国际著名数学图书——影印版

Accuracy and Stability of Numerical Algorithms

Second Edition

数值算法的精确性与稳定性

（第2版）

Nicholas J. Higham 著

TSINGHUA UNIVERSITY PRESS

siam

Accuracy and Stability of Numerical Algorithms

Second Edition

数值算法的精确性与稳定性

（第2版）

Nicholas J. Higham　著

TSINGHUA UNIVERSITY PRESS
北京

siam

Nicholas J. Higham
Accuracy and Stability of Numerical Algorithms, Second Edition
ISBN:0-89871-521-0

图书在版编目（CIP）数据

数值算法的精确性与稳定性 = Accuracy and Stability of Numerical Algorithms：第2版：英文/（美）汉安（Higham，N. J.）著.-- 影印本. -- 北京：清华大学出版社，2011.2
（国际著名数学图书）
ISBN 978-7-302-24493-6

Ⅰ. ①数…　Ⅱ. ①汉…　Ⅲ. ①数值计算-计算方法-英文　Ⅳ. ①O241

中国版本图书馆CIP数据核字（2011）第006597号

责任编辑：陈朝晖
责任印制：王秀菊
出版发行：清华大学出版社　　**地　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　销：全国新华书店
开　本：175×245　**印　张**：44.75
版　次：2011 年 2 月第 1 版　　**印　次**：2011 年 2 月第1 次印刷
印　数：1～3000
定　价：80.00 元

产品编号：027803-01

Accuracy and Stability of Numerical Algorithms

Dedicated to
Alan M. Turing
and
James H. Wilkinson

Preface to Second Edition

We dare not lengthen this book much more,
lest it be out of moderation and should
stir up men's apathy because of its size.
— AELFRIC, schoolteacher of Cerne Abbas,
later Abbot of Eynsham (c. 995–1020)

In the nearly seven years since I finished writing the first edition of this book research on the accuracy and stability of numerical algorithms has continued to flourish and mature. Our understanding of algorithms has steadily improved, and in some areas new or improved algorithms have been derived.

Three developments during this period deserve particular note. First, the widespread adoption of electronic publication of journals and the increased practice of posting technical reports and preprints on the Web have both made research results more quickly available than before. Second, the inclusion of routines from state-of-the-art numerical software libraries such as LAPACK in packages such as MATLAB* and Maple† has brought the highest-quality algorithms to a very wide audience. Third, IEEE arithmetic is now ubiquitous—indeed, it is hard to find a computer whose arithmetic does not comply with the standard.

This new edition is a major revision of the book that brings it fully up to date, expands the coverage, and includes numerous improvements to the original material. The changes reflect my own experiences in using the book, as well as suggestions received from readers.

The changes to the book can be summarized as follows.

New Chapters

- *Symmetric Indefinite and Skew-Symmetric Systems* (Chapter 11). A greatly expanded treatment is given of symmetric indefinite systems (previously contained in the chapter *Cholesky Factorization*) and a new section treats skew-symmetric systems.

- *Nonlinear Systems and Newton's Method* (Chapter 25). Results on the limiting accuracy and limiting residual of Newton's method are given under general assumptions that permit the use of extended precision in calculating residuals. The conditioning of nonlinear systems, and termination criteria for iterative methods, are also investigated.

*MATLAB is a registered trademark of The MathWorks, Inc.
†Maple is a registered trademark of Waterloo Maple Software.

New Sections

- *Fused Multiply-Add Operation* (§2.6). The advantages of this operation, which is included in the Intel IA-64 architecture, are discussed, along with some subtle issues that it raises.

- *Elementary Functions* (§2.10). We explain why it is difficult to compute elementary functions in a way that satisfies all the natural requirements, and give pointers to relevant work.

- *Matrix Polynomials* (§5.4). How to evaluate three different matrix generalizations of a scalar polynomial is discussed.

- *More Error Bounds* (§9.7). Some additional backward and forward error bounds for Gaussian elimination (GE) without pivoting are given, leading to the new result that GE is row-wise backward stable for row diagonally dominant matrices.

- *Variants of Gaussian Elimination* (§9.9). Some lesser-known variants of GE with partial pivoting are described.

