Assem Deif

Sensitivity Analysis in Linear Systems

Springer-Verlag

Assem Deif

Sensitivity Analysis in Linear Systems

With 15 Figures

Springer-Verlag Berlin Heidelberg NewYork
London Paris Tokyo

Prof. Assem Deif
Department of Engineering Mathematics
Cairo University
Giza/Egypt

ISBN 3-540-16312-3 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-16312-3 Springer-Verlag New York Heidelberg Berlin

Library of Congress Cataloging in Publication Data:
Deif, Assem:
Sensitivity analysis in linear systems / Assem Deif.
Berlin, Heidelberg, New York: Springer, 1986.
ISBN 3-540-16312-3 (Berlin . . .)
ISBN 0-387-16312-3 (New York . . .)

Printing: Sala-Druck, Berlin
Bookbinding: Schöneberger Buchbinderei, Berlin
2161/3020-543210

Preface

A text surveying perturbation techniques and sensitivity analysis of linear systems is an ambitious undertaking, considering the lack of basic comprehensive texts on the subject. A wide-ranging and global coverage of the topic is as yet missing, despite the existence of numerous monographs dealing with specific topics but generally of use to only a narrow category of people. In fact, most works approach this subject from the numerical analysis point of view. Indeed, researchers in this field have been most concerned with this topic, although engineers and scholars in all fields may find it equally interesting.

One can state, without great exaggeration, that a great deal of engineering work is devoted to testing systems' sensitivity to changes in design parameters. As a rule, high-sensitivity elements are those which should be designed with utmost care. On the other hand, as the mathematical modelling serving for the design process is usually idealized and often inaccurately formulated, some unforeseen alterations may cause the system to behave in a slightly different manner. Sensitivity analysis can help the engineer innovate ways to minimize such system discrepancy, since it starts from the assumption of such a discrepancy between the ideal and the actual system.

All in all, methods of mathematical optimization rely one way or the other on relative sensitivities, under a different title in each, ranging from gradient methods to model tracking or self-learning systems. Even the simple task of fitting data to a curve usually involves sensitivity calculations. As for social scientists, economists, as well as for many other disciplines, sensitivity and perturbation techniques can provide valuable information about the amount of inaccuracy in the behaviour of a model as related to the inaccuracies in the system's data. If the data gathered by field study or experimental testing falls within certain tolerance limits, the tolerances may well be amplified and widened in the output results obtained. The question might then arise as to how uncertain the results are — or how unrealiable — in relation to the data's uncertainties. In this instance, perturbation analysis can provide valuable information about regions of compatibility and admissibility of solutions. An alternate use might also be to determine the allowable data tolerances in a parameter for the results to sustain a certain level of accuracy; and so forth.

As rewarding a subject as it may be, sensitivity analysis still imposes a tedious job when it comes to organizing a text on it. As the text is intended to serve a wide audience, applications of various kinds had to be included, and a huge effort had to be devoted to ensuring as comprehensive a discussion as possible of the area of linear systems. Some texts have tried to attract the widest readership by choosing some topics of the linear systems and some of the nonlinear ones. Practical experience has shown that such texts have little to add to the real user's knowledge, and only serve to provide an idea about the subject. The present line of approach is stronger, in that it provides not only practical applications but mainly a coherent mathematical justification that brings still further applications within range.

Perturbation techniques and sensitivity analysis are of course no new terms, nor are they recently explored fields. As a mathematical discipline, however, a unified body of knowledge rather than an elementary application, they are rather young. Only during this last century have celestial mechanics witnessed an era of rapid progress; the three-body problem — in contrast with Newton's two body problem — becoming the new challenge. Workers in the field opted to consider this third body as a perturbation in the field. In this context also, Lagrange's ingenious method of variation of elements was introduced, and Poincaré's theory of asymptotic expansions enabled the summing up of a few terms of a divergent series to yield almost exactly its sum.

