

# Linear Algebra: Foundations to Frontiers

## A Collection of Notes on Numerical Linear Algebra

Robert A. van de Geijn

Release Date March 28, 2015

**Kindly do not share this PDF**

Point others to  <http://www.ulaff.net> instead

This is a work in progress

Copyright © 2014 by Robert A. van de Geijn.

10 9 8 7 6 5 4 3 2 1

All rights reserved. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, contact any of the authors.

No warranties, express or implied, are made by the publisher, authors, and their employers that the programs contained in this volume are free of error. They should not be relied on as the sole basis to solve a problem whose incorrect solution could result in injury to person or property. If the programs are employed in such a manner, it is at the user's own risk and the publisher, authors, and their employers disclaim all liability for such misuse.

Trademarked names may be used in this book without the inclusion of a trademark symbol. These names are used in an editorial context only; no infringement of trademark is intended.

**Library of Congress Cataloging-in-Publication Data** not yet available

Draft Edition, November 2014

This "Draft Edition" allows this material to be used while we sort out through what mechanism we will publish the book.

# Contents

<b>Preface</b>	<b>ix</b>
Video	xi
<b>1. Notes on Simple Vector and Matrix Operations</b>	<b>1</b>
Video	1
Outline	2
1.1. Notation	3
1.2. (Hermitian) Transposition	5
1.2.1. Conjugating a complex scalar	5
1.2.2. Conjugate of a vector	5
1.2.3. Conjugate of a matrix	5
1.2.4. Transpose of a vector	6
1.2.5. Hermitian transpose of a vector	6
1.2.6. Transpose of a matrix	6
1.2.7. Hermitian transpose (adjoint) of a matrix	6
1.2.8. Exercises	7
1.3. Vector-vector Operations	9
1.3.1. Scaling a vector (scal)	9
1.3.2. Scaled vector addition (axpy)	10
1.3.3. Dot (inner) product (dot)	11
1.4. Matrix-vector Operations	12
1.4.1. Matrix-vector multiplication (product)	12
1.4.2. Rank-1 update	15
1.5. Matrix-matrix multiplication (product)	18
1.5.1. Element-by-element computation	18
1.5.2. Via matrix-vector multiplications	19
1.5.3. Via row-vector times matrix multiplications	19
1.5.4. Via rank-1 updates	20
1.5.5. Cost	20
1.6. Summarizing All	20

<b>2. Notes on Vector and Matrix Norms</b>	<b>23</b>
Video	23
Outline	24
2.1. Absolute Value	25
2.2. Vector Norms	25
2.2.1. Vector 2-norm (length)	25
2.2.2. Vector 1-norm	27
2.2.3. Vector $\infty$ -norm (infinity norm)	27
2.2.4. Vector $p$ -norm	27
2.3. Matrix Norms	27
2.3.1. Frobenius norm	27
2.3.2. Induced matrix norms	28
2.3.3. Special cases used in practice (matrix $p$ -norm)	29
2.3.4. Discussion	31
2.3.5. Submultiplicative matrix norms	31
2.4. An Application to Conditioning of Linear Systems	32
2.5. Equivalence of Norms	33
<b>3. Notes on Orthogonality and the Singular Value Decomposition</b>	<b>35</b>
Video	35
Outline	36
3.1. Orthogonality and Unitary Matrices	37
3.2. Toward the Singular Value Decomposition	39
3.3. The Singular Value Decomposition Theorem	41
3.4. Geometric Interpretation	41
3.5. Consequences of the SVD Theorem	45
3.6. Projection onto the Column Space	49
3.7. Low-rank Approximation of a Matrix	50
3.8. An Application	51
3.9. SVD and the Condition Number of a Matrix	54
3.10. An Algorithm for Computing the SVD?	55
<b>4. Notes on Gram-Schmidt QR Factorization</b>	<b>57</b>
Video	57
Outline	58
4.1. Classical Gram-Schmidt (CGS) Process	59
4.2. Modified Gram-Schmidt (MGS) Process	64
4.3. In Practice, MGS is More Accurate	68
4.4. Cost	70
4.4.1. Cost of CGS	71
4.4.2. Cost of MGS	72
<b>5. Notes on the FLAME APIs</b>	<b>73</b>
Video	73
Outline	74
5.1. Motivation	75

5.2.	Install FLAME@lab	75
5.3.	An Example: Gram-Schmidt Orthogonalization	75
5.3.1.	The Spark Webpage	75
5.3.2.	Implementing CGS with FLAME@lab	76
5.3.3.	Editing the code skeleton	78
5.3.4.	Testing	79
5.4.	Implementing the Other Algorithms	80
<b>6.</b>	<b>Notes on Householder QR Factorization</b>	<b>83</b>
	Video	83
	Outline	84
6.1.	Motivation	85
6.2.	Householder Transformations (Reflectors)	85
6.2.1.	The general case	85
6.2.2.	As implemented for the Householder QR factorization (real case)	87
6.2.3.	The complex case (optional)	87
6.2.4.	A routine for computing the Householder vector	88
6.3.	Householder QR Factorization	89
6.4.	Forming $Q$	92
6.5.	Applying $Q^H$	97
6.6.	Blocked Householder QR Factorization	99
6.6.1.	The UT transform: Accumulating Householder transformations	99
6.6.2.	A blocked algorithm	102
6.6.3.	Variations on a theme	104
<b>7.</b>	<b>Notes on Solving Linear Least-Squares Problems</b>	<b>109</b>
	Video	109
	Outline	110
7.1.	The Linear Least-Squares Problem	111
7.2.	Method of Normal Equations	111
7.3.	Solving the LLS Problem Via the QR Factorization	112
7.3.1.	Simple derivation of the solution	112
7.3.2.	Alternative derivation of the solution	113
7.4.	Via Householder QR Factorization	114
7.5.	Via the Singular Value Decomposition	115
7.5.1.	Simple derivation of the solution	115
7.5.2.	Alternative derivation of the solution	116
7.6.	What If $A$ Does Not Have Linearly Independent Columns?	116
7.7.	Exercise: Using the the $LQ$ factorization to solve underdetermined systems	123
<b>8.</b>	<b>Notes on the Condition of a Problem</b>	<b>127</b>
	Video	127
	Outline	128
8.1.	Notation	129
8.2.	The Prototypical Example: Solving a Linear System	129
8.3.	Condition Number of a Rectangular Matrix	133

8.4. Why Using the Method of Normal Equations Could be Bad . . . . .	134
8.5. Why Multiplication with Unitary Matrices is a Good Thing . . . . .	135
<b>9. Notes on the Stability of an Algorithm</b>	<b>137</b>
Video . . . . .	137
Outline . . . . .	138
9.1. Motivation . . . . .	139
9.2. Floating Point Numbers . . . . .	140
9.3. Notation . . . . .	142
9.4. Floating Point Computation . . . . .	142
9.4.1. Model of floating point computation . . . . .	142
9.4.2. Stability of a numerical algorithm . . . . .	143
9.4.3. Absolute value of vectors and matrices . . . . .	143
9.5. Stability of the Dot Product Operation . . . . .	144
9.5.1. An algorithm for computing DOT . . . . .	144
9.5.2. A simple start . . . . .	144
9.5.3. Preparation . . . . .	146
9.5.4. Target result . . . . .	148
9.5.5. A proof in traditional format . . . . .	149
9.5.6. A weapon of math induction for the war on error (optional) . . . . .	149
9.5.7. Results . . . . .	152
9.6. Stability of a Matrix-Vector Multiplication Algorithm . . . . .	152
9.6.1. An algorithm for computing GEMV . . . . .	152
9.6.2. Analysis . . . . .	152
9.7. Stability of a Matrix-Matrix Multiplication Algorithm . . . . .	154
9.7.1. An algorithm for computing GEMM . . . . .	154
9.7.2. Analysis . . . . .	154
9.7.3. An application . . . . .	155
<b>10. Notes on Performance</b>	<b>157</b>
<b>11. Notes on Gaussian Elimination and LU Factorization</b>	<b>159</b>
Video . . . . .	159
Outline . . . . .	160
11.1. Definition and Existence . . . . .	161
11.2. LU Factorization . . . . .	161
11.2.1. First derivation . . . . .	161
11.2.2. Gauss transforms . . . . .	162
11.2.3. Cost of LU factorization . . . . .	164
11.3. LU Factorization with Partial Pivoting . . . . .	165
11.3.1. Permutation matrices . . . . .	165
11.3.2. The algorithm . . . . .	167
11.4. Proof of Theorem 11.3 . . . . .	173
11.5. LU with Complete Pivoting . . . . .	174
11.6. Solving $Ax = y$ Via the LU Factorization with Pivoting . . . . .	175
11.7. Solving Triangular Systems of Equations . . . . .	175

11.7.1. $Lz = y$	175
11.7.2. $Ux = z$	178
11.8. Other LU Factorization Algorithms	178
11.8.1. Variant 1: Bordered algorithm	183
11.8.2. Variant 2: Left-looking algorithm	183
11.8.3. Variant 3: Up-looking variant	184
11.8.4. Variant 4: Crout variant	184
11.8.5. Variant 5: Classical LU factorization	185
11.8.6. All algorithms	185
11.8.7. Formal derivation of algorithms	185
11.9. Numerical Stability Results	187
11.10. Is LU with Partial Pivoting Stable?	188
11.11. Blocked Algorithms	188
11.11.1. Blocked classical LU factorization (Variant 5)	188
11.11.2. Blocked classical LU factorization with pivoting (Variant 5)	191
11.12. Variations on a Triple-Nested Loop	192

## 12. Notes on Cholesky Factorization 195

Video	195
Outline	196
12.1. Definition and Existence	197
12.2. Application	197
12.3. An Algorithm	198
12.4. Proof of the Cholesky Factorization Theorem	199
12.5. Blocked Algorithm	200
12.6. Alternative Representation	200
12.7. Cost	203
12.8. Solving the Linear Least-Squares Problem via the Cholesky Factorization	204
12.9. Other Cholesky Factorization Algorithms	204
12.10. Implementing the Cholesky Factorization with the (Traditional) BLAS	206
12.10.1. What are the BLAS?	206
12.10.2. A simple implementation in Fortran	209
12.10.3. Implementation with calls to level-1 BLAS	209
12.10.4. Matrix-vector operations (level-2 BLAS)	209
12.10.5. Matrix-matrix operations (level-3 BLAS)	213
12.10.6. Impact on performance	213
12.11. Alternatives to the BLAS	214
12.11.1. The FLAME/C API	214
12.11.2. BLIS	214

## 13. Notes on Eigenvalues and Eigenvectors 215

Video	215
Outline	216
13.1. Definition	217
13.2. The Schur and Spectral Factorizations	220
13.3. Relation Between the SVD and the Spectral Decomposition	222

<b>14. Notes on the Power Method and Related Methods</b>	<b>223</b>
Video	223
Outline	224
14.1. The Power Method	225
14.1.1. First attempt	225
14.1.2. Second attempt	226
14.1.3. Convergence	227
14.1.4. Practical Power Method	230
14.1.5. The Rayleigh quotient	230
14.1.6. What if $ \lambda_0  \geq  \lambda_1 $ ?	231
14.2. The Inverse Power Method	231
14.3. Rayleigh-quotient Iteration	232
<b>15. Notes on the QR Algorithm and other Dense Eigensolvers</b>	<b>235</b>
Video	235
Outline	236
15.1. Preliminaries	237
15.2. Subspace Iteration	237
15.3. The QR Algorithm	242
15.3.1. A basic (unshifted) QR algorithm	242
15.3.2. A basic shifted QR algorithm	242
15.4. Reduction to Tridiagonal Form	244
15.4.1. Householder transformations (reflectors)	244
15.4.2. Algorithm	245
15.5. The QR algorithm with a Tridiagonal Matrix	247
15.5.1. Givens' rotations	247
15.6. QR Factorization of a Tridiagonal Matrix	248
15.7. The Implicitly Shifted QR Algorithm	250
15.7.1. Upper Hessenberg and tridiagonal matrices	250
15.7.2. The Implicit Q Theorem	251
15.7.3. The Francis QR Step	252
15.7.4. A complete algorithm	254
15.8. Further Reading	258
15.8.1. More on reduction to tridiagonal form	258
15.8.2. Optimizing the tridiagonal QR algorithm	258
15.9. Other Algorithms	258
15.9.1. Jacobi's method for the symmetric eigenvalue problem	258
15.9.2. Cuppen's Algorithm	261
15.9.3. The Method of Multiple Relatively Robust Representations (MRRR)	261
15.10. The Nonsymmetric QR Algorithm	261
15.10.1. A variant of the Schur decomposition	261
15.10.2. Reduction to upper Hessenberg form	262
15.10.3. The implicitly double-shifted QR algorithm	265



<b>16. Notes on the Method of Relatively Robust Representations (MRRR)</b>	<b>267</b>
Outline	268
16.1. MRRR, from 35,000 Feet	269
16.2. Cholesky Factorization, Again	269
16.3. The $LDL^T$ Factorization	272
16.4. The $UDU^T$ Factorization	274
16.5. The $UDU^T$ Factorization	274
16.6. The Twisted Factorization	278
16.7. Computing an Eigenvector from the Twisted Factorization	279
<b>17. Notes on Computing the SVD of a Matrix</b>	<b>281</b>
Outline	282
17.1. Background	283
17.2. Reduction to Bidiagonal Form	283
17.3. The QR Algorithm with a Bidiagonal Matrix	287
17.4. Putting it all together	290
<b>18. Notes on Splitting Methods</b>	<b>293</b>
Video	293
Outline	294
18.1. A Simple Example: One-Dimensional Boundary Value Problem	295
18.2. A Two-dimensional Example	296
18.2.1. Discretization	296
18.3. Direct solution	298
18.4. Iterative solution: Jacobi iteration	298
18.4.1. Motivating example	299
18.4.2. More generally	300
18.5. The general case	301
18.6. Theory	307
18.6.1. Some useful facts	307
<b>19. Notes on Descent Methods and the Conjugate Gradient Method</b>	<b>313</b>
Outline	314
19.1. Basics	315
19.2. Descent Methods	315
19.3. Relation to Splitting Methods	316
19.4. Method of Steepest Descent	317
19.5. Preconditioning	318
19.6. Methods of A-conjugate Directions	318
19.7. Conjugate Gradient Method	320
<b>20. Notes on Lanczos Methods</b>	<b>323</b>
Outline	324
20.1. Krylov Subspaces	325
20.2. The Lanczos Method	325

<b>Answers</b>	<b>333</b>
1. Notes on Simple Vector and Matrix Operations . . . . .	333
2. Notes on Vector and Matrix Norms . . . . .	337
3. Notes on Orthogonality and the SVD . . . . .	342
4. Notes on Gram-Schmidt QR Factorization . . . . .	347
6. Notes on Householder QR Factorization . . . . .	349
7. Notes on Solving Linear Least-squares Problems . . . . .	352
8. Notes on the Condition of a Problem . . . . .	354
9. Notes on the Stability of an Algorithm . . . . .	356
10. Notes on Performance . . . . .	361
11. Notes on Gaussian Elimination and LU Factorization . . . . .	362
12. Notes on Cholesky Factorization . . . . .	369
13. Notes on Eigenvalues and Eigenvectors . . . . .	373
14. Notes on the Power Method and Related Methods . . . . .	378
15. Notes on the Symmetric QR Algorithm . . . . .	379
16. Notes on the Method of Relatively Robust Representations . . . . .	383
17. Notes on Computing the SVD . . . . .	393
18. Notes on Splitting Methods . . . . .	394
 <b>A. How to Download</b>	 <b>395</b>
 <b>Bibliography</b>	 <b>397</b>
 <b>Index</b>	 <b>401</b>

# Preface

This document was created over the course of many years, as I periodically taught an introductory course titled “Numerical Analysis: Linear Algebra,” cross-listed in the departments of Computer Science, Math, and Statistics and Data Sciences (SDS), as well as the Computational Science Engineering Mathematics (CSEM) graduate program.

Over the years, my colleagues and I have used many different books for this course: *Matrix Computations* by Golub and Van Loan [21], *Fundamentals of Matrix Computation* by Watkins [45], *Numerical Linear Algebra* by Trefethen and Bau [36], and *Applied Numerical Linear Algebra* by Demmel [11]. All are books with tremendous strengths and depth. Nonetheless, I found myself writing materials for my students that add insights that are often drawn from our own research experiences in the field. These became a series of notes that are meant to supplement rather than replace any of the mentioned books.

Fundamental to our exposition is the *FLAME notation* [24, 38], which we use to present algorithms hand-in-hand with theory. For example, in Figure 1 (left), we present a commonly encountered LU factorization algorithm, which we will see performs exactly the same computations as does Gaussian elimination. By abstracting away from the detailed indexing that are required to implement an algorithm in, for example, Matlab’s M-script language, the reader can focus on the mathematics that justifies the algorithm rather than the indices used to express the algorithm. The algorithms can be easily translated into code with the help of a FLAME Application Programming Interface (API). Such interfaces have been created for C, M-script, and Python, to name a few [24, 5, 29]. In Figure 1 (right), we show the LU factorization algorithm implemented with the FLAME@lab API for M-script. The C API is used extensively in our implementation of the `libflame` dense linear algebra library [39, 40] for sequential and shared-memory architectures and the notation also inspired the API used to implement the Elemental dense linear algebra library [30] that targets distributed memory architectures. Thus, the reader is exposed to the abstractions that experts use to translate algorithms to high-performance software.

These notes may at times appear to be a vanity project, since I often point the reader to our research papers for a glimpse at the cutting edge of the field. The fact is that over the last two decades, we have helped further the understanding of many of the topics discussed in a typical introductory course on numerical linear algebra. Since we use the FLAME notation in many of our papers, they should be relatively easy to read once one familiarizes oneself with these notes. Let’s be blunt: these notes do not do the field justice when it comes to also giving a historic perspective. For that, we recommend any of the above mentioned texts, or the wonderful books by G.W. Stewart [34, 35]. This is yet another reason why they should be used to supplement other texts.

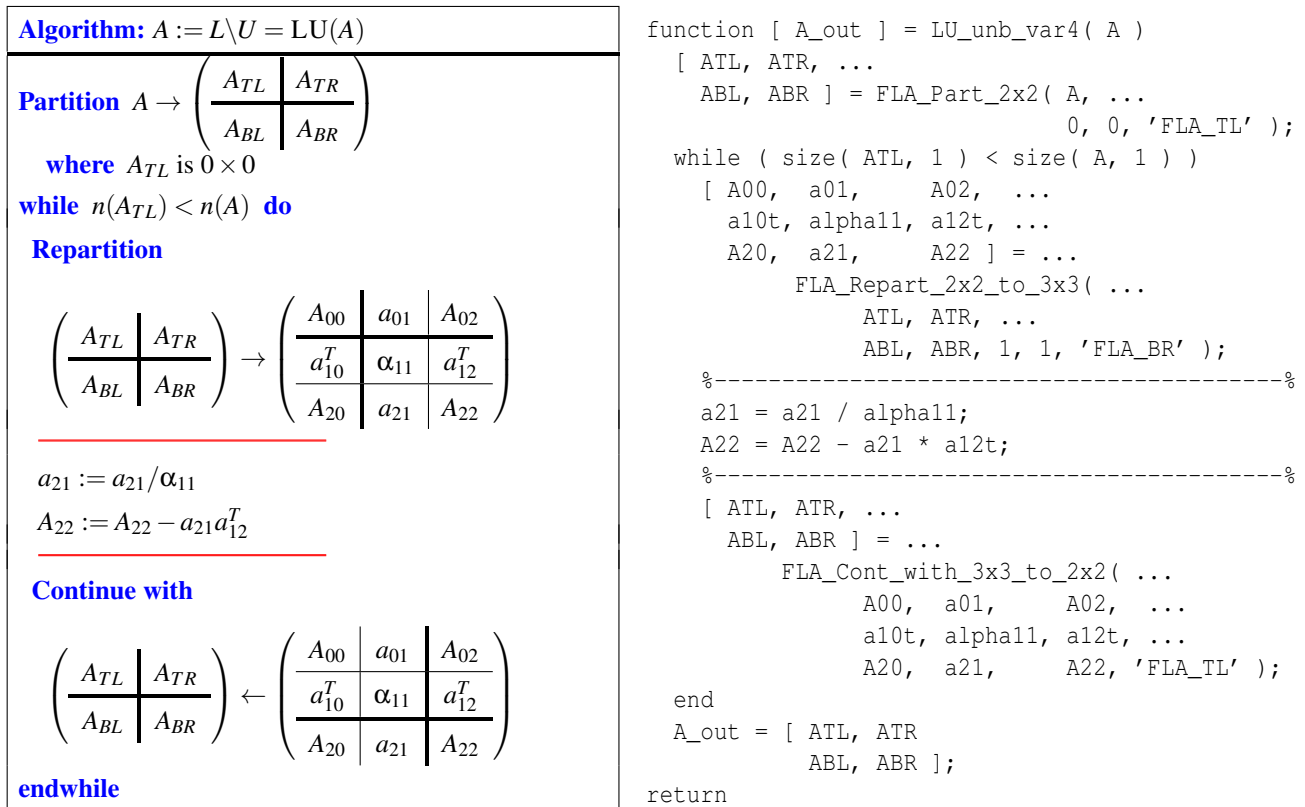


Figure 1: LU factorization represented with the FLAME notation and the FLAME@lab API.

The order of the chapters should not be taken too seriously. They were initially written as separate, relatively self-contained notes. The sequence on which I settled roughly follows the order that the topics are encountered in *Numerical Linear Algebra* by Trefethen and Bau [36]. The reason is the same as the one they give: It introduces orthogonality and the Singular Value Decomposition early on, leading with important material that many students will not yet have seen in the undergraduate linear algebra classes they took. However, one could just as easily rearrange the chapters so that one starts with a more traditional topic: solving dense linear systems.

The notes frequently refer the reader to another resource of ours titled *Linear Algebra: Foundations to Frontiers - Notes to LAFF With* (LAFF Notes) [29]. This is a 900+ page document with more than 270 videos that was created for the Massive Open Online Course (MOOC) *Linear Algebra: Foundations to Frontiers* (LAFF), offered by the [edX platform](#). That course provides an appropriate undergraduate background for these notes.

I (tried to) video tape my lectures during Fall 2014. Unlike the many short videos that we created for the Massive Open Online Course (MOOC) titled “Linear Algebra: Foundations to Frontiers” that are now part of the notes for that course, I simply set up a camera, taped the entire lecture, spent minimal time editing, and uploaded the result for the world to see. Worse, I did not prepare particularly well for the lectures, other than feverishly writing these notes in the days prior to the presentation. Sometimes, I forgot to turn on the microphone and/or the camera. Sometimes the memory of the camcorder was full. Sometimes I

forgot to shave. Often I forgot to comb my hair. You are welcome to watch, but don't expect too much!

One should consider this a living document. As time goes on, I will be adding more material. Ideally, people with more expertise than I have on, for example, solving sparse linear systems will contribute notes of their own.

## Video

Here is the video for the first lecture, which is meant as an introduction.

 [YouTube](#)

 [Download from UT Box](#)

 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

(In the downloadable version, the lecture starts about 27 seconds into the video. I was still learning how to do the editing...)



# Acknowledgments

These notes use notation and tools developed by the FLAME project at The University of Texas at Austin (USA), Universidad Jaume I (Spain), and RWTH Aachen University (Germany). This project involves a large and ever expanding number of very talented people, to whom I am indebted. Over the years, it has been supported by a number of grants from the National Science Foundation and industry. The most recent and most relevant funding came from NSF Award ACI-1148125 titled “SI2-SSI: A Linear Algebra Software Infrastructure for Sustained Innovation in Computational Chemistry and other Sciences”<sup>1</sup>

In Texas, behind every successful man there is a woman who really pulls the strings. For more than thirty years, my research, teaching, and other pedagogical activities have been greatly influenced by my wife, Dr. Maggie Myers. For parts of these notes that are particularly successful, the credit goes to her. Where they fall short, the blame is all mine!

---

<sup>1</sup> Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).





# Notes on Simple Vector and Matrix Operations

We assume that the reader is quite familiar with vectors, linear transformations, and matrices. If not, we suggest the reader reviews the first five weeks of

[Linear Algebra: Foundations to Frontiers - Notes to LAFF With \[29\]](#).

Undergraduate courses tend to focus on real valued matrices and vectors. In this note, we mostly focus on the case where they are complex valued.

## Video

The video for this particular note didn't turn out, so you will want to read instead. As I pointed out in the preface: these videos aren't refined by any measure!

## Outline

<b>Video</b>	<b>1</b>
<b>Outline</b>	<b>2</b>
<b>1.1. Notation</b>	<b>3</b>
<b>1.2. (Hermitian) Transposition</b>	<b>5</b>
1.2.1. Conjugating a complex scalar	5
1.2.2. Conjugate of a vector	5
1.2.3. Conjugate of a matrix	5
1.2.4. Transpose of a vector	6
1.2.5. Hermitian transpose of a vector	6
1.2.6. Transpose of a matrix	6
1.2.7. Hermitian transpose (adjoint) of a matrix	6
1.2.8. Exercises	7
<b>1.3. Vector-vector Operations</b>	<b>9</b>
1.3.1. Scaling a vector ( <code>scal</code> )	9
1.3.2. Scaled vector addition ( <code>axpy</code> )	10
1.3.3. Dot (inner) product ( <code>dot</code> )	11
<b>1.4. Matrix-vector Operations</b>	<b>12</b>
1.4.1. Matrix-vector multiplication ( <code>product</code> )	12
1.4.2. Rank-1 update	15
<b>1.5. Matrix-matrix multiplication (product)</b>	<b>18</b>
1.5.1. Element-by-element computation	18
1.5.2. Via matrix-vector multiplications	19
1.5.3. Via row-vector times matrix multiplications	19
1.5.4. Via rank-1 updates	20
1.5.5. Cost	20
<b>1.6. Summarizing All</b>	<b>20</b>

## 1.1 Notation

Throughout our notes we will adopt notation popularized by Alston Householder, and that is therefore sometimes called *Householder notation*. As a rule, we will use lower case Greek letters ( $\alpha$ ,  $\beta$ , etc.) to denote scalars. For vectors, we will use lower case Roman letters ( $a$ ,  $b$ , etc.). Matrices are denoted by upper case Roman letters ( $A$ ,  $B$ , etc.). A table of how symbols are often used in these notes is given in Figure 1.1.

The set of all real numbers will be denoted by  $\mathbb{R}$  and the set of all complex numbers by  $\mathbb{C}$ . The set of all vectors of size  $n$  is denoted by  $\mathbb{R}^n$  or  $\mathbb{C}^n$ , depending on whether its elements are real or complex valued. The set of all  $m \times n$  matrices is denoted by  $\mathbb{R}^{m \times n}$  or  $\mathbb{C}^{m \times n}$ .

If  $x \in \mathbb{C}^n$ , and  $A \in \mathbb{C}^{m \times n}$  then the elements of  $x$  and  $A$  can be exposed as

$$x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix}.$$

Notice how the symbols used to represent the elements in  $x$  and  $A$  are the Greek letters that (roughly) correspond to the Roman letters used to denote the vector and matrix.

If vectors  $x$  is partitioned into  $N$  subvectors, we may denote this by

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} \quad \text{or} \quad x = \begin{pmatrix} \overline{x_0} \\ \overline{x_1} \\ \vdots \\ \overline{x_{N-1}} \end{pmatrix},$$

where the horizontal lines should not be mistaken for division. They instead emphasize how the vector is partitioned into subvectors. It is possible for a subvector to be of size zero (no elements) or one (a scalar). If we want to emphasize that a specific subvector is a scalar, then we may choose to use a lower case Greek letter for that scalar, as in

$$x = \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix} \quad \text{or} \quad x = \begin{pmatrix} \overline{x_0} \\ \overline{\chi_1} \\ \overline{hlinex_2} \end{pmatrix}.$$

We will see frequent examples where a matrix,  $A \in \mathbb{C}^{m \times n}$ , is partitioned into columns or rows:

$$A = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) = \begin{pmatrix} \overline{\hat{a}_0^T} \\ \overline{\hat{a}_1^T} \\ \vdots \\ \overline{\hat{a}_{m-1}^T} \end{pmatrix}.$$

The horizontal and/or vertical lines are optional. We add the  $\hat{\phantom{x}}$  to be able to distinguish between the identifiers for the columns and rows. When from context it is obvious whether we refer to a row or

Matrix	Vector	Scalar			Note
		Symbol	L <sup>A</sup> T <sub>E</sub> X	Code	
<i>A</i>	<i>a</i>	$\alpha$	<code>\alpha</code>	alpha	
<i>B</i>	<i>b</i>	$\beta$	<code>\beta</code>	beta	
<i>C</i>	<i>c</i>	$\gamma$	<code>\gamma</code>	gamma	
<i>D</i>	<i>d</i>	$\delta$	<code>\delta</code>	delta	
<i>E</i>	<i>e</i>	$\epsilon$	<code>\epsilon</code>	epsilon	$e_j = j$ th unit basis vector.
<i>F</i>	<i>f</i>	$\phi$	<code>\phi</code>	phi	
<i>G</i>	<i>g</i>	$\xi$	<code>\xi</code>	xi	
<i>H</i>	<i>h</i>	$\eta$	<code>\eta</code>	eta	
<i>I</i>					Used for identity matrix.
<i>K</i>	<i>k</i>	$\kappa$	<code>\kappa</code>	kappa	
<i>L</i>	<i>l</i>	$\lambda$	<code>\lambda</code>	lambda	
<i>M</i>	<i>m</i>	$\mu$	<code>\mu</code>	mu	$m(\cdot) =$ row dimension.
<i>N</i>	<i>n</i>	$\nu$	<code>\nu</code>	nu	$\nu$ is shared with V. $n(\cdot) =$ column dimension.
<i>P</i>	<i>p</i>	$\pi$	<code>\pi</code>	pi	
<i>Q</i>	<i>q</i>	$\theta$	<code>\theta</code>	theta	
<i>R</i>	<i>r</i>	$\rho$	<code>\rho</code>	rho	
<i>S</i>	<i>s</i>	$\sigma$	<code>\sigma</code>	sigma	
<i>T</i>	<i>t</i>	$\tau$	<code>\tau</code>	tau	
<i>U</i>	<i>u</i>	$\upsilon$	<code>\upsilon</code>	upsilon	
<i>V</i>	<i>v</i>	$\nu$	<code>\nu</code>	nu	$\nu$ shared with N.
<i>W</i>	<i>w</i>	$\omega$	<code>\omega</code>	omega	
<i>X</i>	<i>x</i>	$\chi$	<code>\chi</code>	chi	
<i>Y</i>	<i>y</i>	$\psi$	<code>\psi</code>	psi	
<i>Z</i>	<i>z</i>	$\zeta$	<code>\zeta</code>	zeta	

Figure 1.1: Correspondence between letters used for matrices (uppercase Roman)/vectors (lowercase Roman) and the symbols used to denote their scalar entries (lowercase Greek letters).

column, we may choose to skip the  $\hat{\cdot}$ . The  $T$  is used to indicate transposition of the column vector  $\hat{a}_i$ , to make it into the row vector  $\hat{a}_i^T$ .

Sometimes, we partition matrices into submatrices:

$$A = \left( A_0 \mid A_1 \mid \cdots \mid A_{N-1} \right) = \left( \begin{array}{c} \hat{A}_0 \\ \hat{A}_1 \\ \vdots \\ \hat{A}_{M-1} \end{array} \right) = \left( \begin{array}{c|c|c|c} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ \hline A_{1,0} & A_{1,1} & \cdots & A_{1,N-1} \\ \hline \vdots & \vdots & & \vdots \\ \hline A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,N-1} \end{array} \right).$$

For these partitionings of  $A \in \mathbb{C}^{m \times n}$ ,  $A_i \in \mathbb{C}^{m_i \times n}$ ,  $\hat{A}_j \in \mathbb{C}^{m \times n_j}$ , and  $A_{i,j} \in \mathbb{C}^{m_i \times n_j}$ , with  $\sum_i m_i = m$  and  $\sum_j n_j = n$ .

## 1.2 (Hermitian) Transposition

### 1.2.1 Conjugating a complex scalar

Recall that if  $\alpha = \alpha_r + i\alpha_c$ , then its (complex) conjugate is given by

$$\bar{\alpha} = \alpha_r - i\alpha_c$$

and its length (absolute value) by

$$|\alpha| = |\alpha_r + i\alpha_c| = \sqrt{\alpha_r^2 + \alpha_c^2} = \sqrt{(\alpha_r + i\alpha_c)(\alpha_r - i\alpha_c)} = \sqrt{\alpha\bar{\alpha}} = \sqrt{\bar{\alpha}\alpha} = |\bar{\alpha}|.$$

### 1.2.2 Conjugate of a vector

The (complex) conjugate of  $x$  is given by

$$\bar{x} = \begin{pmatrix} \bar{x}_0 \\ \bar{x}_1 \\ \vdots \\ \bar{x}_{n-1} \end{pmatrix}.$$

### 1.2.3 Conjugate of a matrix

The (complex) conjugate of  $A$  is given by

$$\bar{A} = \begin{pmatrix} \bar{\alpha}_{0,0} & \bar{\alpha}_{0,1} & \cdots & \bar{\alpha}_{0,n-1} \\ \bar{\alpha}_{1,0} & \bar{\alpha}_{1,1} & \cdots & \bar{\alpha}_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ \bar{\alpha}_{m-1,0} & \bar{\alpha}_{m-1,1} & \cdots & \bar{\alpha}_{m-1,n-1} \end{pmatrix}.$$

### 1.2.4 Transpose of a vector

The *transpose* of  $x$  is given by

$$x^T = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}^T = \left( \chi_0 \mid \chi_1 \mid \cdots \mid \chi_{n-1} \right).$$

Notice that transposing a (column) vector rearranges its elements to make a row vector.

### 1.2.5 Hermitian transpose of a vector

The *Hermitian transpose* of  $x$  is given by

$$x^H (= x^c) = (\bar{x})^T = \overline{\begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}}^T = \begin{pmatrix} \bar{\chi}_0 \\ \bar{\chi}_1 \\ \vdots \\ \bar{\chi}_{n-1} \end{pmatrix}^T = \left( \bar{\chi}_0 \mid \bar{\chi}_1 \mid \cdots \mid \bar{\chi}_{n-1} \right).$$

### 1.2.6 Transpose of a matrix

The *transpose* of  $A$  is given by

$$A^T = \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix}^T = \begin{pmatrix} \alpha_{0,0} & \alpha_{1,0} & \cdots & \alpha_{m-1,0} \\ \alpha_{0,1} & \alpha_{1,1} & \cdots & \alpha_{m-1,1} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha_{0,n-1} & \alpha_{1,n-1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix}.$$

### 1.2.7 Hermitian transpose (adjoint) of a matrix

The *Hermitian transpose* of  $A$  is given by

$$A^H (= A^c) = \bar{A}^T = \overline{\begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix}}^T = \begin{pmatrix} \bar{\alpha}_{0,0} & \bar{\alpha}_{1,0} & \cdots & \bar{\alpha}_{m-1,0} \\ \bar{\alpha}_{0,1} & \bar{\alpha}_{1,1} & \cdots & \bar{\alpha}_{m-1,1} \\ \vdots & \vdots & \cdots & \vdots \\ \bar{\alpha}_{0,n-1} & \bar{\alpha}_{1,n-1} & \cdots & \bar{\alpha}_{m-1,n-1} \end{pmatrix}.$$

## 1.2.8 Exercises

Homework 1.1 Partition  $A$ 

$$A = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) = \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix}.$$

*Convince yourself that the following hold:*

$$\bullet \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right)^T = \begin{pmatrix} a_0^T \\ a_1^T \\ \vdots \\ a_{m-1}^T \end{pmatrix}.$$

$$\bullet \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix}^T = \left( \widehat{a}_0 \mid \widehat{a}_1 \mid \cdots \mid \widehat{a}_{n-1} \right).$$

$$\bullet \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right)^H = \begin{pmatrix} a_0^H \\ a_1^H \\ \vdots \\ a_{m-1}^H \end{pmatrix}.$$

$$\bullet \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix}^H = \left( \overline{\widehat{a}_0} \mid \overline{\widehat{a}_1} \mid \cdots \mid \overline{\widehat{a}_{n-1}} \right).$$

 [SEE ANSWER](#)

Homework 1.2 Partition  $x$  into subvectors:

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}.$$

*Convince yourself that the following hold:*

$$\bullet \bar{x} = \begin{pmatrix} \overline{x_0} \\ \overline{x_1} \\ \vdots \\ \overline{x_{N-1}} \end{pmatrix}.$$

$$\bullet x^T = \left( x_0^T \mid x_1^T \mid \cdots \mid x_{N-1}^T \right).$$

$$\bullet x^H = \left( x_0^H \mid x_1^H \mid \cdots \mid x_{N-1}^H \right).$$

👉 SEE ANSWER

### Homework 1.3 Partition A

$$A = \begin{pmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,N-1} \end{pmatrix},$$

where  $A_{i,j} \in \mathbb{C}^{m_i \times n_j}$ . Here  $\sum_{i=0}^{M-1} m_i = m$  and  $\sum_{j=0}^{N-1} n_j = n$ .

Convince yourself that the following hold:

$$\bullet \begin{pmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,N-1} \end{pmatrix}^T = \begin{pmatrix} A_{0,0}^T & A_{1,0}^T & \cdots & A_{M-1,0}^T \\ A_{0,1}^T & A_{1,1}^T & \cdots & A_{M-1,1}^T \\ \vdots & \vdots & \cdots & \vdots \\ A_{0,N-1}^T & A_{1,N-1}^T & \cdots & A_{M-1,N-1}^T \end{pmatrix}.$$

$$\bullet \begin{pmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,N-1} \end{pmatrix}^H = \begin{pmatrix} A_{0,0}^H & A_{1,0}^H & \cdots & A_{M-1,0}^H \\ A_{0,1}^H & A_{1,1}^H & \cdots & A_{M-1,1}^H \\ \vdots & \vdots & \cdots & \vdots \\ A_{0,N-1}^H & A_{1,N-1}^H & \cdots & A_{M-1,N-1}^H \end{pmatrix}.$$

👉 SEE ANSWER



## 1.3 Vector-vector Operations

### 1.3.1 Scaling a vector (scal)

Let  $x \in \mathbb{C}^n$  and  $\alpha \in \mathbb{C}$ , with

$$x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}.$$

Then  $\alpha x$  equals the vector  $x$  scaled (stretched) by a factor  $\alpha$ :

$$\alpha x = \alpha \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha \chi_0 \\ \alpha \chi_1 \\ \vdots \\ \alpha \chi_{n-1} \end{pmatrix}.$$

If  $y := \alpha x$  with

$$y = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix},$$

then the following loop computes  $y$ :

```
for  $i := 0, \dots, n-1$ 
     $\psi_i := \alpha \chi_i$ 
endfor
```

**Homework 1.4** *Convince yourself of the following:*

- $\alpha x^T = \left( \alpha \chi_0 \mid \alpha \chi_1 \mid \cdots \mid \alpha \chi_{n-1} \right).$
- $(\alpha x)^T = \alpha x^T.$
- $(\alpha x)^H = \bar{\alpha} x^H.$

$$\bullet \alpha \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} = \begin{pmatrix} \alpha x_0 \\ \alpha x_1 \\ \vdots \\ \alpha x_{N-1} \end{pmatrix}$$

**Cost**

Scaling a vector of size  $n$  requires, approximately,  $n$  multiplications. Each of these becomes a floating point operation (flop) when the computation is performed as part of an algorithm executed on a computer that performs floating point computation. We will thus say that scaling a vector costs  $n$  flops.

It should be noted that arithmetic with complex numbers is roughly 4 times as expensive as is arithmetic with real numbers. In the chapter “Notes on Performance” (page ??) we also discuss that the cost of moving data impacts the cost of a flop. Thus, not all flops are created equal!

