Journal of the Nigerian Association of Mathematical Physics, Volume 9 (November 2005)

On the compatibility of approximate solution of linear equations with given error bounds for coefficients and right sides.

L. A. Fakande, and **I. F. Arunaye**
Department of Mathematics and Computer Science
Delta State University, Abraka, Nigeria

Abstract

In the consideration of solutions to systems of linear equations with inaccuracies in their entries, Oettli and Prager (1964) [1] ***considered extraction of approximate solution to modified systems with various residues, and established the error tolerance for coefficients and right – hand sides. This paper applied the condition number and relative error principles of*** Kreyszig (1999) [2] and Kopchenova and Maron (1984) [3] ***to establish conditions for exact solutions to modified systems of linear equations*** (1.4) ***discussed by*** Oettli and Prager (1964) [1].

pp 265- 268

1.0 Introduction

Consider a system of n linear equations

$$Ax = b \tag{1.1}$$

Efforts for obtaining the solution vector x which satisfy the system have led to many considerations. Besides, there is the fact that the computed solution vector x^0 may not be exact solution *x*, due to one form of error or the other, (Forsythe 1953 [4], Golub and Loan 1996 [5], Fox, Huskey and Wilkinson 1948 [6], Wilkinson 1961[7], 1963 [8]). Possible considerations of residuals and inaccuracies in measurement are $r = Ax^0 - b$, where $r = 0$ is the trivial case; and where $r \neq 0$, $x^0 = x + \delta x$, $\|\delta x\| > 0$

$$A(x + \delta x) = b + \delta \text{ b} \tag{1.2}$$

$$(A + \delta A)(x + \delta x) = b \tag{1.3}$$

$$(A + \delta A)(x + \delta x) = b + \delta b \tag{1.4}$$

Where $\delta A, \delta b$ are errors in measurement in A_{jk} of A and b_j of b respectively, and $\|\delta A\| \leq |\Delta A|$; $\|\delta b\| \leq |\Delta b|$ are given tolerance respectively. For the modified systems, the conditions for x^0 to be the exact solution is a subject of interest in this paper. See Oettli, and Prager (1964) [1], Rice (1966). [9], Kreyszig (1999) [2], established the condition number principles, which indicate that systems which satisfy these conditions strongly indicate solution for the system, and inaccuracies in entries of both coefficient or right–hand sides allow solution to be determined within desired error tolerance. Kreyszig (1999) [2], Rigal and Gaches (1967) [10], Todd (1949) [11], considered the condition number of the coefficient matrix ($K(A) \geq 1$), and established values $K(A) \leq 20$ and $K(A) \geq 20$ to be respectively small and large value in measuring the condition number. It also informed that for such systems, it is possible to obtain modifications to system (1.1) so that its approximate solution is the exact solution of the modified system. Oethli and Prager (1964) [1] considered system (1.4) and established tolerance regions for coefficient and right–hand sides where a given vector x^0 may be regarded as the exact solution of the modified systems. We hereby applied this principle to the modified systems (1.3) and (1.4) in the aforementioned. We also justify Oettli and Prager (1964) [1] results by utilizing the condition number principle; and illustrates this for system (1.3) where b is precisely known but the coefficient matrix may have any value in the range $A_{jk} \pm \delta A_{jk}$; $j, \ k = 1,2,\ldots, n$; where $|\delta A_{jk}| < \|\delta A\| \leq |\Delta A|$.

2.0 Application of the condition number principle to modified systems.

Theorem 2.1

A linear system of equations (1.1) ***whose condition number***

$$K(A) = \left\|A^{-1}\right\|\left\|A\right\|$$

is small is well conditioned. Kreyszig, (1999) [2].

Proof

(i) The proof for system (1.2) can be found in Kreyszig (1999) [2].

(ii) We shall state the proof for the systems (1.3) and (1.4) as following.

Considering system (1.3) we have $(A+\delta A)(x+\delta x)=b$, $\delta x=-A^{-1}\delta A(x+\delta x)$

Hence $\|\delta x\|=\left\|A^{-1}\delta A(x+\delta x)\right\| \le \left\|A^{-1}\right\|\left\|\delta A(x+\delta x)\right\| \le \left\|A^{-1}\right\|\|\delta A\|\|x+\delta x\|$

$\frac{\|\delta x\|}{\|x+\delta x\|} \le \left\|A^{-1}\right\|\|\delta A\|$ Since k (A) $=\left\|A^{-1}\right\|\left\|A\right\|$, then $\frac{\|\delta x\|}{\|x+\delta x\|} \le k(A)\frac{\|\delta A\|}{\|A\|}$. Hence $\frac{\|\delta x\|}{\|x\|}$