Perturbation theory in linear algebra is an even more recent branch. In 1948, Turing's famous paper triggered interest in the problem of sensitivity of solutions of linear equations to round-off errors. In this paper, Turing laid down the definition of a condition number by which a small input error in the data can be drastically amplified in the solution. Numerical analysts then acknowledged this number as the major factor affecting computational accuracy, and have tried since then to control it while working out any new numerical procedure. But an ill-conditioned system can only be cured up to a certain extent, and no matter how cunning, skilled or elaborate one is, Turing's number or a variant of it will eventually hinder our illusion.

This work covers the subject of sensitivity analysis as related to linear equations. At first, the plan was to furnish one for linear systems in general, but as the work grew it was found impossible to survey the whole theory in one volume. It was then decided to release the available material as it constitutes a unified body of knowledge. Naturally, we started with the first basic problem in linear systems, that of linear equations, for we shall need many of the results if we are to proceed further. As the reader might have noticed, it was not in our plan to either survey or compare the different numerical methods for solving the equations. For this, he may refer to current literature in numerical analysis, which is plentiful. Rather, our task was only concerned with the problem of executing perturbation analysis of the equations, while making reference to some relevent applications.

The desire on the part of the author to provide such a text grew up incidentally from his training in engineering and related disciplines. Workers in theses fields sometimes encounter problems in which it is important to perceive the accuracy of the results when the input data is subject to uncertainty or errors. This they seek to determine irrespective of any numerical treatment of the problem. They may also wish to pursue a sensitivity analysis of their models under newly varying conditions. It was this philosophy that inspired the writing of the text and which led the author to prepare the rest of the manuscript. This explains why rounding errors are incorporated into the larger context of perturbations, and why no effort has been made to discuss error analysis from the view point of comparing different numerical strategies. In any case, there is a vast literature on this subject alone. And instead of making entangling the reader so overwhelmingly in the different numerical procedures for solving the equations, it concentrates on deepening his working knowledge in this fruitful area. This does not at all mean that numerical analysts cannot profit from it. On the contrary, many up-to-date error bounds have been included and compared. Furthermore, criteria for validating the solutions and ways of improving them are to be found therein.

The text is therefore a survey and a working knowledge book on perturbation techniques and sensitivity analysis as applied to linear equations and linear programming. It uses a moderate language which appeals to engineers and applied mathematicians. Many workers in various disciplines will find it equally valuable. And as a text, it can easily fit — from experience — into a first course on the subject to be taught in one lecture per week over one semester. As to its rigor, it will soon be realized that there is some overlap in levels, in the sense that some knowledge is standard while some is culled from original papers. Our intention was to encourage readers of different backgrounds and training to approach the subject.

The book consists of five chapters, each related to a specific case so as to make it self-contained. Hence, it will be found that it is not strictly necessary to read the text from the beginning. In other words, it looks as though each chapter is independently written for the reader interested in one specific subject. And we certainly make no claim to completeness in any one of them.

In conclusion, the author feels that such a subject deserves a global coverage which communicates efficiently with the different audiences. And with this specific idea in mind, we hope to have fulfilled this aim and to have filled a gap in the available literature.

Indeed, a word of gratitude should be addressed to all who have contributed to the appearance of this book. Prof. P. Spellucci at the Technische Hochschule in Darmstadt, Prof. T. Yamamoto at Ehime University and Dr. J. Garloff at the Universität Freiburg read various parts of the manuscript and made several suggestions which greatly improved the text and rescued me from many blunders. Any remaining pit-falls become naturally those of the author. Dr. J. Rohn at Charles University provided the author with a mathematical notion which helped in some proofs in section 2.6. The author is also grateful to Prof. N. Makary and Prof. E. Rasmy for fruitful discussions related respectively to sections 4.5 and 5.3, and also to Dr. A. Hussean and Dr. T. El-Mistikawy who helped in calculations of sections 3.4 and 4.3; all are based at Cairo University. The whole project of writing this text would have indeed been foundered without permission from Cairo University to allow me to visit the University of California-Berkeley on my sabbatical leave, thus making use of its rich library. The same also applied to the Alexander von Humboldt Foundation, which gave me the opportunity to visit the Technische Hochschule in Darmstadt, enabling me also to profit from the staff there as well as its different libraries. I would like also to take the opportunity of thanking Prof. Å. Björck for his valuable comments on part of the manuscript. My sincere thanks are also addressed to Eng. A. Assaad for editing the text; the effort he put into this work is greatly appreciated. Finally, I am grateful to my wife for allowing me ample time on the manuscript.

