NUMERICAL ANALYSIS AND COMPUTATION

Yasuhiko Ikebe
Information Science Research Center, Meisei University, Hino City 191-8506, Japan

Keywords: Linear systems of equations, accuracy, condition number, error analysis, norms, vector spaces, stable algorithm, stable problems, eigenvalues, singular values and software packages

Contents

Summary

The study of physical phenomena usually requires mathematical modeling. For the computer solution the exact mathematical model has to be approximated by a suitable numerical model. By far the most frequently used numerical models take the form of a linear system of equations.

This article is dedicated to the elementary exposition of several important concepts needed for understanding and appreciating the surprising depth of the numerical procedures for solving this seemingly well understood simple system. As we will show in the following, even the simplest linear system of two equations in two unknowns is instructive.

The types of problems we consider are linear system of n equations in n unknowns $(n = 1, 2, 3, ...)$ and eigenvalue problems.

1. Linear Systems of Equations

Mathematical modeling of physical phenomena leads to mathematical equations from which the unknown quantities are to be computed. The subject of *numerical analysis* includes the development of numerical procedures suitable for the computer solution and the relevant error analysis. The most basic equation of *Mathematical modeling* is a

linear system of n equations in n unknowns. (n may be any natural number sometimes as large as several millions),

$$\begin{cases} a_{11}x_1 + \cdots a_{1n}x_n = b_1 \\ \vdots \\ a_{n1}x_1 + \cdots a_{nn}x_n = b_n \end{cases}, \tag{1.1}$$

or in matrix form

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ & \vdots & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \tag{1.2}$$

where the coefficient *a*'s and the right hand side *b*'s are known and the *x*'s represent the unknowns to be found, or more compactly as,

$$\mathbf{Ax} = \mathbf{b}\,, \tag{1.3}$$

where

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ & \vdots & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \quad . \tag{1.4}$$

A more general linear system of m equations in n unknowns,

$$\mathbf{Ax} = \mathbf{b} \tag{1.5}$$

may be considered, where **A** is a given $m \times n$ matrix, $\boldsymbol{x}$ is the unknown $n \times 1$ matrix and **b** is a known $m \times 1$ matrix (column vector):

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ & \vdots & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \tag{1.6}$$

2. An Example

Consider the following simple linear system in two equations and two unknowns,

$$\begin{cases} 0.780x + 0.563y = 0.217 \\ 0.913x + 0.659y = 0.254. \end{cases} \tag{2.1}$$

This example is due to G.E.Forsythe. The system has a unique solution since the determinant of the coefficient matrix is nonzero:

$$\begin{vmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{vmatrix} = (0.780)(0.659) - (0.563)(0.913) = 0.000001$$

In fact, the reader can verify that $x = 1$ and $y = -1$ is the solution.

Suppose that a person A proposed an approximate solution $x = 0.999$, $y = -1.001$ and another person B $x = 0.341$, $y = -0.087$:

Unknown	Exact solution	Approximate Solution by Person A	Approximate Solution by Person B
x	1	0.999	0.341
y	-1	-1.001	-0.087

(2.2)

Let us consider the following problem: which is a better approximate solution, that of person A or that of person B?

Comparison with the exact solution shows that the first approximate solution looks better than the second solution. However, substitution of approximate solutions into the given system shows that the second approximate solution yields less discrepancy (*residuals*) between the value of the left hand side and the value of the right hand side. Indeed, for the case of the approximate solution by person A,

$$\begin{aligned} (0.780)(0.999) + (0.563)(-1.001) - 0.217 &= -0.001343 \\ (0.913)(0.999) + (0.659)(-1.001) - 0.254 &= -0.001572, \end{aligned} \tag{2.3}$$

and for the case of the approximate solution by person B,

$$\begin{aligned} (0.780)(0.341) + (0.563)(-0.087) - 0.217 &= -0.000001 \\ (0.913)(0.341) + (0.659)(-0.087) - 0.254 &= 0. \end{aligned} \tag{2.4}$$

The question, then, is: which one is really a better approximation, the first pair or the second pair? The answer: It depends on how one measures the error. Indeed, either one may be taken as a measure of error. The criterion or the choice comes from the nature of the given problem. Let us delve into this question later. For the present, we will first reformulate the question or we consider the question from a different viewpoint.

