
Nicholas J. Higham

University of Manchester
Manchester, England

Accuracy and Stability of Numerical Algorithms

SECOND EDITION

siam

Society for Industrial and Applied Mathematics
Philadelphia

Copyright © 2002 by the Society for Industrial and Applied Mathematics.

10 9 8 7 6 5 4 3 2

All rights reserved. Printed in the United States of America. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, write to the Society for Industrial and Applied Mathematics, 3600 University City Science Center, Philadelphia, PA 19104-2688.

Library of Congress Cataloging-in-Publication Data

Higham, Nicholas J., 1961-

Accuracy and stability of numerical algorithms / Nicholas J. Higham.—
2nd ed.

p. cm.

Includes bibliographical references.

ISBN-10: 0-89871-521-0

ISBN-13: 978-0-898715-21-7

1. Numerical analysis—Data processing. 2. Computer algorithms. I.
Title.

QA297 .H53 2002

519.4'0285'51—dc21

2002075848

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

6	Norms	105
6.1	Vector Norms	106
6.2	Matrix Norms	107
6.3	The Matrix p -Norm	112
6.4	Singular Value Decomposition	114
6.5	Notes and References	114
	Problems	115
7	Perturbation Theory for Linear Systems	119
7.1	Normwise Analysis	120
7.2	Componentwise Analysis	122
7.3	Scaling to Minimize the Condition Number	125
7.4	The Matrix Inverse	127
7.5	Extensions	128
7.6	Numerical Stability	129
7.7	Practical Error Bounds	130
7.8	Perturbation Theory by Calculus	132
7.9	Notes and References	132
	Problems	134
8	Triangular Systems	139
8.1	Backward Error Analysis	140
8.2	Forward Error Analysis	142
8.3	Bounds for the Inverse	147
8.4	A Parallel Fan-In Algorithm	149
8.5	Notes and References	151
	8.5.1 LAPACK	153
	Problems	153
9	LU Factorization and Linear Equations	157
9.1	Gaussian Elimination and Pivoting Strategies	158
9.2	LU Factorization	160
9.3	Error Analysis	163
9.4	The Growth Factor	166
9.5	Diagonally Dominant and Banded Matrices	170
9.6	Tridiagonal Matrices	174
9.7	More Error Bounds	176
9.8	Scaling and Choice of Pivoting Strategy	177
9.9	Variants of Gaussian Elimination	179
9.10	A Posteriori Stability Tests	180
9.11	Sensitivity of the LU Factorization	181
9.12	Rank-Revealing LU Factorizations	182
9.13	Historical Perspective	183
9.14	Notes and References	187
	9.14.1 LAPACK	191
	Problems	192

10 Cholesky Factorization	195
10.1 Symmetric Positive Definite Matrices	196
10.1.1 Error Analysis	197
10.2 Sensitivity of the Cholesky Factorization	201
10.3 Positive Semidefinite Matrices	201
10.3.1 Perturbation Theory	203
10.3.2 Error Analysis	205
10.4 Matrices with Positive Definite Symmetric Part	208
10.5 Notes and References	209
10.5.1 LAPACK	210
Problems	211
11 Symmetric Indefinite and Skew-Symmetric Systems	213
11.1 Block LDL ^T Factorization for Symmetric Matrices	214
11.1.1 Complete Pivoting	215
11.1.2 Partial Pivoting	216
11.1.3 Rook Pivoting	219
11.1.4 Tridiagonal Matrices	221
11.2 Aasen's Method	222
11.2.1 Aasen's Method Versus Block LDL ^T Factorization	224
11.3 Block LDL ^T Factorization for Skew-Symmetric Matrices	225
11.4 Notes and References	226
11.4.1 LAPACK	228
Problems	228
12 Iterative Refinement	231
12.1 Behaviour of the Forward Error	232
12.2 Iterative Refinement Implies Stability	235
12.3 Notes and References	240
12.3.1 LAPACK	242
Problems	242
13 Block LU Factorization	245
13.1 Block Versus Partitioned LU Factorization	246
13.2 Error Analysis of Partitioned LU Factorization	249
13.3 Error Analysis of Block LU Factorization	250
13.3.1 Block Diagonal Dominance	251
13.3.2 Symmetric Positive Definite Matrices	255
13.4 Notes and References	256
13.4.1 LAPACK	257
Problems	257
14 Matrix Inversion	259
14.1 Use and Abuse of the Matrix Inverse	260
14.2 Inverting a Triangular Matrix	262
14.2.1 Unblocked Methods	262
14.2.2 Block Methods	265
14.3 Inverting a Full Matrix by LU Factorization	267

14.3.1	Method A	267
14.3.2	Method B	268
14.3.3	Method C	269
14.3.4	Method D	270
14.3.5	Summary	271
14.4	Gauss–Jordan Elimination	273
14.5	Parallel Inversion Methods	278
14.6	The Determinant	279
14.6.1	Hyman’s Method	280
14.7	Notes and References	281
14.7.1	LAPACK	282
	Problems	283
15	Condition Number Estimation	287
15.1	How to Estimate Componentwise Condition Numbers	288
15.2	The p -Norm Power Method	289
15.3	LAPACK 1-Norm Estimator	292
15.4	Block 1-Norm Estimator	294
15.5	Other Condition Estimators	295
15.6	Condition Numbers of Tridiagonal Matrices	299
15.7	Notes and References	301
15.7.1	LAPACK	303
	Problems	303
16	The Sylvester Equation	305
16.1	Solving the Sylvester Equation	307
16.2	Backward Error	308
16.2.1	The Lyapunov Equation	311
16.3	Perturbation Result	313
16.4	Practical Error Bounds	315
16.5	Extensions	316
16.6	Notes and References	317
16.6.1	LAPACK	318
	Problems	318
17	Stationary Iterative Methods	321
17.1	Survey of Error Analysis	323
17.2	Forward Error Analysis	325
17.2.1	Jacobi’s Method	328
17.2.2	Successive Overrelaxation	329
17.3	Backward Error Analysis	330
17.4	Singular Systems	331
17.4.1	Theoretical Background	331
17.4.2	Forward Error Analysis	333
17.5	Stopping an Iterative Method	335
17.6	Notes and References	337
	Problems	337