**1.3.2 Scaled vector addition (axpy)**

Let  $x, y \in \mathbb{C}^n$  and  $\alpha \in \mathbb{C}$ , with

$$x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix}.$$

Then  $\alpha x + y$  equals the vector

$$\alpha x + y = \alpha \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} + \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha\chi_0 \\ \alpha\chi_1 \\ \vdots \\ \alpha\chi_{n-1} \end{pmatrix} + \begin{pmatrix} \alpha\psi_0 \\ \alpha\psi_1 \\ \vdots \\ \alpha\psi_{n-1} \end{pmatrix}.$$

This operation is known as the `axpy` operation: scalar alpha times x plus y. Typically, the vector  $y$  is overwritten with the result:

$$y := \alpha x + y = \alpha \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} + \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha\chi_0 \\ \alpha\chi_1 \\ \vdots \\ \alpha\chi_{n-1} \end{pmatrix} + \begin{pmatrix} \alpha\psi_0 \\ \alpha\psi_1 \\ \vdots \\ \alpha\psi_{n-1} \end{pmatrix}$$

so that the following loop updates  $y$ :

```

for  $i := 0, \dots, n-1$ 
     $\psi_i := \alpha\chi_i + \psi_i$ 
endfor

```

**Homework 1.5** *Convince yourself of the following:*

$$\bullet \alpha \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} + \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} \alpha x_0 + y_0 \\ \alpha x_1 + y_1 \\ \vdots \\ \alpha x_{N-1} + y_{N-1} \end{pmatrix}. \quad (\text{Provided } x_i, y_i \in \mathbb{C}^{n_i} \text{ and } \sum_{i=0}^{N-1} n_i = n.)$$

 [SEE ANSWER](#)

**Cost**

The axpy with two vectors of size  $n$  requires, approximately,  $n$  multiplies and  $n$  additions or  $2n$  flops.

**1.3.3 Dot (inner) product (dot)**

Let  $x, y \in \mathbb{C}^n$  with

$$x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix}.$$

Then the dot product of  $x$  and  $y$  is defined by

$$\begin{aligned} x^H y &= \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}^H \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix} = \begin{pmatrix} \overline{\chi_0} & \overline{\chi_1} & \cdots & \overline{\chi_{n-1}} \end{pmatrix} \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix} \\ &= \overline{\chi_0} \psi_0 + \overline{\chi_1} \psi_1 + \cdots + \overline{\chi_{n-1}} \psi_{n-1} = \sum_{i=0}^{n-1} \overline{\chi_i} \psi_i. \end{aligned}$$

The following loop computes  $\alpha := x^H y$ :

```

α := 0
for i := 0, ..., n-1
    α :=  $\overline{\chi_i} \psi_i$  + α
endfor

```

**Homework 1.6** Convince yourself of the following:

$$\bullet \quad \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}^H \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} = \sum_{i=0}^{N-1} \overline{x_i} y_i. \quad (\text{Provided } x_i, y_i \in \mathbb{C}^{n_i} \text{ and } \sum_{i=0}^{N-1} n_i = n.)$$

🔗 [SEE ANSWER](#)

**Homework 1.7** Prove that  $x^H y = \overline{y^H x}$ .

🔗 [SEE ANSWER](#)

As we discuss matrix-vector multiplication and matrix-matrix multiplication, the closely related operation  $x^T y$  is also useful:

$$\begin{aligned} x^T y &= \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}^T \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix} = \begin{pmatrix} \chi_0 & \chi_1 & \cdots & \chi_{n-1} \end{pmatrix} \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix} \\ &= \chi_0 \psi_0 + \chi_1 \psi_1 + \cdots + \chi_{n-1} \psi_{n-1} = \sum_{i=0}^{n-1} \chi_i \psi_i. \end{aligned}$$

We will sometimes refer to this operation as a “dot” product since it is like the dot product  $x^H y$ , but without conjugation. In the case of real valued vectors, it *is* the dot product.

### Cost

The dot product of two vectors of size  $n$  requires, approximately,  $n$  multiplies and  $n$  additions. Thus, a dot product cost, approximately,  $2n$  flops.

## 1.4 Matrix-vector Operations

### 1.4.1 Matrix-vector multiplication (product)

Be sure to understand the relation between linear transformations and matrix-vector multiplication by reading Week 2 of

Linear Algebra: Foundations to Frontiers - Notes to LAFF With [29].

Let  $y \in \mathbb{C}^m$ ,  $A \in \mathbb{C}^{m \times n}$ , and  $x \in \mathbb{C}^n$  with

$$y = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix}, \quad A = \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix}, \quad \text{and} \quad x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}.$$

Then  $y = Ax$  means that

$$\begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix} = \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix} \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}$$


---

$$= \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix} \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}.$$

This is the definition of the matrix-vector product  $Ax$ , sometimes referred to as a general matrix-vector multiplication (`gemv`) when no special structure or properties of  $A$  are assumed.

Now, partition

$$A = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) = \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix}.$$

Focusing on how  $A$  can be partitioned by columns, we find that

$$\begin{aligned} y &= Ax = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \\ &= a_0\chi_0 + a_1\chi_1 + \cdots + a_{n-1}\chi_{n-1} \\ &= \chi_0 a_0 + \chi_1 a_1 + \cdots + \chi_{n-1} a_{n-1} \\ &= \chi_{n-1} a_{n-1} + (\cdots + (\chi_1 a_1 + (\chi_0 a_0 + 0)) \cdots), \end{aligned}$$

where 0 denotes the zero vector of size  $m$ . This suggests the following loop for computing  $y := Ax$ :

```

y := 0
for j := 0, ..., n-1
    y :=  $\chi_j a_j$  + y    (axpy)
endfor

```

In Figure 1.2 (left), we present this algorithm using the FLAME notation ([LAFF Notes Week 3 \[29\]](#)).

Focusing on how  $A$  can be partitioned by rows, we find that

$$y = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix} = Ax = \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix} x = \begin{pmatrix} \widehat{a}_0^T x \\ \widehat{a}_1^T x \\ \vdots \\ \widehat{a}_{m-1}^T x \end{pmatrix}.$$

This suggests the following loop for computing  $y := Ax$ :

Algorithm: $[y] := \text{MVMULT\_UNB\_VAR1}(A, x, y)$	Algorithm: $[y] := \text{MVMULT\_UNB\_VAR2}(A, x, y)$
<b>Partition</b> $A \rightarrow \left( A_L \middle  A_R \right), x \rightarrow \begin{pmatrix} x_T \\ x_B \end{pmatrix}$ <b>where</b> $A_L$ is 0 columns, $x_T$ has 0 elements <b>while</b> $n(A_L) < n(A)$ <b>do</b> <b>Repartition</b> $\left( A_L \middle  A_R \right) \rightarrow \left( A_0 \middle  a_1 \middle  A_2 \right), \begin{pmatrix} x_T \\ x_B \end{pmatrix} \rightarrow \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix}$ <hr/> $y := \chi_1 a_1 + y$ <hr/> <b>Continue with</b> $\left( A_L \middle  A_R \right) \leftarrow \left( A_0 \middle  a_1 \middle  A_2 \right), \begin{pmatrix} x_T \\ x_B \end{pmatrix} \leftarrow \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix}$ <b>endwhile</b>	<b>Partition</b> $A \rightarrow \begin{pmatrix} A_T \\ A_B \end{pmatrix}, y \rightarrow \begin{pmatrix} y_T \\ y_B \end{pmatrix}$ <b>where</b> $A_T$ has 0 rows, $y_T$ has 0 elements <b>while</b> $m(A_T) < m(A)$ <b>do</b> <b>Repartition</b> $\begin{pmatrix} A_T \\ A_B \end{pmatrix} \rightarrow \begin{pmatrix} A_0 \\ a_1^T \\ A_2 \end{pmatrix}, \begin{pmatrix} y_T \\ y_B \end{pmatrix} \rightarrow \begin{pmatrix} y_0 \\ \psi_1 \\ y_2 \end{pmatrix}$ <hr/> $\psi_1 := a_1^T x + \psi_1$ <hr/> <b>Continue with</b> $\begin{pmatrix} A_T \\ A_B \end{pmatrix} \leftarrow \begin{pmatrix} A_0 \\ a_1^T \\ A_2 \end{pmatrix}, \begin{pmatrix} y_T \\ y_B \end{pmatrix} \leftarrow \begin{pmatrix} y_0 \\ \psi_1 \\ y_2 \end{pmatrix}$ <b>endwhile</b>

Figure 1.2: Matrix-vector multiplication algorithms for  $y := Ax + y$ . Left: via `axpy` operations (by columns). Right: via “dot products” (by rows).

```

for  $i := 0, \dots, m-1$ 
     $\psi_i := \hat{a}_i^T x + \psi_i$     (``dot'')
endfor

```

Here we use the term “dot” because for complex valued matrices it is not really a dot product. In Figure 1.2 (right), we present this algorithm using the FLAME notation ([LAFF Notes Week 3 \[29\]](#)).

It is important to notice that this first “matrix-vector” operation (matrix-vector multiplication) can be “layered” upon vector-vector operations (`axpy` or “dot”).

## Cost

Matrix-vector multiplication with a  $m \times n$  matrix costs, approximately,  $2mn$  flops. This can be easily argued in three different ways:

- The computation requires a multiply and an add with each element of the matrix. There are  $mn$  such elements.

- The operation can be computed via  $n$  `axpy` operations, each of which requires  $2m$  flops.
- The operation can be computed via  $m$  `dot` operations, each of which requires  $2n$  flops.

### 1.4.2 Rank-1 update

Let  $y \in \mathbb{C}^m$ ,  $A \in \mathbb{C}^{m \times n}$ , and  $x \in \mathbb{C}^n$  with

$$y = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix}, \quad A = \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix}, \quad \text{and} \quad x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}.$$

The outerproduct of  $y$  and  $x$  is given by

$$\begin{aligned} yx^T &= \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix} \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix}^T = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix} \begin{pmatrix} \chi_0 & \chi_1 & \cdots & \chi_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} \psi_0\chi_0 & \psi_0\chi_1 & \cdots & \psi_0\chi_{n-1} \\ \psi_1\chi_0 & \psi_1\chi_1 & \cdots & \psi_1\chi_{n-1} \\ \vdots & \vdots & & \vdots \\ \psi_{m-1}\chi_0 & \psi_{m-1}\chi_1 & \cdots & \psi_{m-1}\chi_{n-1} \end{pmatrix}. \end{aligned}$$

Also,

$$yx^T = y \begin{pmatrix} \chi_0 & \chi_1 & \cdots & \chi_{n-1} \end{pmatrix} = \begin{pmatrix} \chi_0 y & \chi_1 y & \cdots & \chi_{n-1} y \end{pmatrix}.$$

This shows that all columns are a multiple of vector  $y$ . Finally,

$$yx^T = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix} x^T = \begin{pmatrix} \psi_0 x^T \\ \psi_1 x^T \\ \vdots \\ \psi_{m-1} x^T \end{pmatrix},$$

which shows that all columns are a multiple of row vector  $x^T$ . This motivates the observation that the matrix  $yx^T$  has rank at most equal to one ([LAFB Notes Week 10 \[29\]](#)).

The operation  $A := yx^T + A$  is called a rank-1 update to matrix  $A$  and is often referred to as underline-general rank-1 update (`ger`):

$$\begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,n-1} \end{pmatrix} :=$$

$$\begin{pmatrix} \Psi_0\chi_0 + \alpha_{0,0} & \Psi_0\chi_1 + \alpha_{0,1} & \cdots & \Psi_0\chi_{n-1} + \alpha_{0,n-1} \\ \Psi_1\chi_0 + \alpha_{1,0} & \Psi_1\chi_1 + \alpha_{1,1} & \cdots & \Psi_1\chi_{n-1} + \alpha_{1,n-1} \\ \vdots & \vdots & & \vdots \\ \Psi_{m-1}\chi_0 + \alpha_{m-1,0} & \Psi_{m-1}\chi_1 + \alpha_{m-1,1} & \cdots & \Psi_{m-1}\chi_{n-1} + \alpha_{m-1,n-1} \end{pmatrix}.$$

Now, partition

$$A = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) = \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix}.$$

Focusing on how  $A$  can be partitioned by columns, we find that

$$\begin{aligned} yx^T + A &= \left( \chi_0 y \mid \chi_1 y \mid \cdots \mid \chi_{n-1} y \right) + \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) \\ &= \left( \chi_0 y + a_0 \mid \chi_1 y + a_1 \mid \cdots \mid \chi_{n-1} y + a_{n-1} \right). \end{aligned}$$

Notice that each column is updated with an `axpy` operation. This suggests the following loop for computing  $A := yx^T + A$ :

```

for  $j := 0, \dots, n-1$ 
     $a_j := \chi_j y + a_j$     (axpy)
endfor

```

In Figure 1.3 (left), we present this algorithm using the FLAME notation ([LAFF Notes Week 3](#)).

Focusing on how  $A$  can be partitioned by rows, we find that

$$\begin{aligned} yx^T + A &= \begin{pmatrix} \Psi_0 x^T \\ \Psi_1 x^T \\ \vdots \\ \Psi_{m-1} x^T \end{pmatrix} + \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix} \\ &= \begin{pmatrix} \Psi_0 x^T + \widehat{a}_0^T \\ \Psi_1 x^T + \widehat{a}_1^T \\ \vdots \\ \Psi_{m-1} x^T + \widehat{a}_{m-1}^T \end{pmatrix}. \end{aligned}$$

Notice that each row is updated with an `axpy` operation. This suggests the following loop for computing  $A := yx^T + A$ :

```

for  $i := 0, \dots, m-1$ 
     $\widehat{a}_i^T := \Psi_i x^T + \widehat{a}_i^T$     (axpy)
endfor

```



Algorithm: $[A] := \text{RANK1\_UNB\_VAR1}(y, x, A)$	Algorithm: $[A] := \text{RANK1\_UNB\_VAR2}(y, x, A)$
<b>Partition</b> $x \rightarrow \begin{pmatrix} x_T \\ \frac{x_T}{x_B} \end{pmatrix}, A \rightarrow (A_L   A_R)$ <b>where</b> $x_T$ has 0 elements, $A_L$ has 0 columns <b>while</b> $m(x_T) < m(x)$ <b>do</b> <b>Repartition</b> $\begin{pmatrix} x_T \\ \frac{x_T}{x_B} \end{pmatrix} \rightarrow \begin{pmatrix} x_0 \\ \frac{\chi_1}{x_2} \end{pmatrix}, (A_L   A_R) \rightarrow (A_0   a_1   A_2)$ <hr/> $a_1 := y\chi_1 + a_1$ <hr/> <b>Continue with</b> $\begin{pmatrix} x_T \\ \frac{x_T}{x_B} \end{pmatrix} \leftarrow \begin{pmatrix} x_0 \\ \frac{\chi_1}{x_2} \end{pmatrix}, (A_L   A_R) \leftarrow (A_0   a_1   A_2)$ <b>endwhile</b>	<b>Partition</b> $y \rightarrow \begin{pmatrix} y_T \\ \frac{y_T}{y_B} \end{pmatrix}, A \rightarrow \begin{pmatrix} A_T \\ A_B \end{pmatrix}$ <b>where</b> $y_T$ has 0 elements, $A_T$ has 0 rows <b>while</b> $m(y_T) < m(y)$ <b>do</b> <b>Repartition</b> $\begin{pmatrix} y_T \\ \frac{y_T}{y_B} \end{pmatrix} \rightarrow \begin{pmatrix} y_0 \\ \frac{\psi_1}{y_2} \end{pmatrix}, \begin{pmatrix} A_T \\ A_B \end{pmatrix} \rightarrow \begin{pmatrix} A_0 \\ \frac{a_1^T}{A_2} \end{pmatrix}$ <hr/> $a_1^T := \psi_1 x^T + a_1^T$ <hr/> <b>Continue with</b> $\begin{pmatrix} y_T \\ \frac{y_T}{y_B} \end{pmatrix} \leftarrow \begin{pmatrix} y_0 \\ \frac{\psi_1}{y_2} \end{pmatrix}, \begin{pmatrix} A_T \\ A_B \end{pmatrix} \leftarrow \begin{pmatrix} A_0 \\ \frac{a_1^T}{A_2} \end{pmatrix}$ <b>endwhile</b>

Figure 1.3: Rank-1 update algorithms for computing  $A := yx^T + A$ . Left: by columns. Right: by rows.

In Figure 1.3 (right), we present this algorithm using the FLAME notation ([LAFF Notes Week 3](#)).

Again, it is important to notice that this “matrix-vector” operation (rank-1 update) can be “layered” upon the `axpy` vector-vector operation.

### Cost

A rank-1 update of a  $m \times n$  matrix costs, approximately,  $2mn$  flops. This can be easily argued in three different ways:

- The computation requires a multiply and an add with each element of the matrix. There are  $mn$  such elements.
- The operation can be computed one column at a time via  $n$  `axpy` operations, each of which requires  $2m$  flops.
- The operation can be computed one row at a time via  $m$  `axpy` operations, each of which requires  $2n$  flops.

## 1.5 Matrix-matrix multiplication (product)

Be sure to understand the relation between linear transformations and matrix-matrix multiplication ([LAFF Notes Weeks 3 and 4](#) [29]).

We will now discuss the computation of  $C := AB + C$ , where  $C \in \mathbb{C}^{m \times n}$ ,  $A \in \mathbb{C}^{m \times k}$ , and  $B \in \mathbb{C}^{k \times n}$ . (If one wishes to compute  $C := AB$ , one can always start by setting  $C := 0$ , the zero matrix.) This is the definition of the matrix-matrix product  $AB$ , sometimes referred to as a general matrix-matrix multiplication (gemm) when no special structure or properties of  $A$  and  $B$  are assumed.

### 1.5.1 Element-by-element computation

Let

$$C = \begin{pmatrix} \gamma_{0,0} & \gamma_{0,1} & \cdots & \gamma_{0,n-1} \\ \gamma_{1,0} & \gamma_{1,1} & \cdots & \gamma_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m-1,0} & \gamma_{m-1,1} & \cdots & \gamma_{m-1,n-1} \end{pmatrix}, A = \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \cdots & \alpha_{0,k-1} \\ \alpha_{1,0} & \alpha_{1,1} & \cdots & \alpha_{1,k-1} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha_{m-1,0} & \alpha_{m-1,1} & \cdots & \alpha_{m-1,k-1} \end{pmatrix}$$

$$B = \begin{pmatrix} \beta_{0,0} & \beta_{0,1} & \cdots & \beta_{0,n-1} \\ \beta_{1,0} & \beta_{1,1} & \cdots & \beta_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{k-1,0} & \beta_{k-1,1} & \cdots & \beta_{k-1,n-1} \end{pmatrix}.$$

Then

$$C := AB + C$$

$$= \begin{pmatrix} \sum_{p=0}^{k-1} \alpha_{0,p} \beta_{p,0} + \gamma_{0,0} & \sum_{p=0}^{k-1} \alpha_{0,p} \beta_{p,1} + \gamma_{0,1} & \cdots & \sum_{p=0}^{k-1} \alpha_{0,p} \beta_{p,n-1} + \gamma_{0,n-1} \\ \sum_{p=0}^{k-1} \alpha_{1,p} \beta_{p,0} + \gamma_{1,0} & \sum_{p=0}^{k-1} \alpha_{1,p} \beta_{p,1} + \gamma_{1,1} & \cdots & \sum_{p=0}^{k-1} \alpha_{1,p} \beta_{p,n-1} + \gamma_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{p=0}^{k-1} \alpha_{m-1,p} \beta_{p,0} + \gamma_{m-1,0} & \sum_{p=0}^{k-1} \alpha_{m-1,p} \beta_{p,1} + \gamma_{m-1,1} & \cdots & \sum_{p=0}^{k-1} \alpha_{m-1,p} \beta_{p,n-1} + \gamma_{m-1,n-1} \end{pmatrix}.$$

This can be more elegantly stated by partitioning

$$A = \begin{pmatrix} \hat{a}_0^T \\ \hat{a}_1^T \\ \vdots \\ \hat{a}_{m-1}^T \end{pmatrix} \quad \text{and} \quad B = \left( b_0 \mid b_1 \mid \cdots \mid b_{n-1} \right).$$

Then

$$C := AB + C = \begin{pmatrix} \hat{a}_0^T \\ \hat{a}_1^T \\ \vdots \\ \hat{a}_{m-1}^T \end{pmatrix} \left( b_0 \mid b_1 \mid \cdots \mid b_{n-1} \right)$$

$$= \begin{pmatrix} \hat{a}_0^T b_0 + \gamma_{0,0} & \hat{a}_0^T b_1 + \gamma_{0,1} & \cdots & \hat{a}_0^T b_{n-1} + \gamma_{0,n-1} \\ \hat{a}_1^T b_0 + \gamma_{1,0} & \hat{a}_1^T b_1 + \gamma_{1,1} & \cdots & \hat{a}_1^T b_{n-1} + \gamma_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{a}_{m-1}^T b_0 + \gamma_{m-1,0} & \hat{a}_{m-1}^T b_1 + \gamma_{m-1,1} & \cdots & \hat{a}_{m-1}^T b_{n-1} + \gamma_{m-1,n-1} \end{pmatrix}.$$

### 1.5.2 Via matrix-vector multiplications

Partition

$$C = \left( c_0 \mid c_1 \mid \cdots \mid c_{n-1} \right) \quad \text{and} \quad B = \left( b_0 \mid b_1 \mid \cdots \mid b_{n-1} \right).$$

Then

$$C := AB + C = A \left( b_0 \mid b_1 \mid \cdots \mid b_{n-1} \right) + \left( c_0 \mid c_1 \mid \cdots \mid c_{n-1} \right) = \left( Ab_0 + c_0 \mid Ab_1 + c_1 \mid \cdots \mid Ab_{n-1} + c_{n-1} \right)$$

which shows that each column of  $C$  is updated with a matrix-vector multiplication:  $c_j := Ab_j + c_j$ :

```

for  $j := 0, \dots, n-1$ 
     $c_j := Ab_j + c_j$     (matrix-vector multiplication)
endfor

```

### 1.5.3 Via row-vector times matrix multiplications

Partition

$$C = \begin{pmatrix} \hat{c}_0^T \\ \hat{c}_1^T \\ \vdots \\ \hat{c}_{m-1}^T \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} \hat{a}_0^T \\ \hat{a}_1^T \\ \vdots \\ \hat{a}_{m-1}^T \end{pmatrix}.$$

Then

$$C := AB + C = \begin{pmatrix} \hat{a}_0^T \\ \hat{a}_1^T \\ \vdots \\ \hat{a}_{m-1}^T \end{pmatrix} B + \begin{pmatrix} \hat{c}_0^T \\ \hat{c}_1^T \\ \vdots \\ \hat{c}_{m-1}^T \end{pmatrix} = \begin{pmatrix} \hat{a}_0^T B + \hat{c}_0^T \\ \hat{a}_1^T B + \hat{c}_1^T \\ \vdots \\ \hat{a}_{m-1}^T B + \hat{c}_{m-1}^T \end{pmatrix}$$

which shows that each row of  $C$  is updated with a row-vector time matrix multiplication:  $\hat{c}_i^T := \hat{a}_i^T B + \hat{c}_i^T$ :

```

for  $i := 0, \dots, m-1$ 
     $\hat{c}_i^T := \hat{a}_i^T B + \hat{c}_i^T$     (row-vector times matrix-vector multiplication)
endfor

```

### 1.5.4 Via rank-1 updates

Partition

$$A = \left( a_0 \mid a_1 \mid \cdots \mid a_{k-1} \right) \quad \text{and} \quad B = \begin{pmatrix} \widehat{b}_0^T \\ \widehat{b}_1^T \\ \vdots \\ \widehat{b}_{k-1}^T \end{pmatrix}.$$

The

$$\begin{aligned} C := AB + C &= \left( a_0 \mid a_1 \mid \cdots \mid a_{k-1} \right) \begin{pmatrix} \widehat{b}_0^T \\ \widehat{b}_1^T \\ \vdots \\ \widehat{b}_{k-1}^T \end{pmatrix} + C \\ &= a_0 \widehat{b}_0^T + a_1 \widehat{b}_1^T + \cdots + a_{k-1} \widehat{b}_{k-1}^T + C \\ &= a_{k-1} \widehat{b}_{k-1}^T + (\cdots (a_1 \widehat{b}_1^T + (a_0 \widehat{b}_0^T + C)) \cdots) \end{aligned}$$

which shows that  $C$  can be updated with a sequence of rank-1 update, suggesting the loop

```

for  $p := 0, \dots, k-1$ 
     $C := a_p \widehat{b}_p^T + C$     (rank-1 update)
endfor

```

### 1.5.5 Cost

A matrix-matrix multiplication  $C := AB$ , where  $C$  is  $m \times n$ ,  $A$  is  $m \times k$ , and  $B$  is  $k \times n$ , costs approximately  $2mnk$  flops. This can be easily argued in four different ways:

- The computation requires a dot product for each element in  $C$ , at a cost of  $2k$  flops per dot product. There are  $mn$  such elements.
- The operation can be computed one column at a time via  $n$  gemv operations, each of which requires  $2mk$  flops.
- The operation can be computed one row at a time via  $m$  gemv operations (row-vector times matrix), each of which requires  $2nk$  flops.
- The operation can be computed via  $k$  rank-1 updates,  $2mn$  flops.

## 1.6 Summarizing All

It is important to realize that almost all operations that we discussed in this note are special cases of matrix-matrix multiplication. This is summarized in Figure 1.4. A few notes:

$m$	$n$	$k$	Illustration	Label
large	large	large		gemm
large	large	1		ger
large	1	large		gemv
1	large	large		gemv
1	large	1		axpy
large	1	1		axpy
1	1	large		dot
1	1	1		MAC

Figure 1.4: Special shapes of `gemm`  $C := AB + C$ . Here  $C$ ,  $A$ , and  $B$  are  $m \times n$ ,  $m \times k$ , and  $k \times n$  matrices, respectively.

- Row-vector times matrix is matrix-vector multiplication in disguise: If  $y^T := x^T A$  then  $y := A^T x$ .
- For a similar reason  $y^T := \alpha x^T + y^T$  is an `axpy` in disguise:  $y := \alpha x + y$ .
- The operation  $y := x\alpha + y$  is the same as  $y := \alpha x + y$  since multiplication of a vector by a scalar commutes.
- The operation  $\gamma := \alpha\beta + \gamma$  is known as a Multiply-Accumulate (MAC) operation. Often floating point hardware implements this as an integrated operation.

Observations made about operations with partitioned matrices and vectors can be summarized by partitioning matrices into blocks: Let

$$C = \begin{pmatrix} C_{0,0} & C_{0,1} & \cdots & C_{0,N-1} \\ C_{1,0} & C_{1,1} & \cdots & C_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ C_{M-1,0} & C_{M-1,1} & \cdots & C_{M-1,N-1} \end{pmatrix}, A = \begin{pmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,K-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,K-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,K-1} \end{pmatrix}$$

$$B = \begin{pmatrix} B_{0,1} & B_{0,1} & \cdots & B_{0,N-1} \\ B_{1,0} & B_{1,1} & \cdots & B_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ B_{K-1,0} & B_{K-1,1} & \cdots & B_{K-1,N-1} \end{pmatrix}.$$

Then

$$C := AB + C = \begin{pmatrix} \sum_{p=0}^{K-1} A_{0,p} B_{p,0} + C_{0,0} & \sum_{p=0}^{K-1} A_{0,p} B_{p,1} + C_{0,1} & \cdots & \sum_{p=0}^{K-1} A_{0,p} B_{p,N-1} + C_{0,N-1} \\ \sum_{p=0}^{K-1} A_{1,p} B_{p,0} + C_{1,0} & \sum_{p=0}^{K-1} A_{1,p} B_{p,1} + C_{1,1} & \cdots & \sum_{p=0}^{K-1} A_{1,p} B_{p,N-1} + C_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{p=0}^{K-1} A_{M-1,p} B_{p,0} + C_{M-1,0} & \sum_{p=0}^{K-1} A_{M-1,p} B_{p,1} + C_{M-1,1} & \cdots & \sum_{p=0}^{K-1} A_{M-1,p} B_{p,N-1} + C_{M-1,N-1} \end{pmatrix}.$$

Provided the partitionings of  $C$ ,  $A$ , and  $B$  are “conformal”. Loosely speaking, partitionings of vectors and/or matrices are *conformal* if they match in a way that makes operations with the subvectors and/or submatrices legal.

Notice that multiplication with partitioned matrices is exactly like regular matrix-matrix multiplication with scalar elements, *except that multiplication of two blocks does not necessarily commute*.

## Notes on Vector and Matrix Norms

### Video

Read disclaimer regarding the videos in the preface!

👉 [YouTube](#)

👉 [Download from UT Box](#)

👉 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

## Outline

<b>Video</b> . . . . .	<b>23</b>
<b>Outline</b> . . . . .	<b>24</b>
<b>2.1. Absolute Value</b> . . . . .	<b>25</b>
<b>2.2. Vector Norms</b> . . . . .	<b>25</b>
2.2.1. Vector 2-norm (length) . . . . .	25
2.2.2. Vector 1-norm . . . . .	27
2.2.3. Vector $\infty$ -norm (infinity norm) . . . . .	27
2.2.4. Vector $p$ -norm . . . . .	27
<b>2.3. Matrix Norms</b> . . . . .	<b>27</b>
2.3.1. Frobenius norm . . . . .	27
2.3.2. Induced matrix norms . . . . .	28
2.3.3. Special cases used in practice (matrix $p$ -norm) . . . . .	29
2.3.4. Discussion . . . . .	31
2.3.5. Submultiplicative matrix norms . . . . .	31
<b>2.4. An Application to Conditioning of Linear Systems</b> . . . . .	<b>32</b>
<b>2.5. Equivalence of Norms</b> . . . . .	<b>33</b>



## 2.1 Absolute Value

Recall that if  $\alpha \in \mathbb{C}$ , then  $|\alpha|$  equals its absolute value. In other words, if  $\alpha = \alpha_r + i\alpha_c$ , then  $|\alpha| = \sqrt{\alpha_r^2 + \alpha_c^2} = \sqrt{\bar{\alpha}\alpha}$ .

This absolute value function has the following properties:

- $\alpha \neq 0 \Rightarrow |\alpha| > 0$  ( $|\cdot|$  is positive definite),
- $|\alpha\beta| = |\alpha||\beta|$  ( $|\cdot|$  is homogeneous), and
- $|\alpha + \beta| \leq |\alpha| + |\beta|$  ( $|\cdot|$  obeys the triangle inequality).

## 2.2 Vector Norms

A (vector) norm extends the notion of an absolute value (length or size) to vectors:

**Definition 2.1** Let  $v : \mathbb{C}^n \rightarrow \mathbb{R}$ . Then  $v$  is a vector norm if for all  $x, y \in \mathbb{C}^n$

- $x \neq 0 \Rightarrow v(x) > 0$  ( $v$  is positive definite),
- $v(\alpha x) = |\alpha|v(x)$  ( $v$  is homogeneous), and
- $v(x + y) \leq v(x) + v(y)$  ( $v$  obeys the triangle inequality).

**Homework 2.2** Prove that if  $v : \mathbb{C}^n \rightarrow \mathbb{R}$  is a norm, then  $v(0) = 0$  (where the first 0 denotes the zero vector in  $\mathbb{C}^n$ ).

[SEE ANSWER](#)

Note: often we will use  $\|\cdot\|$  to denote a vector norm.

### 2.2.1 Vector 2-norm (length)

**Definition 2.3** The vector 2-norm,  $\|\cdot\|_2 : \mathbb{C}^n \rightarrow \mathbb{R}$ , is defined for  $x \in \mathbb{C}^n$  by

$$\|x\|_2 = \sqrt{x^H x} = \sqrt{\bar{\chi}_0 \chi_0 + \cdots + \bar{\chi}_{n-1} \chi_{n-1}} = \sqrt{|\chi_0|^2 + \cdots + |\chi_{n-1}|^2}.$$

To show that the vector 2-norm is a norm, we will need the following theorem:

**Theorem 2.4** (Cauchy-Schwartz inequality) Let  $x, y \in \mathbb{C}^n$ . Then  $|x^H y| \leq \|x\|_2 \|y\|_2$ .

**Proof:** Assume that  $x \neq 0$  and  $y \neq 0$ , since otherwise the inequality is trivially true. We can then choose  $\hat{x} = x/\|x\|_2$  and  $\hat{y} = y/\|y\|_2$ . This leaves us to prove that  $|\hat{x}^H \hat{y}| \leq 1$ , with  $\|\hat{x}\|_2 = \|\hat{y}\|_2 = 1$ .

Pick  $\alpha \in \mathbb{C}$  with  $|\alpha| = 1$  s that  $\alpha \hat{x}^H \hat{y}$  is real and nonnegative. Note that since it is real,  $\alpha \hat{x}^H \hat{y} = \overline{\alpha \hat{x}^H \hat{y}} = \overline{\alpha} \hat{y}^H \hat{x}$ .

Now,

$$\begin{aligned}
0 &\leq \|\hat{x} - \alpha\hat{y}\|_2^2 \\
&= (x - \alpha\hat{y})^H (\hat{x} - \alpha\hat{y}) && (\|z\|_2^2 = z^H z) \\
&= \hat{x}^H \hat{x} - \overline{\alpha}\hat{y}^H \hat{x} - \alpha\hat{x}^H \hat{y} + \overline{\alpha}\alpha\hat{y}^H \hat{y} && (\text{multiplying out}) \\
&= 1 - 2\alpha\hat{x}^H \hat{y} + |\alpha|^2 && (\|\hat{x}\|_2 = \|\hat{y}\|_2 = 1 \text{ and } \alpha\hat{x}^H \hat{y} = \overline{\alpha\hat{x}^H \hat{y}} = \overline{\alpha}\hat{y}^H \hat{x}) \\
&= 2 - 2\alpha\hat{x}^H \hat{y} && (|\alpha| = 1).
\end{aligned}$$

Thus  $1 \geq \alpha\hat{x}^H \hat{y}$  and, taking the absolute value of both sides,

$$1 \geq |\alpha\hat{x}^H \hat{y}| = |\alpha| |\hat{x}^H \hat{y}| = |\hat{x}^H \hat{y}|,$$

which is the desired result. QED

**Theorem 2.5** *The vector 2-norm is a norm.*

**Proof:** To prove this, we merely check whether the three conditions are met:

Let  $x, y \in \mathbb{C}^n$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $x \neq 0 \Rightarrow \|x\|_2 > 0$  ( $\|\cdot\|_2$  is positive definite):

Notice that  $x \neq 0$  means that at least one of its components is nonzero. Let's assume that  $\chi_j \neq 0$ . Then

$$\|x\|_2 = \sqrt{|\chi_0|^2 + \cdots + |\chi_{n-1}|^2} \geq \sqrt{|\chi_j|^2} = |\chi_j| > 0.$$

- $\|\alpha x\|_2 = |\alpha| \|x\|_2$  ( $\|\cdot\|_2$  is homogeneous):

$$\begin{aligned}
\|\alpha x\|_2 &= \sqrt{|\alpha\chi_0|^2 + \cdots + |\alpha\chi_{n-1}|^2} \\
&= \sqrt{|\alpha|^2 |\chi_0|^2 + \cdots + |\alpha|^2 |\chi_{n-1}|^2} \\
&= \sqrt{|\alpha|^2 (|\chi_0|^2 + \cdots + |\chi_{n-1}|^2)} \\
&= |\alpha| \sqrt{|\chi_0|^2 + \cdots + |\chi_{n-1}|^2} \\
&= |\alpha| \|x\|_2.
\end{aligned}$$

- $\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$  ( $\|\cdot\|_2$  obeys the triangle inequality).

$$\begin{aligned}
\|x + y\|_2^2 &= (x + y)^H (x + y) \\
&= x^H x + y^H x + x^H y + y^H y \\
&\leq \|x\|_2^2 + 2\|x\|_2 \|y\|_2 + \|y\|_2^2 \\
&= (\|x\|_2 + \|y\|_2)^2.
\end{aligned}$$

Taking the square root of both sides yields the desired result.

QED

### 2.2.2 Vector 1-norm

**Definition 2.6** The vector 1-norm  $\|\cdot\|_1 : \mathbb{C}^n \rightarrow \mathbb{R}$  is defined for  $x \in \mathbb{C}^n$  by

$$\|x\|_1 = |\chi_0| + |\chi_1| + \cdots + |\chi_{n-1}|.$$

**Homework 2.7** The vector 1-norm is a norm.

 [SEE ANSWER](#)

The vector 1-norm is sometimes referred to as the “taxi-cab norm”. It is the distance that a taxi travels along the streets of a city that has square blocks.

### 2.2.3 Vector $\infty$ -norm (infinity norm)

**Definition 2.8** The vector  $\infty$ -norm,  $\|\cdot\|_\infty : \mathbb{C}^n \rightarrow \mathbb{R}$ , is defined for  $x \in \mathbb{C}^n$  by  $\|x\|_\infty = \max_i |\chi_i|$ .

**Homework 2.9** The vector  $\infty$ -norm is a norm.

 [SEE ANSWER](#)

### 2.2.4 Vector $p$ -norm

**Definition 2.10** The vector  $p$ -norm  $\|\cdot\|_p : \mathbb{C}^n \rightarrow \mathbb{R}$  is defined for  $x \in \mathbb{C}^n$  by

$$\|x\|_p = \sqrt[p]{|\chi_0|^p + |\chi_1|^p + \cdots + |\chi_{n-1}|^p}.$$

Proving that the  $p$ -norm is a norm is a little tricky and not particularly relevant to this course. To prove the triangle inequality requires the following classical result:

**Theorem 2.11** (Hölder inequality) Let  $x, y \in \mathbb{C}^n$  and  $\frac{1}{p} + \frac{1}{q} = 1$  with  $1 \leq p, q \leq \infty$ . Then  $|x^H y| \leq \|x\|_p \|y\|_q$ .

Clearly, the 1-norm and 2 norms are special cases of the  $p$ -norm. Also,  $\|x\|_\infty = \lim_{p \rightarrow \infty} \|x\|_p$ .

## 2.3 Matrix Norms

It is not hard to see that vector norms are all measures of how “big” the vectors are. Similarly, we want to have measures for how “big” matrices are. We will start with one that are somewhat artificial and then move on to the important class of induced matrix norms.

### 2.3.1 Frobenius norm

**Definition 2.12** The Frobenius norm,  $\|\cdot\|_F : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$ , is defined for  $A \in \mathbb{C}^{m \times n}$  by

$$\|A\|_F = \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}|^2}.$$

Notice that one can think of the Frobenius norm as taking the columns of the matrix, stacking them on top of each other to create a vector of size  $m \times n$ , and then taking the vector 2-norm of the result.

**Homework 2.13** Show that the Frobenius norm is a norm.

 [SEE ANSWER](#)

Similarly, other matrix norms can be created from vector norms by viewing the matrix as a vector. It turns out that other than the Frobenius norm, these aren't particularly interesting in practice.

## 2.3.2 Induced matrix norms

**Definition 2.14** Let  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  be vector norms. Define  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_{\mu,\nu} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu}.$$

Let us start by interpreting this. How “big”  $A$  is, as measured by  $\|A\|_{\mu,\nu}$ , is defined as the most that  $A$  magnifies the length of nonzero vectors, where the length of the vectors ( $x$ ) is measured with norm  $\|\cdot\|_\nu$  and the length of the transformed vector ( $Ax$ ) is measured with norm  $\|\cdot\|_\mu$ .

Two comments are in order. First,

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu} = \sup_{\|x\|_\nu=1} \|Ax\|_\mu.$$

This follows immediately from the fact this sequence of equivalences:

$$\sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \left\| \frac{Ax}{\|x\|_\nu} \right\|_\mu = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \left\| A \frac{x}{\|x\|_\nu} \right\|_\mu = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0 \\ y = \frac{x}{\|x\|_\nu}}} \|Ay\|_\mu = \sup_{\|y\|_\nu=1} \|Ay\|_\mu = \sup_{\|x\|_\nu=1} \|Ax\|_\mu.$$

Also the “sup” (which stands for supremum) is used because we can't claim yet that there is a vector  $x$  with  $\|x\|_\nu = 1$  for which

$$\|A\|_{\mu,\nu} = \|Ax\|_\mu.$$

The fact is that there is always such a vector  $x$ . The proof depends on a result from real analysis (sometimes called “advanced calculus”) that states that  $\sup_{x \in S} f(x)$  is attained for some vector  $x \in S$  as long as  $f$  is continuous and  $S$  is a compact set. Since real analysis is not a prerequisite for this course, the reader may have to take this on faith!