Cairo, August, 1986

A. S. Deif

Table of Contents

Chapter 1

Perturbation of Linear Equations

1.1 Introduction

This chapter will discuss the behaviour of the system of linear simultaneous equations

$$Ax = b$$

when the matrix A and the vector b are subjected to small order perturbations ΔA and Δb respectively. The problem then becomes

$$(A + \Delta A)(x + \Delta x) = b + \Delta b$$

and we are mainly concerned with studying the deviation Δx of the solution with the perturbation. Such an exercise is called *sensitivity analysis*, for the extent of the deviation Δx relative to ΔA and Δb defines the *sensitivity* of the system. A highly sensitive system is roughly one where a relatively large deviation Δx is incurred by small perturbations ΔA and Δb. As we shall see, highly sensitive systems are generally to be avoided in practice. They are referred to as *ill-conditioned*, for a highly sensitive system would yield large variations in the results for only small uncertainties in the data. To clarify this fact, let us study the behaviour of the system of equations

$$\begin{aligned} x + y &= 2 \\ 0.49x + 0.51y &= 1 \end{aligned}$$

— representing a pair of straight lines intersecting at $x = 1$, $y = 1$ — when a small term ε is added to the equations. Surprisingly enough, the set of equations obtained, namely

$$\begin{aligned} x + y &= 2 + \varepsilon \\ 0.49x + 0.51y &= 1 \end{aligned}$$

represents a pair of straight lines meeting at a point (x, y) given by $x = 1 + 25.5\varepsilon$; $y = 1 - 24.5\varepsilon$, and rather distant from $x = 1, y = 1$. Here, a small change of order ε in the equations has produced a change of 25ε in the solution. This system has definitely a high sensitivity to perturbations in its matrix A and vector b.

It is indeed worth noting that sensitivity analysis is usually performed after a problem has been solved for x. It is not intended to find an adjusted solution for the system

$$(A + \Delta A)\, x = b + \Delta b$$

for then, the effect of the perturbations ΔA and Δb on x will have remained unexamined. Rather, it aims at representing Δx as a function of the perturbations ΔA and Δb, in order to elicit their effect on the original solution. For the forementioned example, this function would be given by

$$\frac{\Delta x}{x} \cong 50 \frac{\Delta b_1}{b_1}$$

At this stage, let us examine the origins of the perturbations ΔA and Δb of a system of linear equations and their possible practical meanings. In the field of economy, social sciences, etc . . ., system perturbations usually stem from the lack of precision of the data collected through field observations, this lack being termed *uncertainty of the data*. For the engineer they may be intentionally induced into one of the design parameters to investigate the behaviour in service of the designed system, be it an electrical system, a chemical plant or a building's structure. For the mathematician, perturbations might appear as the result of truncating some infinite series, say π or e.

Finally, coming last though not least, perturbations represent in numerical analysis the effect of round-off errors. Such errors might arise during data reading, or in the course of computation, at the level of an intermediate result. Of all disciplines, numerical analysts have been most interested in this area for the sake of precision in the results they obtain.

This text will not assign to ΔA and Δb any one of the foregoing interpretations, dealing with them in a most general manner through their symbols. This approach does not void the symbols from their factual content. Instead, it keeps them in a form so general that they can account at the same time for all perturbations inherent to — or induced in the system. On some occasions, some interpretation or the other will be emphasized, only for the sake of illustrating the concept. Being far from specialization, this text is not intended for the sole use of numerical analysts. For the more specialized application of error analysis to various computational algorithms, the reader is referred to Wilkinson's treatises (1963, 65), which form only a part of the literature on the subject.