Using matrix notation,

$$\begin{cases} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{A}, \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \mathbf{x}, \\ \begin{bmatrix} x_A \\ y_A \end{bmatrix} = \begin{bmatrix} 0.999 \\ -1.001 \end{bmatrix} = \mathbf{x}_A, \begin{bmatrix} x_B \\ y_B \end{bmatrix} = \begin{bmatrix} 0.341 \\ -0.087 \end{bmatrix} = \mathbf{x}_B, \\ \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix} = \mathbf{b}, \\ \begin{bmatrix} e_A \\ f_A \end{bmatrix} = \begin{bmatrix} -0.001343 \\ -0.001572 \end{bmatrix} = \Delta\mathbf{b}_A, \begin{bmatrix} e_B \\ f_B \end{bmatrix} = \begin{bmatrix} -0.000001 \\ 0 \end{bmatrix} = \Delta\mathbf{b}_B, \end{cases} \tag{2.5}$$

the relations (2.1), (2.3) and (2.4) are written respectively

$$\mathbf{Ax} = \mathbf{b}, \mathbf{Ax}_A = \mathbf{b} + \Delta\mathbf{b}_A, \mathbf{Ax}_B = \mathbf{b} + \Delta\mathbf{b}_B \tag{2.6}$$

The second equation of (2.6) shows that $\mathbf{x}_A$, the approximate solution proposed by person A to the given equation

$$\mathbf{Ax} = \mathbf{b} \tag{2.7}$$

gives the *exact* solution of the slightly *perturbed* equation.

$$\mathbf{Ax}_A = \mathbf{b} + \Delta\mathbf{b}_A \quad . \tag{2.8}$$

Similarly , $\mathbf{x}_B$ gives the exact solution of the perturbed equation

$$\mathbf{Ax}_B = \mathbf{b} + \Delta\mathbf{b}_B \tag{2.9}$$

In other words, by perturbing the right-hand side $\mathbf{b}$ of the given equation from $\mathbf{b}$ to $\mathbf{b} + \Delta\mathbf{b}_A$ the exact solution changes from $\mathbf{x}$ to $\mathbf{x}_A$.

This viewpoint of regarding approximate solutions as exact solutions of slightly perturbed systems leads to what is known as the *backward* error analysis. Its virtue is that is allows us to introduce the concept of *stability*. This generic term refers to how the solution changes according to changes in data. In other words, we are concerned with how the solutions x and y change according to changes in data *a, b, c, d, e, f* in

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} e \\ f \end{bmatrix} \qquad \text{(recap of the first equation of (2.6))} \tag{2.10}$$

It is J.H. Wilkinson, British numerical analyst and the 1970 Turing Award recipient, who should be credited as being the earliest to use this viewpoint most effectively. His many books are still considered as classics, sometimes even as a bible. It is this viewpoint that we adopt in order to answer the question at the beginning.

3. Condition Number

Returning to the main stream, we have the following general theorem:

Theorem 3.1 Consider two sets of linear systems

$$\left.\begin{aligned} ax+by=e \\ cx+dy=f \end{aligned}\right\}, \quad \left.\begin{aligned} ax'+by'=e+e' \\ cx'+dy'=f+f' \end{aligned}\right\} \tag{3.1}$$

where $ad-bc\neq 0$. The second system is the system obtained from the first by perturbing the right-hand side, e to $e+e'$ and f to $f+f'$, respectively. The solution x and y change to x' and y' according to this perturbation. Then, the following inequality is known:

$$\frac{1}{C}\frac{|e'|+|f'|}{|e|+|f|} \leq \frac{|x-x'|+|y-y'|}{|x|+|y|} \leq C\frac{|e'|+|f'|}{|e|+|f|}, \tag{3.2}$$

where

$$C=\frac{\max\{|a|+|c|,|b|+|d|\}\cdot\max\{|c|+|d|,|a|+|b|\}}{|ad-bc|}. \tag{3.3}$$

The theorem may be interpreted as follows. If we measure the rate of change in solution by $(|x-x'|+|y-y'|)/(|x|+|y|)$, the rate of change in data by $(|e'|+|f'|)/(|e|+|f|)$, then the rate of change in solution may be as large as *C* times the rate of change in data.