$\approx \frac{\|\delta x\|}{\|x+\delta x\|} \le k\ (A)\ \frac{\|\delta A\|}{\|A\|}$. Since A and x are not known we note that Kopchenova and Maron (1984) [3], numerically characterized the accuracy of approximate values by their relative errors. They established relative error of the order of 1% as corresponding with the presence of 2 correct digits, and relative error of 0.01% corresponds with 4 correct digits in the results. When relative error of an approximate value is of order less than 5%, the effect is only in the second correct digit of the absolute error which is insignificant. Kophenova and Maron (1984) [3]. Therefore a small K(A) and $\frac{\|\delta A\|}{\|A\|}$ implies a small $\frac{\|\delta x\|}{\|x\|}$ and the system (1.3) is well conditioned. See Kophenova and Maron (1984) [3], Kreyszig (1999) [2], Todd (1949) [11], Rice (1966) [9], Gautschi (1978) [12]. We now consider system (1.4) $(A+\delta A)(x+\delta x)=(b+\delta b)\Rightarrow \delta x-A^{-1}\delta b=-A^{-1}\delta A(x+\delta x)$. Therefore

$\left\|\delta x-A^{-1}\delta b\right\|=\left\|A^{-1}\delta A(x+\delta x)\right\| \le \left\|A^{-1}\right\|\|\delta A(x+\delta x)\| \le \left\|A^{-1}\right\|\|\delta A\|\ \|(x+\delta x)\|$. Hence $\frac{\left\|\delta x-A^{-1}\delta b\right\|}{\|x+\delta x\|}$

$\le \left\|A^{-1}\right\|\|\delta A\|$ and $\frac{\|\delta x\|}{\|x+\delta x\|}-\frac{\left\|A^{-1}\right\|\|\delta b\|}{\|x+\delta x\|} \le \frac{\|\delta x\|}{\|x+\delta x\|}-\frac{\left\|A^{-1}\delta b\right\|}{\|x+\delta x\|} \le \frac{\left\|\delta x-A^{-1}\delta b\right\|}{\|x+\delta x\|} \le \left\|A^{-1}\right\|\|\delta A\|$.

Therefore

$$\frac{\|\delta x\|}{\|x+\delta x\|} \le \left\|A^{-1}\right\|\|\delta A\|+\left\|A^{-1}\right\|\|\delta b\| \Rightarrow \frac{\|\delta x\|}{\|x+\delta x\|} \le \left\|A^{-1}\right\|\left(\|\delta A\|+\|\delta b\|\right) \Rightarrow \frac{\|\delta x\|}{\|x+\delta x\|} \le k(A)\frac{\left(\|\delta A\|+\|\delta b\|\right)}{\|A\|}$$

$$\Rightarrow \frac{\|\delta x\|}{\|x\|} \approx \frac{\|\delta x\|}{\|x+\delta x\|} \le k(A)\frac{\left(\|\delta A\|+\|\delta b\|\right)}{\|A\|}.$$

For small k (A) and $\frac{\|\delta A\|+\|\delta b\|}{\|A\|}$

The system (1.4) is well conditioned. Thus establishing the fact that for modified systems (1.2), (1.3) and (1.4) satisfying Theorem 2.1, an approximate solution of (1.1) can be made the exact solution.

3.0 Construction of a modified system (1.3) with the approximate solution of system (1.1) as its exact solution.

Consider the system (1.1) where x is the exact solution, let $x^0 = x + \delta x$ be an approximate solution of (1.1).

Then

$$r = Ax^0 - b; \quad r \neq 0 \tag{3.1}$$

Using (1.3)

$$(A + \delta A)x^0 = b \tag{3.2}$$

$$Ax^0 + \delta A x^0 = b$$

$$\Rightarrow \delta A x^0 = -r \tag{3.3}$$

Considering the *jth* row of (3.3)

$$\sum_{k=1}^{n} \delta A_{jk} x_k^0 = -r_j$$

where $\delta A_{jk} \cdots$ is error in each element of the *jth* row of A.