However, it may be of interest to illustrate here how rounding-off can be accounted for as a perturbation. Let us therefore consider the solution for $x = (x_1, x_2)^T$ of the system

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

when performed on a four-digit machine with fixed-point arithmetic. The solution comes as

$$\hat{x}_1 = 1.667\,, \qquad \hat{x}_2 = -1.333$$

This is only an approximation to the exact solution, since substitution into the set of equations yields the residual

$$r = A\hat{x} - b = (0.001, 0.001)^T$$

The error residual is due to the machine's approximation of the solution which is more exactly

$$x_1 = 5/3 \qquad x_2 = -4/3$$

Alternatively, one can state that the vector $\hat{x}$ obtained is an exact solution of the perturbed system

$$A\hat{x} = b + \Delta b$$

where $\Delta b = r$.

In any case, whether ΔA and Δb account for errors due to rounding, truncation or inaccuracy in the input data, the solution obtained will still deviate from the ideal case. For the above example, the solution's deviation can be expressed by the vector

$$\begin{aligned}\Delta x &= (1.667 - 5/3\,, \qquad -1.333 + 4/3)^T\\ &= (0.001/3\,, \qquad 0.001/3)^T\end{aligned}$$

Therefore, the value 0.001/3 could be taken as a measure of the deviation, or, alternatively, the sum of deviations in both variables, namely 0.002/3, could be chosen, etc . . . In general, how to measure this deviation and in what form is one question, the answer to which necessitates the introduction of the concept of norm.

1.2 Norms of Vectors and Matrices

The norm of a vector is introduced here to provide a measure of the vector's magnitude exactly analogous in concept to that of absolute value for a complex number. The *norm* of a vector x, denoted $\|x\|$, is a non-negative scalar function of x satisfying the following set of axioms

$$\begin{aligned}&\|x\| > 0\,, \quad \forall x \neq 0 && \text{(Positivity)}\\ &\|cx\| = |c| \cdot \|x\| && \text{(Homogeneity)}\\ &\|x + y\| \leqq \|x\| + \|y\| && \text{(Triangular inequality)}\end{aligned}$$

In general, a scalar function $\|\cdot\|$ satisfying the above axioms qualifies as a norm. The reader can exercise in showing that the following function, termed the *Hölder norm*, and stated as

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \qquad p \geqq 1$$

qualifies as a norm of x. For a proof of the validity of the triangular inequality for this norm, the reader is referred to Deif (1982) for p assuming integer values. For p assuming general values, the reader may refer to Beckenback and Bellman (1965). In the special case where $p = 2$, the general norm function yields the well known Euclidean norm $\|\cdot\|_2$ or $\|\cdot\|_E$, which gives the length of a vector in analytical geometry. Other widely used forms of the norm function are

$$\|x\|_1 = \sum_i |x_i|, \qquad \text{for} \quad p = 1$$

and

$$\|x\|_\infty = \max_i |x_i|, \quad \text{for} \quad p \to \infty$$

Applying each of the foregoing versions of norm to a vector given by

$$x = (1, -2, 3 + i)^T, \qquad i = \sqrt{-1}$$

would yield

$$\begin{aligned} \|x\|_1 &= 1 + 2 + \sqrt{10} = 3 + \sqrt{10} \\ \|x\|_E &= \sqrt{1 + 4 + 10} = \sqrt{15} \\ \|x\|_\infty &= \sqrt{10} \end{aligned}$$

Likewise, the norm of a square matrix A is defined as a nonnegative scalar function noted $\|A\|$ and satisfying the following axioms

$$\begin{aligned} &\|A\| > 0, \qquad \forall A \neq 0 \\ &\|cA\| = |c| \; \|A\| \\ &\|A + B\| \leqq \|A\| + \|B\| \\ &\|AB\| \leqq \|A\| \; \|B\| \end{aligned}$$