- *Rank-Revealing LU Factorizations* (§9.12). This section explains why LU factorization with an appropriate pivoting strategy leads to a factorization that is usually rank revealing.

- *Parallel Inversion Methods* (§14.5). Several methods for matrix inversion on parallel machines are described, including the Schulz iteration, which is of wider interest.

- *Block* 1*-Norm Estimator* (§15.4). An improved version of the LAPACK condition estimator, implemented in MATLAB's **condest** function, is outlined.

- *Pivoting and Row-Wise Stability* (§19.4). The behaviour of Householder QR factorization for matrices whose rows are poorly scaled is examined. The backward error result herein is the only one I know that requires a particular choice of sign when constructing Householder matrices.

- *Weighted Least Squares Problems* (§20.8). Building on §19.4, an overall row-wise backward error result is given for solution of the least squares problem by Householder QR factorization with column pivoting.

- *The Equality Constrained Least Squares Problem* (§20.9). This section treats the least squares problem subject to linear equality constraints. It gives a perturbation result and describes three classes of methods (the method of weighting, null space methods, and elimination methods) and their numerical stability.

- *Extended and Mixed Precision BLAS* (§27.10). A brief description is given of these important new aids to carrying out extended precision computations in a portable way.

Other Changes

In the error analysis of QR factorization in the first edition of the book, backward error bounds were given in normwise form and in a componentwise form that essentially provided columnwise bounds. I now give just columnwise bounds, as they are the natural result of the analysis and trivially imply both normwise and componentwise bounds. The basic lemma on construction of the Householder vector has been modified so that most of the ensuing results apply for either choice of sign in constructing the vector. These and other results are expressed using the error constant $\tilde{\gamma}_n$, which replaces the more clumsy notation γ_{cn} used in the first edition (see §3.4).

Rook pivoting is a pivoting strategy that is applicable to both GE for general matrices and block LDL^T factorization for symmetric indefinite matrices, and it is of pedagogical interest because it is intermediate between partial pivoting and complete pivoting in both cost and stability. Rook pivoting is described in detail and its merits for practical computation are explained. A thorough discussion is given of the choice of pivoting strategy for GE and of the effects on the method of scaling. Some new error bounds are included, as well as several other results that help to provide a comprehensive picture of current understanding of GE.

This new edition has a more thorough treatment of block LDL^T factorization for symmetric indefinite matrices, including recent error analysis, rook pivoting, and Bunch's pivoting strategy for tridiagonal matrices. Aasen's method and Bunch's block LDL^T factorization method for skew-symmetric matrices are also treated.

Strengthened error analysis includes results for Gauss–Jordan elimination (Theorem 14.5, Corollary 14.7), fast solution of Vandermonde systems (Corollary 22.7), the fast Fourier transform (FFT) (Theorem 24.2), and solution of circulant linear systems via the FFT (Theorem 24.3).

All the numerical experiments have been redone in the latest version, 6.1, of MATLAB. The figures have been regenerated and their design improved, where possible. Discussions of LAPACK reflect the current release, 3.0.

A major effort has gone into updating the bibliography, with the aim of referring to the most recent results and ensuring that the latest editions of books are referenced and papers are cited in their final published form. Over 190 works published since the first edition are cited. See page 587 for a histogram that shows the distribution of dates of publication of the works cited.

In revising the book I took the opportunity to rewrite and rearrange material, improve the index, and fine tune the typesetting (in particular, using ideas of Knuth [745, 1999, Chap. 33]). Several research problems from the first edition have been solved and are now incorporated into the text, and new research problems and general problems have been added.

In small ways the emphasis of the book has been changed. For example, when the first edition was written IEEE arithmetic was not so prevalent, so a number of results were stated with the proviso that a guard digit was present. Now it is implicitly assumed throughout that the arithmetic is "well behaved" and unfortunate consequences of lack of a guard digit are given less prominence.

A final change concerns the associated MATLAB toolbox. The Test Matrix Toolbox from the first edition is superseded by the new Matrix Computation Toolbox, described in Appendix D. The new toolbox includes functions implementing a number of algorithms in the book—in particular, GE with rook pivoting and block LDL^T factorization for symmetric and skew-symmetric matrices. The toolbox should be of use for both teaching and research.

