

DOI: <http://dx.doi.org/10.21123/bsj.2016.13.1.0190>

A Note on the Perturbation of arithmetic expressions

Shawki A.M. Abbas

Department of Computer technologies Engineering of AL-Nsour University College,
Baghdad, Iraq

Received 21, September, 2014

Accepted 16, March, 2015



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](http://creativecommons.org/licenses/by-nc-nd/4.0/)

Abstract:

In this paper we present the theoretical foundation of forward error analysis of numerical algorithms under;

- Approximations in "built-in" functions.
- Rounding errors in arithmetic floating-point operations.
- Perturbations of data.

The error analysis is based on linearization method. The fundamental tools of the forward error analysis are system of linear absolute and relative a priori and a posteriori error equations and associated condition numbers constituting optimal of possible cumulative round – off errors. The condition numbers enable simple general, quantitative bounds definitions of numerical stability. The theoretical results have been applied a Gaussian elimination, and have proved to be very effective means of both a priori and a posteriori error analysis.

Key words: built in function, Rounding errors, perturbations data, boundary value problem.

Introduction:

Evaluation algorithms are defined by finite sequences $F = (F_0, \dots, F_n)$ of input operations, evaluations of 'built-in' functions, and arithmetic operation for determining sequences.

Let $U = (U_0, \dots, U_n)$ of input data, Intermediate and final results [1- 4].

$$U_t = F_t(u), t = 0, \dots, n, \dots \quad (1)$$

Under perturbations, an evaluation algorithm yields approximations V_t of U_t that can be written in the form

$$V_t = (1 + e_t)F_t(V), t = 0, \dots, n, \dots \quad (2)$$

e_i Called the local error are the relative errors of data input.

We shall assume that the local errors are bounded by $|e_t| \leq \gamma_t \eta$ for $t=0, \dots, n$, where γ_t are suitable non

negative weights and η is an accuracy, constant.

Remark

Many of these theoretical properties do not hold in the presence of rounding errors. Where:

Cond (A) = $\|A\| \|A^{-1}\|$ condition number. [4],

Cond $\|A\|$ = denote the spectral norm of A (The norm of a matrix is in some sense a measure of the magnitude of the matrix).

$\|A\| = \text{MAX}_i |\lambda_i(A A^T)|^{1/2}$ The spectral norm. (The notation $\lambda_i(A A^T)$ denotes an eigenvalue of AA^T note that for any real matrix A the matrix AA^T is symmetric and nonnegative definite.

1. Error Propagation:

Starting points are representations and elements of the errors in elementary arithmetic operations, +, -, *, /, and "built-in" function occurring in the floating – point arithmetic of computers. Condition numbers are defined for the simplest algorithms, consisting of the input of one or two operands followed by an arithmetic operation or function.

Estimates for the remainder terms of Taylor’s formula are established. By neglecting remainder terms in the general error equations, Propagation of error is the effect of errors on the uncertainty of a function based on them. When the variables are the values of experimental measurements they have uncertainties due to measurement limitations which propagate to the combination of variables in the function.

The algorithm (A) determines uniquely mapping A in R^{n+1} . The solution of the linear absolute error equations are obtained by means of the solution operators $L = (A'W)^{-1}$, and of the relative error equation by $L = Jw^{-1} (A'W)^{-1} Jw$. this is in correspondence with the use of so-called relative or logarithmic derivatives [5].

2. Error estimates for numerical algorithms:

Round – off errors arise because it is impossible to represent all real numbers exactly on a machine with finite memory (which is what all practical digital computers are)

Given two numbers a, b, and arbitrary approximations a',b', of a,b the following absolute and relative a priori and a posteriori errors are defined:

Absolute	Relative
A priori $\Delta a = a' - a$	$P_a = (a' - a) / a, a \neq 0$
A posteriori $\Delta a' = a - a'$	$pa' = (a - a') / a', a' \neq 0$
$\Delta (aob) = a' ob' - aob,$	$P(aob) = (a'ob' - aob),$

O = +, -, *, / of sums, differences, products, and quotients under perturbations of the operands a, b. It is well known that:

$$\Delta (a \pm b) = \Delta a \pm \Delta b,$$

$$\Delta (ab) = (b \Delta a) + (a \Delta b) + (\Delta a \Delta b)$$

$$\Delta (a/b) = 1 / (b + \Delta b) (\Delta a - (a/b) \Delta b)$$

$$P (a \pm b) = (a/c) P_a \pm (b/c) P_b,$$

$$P (ab) = P_a + P_b + (P_a P_b)$$

$$P (a/b) = (P_a - P_b) / (1 + P_b)$$

Where $c = a \pm b$. The numerical computation of aob first requires input or computing of operands a, b carried out approximately.