The number *C* is called the *condition number*. It depends only on the coefficients *a, b, c* and *d* of the common left-hand side of Eq.(3.1) The reader can verify that $C\geq 1$ (always). The condition number of the system (2.1) turns out to be C=2,661,396, meaning that as much as 2.66 million times the change in data may be transmitted into the solution. The reader can verify the theorem by applying Theorem 6.1 where $\|\cdot\|$ is the $\|\cdot\|_1$ -norm

4. Norms and Vector spaces

Norms are measures for measuring the size of vectors or matrices. Norms are defined on a *vector space*. The concept of norm and vector space is a standard conceptual framework on which linear algebra and analysis stand. The main object of study is linear transformations, of which matrices are typical examples, from one vector space to another. The reader is referred to any standard text in linear algebra for details.

A norm is a real-valued function on a vector space. The norm is defined on a vector space (call it ***X***). The norm of $\mathbf{x}$ is denoted by $\|\mathbf{x}\|$. An infinite number of norms may be

defined in the same vector space $\boldsymbol{X}$. Any norm of the vector space $\boldsymbol{X}$ is to satisfy the following properties, where $\mathbf{x}, \mathbf{y}$ are vectors and a any scalar:

$$\begin{cases} \text{(a)} \ \|\mathbf{x}\| \geq 0; \|\mathbf{x}\| = 0 \text{ if and only if } \mathbf{x} = \mathbf{0} \\ \text{(b)} \ \|a\mathbf{x}\| = |a|\|\mathbf{x}\| \\ \text{(c)} \ \text{Triangle Inequality}: \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \end{cases} \tag{4.1}$$

Example 4.1 $\boldsymbol{X} = \mathbb{R}^2 =$ the vector space of all column vectors $\begin{bmatrix} x \\ y \end{bmatrix}$ (x, y are real numbers)

$$\text{1-norm:} \quad \|\mathbf{x}\|_1 = |x| + |y| \tag{4.2}$$

$$\text{2-norm:} \quad \|\mathbf{x}\|_2 = \sqrt{x^2 + y^2} \tag{4.3}$$

$$\infty\text{-norm:} \quad \|\mathbf{x}\|_\infty = \max\{|x|, |y|\} \tag{4.4}$$

The reader can verify the properties (a), (b), (c) listed above for any one of these norms. Proving this for 2-norm requires the use of Cauchy-Schwarz inequality

$$|ax + by| \leq \sqrt{a^2 + b^2}\sqrt{x^2 + y^2}\,. \tag{4.5}$$

For example, let $\mathbf{x} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$. Then

$\|\mathbf{x}\|_1 = |3| + |-4| = 7$

$\|\mathbf{x}\|_2 = \sqrt{3^2 + (-4)^2} = 5$

$\|\mathbf{x}\|_\infty = \max\{3, |-4|\} = 4$

The reader can generalize the concept of 1-, 2- and ∞-norms to $\mathbb{R}^n$, the vector space of all n-component real column vectors.

Example 4.2 Matrix Norm

Let $\mathbf{A}$ be an $n \times n$ real matrix. Take nay norm $\|\cdot\|$ defined on $\mathbb{R}^n$ (see Example 4.1). The definition

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| \tag{4.6}$$

gives a norm on the vector space $\mathbb{R}^{n\times n}$ of all $n \times n$ real matrices, where the right-hand side denotes the maximum value of $\|\mathbf{A}\mathbf{x}\|$ when $\mathbf{x}$ varies over all vectors $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$

The norm defined in this way is called the *operator norm* corresponding to the given vector norm. The operator norm does satisfy the axiom for norms. For any **A** and **B** in $\mathbb{R}^{n\times n}$ and any scalar *a*,

$$\begin{cases} \text{(a)} & \|\mathbf{A}\| \geq 0; \|\mathbf{A}\| = 0 \text{ if and only if } \mathbf{A} = \mathbf{0} \\ \text{(b)} & \|a\mathbf{A}\| = |a| \ \|\mathbf{A}\| \\ \text{(c)} & \|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| \end{cases} \tag{4.7}$$

In other words, the operator norm behaves like a norm and hence is a norm.

A very important inequality follows from the definition in Eq.(4.6):

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \|\mathbf{x}\| \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n \tag{4.8}$$

Replacing **A** by the product **AB** we have

$$\|\mathbf{ABx}\| \leq \|\mathbf{A}\| \ \|\mathbf{Bx}\| \leq \|\mathbf{A}\| \ \|\mathbf{B}\| \ \|\mathbf{x}\|$$

which implies

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \ \|\mathbf{B}\| \tag{4.9}$$

In other words, *the norm of the product does not exceed the product of the norms*.