I am grateful to Bill Gragg, Beresford Parlett, Colin Percival, Siegfried Rump, Françoise Tisseur, Nick Trefethen, and Tjalling Ypma for comments that influenced the second edition.

It has been a pleasure working once again with the SIAM publication staff, in particular Linda Thiel, Sara Triller Murphy, Marianne Will, and my copy editor, Beth Gallagher.

Research leading to this book has been supported by grants from the Engineering and Physical Sciences Research Council and by a Royal Society Leverhulme Trust Senior Research Fellowship.

The tools used to prepare the book were the same as for the first edition, except that for TeX-related tasks I used MikTeX (`http://www.miktex.org/`), including its excellent YAP previewer.

Manchester
February 2002

Nicholas J. Higham

Preface to First Edition

It has been 30 years since the publication of Wilkinson's books *Rounding Errors in Algebraic Processes* [1232, 1963] and *The Algebraic Eigenvalue Problem* [1233, 1965]. These books provided the first thorough analysis of the effects of rounding errors on numerical algorithms, and they rapidly became highly influential classics in numerical analysis. Although a number of more recent books have included analysis of rounding errors, none has treated the subject in the same depth as Wilkinson.

This book gives a thorough, up-to-date treatment of the behaviour of numerical algorithms in finite precision arithmetic. It combines algorithmic derivations, perturbation theory, and rounding error analysis. Software practicalities are emphasized throughout, with particular reference to LAPACK. The best available error bounds, some of them new, are presented in a unified format with a minimum of jargon. Historical perspective is given to provide insight into the development of the subject, and further information is provided in the many quotations. Perturbation theory is treated in detail, because of its central role in revealing problem sensitivity and providing error bounds. The book is unique in that algorithmic derivations and motivation are given succinctly, and implementation details minimized, so that attention can be concentrated on accuracy and stability results. The book was designed to be a comprehensive reference and contains extensive citations to the research literature.

Although the book's main audience is specialists in numerical analysis, it will be of use to all computational scientists and engineers who are concerned about the accuracy of their results. Much of the book can be understood with only a basic grounding in numerical analysis and linear algebra.

This first two chapters are very general. Chapter 1 describes fundamental concepts of finite precision arithmetic, giving many examples for illustration and dispelling some misconceptions. Chapter 2 gives a thorough treatment of floating point arithmetic and may well be the single most useful chapter in the book. In addition to describing models of floating point arithmetic and the IEEE standard, it explains how to exploit "low-level" features not represented in the models and contains a large set of informative exercises.

In the rest of the book the focus is, inevitably, on numerical linear algebra, because it is in this area that rounding errors are most influential and have been most extensively studied. However, I found that it was impossible to cover the whole of numerical linear algebra in a single volume. The main omission is the area of eigenvalue and singular value computations, which is still the subject of intensive research and requires a book of its own to summarize algorithms, perturbation theory, and error analysis. This book is therefore certainly not a replacement for *The Algebraic Eigenvalue Problem.*

Two reasons why rounding error analysis can be hard to understand are that, first, there is no standard notation and, second, error analyses are often cluttered with re-derivations of standard results. In this book I have used notation that I find nearly always to be the most convenient for error analysis: the key ingredient is the symbol $\gamma_n = nu/(1-nu)$, explained in §3.1. I have also summarized many basic error analysis results (for example, in Chapters 3 and 8) and made use of them throughout the book. I like to think of these basic results as analogues of the Fortran BLAS (Basic Linear Algebra Subprograms): once available in a standard form they can be used as black boxes and need not be reinvented.

A number of the topics included here have not been treated in depth in previous numerical analysis textbooks. These include floating point summation, block LU factorization, condition number estimation, the Sylvester equation, powers of matrices, finite precision behaviour of stationary iterative methods, Vandermonde systems, and fast matrix multiplication, each of which has its own chapter. But there are also some notable omissions. I would have liked to include a chapter on Toeplitz systems, but this is an area in which stability and accuracy are incompletely understood and where knowledge of the underlying applications is required to guide the investigation. The important problems of updating and downdating matrix factorizations when the matrix undergoes a "small" change have also been omitted due to lack of time and space. A further omission is analysis of parallel algorithms for all the problems considered in the book (though blocked and partitioned algorithms and one particular parallel method for triangular systems are treated). Again, there are relatively few results and this is an area of active research.