We conclude that the following two definitions are equivalent definitions to the one we already gave:

**Definition 2.15** Let  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  be vector norms. Define  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_{\mu,\nu} = \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu}.$$

and

**Definition 2.16** Let  $\|\cdot\|_\mu : \mathbb{C}^m \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  be vector norms. Define  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_{\mu,\nu} = \max_{\|x\|_\nu=1} \|Ax\|_\mu.$$

**Theorem 2.17**  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is a norm.

**Proof:** To prove this, we merely check whether the three conditions are met:

Let  $A, B \in \mathbb{C}^{m \times n}$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $A \neq 0 \Rightarrow \|A\|_{\mu,\nu} > 0$  ( $\|\cdot\|_{\mu,\nu}$  is positive definite):

Notice that  $A \neq 0$  means that at least one of its columns is not a zero vector (since at least one element). Let us assume it is the  $j$ th column,  $a_j$ , that is nonzero. Then

$$\|A\|_{\mu,\nu} = \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu} \geq \frac{\|Ae_j\|_\mu}{\|e_j\|_\nu} = \frac{\|a_j\|_\mu}{\|e_j\|_\nu} > 0.$$

- $\|\alpha A\|_{\mu,\nu} = |\alpha| \|A\|_{\mu,\nu}$  ( $\|\cdot\|_{\mu,\nu}$  is homogeneous):

$$\|\alpha A\|_{\mu,\nu} = \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|\alpha Ax\|_\mu}{\|x\|_\nu} = \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} |\alpha| \frac{\|Ax\|_\mu}{\|x\|_\nu} = |\alpha| \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu} = |\alpha| \|A\|_{\mu,\nu}.$$

- $\|A + B\|_{\mu,\nu} \leq \|A\|_{\mu,\nu} + \|B\|_{\mu,\nu}$  ( $\|\cdot\|_{\mu,\nu}$  obeys the triangle inequality).

$$\begin{aligned} \|A + B\|_{\mu,\nu} &= \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|(A + B)x\|_\mu}{\|x\|_\nu} = \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax + Bx\|_\mu}{\|x\|_\nu} \leq \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu + \|Bx\|_\mu}{\|x\|_\nu} \\ &\leq \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \left( \frac{\|Ax\|_\mu}{\|x\|_\nu} + \frac{\|Bx\|_\mu}{\|x\|_\nu} \right) \leq \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_\mu}{\|x\|_\nu} + \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Bx\|_\mu}{\|x\|_\nu} = \|A\|_{\mu,\nu} + \|B\|_{\mu,\nu}. \end{aligned}$$

QED

### 2.3.3 Special cases used in practice (matrix p-norm)

The most important case of  $\|\cdot\|_{\mu,\nu} : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  uses the same norm for  $\|\cdot\|_\mu$  and  $\|\cdot\|_\nu$  (except that  $m$  may not equal  $n$ ).

**Definition 2.18** Define  $\|\cdot\|_p : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  by

$$\|A\|_p = \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p=1} \|Ax\|_p.$$

Special cases are the matrix 2-norm ( $\|\cdot\|_2$ ), matrix 1-norm ( $\|\cdot\|_1$ ), and matrix  $\infty$ -norm ( $\|\cdot\|_\infty$ ).

**Theorem 2.19** Let  $A \in \mathbb{C}^{m \times n}$  and partition  $A = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right)$ . Show that

$$\|A\|_1 = \max_{0 \leq j < n} \|a_j\|_1.$$

**Proof:** Let  $\bar{j}$  be chosen so that  $\max_{0 \leq j < n} \|a_j\|_1 = \|a_{\bar{j}}\|_1$ . Then

$$\begin{aligned} \max_{\|x\|_1=1} \|Ax\|_1 &= \max_{\|x\|_1=1} \left\| \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \right\|_1 \\ &= \max_{\|x\|_1=1} \|\chi_0 a_0 + \chi_1 a_1 + \cdots + \chi_{n-1} a_{n-1}\|_1 \\ &\leq \max_{\|x\|_1=1} (\|\chi_0 a_0\|_1 + \|\chi_1 a_1\|_1 + \cdots + \|\chi_{n-1} a_{n-1}\|_1) \\ &= \max_{\|x\|_1=1} (|\chi_0| \|a_0\|_1 + |\chi_1| \|a_1\|_1 + \cdots + |\chi_{n-1}| \|a_{n-1}\|_1) \\ &\leq \max_{\|x\|_1=1} (|\chi_0| \|a_{\bar{j}}\|_1 + |\chi_1| \|a_{\bar{j}}\|_1 + \cdots + |\chi_{n-1}| \|a_{\bar{j}}\|_1) \\ &= \max_{\|x\|_1=1} (|\chi_0| + |\chi_1| + \cdots + |\chi_{n-1}|) \|a_{\bar{j}}\|_1 \\ &= \|a_{\bar{j}}\|_1. \end{aligned}$$

Also,

$$\|a_{\bar{j}}\|_1 = \|Ae_{\bar{j}}\|_1 \leq \max_{\|x\|_1=1} \|Ax\|_1.$$

Hence

$$\|a_{\bar{j}}\|_1 \leq \max_{\|x\|_1=1} \|Ax\|_1 \leq \|a_{\bar{j}}\|_1$$

which implies that

$$\max_{\|x\|_1=1} \|Ax\|_1 = \|a_{\bar{j}}\|_1 = \max_{0 \leq j < n} \|a_j\|_1.$$

QED

**Homework 2.20** Let  $A \in \mathbb{C}^{m \times n}$  and partition  $A = \begin{pmatrix} \hat{a}_0^T \\ \hat{a}_1^T \\ \vdots \\ \hat{a}_{m-1}^T \end{pmatrix}$ . Show that

$$\|A\|_\infty = \max_{0 \leq i < m} \|\hat{a}_i\|_1 = \max_{0 \leq i < m} (|\alpha_{i,0}| + |\alpha_{i,1}| + \cdots + |\alpha_{i,n-1}|)$$

SEE ANSWER

Notice that in the above exercise  $\hat{a}_i$  is really  $(\hat{a}_i^T)^T$  since  $\hat{a}_i^T$  is the label for the  $i$ th row of matrix  $A$ .

**Homework 2.21** Let  $y \in \mathbb{C}^m$  and  $x \in \mathbb{C}^n$ . Show that  $\|yx^H\|_2 = \|y\|_2 \|x\|_2$ .

SEE ANSWER

### 2.3.4 Discussion

While  $\|\cdot\|_2$  is a very important matrix norm, it is in practice often difficult to compute. The matrix norms,  $\|\cdot\|_F$ ,  $\|\cdot\|_1$ , and  $\|\cdot\|_\infty$  are more easily computed and hence more practical in many instances.

### 2.3.5 Submultiplicative matrix norms

**Definition 2.22** A matrix norm  $\|\cdot\|_v : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is said to be submultiplicative (consistent) if it also satisfies

$$\|AB\|_v \leq \|A\|_v \|B\|_v.$$

**Theorem 2.23** Let  $\|\cdot\|_v : \mathbb{C}^n \rightarrow \mathbb{R}$  be a vector norm and given any matrix  $C \in \mathbb{C}^{m \times n}$  define the corresponding induced matrix norm as

$$\|C\|_v = \max_{x \neq 0} \frac{\|Cx\|_v}{\|x\|_v} = \max_{\|x\|_v=1} \|Cx\|_v.$$

Then for any  $A \in \mathbb{C}^{m \times k}$  and  $B \in \mathbb{C}^{k \times n}$  the inequality  $\|AB\|_v \leq \|A\|_v \|B\|_v$  holds.

In other words, induced matrix norms are submultiplicative.

To prove the above, it helps to first prove a simpler result:

**Lemma 2.24** Let  $\|\cdot\|_v : \mathbb{C}^n \rightarrow \mathbb{R}$  be a vector norm and given any matrix  $C \in \mathbb{C}^{m \times n}$  define the induced matrix norm as

$$\|C\|_v = \max_{x \neq 0} \frac{\|Cx\|_v}{\|x\|_v} = \max_{\|x\|_v=1} \|Cx\|_v.$$

Then for any  $A \in \mathbb{C}^{m \times n}$  and  $y \in \mathbb{C}^n$  the inequality  $\|Ay\|_v \leq \|A\|_v \|y\|_v$  holds.

**Proof:** If  $y = 0$ , the result obviously holds since then  $\|Ay\|_v = 0$  and  $\|y\|_v = 0$ . Let  $y \neq 0$ . Then

$$\|A\|_v = \max_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v} \geq \frac{\|Ay\|_v}{\|y\|_v}.$$

Rearranging this yields  $\|Ay\|_v \leq \|A\|_v \|y\|_v$ . QED

We can now prove the theorem:

**Proof:**

$$\|AB\|_v = \max_{\|x\|_v=1} \|ABx\|_v = \max_{\|x\|_v=1} \|A(Bx)\|_v \leq \max_{\|x\|_v=1} \|A\|_v \|Bx\|_v \leq \max_{\|x\|_v=1} \|A\|_v \|B\|_v \|x\|_v = \|A\|_v \|B\|_v.$$

QED

**Homework 2.25** Show that  $\|Ax\|_\mu \leq \|A\|_{\mu,v} \|x\|_v$ .

👉 SEE ANSWER

**Homework 2.26** Show that  $\|AB\|_\mu \leq \|A\|_{\mu,v} \|B\|_v$ .

👉 SEE ANSWER

**Homework 2.27** Show that the Frobenius norm,  $\|\cdot\|_F$ , is submultiplicative.

👉 SEE ANSWER

## 2.4 An Application to Conditioning of Linear Systems

A question we will run into later in the course asks how accurate we can expect the solution of a linear system to be if the right-hand side of the system has error in it.

Formally, this can be stated as follows: We wish to solve  $Ax = b$ , where  $A \in \mathbb{C}^{m \times m}$  but the right-hand side has been perturbed by a small vector so that it becomes  $b + \delta b$ . (Notice how that  $\delta$  touches the  $b$ . This is meant to convey that this is a symbol that represents a vector rather than the vector  $b$  that is multiplied by a scalar  $\delta$ .) The question now is how a relative error in  $b$  propagates into a potential error in the solution  $x$ .

This is summarized as follows:

$$\begin{array}{ll} Ax = b & \text{Exact equation} \\ A(x + \delta x) = b + \delta b & \text{Perturbed equation} \end{array}$$

We would like to determine a formula,  $\kappa(A, b, \delta b)$ , that tells us how much a relative error in  $b$  is potentially amplified into an error in the solution  $b$ :

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A, b, \delta b) \frac{\|\delta b\|}{\|b\|}.$$

We will assume that  $A$  has an inverse. To find an expression for  $\kappa(A, b, \delta b)$ , we notice that

$$\begin{array}{rcl} Ax + A\delta x & = & b + \delta b \\ Ax & = & b \quad - \\ \hline A\delta x & = & \delta b \end{array}$$

and from this

$$\begin{array}{l} Ax = b \\ \delta x = A^{-1}\delta b. \end{array}$$

If we now use a vector norm  $\|\cdot\|$  and induced matrix norm  $\|\cdot\|$ , then

$$\begin{array}{l} \|b\| = \|Ax\| \leq \|A\|\|x\| \\ \|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|. \end{array}$$

From this we conclude that

$$\begin{array}{l} \frac{1}{\|x\|} \leq \|A\| \frac{1}{\|b\|} \\ \|\delta x\| \leq \|A^{-1}\|\|\delta b\|. \end{array}$$

so that

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\| \frac{\|\delta b\|}{\|b\|}.$$

Thus, the desired expression  $\kappa(A, b, \delta b)$  doesn't depend on anything but the matrix  $A$ :

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\|A\|\|A^{-1}\|}_{\kappa(A)} \frac{\|\delta b\|}{\|b\|}.$$



$\kappa(A) = \|A\| \|A^{-1}\|$  is called the *condition number* of matrix  $A$ .

A question becomes whether this is a pessimistic result or whether there are examples of  $b$  and  $\delta b$  for which the relative error in  $b$  is amplified by exactly  $\kappa(A)$ . The answer is, unfortunately, “yes!”, as we will show next.

Notice that

- There is an  $\hat{x}$  for which

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \|A\hat{x}\|,$$

namely the  $x$  for which the maximum is attained. Pick  $\hat{b} = A\hat{x}$ .

- There is an  $\hat{\delta b}$  for which

$$\|A^{-1}\| = \max_{\|x\| \neq 0} \frac{\|A^{-1}x\|}{\|x\|} = \frac{\|A^{-1}\hat{\delta b}\|}{\|\hat{\delta b}\|},$$

again, the  $x$  for which the maximum is attained.

It is when solving the perturbed system

$$A(x + \delta x) = \hat{b} + \hat{\delta b}$$

that the maximal magnification by  $\kappa(A)$  is attained.

**Homework 2.28** Let  $\|\cdot\|$  be a matrix norm induced by the  $\|\cdots\|$  vector norm. Show that  $\kappa(A) = \|A\| \|A^{-1}\| \geq 1$ .

 [SEE ANSWER](#)

This last exercise shows that there will always be choices for  $b$  and  $\delta b$  for which the relative error is at best directly translated into an equal relative error in the solution (if  $\kappa(A) = 1$ ).

## 2.5 Equivalence of Norms

Many results we encounter show that the norm of a particular vector or matrix is small. Obviously, it would be unfortunate if a vector or matrix is large in one norm and small in another norm. The following result shows that, modulo a constant, all norms are equivalent. Thus, if the vector is small in one norm, it is small in other norms as well.

**Theorem 2.29** Let  $\|\cdot\|_\mu : \mathbb{C}^n \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^n \rightarrow \mathbb{R}$  be vector norms. Then there exist constants  $\alpha_{\mu,\nu}$  and  $\beta_{\mu,\nu}$  such that for all  $x \in \mathbb{C}^n$

$$\alpha_{\mu,\nu} \|x\|_\mu \leq \|x\|_\nu \leq \beta_{\mu,\nu} \|x\|_\mu.$$

A similar result holds for matrix norms:

**Theorem 2.30** Let  $\|\cdot\|_\mu : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  and  $\|\cdot\|_\nu : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  be matrix norms. Then there exist constants  $\alpha_{\mu,\nu}$  and  $\beta_{\mu,\nu}$  such that for all  $A \in \mathbb{C}^{m \times n}$

$$\alpha_{\mu,\nu} \|A\|_\mu \leq \|A\|_\nu \leq \beta_{\mu,\nu} \|A\|_\mu.$$



# Chapter 3

## Notes on Orthogonality and the Singular Value Decomposition

The reader may wish to review Weeks 9-11 of

[Linear Algebra: Foundations to Frontiers - Notes to LAFF With \[29\]](#).

### Video

Read disclaimer regarding the videos in the preface!

 [YouTube Part 1](#)

 [YouTube Part 2](#)

 [Download Part 1 from UT Box](#)

 [Download Part 2 from UT Box](#)

 [View Part 1 After Local Download](#)

 [View Part 2 After Local Download](#)

(For help on viewing, see Appendix [A](#).)

## Outline

<b>Video</b> . . . . .	<b>35</b>
<b>Outline</b> . . . . .	<b>36</b>
<b>3.1. Orthogonality and Unitary Matrices</b> . . . . .	<b>37</b>
<b>3.2. Toward the Singular Value Decomposition</b> . . . . .	<b>39</b>
<b>3.3. The Singular Value Decomposition Theorem</b> . . . . .	<b>41</b>
<b>3.4. Geometric Interpretation</b> . . . . .	<b>41</b>
<b>3.5. Consequences of the SVD Theorem</b> . . . . .	<b>45</b>
<b>3.6. Projection onto the Column Space</b> . . . . .	<b>49</b>
<b>3.7. Low-rank Approximation of a Matrix</b> . . . . .	<b>50</b>
<b>3.8. An Application</b> . . . . .	<b>51</b>
<b>3.9. SVD and the Condition Number of a Matrix</b> . . . . .	<b>54</b>
<b>3.10. An Algorithm for Computing the SVD?</b> . . . . .	<b>55</b>

---

### 3.1 Orthogonality and Unitary Matrices

**Definition 3.1** Let  $u, v \in \mathbb{C}^m$ . These vectors are orthogonal (perpendicular) if  $u^H v = 0$ .

**Definition 3.2** Let  $q_0, q_1, \dots, q_{n-1} \in \mathbb{C}^m$ . These vectors are said to be mutually orthonormal if for all  $0 \leq i, j < n$

$$q_i^H q_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

Notice that for  $n$  vectors of length  $m$  to be mutually orthonormal,  $n$  must be less than or equal to  $m$ . This is because  $n$  mutually orthonormal vectors are linearly independent and there can be at most  $m$  linearly independent vectors of length  $m$ .

**Definition 3.3** Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Then  $Q$  is said to be an orthonormal matrix if  $Q^H Q = I$  (the identity).

**Homework 3.4** Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Partition  $Q = \left( q_0 \mid q_1 \mid \dots \mid q_{n-1} \right)$ . Show that  $Q$  is an orthonormal matrix if and only if  $q_0, q_1, \dots, q_{n-1}$  are mutually orthonormal.

🔗 [SEE ANSWER](#)

**Definition 3.5** Let  $Q \in \mathbb{C}^{m \times m}$ . Then  $Q$  is said to be a unitary matrix if  $Q^H Q = I$  (the identity).

Notice that unitary matrices are always square and only square matrices can be unitary. Sometimes the term *orthogonal matrix* is used instead of unitary matrix, especially if the matrix is real valued.

**Homework 3.6** Let  $Q \in \mathbb{C}^{m \times m}$ . Show that if  $Q$  is unitary then  $Q^{-1} = Q^H$  and  $QQ^H = I$ .

🔗 [SEE ANSWER](#)

**Homework 3.7** Let  $Q_0, Q_1 \in \mathbb{C}^{m \times m}$  both be unitary. Show that their product,  $Q_0 Q_1$ , is unitary.

🔗 [SEE ANSWER](#)

**Homework 3.8** Let  $Q_0, Q_1, \dots, Q_{k-1} \in \mathbb{C}^{m \times m}$  all be unitary. Show that their product,  $Q_0 Q_1 \dots Q_{k-1}$ , is unitary.

🔗 [SEE ANSWER](#)

The following is a very important observation: Let  $Q$  be a unitary matrix with

$$Q = \left( q_0 \mid q_1 \mid \dots \mid q_{m-1} \right).$$

Let  $x \in \mathbb{C}^m$ . Then

$$\begin{aligned} x &= QQ^H x = \left( q_0 \mid q_1 \mid \dots \mid q_{m-1} \right) \left( q_0 \mid q_1 \mid \dots \mid q_{m-1} \right)^H x \\ &= \left( q_0 \mid q_1 \mid \dots \mid q_{m-1} \right) \begin{pmatrix} q_0^H \\ q_1^H \\ \vdots \\ q_{m-1}^H \end{pmatrix} x \end{aligned}$$

$$\begin{aligned}
&= \left( q_0 \mid q_1 \mid \cdots \mid q_{m-1} \right) \begin{pmatrix} q_0^H x \\ q_1^H x \\ \vdots \\ q_{m-1}^H x \end{pmatrix} \\
&= (q_0^H x)q_0 + (q_1^H x)q_1 + \cdots + (q_{m-1}^H x)q_{m-1}.
\end{aligned}$$

What does this mean?

- The vector  $x = \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{m-1} \end{pmatrix}$  gives the coefficients when the vector  $x$  is written as a linear combination of the unit basis vectors:

$$x = \chi_0 e_0 + \chi_1 e_1 + \cdots + \chi_{m-1} e_{m-1}.$$

- The vector

$$Q^H x = \begin{pmatrix} q_0^H x \\ q_1^H x \\ \vdots \\ q_{m-1}^H x \end{pmatrix}$$

gives the coefficients when the vector  $x$  is written as a linear combination of the orthonormal vectors  $q_0, q_1, \dots, q_{m-1}$ :

$$x = (q_0^H x)q_0 + (q_1^H x)q_1 + \cdots + (q_{m-1}^H x)q_{m-1}.$$

- The vector  $(q_i^H x)q_i$  equals the component of  $x$  in the direction of vector  $q_i$ .

Another way of looking at this is that if  $q_0, q_1, \dots, q_{m-1}$  is an orthonormal basis for  $\mathbb{C}^m$ , then any  $x \in \mathbb{C}^m$  can be written as a linear combination of these vectors:

$$x = \alpha_0 q_0 + \alpha_1 q_1 + \cdots + \alpha_{m-1} q_{m-1}.$$

Now,

$$\begin{aligned}
q_i^H x &= q_i^H (\alpha_0 q_0 + \alpha_1 q_1 + \cdots + \alpha_{i-1} q_{i-1} + \alpha_i q_i + \alpha_{i+1} q_{i+1} + \cdots + \alpha_{m-1} q_{m-1}) \\
&= \alpha_0 \underbrace{q_i^H q_0}_0 + \alpha_1 \underbrace{q_i^H q_1}_0 + \cdots + \alpha_{i-1} \underbrace{q_i^H q_{i-1}}_0 \\
&\quad + \alpha_i \underbrace{q_i^H q_i}_1 + \alpha_{i+1} \underbrace{q_i^H q_{i+1}}_0 + \cdots + \alpha_{m-1} \underbrace{q_i^H q_{m-1}}_0 \\
&= \alpha_i.
\end{aligned}$$

Thus  $q_i^H x = \alpha_i$ , the coefficient that multiplies  $q_i$ .

**Homework 3.9** Let  $U \in \mathbb{C}^{m \times m}$  be unitary and  $x \in \mathbb{C}^m$ , then  $\|Ux\|_2 = \|x\|_2$ .

🔗 SEE ANSWER

**Homework 3.10** Let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary matrices and  $A \in \mathbb{C}^{m \times n}$ . Then

$$\|UA\|_2 = \|AV\|_2 = \|A\|_2.$$

🔗 SEE ANSWER

**Homework 3.11** Let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary matrices and  $A \in \mathbb{C}^{m \times n}$ . Then  $\|UA\|_F = \|AV\|_F = \|A\|_F$ .

🔗 SEE ANSWER

## 3.2 Toward the Singular Value Decomposition

**Lemma 3.12** Given  $A \in \mathbb{C}^{m \times n}$  there exists unitary  $U \in \mathbb{C}^{m \times m}$ , unitary  $V \in \mathbb{C}^{n \times n}$ , and diagonal  $D \in \mathbb{R}^{m \times n}$  such that  $A = UDV^H$  where  $D = \left( \begin{array}{c|c} D_{TL} & 0 \\ \hline 0 & 0 \end{array} \right)$  with  $D_{TL} = \text{diag}(\delta_0, \dots, \delta_{r-1})$  and  $\delta_i > 0$  for  $0 \leq i < r$ .

**Proof:** First, let us observe that if  $A = 0$  (the zero matrix) then the theorem trivially holds:  $A = UDV^H$  where  $U = I_{m \times m}$ ,  $V = I_{n \times n}$ , and  $D = \left( \begin{array}{c|c} \text{---} & 0 \\ \hline 0 & 0 \end{array} \right)$ , so that  $D_{TL}$  is  $0 \times 0$ . Thus, w.l.o.g. assume that  $A \neq 0$ .

We will prove this for  $m \geq n$ , leaving the case where  $m \leq n$  as an exercise, employing a proof by induction on  $n$ .

- **Base case:**  $n = 1$ . In this case  $A = \begin{pmatrix} a_0 \end{pmatrix}$  where  $a_0 \in \mathbb{R}^m$  is its only column. By assumption,  $a_0 \neq 0$ . Then

$$A = \begin{pmatrix} a_0 \end{pmatrix} = \begin{pmatrix} u_0 \end{pmatrix} (\|a_0\|_2) \begin{pmatrix} 1 \end{pmatrix}^H$$

where  $u_0 = a_0 / \|a_0\|_2$ . Choose  $U_1 \in \mathbb{C}^{m \times (m-1)}$  so that  $U = \begin{pmatrix} u_0 & U_1 \end{pmatrix}$  is unitary. Then

$$A = \begin{pmatrix} a_0 \end{pmatrix} = \begin{pmatrix} u_0 \end{pmatrix} (\|a_0\|_2) \begin{pmatrix} 1 \end{pmatrix}^H = \begin{pmatrix} u_0 & U_1 \end{pmatrix} \left( \begin{array}{c|c} \|a_0\|_2 & \\ \hline 0 & \end{array} \right) \begin{pmatrix} 1 \end{pmatrix}^H = UDV^H$$

where  $D_{TL} = \begin{pmatrix} \delta_0 \end{pmatrix} = \begin{pmatrix} \|a_0\|_2 \end{pmatrix}$  and  $V = \begin{pmatrix} 1 \end{pmatrix}$ .

- **Inductive step:** Assume the result is true for all matrices with  $1 \leq k < n$  columns. Show that it is true for matrices with  $n$  columns.

Let  $A \in \mathbb{C}^{m \times n}$  with  $n \geq 2$ . W.l.o.g.,  $A \neq 0$  so that  $\|A\|_2 \neq 0$ . Let  $\delta_0$  and  $v_0 \in \mathbb{C}^n$  have the property that  $\|v_0\|_2 = 1$  and  $\delta_0 = \|Av_0\|_2 = \|A\|_2$ . (In other words,  $v_0$  is the vector that maximizes

$\max_{\|x\|_2=1} \|Ax\|_2$ .) Let  $u_0 = Av_0/\delta_0$ . Note that  $\|u_0\|_2 = 1$ . Choose  $U_1 \in \mathbb{C}^{m \times (m-1)}$  and  $V_1 \in \mathbb{C}^{n \times (n-1)}$  so that  $\tilde{U} = \begin{pmatrix} u_0 & U_1 \end{pmatrix}$  and  $\tilde{V} = \begin{pmatrix} v_0 & V_1 \end{pmatrix}$  are unitary. Then

$$\begin{aligned} \tilde{U}^H A \tilde{V} &= \begin{pmatrix} u_0 & U_1 \end{pmatrix}^H A \begin{pmatrix} v_0 & V_1 \end{pmatrix} \\ &= \begin{pmatrix} u_0^H A v_0 & u_0^H A V_1 \\ U_1^H A v_0 & U_1^H A V_1 \end{pmatrix} = \begin{pmatrix} \delta_0 u_0^H u_0 & u_0^H A V_1 \\ \delta U_1^H u_0 & U_1^H A V_1 \end{pmatrix} = \begin{pmatrix} \delta_0 & w^H \\ 0 & B \end{pmatrix}, \end{aligned}$$

where  $w = V_1^H A^H u_0$  and  $B = U_1^H A V_1$ . Now, we will argue that  $w = 0$ , the zero vector of appropriate size:

$$\begin{aligned} \delta_0^2 = \|A\|_2^2 &= \|U^H A V\|_2^2 = \max_{x \neq 0} \frac{\|U^H A V x\|_2^2}{\|x\|_2^2} = \max_{x \neq 0} \frac{\left\| \begin{pmatrix} \delta_0 & w^H \\ 0 & B \end{pmatrix} x \right\|_2^2}{\|x\|_2^2} \\ &\geq \frac{\left\| \begin{pmatrix} \delta_0 & w^H \\ 0 & B \end{pmatrix} \begin{pmatrix} \delta_0 \\ w \end{pmatrix} \right\|_2^2}{\left\| \begin{pmatrix} \delta_0 \\ w \end{pmatrix} \right\|_2^2} = \frac{\left\| \begin{pmatrix} \delta_0^2 + w^H w \\ B w \end{pmatrix} \right\|_2^2}{\left\| \begin{pmatrix} \delta_0 \\ w \end{pmatrix} \right\|_2^2} \\ &\geq \frac{(\delta_0^2 + w^H w)^2}{\delta_0^2 + w^H w} = \delta_0^2 + w^H w. \end{aligned}$$

Thus  $\delta_0^2 \geq \delta_0^2 + w^H w$  which means that  $w = 0$  and  $\tilde{U}^H A \tilde{V} = \begin{pmatrix} \delta_0 & 0 \\ 0 & B \end{pmatrix}$ .

By the induction hypothesis, there exists unitary  $\check{U} \in \mathbb{C}^{(m-1) \times (m-1)}$ , unitary  $\check{V} \in \mathbb{C}^{(n-1) \times (n-1)}$ , and  $\check{D} \in \mathbb{R}^{(m-1) \times (n-1)}$  such that  $B = \check{U} \check{D} \check{V}^H$  where  $\check{D} = \begin{pmatrix} \check{D}_{TL} & 0 \\ 0 & 0 \end{pmatrix}$  with  $\check{D}_{TL} = \text{diag}(\delta_1, \dots, \delta_{r-1})$ .

Now, let

$$U = \tilde{U} \begin{pmatrix} 1 & 0 \\ 0 & \check{U} \end{pmatrix}, V = \tilde{V} \begin{pmatrix} 1 & 0 \\ 0 & \check{V} \end{pmatrix}, \text{ and } D = \begin{pmatrix} \delta_0 & 0 \\ 0 & \check{D} \end{pmatrix}.$$

(There are some really tough to see "checks" in the definition of  $U$ ,  $V$ , and  $D$ !!) Then  $A = U D V^H$  where  $U$ ,  $V$ , and  $D$  have the desired properties.

- **By the Principle of Mathematical Induction** the result holds for all matrices  $A \in \mathbb{C}^{m \times n}$  with  $m \geq n$ .

**Homework 3.13** Let  $D = \text{diag}(\delta_0, \dots, \delta_{n-1})$ . Show that  $\|D\|_2 = \max_{i=0}^{n-1} |\delta_i|$ .

➡ SEE ANSWER



**Homework 3.14** Let  $A = \begin{pmatrix} A_T \\ 0 \end{pmatrix}$ . Use the SVD of  $A$  to show that  $\|A\|_2 = \|A_T\|_2$ .

🔗 [SEE ANSWER](#)

**Homework 3.15** Assume that  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary matrices. Let  $A, B \in \mathbb{C}^{m \times n}$  with  $B = UAV^H$ . Show that the singular values of  $A$  equal the singular values of  $B$ .

🔗 [SEE ANSWER](#)

**Homework 3.16** Let  $A \in \mathbb{C}^{m \times n}$  with  $A = \left( \begin{array}{c|c} \sigma_0 & 0 \\ \hline 0 & B \end{array} \right)$  and assume that  $\|A\|_2 = \sigma_0$ . Show that  $\|B\|_2 \leq \|A\|_2$ . (Hint: Use the SVD of  $B$ .)

🔗 [SEE ANSWER](#)

**Homework 3.17** Prove Lemma 3.12 for  $m \leq n$ .

🔗 [SEE ANSWER](#)

### 3.3 The Singular Value Decomposition Theorem

**Theorem 3.18 (Singular Value Decomposition)** Given  $A \in \mathbb{C}^{m \times n}$  there exists unitary  $U \in \mathbb{C}^{m \times m}$ , unitary  $V \in \mathbb{C}^{n \times n}$ , and  $\Sigma \in \mathbb{R}^{m \times n}$  such that  $A = U\Sigma V^H$  where  $\Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right)$  with  $\Sigma_{TL} = \text{diag}(\sigma_0, \dots, \sigma_{r-1})$  and  $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_{r-1} > 0$ . The  $\sigma_0, \dots, \sigma_{r-1}$  are known as the singular values of  $A$ .

**Proof:** Notice that the proof of the above theorem is identical to that of Lemma 3.12. However, thanks to the above exercises, we can conclude that  $\|B\|_2 \leq \sigma_0$  in the proof, which then can be used to show that the singular values are found in order.

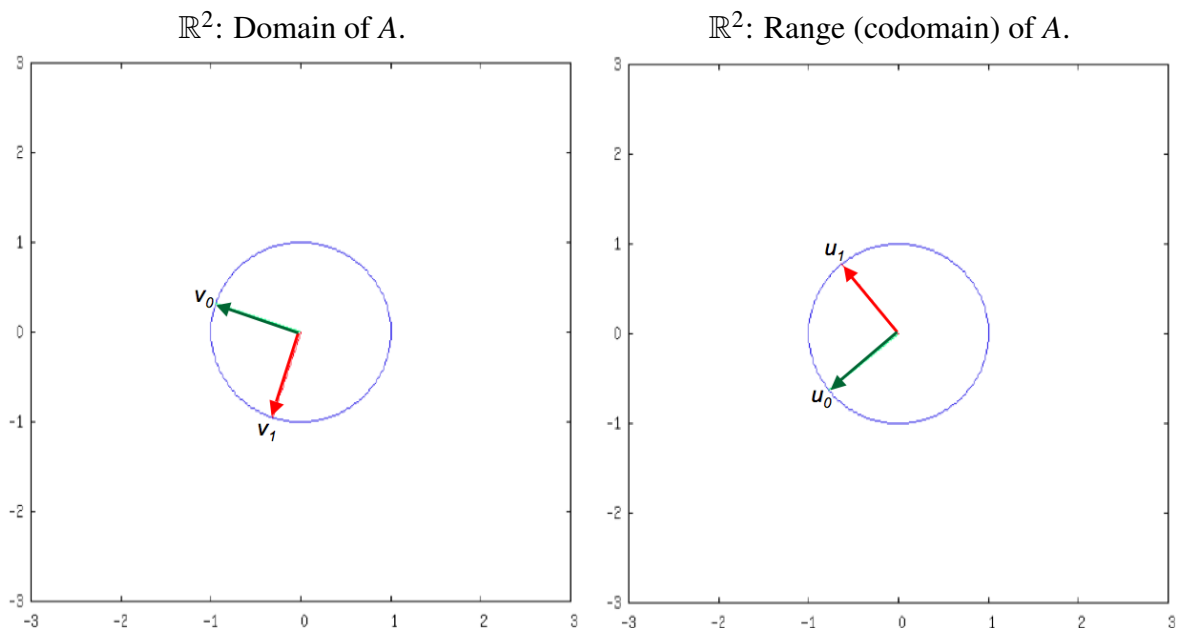
**Proof: (Alternative)** An alternative proof uses Lemma 3.12 to conclude that  $A = UDV^H$ . If the entries on the diagonal of  $D$  are not ordered from largest to smallest, then this can be fixed by permuting the rows and columns of  $D$ , and correspondingly permuting the columns of  $U$  and  $V$ .

### 3.4 Geometric Interpretation

We will now quickly illustrate what the SVD Theorem tells us about matrix-vector multiplication (linear transformations) by examining the case where  $A \in \mathbb{R}^{2 \times 2}$ . Let  $A = U\Sigma V^T$  be its SVD. (Notice that all matrices are now real valued, and hence  $V^H = V^T$ .) Partition

$$A = \left( \begin{array}{c|c} u_0 & u_1 \end{array} \right) \left( \begin{array}{c|c} \sigma_0 & 0 \\ \hline 0 & \sigma_1 \end{array} \right) \left( \begin{array}{c|c} v_0 & v_1 \end{array} \right)^T.$$

Since  $U$  and  $V$  are unitary matrices,  $\{u_0, u_1\}$  and  $\{v_0, v_1\}$  form orthonormal bases for the range and domain of  $A$ , respectively:



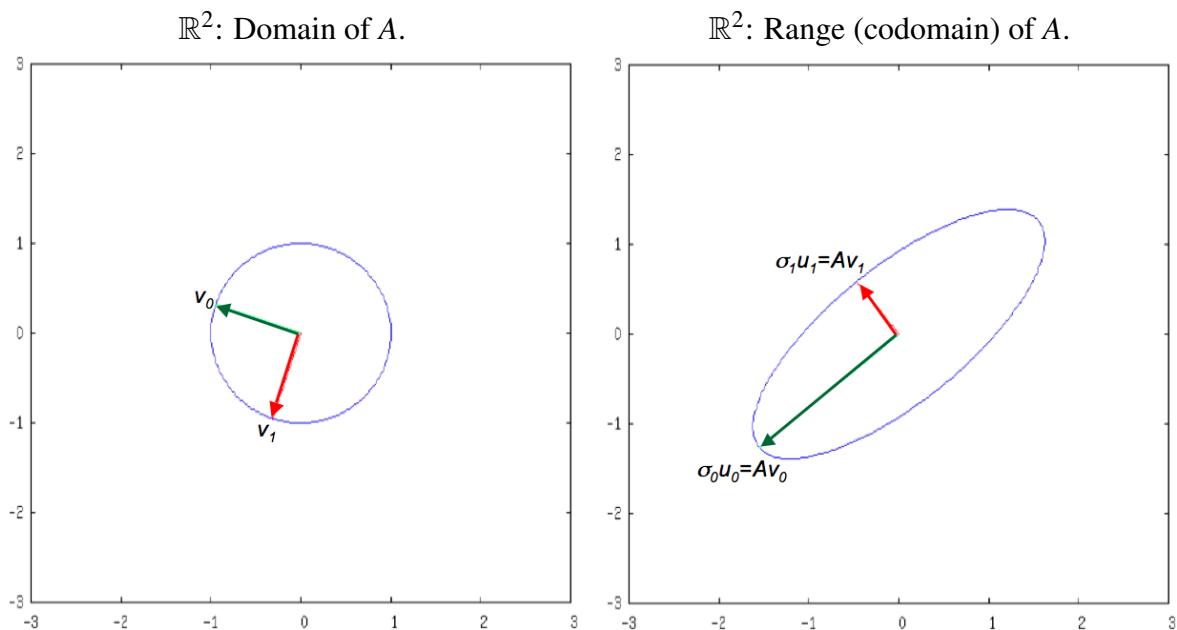
Let us manipulate the decomposition a little:

$$\begin{aligned}
 A &= \begin{pmatrix} u_0 & u_1 \end{pmatrix} \begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix} \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T = \left[ \begin{pmatrix} u_0 & u_1 \end{pmatrix} \begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix} \right] \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T \\
 &= \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T.
 \end{aligned}$$

Now let us look at how  $A$  transforms  $v_0$  and  $v_1$ :

$$Av_0 = \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T v_0 = \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \sigma_0 u_0$$

and similarly  $Av_1 = \sigma_1 u_1$ . This motivates the pictures



Now let us look at how  $A$  transforms any vector with (Euclidean) unit length. Notice that  $x = \begin{pmatrix} \chi_0 \\ \chi_1 \end{pmatrix}$  means that

$$x = \chi_0 e_0 + \chi_1 e_1,$$

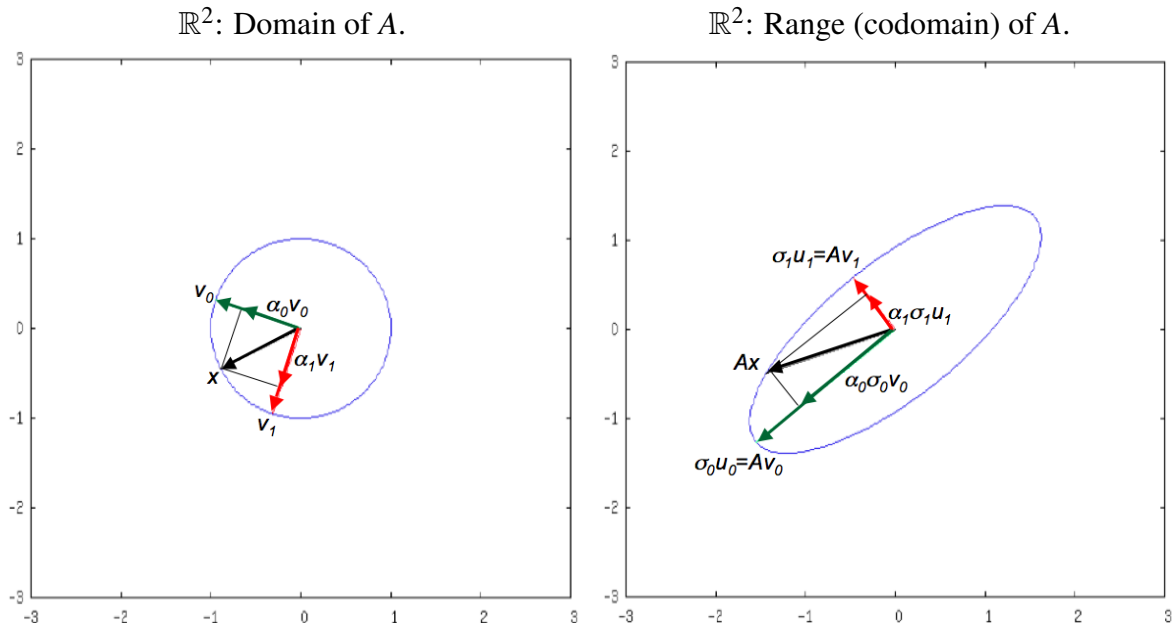
where  $e_0$  and  $e_1$  are the unit basis vectors. Thus,  $\chi_0$  and  $\chi_1$  are the coefficients when  $x$  is expressed using  $e_0$  and  $e_1$  as basis. However, we can also express  $x$  in the basis given by  $v_0$  and  $v_1$ :

$$\begin{aligned} x &= \underbrace{VV^T}_I x = \begin{pmatrix} v_0 & v_1 \end{pmatrix} \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T x = \begin{pmatrix} v_0 & v_1 \end{pmatrix} \begin{pmatrix} \frac{v_0^T x}{v_1^T x} \end{pmatrix} \\ &= \underbrace{v_0^T x}_{\alpha_0} v_0 + \underbrace{v_1^T x}_{\alpha_1} v_1 = \alpha_0 v_0 + \alpha_1 v_1 = \begin{pmatrix} v_0 & v_1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}. \end{aligned}$$

Thus, in the basis formed by  $v_0$  and  $v_1$ , its coefficients are  $\alpha_0$  and  $\alpha_1$ . Now,

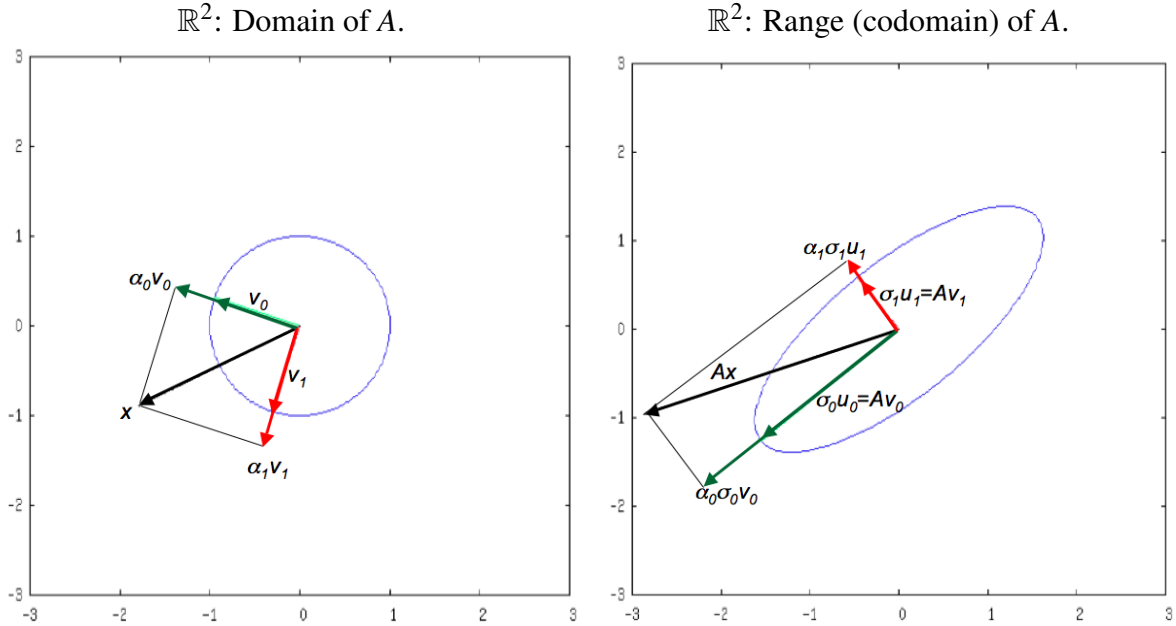
$$\begin{aligned} Ax &= \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T x = \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} v_0 & v_1 \end{pmatrix}^T \begin{pmatrix} v_0 & v_1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_0 u_0 & \sigma_1 u_1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \alpha_0 \sigma_0 u_0 + \alpha_1 \sigma_1 u_1. \end{aligned}$$

This is illustrated by the following picture, which also captures the fact that the unit ball is mapped to an “ellipse”<sup>1</sup> with major axis equal to  $\sigma_0 = \|A\|_2$  and minor axis equal to  $\sigma_1$ :



<sup>1</sup>It is not clear that it is actually an ellipse and this is not important to our observations.