Example 4.3 Given $\mathbf{A} = \left[a_{ij} \right]$,

$$\|\mathbf{A}\|_1 \equiv \max_{\|\mathbf{x}\|_1 = 1} \|\mathbf{Ax}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$$

This norm is called the “maximum column sum”.

$$\|\mathbf{A}\|_\infty \equiv \max_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{Ax}\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$$

This norm is called the “maximum row sum norm”.

$$\|\mathbf{A}\|_2 \equiv \max_{\|\mathbf{x}\|_2 = 1} \|\mathbf{Ax}\|_2 = \sqrt{\text{the maximum eigenvalue of } \mathbf{A}'\mathbf{A}},$$

where $\mathbf{A}'$ denotes the *transpose* of **A**. This norm is called the “spectral norm”.

Example 4.4 $\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix}$

$$\|\mathbf{A}\|_1 = \max\{1+3, |-2|+4\} = 6$$

$$\|\mathbf{A}\|_\infty = \max\{1+|-2|, 3+4\} = 7$$

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} 1 & 3 \\ -2 & 4 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 10 & 10 \\ 10 & 20 \end{bmatrix}$$

The eigenvalues of $\mathbf{A}'\mathbf{A}$ are given by the roots of characteristic equation

$$|\boldsymbol{A}'\boldsymbol{A} - \lambda \boldsymbol{I}| = \begin{vmatrix} 10-\lambda & 10 \\ 10 & 20-\lambda \end{vmatrix} = (10-\lambda)(20-\lambda) = 100 - 30\lambda + \lambda^2 ,$$

or

$$\lambda = 15 \pm 5\sqrt{5} \qquad .$$

Hence

$$\|\mathbf{A}\|_2 = \sqrt{15+5\sqrt{5}} = 5.116...$$

Example 4.5 *Induced Norm.*

Consider the vector space $\mathbb{R}^n$ of all n-component real column vectors and let $\|\cdot\|$ be a norm defined on it. Take any invertible $n \times n$ matrix $\mathbf{A}$. Invertible matrices are those which have an inverse. Then, the norm defined by

$$\|\mathbf{x}\|_{\mathbf{A}} = \|\mathbf{A}\mathbf{x}\|$$

gives another norm on $\mathbf{R}^n$, since it satisfies the axiom (4.1) for norms as can be verified from the definition, where invertibility of $\mathbf{A}$ is needed to have the equivalence of $\mathbf{x} = \mathbf{0}$ and $\mathbf{Ax} = \mathbf{0}$.

-
-
-

Bibliography

L.S. Blackford, J. Choi, a. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R.C. Whaley, *ScaLAPACK User's Guide*, SIAM, 1997 [The scalable LAPACK package for vector and parallel computers.]

James W. Demmel, *Applied Numerical Linear Algebra*, SIAM, 1997 [A standard graduate text in numerical linear algebra.]

Jack J. Dongarra, Iain S. Duff, Danny C. Sorensen, and Henk A. van der Vorst, *Numerical Linear Algebra for High Performance Computers*, SIAM, 1998 [An up-to-date text in numerical linear algebra

for high performance computing on vector and parallel computers, dealing with both systems of linear equations and eigenvalue problems.]

Gene. H. Golub and Charles F. Van Loan, *Matrix Computations: Third edition*, The Johns Hopkins University Press, 1996 [An up-to-date encyclopedic treatise in numerical linear algebra.]

G.W. Stewart, *Matrix Algorithms: Volume I: Basic Decompositions; Volume II:Eigensystems*, SIAM, 1998 & 2001 [The first two volumes in a projected five volumes survey of numerical linear algebra and matrix algorithms. Detailed proofs and many numerical examples are provided.]

Lloyd N. Trefethen and David Bau III, *Numerical Linear Algebra*, SIAM, 1997 [A readable introductory text in numerical examples are provided.]

J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965 [A well-known comprehensive text in numerical matrix eigenvalue problems by the late 1970 Turing award recipient. Readable and self-contained.]

Biographical Sketch

Yasuhiko Ikebe was born on Dec. 8, 1934. Dr. Ikebe received the BS and MS degrees respectively in 1957 and 57 both from Kyoto University and received the Ph. D. degree from The University of Texas at Austin in 1966. Dr. Ikebe served the following universities: Kyoto Sangyo University(1966-68), Kyoto University(1968-69), The University of Texas at Austin(1969-74, 1975-76), The University of Utah(1974-75), Northwestern University(1976-78), The University of Tsukuba(1978-95), The University of Aizu(1995-02), and Meisei University(2002-present).