3. Forward Elimination:

The following investigation deals with the error analysis of Gaussian elimination for the solution of regular linear systems.

1- $Ax = y : \sum_{k=1}^n a_{ik} x_k = y_i \quad i= 1$
 (1)n. with real coefficients a_{ik} and right-hand sides y_i first the common forward elimination is analyzed which reduces the given system with one or several right –hand sides to a triangular linear system it is well known that the determinant of A can also computed in this way.

Thus let $\hat{A}_1 = (a_{ik}^1)$ be a rectangular $n \times (n+h)$ matrix such that

2- $a_{ik}^1 = a_{ik} \quad i= 1 (1) n.$

For the solution of the above linear system (1) set $h=1$ and

3- $a_{i,n+1}^1 = y_i \quad i = 1(1) n.$

For the error analysis of computing the determinant put, $h = 0$ such that $\hat{A}_1 = = A$ by forward elimination a sequence of matrices

$\hat{A}_{i+1} = (a_{ik}^{t+1})$ is

determined according to

4- $m_{it} = (a_{it}^t / t_{a_u}), a_{ik}^{t+1} = t_{a_{ik}} - m_{it} a_{ik}^t$

($i=t+1, \dots, n,$ & $k=t+1, \dots, n+h$) for $t=1, \dots, n-1$. Due to data perturbations, rounding errors, or a preceding computation of the coefficients and right hand sides this algorithm is

performed numerically with approximations a_{ik}^1 of a_{ik}^1 . The floating-point arithmetic of computer computes, instead of the exact a_{ik}^{t+1} , the approximations [6].

$$5- m_{it} = F1(a_{it}^{-t}/a_u^t), a_{ik}^{-t+1} = F1(a_{ik}^{-t} - F1(m_{it} a_{ik}^{-t})).$$

The linear absolute a priori error approximations s_{ik}^t, s_{it}^m of a_{ik}^{-t}, m_{it} satisfy the linear error equations:

$$s_{it}^m = (1/a_u^t) s_{it}^t - (m_{it}/a_u^t) s_{it}^t + m_{it} e_{li},$$

$$6- s_{ik}^{t+1} = s_{ik}^t - m_{it} s_{tk}^t - a_{tk}^t s_{it}^m - m_{it} a_{tk}^t e_{tik}^x + a_{tk}^{t+1} e_{tik}$$

By $e_u, e_{tik}^*, e_{tik}^-$ are meant the relative rounding errors of the floating-point division/ multiplication/ and subtraction in the computation of a_{ik}^{-t+1} according to (5).

By these equations the linear error approximation s_{ik}^t are uniquely determined as functions of the absolute data errors.

$$s_{ik}^1 = F_{ik}^1 = \Delta a_{ik}^1 \quad (I = 1 \dots n, k = 1, \dots, n+1)$$

and the relative rounding errors $e_{tik}^/, e_{tik}^*, e_{tik}^-$

4. Back substitution:

Forward elimination reduces the given linear system 1. To the triangular system [7].

$$1- Ux = z: \sum_{k=1}^n U_{ki} X_k = Zi \quad (i = 1 \dots n)$$

With the coefficients

$$2. U_{ki} = a_{ik}^1 (k \geq i), Zi = a_{i,n+1}^1 \quad (i, k = 1 \dots n).$$

simultaneously, the lower triangular matrix $L = (L_{ik})$.

3. $L_{ik} = m_{ik} (k < i), L_{ik} = s_{ik} (i \leq k)$, and thus the triangular factorization $A=Lu$ of A is obtained. The solutions X_i of the linear system are then determined successively by:

$$4. X_i = (-\sum_{k=i+1}^{n+1} U_k)/U_{ii} \quad (i=n, \dots, 1).$$

For each index I choose a suitable permutation $J_{i+1}, \dots, J_n + 1$ of the natural numbers $i + 1, \dots, n + 1$, In

computing the solution x_i in the floating-point arithmetic of a computer, becomes

$$5. Z_i^{-n+1} = -F1(U_{i,j_{n+1}} x_{j_{n+1}}),$$

$$6. Z_i^{-k} = F1(Z_i^{-k+1} - F1(U_{ijk} X_{jk})) \quad (k=n, \dots, i+1)$$

$$7. X_i = F1(Z_i^{i+1}/U_{ii}) \quad (i = n, \dots, 1).$$

The linear error approximation S in (6) is a first order approximation of the absolute error $\Delta x = X - X$ of the computed solution X , that is, the first order approximation S of the absolute error of the computed solution vector X permits the representation [8].

