. Indeed ignoring rounding errors we have

$$I - AX^{(n+1)} = (I - AX^{(n)})^2.$$

(4)

In [1] several different procedures were given based on symmetric decompositions of $A$; here we give only the procedures related to choldet1, cholsol1 and cholinv1. Analogous procedures could be made based on symdet, symsol and syminv1. Since $A$ must be retained in order to form the residuals the decompositions in which the upper half of $A$ is stored as a linear array of $\frac{1}{2}n(n+1)$ elements are not relevant.

2. Applicability

These procedures may be used to compute accurate solutions of systems of equations with positive definite matrices or to compute accurate inverses of such matrices. An essential feature of the algorithms is that in each iteration an accurate residual must be determined corresponding to the current solution. The $i$-th components $r_i$ of a residual is given by

$$r_i = b_i - \sum_{j=1}^{n} a_{ij}x_j,$$

(5)

and since only the upper-half of $A$ is stored this must be written in the form

$$r_i = b_i - \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i}^{n} a_{ij}x_j,$$

(6)

taking advantage of the symmetry of $A$. Similar remarks apply to the computation of $I - AX$ in the procedure for an accurate inverse.