Finally, we show the same insights for general vector  $x$  (not necessarily of unit length).



Another observation is that *if* one picks the right basis for the domain and codomain, then the computation  $Ax$  simplifies to a matrix multiplication with a diagonal matrix. Let us again illustrate this for nonsingular  $A \in \mathbb{R}^{2 \times 2}$  with

$$A = \underbrace{\begin{pmatrix} u_0 & u_1 \end{pmatrix}}_U \underbrace{\begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix}}_{\Sigma} \underbrace{\begin{pmatrix} v_0 & v_1 \end{pmatrix}^T}_V.$$

Now, if we chose to express  $y$  using  $u_0$  and  $u_1$  as the basis and express  $x$  using  $v_0$  and  $v_1$  as the basis, then

$$\begin{aligned} \hat{y} &= \underbrace{UU^T}_I y = (u_0^T y)u_0 + (u_1^T y)u_1 = \begin{pmatrix} \hat{\psi}_0 \\ \hat{\psi}_1 \end{pmatrix} \\ \hat{x} &= \underbrace{VV^T}_I x = (v_0^T x)v_0 + (v_1^T x)v_1 = \begin{pmatrix} \hat{\chi}_0 \\ \hat{\chi}_1 \end{pmatrix}. \end{aligned}$$

If  $y = Ax$  then

$$\underbrace{U^T y}_{\hat{y}} = \underbrace{U^T Ax}_{Ax} = U^T \Sigma V^T x = U^T \Sigma \hat{x}$$

so that  $\hat{y} = \Sigma \hat{x}$  and

$$\begin{pmatrix} \hat{\psi}_0 \\ \hat{\psi}_1 \end{pmatrix} = \begin{pmatrix} \sigma_0 \hat{\chi}_0 \\ \sigma_1 \hat{\chi}_1 \end{pmatrix}.$$

These observation generalize to  $A \in \mathbb{C}^{m \times m}$ .

### 3.5 Consequences of the SVD Theorem

Throughout this section we will assume that

- $A = U\Sigma V^H$  is the SVD of  $A \in \mathbb{C}^{m \times n}$ , with  $U$  and  $V$  unitary and  $\Sigma$  diagonal.
- $\Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right)$  where  $\Sigma_{TL} = \text{diag}(\sigma_0, \dots, \sigma_{r-1})$  with  $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_{r-1} > 0$ .
- $U = \left( \begin{array}{c|c} U_L & U_R \end{array} \right)$  with  $U_L \in \mathbb{C}^{m \times r}$ .
- $V = \left( \begin{array}{c|c} V_L & V_R \end{array} \right)$  with  $V_L \in \mathbb{C}^{n \times r}$ .

We first generalize the observations we made for  $A \in \mathbb{R}^{2 \times 2}$ . Let us track what the effect of  $Ax = U\Sigma V^H x$  is on vector  $x$ . We assume that  $m \geq n$ .

- Let  $U = \left( \begin{array}{c|c|c} u_0 & \dots & u_{m-1} \end{array} \right)$  and  $V = \left( \begin{array}{c|c|c} v_0 & \dots & v_{n-1} \end{array} \right)$ .
- Let

$$\begin{aligned} x &= VV^H x = \left( \begin{array}{c|c|c} v_0 & \dots & v_{n-1} \end{array} \right) \left( \begin{array}{c|c|c} v_0 & \dots & v_{n-1} \end{array} \right)^H x = \left( \begin{array}{c|c|c} v_0 & \dots & v_{n-1} \end{array} \right) \left( \begin{array}{c} v_0^H x \\ \vdots \\ v_{n-1}^H x \end{array} \right) \\ &= v_0^H x v_0 + \dots + v_{n-1}^H x v_{n-1}. \end{aligned}$$

This can be interpreted as follows: vector  $x$  can be written in terms of the usual basis of  $\mathbb{C}^n$  as  $\chi_0 e_0 + \dots + \chi_1 e_{n-1}$  or in the orthonormal basis formed by the columns of  $V$  as  $v_0^H x v_0 + \dots + v_{n-1}^H x v_{n-1}$ .

- Notice that  $Ax = A(v_0^H x v_0 + \dots + v_{n-1}^H x v_{n-1}) = v_0^H x A v_0 + \dots + v_{n-1}^H x A v_{n-1}$  so that we next look at how  $A$  transforms each  $v_i$ :  $Av_i = U\Sigma V^H v_i = U\Sigma e_i = \sigma_i U e_i = \sigma_i u_i$ .
- Thus, another way of looking at  $Ax$  is

$$\begin{aligned} Ax &= v_0^H x A v_0 + \dots + v_{n-1}^H x A v_{n-1} \\ &= v_0^H x \sigma_0 u_0 + \dots + v_{n-1}^H x \sigma_{n-1} u_{n-1} \\ &= \sigma_0 u_0 v_0^H x + \dots + \sigma_{n-1} u_{n-1} v_{n-1}^H x \\ &= (\sigma_0 u_0 v_0^H + \dots + \sigma_{n-1} u_{n-1} v_{n-1}^H) x. \end{aligned}$$

**Corollary 3.19**  $A = U_L \Sigma_{TL} V_L^H$ . This is called the reduced SVD of  $A$ .

**Proof:**

$$A = U\Sigma V^H = \left( \begin{array}{c|c} U_L & U_R \end{array} \right) \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right) \left( \begin{array}{c|c} V_L & V_R \end{array} \right)^H = U_L \Sigma_{TL} V_L^H.$$

**Corollary 3.20** Let  $A = U_L \Sigma_{TL} V_L^H$  be the reduced SVD with

$$U_L = \left( u_0 \mid \cdots \mid u_{r-1} \right), \quad \Sigma_{TL} = \text{diag}(\sigma_0, \dots, \sigma_{r-1}), \quad \text{and} \quad V_L = \left( v_0 \mid \cdots \mid v_{r-1} \right).$$

Then

$$A = \underbrace{\sigma_0 u_0 v_0^H}_{\sigma_0} + \underbrace{\sigma_1 u_1 v_1^H}_{\sigma_1} + \cdots + \underbrace{\sigma_{r-1} u_{r-1} v_{r-1}^H}_{\sigma_{r-1}}.$$

(Each term nonzero and an outer product, and hence a rank-1 matrix.)

**Proof:** We leave the proof as an exercise.

**Corollary 3.21**  $C(A) = C(U_L)$ .

**Proof:**

- Let  $y \in C(A)$ . Then there exists  $x \in \mathbb{C}^n$  such that  $y = Ax$  (by the definition of  $y \in C(A)$ ). But then

$$y = Ax = U_L \underbrace{\Sigma_{TL} V_L^H x}_z = U_L z,$$

i.e., there exists  $z \in \mathbb{C}^r$  such that  $y = U_L z$ . This means  $y \in C(U_L)$ .

- Let  $y \in C(U_L)$ . Then there exists  $z \in \mathbb{C}^r$  such that  $y = U_L z$ . But then

$$y = U_L z = U_L \underbrace{\Sigma_{TL} \Sigma_{TL}^{-1}}_I z = U_L \Sigma_{TL} \underbrace{V_L^H V_L}_I \Sigma_{TL}^{-1} z = A \underbrace{V_L \Sigma_{TL}^{-1} z}_x = Ax$$

so that there exists  $x \in \mathbb{C}^n$  such that  $y = Ax$ , i.e.,  $y \in C(A)$ .

**Corollary 3.22** Let  $A = U_L \Sigma_{TL} V_L^H$  be the reduced SVD of  $A$  where  $U_L$  and  $V_L$  have  $r$  columns. Then the rank of  $A$  is  $r$ .

**Proof:** The rank of  $A$  equals the dimension of  $C(A) = C(U_L)$ . But the dimension of  $C(U_L)$  is clearly  $r$ .

**Corollary 3.23**  $\mathcal{N}(A) = C(V_R)$ .

**Proof:**

- Let  $x \in \mathcal{N}(A)$ . Then

$$\begin{aligned} x &= \underbrace{V V^H}_I x = \left( V_L \mid V_R \right) \left( V_L \mid V_R \right)^H x = \left( V_L \mid V_R \right) \left( \frac{V_L^H x}{V_R^H x} \right) \\ &= \left( V_L \mid V_R \right) \left( \frac{V_L^H x}{V_R^H x} \right) = V_L V_L^H x + V_R V_R^H x. \end{aligned}$$

If we can show that  $V_L^H x = 0$  then  $x = V_R z$  where  $z = V_R^H x$ . Assume that  $V_L^H x \neq 0$ . Then  $\Sigma_{TL}(V_L^H x) \neq 0$  (since  $\Sigma_{TL}$  is nonsingular) and  $U_L(\Sigma_{TL}(V_L^H x)) \neq 0$  (since  $U_L$  has linearly independent columns). But that contradicts the fact that  $Ax = U_L \Sigma_{TL} V_L^H x = 0$ .

- Let  $x \in C(V_R)$ . Then  $x = V_R z$  for some  $z \in \mathbb{C}^r$  and  $Ax = U_L \Sigma_{TL} \underbrace{V_L^H V_R}_{0} z = 0$ .

**Corollary 3.24** For all  $x \in \mathbb{C}^n$  there exists  $z \in C(V_L)$  such that  $Ax = Az$ .

**Proof:**

$$\begin{aligned}
 Ax &= A \underbrace{VV^H}_I x = A \left( \begin{array}{c|c} V_L & V_R \end{array} \right) \left( \begin{array}{c|c} V_L & V_R \end{array} \right)^H x \\
 &= A (V_L V_L^H x + V_R V_R^H x) = AV_L V_L^H x + AV_R V_R^H x \\
 &= AV_L V_L^H x + U_L \Sigma_{TL} \underbrace{V_L^H V_R}_{0} V_R^H x = A \underbrace{V_L V_L^H}_z x.
 \end{aligned}$$

Alternative proof (which uses the last corollary):

$$Ax = A (V_L V_L^H x + V_R V_R^H x) = AV_L V_L^H x + A \underbrace{V_R V_R^H x}_{\in \mathcal{N}(A)} = A \underbrace{V_L V_L^H}_z x.$$

The proof of the last corollary also shows that

**Corollary 3.25** Any vector  $x \in \mathbb{C}^n$  can be written as  $x = z + x_n$  where  $z \in C(V_L)$  and  $x_n \in \mathcal{N}(A) = C(V_R)$ .

**Corollary 3.26**  $A^H = V_L \Sigma_{TL} U_L^H$  so that  $C(A^H) = C(V_L)$  and  $\mathcal{N}(A^H) = C(U_R)$ .

The above corollaries are summarized in Figure 3.1.

**Theorem 3.27** Let  $A \in \mathbb{C}^{n \times n}$  be nonsingular. Let  $A = U \Sigma V^H$  be its SVD. Then

1. The SVD is the reduced SVD.
2.  $\sigma_{n-1} \neq 0$ .
3. If

$$U = \left( \begin{array}{c|c|c} u_0 & \cdots & u_{n-1} \end{array} \right), \Sigma = \text{diag}(\sigma_0, \dots, \sigma_{n-1}), \text{ and } V = \left( \begin{array}{c|c|c} v_0 & \cdots & v_{n-1} \end{array} \right),$$

then

$$A^{-1} = (V P^T)(P \Sigma^{-1} P^T)(U P^T)^H = \left( \begin{array}{c|c|c} v_{n-1} & \cdots & v_0 \end{array} \right) \text{diag}\left(\frac{1}{\sigma_{n-1}}, \dots, \frac{1}{\sigma_0}\right) \left( \begin{array}{c|c|c} u_{n-1} & \cdots & u_0 \end{array} \right),$$

$$\text{where } P = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{pmatrix} \text{ is the permutation matrix such that } Px \text{ reverses the order of the entries}$$

in  $x$ . (Note: for this permutation matrix,  $P^T = P$ . In general, this is not the case. What is the case for all permutation matrices  $P$  is that  $P^T P = P P^T = I$ .)

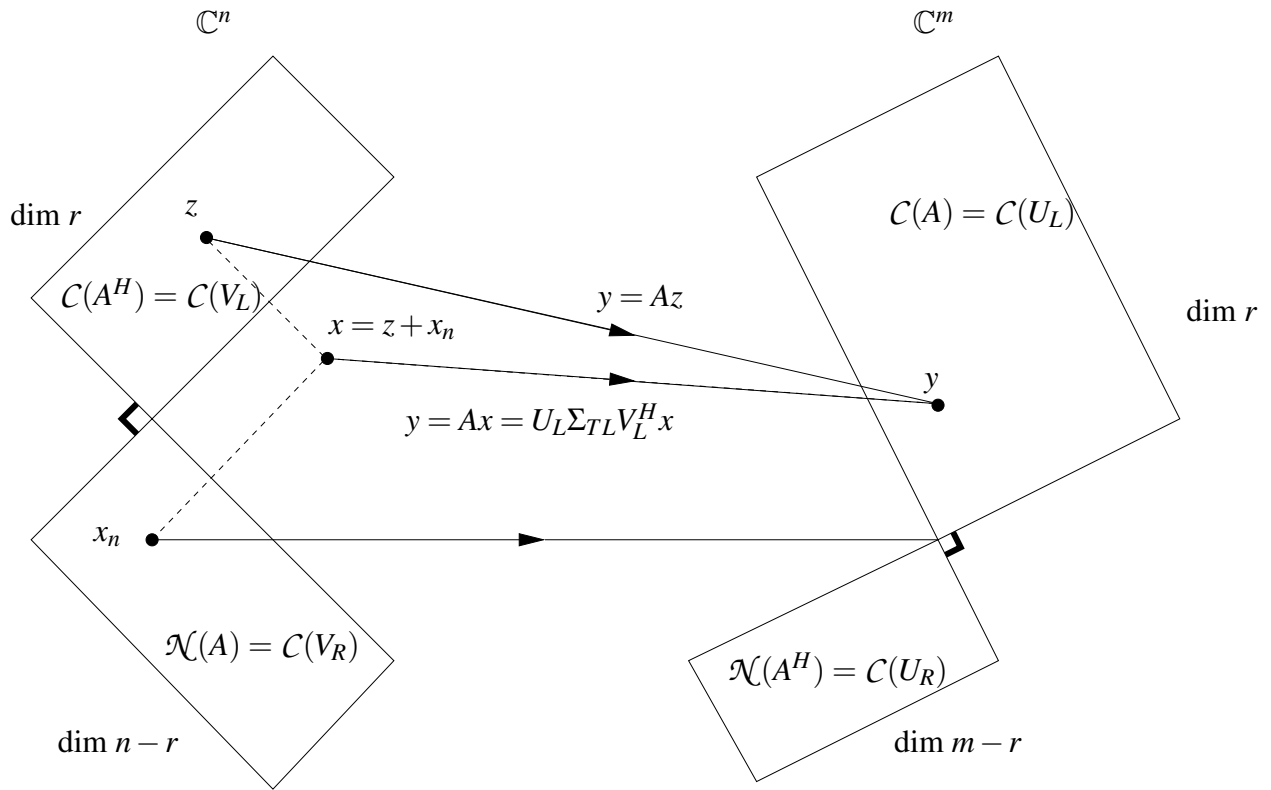


Figure 3.1: A pictorial description of how  $x = z + x_n$  is transformed by  $A \in \mathbb{C}^{m \times n}$  into  $y = Ax = A(z + x_n)$ . We see that  $C(V_L)$  and  $C(V_R)$  are orthogonal complements of each other within  $\mathbb{C}^n$ . Similarly,  $C(U_L)$  and  $C(U_R)$  are orthogonal complements of each other within  $\mathbb{C}^m$ . Any vector  $x$  can be written as the sum of a vector  $z \in C(V_L)$  and  $x_n \in C(V_R) = \mathcal{N}(A)$ .

4.  $\|A^{-1}\|_2 = 1/\sigma_{n-1}$ .

**Proof:** The only item that is less than totally obvious is (3). Clearly  $A^{-1} = V\Sigma^{-1}U^H$ . The problem is that in  $\Sigma^{-1}$  the diagonal entries are not ordered from largest to smallest. The permutation fixes this.

**Corollary 3.28** If  $A \in \mathbb{C}^{m \times n}$  has linearly independent columns then  $A^H A$  is invertible (nonsingular) and  $(A^H A)^{-1} = V_L (\Sigma_{TL}^2)^{-1} V_L^H$ .

**Proof:** Since  $A$  has linearly independent columns,  $A = U_L \Sigma_{TL} V_L^H$  is the reduced SVD where  $U_L$  has  $n$  columns and  $V_L$  is unitary. Hence

$$A^H A = (U_L \Sigma_{TL} V_L^H)^H U_L \Sigma_{TL} V_L^H = V_L \Sigma_{TL}^H U_L^H U_L \Sigma_{TL} V_L^H = V_L \Sigma_{TL} \Sigma_{TL} V_L^H = V_L \Sigma_{TL}^2 V_L^H.$$

Since  $V_L$  is unitary and  $\Sigma_{TL}$  is diagonal with nonzero diagonal entries, they are both nonsingular. Thus

$$(V_L \Sigma_{TL}^2 V_L^H) (V_L (\Sigma_{TL}^2)^{-1} V_L^H) = I.$$

This means  $A^T A$  is invertible and  $(A^T A)^{-1}$  is as given.



### 3.6 Projection onto the Column Space

**Definition 3.29** Let  $U_L \in \mathbb{C}^{m \times k}$  have orthonormal columns. The projection of a vector  $y \in \mathbb{C}^m$  onto  $C(U_L)$  is the vector  $U_L x$  that minimizes  $\|y - U_L x\|_2$ , where  $x \in \mathbb{C}^k$ . We will also call this vector  $y$  the component of  $x$  in  $C(U_L)$ .

**Theorem 3.30** Let  $U_L \in \mathbb{C}^{m \times k}$  have orthonormal columns. The projection of  $y$  onto  $C(U_L)$  is given by  $U_L U_L^H y$ .

**Proof:** The vector  $U_L x$  that we want must satisfy

$$\|U_L x - y\|_2 = \min_{w \in \mathbb{C}^k} \|U_L w - y\|_2.$$

Now, the 2-norm is invariant under multiplication by the unitary matrix  $U^H = \begin{pmatrix} U_L & U_R \end{pmatrix}^H$

$$\begin{aligned} \|U_L x - y\|_2^2 &= \min_{w \in \mathbb{C}^k} \|U_L w - y\|_2^2 \\ &= \min_{w \in \mathbb{C}^k} \|U^H (U_L w - y)\|_2^2 \quad (\text{since the two norm is preserved}) \\ &= \min_{w \in \mathbb{C}^k} \left\| \begin{pmatrix} U_L & U_R \end{pmatrix}^H (U_L w - y) \right\|_2^2 \\ &= \min_{w \in \mathbb{C}^k} \left\| \begin{pmatrix} U_L^H \\ U_R^H \end{pmatrix} (U_L w - y) \right\|_2^2 \\ &= \min_{w \in \mathbb{C}^k} \left\| \begin{pmatrix} U_L^H \\ U_R^H \end{pmatrix} U_L w - \begin{pmatrix} U_L^H \\ U_R^H \end{pmatrix} y \right\|_2^2 \\ &= \min_{w \in \mathbb{C}^k} \left\| \begin{pmatrix} U_L^H U_L w \\ U_R^H U_L w \end{pmatrix} - \begin{pmatrix} U_L^H y \\ U_R^H y \end{pmatrix} \right\|_2^2 \\ &= \min_{w \in \mathbb{C}^k} \left\| \begin{pmatrix} w \\ 0 \end{pmatrix} - \begin{pmatrix} U_L^H y \\ U_R^H y \end{pmatrix} \right\|_2^2 \\ &= \min_{w \in \mathbb{C}^k} \left\| \begin{pmatrix} w - U_L^H y \\ -U_R^H y \end{pmatrix} \right\|_2^2 \\ &= \min_{w \in \mathbb{C}^k} \left( \|w - U_L^H y\|_2^2 + \|-U_R^H y\|_2^2 \right) \quad (\text{since } \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_2^2 = \|u\|_2^2 + \|v\|_2^2) \\ &= \left( \min_{w \in \mathbb{C}^k} \|w - U_L^H y\|_2^2 \right) + \|U_R^H y\|_2^2. \end{aligned}$$

This is minimized when  $w = U_L^H y$ . Thus, the vector that is closest to  $y$  in the space spanned by  $U_L$  is given by  $x = U_L U_L^H y$ .

**Corollary 3.31** Let  $A \in \mathbb{C}^{m \times n}$  and  $A = U_L \Sigma_{TL} V_L^H$  be its reduced SVD. Then the projection of  $y \in \mathbb{C}^m$  onto  $C(A)$  is given by  $U_L U_L^H y$ .

**Proof:** This follows immediately from the fact that  $C(A) = C(U_L)$ .

**Corollary 3.32** Let  $A \in \mathbb{C}^{m \times n}$  have linearly independent columns. Then the projection of  $y \in \mathbb{C}^m$  onto  $C(A)$  is given by  $A(A^H A)^{-1} A^H y$ .

**Proof:** From Corollary 3.28, we know that  $A^H A$  is nonsingular and that  $(A^H A)^{-1} = V_L (\Sigma_{TL}^2)^{-1} V_L^H$ . Now,

$$\begin{aligned} A(A^H A)^{-1} A^H y &= (U_L \Sigma_{TL} V_L^H) (V_L (\Sigma_{TL}^2)^{-1} V_L^H) (U_L \Sigma_{TL} V_L^H)^H y \\ &= U_L \Sigma_{TL} \underbrace{V_L^H V_L}_I \Sigma_{TL}^{-1} \Sigma_{TL}^{-1} \underbrace{V_L^H V_L}_I \Sigma_{TL} U_L^H y = U_L U_L^H y. \end{aligned}$$

Hence the projection of  $y$  onto  $C(A)$  is given by  $A(A^H A)^{-1} A^H y$ .

**Definition 3.33** Let  $A$  have linearly independent columns. Then  $(A^H A)^{-1} A^H$  is called the pseudo-inverse or Moore-Penrose generalized inverse of matrix  $A$ .

### 3.7 Low-rank Approximation of a Matrix

**Theorem 3.34** Let  $A \in \mathbb{C}^{m \times n}$  have SVD  $A = U \Sigma V^H$  and assume  $A$  has rank  $r$ . Partition

$$U = \left( U_L \mid U_R \right), \quad V = \left( V_L \mid V_R \right), \quad \text{and} \quad \Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & \Sigma_{BR} \end{array} \right),$$

where  $U_L \in \mathbb{C}^{m \times k}$ ,  $V_L \in \mathbb{C}^{n \times k}$ , and  $\Sigma_{TL} \in \mathbb{R}^{k \times k}$  with  $k \leq r$ . Then  $B = U_L \Sigma_{TL} V_L^H$  is the matrix in  $\mathbb{C}^{m \times n}$  closest to  $A$  in the following sense:

$$\|A - B\|_2 = \min_{\substack{C \in \mathbb{C}^{m \times n} \\ \text{rank}(C) \leq k}} \|A - C\|_2 = \sigma_k.$$

**Proof:** First, if  $B$  is as defined, then clearly  $\|A - B\|_2 = \sigma_k$ :

$$\begin{aligned} \|A - B\|_2 &= \|U^H (A - B) V\|_2 = \|U^H A V - U^H B V\|_2 \\ &= \left\| \Sigma - \left( U_L \mid U_R \right)^H B \left( V_L \mid V_R \right) \right\|_2 = \left\| \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & \Sigma_{BR} \end{array} \right) - \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right) \right\|_2 \\ &= \left\| \left( \begin{array}{c|c} 0 & 0 \\ \hline 0 & \Sigma_{BR} \end{array} \right) \right\|_2 = \|\Sigma_{BR}\|_2 = \sigma_k \end{aligned}$$

Next, assume that  $C$  has rank  $t \leq k$  and  $\|A - C\|_2 < \|A - B\|_2$ . We will show that this leads to a contradiction.

- The null space of  $C$  has dimension at least  $n - k$  since  $\dim(\mathcal{N}(C)) + \text{rank}(C) = n$ .
- If  $x \in \mathcal{N}(C)$  then

$$\|Ax\|_2 = \|(A - C)x\|_2 \leq \|A - C\|_2 \|x\|_2 < \sigma_k \|x\|_2.$$

- Partition  $U = \left( \begin{array}{c|c|c} u_0 & \cdots & u_{m-1} \end{array} \right)$  and  $V = \left( \begin{array}{c|c|c} v_0 & \cdots & v_{n-1} \end{array} \right)$ . Then  $\|Av_j\|_2 = \|\sigma_j u_j\|_2 = \sigma_j \geq \sigma_s$  for  $j = 0, \dots, k$ . Now, let  $x$  be any linear combination of  $v_0, \dots, v_k$ :  $x = \alpha_0 v_0 + \cdots + \alpha_k v_k$ . Notice that

$$\|x\|_2^2 = \|\alpha_0 v_0 + \cdots + \alpha_k v_k\|_2^2 \leq |\alpha_0|^2 + \cdots + |\alpha_k|^2.$$

Then

$$\begin{aligned} \|Ax\|_2^2 &= \|A(\alpha_0 v_0 + \cdots + \alpha_k v_k)\|_2^2 = \|\alpha_0 A v_0 + \cdots + \alpha_k A v_k\|_2^2 \\ &= \|\alpha_0 \sigma_0 u_0 + \cdots + \alpha_k \sigma_k u_k\|_2^2 = \|\alpha_0 \sigma_0 u_0\|_2^2 + \cdots + \|\alpha_k \sigma_k u_k\|_2^2 \\ &= |\alpha_0|^2 \sigma_0^2 + \cdots + |\alpha_k|^2 \sigma_k^2 \geq (|\alpha_0|^2 + \cdots + |\alpha_k|^2) \sigma_k^2 \end{aligned}$$

so that  $\|Ax\|_2 \geq \sigma_k \|x\|_2$ . In other words, vectors in the subspace of all linear combinations of  $\{v_0, \dots, v_k\}$  satisfy  $\|Ax\|_2 \geq \sigma_k \|x\|_2$ . The dimension of this subspace is  $k + 1$  (since  $\{v_0, \dots, v_k\}$  form an orthonormal basis).

- Both these subspaces are subspaces of  $\mathbb{C}^n$ . Since their dimensions add up to more than  $n$  there must be at least one nonzero vector  $z$  that satisfies both  $\|Az\|_2 < \sigma_k \|z\|_2$  and  $\|Az\|_2 \geq \sigma_k \|z\|_2$ , which is a contradiction.

The above theorem tells us how to pick the best approximation with given rank to a given matrix.

## 3.8 An Application

Let  $Y \in \mathbb{R}^{m \times n}$  be a matrix that, for example, stores a picture. In this case, the  $(i, j)$  entry in  $Y$  is, for example, a number that represents the grayscale value of pixel  $(i, j)$ . The following instructions, executed in octave or matlab, generate the picture of Mexican artist Frida Kahlo in Figure 3.2(top-left). The file `FridaPNG.png` can be found at <http://www.cs.utexas.edu/users/flame/Notes/FridaPNG.png>.

```
octave> IMG = imread( 'FridaPNG.png' ); % this reads the image
octave> Y = IMG( :, :, 1 );
octave> imshow( Y ) % this displays the image
```

Although the picture is black and white, it was read as if it is a color image, which means a  $m \times n \times 3$  array of pixel information is stored. Setting  $Y = \text{IMG}( :, :, 1 )$  extracts a single matrix of pixel information. (If you start with a color picture, you will want to approximate  $\text{IMG}( :, :, 1 )$ ,  $\text{IMG}( :, :, 2 )$ , and  $\text{IMG}( :, :, 3 )$  separately.)

Now, let  $Y = U\Sigma V^T$  be the SVD of matrix  $Y$ . Partition, conformally,

$$U = \left( \begin{array}{c|c} U_L & U_R \end{array} \right), \quad V = \left( \begin{array}{c|c} V_L & V_R \end{array} \right), \quad \text{and} \quad \Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & \Sigma_{BR} \end{array} \right),$$

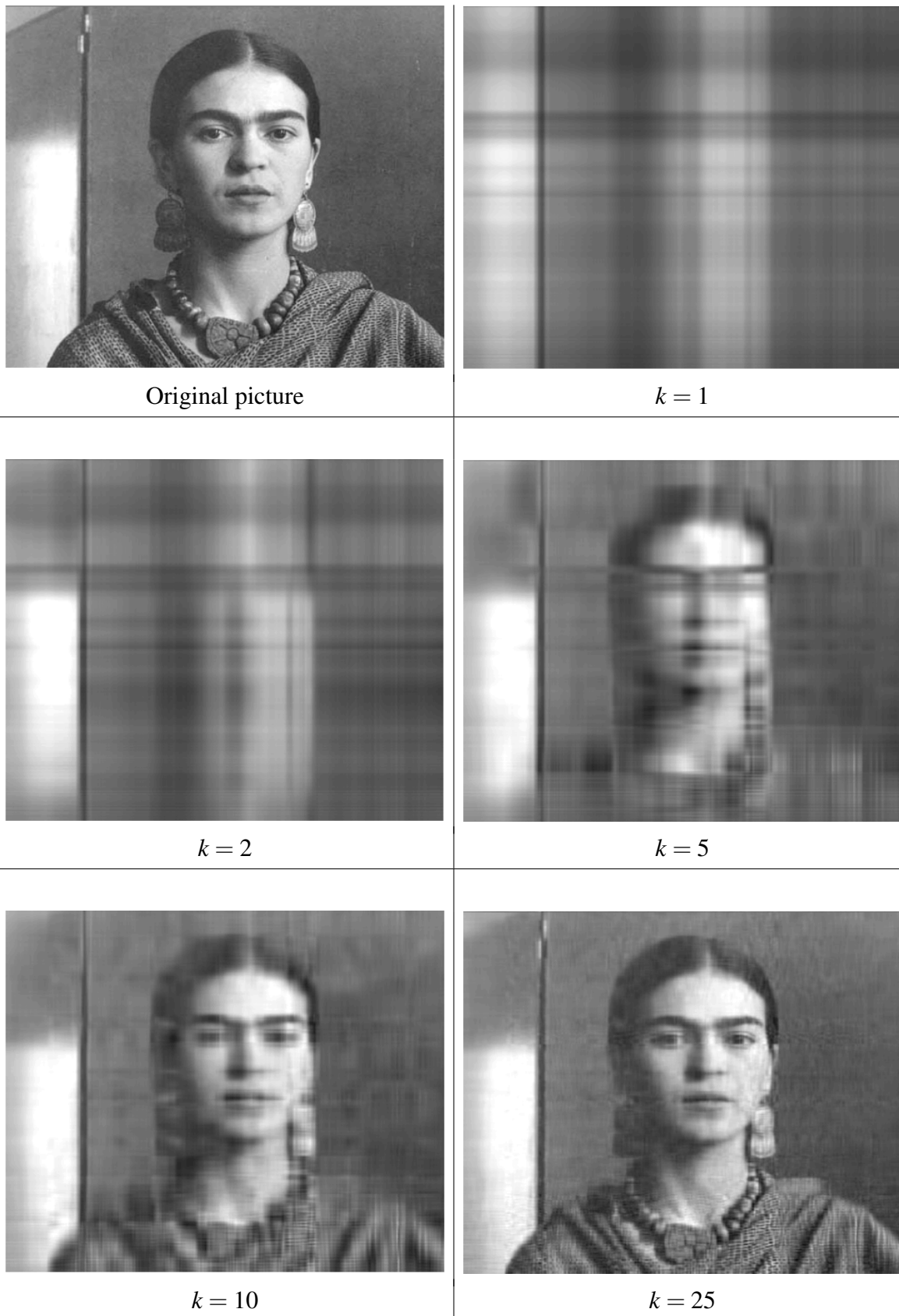


Figure 3.2: Multiple pictures as generated by the code

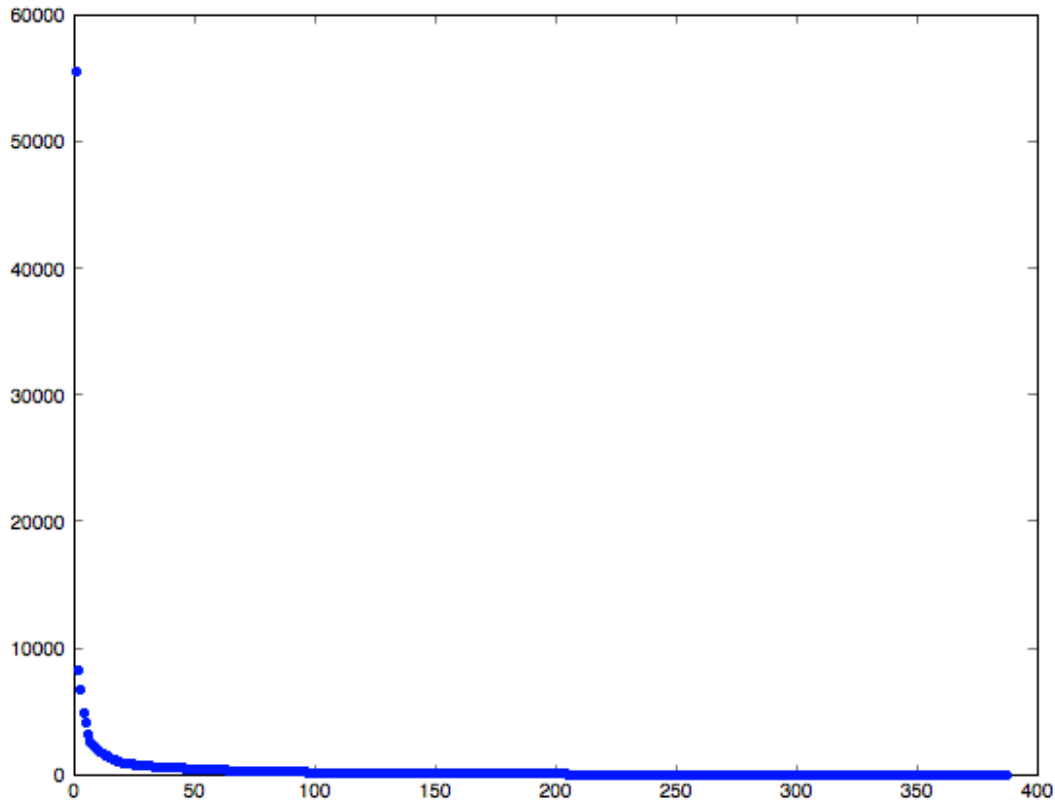


Figure 3.3: Distribution of singular values for the picture.

where  $U_L$  and  $V_L$  have  $k$  columns and  $\Sigma_{TL}$  is  $k \times k$ . so that

$$\begin{aligned}
 Y &= \begin{pmatrix} U_L & U_R \end{pmatrix} \begin{pmatrix} \Sigma_{TL} & 0 \\ 0 & \Sigma_{BR} \end{pmatrix} \begin{pmatrix} V_L & V_R \end{pmatrix}^T \\
 &= \begin{pmatrix} U_L & U_R \end{pmatrix} \begin{pmatrix} \Sigma_{TL} & 0 \\ 0 & \Sigma_{BR} \end{pmatrix} \begin{pmatrix} V_L^T \\ V_R^T \end{pmatrix} \\
 &= \begin{pmatrix} U_L & U_R \end{pmatrix} \begin{pmatrix} \Sigma_{TL} V_L^T \\ \Sigma_{BR} V_R^T \end{pmatrix} \\
 &= U_L \Sigma_{TL} V_L^T + U_R \Sigma_{BR} V_R^T.
 \end{aligned}$$

Recall that then  $U_L \Sigma_{TL} V_L^T$  is the best rank- $k$  approximation to  $Y$ .

Let us approximate the matrix that stores the picture with  $U_L \Sigma_{TL} V_L^T$ :

```

>> IMG = imread( 'FridaPNG.png' ); % read the picture
>> Y = IMG( :, :, 1 );
>> imshow( Y ); % this dispays the image
>> k = 1;
>> [ U, Sigma, V ] = svd( Y );

```

```

>> UL = U( :, 1:k );           % first k columns
>> VL = V( :, 1:k );           % first k columns
>> SigmaTL = Sigma( 1:k, 1:k ); % TL submatrix of Sigma
>> Yapprox = uint8( UL * SigmaTL * VL' );
>> imshow( Yapprox );

```

As one increases  $k$ , the approximation gets better, as illustrated in Figure 3.2. The graph in Figure 3.3 helps explain. The original matrix  $Y$  is  $387 \times 469$ , with 181,503 entries. When  $k = 10$ , matrices  $U$ ,  $V$ , and  $\Sigma$  are  $387 \times 10$ ,  $469 \times 10$  and  $10 \times 10$ , respectively, requiring only 8,660 entries to be stores.