$$8. S = U^{-1} F, F = F^0 + F^1, \text{ using the error terms.}$$

$$9. F_i^0 = -\sum_{k=i}^{n+1} S_{ik}^u X_k$$

$$10. F_i^1 = \sum_{k=i+1}^n Z_i^k e_{ik} + \sum_{k=i}^n U_{ik} X_k e_{ik}$$

and the local rounding error e_{ik} ,

$$11. e_{ii}^* = e_i, e_{ik}^* = e_{ik}^x (i < k), I = 1, \dots, n$$

of arithmetic floating-point operations in (5).

$$12. \Delta x = s + O(\eta)^2$$

Accordingly, the residual of the linear system $Ax=y$ has the form[8,9].

$$13. Ax - y = A\Delta x = As + O(\eta)^2$$

By (6) and triangular factorization $A = Lu$, the linear residual approximation $t = As$ has the representation.

$$14. t = As = LF.$$

Thus s, t can be decomposed into.

$$15. S = S^0 + S^1, S^0 = U^{-1} F^0, S^1 = U^{-1} F^1, t = t^0 + t^1, t^0 = LF^0, t^1 = LF^1.$$

The terms S^0, t^0 represent the error and residual contributions of all data error and all rounding errors of the floating-point arithmetic in forward elimination, and S^1, t^1 the error and residual contributions of rounding error occurring in back substitution.

5. Condition Numbers [9].

It is presupposed in the following that the absolute data errors $\Delta a_{ik}, \Delta y_i$

of the coefficients a_{ik} and the right-hand sides y_i of the linear system are bounded by:

- 1- $|\Delta a_{ik}| \leq \alpha_{ik} \eta_D, |\Delta y_i| \leq \beta_i \eta_D (i, k = 1(1)n)$. Where η_D is a data accuracy and α_{ik}, β_i are non-relative weights. Analogously is assumed that the relative rounding errors of the arithmetic floating-point operations of the numerical solution of the linear system are bounded by, (the following inequalities is the theoretical background for the condition numbers).

$$|e^-_{tik}| \leq \varepsilon^-_{tik} \eta_R$$

$$|e^*_{tik}| \leq \varepsilon^*_{tik} \eta_R$$

$$2- |e^-_{ik}| \leq \varepsilon^-_{ik} \eta_R$$

$$|e^*_{ik}| \leq \varepsilon^*_{ik} \eta_R$$

For more information see MAT LAB COND. Where η_R is the relative accuracy of the floating-point rounding function and $\varepsilon_{tik}, \dots, \varepsilon_{ik}$ are suitable nonnegative weights. Further let

$$3- \eta = \max(\eta_D, \eta_R) > 0.$$

The linearization method [10]. guarantees that the linear error and residual approximations s, t are first order approximations of the absolute error $\Delta x = x - \bar{x}$ and the residual $\Delta \bar{x} - y$ in the form.

- 4- $\Delta x = S + O(\eta^2), \Delta \bar{x} - y = t + O(\eta^2)$. Provided that η^2 sufficiently small. The linear error and residual approximations may be decomposed with respect to data rounding errors by.

- 5- $S = S^D + S^R, t = t^D + t^R$, where t^D, S^D, t^R, S^R , are given above the error and residual approximations S^D, t^D are linear forms in the data errors $\Delta A, \Delta Y$, for all data errors of the error distribution these linear forms are bounded component-wise by:

- 6- $|S_i^D| \leq \sigma_i^D \eta_D, |t_j^D| \leq \tau_j^D \eta_D (i, j = 1(1)n)$. Using the following data condition numbers σ_i^D of the solution x_i and residual condition numbers τ_j^D with respect to data perturbations.

$$\sigma_i^D = \sum_{j=1}^n |a_{ij}^{(-1)}| \tau_j^D$$

- 7- $\tau_j^D = \sum_{k=1}^n a_{jk} |X_k| + \beta_i (i, j = 1(1)n)$. For each i, j there is a data perturbation in the distribution such that the bounds $\sigma_i^D \eta_D, \tau_j^D \eta_D$ are attained this sense the estimates are optimal.