Throughout the history of numerical linear algebra, theoretical advances have gone hand in hand with software development. This tradition has continued with LAPACK (1987–), a project to develop a state-of-the-art Fortran package for solving linear equations and eigenvalue problems. LAPACK has enjoyed a synergy with research that has led to a number of important breakthroughs in the design and analysis of algorithms, from the standpoints of both performance and accuracy. A key feature of this book is that it provides the material needed to understand the numerical properties of many of the algorithms in LAPACK, the exceptions being the routines for eigenvalue and singular value problems. In particular, the error bounds computed by the LAPACK linear equation solvers are explained, the LAPACK condition estimator is described in detail, and some of the software issues confronted by the LAPACK developers are highlighted. Chapter 27 examines the influence of floating point arithmetic on general numerical software, offering salutary stories, useful techniques, and brief descriptions of relevant codes.

This book has been written with numerical analysis courses in mind, although it is not designed specifically as a textbook. It would be a suitable reference for an advanced course (for example, for a graduate course on numerical linear algebra following the syllabus recommended by the ILAS Education Committee [661, 1993]), and could be used by instructors at all levels as a supplementary text from which to draw examples, historical perspective, statements of results, and exercises. The exercises (actually labelled "problems") are an important part of the book, and many of them have not, to my knowledge, appeared in textbooks before.

Where appropriate I have indicated the source of an exercise; a name without a citation means that the exercise came from private communication or unpublished notes. Research problems given at the end of some sets of exercises emphasize that most of the areas covered are still active.

In addition to surveying and unifying existing results (including some that have not appeared in the mainstream literature) and sometimes improving upon their presentation or proof, this book contains new results. Some of particular note are as follows.

1. The error analysis in §5.3 for evaluation of the Newton interpolating polynomial.

2. The forward error analysis for iterative refinement in §12.1.

3. The error analysis of Gauss–Jordan elimination in §14.4.

4. The unified componentwise error analysis of QR factorization methods in Chapter 19, and the corresponding analysis of their use for solving the least squares problem in Chapter 20.

5. Theorem 21.4, which shows the backward stability of the QR factorization method for computing the minimum 2-norm solution to an underdetermined system.

The Notes and References are an integral part of each chapter. In addition to containing references, historical information, and further details, they include material not covered elsewhere in the chapter, and should always be consulted, in conjunction with the index, to obtain the complete picture.

I have included relatively few numerical examples except in the first chapter. There are two reasons. One is to reduce the length of the book. The second reason is because today it so easy for the reader to perform experiments in MATLAB or some other interactive system. To this end I have made available the Test Matrix Toolbox, which contains MATLAB M-files for many of the algorithms and special matrices described in the book; see Appendix D.

This book has been designed to be as easy to use as possible. There are thorough name and subject indexes, page headings show chapter and section titles and numbers, and there is extensive cross-referencing. I have adopted the unusual policy of giving with (nearly) every citation not only its numerical location in the bibliography but also the names of the authors and the year of publication. This provides as much information as possible in a citation and reduces the need for the reader to turn to the bibliography.

A BIBTEX database `acc-stab-num-alg.bib` containing all the references in the bibliography is available over the Internet from the bibnet project (which can be accessed via netlib, described in §B.2).

Special care has been taken to minimize the number of typographical and other errors, but no doubt some remain. I will be happy to receive notification of errors, as well as comments and suggestions for improvement.

Acknowledgements

Three books, in addition to Wilkinson's, have strongly influenced my research in numerical linear algebra and have provided inspiration for this book: Golub and Van Loan's *Matrix Computations* [509, 1996] (first edition 1983), Parlett's *The Symmetric Eigenvalue Problem* [926, 1998] (first published 1980) and Stewart's *Introduction to Matrix Computations* [1065, 1973]. Knuth's *The Art of Computer Programming* books (1973–) [743], [744], have also influenced my style and presentation.