### 3.9 SVD and the Condition Number of a Matrix

In “Notes on Norms” we saw that if  $Ax = b$  and  $A(x + \delta x) = b + \delta b$ , then

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \kappa_2(A) \frac{\|\delta b\|_2}{\|b\|_2},$$

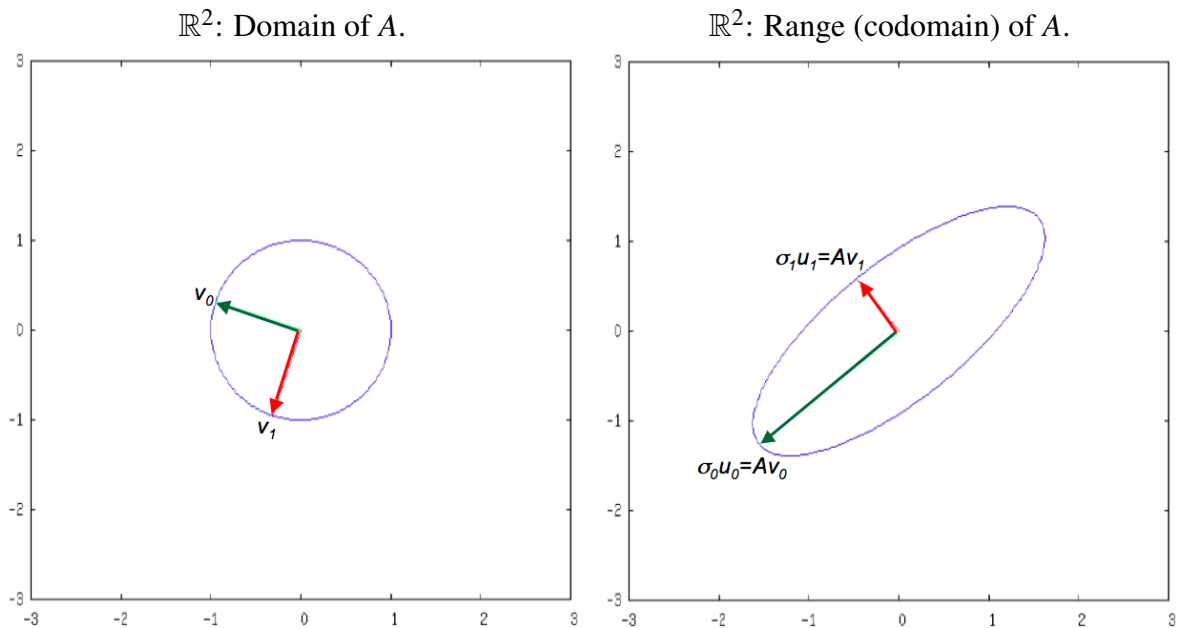
where  $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$  is the condition number of  $A$ , using the 2-norm.

**Homework 3.35** Show that if  $A \in \mathbb{C}^{m \times m}$  is nonsingular, then

- $\|A\|_2 = \sigma_0$ , the largest singular value;
- $\|A^{-1}\|_2 = 1/\sigma_{m-1}$ , the inverse of the smallest singular value; and
- $\kappa_2(A) = \sigma_0/\sigma_{m-1}$ .

SEE ANSWER

If we go back to the example of  $A \in \mathbb{R}^{2 \times 2}$ , recall the following pictures that shows how  $A$  transforms the unit circle:



In this case, the ratio  $\sigma_0/\sigma_{n-1}$  represents the ratio between the major and minor axes of the “ellipse” on the right.

## 3.10 An Algorithm for Computing the SVD?

It would seem that the proof of the existence of the SVD is constructive in the sense that it provides an algorithm for computing the SVD of a given matrix  $A \in \mathbb{C}^{m \times m}$ . Not so fast! Observe that

- Computing  $\|A\|_2$  is nontrivial.
- Computing the vector that maximizes  $\max_{\|x\|_2=1} \|Ax\|_2$  is nontrivial.
- Given a vector  $q_0$  computing vectors  $q_0, \dots, q_{m-1}$  is expensive (as we will see when we discuss the QR factorization).

Towards the end of the course we will discuss algorithms for computing the eigenvalues and eigenvectors of a matrix, and related algorithms for computing the SVD.

---





## Notes on Gram-Schmidt QR Factorization

A classic problem in linear algebra is the computation of an orthonormal basis for the space spanned by a given set of linearly independent vectors: Given a linearly independent set of vectors  $\{a_0, \dots, a_{n-1}\} \subset \mathbb{C}^m$  we would like to find a set of mutually orthonormal vectors  $\{q_0, \dots, q_{n-1}\} \subset \mathbb{C}^m$  so that

$$\text{Span}(\{a_0, \dots, a_{n-1}\}) = \text{Span}(\{q_0, \dots, q_{n-1}\}).$$

This problem is equivalent to the problem of, given a matrix  $A = \left( a_0 \mid \cdots \mid a_{n-1} \right)$ , computing a matrix  $Q = \left( q_0 \mid \cdots \mid q_{n-1} \right)$  with  $Q^H Q = I$  so that  $\mathcal{C}(A) = \mathcal{C}(Q)$ , where  $\mathcal{C}(A)$  denotes the column space of  $A$ .

A review at the undergraduate level of this topic (with animated illustrations) can be found in Week 11 of

Linear Algebra: Foundations to Frontiers - Notes to LAFF With [29].

### Video

Read disclaimer regarding the videos in the preface!

👉 YouTube

👉 Download from UT Box

👉 View After Local Download

(For help on viewing, see Appendix A.)

## Outline

<b>Video</b>	<b>57</b>
<b>Outline</b>	<b>58</b>
<b>4.1. Classical Gram-Schmidt (CGS) Process</b>	<b>59</b>
<b>4.2. Modified Gram-Schmidt (MGS) Process</b>	<b>64</b>
<b>4.3. In Practice, MGS is More Accurate</b>	<b>68</b>
<b>4.4. Cost</b>	<b>70</b>
4.4.1. Cost of CGS	71
4.4.2. Cost of MGS	72

---

## 4.1 Classical Gram-Schmidt (CGS) Process

Given a set of linearly independent vectors  $\{a_0, \dots, a_{n-1}\} \subset \mathbb{C}^m$ , the Gram-Schmidt process computes an orthonormal basis  $\{q_0, \dots, q_{n-1}\}$  that span the same subspace, i.e.

$$\text{Span}(\{a_0, \dots, a_{n-1}\}) = \text{Span}(\{q_0, \dots, q_{n-1}\}).$$

The process proceeds as described in Figure 4.1 and in the algorithms in Figure 4.2.

**Homework 4.1** • What happens in the Gram-Schmidt algorithm if the columns of  $A$  are NOT linearly independent?

- How might one fix this?
- How can the Gram-Schmidt algorithm be used to identify which columns of  $A$  are linearly independent?

➡ SEE ANSWER

**Homework 4.2** **Homework 4.3** Convince yourself that the relation between the vectors  $\{a_j\}$  and  $\{q_j\}$  in the algorithms in Figure 4.2 is given by

$$\left( \begin{array}{c|c|c|c} a_0 & a_1 & \cdots & a_{n-1} \end{array} \right) = \left( \begin{array}{c|c|c|c} q_0 & q_1 & \cdots & q_{n-1} \end{array} \right) \begin{pmatrix} \rho_{0,0} & \rho_{0,1} & \cdots & \rho_{0,n-1} \\ 0 & \rho_{1,1} & \cdots & \rho_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_{n-1,n-1} \end{pmatrix},$$

where

$$q_i^H q_j = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \rho_{i,j} = \begin{cases} q_i^H a_j & \text{for } i < j \\ \|a_j - \sum_{i=0}^{j-1} \rho_{i,j} q_i\|_2 & \text{for } i = j \\ 0 & \text{otherwise.} \end{cases}$$

➡ SEE ANSWER

Thus, this relationship between the linearly independent vectors  $\{a_j\}$  and the orthonormal vectors  $\{q_j\}$  can be concisely stated as

$$A = QR,$$

where  $A$  and  $Q$  are  $m \times n$  matrices ( $m \geq n$ ),  $Q^H Q = I$ , and  $R$  is an  $n \times n$  upper triangular matrix.

**Theorem 4.4** Let  $A$  have linearly independent columns,  $A = QR$  where  $A, Q \in \mathbb{C}^{m \times n}$  with  $n \leq m$ ,  $R \in \mathbb{C}^{n \times n}$ ,  $Q^H Q = I$ , and  $R$  is an upper triangular matrix with nonzero diagonal entries. Then, for  $0 < k < n$ , the first  $k$  columns of  $A$  span the same space as the first  $k$  columns of  $Q$ .

**Proof:** Partition

$$A \rightarrow \left( \begin{array}{c|c} A_L & A_R \end{array} \right), \quad Q \rightarrow \left( \begin{array}{c|c} Q_L & Q_R \end{array} \right), \quad \text{and} \quad R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right),$$

where  $A_L, Q_L \in \mathbb{C}^{m \times k}$  and  $R_{TL} \in \mathbb{C}^{k \times k}$ . Then  $R_{TL}$  is nonsingular (since it is upper triangular and has no zero on its diagonal),  $Q_L^H Q_L = I$ , and  $A_L = Q_L R_{TL}$ . We want to show that  $C(A_L) = C(Q_L)$ :

Steps	Comment
$\rho_{0,0} := \ a_0\ _2$ $q_0 := a_0 / \rho_{0,0}$	<p>Compute the length of vector <math>a_0</math>, <math>\rho_{0,0} := \ a_0\ _2</math>.</p> <p>Set <math>q_0 := a_0 / \rho_{0,0}</math>, creating a unit vector in the direction of <math>a_0</math>.</p> <p>Clearly, <math>\text{Span}(\{a_0\}) = \text{Span}(\{q_0\})</math>. (Why?)</p>
$\rho_{0,1} = q_0^H a_1$ $a_1^\perp = a_1 - \rho_{0,1} q_0$ $\rho_{1,1} = \ a_1^\perp\ _2$ $q_1 = a_1^\perp / \rho_{1,1}$	<p>Compute <math>a_1^\perp</math>, the component of vector <math>a_1</math> orthogonal to <math>q_0</math>.</p> <p>Compute <math>\rho_{1,1}</math>, the length of <math>a_1^\perp</math>.</p> <p>Set <math>q_1 = a_1^\perp / \rho_{1,1}</math>, creating a unit vector in the direction of <math>a_1^\perp</math>.</p> <p>Now, <math>q_0</math> and <math>q_1</math> are mutually orthonormal and <math>\text{Span}(\{a_0, a_1\}) = \text{Span}(\{q_0, q_1\})</math>. (Why?)</p>
$\rho_{0,2} = q_0^H a_2$ $\rho_{1,2} = q_1^H a_2$ $a_2^\perp = a_2 - \rho_{0,2} q_0 - \rho_{1,2} q_1$ $\rho_{2,2} = \ a_2^\perp\ _2$ $q_2 = a_2^\perp / \rho_{2,2}$	<p>Compute <math>a_2^\perp</math>, the component of vector <math>a_2</math> orthogonal to <math>q_0</math> and <math>q_1</math>.</p> <p>Compute <math>\rho_{2,2}</math>, the length of <math>a_2^\perp</math>.</p> <p>Set <math>q_2 = a_2^\perp / \rho_{2,2}</math>, creating a unit vector in the direction of <math>a_2^\perp</math>.</p> <p>Now, <math>\{q_0, q_1, q_2\}</math> is an orthonormal basis and <math>\text{Span}(\{a_0, a_1, a_2\}) = \text{Span}(\{q_0, q_1, q_2\})</math>. (Why?)</p>
And so forth.	

Figure 4.1: Gram-Schmidt orthogonalization.

<pre> <b>for</b> <math>j = 0, \dots, n-1</math>   <math>a_j^\perp := a_j</math>   <b>for</b> <math>k = 0, \dots, j-1</math>     <math>\rho_{k,j} := q_k^H a_j</math>     <math>a_j^\perp := a_j^\perp - \rho_{k,j} q_k</math>   <b>end</b>   <math>\rho_{j,j} := \ a_j^\perp\ _2</math>   <math>q_j := a_j^\perp / \rho_{j,j}</math> <b>end</b> </pre>	<pre> <b>for</b> <math>j = 0, \dots, n-1</math>   <b>for</b> <math>k = 0, \dots, j-1</math>     <math>\rho_{k,j} := q_k^H a_j</math>   <b>end</b>   <math>a_j^\perp := a_j</math>   <b>for</b> <math>k = 0, \dots, j-1</math>     <math>a_j^\perp := a_j^\perp - \rho_{k,j} q_k</math>   <b>end</b>   <math>\rho_{j,j} := \ a_j^\perp\ _2</math>   <math>q_j := a_j^\perp / \rho_{j,j}</math> <b>end</b> </pre>	<pre> <b>for</b> <math>j = 0, \dots, n-1</math>   <math>\begin{pmatrix} \rho_{0,j} \\ \vdots \\ \rho_{j-1,j} \end{pmatrix} := \begin{pmatrix} q_0^H a_j \\ \vdots \\ q_{j-1}^H a_j \end{pmatrix} = \left( q_0 \mid \dots \mid q_{j-1} \right)^H a_j</math>   <math>a_j^\perp := a_j - \left( q_0 \mid \dots \mid q_{j-1} \right) \begin{pmatrix} \rho_{0,j} \\ \vdots \\ \rho_{j-1,j} \end{pmatrix}</math>   <math>\rho_{j,j} := \ a_j^\perp\ _2</math>   <math>q_j := a_j^\perp / \rho_{j,j}</math> <b>end</b> </pre>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 4.2: Three equivalent (Classical) Gram-Schmidt algorithms.

- We first show that  $C(A_L) \subseteq C(Q_L)$ . Let  $y \in C(A_L)$ . Then there exists  $x \in \mathbb{C}^k$  such that  $y = A_L x$ . But then  $y = Q_L z$ , where  $z = R_T L x \neq 0$ , which means that  $y \in C(Q_L)$ . Hence  $C(A_L) \subseteq C(Q_L)$ .
- We next show that  $C(Q_L) \subseteq C(A_L)$ . Let  $y \in C(Q_L)$ . Then there exists  $z \in \mathbb{C}^k$  such that  $y = Q_L z$ . But then  $y = A_L x$ , where  $x = R_T^{-1} L z$ , from which we conclude that  $y \in C(A_L)$ . Hence  $C(Q_L) \subseteq C(A_L)$ .

Since  $C(A_L) \subseteq C(Q_L)$  and  $C(Q_L) \subseteq C(A_L)$ , we conclude that  $C(Q_L) = C(A_L)$ .

**Theorem 4.5** Let  $A \in \mathbb{C}^{m \times n}$  have linearly independent columns. Then there exist  $Q \in \mathbb{C}^{m \times n}$  with  $Q^H Q = I$  and upper triangular  $R$  with no zeroes on the diagonal such that  $A = QR$ . **This is known as the QR factorization.** If the diagonal elements of  $R$  are chosen to be real and positive, the QR factorization is unique.

**Proof:** (By induction). Note that  $n \leq m$  since  $A$  has linearly independent columns.

- **Base case:**  $n = 1$ . In this case  $A = \begin{pmatrix} a_0 \end{pmatrix}$  where  $a_0$  is its only column. Since  $A$  has linearly independent columns,  $a_0 \neq 0$ . Then

$$A = \begin{pmatrix} a_0 \end{pmatrix} = (q_0) (\rho_{00}),$$

where  $\rho_{00} = \|a_0\|_2$  and  $q_0 = a_0 / \rho_{00}$ , so that  $Q = (q_0)$  and  $R = (\rho_{00})$ .

- **Inductive step:** Assume that the result is true for all  $A$  with  $n-1$  linearly independent columns. We will show it is true for  $A \in \mathbb{C}^{m \times n}$  with linearly independent columns.

Let  $A \in \mathbb{C}^{m \times n}$ . Partition  $A \rightarrow \begin{pmatrix} A_0 & a_1 \end{pmatrix}$ . By the induction hypothesis, there exist  $Q_0$  and  $R_{00}$  such that  $Q_0^H Q_0 = I$ ,  $R_{00}$  is upper triangular with nonzero diagonal entries and  $A_0 = Q_0 R_{00}$ . Now,

**Algorithm:**  $[Q, R] := \text{QR}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_L & A_R \end{array} \right),$   
 $Q \rightarrow \left( \begin{array}{c|c} Q_L & Q_R \end{array} \right),$   
 $R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right)$   
**where**  $A_L$  and  $Q_L$  has 0 columns and  $R_{TL}$  is  $0 \times 0$   
**while**  $n(A_L) \neq n(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_L & A_R \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_0 & a_1 & A_2 \end{array} \right),$$

$$\left( \begin{array}{c|c} Q_L & Q_R \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} Q_0 & q_1 & Q_2 \end{array} \right),$$

$$\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$$


---


$$r_{01} := Q_0^T a_1$$

$$a_1^\perp := a_1 - Q_0 r_{01}$$

$$\rho_{11} := \|a_1^\perp\|_2$$

$$q_1 := a_1^\perp / \rho_{11}$$


---

**Continue with**

$$\left( \begin{array}{c|c} A_L & A_R \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_0 & a_1 & A_2 \end{array} \right),$$

$$\left( \begin{array}{c|c} Q_L & Q_R \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} Q_0 & q_1 & Q_2 \end{array} \right),$$

$$\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$$

**endwhile**

**for**  $j = 0, \dots, n-1$

$$\underbrace{\begin{pmatrix} \rho_{0,j} \\ \vdots \\ \rho_{j-1,j} \end{pmatrix}}_{r_{01}} := \underbrace{\begin{pmatrix} q_0 & \cdots & q_{j-1} \end{pmatrix}^H}_{Q_0^H} \underbrace{\begin{pmatrix} a_j \end{pmatrix}}_{a_1}$$

$$\underbrace{a_j^\perp}_{a_1^\perp} := \underbrace{a_j}_{a_1} - \underbrace{\begin{pmatrix} q_0 & \cdots & q_{j-1} \end{pmatrix}}_{Q_0} \underbrace{\begin{pmatrix} \rho_{0,j} \\ \vdots \\ \rho_{j-1,j} \end{pmatrix}}_{r_{01}}$$

$$\rho_{j,j} := \|a_j^\perp\|_2 \quad (\rho_{11} := \|a_1^\perp\|_2)$$

$$q_j := a_j^\perp / \rho_{j,j} \quad (q_1 := a_1^\perp / \rho_{11})$$

**end**

Figure 4.3: (Classical) Gram-Schmidt algorithm for computing the QR factorization of a matrix  $A$ .

compute  $r_{01} = Q_0^H a_1$  and  $a_1^\perp = a_1 - Q_0 r_{01}$ , the component of  $a_1$  orthogonal to  $\mathcal{C}(Q_0)$ . Because the columns of  $A$  are linearly independent,  $a_1^\perp \neq 0$ . Let  $\rho_{11} = \|a_1^\perp\|_2$  and  $q_1 = a_1^\perp / \rho_{11}$ . Then

$$\begin{aligned} \left( \begin{array}{c|c} Q_0 & q_1 \end{array} \right) \left( \begin{array}{c|c} R_{00} & r_{01} \\ \hline 0 & \rho_{11} \end{array} \right) &= \left( \begin{array}{c|c} Q_0 R_{00} & Q_0 r_{01} + q_1 \rho_{11} \end{array} \right) \\ &= \left( \begin{array}{c|c} A_0 & Q_0 r_{01} + a_1^\perp \end{array} \right) = \left( \begin{array}{c|c} A_0 & a_1 \end{array} \right) = A. \end{aligned}$$

$$\text{Hence } Q = \left( Q_0 \mid q_1 \right) \text{ and } R = \left( \begin{array}{c|c} R_{00} & r_{01} \\ \hline 0 & \rho_{11} \end{array} \right).$$

- **By the Principle of Mathematical Induction** the result holds for all matrices  $A \in \mathbb{C}^{m \times n}$  with  $m \geq n$ .

The proof motivates the algorithm in Figure 4.3 (left) in FLAME notation<sup>1</sup>.

An alternative for motivating that algorithm is as follows: Consider  $A = QR$ . Partition  $A$ ,  $Q$ , and  $R$  to yield

$$\left( A_0 \mid a_1 \mid A_2 \right) = \left( Q_0 \mid q_1 \mid Q_2 \right) \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right).$$

Assume that  $Q_0$  and  $R_{00}$  have already been computed. Since corresponding columns of both sides must be equal, we find that

$$a_1 = Q_0 r_{01} + q_1 \rho_{11}. \quad (4.1)$$

Also,  $Q_0^H Q_0 = I$  and  $Q_0^H q_1 = 0$ , since the columns of  $Q$  are mutually orthonormal. Hence  $Q_0^H a_1 = Q_0^H Q_0 r_{01} + Q_0^H q_1 \rho_{11} = r_{01}$ . This shows how  $r_{01}$  can be computed from  $Q_0$  and  $a_1$ , which are already known. Next,  $a_1^\perp = a_1 - Q_0 r_{01}$  is computed from (4.1). This is the component of  $a_1$  that is perpendicular to the columns of  $Q_0$ . We know it is nonzero since the columns of  $A$  are linearly independent. Since  $\rho_{11} q_1 = a_1^\perp$  and we know that  $q_1$  has unit length, we now compute  $\rho_{11} = \|a_1^\perp\|_2$  and  $q_1 = a_1^\perp / \rho_{11}$ , which completes a derivation of the algorithm in Figure 4.3.

**Homework 4.6 Homework 4.7** Let  $A$  have linearly independent columns and let  $A = QR$  be a QR factorization of  $A$ . Partition

$$A \rightarrow \left( A_L \mid A_R \right), \quad Q \rightarrow \left( Q_L \mid Q_R \right), \quad \text{and} \quad R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right),$$

where  $A_L$  and  $Q_L$  have  $k$  columns and  $R_{TL}$  is  $k \times k$ . Show that

1.  $A_L = Q_L R_{TL}$ :  $Q_L R_{TL}$  equals the QR factorization of  $A_L$ ,
2.  $C(A_L) = C(Q_L)$ : the first  $k$  columns of  $Q$  form an orthonormal basis for the space spanned by the first  $k$  columns of  $A$ .
3.  $R_{TR} = Q_L^H A_R$ ,
4.  $(A_R - Q_L R_{TR})^H Q_L = 0$ ,
5.  $A_R - Q_L R_{TR} = Q_R R_{BR}$ , and
6.  $C(A_R - Q_L R_{TR}) = C(Q_R)$ .

➡ SEE ANSWER

<sup>1</sup> The FLAME notation should be intuitively obvious. If it is not, you may want to review the earlier weeks in [Linear Algebra: Foundations to Frontiers - Notes to LAFF With](#).

$[y^\perp, r] = \text{Proj\_orthog\_to\_Q}_{\text{CGS}}(Q, y)$ (used by classical Gram-Schmidt)	$[y^\perp, r] = \text{Proj\_orthog\_to\_Q}_{\text{MGS}}(Q, y)$ (used by modified Gram-Schmidt)
$y^\perp = y$ for $i = 0, \dots, k-1$ $\rho_i := q_i^H y$ $y^\perp := y^\perp - \rho_i q_i$ endfor	$y^\perp = y$ for $i = 0, \dots, k-1$ $\rho_i := q_i^H y^\perp$ $y^\perp := y^\perp - \rho_i q_i$ endfor

Figure 4.4: Two different ways of computing  $y^\perp = (I - QQ^H)y$ , the component of  $y$  orthogonal to  $C(Q)$ , where  $Q$  has  $k$  orthonormal columns.

## 4.2 Modified Gram-Schmidt (MGS) Process

We start by considering the following problem: Given  $y \in \mathbb{C}^m$  and  $Q \in \mathbb{C}^{m \times k}$  with orthonormal columns, compute  $y^\perp$ , the component of  $y$  orthogonal to the columns of  $Q$ . This is a key step in the Gram-Schmidt process in Figure 4.3.

Recall that if  $A$  has linearly independent columns, then  $A(A^H A)^{-1} A^H y$  equals the projection of  $y$  onto the columns space of  $A$  (i.e., the component of  $y$  in  $C(A)$ ) and  $y - A(A^H A)^{-1} A^H y = (I - A(A^H A)^{-1} A^H)y$  equals the component of  $y$  orthogonal to  $C(A)$ . If  $Q$  has orthonormal columns, then  $Q^H Q = I$  and hence  $QQ^H y$  equals the projection of  $y$  onto the columns space of  $Q$  (i.e., the component of  $y$  in  $C(Q)$ ) and  $y - QQ^H y = (I - QQ^H)y$  equals the component of  $y$  orthogonal to  $C(Q)$ .

Thus, mathematically, the solution to the stated problem is given by

$$\begin{aligned}
y^\perp &= (I - QQ^H)y = y - QQ^H y \\
&= y - \left( q_0 \mid \cdots \mid q_{k-1} \right) \left( q_0 \mid \cdots \mid q_{k-1} \right)^H y \\
&= y - \left( q_0 \mid \cdots \mid q_{k-1} \right) \begin{pmatrix} q_0^H y \\ \vdots \\ q_{k-1}^H y \end{pmatrix} \\
&= y - \left( q_0 \mid \cdots \mid q_{k-1} \right) \begin{pmatrix} q_0^H y \\ \vdots \\ q_{k-1}^H y \end{pmatrix} \\
&= y - [(q_0^H y)q_0 + \cdots + (q_{k-1}^H y)q_{k-1}] \\
&= y - (q_0^H y)q_0 - \cdots - (q_{k-1}^H y)q_{k-1}.
\end{aligned}$$

This can be computed by the algorithm in Figure 4.4 (left) and is used by what is often called the *Classical* Gram-Schmidt (CGS) algorithm given in Figure 4.3.

An alternative algorithm for computing  $y^\perp$  is given in Figure 4.4 (right) and is used by the *Modified* Gram-Schmidt (MGS) algorithm also given in Figure 4.5. This approach is mathematically equivalent to



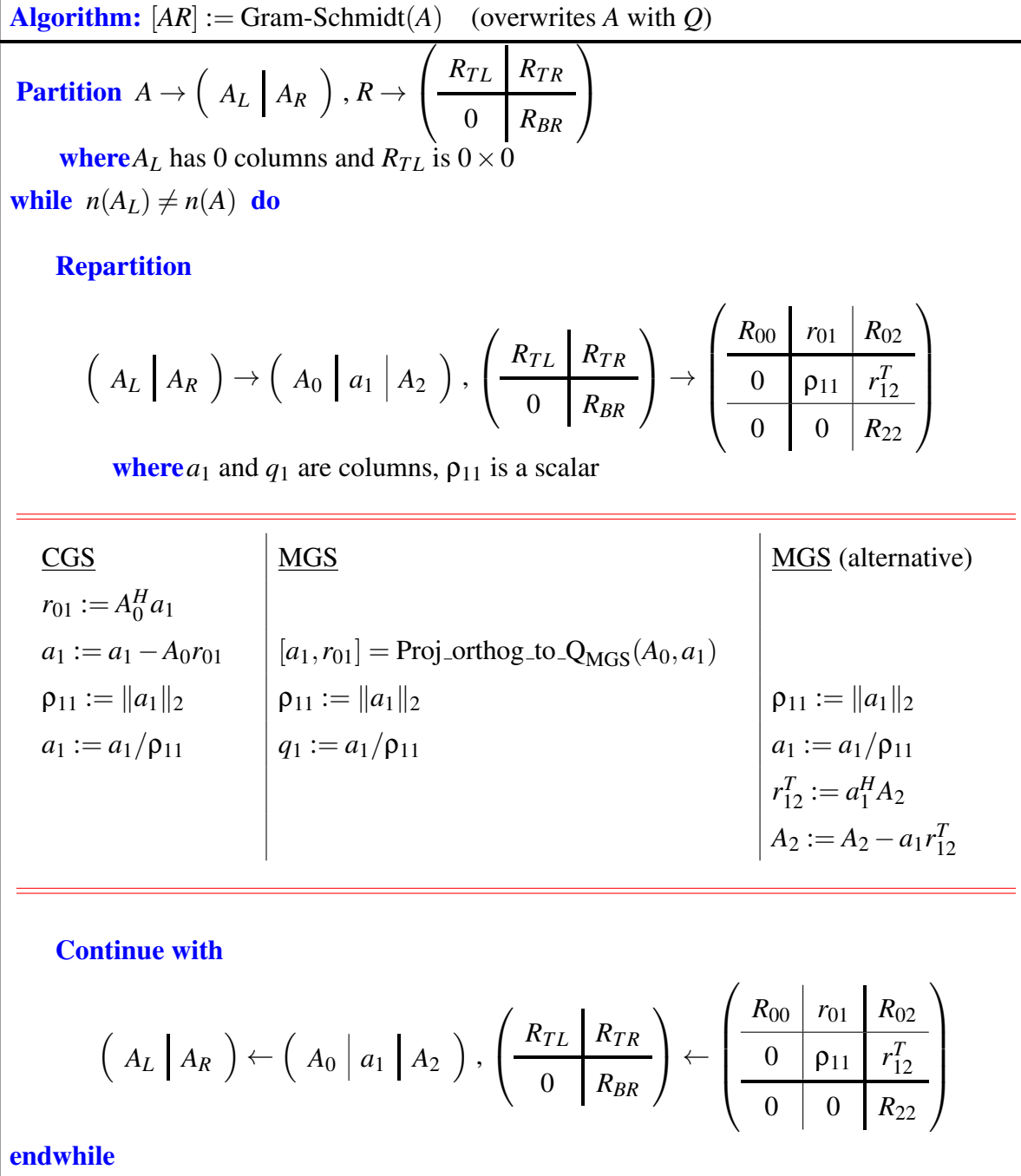


Figure 4.5: Left: Classical Gram-Schmidt algorithm. Middle: Modified Gram-Schmidt algorithm. Right: Modified Gram-Schmidt algorithm where every time a new column of  $Q$ ,  $q_1$  is computed the component of all future columns in the direction of this new vector are subtracted out.

the algorithm to its left for the following reason:

The algorithm on the left in Figure 4.4 computes

$$y^\perp := y - (q_0^H y) q_0 - \cdots - (q_{k-1}^H y) q_{k-1}$$

by in the  $i$ th step computing the component of  $y$  in the direction of  $q_i$ ,  $(q_i^H y) q_i$ , and then subtracting this

```

for  $j = 0, \dots, n-1$ 
   $a_j^\perp := a_j$ 
  for  $k = 0, \dots, j-1$ 
     $\rho_{k,j} := q_k^H a_j^\perp$ 
     $a_j^\perp := a_j^\perp - \rho_{k,j} q_k$ 
  end
   $\rho_{j,j} := \|a_j^\perp\|_2$ 
   $q_j := a_j^\perp / \rho_{j,j}$ 
end

```

(a) MGS algorithm that computes  $Q$  and  $R$  from  $A$ .

```

for  $j = 0, \dots, n-1$ 
  for  $k = 0, \dots, j-1$ 
     $\rho_{k,j} := a_k^H a_j$ 
     $a_j := a_j - \rho_{k,j} a_k$ 
  end
   $\rho_{j,j} := \|a_j\|_2$ 
   $a_j := a_j / \rho_{j,j}$ 
end

```

(b) MGS algorithm that computes  $Q$  and  $R$  from  $A$ , overwriting  $A$  with  $Q$ .

```

for  $j = 0, \dots, n-1$ 
   $\rho_{j,j} := \|a_j\|_2$ 
   $a_j := a_j / \rho_{j,j}$ 
  for  $k = j+1, \dots, n-1$ 
     $\rho_{j,k} := a_j^H a_k$ 
     $a_k := a_k - \rho_{j,k} a_j$ 
  end
end

```

(c) MGS algorithm that normalizes the  $j$ th column to have unit length to compute  $q_j$  (overwriting  $a_j$  with the result) and then subtracts the component in the direction of  $q_j$  off the rest of the columns ( $a_{j+1}, \dots, a_{n-1}$ ).

```

for  $j = 0, \dots, n-1$ 
   $\rho_{j,j} := \|a_j\|_2$ 
   $a_j := a_j / \rho_{j,j}$ 
  for  $k = j+1, \dots, n-1$ 
     $\rho_{j,k} := a_j^H a_k$ 
    end
    for  $k = j+1, \dots, n-1$ 
       $a_k := a_k - \rho_{j,k} a_j$ 
    end
  end

```

(d) Slight modification of the algorithm in (c) that computes  $\rho_{j,k}$  in a separate loop.

```

for  $j = 0, \dots, n-1$ 
   $\rho_{j,j} := \|a_j\|_2$ 
   $a_j := a_j / \rho_{j,j}$ 
   $\begin{pmatrix} \rho_{j,j+1} & \dots & \rho_{j,n-1} \end{pmatrix} := \begin{pmatrix} a_j^H a_{j+1} & \dots & a_j^H a_{n-1} \end{pmatrix}$ 
   $\begin{pmatrix} a_{j+1} & \dots & a_{n-1} \end{pmatrix} :=$ 
     $\begin{pmatrix} a_{j+1} - \rho_{j,j+1} a_j & \dots & a_{n-1} - \rho_{j,n-1} a_j \end{pmatrix}$ 
end

```

(e) Algorithm in (d) rewritten without loops.

```

for  $j = 0, \dots, n-1$ 
   $\rho_{j,j} := \|a_j\|_2$ 
   $a_j := a_j / \rho_{j,j}$ 
   $\begin{pmatrix} \rho_{j,j+1} & \dots & \rho_{j,n-1} \end{pmatrix} := a_j^H \begin{pmatrix} a_{j+1} & \dots & a_{n-1} \end{pmatrix}$ 
   $\begin{pmatrix} a_{j+1} & \dots & a_{n-1} \end{pmatrix} := \begin{pmatrix} a_{j+1} & \dots & a_{n-1} \end{pmatrix} -$ 
     $a_j \begin{pmatrix} \rho_{j,j+1} & \dots & \rho_{j,n-1} \end{pmatrix}$ 
end

```

(f) Algorithm in (e) rewritten to expose the row-vector-times matrix multiplication  $a_j^H \begin{pmatrix} a_{j+1} & \dots & a_{n-1} \end{pmatrix}$  and rank-1 update  $\begin{pmatrix} a_{j+1} & \dots & a_{n-1} \end{pmatrix} - a_j \begin{pmatrix} \rho_{j,j+1} & \dots & \rho_{j,n-1} \end{pmatrix}$ .

Figure 4.6: Various equivalent MGS algorithms.

**Algorithm:**  $[A, R] := \text{QR}(A)$

**Partition**  $A \rightarrow \left( A_L \mid A_R \right),$   
 $R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right)$   
**where**  $A_L$  and  $Q_L$  has 0 columns and  $R_{TL}$  is  $0 \times 0$   
**while**  $n(A_L) \neq n(A)$  **do**

**Repartition**

$$\left( A_L \mid A_R \right) \rightarrow \left( A_0 \mid a_1 \mid A_2 \right),$$

$$\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$$


---

$\rho_{11} := \|a_1\|_2$   
 $a_1 := a_1 / \rho_{11}$   
 $r_{12}^T := a_1^H A_2$   
 $A_2 := A_2 - a_1 r_{12}^T$

---

**Continue with**

$$\left( A_L \mid A_R \right) \leftarrow \left( A_0 \mid a_1 \mid A_2 \right),$$

$$\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$$

**endwhile**

**for**  $j = 0, \dots, n-1$

$$\rho_{j,j} := \|a_j\|_2 \quad (\rho_{11} := \|a_1^\perp\|_2)$$

$$a_j := a_j / \rho_{j,j} \quad (a_1 := a_1 / \rho_{11})$$

$$\overbrace{\left( \begin{array}{c|c|c} r_{12}^T & & \\ \hline \rho_{j,j+1} & \cdots & \rho_{j,n-1} \end{array} \right)}^{A_2} :=$$

$$\underbrace{a_j^H}_{a_1^H} \underbrace{\left( \begin{array}{c|c|c} a_{j+1} & \cdots & a_{n-1} \end{array} \right)}_{A_2}$$

$$\overbrace{\left( \begin{array}{c|c|c} a_{j+1} & \cdots & a_{n-1} \end{array} \right)}^{A_2} := \overbrace{\left( \begin{array}{c|c|c} a_{j+1} & \cdots & a_{n-1} \end{array} \right)}^{A_2}$$

$$- \underbrace{a_j}_{a_1} \underbrace{\left( \begin{array}{c|c|c} \rho_{j,j+1} & \cdots & \rho_{j,n-1} \end{array} \right)}_{r_{12}^T}$$

**end**

Figure 4.7: Modified Gram-Schmidt algorithm for computing the QR factorization of a matrix  $A$ .

off the vector  $y^\perp$  that already contains

$$y^\perp = y - (q_0^H y) q_0 - \cdots - (q_{i-1}^H y) q_{i-1},$$

leaving us with

$$y^\perp = y - (q_0^H y) q_0 - \cdots - (q_{i-1}^H y) q_{i-1} - (q_i^H y) q_i.$$

Now, notice that

$$\begin{aligned} q_i^H [y - (q_0^H y) q_0 - \cdots - (q_{i-1}^H y) q_{i-1}] &= q_i^H y - q_i^H (q_0^H y) q_0 - \cdots - q_i^H (q_{i-1}^H y) q_{i-1} \\ &= q_i^H y - (q_0^H y) \underbrace{q_i^H q_0}_0 - \cdots - (q_{i-1}^H y) \underbrace{q_i^H q_{i-1}}_0 \\ &= q_i^H y. \end{aligned}$$

What this means is that we can use  $y^\perp$  in our computation of  $\rho_i$  instead:

$$\rho_i := q_i^H y^\perp = q_i^H y,$$

an insight that justifies the equivalent algorithm in Figure 4.4 (right).

Next, we massage the MGS algorithm into the third (right-most) algorithm given in Figure 4.5. For this, consider the equivalent algorithms in Figure 4.6 and 4.7.

### 4.3 In Practice, MGS is More Accurate

In theory, all Gram-Schmidt algorithms discussed in the previous sections are equivalent: they compute the exact same QR factorizations. In practice, in the presense of round-off error, MGS is more accurate than CGS. We will (hopefully) get into detail about this later, but for now we will illustrate it with a classic example.

When storing real (or complex for that matter) valued numbers in a computer, a limited accuracy can be maintained, leading to round-off error when a number is stored and/or when computation with numbers are performed. The *machine epsilon* or *unit roundoff error* is defined as the largest positive number  $\epsilon_{\text{mach}}$  such that the stored value of  $1 + \epsilon_{\text{mach}}$  is rounded to 1. Now, let us consider a computer where the **only** error that is ever incurred is when  $1 + \epsilon_{\text{mach}}$  is computed and rounded to 1. Let  $\epsilon = \sqrt{\epsilon_{\text{mach}}}$  and consider the matrix

$$A = \left( \begin{array}{c|c|c} 1 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{array} \right) = \left( \begin{array}{c|c|c} a_0 & a_1 & a_2 \end{array} \right) \quad (4.2)$$

In Figure 4.8 (left) we execute the CGS algorithm. It yields the approximate matrix

$$Q \approx \left( \begin{array}{c|c|c} 1 & 0 & 0 \\ \epsilon & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} \end{array} \right)$$

If we now ask the question “Are the columns of  $Q$  orthonormal?” we can check this by computing  $Q^H Q$ , which should equal  $I$ , the identity. But

$$Q^H Q = \left( \begin{array}{c|c|c} 1 & 0 & 0 \\ \epsilon & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} \end{array} \right)^H \left( \begin{array}{c|c|c} 1 & 0 & 0 \\ \epsilon & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} \end{array} \right) = \left( \begin{array}{ccc} 1 + \epsilon_{\text{mach}} & -\frac{\sqrt{2}}{2}\epsilon & -\frac{\sqrt{2}}{2}\epsilon \\ -\frac{\sqrt{2}}{2}\epsilon & 1 & \frac{1}{2} \\ -\frac{\sqrt{2}}{2}\epsilon & \frac{1}{2} & 1 \end{array} \right).$$

Clearly, the computed columns of  $Q$  are **not** mutually orthogonal.