18 Matrix Powers	339
18.1 Matrix Powers in Exact Arithmetic	340
18.2 Bounds for Finite Precision Arithmetic	346
18.3 Application to Stationary Iteration	351
18.4 Notes and References	351
Problems	352
19 QR Factorization	353
19.1 Householder Transformations	354
19.2 QR Factorization	355
19.3 Error Analysis of Householder Computations	357
19.4 Pivoting and Row-Wise Stability	362
19.5 Aggregated Householder Transformations	363
19.6 Givens Rotations	365
19.7 Iterative Refinement	368
19.8 Gram–Schmidt Orthogonalization	369
19.9 Sensitivity of the QR Factorization	373
19.10 Notes and References	374
19.10.1 LAPACK	377
Problems	378
20 The Least Squares Problem	381
20.1 Perturbation Theory	382
20.2 Solution by QR Factorization	384
20.3 Solution by the Modified Gram–Schmidt Method	386
20.4 The Normal Equations	386
20.5 Iterative Refinement	388
20.6 The Seminormal Equations	391
20.7 Backward Error	392
20.8 Weighted Least Squares Problems	395
20.9 The Equality Constrained Least Squares Problem	396
20.9.1 Perturbation Theory	396
20.9.2 Methods	397
20.10 Proof of Wedin’s Theorem	400
20.11 Notes and References	402
20.11.1 LAPACK	405
Problems	405
21 Underdetermined Systems	407
21.1 Solution Methods	408
21.2 Perturbation Theory and Backward Error	409
21.3 Error Analysis	411
21.4 Notes and References	413
21.4.1 LAPACK	414
Problems	414

22 Vandermonde Systems	415
22.1 Matrix Inversion	416
22.2 Primal and Dual Systems	418
22.3 Stability	423
22.3.1 Forward Error	424
22.3.2 Residual	425
22.3.3 Dealing with Instability	426
22.4 Notes and References	428
Problems	430
23 Fast Matrix Multiplication	433
23.1 Methods	434
23.2 Error Analysis	438
23.2.1 Winograd's Method	439
23.2.2 Strassen's Method	440
23.2.3 Bilinear Noncommutative Algorithms	443
23.2.4 The 3M Method	444
23.3 Notes and References	446
Problems	448
24 The Fast Fourier Transform and Applications	451
24.1 The Fast Fourier Transform	452
24.2 Circulant Linear Systems	454
24.3 Notes and References	456
Problems	457
25 Nonlinear Systems and Newton's Method	459
25.1 Newton's Method	460
25.2 Error Analysis	461
25.3 Special Cases and Experiments	462
25.4 Conditioning	464
25.5 Stopping an Iterative Method	467
25.6 Notes and References	468
Problems	469
26 Automatic Error Analysis	471
26.1 Exploiting Direct Search Optimization	472
26.2 Direct Search Methods	474
26.3 Examples of Direct Search	477
26.3.1 Condition Estimation	477
26.3.2 Fast Matrix Inversion	478
26.3.3 Roots of a Cubic	479
26.4 Interval Analysis	481
26.5 Other Work	484
26.6 Notes and References	486
Problems	487

27 Software Issues in Floating Point Arithmetic	489
27.1 Exploiting IEEE Arithmetic	490
27.2 Subtleties of Floating Point Arithmetic	493
27.3 Cray Peculiarities	493
27.4 Compilers	494
27.5 Determining Properties of Floating Point Arithmetic	494
27.6 Testing a Floating Point Arithmetic	495
27.7 Portability	496
27.7.1 Arithmetic Parameters	496
27.7.2 2×2 Problems in LAPACK	497
27.7.3 Numerical Constants	498
27.7.4 Models of Floating Point Arithmetic	498
27.8 Avoiding Underflow and Overflow	499
27.9 Multiple Precision Arithmetic	501
27.10 Extended and Mixed Precision BLAS	503
27.11 Patriot Missile Software Problem	503
27.12 Notes and References	504
Problems	505
28 A Gallery of Test Matrices	511
28.1 The Hilbert and Cauchy Matrices	512
28.2 Random Matrices	515
28.3 “Randsvd” Matrices	517
28.4 The Pascal Matrix	518
28.5 Tridiagonal Toeplitz Matrices	521
28.6 Companion Matrices	522
28.7 Notes and References	523
28.7.1 LAPACK	525
Problems	525
A Solutions to Problems	527
B Acquiring Software	573
B.1 Internet	574
B.2 Netlib	574
B.3 MATLAB	575
B.4 NAG Library and NAGWare F95 Compiler	575
C Program Libraries	577
C.1 Basic Linear Algebra Subprograms	578
C.2 EISPACK	579
C.3 LINPACK	579
C.4 LAPACK	579
C.4.1 Structure of LAPACK	580
D The Matrix Computation Toolbox	583
Bibliography	587

CONTENTS

xv

Name Index

657

Subject Index

667