Jim Demmel has contributed greatly to my understanding of the subject of this book and provided valuable technical help and suggestions. The first two chapters owe much to the work of Velvel Kahan; I am grateful to him for giving me access to unpublished notes and for suggesting improvements to early versions of Chapters 2 and 27. Des Higham read various drafts of the book, offering sound advice and finding improvements that had eluded me.

Other people who have given valuable help, suggestions, or advice are

> Zhaojun Bai, Brad Baxter, Åke Björck, Martin Campbell-Kelly, Shivkumar Chandrasekaran, Alan Edelman, Warren Ferguson, Philip Gill, Gene Golub, George Hall, Sven Hammarling, Andrzej Kiełbasiński, Philip Knight, Beresford Parlett, David Silvester, Michael Saunders, Ian Smith, Doron Swade, Nick Trefethen, Jack Williams, and Hongyuan Zha.

David Carlisle provided invaluable help and advice concerning LaTeX 2ε.

Working with SIAM on the publication of this book was a pleasure. Special thanks go to Nancy Abbott (design), Susan Ciambrano (acquisition), Ed Cilurso (production), Beth Gallagher (copy editing), Corey Gray (production), Mary Rose Muccie (copy editing and indexing), Colleen Robishaw (design), and Sam Young (production).

Research leading to this book has been supported by grants from the Engineering and Physical Sciences Research Council, by a Nuffield Science Research Fellowship from the Nuffield Foundation, and by a NATO Collaborative Research Grant held with J. W. Demmel. I was fortunate to be able to make extensive use of the libraries of the University of Manchester, the University of Dundee, Stanford University, and the University of California, Berkeley.

This book was typeset in LaTeX 2ε using the book document style. The references were prepared in BibTeX and the index with MakeIndex. It is difficult to imagine how I could have written the book without these wonderful tools. I used the "big" software from the emTeX distribution, running on a 486DX workstation. I used text editors The Semware Editor (Semware Corporation) and GNU Emacs (Free Software Foundation) and checked spelling with PC-Write (Quicksoft).

Manchester
April 1995

Nicholas J. Higham

About the Dedication

This book is dedicated to the memory of two remarkable English mathematicians, James Hardy Wilkinson (1919–1986), FRS, and Alan Mathison Turing (1912–1954), FRS, both of whom made immense contributions to scientific computation.

Turing's achievements include his paper "On Computable Numbers, with an Application to the Entscheidungsproblem", which answered Hilbert's decidability question using the abstract device now known as a Turing machine [1164, 1936]; his work at Bletchley Park during World War II on breaking the ciphers of the Enigma machine; his 1945 report proposing a design for the Automatic Computing Engine (ACE) at the National Physical Laboratory [1165, 1945]; his 1948 paper on LU factorization and its rounding error analysis [1166, 1948]; his consideration of fundamental questions in artificial intelligence (including his proposal of the "Turing test"); and, during the last part of his life, spent at the University of Manchester, his work on morphogenesis (the development of structure and form in an organism). Turing is remembered through the Turing Award of the Association for Computing Machinery (ACM), which has been awarded yearly since 1966 [3, 1987]. For more about Turing, read the superb biography by Hodges [631, 1983], described by a reviewer as "one of the finest pieces of scholarship to appear in the history of computing" [201, 1984]. Hodges maintains the Alan Turing Home Page at `http://www.turing.org.uk/turing/`

Wilkinson, like Turing a Cambridge-trained mathematician, was Turing's assistant at the National Physical Laboratory. When Turing left, Wilkinson managed the group that built the Pilot ACE, contributing to the design and construction of the machine and its software. Subsequently, he used the machine to develop and study a variety of numerical methods. He developed backward error analysis in the 1950s and 1960s, publishing the books *Rounding Errors in Algebraic Processes* [1232, 1963][‡] (REAP) and *The Algebraic Eigenvalue Problem* [1233, 1965][§] (AEP), both of which rapidly achieved the status of classics. (AEP was reprinted in paperback in 1988 and, after being out of print for many years, REAP is now also available in paperback.) The AEP was described by the late Professor Leslie Fox as "almost certainly the most important and widely read title in numerical analysis". Wilkinson also contributed greatly to the development of mathematical software. The volume *Handbook for Automatic Computation, Volume* II: *Linear Algebra* [1246, 1971], co-edited with Reinsch, contains high-quality, properly documented software and has strongly influenced subsequent software projects such as the NAG Library, EISPACK, LINPACK, and LAPACK.