<p><u>First iteration</u></p> $\rho_{0,0} = \ a_0\ _2 = \sqrt{1 + \epsilon^2} = \sqrt{1 + \epsilon_{\text{mach}}}$ <p><b>which is rounded to 1.</b></p> $q_0 = a_0 / \rho_{0,0} = \begin{pmatrix} 1 \\ \epsilon \\ 0 \\ 0 \end{pmatrix} / 1 = \begin{pmatrix} 1 \\ \epsilon \\ 0 \\ 0 \end{pmatrix}$ <p><u>Second iteration</u></p> $\rho_{0,1} = q_0^H a_1 = 1$ $a_1^\perp = a_1 - \rho_{0,1} q_0 = \begin{pmatrix} 0 \\ -\epsilon \\ \epsilon \\ 0 \end{pmatrix}$ $\rho_{1,1} = \ a_1^\perp\ _2 = \sqrt{2\epsilon^2} = \sqrt{2}\epsilon$ $q_1 = a_1^\perp / \rho_{1,1} = \begin{pmatrix} 0 \\ -\epsilon \\ \epsilon \\ 0 \end{pmatrix} / (\sqrt{2}\epsilon) = \begin{pmatrix} 0 \\ -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}$ <p><u>Third iteration</u></p> $\rho_{0,2} = q_0^H a_2 = 1$ $\rho_{1,2} = q_1^H a_2 = 0$ $a_2^\perp = a_2 - \rho_{0,2} q_0 - \rho_{1,2} q_1 = \begin{pmatrix} 0 \\ -\epsilon \\ 0 \\ \epsilon \end{pmatrix}$ $\rho_{2,2} = \ a_2^\perp\ _2 = \sqrt{2\epsilon^2} = \sqrt{2}\epsilon$ $q_2 = a_2^\perp / \rho_{2,2} = \begin{pmatrix} 0 \\ -\epsilon \\ 0 \\ \epsilon \end{pmatrix} / (\sqrt{2}\epsilon) = \begin{pmatrix} 0 \\ -\frac{\sqrt{2}}{2} \\ 0 \\ \frac{\sqrt{2}}{2} \end{pmatrix}$	<p><u>First iteration</u></p> $\rho_{0,0} = \ a_0\ _2 = \sqrt{1 + \epsilon^2} = \sqrt{1 + \epsilon_{\text{mach}}}$ <p><b>which is rounded to 1.</b></p> $q_0 = a_0 / \rho_{0,0} = \begin{pmatrix} 1 \\ \epsilon \\ 0 \\ 0 \end{pmatrix} / 1 = \begin{pmatrix} 1 \\ \epsilon \\ 0 \\ 0 \end{pmatrix}$ <p><u>Second iteration</u></p> $\rho_{0,1} = q_0^H a_1 = 1$ $a_1^\perp = a_1 - \rho_{0,1} q_0 = \begin{pmatrix} 0 \\ -\epsilon \\ \epsilon \\ 0 \end{pmatrix}$ $\rho_{1,1} = \ a_1^\perp\ _2 = \sqrt{2\epsilon^2} = \sqrt{2}\epsilon$ $q_1 = a_1^\perp / \rho_{1,1} = \begin{pmatrix} 0 \\ -\epsilon \\ \epsilon \\ 0 \end{pmatrix} / (\sqrt{2}\epsilon) = \begin{pmatrix} 0 \\ -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}$ <p><u>Third iteration</u></p> $\rho_{0,2} = q_0^H a_2 = 1$ $a_2^\perp = a_2 - \rho_{0,2} q_0 = \begin{pmatrix} 0 \\ -\epsilon \\ 0 \\ \epsilon \end{pmatrix}$ $\rho_{1,2} = q_1^H a_2^\perp = (\sqrt{2}/2)\epsilon$ $a_2^\perp = a_2^\perp - \rho_{1,2} q_1 = \begin{pmatrix} 0 \\ -\epsilon/2 \\ -\epsilon/2 \\ \epsilon \end{pmatrix}$ $\rho_{2,2} = \ a_2^\perp\ _2 = \sqrt{(6/4)\epsilon^2} = (\sqrt{6}/2)\epsilon$ $q_2 = a_2^\perp / \rho_{2,2} = \begin{pmatrix} 0 \\ -\frac{\epsilon}{2} \\ -\frac{\epsilon}{2} \\ \epsilon \end{pmatrix} / \left(\frac{\sqrt{6}}{2}\epsilon\right) = \begin{pmatrix} 0 \\ \frac{\sqrt{6}}{6} \\ -\frac{\sqrt{6}}{6} \\ \frac{2\sqrt{6}}{6} \end{pmatrix}$
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 4.8: Execution of the CGS algorithm (left) and MGS algorithm (right) on the example in Eqn. (4.2).

Similarly, in Figure 4.8 (right) we execute the MGS algorithm. It yields the approximate matrix

$$Q \approx \left( \begin{array}{c|c|c} 1 & 0 & 0 \\ \epsilon & -\frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{6} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{6}}{6} \\ 0 & 0 & \frac{2\sqrt{6}}{6} \end{array} \right).$$

If we now ask the question “Are the columns of  $Q$  orthonormal?” we can check if  $Q^H Q = I$ . The answer:

$$Q^H Q = \left( \begin{array}{c|c|c} 1 & 0 & 0 \\ \epsilon & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{6}}{6} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{6}}{6} \\ 0 & 0 & \frac{2\sqrt{6}}{6} \end{array} \right)^H \left( \begin{array}{c|c|c} 1 & 0 & 0 \\ \epsilon & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{6}}{6} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{6}}{6} \\ 0 & 0 & \frac{2\sqrt{6}}{6} \end{array} \right) = \left( \begin{array}{ccc} 1 + \epsilon_{\text{mach}} & -\frac{\sqrt{2}}{2}\epsilon & -\frac{\sqrt{6}}{6}\epsilon \\ -\frac{\sqrt{2}}{2}\epsilon & 1 & 0 \\ -\frac{\sqrt{6}}{6}\epsilon & 0 & 1 \end{array} \right),$$

which shows that for this example MGS yields better orthogonality than does CGS. What is going on? The answer lies with how  $a_2^\perp$  is computed in the last step of each of the algorithms.

- In the CGS algorithm, we find that

$$a_2^\perp := a_2 - (q_0^H a_2)q_0 - (q_1^H a_2)q_1.$$

Now,  $q_0$  has a relatively small error in it and hence  $q_0^H a_2 q_0$  has a relatively small error in it. It is likely that a part of that error is in the direction of  $q_1$ . Relative to  $q_0^H a_2 q_0$ , that error in the direction of  $q_1$  is small, but relative to  $a_2 - q_0^H a_2 q_0$  it is not. The point is that then  $a_2 - q_0^H a_2 q_0$  has a relatively large error in it in the direction of  $q_1$ . Subtracting  $q_1^H a_2 q_1$  does not fix this and since in the end  $a_2^\perp$  is small, it has a relatively large error in the direction of  $q_1$ . This error is amplified when  $q_2$  is computed by normalizing  $a_2^\perp$ .

- In the MGS algorithm, we find that

$$a_2^\perp := a_2 - (q_0^H a_2)q_0$$

after which

$$a_2^\perp := a_2^\perp - q_1^H a_2^\perp q_1 = [a_2 - (q_0^H a_2)q_0] - (q_1^H [a_2 - (q_0^H a_2)q_0])q_1.$$

This time, if  $a_2 - q_0^H a_2 q_0$  has an error in the direction of  $q_1$ , this error is subtracted out when  $(q_1^H a_2^\perp)q_1$  is subtracted from  $a_2^\perp$ . This explains the better orthogonality between the computed vectors  $q_1$  and  $q_2$ .

Obviously, we have argued via an example that MGS is more accurate than CGS. A more thorough analysis is needed to explain why this is generally so. This is beyond the scope of this note.

## 4.4 Cost

Let us examine the cost of computing the QR factorization of an  $m \times n$  matrix  $A$ . We will count multiplies and adds as each as one floating point operation.

We start by reviewing the cost, in floating point operations (flops), of various vector-vector and matrix-vector operations:

Name	Operation	Approximate cost (in flops)
Vector-vector operations ( $x, y \in \mathbb{C}^n, \alpha \in \mathbb{C}$ )		
Dot	$\alpha := x^H y$	$2n$
Axpy	$y := \alpha x + y$	$2n$
Scal	$x := \alpha x$	$n$
Nrm2	$\alpha := \ a_1\ _2$	$2n$
Matrix-vector operations ( $A \in \mathbb{C}^{m \times n}, \alpha, \beta \in \mathbb{C}$ , with $x$ and $y$ vectors of appropriate size)		
Matrix-vector multiplication (Gemv)	$y := \alpha A x + \beta y$	$2mn$
	$y := \alpha A^H x + \beta y$	$2mn$
Rank-1 update (Ger)	$A := \alpha y x^H + A$	$2mn$

Now, consider the algorithms in Figure 4.5. Notice that the columns of  $A$  are of size  $m$ . During the  $k$ th iteration ( $0 \leq k < n$ ),  $A_0$  has  $k$  columns and  $A_2$  has  $n - k - 1$  columns.

#### 4.4.1 Cost of CGS

Operation	Approximate cost (in flops)
$r_{01} := A_0^H a_1$	$2mk$
$a_1 := a_1 - A_0 r_{01}$	$2mk$
$\rho_{11} := \ a_1\ _2$	$2m$
$a_1 := a_1 / \rho_{11}$	$m$

Thus, the total cost is (approximately)

$$\begin{aligned}
& \sum_{k=0}^{n-1} [2mk + 2mk + 2m + m] \\
&= \sum_{k=0}^{n-1} [3m + 4mk] \\
&= 3mn + 4m \sum_{k=0}^{n-1} k \\
&\approx 3mn + 4m \frac{n^2}{2} & (\sum_{k=0}^{n-1} k = n(n-1)/2 \approx n^2/2) \\
&= 3mn + 2mn^2 \\
&\approx 2mn^2 & (3mn \text{ is of lower order}).
\end{aligned}$$

### 4.4.2 Cost of MGS

Operation	Approximate cost (in flops)
$\rho_{11} := \ a_1\ _2$	$2m$
$a_1 := a_1/\rho_{11}$	$m$
$r_{12}^T := a_1^H A_2$	$2m(n-k-1)$
$A_2 := A_2 - a_1 r_{12}^T$	$2m(n-k-1)$

Thus, the total cost is (approximately)

$$\begin{aligned}
 & \sum_{k=0}^{n-1} [2m + m + 2m(n-k-1) + 2m(n-k-1)] \\
 &= \sum_{k=0}^{n-1} [3m + 4m(n-k-1)] \\
 &= 3mn + 4m \sum_{k=0}^{n-1} (n-k-1) \\
 &= 3mn + 4m \sum_{i=0}^{n-1} i && \text{(Change of variable: } i = n-k-1) \\
 &\approx 3mn + 4m \frac{n^2}{2} && (\sum_{i=0}^{n-1} i = n(n-1)/2 \approx n^2/2) \\
 &= 3mn + 2mn^2 \\
 &\approx 2mn^2 && (3mn \text{ is of lower order}).
 \end{aligned}$$



## Notes on the FLAME APIs

### Video

Read disclaimer regarding the videos in the preface!

No video.

## Outline

<b>Video</b> . . . . .	<b>73</b>
<b>Outline</b> . . . . .	<b>74</b>
<b>5.1. Motivation</b> . . . . .	<b>75</b>
<b>5.2. Install FLAME@lab</b> . . . . .	<b>75</b>
<b>5.3. An Example: Gram-Schmidt Orthogonalization</b> . . . . .	<b>75</b>
5.3.1. The Spark Webpage . . . . .	75
5.3.2. Implementing CGS with FLAME@lab . . . . .	76
5.3.3. Editing the code skeleton . . . . .	78
5.3.4. Testing . . . . .	79
<b>5.4. Implementing the Other Algorithms</b> . . . . .	<b>80</b>

$[y^\perp, r] = \text{Proj\_orthog\_to\_Q}_{\text{CGS}}(Q, y)$ (used by classical Gram-Schmidt)	$[y^\perp, r] = \text{Proj\_orthog\_to\_Q}_{\text{MGS}}(Q, y)$ (used by modified Gram-Schmidt)
$y^\perp = y$ for $i = 0, \dots, k-1$ $\rho_i := q_i^H y$ $y^\perp := y^\perp - \rho_i q_i$ endfor	$y^\perp = y$ for $i = 0, \dots, k-1$ $\rho_i := q_i^H y^\perp$ $y^\perp := y^\perp - \rho_i q_i$ endfor

Figure 5.1: Two different ways of computing  $y^\perp = (I - QQ^H)y$ , the component of  $y$  orthogonal to  $\mathcal{C}(Q)$ , where  $Q$  has  $k$  orthonormal columns.

## 5.1 Motivation

In the course so far, we have frequently used the “FLAME Notation” to express linear algebra algorithms. In this note we show how to translate such algorithms into code, using various QR factorization algorithms as examples.

## 5.2 Install FLAME@lab

The API we will use we refer to as the “FLAME@lab” API, which is an API that targets the M-script language used by **Matlab** and **Octave** (an Open Source Matlab implementation). This API is very intuitive, and hence we will spend (almost) no time explaining it.

Download all files from <http://www.cs.utexas.edu/users/flame/Notes/FLAMEatlab/> and place them in the same directory as you will the remaining files that you will create as part of the exercises in this document. (Unless you know how to set up paths in Matlab/Octave, in which case you can put it wherever you please, and set the path.)

## 5.3 An Example: Gram-Schmidt Orthogonalization

Let us start by considering the various Gram-Schmidt based QR factorization algorithms from “Notes on Gram-Schmidt QR Factorization”, typeset using the FLAME Notation in Figure 5.2.

### 5.3.1 The Spark Webpage

We wish to typeset the code so that it closely resembles the algorithms in Figure 5.2. The FLAME notation itself uses “white space” to better convey the algorithms. We want to do the same for the codes that implement the algorithms. However, typesetting that code is somewhat bothersome because of the careful spacing that is required. For this reason, we created a webpage that creates a “code skeleton.”. We call this page the “Spark” page:

<http://www.cs.utexas.edu/users/flame/Spark/>.

**Algorithm:**  $[AR] := \text{Gram-Schmidt}(A)$  (overwrites  $A$  with  $Q$ )

**Partition**  $A \rightarrow \left( A_L \mid A_R \right), R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right)$

**where**  $A_L$  has 0 columns and  $R_{TL}$  is  $0 \times 0$

**while**  $n(A_L) \neq n(A)$  **do**

**Repartition**

$\left( A_L \mid A_R \right) \rightarrow \left( A_0 \mid a_1 \mid A_2 \right), \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$

**where**  $a_1$  and  $q_1$  are columns,  $\rho_{11}$  is a scalar

CGS

$$r_{01} := A_0^H a_1$$

$$a_1 := a_1 - A_0 r_{01}$$

$$\rho_{11} := \|a_1\|_2$$

$$a_1 := a_1 / \rho_{11}$$

MGS

$$[a_1, r_{01}] = \text{Proj\_orthog\_to\_Q}_{\text{MGS}}(A_0, a_1)$$

$$\rho_{11} := \|a_1\|_2$$

$$q_1 := a_1 / \rho_{11}$$

MGS (alternative)

$$\rho_{11} := \|a_1\|_2$$

$$a_1 := a_1 / \rho_{11}$$

$$r_{12}^T := a_1^H A_2$$

$$A_2 := A_2 - a_1 r_{12}^T$$

**Continue with**

$\left( A_L \mid A_R \right) \leftarrow \left( A_0 \mid a_1 \mid A_2 \right), \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$

**endwhile**

Figure 5.2: Left: Classical Gram-Schmidt algorithm. Middle: Modified Gram-Schmidt algorithm. Right: Modified Gram-Schmidt algorithm where every time a new column of  $Q$ ,  $q_1$  is computed the component of all future columns in the direction of this new vector are subtracted out.

When you open the link, you will get a page that looks something like the picture in Figure 5.3.

### 5.3.2 Implementing CGS with FLAME@lab

The screenshot shows the Spark web interface. The left sidebar (yellow) contains the following fields:

- Name of the function to be generated:** Function Name (text input)
- Type of function:** blocked (dropdown menu)
- Variant number:** none (dropdown menu)
- Number of operands:** 1 (text input)
- Pick properties of the operands:**

Operand	Tag	Type	Direction	Input/Output
1:	A	matrix	TL->BR	input/output
- Pick an output language:** FLAME@lab (dropdown menu)
- Additional Information:** Name of author (text input)

The right sidebar (white) is titled "Spark" and "FLAME code-skeleton generator". It contains introductory text and a list of instructions for the user.

Figure 5.3: The Spark webpage.

We will focus on the Classical Gram-Schmidt algorithm on the left, which we show by itself in Figure 5.4 (left). To its right, we show how the menu on the left side of the Spark webpage needs to be filled out.

Some comments:

**Name:** Choose a name that describes the algorithm/operation being implemented.

**Type of function:** Later you will learn about “blocked” algorithms. For now, we implement “unblocked” algorithms.

**Variant number:** Notice that there are a number of algorithmic variants for implementing the Gram-Schmidt algorithm. We choose to call the first one “Variant 1”.

**Number of operands:** This routine requires two operands: one each for matrices  $A$  and  $R$ . ( $A$  will be overwritten by the matrix  $Q$ .)

**Operand 1:** We indicate that  $A$  is a matrix through which we “march” from left to right ( $L \rightarrow R$ ) and it is both input and output.

**Operand 2:** We indicate that  $R$  is a matrix through which we “march” from to-left to bottom-right ( $TL \rightarrow BR$ ) and it is both input and output. Our API requires you to pass in the array in which to put an output, so an appropriately sized  $R$  must be passed in.

**Pick and output language:** A number of different representations are supported, including APIs for M-script (FLAME@lab), C (FLAMEC), L<sup>A</sup>T<sub>E</sub>X (FLaTeX), and Python (FlamePy). Pick FLAME@lab.

To the left of the menu, you now find what we call a code skeleton for the implementation, as shown in Figure 5.5. In Figure 5.6 we show the algorithm and generated code skeleton side-by-side.

**Algorithm:**  $[A, R] := \text{Gram-Schmidt}(A)$  (overwrites  $A$  with  $Q$ )

**Partition**  $A \rightarrow \left( A_L \mid A_R \right),$   
 $R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right)$   
**where**  $A_L$  has 0 columns and  $R_{TL}$  is  $0 \times 0$   
**while**  $n(A_L) \neq n(A)$  **do**  
**Repartition**  
 $\left( A_L \mid A_R \right) \rightarrow \left( A_0 \mid a_1 \mid A_2 \right),$   
 $\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$   


---

 $r_{01} := A_0^H a_1$   
 $a_1 := a_1 - A_0 r_{01}$   
 $\rho_{11} := \|a_1\|_2$   
 $a_1 := a_1 / \rho_{11}$   


---

**Continue with**  
 $\left( A_L \mid A_R \right) \leftarrow \left( A_0 \mid a_1 \mid A_2 \right),$   
 $\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$   
**endwhile**

**Name of the function to be generated:**

**Type of function:**

**Variant number:**

---

[learn about this section](#) [introduction to Spark](#)

**Number of operands:**

**Pick properties of the operands**

Operand	Tag	Type	Direction	Input/Output
1:	<input type="button" value="A"/>	<input type="button" value="matrix"/>	<input type="button" value="L-&gt;R"/>	<input type="button" value="input/output"/>
2:	<input type="button" value="R"/>	<input type="button" value="matrix"/>	<input type="button" value="TL-&gt;BR"/>	<input type="button" value="input/output"/>

---

[learn about this section](#) [introduction to Spark](#)

**Pick an output language:**

Figure 5.4: Left: Classical Gram-Schmidt algorithm. Right: Generated code-skeleton for CGS.

### 5.3.3 Editing the code skeleton

At this point, one should copy the code skeleton into one's favorite text editor. (We highly recommend emacs for the serious programmer.) Once this is done, there are two things left to do:

**Fix the code skeleton:** The Spark webpage “guesses” the code skeleton. One detail that it sometimes gets wrong is the “stopping criteria”. In this case, the algorithm should stay in the loop as long as  $n(A_L) \neq n(A)$  (the width of  $A_L$  is not yet the width of  $A$ ). In our example, the Spark webpage guessed that the column size of matrix  $A$  is to be used for the stopping criteria:

```
while ( size( AL, 2 ) < size( A, 2 ) )
```

which happens to be correct. (When you implement the Householder QR factorization, you may not be so lucky...)

**The “update” statements:** The Spark webpage can't guess what the actual updates to the various parts of matrices  $A$  and  $R$  should be. It fills in

Generate Code and/or Update Form
Reset Form

[learn about this section](#)
[introduction to Spark](#)

**Name of the function to be generated:**

**Type of function:**

**Variant number:**

---

[learn about this section](#)
[introduction to Spark](#)

**Number of operands:**

**Pick properties of the operands**

Operand	Tag	Type	Direction	Input/Output
1:	<input type="button" value="A"/>	<input type="button" value="matrix"/>	<input type="button" value="L-&gt;R"/>	<input type="button" value="input/output"/>
2:	<input type="button" value="R"/>	<input type="button" value="matrix"/>	<input type="button" value="TL-&gt;BR"/>	<input type="button" value="input/output"/>

---

[learn about this section](#)
[introduction to Spark](#)

**Pick an output language:**

---

[learn about this section](#)
[introduction to Spark](#)

**Additional Information**

**Name of Author**

```

function [ A_out, R_out ] = CGS_unb_var1( A, R )

[ AL, AR ] = FLA_Part_1x2( A, ...
                          0, 'FLA_LEFT' );

[ RTL, RTR, ...
  RBL, RBR ] = FLA_Part_2x2( R, ...
                          0, 0, 'FLA_TL' );

while ( size( AL, 2 ) < size( A, 2 ) )

  [ A0, a1, A2 ] = FLA_Repart_1x2_to_1x3( AL, AR, ...
                                          1, 'FLA_RIGHT' );

  [ R00, r01, R02, ...
    r10t, rho11, r12t, ...
    R20, r21, R22 ] = FLA_Repart_2x2_to_3x3( RTL, RTR, ...
                                          RBL, RBR, ...
                                          1, 1, 'FLA_BR' );

  %-----%
  %               update line 1               %
  %               :                           %
  %               update line n               %
  %-----%

  [ AL, AR ] = FLA_Cont_with_1x3_to_1x2( A0, a1, A2, ...
                                          'FLA_LEFT' );

  [ RTL, RTR, ...
    RBL, RBR ] = FLA_Cont_with_3x3_to_2x2( R00, r01, R02, ...
                                          r10t, rho11, r12t, ...
                                          R20, r21, R22, ...
                                          'FLA_TL' );

end

```

Figure 5.5: The Spark webpage filled out for CGS Variant 1.

```

%               update line 1               %
%               :                           %
%               update line n               %

```

Thus, one has to manually translate

$$\begin{aligned}
 r_{01} &:= A_0^H a_1 \\
 a_1 &:= a_1 - A_0 r_{01} \\
 \rho_{11} &:= \|a_1\|_2 \\
 a_1 &:= a_1 / \rho_{11}
 \end{aligned}$$

into appropriate M-script code:

```

r01 = A0' * a1;
a1 = a1 - A0 * r01;
rho11 = norm( a1 );
a1 = a1 / rho11;

```

(Notice: if one forgets the “;”, when executed the results of the assignment will be printed by Matlab/Octave.)

At this point, one saves the resulting code in the file `CGS_unb_var1.m`. The “.m” ending is important since the name of the file is used to find the routine when using Matlab/Octave.

### 5.3.4 Testing

To now test the routine, one starts octave and, for example, executes the commands

**Algorithm:**  $[A, R] := \text{Gram-Schmidt}(A)$  (overwrites  $A$  with  $Q$ )

**Partition**  $A \rightarrow \left( A_L \mid A_R \right),$

$$R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right)$$

**where**  $A_L$  has 0 columns and  $R_{TL}$  is  $0 \times 0$

**while**  $n(A_L) \neq n(A)$  **do**

**Repartition**

$$\left( A_L \mid A_R \right) \rightarrow \left( A_0 \mid a_1 \mid A_2 \right),$$

$$\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$$

$$r_{01} := A_0^H a_1$$

$$a_1 := a_1 - A_0 r_{01}$$

$$\rho_{11} := \|a_1\|_2$$

$$a_1 := a_1 / \rho_{11}$$

**Continue with**

$$\left( A_L \mid A_R \right) \leftarrow \left( A_0 \mid a_1 \mid A_2 \right),$$

$$\left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & r_{12}^T \\ \hline 0 & 0 & R_{22} \end{array} \right)$$

**endwhile**

```
function [ A_out, R_out ] = CGS_unb_var1( A, R )

[ AL, AR ] = FLA_Part_1x2( A, ...
    0, 'FLA_LEFT' );

[ RTL, RTR, ...
  RBL, RBR ] = FLA_Part_2x2( R, ...
    0, 0, 'FLA_TL' );

while ( size( AL, 2 ) < size( A, 2 ) )

    [ A0, a1, A2 ] = FLA_Repart_1x2_to_1x3( AL, AR, ...
        1, 'FLA_RIGHT' );

    [ R00, r01, R02, ...
      r10t, rho11, r12t, ...
      R20, r21, R22 ] = FLA_Repart_2x2_to_3x3( RTL, RTR, ...
        RBL, RBR, ...
        1, 1, 'FLA_BR' );

    %-----%
    %               update line 1               %
    %               :                             %
    %               update line n                 %
    %-----%

    [ AL, AR ] = FLA_Cont_with_1x3_to_1x2( A0, a1, A2, ...
        'FLA_LEFT' );

    [ RTL, RTR, ...
      RBL, RBR ] = FLA_Cont_with_3x3_to_2x2( R00, r01, R02, ...
        r10t, rho11, r12t, ...
        R20, r21, R22, ...
        'FLA_TL' );

end
```

Figure 5.6: Left: Classical Gram-Schmidt algorithm. Right: Generated code-skeleton for CGS.

```
> A = rand( 5, 4 )
> R = zeros( 4, 4 )
> [ Q, R ] = CGS_unb_var1( A, R )
> A - Q * triu( R )
```

The result should be (approximately) a  $5 \times 4$  zero matrix.

(The first time you execute the above, you may get a bunch of warnings from Octave. Just ignore those.)

## 5.4 Implementing the Other Algorithms

Next, we leave it to the reader to implement

- Modified Gram Schmidt algorithm, (MGS\_unb\_var1, corresponding to the **right-most** algorithm in Figure 5.2), respectively.



- The Householder QR factorization algorithm and algorithm to form  $Q$  from “Notes on Householder QR Factorization”.

The routine for computing a Householder transformation (similar to Figure 5.1) can be found at

<http://www.cs.utexas.edu/users/flame/Notes/FLAMEatlab/Housev.m>

That routine implements the algorithm on the left in Figure 5.1). Try and see what happens if you replace it with the algorithm to its right.

**Note:** For the Householder QR factorization and “form  $Q$ ” algorithm how to start the algorithm when the matrix is not square is a bit tricky. Thus, you may assume that the matrix *is* square.



## Notes on Householder QR Factorization

### Video

Read disclaimer regarding the videos in the preface!

👉 [YouTube](#)

👉 [Download from UT Box](#)

👉 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

## Outline

<b>Video</b> . . . . .	<b>83</b>
<b>Outline</b> . . . . .	<b>84</b>
<b>6.1. Motivation</b> . . . . .	<b>85</b>
<b>6.2. Householder Transformations (Reflectors)</b> . . . . .	<b>85</b>
6.2.1. The general case . . . . .	85
6.2.2. As implemented for the Householder QR factorization (real case) . . . . .	87
6.2.3. The complex case (optional) . . . . .	87
6.2.4. A routine for computing the Householder vector . . . . .	88
<b>6.3. Householder QR Factorization</b> . . . . .	<b>89</b>
<b>6.4. Forming <math>Q</math></b> . . . . .	<b>92</b>
<b>6.5. Applying <math>Q^H</math></b> . . . . .	<b>97</b>
<b>6.6. Blocked Householder QR Factorization</b> . . . . .	<b>99</b>
6.6.1. The UT transform: Accumulating Householder transformations . . . . .	99
6.6.2. A blocked algorithm . . . . .	102
6.6.3. Variations on a theme . . . . .	104

## 6.1 Motivation

A fundamental problem to avoid in numerical codes is the situation where one starts with large values and one ends up with small values with large relative errors in them. This is known as catastrophic cancellation. The Gram-Schmidt algorithms can inherently fall victim to this: column  $a_j$  is successively reduced in length as components in the directions of  $\{q_0, \dots, q_{j-1}\}$  are subtracted, leaving a small vector if  $a_j$  was almost in the span of the first  $j$  columns of  $A$ . Application of a unitary transformation to a matrix or vector inherently preserves length. Thus, it would be beneficial if the QR factorization can be implemented as the successive application of unitary transformations. The Householder QR factorization accomplishes this.

The first fundamental insight is that the product of unitary matrices is itself unitary. If, given  $A \in \mathbb{C}^{m \times n}$  (with  $m \geq n$ ), one could find a sequence of unitary matrices,  $\{H_0, \dots, H_{n-1}\}$ , such that

$$H_{n-1} \cdots H_0 A = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where  $R \in \mathbb{C}^{m \times n}$  is upper triangular, then

$$H_{n-1} \cdots H_0 A = \underbrace{H_0 \cdots H_{n-1}}_Q \begin{pmatrix} R \\ 0 \end{pmatrix} = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = \left( Q_L \mid Q_R \right) \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_L R,$$

where  $Q_L$  equals the first  $n$  columns of  $A$ . Then  $A = Q_L R$  is the QR factorization of  $A$ . The second fundamental insight will be that the desired unitary transformations  $\{H_0, \dots, H_{n-1}\}$  can be computed and applied cheaply.

## 6.2 Householder Transformations (Reflectors)

### 6.2.1 The general case

In this section we discuss *Householder transformations*, also referred to as *reflectors*.

**Definition 6.1** Let  $u \in \mathbb{C}^n$  be a vector of unit length ( $\|u\|_2 = 1$ ). Then  $H = I - 2uu^H$  is said to be a reflector or Householder transformation.

We observe:

- Any vector  $z$  that is perpendicular to  $u$  is left unchanged:

$$(I - 2uu^H)z = z - 2u(u^H z) = z.$$

- Any vector  $x$  can be written as  $x = z + u^H x u$  where  $z$  is perpendicular to  $u$  and  $u^H x u$  is the component of  $x$  in the direction of  $u$ . Then

$$(I - 2uu^H)x = (I - 2uu^H)(z + u^H x u) = z + u^H x u - 2u \underbrace{u^H z}_0 - 2uu^H u^H x u$$

$$= z + u^H x u - 2u^H x \underbrace{u^H u}_1 u = z - u^H x u.$$

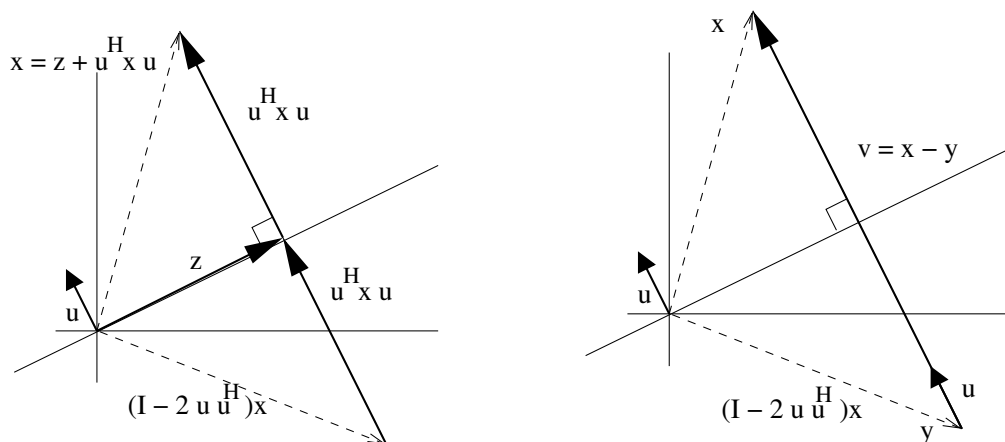


Figure 6.1: Left: Illustration that shows how, given vectors  $x$  and unit length vector  $u$ , the subspace orthogonal to  $u$  becomes a mirror for reflecting  $x$  represented by the transformation  $(I - 2uu^H)$ . Right: Illustration that shows how to compute  $u$  given vectors  $x$  and  $y$  with  $\|x\|_2 = \|y\|_2$ .

This can be interpreted as follows: The space perpendicular to  $u$  acts as a “mirror”: any vector in that space (along the mirror) is not reflected, while any other vector has the component that is orthogonal to the space (the component outside, orthogonal to, the mirror) reversed in direction, as illustrated in Figure 6.1. Notice that a reflection preserves the length of the vector.

**Homework 6.2** Show that if  $H$  is a reflector, then

- $HH = I$  (reflecting the reflection of a vector results in the original vector),
- $H = H^H$ , and
- $H^H H = I$  (a reflection is a unitary matrix and thus preserves the norm).

➡ SEE ANSWER

Next, let us ask the question of how to reflect a given  $x \in \mathbb{C}^n$  into another vector  $y \in \mathbb{C}^n$  with  $\|x\|_2 = \|y\|_2$ . In other words, how do we compute vector  $u$  so that  $(I - 2uu^H)x = y$ . From our discussion above, we need to find a vector  $u$  that is perpendicular to the space with respect to which we will reflect. From Figure 6.1(right) we notice that the vector from  $y$  to  $x$ ,  $v = x - y$ , is perpendicular to the desired space. Thus,  $u$  must equal a unit vector in the direction  $v$ :  $u = v/\|v\|_2$ .

**Remark 6.3** In subsequent discussion we will prefer to give Householder transformations as  $I - uu^H/\tau$ , where  $\tau = u^H u/2$  so that  $u$  needs no longer be a unit vector, just a direction. The reason for this will become obvious later.

In the next subsection, we will need to find a Householder transformation  $H$  that maps a vector  $x$  to a multiple of the first unit basis vector ( $e_0$ ).

Let us first discuss how to find  $H$  in the case where  $x \in \mathbb{R}^n$ . We seek  $v$  so that  $(I - \frac{2}{v^T v} vv^T)x = \pm\|x\|_2 e_0$ . Since the resulting vector that we want is  $y = \pm\|x\|_2 e_0$ , we must choose  $v = x - y = x \mp \|x\|_2 e_0$ .

**Homework 6.4** Show that if  $x \in \mathbb{R}^n$ ,  $v = x \mp \|x\|_2 e_0$ , and  $\tau = v^T v / 2$  then  $(I - \frac{1}{\tau} v v^T)x = \pm \|x\|_2 e_0$ .

🔗 [SEE ANSWER](#)

In practice, we choose  $v = x + \text{sign}(\chi_1)\|x\|_2 e_0$  where  $\chi_1$  denotes the first element of  $x$ . The reason is as follows: the first element of  $v$ ,  $v_1$ , will be  $v_1 = \chi_1 \mp \|x\|_2$ . If  $\chi_1$  is positive and  $\|x\|_2$  is almost equal to  $\chi_1$ , then  $\chi_1 - \|x\|_2$  is a small number and if there is error in  $\chi_1$  and/or  $\|x\|_2$ , this error becomes large *relative* to the result  $\chi_1 - \|x\|_2$ , due to catastrophic cancellation. Regardless of whether  $\chi_1$  is positive or negative, we can avoid this by choosing  $x = \chi_1 + \text{sign}(\chi_1)\|x\|_2 e_0$ .

### 6.2.2 As implemented for the Householder QR factorization (real case)

Next, we discuss a slight variant on the above discussion that is used in practice. To do so, we view  $x$  as a vector that consists of its first element,  $\chi_1$ , and the rest of the vector,  $x_2$ : More precisely, partition

$$x = \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix},$$

where  $\chi_1$  equals the first element of  $x$  and  $x_2$  is the rest of  $x$ . Then we will wish to find a Householder vector  $u = \begin{pmatrix} 1 \\ u_2 \end{pmatrix}$  so that

$$\left( I - \frac{1}{\tau} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^T \right) \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \pm \|x\|_2 \\ 0 \end{pmatrix}.$$

Notice that  $y$  in the previous discussion equals the vector  $\begin{pmatrix} \pm \|x\|_2 \\ 0 \end{pmatrix}$ , so the direction of  $u$  is given by

$$v = \begin{pmatrix} \chi_1 \mp \|x\|_2 \\ x_2 \end{pmatrix}.$$

We now wish to normalize this vector so its first entry equals “1”:

$$u = \frac{v}{v_1} = \frac{1}{\chi_1 \mp \|x\|_2} \begin{pmatrix} \chi_1 \mp \|x\|_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ x_2/v_1 \end{pmatrix}.$$

where  $v_1 = \chi_1 \mp \|x\|_2$  equals the first element of  $v$ . (Note that if  $v_1 = 0$  then  $u_2$  can be set to 0.)

### 6.2.3 The complex case (optional)

Next, let us work out the complex case, dealing explicitly with  $x$  as a vector that consists of its first element,  $\chi_1$ , and the rest of the vector,  $x_2$ : More precisely, partition

$$x = \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix},$$

where  $\chi_1$  equals the first element of  $x$  and  $x_2$  is the rest of  $x$ . Then we will wish to find a Householder vector  $u = \begin{pmatrix} 1 \\ u_2 \end{pmatrix}$  so that

$$\left( I - \frac{1}{\tau} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \right) \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \oplus \|x\|_2 \\ 0 \end{pmatrix}.$$

Here  $\oplus$  denotes a complex scalar on the complex unit circle. By the same argument as before

$$v = \begin{pmatrix} \chi_1 - \oplus \|x\|_2 \\ x_2 \end{pmatrix}.$$

We now wish to normalize this vector so its first entry equals “1”:

$$u = \frac{v}{\|v\|_2} = \frac{1}{\chi_1 - \oplus \|x\|_2} \begin{pmatrix} \chi_1 - \oplus \|x\|_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ x_2/v_1 \end{pmatrix}.$$

where  $v_1 = \chi_1 - \oplus \|x\|_2$ . (If  $v_1 = 0$  then we set  $u_2$  to 0.)

**Homework 6.5** Verify that

$$\left( I - \frac{1}{\tau} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \right) \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \rho \\ 0 \end{pmatrix}$$

where  $\tau = u^H u / 2 = (1 + u_2^H u_2) / 2$  and  $\rho = \oplus \|x\|_2$ .

Hint:  $\rho \bar{\rho} = |\rho|^2 = \|x\|_2^2$  since  $H$  preserves the norm. Also,  $\|x\|_2^2 = |\chi_1|^2 + \|x_2\|_2^2$  and  $\sqrt{\frac{z}{|z|}} = \frac{z}{|z|}$ .

SEE ANSWER

Again, the choice  $\oplus$  is important. For the complex case we choose  $\oplus = -\text{sign}(\chi_1) = \frac{\chi_1}{|\chi_1|}$

### 6.2.4 A routine for computing the Householder vector

We will refer to the vector

$$\begin{pmatrix} 1 \\ u_2 \end{pmatrix}$$

as the Householder vector that reflects  $x$  into  $\oplus \|x\|_2 e_0$  and introduce the notation

$$\left[ \begin{pmatrix} \rho \\ u_2 \end{pmatrix}, \tau \right] := \text{Housev} \left( \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} \right)$$

as the computation of the above mentioned vector  $u_2$ , and scalars  $\rho$  and  $\tau$ , from vector  $x$ . We will use the notation  $H(x)$  for the transformation  $I - \frac{1}{\tau} u u^H$  where  $u$  and  $\tau$  are computed by  $\text{Housev}(x)$ .



<b>Algorithm:</b> $\left[ \begin{pmatrix} \rho \\ u_2 \end{pmatrix}, \tau \right] = \text{Housev} \left( \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} \right)$	
$\rho = -\text{sign}(\chi_1) \ x\ _2$ $v_1 = \chi_1 + \text{sign}(\chi_1) \ x\ _2$ $u_2 = x_2 / v_1$ $\tau = (1 + u_2^H u_2) / 2$	$\chi_2 := \ x_2\ _2$ $\alpha := \left\  \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} \right\ _2 (= \ x\ _2)$ $\rho := -\text{sign}(\chi_1) \alpha$ $v_1 := \chi_1 - \rho$ $u_2 := x_2 / v_1$ $\chi_2 = \chi_2 /  v_1  (= \ u_2\ _2)$ $\tau = (1 + \chi_2^2) / 2$

Figure 6.2: Computing the Householder transformation. Left: simple formulation. Right: efficient computation. **Note: I have not completely double-checked these formulas for the complex case. They work for the real case.**

## 6.3 Householder QR Factorization

Let  $A$  be an  $m \times n$  with  $m \geq n$ . We will now show how to compute  $A \rightarrow QR$ , the QR factorization, as a sequence of Householder transformations applied to  $A$ , which eventually zeroes out all elements of that matrix below the diagonal. The process is illustrated in Figure 6.3.

In the first iteration, we partition

$$A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right).$$

Let

$$\left[ \begin{pmatrix} \rho_{11} \\ u_{21} \end{pmatrix}, \tau_1 \right] = \text{Housev} \left( \begin{pmatrix} \alpha_{11} \\ a_{21} \end{pmatrix} \right)$$

be the Householder transform computed from the first column of  $A$ . Then applying this Householder transform to  $A$  yields

$$\begin{aligned} \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) &:= \left( I - \frac{1}{\tau_1} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \right) \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) \\ &= \left( \begin{array}{c|c} \rho_{11} & a_{12}^T - w_{12}^T \\ \hline 0 & A_{22} - u_{21} w_{12}^T \end{array} \right), \end{aligned}$$

where  $w_{12}^T = (a_{12}^T + u_{21}^H A_{22}) / \tau_1$ . Computation of a full QR factorization of  $A$  will now proceed with the updated matrix  $A_{22}$ .

Figure 6.3: Illustration of Householder QR factorization.