Again, many functions could be found that qualify as matrix norms, according to the above rules, e.g.

$$\|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2} \qquad \text{(Frobenius norm)}$$

$$\|A\|_M = n \cdot \max_{i,j} |a_{ij}| \qquad \text{(Maximum norm)}$$

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \qquad \text{(Column norm)}$$

$$\|A\|_\infty = \max_i \sum_j |a_{ij}| \qquad \text{(Row norm)}$$

Some have even more properties than those described by the axioms. The last two norms defined above — Row and Column norms — may for instance be *subordinated* to a corresponding vector norm; that is for every matrix A one can always find a vector x such that

$$\|Ax\| = \|A\| \; \|x\|$$

(cf. Young and Gregory (1973); Deif (1982)). Many matrix norms are *consistent* with vector norms, e.g. the Frobenius matrix norm with the Euclidean vector norm, meaning that they satisfy the relation

$$\|Ax\| \leqq \|A\| \; \|x\|$$

However, the equality part of a consistency relation is only verified for specific configurations of A and x, whereas for the Row and Column matrix norms, each matrix A can be subordinated to at least one vector norm $\|x\|$.

In virtue of this additional property, subordination to a vector norm, the row and column norms — and all norms analogous in this respect — are termed *bounds*, or better still, *least upper bounds* (lub). This means that

$$\|A\|_p = \mathrm{lub}_p(A) = \sup_x \frac{\|Ax\|_p}{\|x\|_p}$$

In the cases where $p = 1, \infty$ in the above relation, we obtain respectively the formentioned Column and Row norms. As an example, if

$$A = \begin{bmatrix} 1 & 0 & -1 \\ -3 & 5 & 4 \\ 2 & -2 & 0 \end{bmatrix}$$

then

$$\|A\|_1 = 7\,, \qquad \|A\|_\infty = 12$$

The derivation of lub(A) necessitates the use of a vector x, whence some authors refer to it as the *induced matrix norm.* If we consider the Frobenius norm, noted $\|A\|_F$ (and held as the matrix analogue of the Euclidean length of a vector), as compared to its corresponding induced norm-version (called the *spectral norm*) given by

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^*A)}$$

where $\lambda_{\max}$ stands for the largest eigenvalue, we find that

$$\|A\|_2 \leqq \|A\|_F$$

This characteristic makes induced matrix norms widely used in relation to error analysis, as they set tighter bounds.

Usually, authors do not differentiate in notation between both types of norms, unless the need rises for the exclusive use of one. In this text, the analytical expressions derived will be valid for nearly all norms, the bounds set by their numerical values being tighter or looser according to the type of norm used. Furthermore, as norms were devised to quantify and compare magnitudes of vectors and matrices, we may settle as analysis proceeds for using the popular lazy quotation: for some norm — as many authors incidentally do. Also, $\|A\|$ will denote any matrix norm, including the cumbersome notation lub (A), further replaced by $\|A\|_1$, $\|A\|_2$, ... , $\|A\|_\infty$.

These definitions of norms, already so well known and frequently used in functional analysis, were employed in 1950 by Faddeeva, in the context of proofs of convergence. Faddeeva defined vector and matrix norms independently, linking them with the concepts of consistency and subordination. The Frobenius norm is called the *absolute value of the matrix* by Wedderburn (1934), who in turn traces the idea back to Peano.

Another type of norm, the Dual norm, was introduced by von Neumann (1937) to be treated axiomatically by many authors, see for instance Stoer (1964). The *Dual norm* of a vector u, noted $\|u\|^D$, is defined by

$$\|u\|_p^D = \sup_x \frac{|\langle u \cdot x\rangle|}{\|x\|_p}$$

For example, for the vector

$$u^T = (1, -2, 3 + i)\,, \qquad i = \sqrt{-1}$$

the dual norm is given by

$$\|u\|_1^D = \sup_x \frac{|\langle u \cdot x\rangle|}{\sum_i |x_i|} = \sup_x \frac{|x_1 - 2x_2 + (3 + i)\, x_3|}{|x_1| + |x_2| + |x_3|}$$

$$= \sqrt{10}\,, \qquad \text{taking} \quad x_1 = 0; \qquad x_2 = 0; \qquad x_3 = 1$$