Numerical example and applications:

The error analysis of numerically solving two linear equations in two unknowns,

$$1- ax + by = f, cx + dy = g$$

The relative data and rounding condition numbers of computing the solutions x, y by Cramer's rule and Gaussian elimination are determined. The important results: for non-singular linear system Gramer's rule is always well-conditioned or back-ward stable.

$$2- \rho_x^R / \rho_x^D \leq 2.5, \rho_y^R / \rho_y^D \leq 2.5$$

(stability constants). Gaussian elimination is back ward stable, and

$$3- \rho_x^R / \rho_x^D \leq 2.75, \frac{\rho_y^R}{\rho_y^D} \leq 2,$$

provided that the system is properly pivoted such that $|bc| \leq |ad|$. The algorithms are analyzed further with respect to the behavior of the residuals of the computed solutions. It is shown that Gaussian elimination is, additionally well-conditioned in this sense.

Where Gramer's rule is not. It is proved that the relative condition numbers, the stability constants (2), (3), and the above pivotal strategies are invariant under scaling of the linear system.

As an example.1. [11]. let us consider a discretization of the boundary value problem $-(px') + qx = f$. The coefficients of the linear system then become $b_{i+5} = p(t_{i+5})$; $a_i = h^2q(t_i)$, $y_i = h^2f(t_i)$, $i = 1(1) N$, where $t_i = jh, j = 0(0.5) N+1$, and $h = 1/ (N+1)$. For $N = 20$ and a binary floating-point arithmetic with mantissa of 23 bits, numerical results were

calculated without data errors. Furthermore, error percentages $p_i = 100 |\Delta x_i|/(\sigma_i^R \eta)$ and residual percentages $R_i = 100 |t_i|/(\tau_j^R \eta)$ were computed. Selected results are presented in Table below. The last row exhibits the maximum value that occurred in the respective columns, taking into account all indices $i = 0, 1, \dots, N+1$.

Table1. Relative data and rounding condition numbers, error and residual percentages for Gaussian and two-sided elimination [12].

i	Gaussian elimination			Two-sided elimination			
	e_i^D	e_i^R	P_i	R_i	e_i^R	P_i	R_i
0	2.20	35.4	13	18	16.8	8	2
2	2.26	44.1	11	10	23.4	9	0
8	2.03	42.6	12	13	24.3	11	4
12	2.01	38.8	11	3	42.2	10	6
18	2.22	30.5	11	8	25.1	10	6
21	2.29	16.8	5	18	16.8	9	2
max	2.29	45.4	13	18	25.4	15	14

A Theoretical Result [13].

Using techniques of maximum principle it is easy to prove the following theorems

Theorem: if for the classical $u(x)$ solution of

$$\frac{1}{r} \frac{d}{dr} \left[r \frac{du}{dr} \right] + k_1 u - k_2 u^2 + k_3 u^3 = 0;$$

$$r \in (0,1); \lim_{r \rightarrow 0} r \frac{du}{dr} = 0, u(1) =$$

$$0, \text{ holds } u(x) \in C^2(0,1) \cap$$

$$C[0,1], \text{ So}$$

$$\max_{x \in [0,1]} u(x) \leq \frac{\sqrt{K_2^2 + 4K_1K_3 - K_2}}{2K_3}, \text{ and}$$

$$\min_{x \in [0,1]} u(x) \leq -\frac{\sqrt{K_3^2 + 4K_1K_3 + K_2}}{2K_3},$$

consequently the solution are bounded [14].

The Numerical Results

In our physical problems the constants K_2 and K_3 are characterized by the following: $K_2 \sim 10^{-4}$, $K_3 \sim 7.10^{-4}$. Using the numerical algorithm developed we obtain the following results:

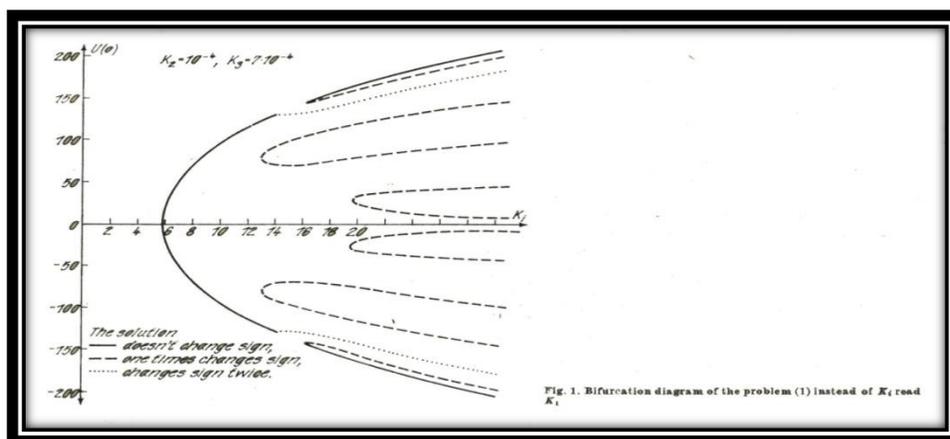


Fig.(1)Bifurcation diagram of the problem instead of R_i read K_i

The nonlinear equation $F^m(a_o^m, k_1, k_2, k_3) = 0$ was solved by the tangent parabolic method. The bifurcation on Fig.1.was constructed

for the values of $m = 8$ and 10 the results didn't differ essentially from each other.

Ex.3,[13] Solve the following square 10×10 matrix by Jacobi method

$$A = \begin{pmatrix} 12.000 & 0.0 & 1.0000 & 0.0 & 3.0000 & 4.0000 & 1.0000 & 4.0000 & 3.0000 & 1.0000 \\ 4.0000 & 3.0000 & 1.0000 & 1.0000 & 2.0000 & 2.0000 & 7.0000 & 0.0 & 7.0000 & 4.0000 \\ 7.0000 & 0.0 & 7.0000 & 4.0000 & 0.0 & 5.0000 & 3.0000 & 1.0000 & 2.0000 & 2.0000 \\ 3.0000 & 2.0000 & 4.0000 & 3.0000 & 1.0000 & 1.0000 & 5.0000 & 2.0000 & 0.0 & 1.0000 \\ 4.0000 & 1.0000 & 0.0 & 3.0000 & 5.0000 & 1.0000 & 12.0000 & 0.0 & 1.0000 & 0.0 \\ 5.0000 & 2.0000 & 0.0 & 1.0000 & 0.0 & 6.0000 & 1.0000 & 1.0000 & 0.0 & 4.0000 \\ 1.0000 & 2.0000 & 2.0000 & 0.0 & 0.0 & 1.0000 & 4.0000 & 1.0000 & 0.0 & 3.0000 \\ 1.0000 & 0.0 & 4.0000 & 0.0 & 0.0 & 1.0000 & 7.0000 & 2.0000 & 4.0000 & 1.0000 \\ 7.0000 & 2.0000 & 4.0000 & 1.0000 & 4.0000 & 2.0000 & 3.0000 & 0.0 & 1.0000 & 0.0 \\ 3.0000 & 0.0 & 1.0000 & 0.0 & 0.0 & 2.0000 & 1.0000 & 3.0000 & 2.0000 & 4.0000 \end{pmatrix}$$

Discussion:

In order to study approximation feasibility in the iterations methods (see references), it is clear from the above that the example under discussion is diverted by applying indirect methods and have given diverted results different from the correct approximate results. This is clear from the oscillated graph as the results are diverted and are opposite to the results of the direct methods noting that the iterations methods are approximate if the sequence of solutions increase in their accuracy and approach a fixed limit, otherwise it is considered diverted. It is possible to apply another idea on the solution and even develop it by increasing the number of decimal places and by

starting with the direct method to solve the system of linear equations, then following it by the iterations method as the numerical value we have achieved by applying the direct method which is the initial value for the iterations method following it. The opposite is not true and in this way we can obtain a more accurate solution by less effort. The iterations methods have been tested and the (JACOPLLOT) program has been executed on many examples. The basic programs have been executed on the personal computer (NEC system). The following is a sample of the outputs together with a printed graph which shows failure of the iterations method to solve this system as the results are oscillated as shown in the graphs (see enclosures).