Now let us assume that after  $k$  iterations of the algorithm matrix  $A$  contains

$$A \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & A_{BR} \end{array} \right) = \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right),$$

where  $R_{TL}$  and  $R_{00}$  are  $k \times k$  upper triangular matrices. Let

$$\left[ \left( \begin{array}{c} \rho_{11} \\ u_{21} \end{array} \right), \tau_1 \right] = \text{Housev} \left( \begin{array}{c} \alpha_{11} \\ a_{21} \end{array} \right).$$

and update

$$\begin{aligned} A &:= \left( \begin{array}{c|c} I & 0 \\ \hline 0 & \left( I - \frac{1}{\tau_1} \begin{pmatrix} 1 \\ u_{21} \end{pmatrix} \begin{pmatrix} 1 \\ u_{21} \end{pmatrix}^H \right) \end{array} \right) \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) \\ &= \left( \begin{array}{c|c} I - \frac{1}{\tau_1} \begin{pmatrix} 0 \\ 1 \\ u_{21} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ u_{21} \end{pmatrix}^H & \begin{pmatrix} R_{00} & r_{01} & R_{02} \\ 0 & \alpha_{11} & a_{12}^T \\ 0 & a_{21} & A_{22} \end{pmatrix} \end{array} \right) \\ &= \left( \begin{array}{c|c|c} R_{00} & r_{01} & R_{02} \\ \hline 0 & \rho_{11} & a_{12}^T - w_{12}^T \\ \hline 0 & 0 & A_{22} - u_{21} w_{12}^T \end{array} \right), \end{aligned}$$

where again  $w_{12}^T = (a_{12}^T + u_{21}^H A_{22})/\tau_1$ .

Let

$$H_k = \left( I - \frac{1}{\tau_1} \begin{pmatrix} 0_k \\ 1 \\ u_{21} \end{pmatrix} \begin{pmatrix} 0_k \\ 1 \\ u_{21} \end{pmatrix}^H \right)$$

be the Householder transform so computed during the  $(k+1)$ st iteration. Then upon completion matrix  $A$  contains

$$R = \left( \begin{array}{c} R_{TL} \\ 0 \end{array} \right) = H_{n-1} \cdots H_1 H_0 \hat{A}$$

where  $\hat{A}$  denotes the original contents of  $A$  and  $R_{TL}$  is an upper triangular matrix. Rearranging this we find that

$$\hat{A} = H_0 H_1 \cdots H_{n-1} R$$

which shows that if  $Q = H_0 H_1 \cdots H_{n-1}$  then  $\hat{A} = QR$ .

**Homework 6.6** Show that

$$\left( \begin{array}{c|c} I & 0 \\ \hline 0 & \left( I - \frac{1}{\tau_1} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \right) \end{array} \right) = \left( I - \frac{1}{\tau_1} \begin{pmatrix} 0 \\ 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ u_2 \end{pmatrix}^H \right).$$

🔗 SEE ANSWER

Typically, the algorithm overwrites the original matrix  $A$  with the upper triangular matrix, and at each step  $u_{21}$  is stored over the elements that become zero, thus overwriting  $a_{21}$ . (It is for this reason that the first element of  $u$  was normalized to equal “1”.) In this case  $Q$  is usually not explicitly formed as it can be stored as the separate Householder vectors below the diagonal of the overwritten matrix. The algorithm that overwrites  $A$  in this manner is given in Fig. 6.4.

We will let

$$[\{U \setminus R\}, t] = \text{HouseholderQR}(A)$$

denote the operation that computes the QR factorization of  $m \times n$  matrix  $A$ , with  $m \geq n$ , via Householder transformations. It returns the Householder vectors and matrix  $R$  in the first argument and the vector of scalars “ $\tau_i$ ” that are computed as part of the Householder transformations in  $t$ .

**Theorem 6.7** Given  $A \in \mathbb{C}^{m \times n}$  the cost of the algorithm in Figure 6.4 is given by

$$C_{\text{HQR}}(m, n) \approx 2mn^2 - \frac{2}{3}n^3 \text{ flops.}$$

**Proof:** The bulk of the computation is in  $w_{12}^T = (a_{12}^T + u_{21}^H A_{22})/\tau_1$  and  $A_{22} - u_{21}w_{12}^T$ . During the  $k$ th iteration (when  $R_{TL}$  is  $k \times k$ ), this means a matrix-vector multiplication ( $u_{21}^H A_{22}$ ) and rank-1 update with matrix  $A_{22}$  which is of size approximately  $(m-k) \times (n-k)$  for a cost of  $4(m-k)(n-k)$  flops. Thus the total cost is approximately

$$\begin{aligned} \sum_{k=0}^{n-1} 4(m-k)(n-k) &= 4 \sum_{j=0}^{n-1} (m-n+j)j = 4(m-n) \sum_{j=0}^{n-1} j + 4 \sum_{j=0}^{n-1} j^2 \\ &= 2(m-n)n(n-1) + 4 \sum_{j=0}^{n-1} j^2 \\ &\approx 2(m-n)n^2 + 4 \int_0^n x^2 dx = 2mn^2 - 2n^3 + \frac{4}{3}n^3 = 2mn^2 - \frac{2}{3}n^3. \end{aligned}$$

## 6.4 Forming $Q$

Given  $A \in \mathbb{C}^{m \times n}$ , let  $[A, t] = \text{HouseholderQR}(A)$  yield the matrix  $A$  with the Householder vectors stored below the diagonal,  $R$  stored on and above the diagonal, and the  $\tau_i$  stored in vector  $t$ . We now discuss how to form the first  $n$  columns of  $Q = H_0 H_1 \cdots H_{n-1}$ . The computation is illustrated in Figure 6.5.

**Algorithm:**  $[A, t] = \text{HOUSEQR\_UNB\_VAR1}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)$  and  $t \rightarrow \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$  and  $t_T$  has 0 elements

**while**  $n(A_{BR}) \neq 0$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right) \text{ and } \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \rightarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right)$$

**where**  $\alpha_{11}$  and  $\tau_1$  are scalars

$$\left[ \left( \begin{array}{c} \alpha_{11} \\ \hline a_{21} \end{array} \right), \tau_1 \right] := \left[ \left( \begin{array}{c} \rho_{11} \\ \hline u_{21} \end{array} \right), \tau_1 \right] = \text{Housev} \left( \begin{array}{c} \alpha_{11} \\ \hline a_{21} \end{array} \right)$$

$$\text{Update } \left( \begin{array}{c} a_{12}^T \\ \hline A_{22} \end{array} \right) := \left( I - \frac{1}{\tau_1} \left( \begin{array}{c} 1 \\ \hline u_{21} \end{array} \right) \left( 1 \mid u_{21}^H \right) \right) \left( \begin{array}{c} a_{12}^T \\ \hline A_{22} \end{array} \right)$$

via the steps

- $w_{12}^T := (a_{12}^T + a_{21}^H A_{22}) / \tau_1$
- $\left( \begin{array}{c} a_{12}^T \\ \hline A_{22} \end{array} \right) := \left( \begin{array}{c} a_{12}^T - w_{12}^T \\ \hline A_{22} - a_{21} w_{12}^T \end{array} \right)$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right) \text{ and } \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \leftarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right)$$

**endwhile**

Figure 6.4: Unblocked Householder transformation based QR factorization.

Original matrix	$\left( \begin{array}{c c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) :=$ $\left( \begin{array}{c c} 1 - 1/\tau_1 & -(u_{21}^H A_{22})/\tau_1 \\ \hline -u_{21}/\tau_1 & A_{22} + u_{21}a_{12}^T \end{array} \right)$	“Move forward”
$\begin{array}{cccc c} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \\ 0 & 0 & 0 & 1 & \\ \hline 0 & 0 & 0 & 0 & \end{array}$	$\begin{array}{cccc c} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \\ \hline 0 & 0 & 0 & \times & \\ 0 & 0 & 0 & \times & \end{array}$	$\begin{array}{cccc c} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \\ \hline 0 & 0 & 0 & \times & \\ 0 & 0 & 0 & \times & \end{array}$
	$\begin{array}{ccc cc} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \\ \hline 0 & 0 & \times & \times & \\ 0 & 0 & \times & \times & \\ 0 & 0 & \times & \times & \end{array}$	$\begin{array}{ccc cc} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \\ \hline 0 & 0 & \times & \times & \\ 0 & 0 & \times & \times & \\ 0 & 0 & \times & \times & \end{array}$
	$\begin{array}{c ccc} 1 & 0 & 0 & 0 \\ \hline 0 & \times & \times & \times \\ \hline 0 & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \end{array}$	$\begin{array}{c cccc} 1 & 0 & 0 & 0 \\ \hline 0 & \times & \times & \times \\ \hline 0 & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \end{array}$
	$\begin{array}{c cccc} \times & \times & \times & \times \\ \hline \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{array}$	$\begin{array}{cccc} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{array}$

Figure 6.5: Illustration of the computation of  $Q$ .

Notice that to pick out the first  $n$  columns we must form

$$Q \begin{pmatrix} I_{n \times n} \\ 0 \end{pmatrix} = H_0 \cdots H_{n-1} \begin{pmatrix} I_{n \times n} \\ 0 \end{pmatrix} = H_0 \cdots H_{k-1} \underbrace{H_k \cdots H_{n-1}}_{B_k} \begin{pmatrix} I_{n \times n} \\ 0 \end{pmatrix}.$$

where  $B_k$  is defined as indicated.

**Lemma 6.8**  $B_k$  has the form

$$B_k = H_k \cdots H_{n-1} \left( \begin{array}{c|c} I_{n \times n} & \\ \hline 0 & \end{array} \right) = \left( \begin{array}{c|c} I_{k \times k} & 0 \\ \hline 0 & \tilde{B}_k \end{array} \right).$$

**Proof:** The proof of this is by induction on  $k$ :

- **Base case:**  $k = n$ . Then  $B_n = \left( \begin{array}{c|c} I_{n \times n} & \\ \hline 0 & \end{array} \right)$ , which has the desired form.
- **Inductive step:** Assume the result is true for  $B_k$ . We show it is true for  $B_{k-1}$ :

$$\begin{aligned} B_{k-1} &= H_{k-1} H_k \cdots H_{n-1} \left( \begin{array}{c|c} I_{n \times n} & \\ \hline 0 & \end{array} \right) = H_{k-1} B_k = H_{k-1} \left( \begin{array}{c|c} I_{k \times k} & 0 \\ \hline 0 & \tilde{B}_k \end{array} \right). \\ &= \left( \begin{array}{c|c} I_{(k-1) \times (k-1)} & 0 \\ \hline 0 & I - \frac{1}{\tau_k} \left( \begin{array}{c} 1 \\ u_k \end{array} \right) \left( \begin{array}{c|c} 1 & u_k^H \end{array} \right) \end{array} \right) \left( \begin{array}{c|c|c} I_{(k-1) \times (k-1)} & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & 0 & \tilde{B}_k \end{array} \right) \\ &= \left( \begin{array}{c|c} I_{(k-1) \times (k-1)} & 0 \\ \hline 0 & \left( I - \frac{1}{\tau_k} \left( \begin{array}{c} 1 \\ u_k \end{array} \right) \left( \begin{array}{c|c} 1 & u_k^H \end{array} \right) \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \tilde{B}_k \end{array} \right) \end{array} \right) \\ &= \left( \begin{array}{c|c} I_{(k-1) \times (k-1)} & 0 \\ \hline 0 & \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \tilde{B}_k \end{array} \right) - \left( \begin{array}{c} 1 \\ u_k \end{array} \right) \left( \begin{array}{c|c} 1/\tau_k & y_k^T \end{array} \right) \end{array} \right) \quad \text{where } y_k^T = u_k^H \tilde{B}_k / \tau_k \\ &= \left( \begin{array}{c|c} I_{(k-1) \times (k-1)} & 0 \\ \hline 0 & \left( \begin{array}{c|c} 1 - 1/\tau_k & -y_k^T \\ \hline -u_k/\tau_k & B_k - u_k y_k^T \end{array} \right) \end{array} \right) \\ &= \left( \begin{array}{c|c|c} I_{(k-1) \times (k-1)} & 0 & 0 \\ \hline 0 & 1 - 1/\tau_k & -y_k^T \\ \hline 0 & -u_k/\tau_k & B_k - u_k y_k^T \end{array} \right) = \left( \begin{array}{c|c} I_{(k-1) \times (k-1)} & 0 \\ \hline 0 & \tilde{B}_{k-1} \end{array} \right). \end{aligned}$$

- **By the Principle of Mathematical Induction** the result holds for  $B_0, \dots, B_n$ .

**Theorem 6.9** Given  $[A, t] = \text{HouseholderQR}(A)$  from Figure 6.4, the algorithm in Figure 6.6 overwrites  $A$  with the first  $n = n(A)$  columns of  $Q$  as defined by the Householder transformations stored below the diagonal of  $A$  and in the vector  $t$ .

**Algorithm:**  $[A] = \text{FORMQ}(A, t)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)$  and  $t \rightarrow \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right)$   
**where**  $A_{TL}$  is  $n(A) \times n(A)$  and  $t_T$  has  $n(A)$  elements

**while**  $n(A_{TR}) \neq 0$  **do**

**Repartitionition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right) \text{ and } \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \rightarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right)$$

**where**  $\alpha_{11}$  and  $\tau_1$  are scalars

$$\text{Update } \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) := \left( I - \frac{1}{\tau_1} \left( \begin{array}{c} 1 \\ \hline u_{21} \end{array} \right) \left( \begin{array}{c|c} 1 & u_{21}^H \end{array} \right) \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & A_{22} \end{array} \right)$$

via the steps

- $\alpha_{11} := 1 - 1/\tau_1$
- $a_{12}^T := -(a_{21}^H A_{22})/\tau_1$
- $A_{22} := A_{22} + a_{21} a_{12}^T$
- $a_{21} := -a_{21}/\tau_1$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right) \text{ and } \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \leftarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right)$$

**endwhile**

Figure 6.6: Algorithm for overwriting  $A$  with  $Q$  from the Householder transformations stored as Householder vectors below the diagonal of  $A$  (as produced by  $[A, t] = \text{HouseholderQR}(A)$  ).



**Proof:** The algorithm is justified by the proof of Lemma 6.8.

**Theorem 6.10** Given  $A \in \mathbb{C}^{m \times n}$  the cost of the algorithm in Figure 6.6 is given by

$$C_{\text{FormQ}}(m, n) \approx 2mn^2 - \frac{2}{3}n^3 \text{ flops.}$$

**Proof:** Hence the proof for Theorem 6.7 can be easily modified to establish this result.

**Homework 6.11** If  $m = n$  then  $Q$  could be accumulated by the sequence

$$Q = (\cdots ((IH_0)H_1) \cdots H_{n-1}).$$

Give a high-level reason why this would be (much) more expensive than the algorithm in Figure 6.6.

SEE ANSWER

## 6.5 Applying $Q^H$

In a future Note, we will see that the QR factorization is used to solve the linear least-squares problem. To do so, we need to be able to compute  $\hat{y} = Q^H y$  where  $Q^H = H_{n-1} \cdots H_0$ .

Let us start by computing  $H_0 y$ :

$$\begin{aligned} \left( I - \frac{1}{\tau_1} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \right) \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix} &= \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \underbrace{\begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix}}_{\omega_1} / \tau_1 \\ &= \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix} - \omega_1 \begin{pmatrix} 1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \psi_1 - \omega_1 \\ y_2 - \omega_1 u_2 \end{pmatrix}. \end{aligned}$$

More generally, let us compute  $H_k y$ :

$$\left( I - \frac{1}{\tau_1} \begin{pmatrix} 0 \\ 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ u_2 \end{pmatrix}^H \right) \begin{pmatrix} y_0 \\ \psi_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ \psi_1 - \omega_1 \\ y_2 - \omega_1 u_2 \end{pmatrix},$$

where  $\omega_1 = (\psi_1 + u_2^H y_2) / \tau_1$ . This motivates the algorithm in Figure 6.7 for computing  $y := H_{n-1} \cdots H_0 y$  given the output matrix  $A$  and vector  $t$  from routine HouseholderQR.

The cost of this algorithm can be analyzed as follows: When  $y_T$  is of length  $k$ , the bulk of the computation is in an inner product with vectors of length  $m - k$  (to compute  $\omega_1$ ) and an axpy operation with vectors of length  $m - k$  to subsequently update  $\psi_1$  and  $y_2$ . Thus, the cost is approximately given by

$$\sum_{k=0}^{n-1} 4(m - k) \approx 4mn - 2n^2.$$

Notice that this is *much* cheaper than forming  $Q$  and then multiplying.

**Algorithm:**  $[y] = \text{APPLYQT}(A, t, y)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)$ ,  $t \rightarrow \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right)$ , and  $y \rightarrow \left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$  and  $t_T, y_T$  has 0 elements

**while**  $n(A_{BR}) \neq 0$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \rightarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right), \text{ and } \left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \rightarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

**where**  $\alpha_{11}$ ,  $\tau_1$ , and  $\psi_1$  are scalars

---


$$\text{Update } \left( \begin{array}{c} \psi_1 \\ \hline y_2 \end{array} \right) := \left( I - \frac{1}{\tau_1} \left( \begin{array}{c} 1 \\ \hline u_{21} \end{array} \right) \left( 1 \mid u_{21}^H \right) \right) \left( \begin{array}{c} \psi_1 \\ \hline y_2 \end{array} \right)$$

via the steps

- $\omega_1 := (\psi_1 + a_{21}^H y_2) / \tau_1$
  - $\left( \begin{array}{c} \psi_1 \\ \hline y_2 \end{array} \right) := \left( \begin{array}{c} \psi_1 - \omega_1 \\ \hline y_2 - \omega_1 u_2 \end{array} \right)$
- 

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \leftarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right), \text{ and } \left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \leftarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

**endwhile**

Figure 6.7: Algorithm for computing  $y := H_{n-1} \cdots H_0 y$  given the output from routine HouseholderQR.

## 6.6 Blocked Householder QR Factorization

### 6.6.1 The UT transform: Accumulating Householder transformations

Through a series of exercises, we will show how the application of a sequence of  $k$  Householder transformations can be accumulated. What we exactly mean that this will become clear.

**Homework 6.12** Assuming all inverses exist, show that

$$\left( \begin{array}{c|c} T_{00} & t_{01} \\ \hline 0 & \tau_1 \end{array} \right)^{-1} = \left( \begin{array}{c|c} T_{00}^{-1} & -T_{00}^{-1}t_{01}/\tau_1 \\ \hline 0 & 1/\tau_1 \end{array} \right).$$

➡ SEE ANSWER

An algorithm that computes the inverse of an upper triangular matrix  $T$  based on the above exercise is given in Figure 6.8.

**Homework 6.13 Homework 6.14** Consider  $u_1 \in \mathbb{C}^m$  with  $u_1 \neq 0$  (the zero vector),  $U_0 \in \mathbb{C}^{m \times k}$ , and non-singular  $T_{00} \in \mathbb{C}^{k \times k}$ . Define  $\tau_1 = (u_1^H u_1)/2$ , so that

$$H_1 = I - \frac{1}{\tau_1} u_1 u_1^H$$

equals a Householder transformation, and let

$$Q_0 = I - U_0 T_{00}^{-1} U_0^H.$$

Show that

$$Q_0 H_1 = (I - U_0 T_{00}^{-1} U_0^H) \left( I - \frac{1}{\tau_1} u_1 u_1^H \right) = I - \left( \begin{array}{c|c} U_0 & u_1 \end{array} \right) \left( \begin{array}{c|c} T_{00} & t_{01} \\ \hline 0 & \tau_1 \end{array} \right)^{-1} \left( \begin{array}{c|c} U_0 & u_1 \end{array} \right)^H,$$

where  $t_{01} = Q_0^H u_1$ .

➡ SEE ANSWER

**Homework 6.15 Homework 6.16** Consider  $u_i \in \mathbb{C}^m$  with  $u_i \neq 0$  (the zero vector). Define  $\tau_i = (u_i^H u_i)/2$ , so that

$$H_i = I - \frac{1}{\tau_i} u_i u_i^H$$

equals a Householder transformation, and let

$$U = \left( \begin{array}{c|c|c|c} u_0 & u_1 & \cdots & u_{k-1} \end{array} \right).$$

Show that

$$H_0 H_1 \cdots H_{k-1} = I - U T^{-1} U^H,$$

where  $T$  is an upper triangular matrix.

➡ SEE ANSWER

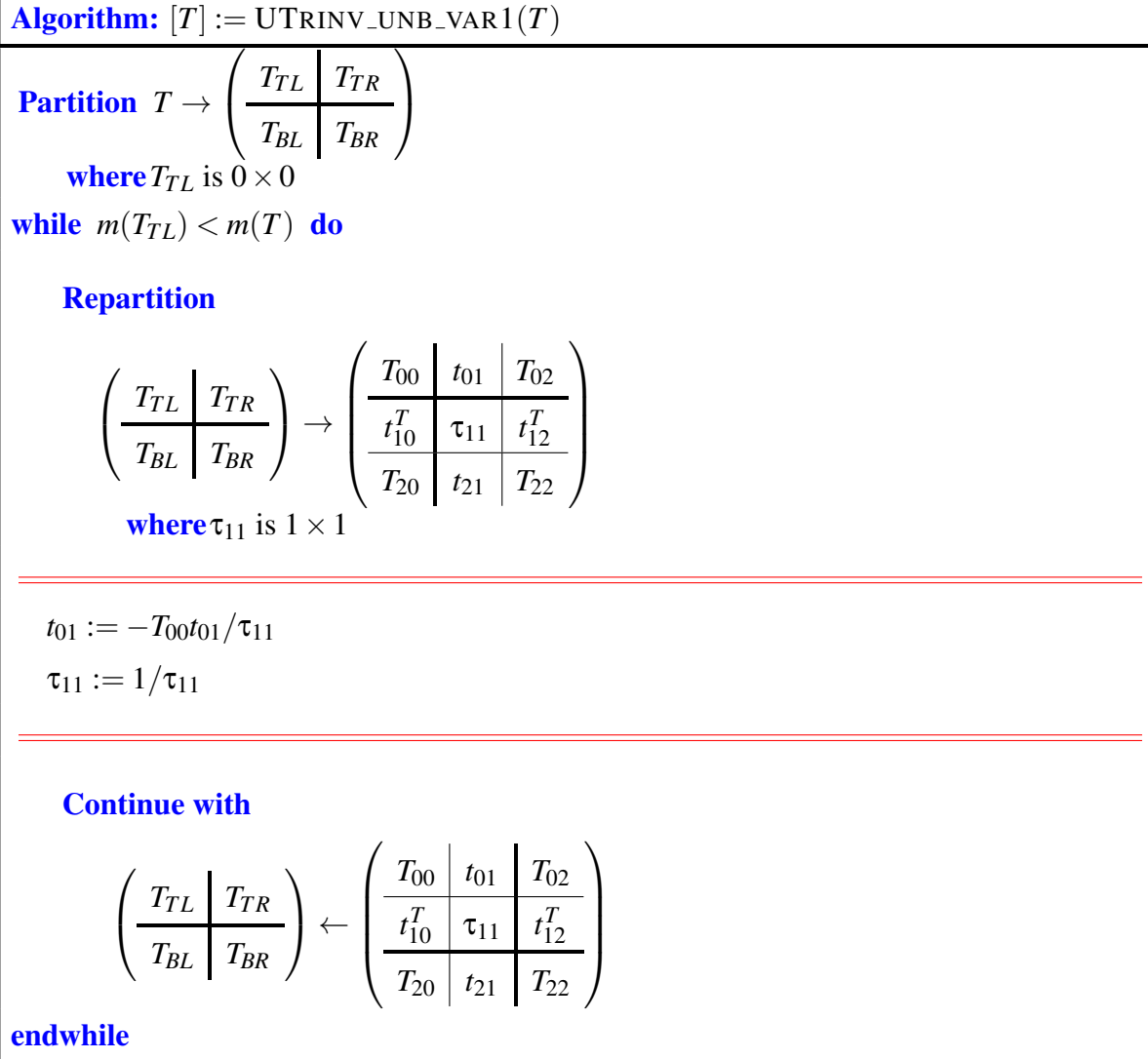


Figure 6.8: Unblocked algorithm for inverting an upper triangular matrix. The algorithm assumes that  $T_{00}$  has already been inverted, and computes the next column of  $T$  in the current iteration.

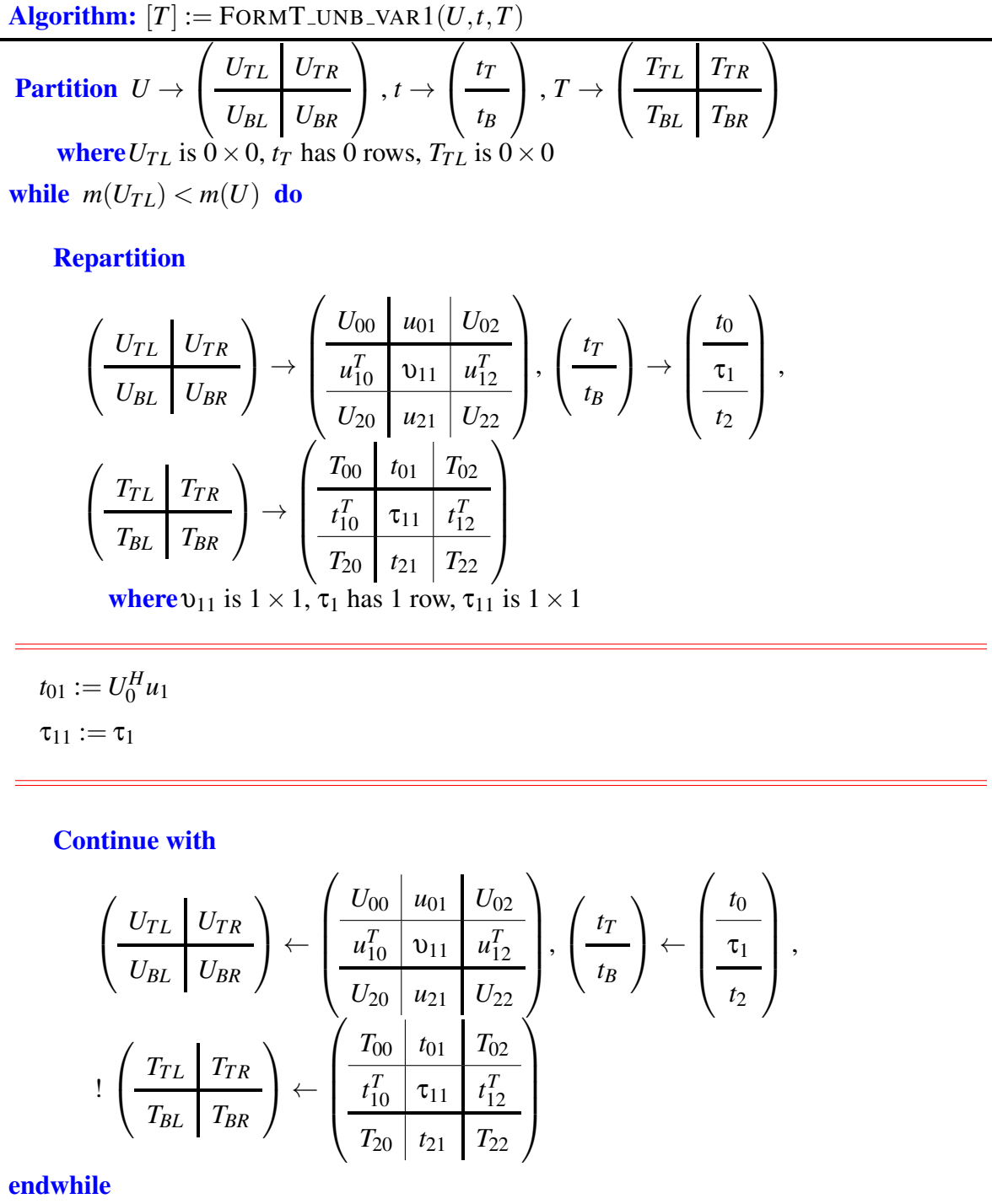


Figure 6.9: Algorithm that computes  $T$  from  $U$  and the vector  $t$  so that  $I - UTU^H$  equals the UT transform. Here  $U$  is assumed to be the output of, for example, an unblocked Householder based QR algorithm. This means that it is a lower trapezoidal matrix, with ones on the diagonal.

The above exercises can be summarized in the algorithm for computing  $T$  from  $U$  in Figure 6.9.

In [27] we call the transformation  $I - UT^{-1}U^H$  that equals the accumulated Householder transformations the *UT transform* and prove that  $T$  can instead be computed as

$$T = \text{triu}(U^H U)$$

(the upper triangular part of  $U^H U$ ) followed by either dividing the diagonal elements by two or setting them to  $\tau_0, \dots, \tau_{k-1}$  (in order). In that paper, we point out similar published results [9, 33, 44, 31].

### 6.6.2 A blocked algorithm

A QR factorization that exploits the insights that resulted in the UT transform can now be described:

- Partition

$$A \rightarrow \left( \begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right)$$

where  $A_{11}$  is  $b \times b$ .

- We can use the unblocked algorithm in Figure 6.4 to factor the panel  $\begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix}$

$$\left[ \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix}, t_1 \right] := \text{HOUSEQR\_UNB\_VAR1} \left( \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix} \right),$$

overwriting the entries below the diagonal with the Householder vectors  $\begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix}$  (with the ones on the diagonal implicitly stored) and the upper triangular part with  $R_{11}$ .

- For  $T_{11}$  from the Householder vectors using the procedure described in Section 6.6.1:

$$T_{11} := \text{FORMAT} \left( \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix}, t_1 \right)$$

- Now we need to also apply the Householder transformations to the rest of the columns:

$$\begin{aligned} \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} &:= \left( I - \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix} T_{11}^{-1} \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix}^H \right)^H \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} \\ &= \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} - \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix} W_{21}^H \\ &= \begin{pmatrix} A_{12} - U_{11} W_{21}^H \\ A_{22} - U_{21} W_{21}^H \end{pmatrix}, \end{aligned}$$

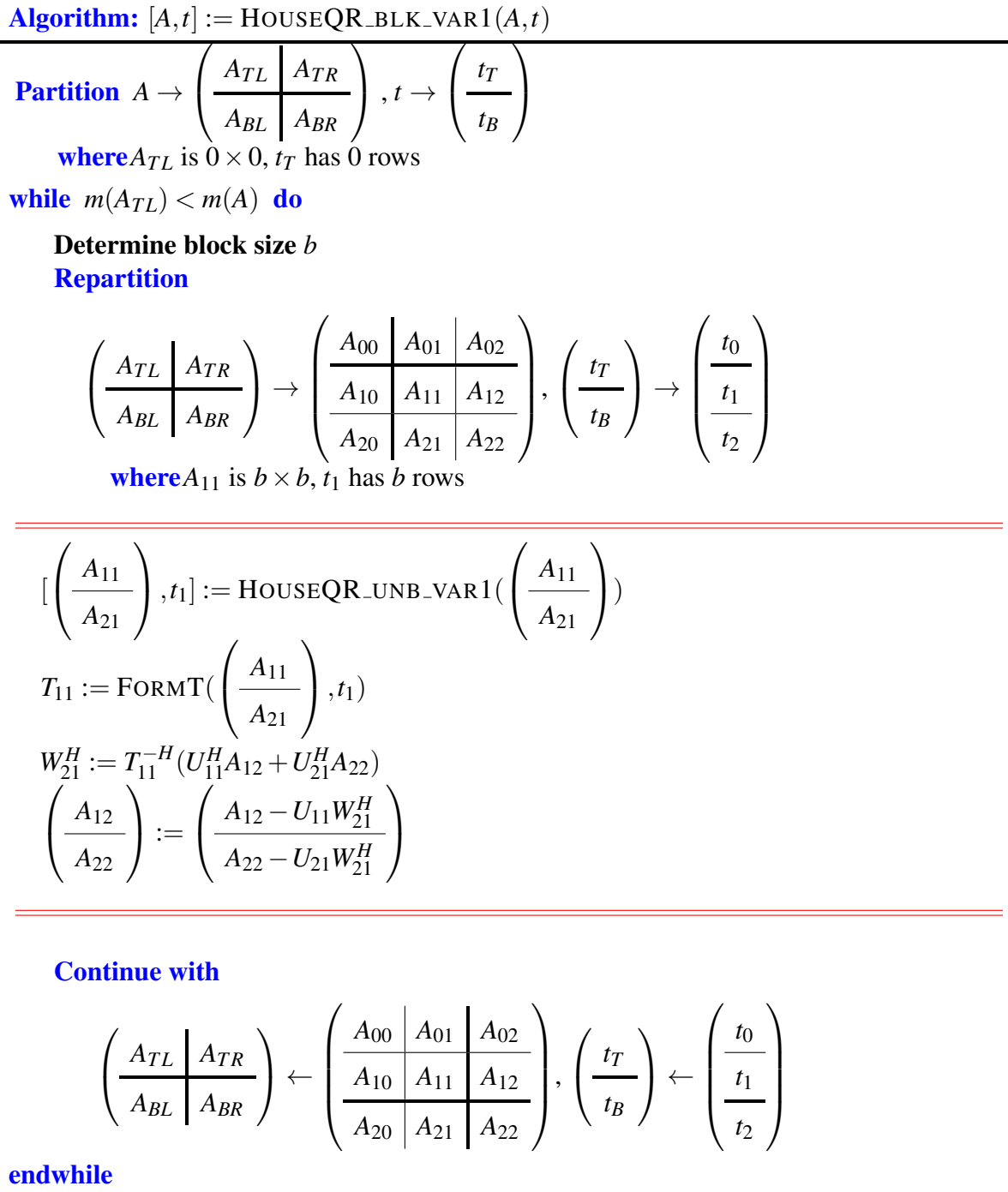


Figure 6.10: Blocked Householder transformation based QR factorization.

where

$$W_{21}^H = T_{11}^{-H}(U_{11}^H A_{12} + U_{21}^H A_{22}).$$

This motivates the blocked algorithm in Figure 6.10.

### 6.6.3 Variations on a theme

#### Merging the unblocked Householder QR factorization and the formation of $T$

There are many possible algorithms for computing the QR factorization. For example, the unblocked algorithm from Figure 6.4 can be merged with the unblocked algorithm for forming  $T$  in Figure 6.9 to yield the algorithm in Figure 6.11.

#### An alternative unblocked merged algorithm

Let us now again compute the QR factorization of  $A$  simultaneous with the forming of  $T$ , but now taking advantage of the fact that  $T$  is partially computed to change the algorithm into what some would consider a “left-looking” algorithm.

Partition

$$A \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right) \quad \text{and} \quad T \rightarrow \left( \begin{array}{c|c|c} T_{00} & t_{01} & T_{02} \\ \hline 0 & \tau_{11} & t_{12}^T \\ \hline 0 & 0 & T_{22} \end{array} \right).$$

Assume that  $\begin{pmatrix} A_{00} \\ \hline a_{10}^T \\ \hline A_{20} \end{pmatrix}$  has been factored and overwritten with  $\begin{pmatrix} U_{00} \\ \hline u_{10}^T \\ \hline U_{20} \end{pmatrix}$  and  $R_{00}$  while also computing

$T_{00}$ . In the next step, we need to apply previous Householder transformations to the next column of  $A$  and then update that column with the next column of  $R$  and  $U$ . In addition, the next column of  $T$  must be computing. This means:

- Update

$$\begin{pmatrix} a_{01} \\ \hline \alpha_{11} \\ \hline a_{21} \end{pmatrix} := \left( I - \begin{pmatrix} U_{00} \\ \hline u_{10}^T \\ \hline U_{20} \end{pmatrix} T_{00}^{-1} \begin{pmatrix} U_{00} \\ \hline u_{10}^T \\ \hline U_{20} \end{pmatrix}^H \right)^H \begin{pmatrix} a_{01} \\ \hline \alpha_{11} \\ \hline a_{21} \end{pmatrix}$$

- Compute the next Householder transform:

$$\left[ \begin{pmatrix} \alpha_{11} \\ \hline a_{21} \end{pmatrix}, \tau_{11} \right] := \text{Housev}\left( \begin{pmatrix} \alpha_{11} \\ \hline a_{21} \end{pmatrix} \right)$$

- Compute the rest of the next column of  $T$

$$t_{01} := A_{20}^H a_{21}$$

This yields the algorithm in Figure 6.12.



**Algorithm:**  $[A, T] := \text{HOUSEQR\_AND\_FORMT\_UNB\_VAR1}(A, T)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), T \rightarrow \left( \begin{array}{c|c} T_{TL} & T_{TR} \\ \hline T_{BL} & T_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$ ,  $T_{TL}$  is  $0 \times 0$

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} T_{TL} & T_{TR} \\ \hline T_{BL} & T_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} T_{00} & t_{01} & T_{02} \\ \hline t_{10}^T & \tau_{11} & t_{12}^T \\ \hline T_{20} & t_{21} & T_{22} \end{array} \right)$$

**where**  $\alpha_{11}$  is  $1 \times 1$ ,  $\tau_{11}$  is  $1 \times 1$

$$\left[ \left( \begin{array}{c} \alpha_{11} \\ a_{21} \end{array} \right), \tau_{11} \right] := \left[ \left( \begin{array}{c} \rho_{11} \\ u_{21} \end{array} \right), \tau_{11} \right] = \text{Housev} \left( \begin{array}{c} \alpha_{11} \\ a_{21} \end{array} \right)$$

Update  $\left( \begin{array}{c} a_{12}^T \\ A_{22} \end{array} \right) := \left( I - \frac{1}{\tau_{11}} \left( \begin{array}{c} 1 \\ u_{21} \end{array} \right) \left( 1 \mid u_{21}^H \right) \right) \left( \begin{array}{c} a_{12}^T \\ A_{22} \end{array} \right)$   
via the steps

- $w_{12}^T := (a_{12}^T + a_{21}^H A_{22}) / \tau_{11}$

- $\left( \begin{array}{c} a_{12}^T \\ A_{22} \end{array} \right) := \left( \begin{array}{c} a_{12}^T - w_{12}^T \\ A_{22} - a_{21} w_{12}^T \end{array} \right)$

$$t_{01} := A_{20}^H a_{21}$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} T_{TL} & T_{TR} \\ \hline T_{BL} & T_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} T_{00} & t_{01} & T_{02} \\ \hline t_{10}^T & \tau_{11} & t_{12}^T \\ \hline T_{20} & t_{21} & T_{22} \end{array} \right)$$

**endwhile**

Figure 6.11: Unblocked Householder transformation based QR factorization merged with the computation of  $T$  for the UT transform.

**Algorithm:**  $[A, T] := \text{HOUSEQR\_AND\_FORMT\_UNB\_VAR2}(A, T)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), T \rightarrow \left( \begin{array}{c|c} T_{TL} & T_{TR} \\ \hline T_{BL} & T_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$ ,  $T_{TL}$  is  $0 \times 0$

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} T_{TL} & T_{TR} \\ \hline T_{BL} & T_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} T_{00} & t_{01} & T_{02} \\ \hline t_{10}^T & \tau_{11} & t_{12}^T \\ \hline T_{20} & t_{21} & T_{22} \end{array} \right)$$

**where**  $\alpha_{11}$  is  $1 \times 1$ ,  $\tau_{11}$  is  $1 \times 1$

$$\left( \begin{array}{c} a_{01} \\ \alpha_{11} \\ a_{02} \end{array} \right) := \left( I - \left( \begin{array}{c} U_{00} \\ u_{10}^T \\ U_{20} \end{array} \right) T_{00}^{-1} \left( \begin{array}{c} U_{00} \\ u_{10}^T \\ U_{20} \end{array} \right)^H \right)^H \left( \begin{array}{c} a_{01} \\ \alpha_{11} \\ a_{21} \end{array} \right)$$

via the steps

$$\bullet w_{01} := T_{00}^{-H} (U_{00}^H a_{01} + \alpha_{11} (u_{10}^T)^H + U_{20}^H a_{21})$$

$$\bullet \left( \begin{array}{c} a_{01} \\ \alpha_{11} \\ a_{02} \end{array} \right) := \left( \begin{array}{c} a_{01} \\ \alpha_{11} \\ a_{02} \end{array} \right) - \left( \begin{array}{c} U_{00} \\ u_{10}^T \\ U_{20} \end{array} \right) w_{01}$$

$$\left[ \left( \begin{array}{c} \alpha_{11} \\ a_{21} \end{array} \right), \tau_{11} \right] := \left[ \left( \begin{array}{c} \rho_{11} \\ u_{21} \end{array} \right), \tau_{11} \right] = \text{Housev} \left( \begin{array}{c} \alpha_{11} \\ a_{21} \end{array} \right)$$

$t_{01} := A_{20}^H a_{21}$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} T_{TL} & T_{TR} \\ \hline T_{BL} & T_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} T_{00} & t_{01} & T_{02} \\ \hline t_{10}^T & \tau_{11} & t_{12}^T \\ \hline T_{20} & t_{21} & T_{22} \end{array} \right)$$

**endwhile**

Figure 6.12: Alternative unblocked Householder transformation based QR factorization merged with the computation of  $T$  for the UT transform.

**Alternative blocked algorithm (Variant 2)**

An alternative blocked variant that uses either of the unblocked factorization routines that merges the formation of  $T$  is given in [Figure 6.13](#).