On the other hand

$$\|u\|_\infty^D = \sup_x \frac{|\langle u \cdot x\rangle|}{\max_i |x_i|} = 1 + 2 + \sqrt{10} = 3 + \sqrt{10}$$

taking $x_1 = 1$; $x_2 = -1$; $x_3 = (3 - i)/\sqrt{10}$; $i = \sqrt{-1}$.

This concept of Dual norms is no more than a direct application of Hölder's inequality, stated as

$$\sum_{i=1}^n |u_i x_i| \leq \left(\sum_{i=1}^n |u_i|^p\right)^{1/p} \cdot \left(\sum_{i=1}^n |x_i|^q\right)^{1/q}$$

where u_i and x_i are two sets of numbers; $i = 1, 2, \ldots, n$; and

$$1/p + 1/q = 1\,, \qquad p \geqq 1$$

In fact, if we allow x_i to vary, there will surely exist a homogeneous configuration of x for which the equality strictly holds. In that case $\|u\|_1^D = \|u\|_\infty$ and vice-versa.

What is most interesting for our purposes is however the application of the concept of Dual norm to find the value of lub (uv^*), uv^* being a matrix

$$\operatorname{lub}(uv^*) = \sup_x \frac{\|uv^*x\|}{\|x\|} = \sup_x \frac{|v^*x|\ \|u\|}{\|x\|} = \|u\|\ \|v\|^D$$

Other similar dual matrix norms can also be derived. The reader interested in further information is referred to Stoer (1964), who also deduced some further properties of dual norms as compared to usual norms.

Returning to the system of linear equations

$$Ax = b$$

we recall having stated that solving for x with an accuracy depending on the used computing machine always yields the residual vector r

$$r = A\hat{x} - b$$

This vector r should in some way give an indication of the accuracy of $\hat{x}$. Writing

$$A(x + \Delta x) - b = r$$

$x = A^{-1}b$ being the system's exact solution, we get

$$A\ \Delta x = r$$

or

$$\Delta x = A^{-1}r$$

i.e.

$$\|\Delta x\| = \|A^{-1}r\| \leqq \|A^{-1}\|\ \|r\|$$

so that

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\ \|r\|}{\|A^{-1}b\|} \leq \frac{\|A^{-1}\|\ \|r\|}{\|A\|^{-1}\ \|b\|} = \|A\|\ \|A^{-1}\|\ \frac{\|r\|}{\|b\|}$$

From this, we can conclude that the magnitude of the relative error in the solution x is bounded by the norm of the residual vector r times the quantity $\|A\|\ \|A^{-1}\|$. This latter quantity is termed the *condition number* of A (noted cond (A)). It plays a vital role in assessing the numerical stability of algorithms, and deserves to be discussed separately. For the time being, let us just clarify the concept by a practical example. Considering the same example discussed before, we will solve

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

using a four-digit machine with fixed-point arithmetic. Then

$$\hat{x}_1 = 1.667\,, \qquad \hat{x}_2 = -1.333$$

and

$$r = A\hat{x} - b = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1.667 \\ -1.333 \end{bmatrix} - \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.001 \\ 0.001 \end{bmatrix}$$

Meanwhile

$$A^{-1} = \begin{bmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}$$

giving

$$\operatorname{cond}(A) = \|A\|\,\|A^{-1}\| = 3$$

Hence, using the l_1-norm, we get

$$\frac{\|\Delta x\|}{\|x\|} \leqq \|A\|\,\|A^{-1}\|\frac{\|r\|}{\|b\|} = 3 \times 0.002/3 = 0.002$$