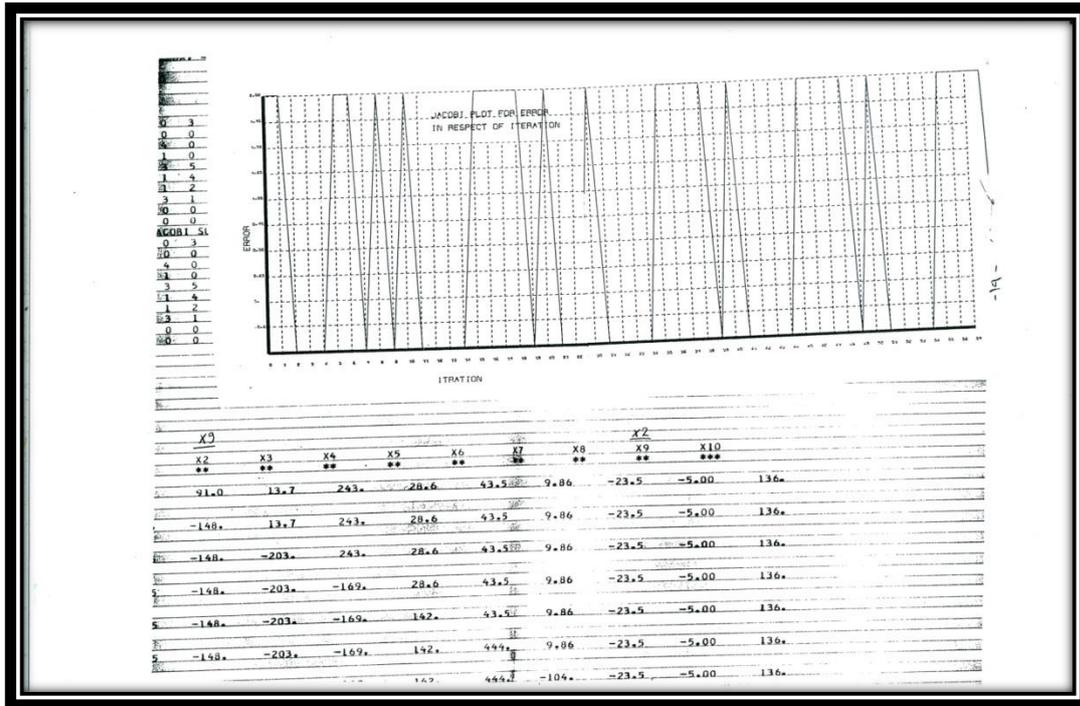


Fig. (2) Jacobi plot for error whit respect of iteration

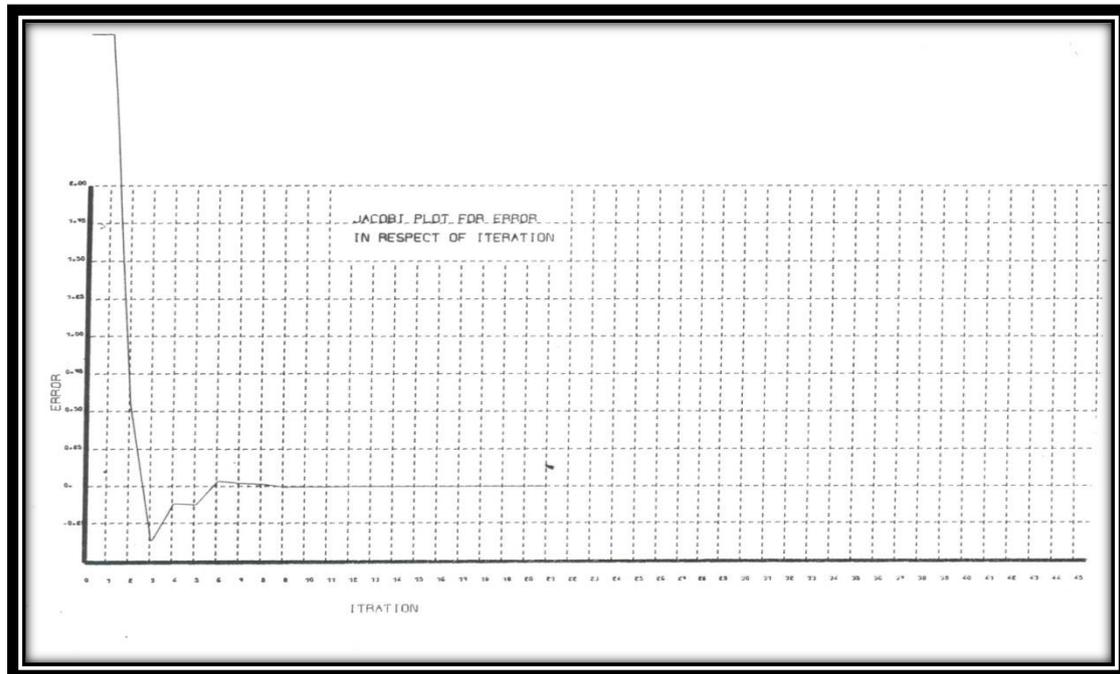


Fig. (2) Jacobi plot for error whit respect of iteration

References:

- [1]Neumaier, A. 1990. Interval methods for systems of equations. Cambridge university press, Cambridge.
- [2]Lipschutz, S. and Lipson, M. 2001. Schaum's outlines: linear Algebra Tata. McGraw-Hill edition, Delhi: 69-80.
- [3]Strangy, G. 2003. Introduction to linear algebra. 3rd edition, Wellesley, Massachusetts: Wallesely- Cambridge press: 74-76.

- [4] Stummel, F. 1985. forward errors analysis of Gaussian elimination., Numer. Math. 46: 365- 416.
- [5] Bauer F.L., 1974. Computational graphs and rounding error., SIAM.J. Numer., Anal.11:87-96.
- [6] K. Makino, M. berz and Y. Kim. Range, 2004. bounding with Taylor models. Some case studies, WSEAS Transaction on Mathematics 3: 137 -145.
- [7] Lange, K. 2010. Numerical analysis for statistic. Depart of Biomathematics, University of California, Losangeles Second edition. USA
- [8] Sadiku, M., 2014. Elements of electromagnetic, Oxford University Press. USA, 896. Covers numerical methods, including Matlab and Vector analysis.
- [9] Legendre, P. and Legendre, L. 2012. Numerical Ecology, 3rd English education, (redundancy analysis canonical correlation analysis), Elsevier science BV, Amsterdam, 1006.
- [10] Stummel, F., 1981. Perturbation theory for evaluation algorithms of arithmetic expressions, Math. Comput. 37: 435- 473.
- [11] Stummel, F. 1980. Rounding error in Gaussian elimination of tri-diagonal linear system, part 1, 11, preprint. U. Frankfurt, numer. Math . 46: 365-395.
- [12] Bobuska, I. 1972. Numerical stability in problems of linear algebra, SIAM J. Numer. Analysis 9:53-77.
- [13] Shawki, A. 1992. Perturbation theory for evaluation algorithms of arithmetic expressions, J. edu. For Women, U. of Baghdad 3:47-50.
- [14] Stummel, F. 1989. Optimal error estimates for Gaussian elimination in floating – point arithmetic, Z. Angew. Math. Mech. 63:355- 357.

ملاحظات على التعابير الرياضياتية القلقة

شوقي عبد المطلب عباس

قسم هندسة تقنيات الحاسبات / كلية النور الجامعة

الخلاصة:

في هذا البحث وضحنا الأسس النظرية لتحليل الأخطاء التقدمية للخوارزميات العددية تحت

- التقريب في دوال البناء
- حسابات أخطاء التقريب في عمليات النقطة السائبة
- التقريب في دوال البناء

عملنا يتألف من نظريات رئيسية في تحليل الأخطاء تطبيقاً إلى خوارزمية عددية بالنسبة إلى البيانات المدخلة المشوشة. تحليل الأخطاء مبني على الطريقة الخطية (Linearization method) والتي اقترحت من قبل عدة مؤلفين وبإشكال مختلفة والذي عينها أولاً العالم نيومن (Neuman) وأعقبه شتومل يعرف مفهوم العدد الشرطي على انه نهاية عظمى إلى الأخطاء الكلية المتجمعة في وقت عمل الحاسوب الذي يبني عليه ثابت الاستقرار النتائج النظرية طبقت على تحليل الأخطاء في مجموعه المعادلات الخطية في المصفوفة المربعة (MXMER) في حالة تشويش البيانات الأولية المدخلة. تعرف فكرة ثابت الاستقرار بالنسبة إلى نظام مشابه ونحصل على نتائج منظورة متكونة من هذه الثوابت بمساعدة هذه المفاهيم ومن ثم تعرف خوارزمية عددية جيدة الشروط وبعد ذلك نحصل على ثابت الاستقرار المتمثل بالنسبة الآتية:

$$6_t^R / 6_t^D = P_t^R / P_t^D$$

الكلمات المفتاحية: دوال البناء، أخطاء التدوير، البيانات القلقة، مسألة ذات قيم حدودية