$$\sum_{k=1}^{n} \delta A_{jk} \frac{\left(-x_k^0\right)}{r_j} = 1 \tag{3.4}$$

Setting

$$\frac{\left(-x_k^0\right)}{r_j} = V_j$$

and δA_{jk} as Cartesian coordinates in R^n. The left hand side of (3.4) represents the volume of parallelopiped in $(n + 1)$–dimensional space, i.e. $V.\left\|A_{j1}x, \cdots, A_{j2}x \cdots, xA_{jn}\right\|\underline{n}$ with lengths of sides $2|\Delta A|$ and supporting hyperplanes

$$\sum_{k=1}^{n} \Delta A_{jk} \left| \frac{-x_k^0}{r_j} \right| = \alpha_j \tag{3.5}$$

$\alpha_j \geq 1$, ΔA_{jk}, is absolute error in each δA_{jk}, $k = 1, 2, \ldots, n$, and center on the origin. Oettli and Prager (1964) [1]. From (3.4), (3.5), and since the absolute error of an algebraic sum of several approximate numbers is equal to the sum of the absolute errors of the numbers, Kopchenova and Maron (1984) [3]. We have

$$\sum_{k=1}^{n} \delta A_{jk} \left(\frac{-x_k^0}{r_j} \right) = \frac{1}{\alpha_j} \sum_{k=1}^{n} \Delta A_{jk} \left| \frac{-x_k^0}{r_j} \right|$$

which implies that

$$\delta A_{jk} = \frac{A_{jk}}{\alpha_j} \operatorname{sgn}\left(\frac{-x_k^0}{r_j} \right) \tag{3.6}$$

We observe that when $\alpha_j \geq 1$, (3.6) are modifications needed in the *j*th row of A for $x^0 = x + \delta x$ to be the exact solution of (3.2), k = 1, 2, ..., n.

3.1 **Remarks**

(i) When α_j = 1, there is only one solution to the modified system since (3.6) yields δA_{jk} = ΔA_{jk} sgn$\left(\frac{-x_k^0}{r_j}\right)$ for k = 1, 2, …, n.

(ii) When $\alpha_j > 1$ there are many solutions since for each value of $\alpha_j > 1$ there corresponds one solution from (3.6).

(iii) When $\alpha_j < 0$, or $0 \leq \alpha_j < 1$ the modified system admits no solution since from (3.5) α_j is non-negative and from (3.4), (3.5).

(i), (ii) and (iii) agree with Oettli and Prager (1964) [1].

4.0 Conclusion

It is the requirement that approximate solutions converge to exact solution, or system with small absolute errors are acceptable from modified systems (1.2), (1.3), (1.4). These systems have feasible solutions, only when $\alpha_j \geq 1$, which would precisely tend to the exact solution and equation (3.6) holds. Hence by obtaining the optimum value of $\alpha_j \geq 1$, we obtained the exact solution to the system (1.1).

References

[1] Oettli, W. and Prager W, (1964), Compatibility of Approximate Solutions of Linear Equations with Given Error Bounds for Coefficients and Right – Hand Sides, Numerische Mathematics 6, 405 – 409.

[2] Kreyszig, E. (1999), Advanced Engineering Mathematics (8th Edition), John Wiley and Sons Inc.

[3] Kopchenova, N.V. and Maron, I.A. (1984), Computational Mathematics, Mir Publishers

[4] Forsythe G.E. (1953) Solving Linear Algebraic Equations can be Interesting Bull. Amer Maths Soc. 59; 299 – 329.

[5] Golub, G.H. and Yam Loan C.F., (1996), Matrix Computations (3rd Edition) John Hopkins University Press.

[6] Fox L, Huskey D.H. & Wilkinson J.H. (1948) Notes on the Solution of algebraic linear Simultaneous Equations. The Quarterly Journal of Mechanics & Applied Maths Vol. 1; 149 – 173.

[7] Wilkinson J.H. (1961). Error analysis of direct methods of matrix inversion. *Journal of the computing machinery*, 8 (3): 281 – 330.

[8] Wilkinson J.H. (1963). Rounding errors in algebraic processes. Prentice hall, Englewood Jersey.

[9] Rice J.R. (1966). A theory of condition. *SIAM Journal on Numerical Analysis*, 3(2): 287 – 310.

[10] Rigal J.L. and Gaches J. (1967). On on the Compatibility of a given solution with the data system. *Journal of the computing machinery* 14(3): 543-548.

[11] Todd T. (1949). The condition of certain matrices, I. *The Quarterly Journal of Mechanics and applied Mathematics 2(4):* 469 – 472.

[12] Gautschi W. (1978). Questions of numerical condition related to polynomials. In C.De B Golub, editors, *recent advances in numerical analysis*, pages 45 – 72, New York University of Wisconsin Mathematics Research Centre, Academic Press. Proceedings of a Symposium May 22 – 24.