Now if we use the exact value of x

$$\Delta x = (0.00033\,\ldots\,,\ 0.00033\,\ldots)^T$$

we get

$$\frac{\|\Delta x\|}{\|x\|} = 0.00033\,\ldots \times 2/3 = 0.00022\,\ldots$$

which is of course smaller than the foregoing upper bound. In both cases, the error was small, because — as we will be explaining shortly — *A is well-conditioned.*

Furthermore, we notice that the error in both $\hat{x}_1$ and $\hat{x}_2$ does not exceed 5×10^{-4}. This is simply due to the fact that we used a three-decimal-place precision with the third decimal place rounded according to whether the fourth decimal place is greater or smaller than 5. In this specific example, the exact solution is

$$x_1 = 5/3 = 1.66666\,\ldots$$
$$x_2 = -4/3 = -1.333333\,\ldots$$

which yielded after rounding-off

$$\hat{x}_1 = 1.667 \qquad \hat{x}_2 = -1.333$$

By cancelling the fourth decimal place without knowing its value, we have induced an error of 0.0005 at most. For a number displayed with t decimal places, the error would likewise not exceed $5 \times 10^{-t-1} = \frac{1}{2} \times 10^{-t}$. Most calculating machines use floating-point arithmetic for minimal error and better accuracy. The reason behind this becomes clear when a number like 125.7235124 is to be represented in a ten-digit machine. The inherent error does not exceed 5×10^{-8}, but its absolute value is variable, depending on the place of the decimal point. For numbers between 1 and 9.99 ... inclusive, the error will not exceed 5×10^{-10}; however it is not so readily determinable for other numbers. The introduction of floating-point arithmetic simplified this issue by setting the accuracy of the machine itself. Any number is stored as

$$c \times 10^b, \qquad 10 > |c| \geqq 1 \qquad (\text{normalized floating-point: } 1 > |c| \geqq .1)$$

The number c is called the *mantissa*, and has as many digits as the machine itself. The number b is the *exponent*. The figure set forth before can thus be represented as 125.7235124 — the error being 5×10^{-8} — or as 1.257235124×10^2 — the error being now $5 \times 10^{-10} \times 10^2 = 5 \times 10^{-8}$. The accuracy of the machine is to the nearest 5×10^{-10}, and for a number a, the error is at most as large as $5 \times 10^{-10} \times |a|$. For a machine with a t-digit mantissa, the error becomes $5 \times 10^{-t} \times |a|$.

Now, supposing a matrix A is to be be processed on the computer, what would be the maximum error in the norm of A due to rounding-off? In other words, what is the bound for $\|A + \Delta A\| - \|A\|$ due to rounding-off? Since the error in a_{ij} is less than $5 \times 10^{-t} \times |a_{ij}|$ (note that sometimes only part of the machine's mantissa is displayed), then $\|\Delta A\| \leqq 5 \times 10^{-t}\|A\|$. While assuming such error to be less than $5 \times 10^{-t} \times \max_{i,j} |a_{ij}|$ for $a_{ij} \neq 0$ and zero for $a_{\mathrm{ij}} = 0$ (see Rice (1981), p. 136), we get

$$\big|\, \|A + \Delta A\| - \|A\| \,\big| \leq \|\Delta A\| \leq 5 \times 10^{-t} \times n \times \max_{i,j} |a_{ij}|$$

But we have that

$$\|A\| \geq \max_{i,j} |a_{ij}|$$

whence

$$\left| \frac{\|A + \Delta A\| - \|A\|}{\|A\|} \right| \leq 5 \times 10^{-t} \times n$$

For instance, a matrix A of order n processed on an HP-15C calculator — having a ten-digit mantissa — has an error $\|\Delta A\|$ less than $5 \times 10^{-10} \times n\|A\|$ (i.e. less than $10^{-9}\, n\|A\|$).

1.3 Condition Number and Nearness to Singularity

The condition number of square matrix A, noted cond (A), is a nonnegative real scalar given by

$$\operatorname{cond}(A) = \|A\| \, \|A^{-1}\|$$