Wilkinson received the 1970 Turing Award. In his Turing Award lecture he

[‡]REAP has been translated into Polish [1235, 1967] and German [1237, 1969].

[§]AEP has been translated into Russian [1238, 1970].

described life with Turing at the National Physical Laboratory in the 1940s [1240, 1971].

Wilkinson is remembered through SIAM's James H. Wilkinson Prize in Numerical Analysis and Scientific Computing, awarded every 4 years; the Wilkinson Prize for Numerical Software, awarded by Argonne National Laboratory, the National Physical Laboratory, and the Numerical Algorithms Group; and the Wilkinson Fellowship in Scientific Computing at Argonne National Laboratory. For more about Wilkinson see the biographical memoir by Fox [439, 1987], Fox's article [438, 1978], Parlett's essay [925, 1990], the prologue and epilogue of the proceedings [279, 1990] of a conference held in honour of Wilkinson at the National Physical Laboratory in 1987, and the tributes in [29, 1987]. Lists of Wilkinson's publications are given in [439, 1987] and in the special volume of the journal *Linear Algebra and Its Applications* (88/89, April 1987) published in his memory.

Contents

List of Figures xvii

List of Tables xix

Preface to Second Edition xxi

Preface to First Edition xxv

About the Dedication xxix

1 Principles of Finite Precision Computation 1

1.1 Notation and Background 2
1.2 Relative Error and Significant Digits 3
1.3 Sources of Errors 5
1.4 Precision Versus Accuracy 6
1.5 Backward and Forward Errors 6
1.6 Conditioning 8
1.7 Cancellation 9
1.8 Solving a Quadratic Equation 10
1.9 Computing the Sample Variance 11
1.10 Solving Linear Equations 12
1.10.1 GEPP Versus Cramer's Rule 13
1.11 Accumulation of Rounding Errors 14
1.12 Instability Without Cancellation 14
1.12.1 The Need for Pivoting 15
1.12.2 An Innocuous Calculation? 15
1.12.3 An Infinite Sum 16
1.13 Increasing the Precision 17
1.14 Cancellation of Rounding Errors 19
1.14.1 Computing $(e^x - 1)/x$ 19
1.14.2 QR Factorization 21
1.15 Rounding Errors Can Be Beneficial 22
1.16 Stability of an Algorithm Depends on the Problem 24
1.17 Rounding Errors Are Not Random 25
1.18 Designing Stable Algorithms 26
1.19 Misconceptions 28
1.20 Rounding Errors in Numerical Analysis 28
1.21 Notes and References 28
Problems 31

2 Floating Point Arithmetic **35**
2.1 Floating Point Number System 36
2.2 Model of Arithmetic 40
2.3 IEEE Arithmetic 41
2.4 Aberrant Arithmetics 43
2.5 Exact Subtraction 45
2.6 Fused Multiply-Add Operation 46
2.7 Choice of Base and Distribution of Numbers 47
2.8 Statistical Distribution of Rounding Errors 48
2.9 Alternative Number Systems 49
2.10 Elementary Functions 50
2.11 Accuracy Tests 51
2.12 Notes and References 52
Problems 57

3 Basics **61**
3.1 Inner and Outer Products 62
3.2 The Purpose of Rounding Error Analysis 65
3.3 Running Error Analysis 65
3.4 Notation for Error Analysis 67
3.5 Matrix Multiplication 69
3.6 Complex Arithmetic 71
3.7 Miscellany 73
3.8 Error Analysis Demystified 74
3.9 Other Approaches 76
3.10 Notes and References 76
Problems 77

4 Summation **79**
4.1 Summation Methods 80
4.2 Error Analysis 81
4.3 Compensated Summation 83
4.4 Other Summation Methods 88
4.5 Statistical Estimates of Accuracy 88
4.6 Choice of Method 89
4.7 Notes and References 90
Problems 91

5 Polynomials **93**
5.1 Horner's Method 94
5.2 Evaluating Derivatives 96
5.3 The Newton Form and Polynomial Interpolation 99
5.4 Matrix Polynomials 102
5.5 Notes and References 102
Problems 104