**Algorithm:**  $[A, t] := \text{HOUSEQR\_BLK\_VAR1}(A, t)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), t \rightarrow \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$ ,  $t_T$  has 0 rows

**while**  $m(A_{TL}) < m(A)$  **do**

**Determine block size**  $b$

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \rightarrow \left( \begin{array}{c} t_0 \\ \hline t_1 \\ \hline t_2 \end{array} \right)$$

**where**  $A_{11}$  is  $b \times b$ ,  $t_1$  has  $b$  rows

$$\left[ \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right), T_{11} \right] := \text{HOUSEQR\_FORMT\_UNB\_VARX} \left( \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right) \right)$$

$$W_{21}^H := T_{11}^{-H} (U_{11}^H A_{12} + U_{21}^H A_{22})$$

$$\left( \begin{array}{c} A_{12} \\ \hline A_{22} \end{array} \right) := \left( \begin{array}{c} A_{12} - U_{11} W_{21}^H \\ \hline A_{22} - U_{21} W_{21}^H \end{array} \right)$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \leftarrow \left( \begin{array}{c} t_0 \\ \hline t_1 \\ \hline t_2 \end{array} \right)$$

**endwhile**

Figure 6.13: Alternative blocked Householder transformation based QR factorization.

## Notes on Solving Linear Least-Squares Problems

For a motivation of the linear least-squares problem, read Week 10 (Sections 10.3-10.5) of [Linear Algebra: Foundations to Frontiers - Notes to LAFF With \[29\]](#).

### Video

Read disclaimer regarding the videos in the preface!

No video... Camera ran out of memory...

## Outline

<b>Video</b> . . . . .	<b>109</b>
<b>Outline</b> . . . . .	<b>110</b>
<b>7.1. The Linear Least-Squares Problem</b> . . . . .	<b>111</b>
<b>7.2. Method of Normal Equations</b> . . . . .	<b>111</b>
<b>7.3. Solving the LLS Problem Via the QR Factorization</b> . . . . .	<b>112</b>
7.3.1. Simple derivation of the solution . . . . .	112
7.3.2. Alternative derivation of the solution . . . . .	113
<b>7.4. Via Householder QR Factorization</b> . . . . .	<b>114</b>
<b>7.5. Via the Singular Value Decomposition</b> . . . . .	<b>115</b>
7.5.1. Simple derivation of the solution . . . . .	115
7.5.2. Alternative derivation of the solution . . . . .	116
<b>7.6. What If <math>A</math> Does Not Have Linearly Independent Columns?</b> . . . . .	<b>116</b>
<b>7.7. Exercise: Using the the <math>LQ</math> factorization to solve underdetermined systems</b> . . . . .	<b>123</b>

## 7.1 The Linear Least-Squares Problem

Let  $A \in \mathbb{C}^{m \times n}$  and  $y \in \mathbb{C}^m$ . Then the linear least-square problem (LLS) is given by

$$\text{Find } x \text{ s.t. } \|Ax - y\|_2 = \min_{z \in \mathbb{C}^n} \|Az - y\|_2.$$

In other words,  $x$  is the vector that minimizes the expression  $\|Ax - y\|_2$ . Equivalently, we can solve

$$\text{Find } x \text{ s.t. } \|Ax - y\|_2^2 = \min_{z \in \mathbb{C}^n} \|Az - y\|_2^2.$$

If  $x$  solves the linear least-squares problem, then  $Ax$  is the vector in  $C(A)$  (the column space of  $A$ ) closest to the vector  $y$ .

## 7.2 Method of Normal Equations

Let  $A \in \mathbb{R}^{m \times n}$  have linearly independent columns (which implies  $m \geq n$ ). Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be defined by

$$\begin{aligned} f(x) &= \|Ax - y\|_2^2 = (Ax - y)^T (Ax - y) = x^T A^T A x - x^T A^T y - y^T A x + y^T y \\ &= x^T A^T A x - 2x^T A^T y + y^T y. \end{aligned}$$

This function is minimized when the gradient is zero,  $\nabla f(x) = 0$ . Now,

$$\nabla f(x) = 2A^T A x - 2A^T y.$$

If  $A$  has linearly independent columns then  $A^T A$  is nonsingular. Hence, the  $x$  that minimizes  $\|Ax - y\|_2$  solves  $A^T A x = A^T y$ . This is known as the method of normal equations. Notice that then

$$x = \underbrace{(A^T A)^{-1} A^T}_{A^\dagger} y,$$

where  $A^\dagger$  is known as the *pseudo inverse* or *Moore-Penrose pseudo inverse*.

In practice, one performs the following steps:

- Form  $B = A^T A$ , a symmetric positive-definite (SPD) matrix.  
Cost: approximately  $mn^2$  floating point operations (flops), if one takes advantage of symmetry.
- Compute the Cholesky factor  $L$ , a lower triangular matrix, so that  $B = LL^T$ .  
This factorization, discussed in Week 8 (Section 8.4.2) of [Linear Algebra: Foundations to Frontiers - Notes to LAFF With](#) and to be revisited later in this course, exists since  $B$  is SPD.  
Cost: approximately  $\frac{1}{3}n^3$  flops.
- Compute  $\hat{y} = A^T y$ .  
Cost:  $2mn$  flops.
- Solve  $Lz = \hat{y}$  and  $L^T x = z$ .  
Cost:  $n^2$  flops each.

Thus, the total cost of solving the LLS problem via normal equations is approximately  $mn^2 + \frac{1}{3}n^3$  flops.

**Remark 7.1** We will later discuss that if  $A$  is not well-conditioned (its columns are nearly linearly dependent), the Method of Normal Equations is numerically unstable because  $A^T A$  is ill-conditioned.

The above discussion can be generalized to the case where  $A \in \mathbb{C}^{m \times n}$ . In that case,  $x$  must solve  $A^H A x = A^H y$ .

A geometric explanation of the method of normal equations (for the case where  $A$  is real valued) can be found in Week 10 (Sections 10.3-10.5) of [Linear Algebra: Foundations to Frontiers - Notes to LAFF With.](#)

## 7.3 Solving the LLS Problem Via the QR Factorization

Assume  $A \in \mathbb{C}^{m \times n}$  has linearly independent columns and let  $A = Q_L R_{TL}$  be its QR factorization. We wish to compute the solution to the LLS problem: Find  $x \in \mathbb{C}^n$  such that

$$\|Ax - y\|_2^2 = \min_{z \in \mathbb{C}^n} \|Az - y\|_2^2.$$

### 7.3.1 Simple derivation of the solution

Notice that we know that, if  $A$  has linearly independent columns, the solution is given by  $x = (A^H A)^{-1} A^H y$  (the solution to the normal equations). Now,

$$\begin{aligned}
 x &= [A^H A]^{-1} A^H y && \text{Solution to the Normal Equations} \\
 &= [(Q_L R_{TL})^H (Q_L R_{TL})]^{-1} (Q_L R_{TL})^H y && A = Q_L R_{TL} \\
 &= [R_{TL}^H Q_L^H Q_L R_{TL}]^{-1} R_{TL}^H Q_L^H y && (BC)^H = (C^H B^H) \\
 &= [R_{TL}^H R_{TL}]^{-1} R_{TL}^H Q_L^H y && Q_L^H Q_L = I \\
 &= R_{TL}^{-1} R_{TL}^{-H} R_{TL}^H Q_L^H y && (BC)^{-1} = C^{-1} B^{-1} \\
 &= R_{TL}^{-1} Q_L^H y && R_{TL}^{-H} R_{TL}^H = I
 \end{aligned}$$

Thus, the  $x$  that solves  $R_{TL} x = Q_L^H y$  solves the LLS problem.



### 7.3.2 Alternative derivation of the solution

We know that then there exists a matrix  $Q_R$  such that  $Q = \begin{pmatrix} Q_L & Q_R \end{pmatrix}$  is unitary. Now,

$$\begin{aligned}
& \min_{z \in \mathbb{C}^n} \|Az - y\|_2^2 \\
&= \min_{z \in \mathbb{C}^n} \|Q_L R_{TL} z - y\|_2^2 && \text{(substitute } A = Q_L R_{TL} \text{)} \\
&= \min_{z \in \mathbb{C}^n} \|Q^H (Q_L R_{TL} z - y)\|_2^2 && \text{(two-norm is preserved since } Q^H \text{ is unitary)} \\
&= \min_{z \in \mathbb{C}^n} \left\| \begin{pmatrix} Q_L^H \\ Q_R^H \end{pmatrix} Q_L R_{TL} z - \begin{pmatrix} Q_L^H \\ Q_R^H \end{pmatrix} y \right\|_2^2 && \text{(partitioning, distributing)} \\
&= \min_{z \in \mathbb{C}^n} \left\| \begin{pmatrix} R_{TL} z \\ 0 \end{pmatrix} - \begin{pmatrix} Q_L^H y \\ Q_R^H y \end{pmatrix} \right\|_2^2 && \text{(partitioned matrix-matrix multiplication)} \\
&= \min_{z \in \mathbb{C}^n} \left\| \begin{pmatrix} R_{TL} z - Q_L^H y \\ -Q_R^H y \end{pmatrix} \right\|_2^2 && \text{(partitioned matrix addition)} \\
&= \min_{z \in \mathbb{C}^n} \left( \|R_{TL} z - Q_L^H y\|_2^2 + \|Q_R^H y\|_2^2 \right) && \text{(property of the 2-norm:} \\
& && \left\| \begin{pmatrix} x \\ y \end{pmatrix} \right\|_2^2 = \|x\|_2^2 + \|y\|_2^2) \\
&= \left( \min_{z \in \mathbb{C}^n} \|R_{TL} z - Q_L^H y\|_2^2 \right) + \|Q_R^H y\|_2^2 && (Q_R^H y \text{ is independent of } z) \\
&= \|Q_R^H y\|_2^2 && \text{(minimized by } x \text{ that satisfies } R_{TL} x = Q_L^H y)
\end{aligned}$$

Thus, the desired  $x$  that minimizes the linear least-squares problem solves  $R_{TL} x = Q_L^H y$ . The solution is unique because  $R_{TL}$  is nonsingular (because  $A$  has linearly independent columns).

In practice, one performs the following steps:

- Compute the QR factorization  $A = Q_L R_{TL}$ .  
If Gram-Schmidt or Modified Gram-Schmidt are used, this costs  $2mn^2$  flops.
- Form  $\hat{y} = Q_L^H y$ .  
Cost:  $2mn$  flops.
- Solve  $R_{TL} x = \hat{y}$  (triangular solve).  
Cost:  $n^2$  flops.

Thus, the total cost of solving the LLS problem via (Modified) Gram-Schmidt QR factorization is approximately  $2mn^2$  flops.

Notice that the solution computed by the Method of Normal Equations (generalized to the complex case) is given by

$$\begin{aligned}
(A^H A)^{-1} A^H y &= ((Q_L R_{TL})^H (Q_L R_{TL}))^{-1} (Q_L R_{TL})^H y = (R_{TL}^H Q_L^H Q_L R_{TL})^{-1} R_{TL}^H Q_L^H y \\
&= (R_{TL}^H R_{TL})^{-1} R_{TL}^H Q_L^H y = R_{TL}^{-1} R_{TL}^{-H} R_{TL}^H Q_L^H y = R_{TL}^{-1} Q_L^H y = R_{TL}^{-1} \hat{y} = x
\end{aligned}$$

where  $R_{TL} x = \hat{y}$ . This shows that the two approaches compute the same solution, generalizes the Method of Normal Equations to complex valued problems, and shows that the Method of Normal Equations computes the desired result without requiring multivariate calculus.

## 7.4 Via Householder QR Factorization

Given  $A \in \mathbb{C}^{m \times n}$  with linearly independent columns, the Householder QR factorization yields  $n$  Householder transformations,  $H_0, \dots, H_{n-1}$ , so that

$$\underbrace{H_{n-1} \cdots H_0}_Q A = \begin{pmatrix} R_{TL} \\ 0 \end{pmatrix}.$$

$$Q = \left( \begin{array}{c|c} Q_L & Q_R \end{array} \right)^H$$

We wish to solve  $R_{TL}x = \underbrace{Q_L^H y}_{\hat{y}}$ . But

$$\begin{aligned} \hat{y} = Q_L^H y &= \left[ \left( \begin{array}{c|c} I & 0 \end{array} \right) \left( \begin{array}{c} Q_L^H \\ Q_R^H \end{array} \right) \right] y = \left( \begin{array}{c|c} I & 0 \end{array} \right) \left( \begin{array}{c|c} Q_L & Q_R \end{array} \right)^H y = \left( \begin{array}{c|c} I & 0 \end{array} \right) Q^H y \\ &= \left( \begin{array}{c|c} I & 0 \end{array} \right) (H_{n-1} \cdots H_0) y = \left( \begin{array}{c|c} I & 0 \end{array} \right) \left( \underbrace{H_{n-1} \cdots H_0 y}_{w} \right) = w_T. \end{aligned}$$

$$w = \begin{pmatrix} w_T \\ w_B \end{pmatrix}$$

This suggests the following approach:

- Compute  $H_0, \dots, H_{n-1}$  so that  $H_{n-1} \cdots H_0 A = \begin{pmatrix} R_{TL} \\ 0 \end{pmatrix}$ , storing the Householder vectors that define  $H_0, \dots, H_{n-1}$  over the elements in  $A$  that they zero out (see “Notes on Householder QR Factorization”).  
Cost:  $2mn^2 - \frac{2}{3}n^3$  flops.
- Form  $w = (H_{n-1}(\cdots(H_0 y)\cdots))$  (see “Notes on Householder QR Factorization”). Partition  $w = \begin{pmatrix} w_T \\ w_B \end{pmatrix}$  where  $w_T \in \mathbb{C}^n$ . Then  $\hat{y} = w_T$ .  
Cost:  $4m^2 - 2n^2$  flops. (See “Notes on Householder QR Factorization” regarding this.)
- Solve  $R_{TL}x = \hat{y}$ .  
Cost:  $n^2$  flops.

Thus, the total cost of solving the LLS problem via Householder QR factorization is approximately  $2mn^2 - \frac{2}{3}n^3$  flops. This is cheaper than using (Modified) Gram-Schmidt QR factorization, and hence preferred (because it is also numerically more stable, as we will discuss later in the course).

## 7.5 Via the Singular Value Decomposition

Given  $A \in \mathbb{C}^{m \times n}$  with linearly independent columns, let  $A = U\Sigma V^H$  be its SVD decomposition. Partition

$$U = \left( U_L \mid U_R \right) \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{TL} \\ 0 \end{pmatrix},$$

where  $U_L \in \mathbb{C}^{m \times n}$  and  $\Sigma_{TL} \in \mathbb{R}^{n \times n}$  so that

$$A = \left( U_L \mid U_R \right) \begin{pmatrix} \Sigma_{TL} \\ 0 \end{pmatrix} V^H = U_L \Sigma_{TL} V^H.$$

We wish to compute the solution to the LLS problem: Find  $x \in \mathbb{C}^n$  such that

$$\|Ax - y\|_2^2 = \min_{z \in \mathbb{C}^n} \|Az - y\|_2^2.$$

### 7.5.1 Simple derivation of the solution

Notice that we know that, if  $A$  has linearly independent columns, the solution is given by  $x = (A^H A)^{-1} A^H y$  (the solution to the normal equations). Now,

$x = [A^H A]^{-1} A^H y$	Solution to the Normal Equations
$= [(U_L \Sigma_{TL} V^H)^H (U_L \Sigma_{TL} V^H)]^{-1} (U_L \Sigma_{TL} V^H)^H y$	$A = U_L \Sigma_{TL} V^H$
$= [(V \Sigma_{TL} U_L^H) (U_L \Sigma_{TL} V^H)]^{-1} (V \Sigma_{TL} U_L^H) y$	$(BCD)^H = (D^H C^H B^H)$ and $\Sigma_{TL}^H = \Sigma_{TL}$
$= [V \Sigma_{TL} \Sigma_{TL}^H V^H]^{-1} V \Sigma_{TL} U_L^H y$	$U_L^H U_L = I$
$= V \Sigma_{TL}^{-1} \Sigma_{TL}^{-1} V^H V \Sigma_{TL} U_L^H y$	$V^{-1} = V^H$ and $(BCD)^{-1} = D^{-1} C^{-1} B^{-1}$
$= V \Sigma_{TL}^{-1} U_L^H y$	$V^H V = I$ and $\Sigma_{TL}^{-1} \Sigma_{TL} = I$

### 7.5.2 Alternative derivation of the solution

We now discuss a derivation of the result that does not depend on the Normal Equations, in preparation for the more general case discussed in the next section.

$$\begin{aligned}
& \min_{z \in \mathbb{C}^n} \|Az - y\|_2^2 \\
&= \min_{z \in \mathbb{C}^n} \|U\Sigma V^H z - y\|_2^2 && \text{(substitute } A = U\Sigma V^H\text{)} \\
&= \min_{z \in \mathbb{C}^n} \|U(\Sigma V^H z - U^H y)\|_2^2 && \text{(substitute } UU^H = I \text{ and factor out } U\text{)} \\
&= \min_{z \in \mathbb{C}^n} \|\Sigma V^H z - U^H y\|_2^2 && \text{(multiplication by a unitary matrix} \\
&&& \text{preserves two-norm)} \\
&= \min_{z \in \mathbb{C}^n} \left\| \begin{pmatrix} \Sigma_{TL} \\ 0 \end{pmatrix} V^H z - \begin{pmatrix} U_L^H y \\ U_R^H y \end{pmatrix} \right\|_2^2 && \text{(partition, partitioned matrix-matrix multiplication)} \\
&= \min_{z \in \mathbb{C}^n} \left\| \begin{pmatrix} \Sigma_{TL} V^H z - U_L^H y \\ -U_R^H y \end{pmatrix} \right\|_2^2 && \text{(partitioned matrix-matrix multiplication and addition)} \\
&= \min_{z \in \mathbb{C}^n} \|\Sigma_{TL} V^H z - U_L^H y\|_2^2 + \|U_R^H y\|_2^2 && \left( \left\| \begin{pmatrix} v_T \\ v_B \end{pmatrix} \right\|_2^2 = \|v_T\|_2^2 + \|v_B\|_2^2 \right)
\end{aligned}$$

The  $x$  that solves  $\Sigma_{TL} V^H x = U_L^H y$  minimizes the expression. That  $x$  is given by  $x = V \Sigma_{TL}^{-1} U_L^H y$ .

This suggests the following approach:

- Compute the reduced SVD:  $A = U_L \Sigma_{TL} V^H$ .  
Cost: Greater than computing the QR factorization! We will discuss this in a future note.
- Form  $\hat{y} = \Sigma_{TL}^{-1} U_L^H y$ .  
Cost: approx.  $2mn$  flops.
- Compute  $z = V \hat{y}$ .  
Cost: approx.  $2mn$  flops.

## 7.6 What If $A$ Does Not Have Linearly Independent Columns?

In the above discussions we assume that  $A$  has linearly independent columns. Things get a bit trickier if  $A$  does not have linearly independent columns. There is a variant of the QR factorization known as the QR factorization with column pivoting that can be used to find the solution. We instead focus on using the SVD.

Given  $A \in \mathbb{C}^{m \times n}$  with  $\text{rank}(A) = r < n$ , let  $A = U\Sigma V^H$  be its SVD decomposition. Partition

$$U = \left( U_L \mid U_R \right), \quad V = \left( V_L \mid V_R \right) \quad \text{and} \quad \Sigma = \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right),$$

where  $U_L \in \mathbb{C}^{m \times r}$ ,  $V_L \in \mathbb{C}^{n \times r}$  and  $\Sigma_{TL} \in \mathbb{R}^{r \times r}$  so that

$$A = \left( \begin{array}{c|c} U_L & U_R \end{array} \right) \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right) \left( \begin{array}{c|c} V_L & V_R \end{array} \right)^H = U_L \Sigma_{TL} V_L^H.$$

Now,

$$\begin{aligned} & \min_{z \in \mathbb{C}^n} \|Az - y\|_2^2 \\ &= \min_{z \in \mathbb{C}^n} \|U \Sigma V^H z - y\|_2^2 && \text{(substitute } A = U \Sigma V^H \text{)} \\ &= \min_{z \in \mathbb{C}^n} \|U \Sigma V^H z - U U^H y\|_2^2 && (U U^H = I) \\ &= \min_{\substack{w \in \mathbb{C}^n \\ z = Vw}} \|U \Sigma V^H V w - U U^H y\|_2^2 && \text{(choosing the max over } w \in \mathbb{C}^n \text{ with } z = Vw \\ &&& \text{is the same as choosing the max over } z \in \mathbb{C}^n \text{.)} \\ &= \min_{\substack{w \in \mathbb{C}^n \\ z = Vw}} \|U (\Sigma w - U^H y)\|_2^2 && \text{(factor out } U \text{ and } V^H V = I) \\ &= \min_{\substack{w \in \mathbb{C}^n \\ z = Vw}} \|\Sigma w - U^H y\|_2^2 && (\|Uv\|_2 = \|v\|_2) \\ &= \min_{\substack{w \in \mathbb{C}^n \\ z = Vw}} \left\| \left( \begin{array}{c|c} \Sigma_{TL} & 0 \\ \hline 0 & 0 \end{array} \right) \left( \begin{array}{c} w_T \\ w_B \end{array} \right) - \left( \begin{array}{c} U_L^H \\ U_R^H \end{array} \right) y \right\|_2^2 && \text{(partition } \Sigma, w, \text{ and } U) \\ &= \min_{\substack{w \in \mathbb{C}^n \\ z = Vw}} \left\| \left( \begin{array}{c} \Sigma_{TL} w_T - U_L^H y \\ -U_R^H y \end{array} \right) \right\|_2^2 && \text{(partitioned matrix-matrix multiplication)} \\ &= \min_{\substack{w \in \mathbb{C}^n \\ z = Vw}} \|\Sigma_{TL} w_T - U_L^H y\|_2^2 + \|U_R^H y\|_2^2 && \left( \left\| \left( \begin{array}{c} v_T \\ v_B \end{array} \right) \right\|_2^2 = \|v_T\|_2^2 + \|v_B\|_2^2 \right) \end{aligned}$$

Since  $\Sigma_{TL}$  is a diagonal with no zeroes on its diagonal, we know that  $\Sigma_{TL}^{-1}$  exists. Choosing  $w_T = \Sigma_{TL}^{-1} U_L^H y$  means that

$$\min_{\substack{w \in \mathbb{C}^n \\ z = Vw}} \|\Sigma_{TL} w_T - U_L^H y\|_2^2 = 0,$$

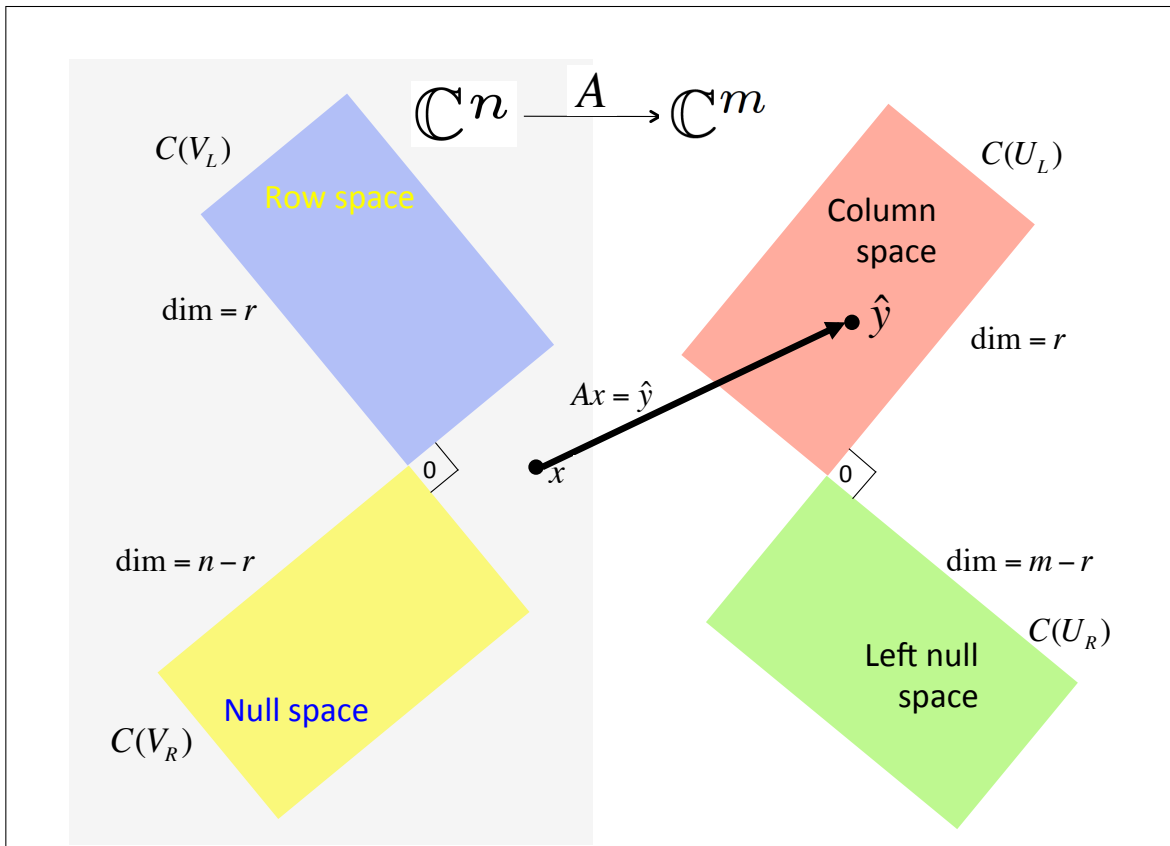
which obviously minimizes the entire expression. We conclude that

$$x = Vw = \left( \begin{array}{c|c} V_L & V_R \end{array} \right) \left( \begin{array}{c} \Sigma_{TL}^{-1} U_L^H y \\ w_B \end{array} \right) = V_L \Sigma_{TL}^{-1} U_L^H y + V_R w_B$$

characterizes all solutions to the linear least-squares problem, where  $w_B$  can be chosen to be any vector of size  $n - r$ . By choosing  $w_B = 0$  and hence  $x = V_L \Sigma_{TL}^{-1} U_L^H y$  we choose the vector  $x$  that itself has minimal 2-norm.

The sequence of pictures on the following pages reasons through the insights that we gained so far (in “Notes on the Singular Value Decomposition” and this note). These pictures can be downloaded as a PowerPoint presentation from

<http://www.cs.utexas.edu/users/flame/Notes/Spaces.pptx>



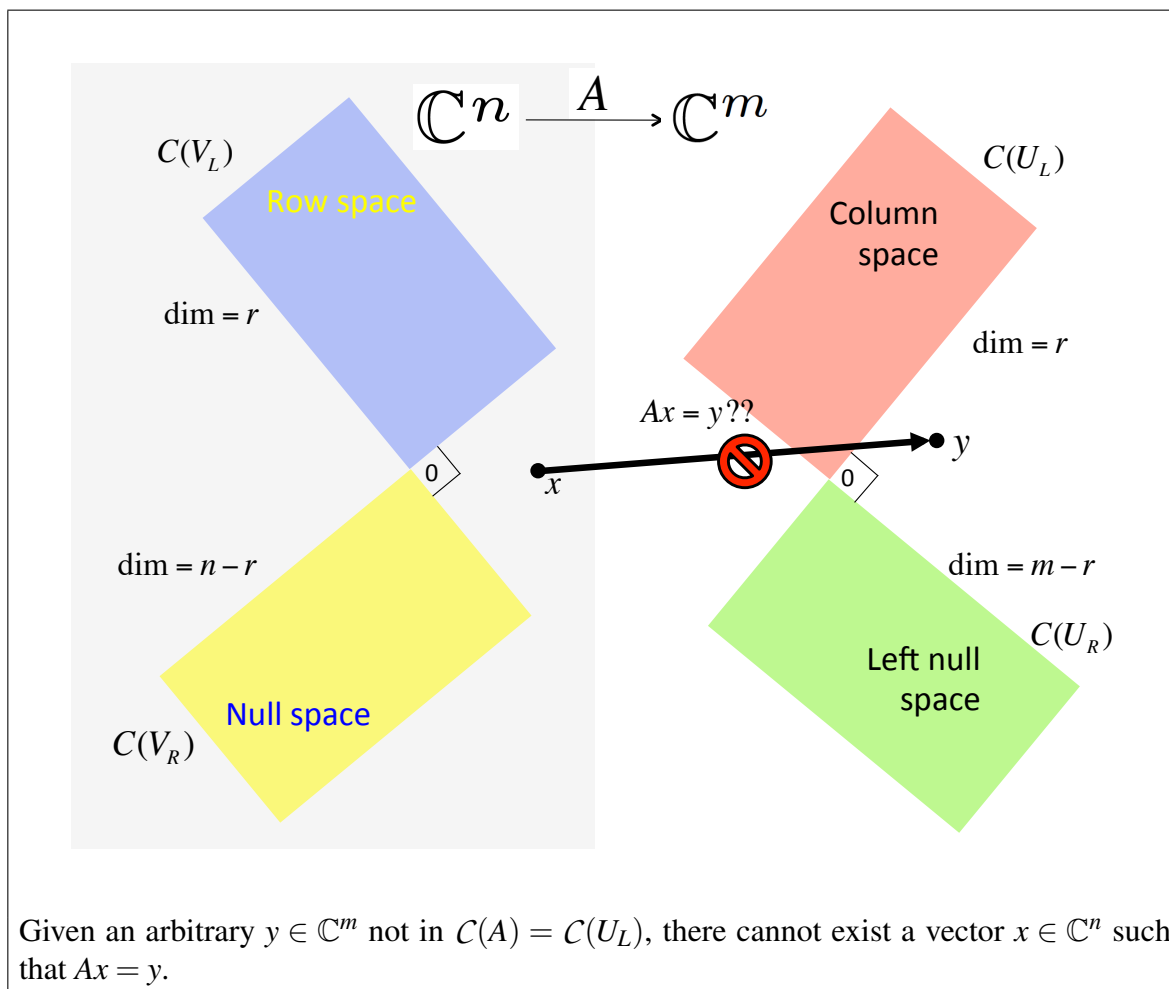
If  $A \in \mathbb{C}^{m \times n}$  and

$$A = \begin{pmatrix} U_L & U_R \end{pmatrix} \begin{pmatrix} \Sigma_{TL} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_L & V_R \end{pmatrix}^H = U_L \Sigma_{TL} V_L^H$$

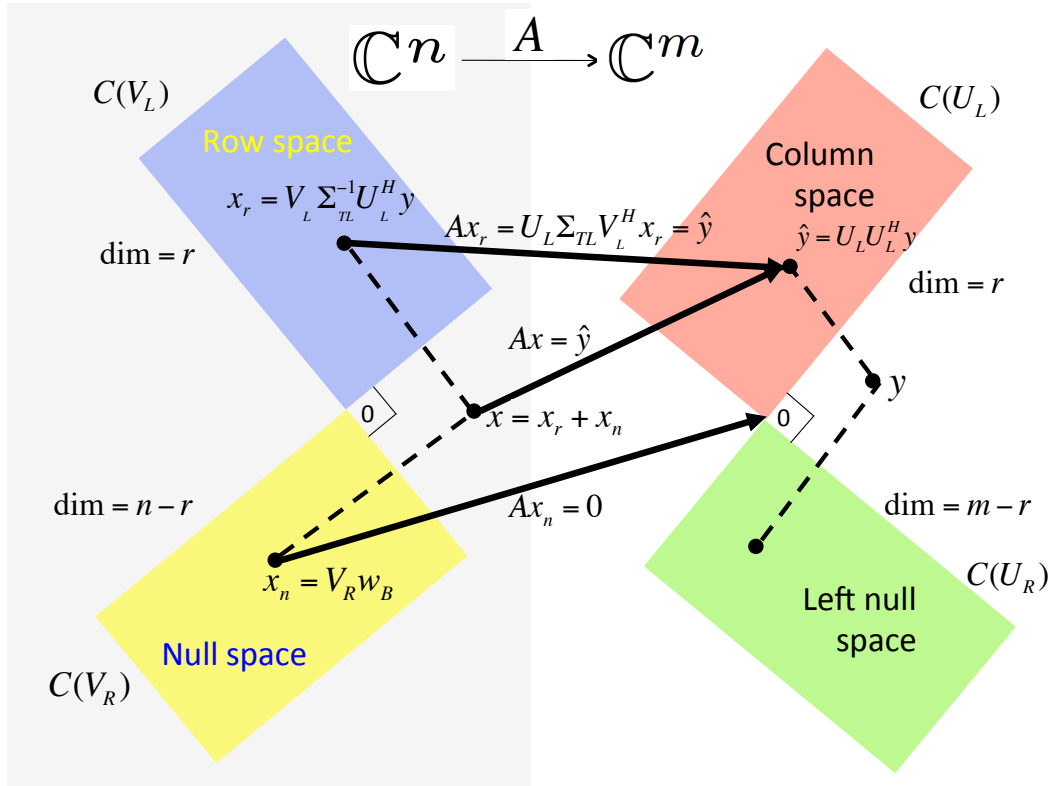
equals the SVD, where  $U_L \in \mathbb{C}^{m \times r}$ ,  $V_L \in \mathbb{C}^{n \times r}$ , and  $\Sigma_{TL} \in \mathbb{C}^{r \times r}$ , then

- The row space of  $A$  equals  $C(V_L)$ , the column space of  $V_L$ ;
- The null space of  $A$  equals  $C(V_R)$ , the column space of  $V_R$ ;
- The column space of  $A$  equals  $C(U_L)$ ; and
- The left null space of  $A$  equals  $C(U_R)$ .

Also, given a vector  $x \in \mathbb{C}^n$ , the matrix  $A$  maps  $x$  to  $\hat{y} = Ax$ , which must be in  $C(A) = C(U_L)$ .







The solution to the Linear Least-Squares problem,  $x$ , equals any vector that is mapped by  $A$  to the projection of  $y$  onto the column space of  $A$ :  $\hat{y} = A(A^H A)^{-1} A^H y = Q Q^H y$ . This solution can be written as the sum of a (unique) vector in the row space of  $A$  and any vector in the null space of  $A$ . The vector in the row space of  $A$  is given by

$$x_r = V_L \Sigma_{TL}^{-1} U_L^H y = V_L \Sigma_{TL}^{-1} U_L^H \hat{y}.$$

The sequence of pictures, and their explanations, suggest a much simpler path towards the formula for solving the LLS problem.

- We know that we are looking for the solution  $x$  to the equation

$$Ax = U_L U_L^H y.$$

- We know that there must be a solution  $x_r$  in the row space of  $A$ . It suffices to find  $w_T$  such that  $x_r = V_L w_T$ .
- Hence we search for  $w_T$  that satisfies

$$A V_L w_T = U_L U_L^H y.$$

- Since there is a one-to-one mapping by  $A$  from the row space of  $A$  to the column space of  $A$ , we know that  $w_T$  is unique. Thus, if we find *a* solution to the above, we have found *the* solution.
- Multiplying both sides of the equation by  $U_L^H$  yields

$$U_L^H A V_L w_T = U_L^H y.$$

- Since  $A = U_L \Sigma_{TL}^{-1} V_L^H$ , we can rewrite the above equation as

$$\Sigma_{TL} w_T = U_L^H y$$

so that  $w_T = \Sigma_{TL}^{-1} U_L^H y$ .

- Hence

$$x_r = V_L \Sigma_{TL}^{-1} U_L^H y.$$

- Adding any vector in the null space of  $A$  to  $x_r$  also yields a solution. Hence all solutions to the LLS problem can be characterized by

$$x = V_L \Sigma_{TL}^{-1} U_L^H y + V_R w_R.$$

**Here is yet another important way of looking at the problem:**

- **We start by considering the LLS problem: Find  $x \in \mathbb{C}^n$  such that**

$$\|Ax - y\|_2^2 = \max_{z \in \mathbb{C}^n} \|Az - y\|_2^2.$$

- **We changed this into the problem of finding  $w_L$  that satisfied**

$$\Sigma_{TL} w_L = v_T$$

**where  $x = V_L w_L$  and  $\hat{y} = U_L U_L^H y = U_L v_T$ .**

- **Thus, by expressing  $x$  in the right basis (the columns of  $V_L$ ) and the projection of  $y$  in the right basis (the columns of  $U_L$ ), the problem became trivial, since the matrix that related the solution to the right-hand side became diagonal.**

## 7.7 Exercise: Using the the $LQ$ factorization to solve underdetermined systems

We next discuss another special case of the LLS problem: Let  $A \in \mathbb{C}^{m \times n}$  where  $m < n$  and  $A$  has linearly independent rows. A series of exercises will lead you to a practical algorithm for solving the problem of describing all solutions to the LLS problem

$$\|Ax - y\|_2 = \min_z \|Az - y\|_2.$$

You may want to review “Notes on the QR Factorization” as you do this exercise.

**Homework 7.2** Let  $A \in \mathbb{C}^{m \times n}$  with  $m < n$  have linearly independent rows. Show that there exist a lower triangular matrix  $L_L \in \mathbb{C}^{m \times m}$  and a matrix  $Q_T \in \mathbb{C}^{m \times n}$  with orthonormal rows such that  $A = L_L Q_T$ , noting that  $L_L$  does not have any zeroes on the diagonal. Letting  $L = \left( L_L \mid 0 \right)$  be  $\mathbb{C}^{m \times n}$  and unitary

$$Q = \begin{pmatrix} Q_T \\ Q_B \end{pmatrix}, \text{ reason that } A = LQ.$$

Don't overthink the problem: use results you have seen before.

➡ [SEE ANSWER](#)

**Homework 7.3** Let  $A \in \mathbb{C}^{m \times n}$  with  $m < n$  have linearly independent rows. Consider

$$\|Ax - y\|_2 = \min_z \|Az - y\|_2.$$

Use the fact that  $A = L_L Q_T$ , where  $L_L \in \mathbb{C}^{m \times m}$  is lower triangular and  $Q_T$  has orthonormal rows, to argue that any vector of the form  $Q_T^H L_L^{-1} y + Q_B^H w_B$  (where  $w_B$  is any vector in  $\mathbb{C}^{n-m}$ ) is a solution to the LLS

$$\text{problem. Here } Q = \begin{pmatrix} Q_T \\ Q_B \end{pmatrix}.$$

➡ [SEE ANSWER](#)

**Homework 7.4** Continuing Exercise 7.2, use Figure 7.1 to give a Classical Gram-Schmidt inspired algorithm for computing  $L_L$  and  $Q_T$ . (The best way to check you got the algorithm right is to implement it!)

➡ [SEE ANSWER](#)

**Homework 7.5** Continuing Exercise 7.2, use Figure 7.2 to give a Householder QR factorization inspired algorithm for computing  $L$  and  $Q$ , leaving  $L$  in the lower triangular part of  $A$  and  $Q$  stored as Householder vectors above the diagonal of  $A$ . (The best way to check you got the algorithm right is to implement it!)

➡ [SEE ANSWER](#)

**Algorithm:**  $[L, Q] := \text{LQ\_CGS\_UNB}(A, L, Q)$

**Partition**  $A \rightarrow \begin{pmatrix} A_T \\ A_B \end{pmatrix}, L \rightarrow \begin{pmatrix} L_{TL} & L_{TR} \\ L_{BL} & L_{BR} \end{pmatrix}, Q \rightarrow \begin{pmatrix} Q_T \\ Q_B \end{pmatrix}$

**where**  $A_T$  has 0 rows,  $L_{TL}$  is  $0 \times 0$ ,  $Q_T$  has 0 rows

**while**  $m(A_T) < m(A)$  **do**

**Repartition**

$$\begin{pmatrix} A_T \\ A_B \end{pmatrix} \rightarrow \begin{pmatrix} A_0 \\ a_1^T \\ A_2 \end{pmatrix}, \begin{pmatrix} L_{TL} & L_{TR} \\ L_{BL} & L_{BR} \end{pmatrix} \rightarrow \begin{pmatrix} L_{00} & l_{01} & L_{02} \\ l_{10}^T & \lambda_{11} & l_{12}^T \\ L_{20} & l_{21} & L_{22} \end{pmatrix}, \begin{pmatrix} Q_T \\ Q_B \end{pmatrix} \rightarrow \begin{pmatrix} Q_0 \\ q_1^T \\ Q_2 \end{pmatrix}$$

**where**  $a_1$  has 1 row,  $\lambda_{11}$  is  $1 \times 1$ ,  $q_1$  has 1 row

**Continue with**

$$\begin{pmatrix} A_T \\ A_B \end{pmatrix} \leftarrow \begin{pmatrix} A_0 \\ a_1^T \\ A_2 \end{pmatrix}, \begin{pmatrix} L_{TL} & L_{TR} \\ L_{BL} & L_{BR} \end{pmatrix} \leftarrow \begin{pmatrix} L_{00} & l_{01} & L_{02} \\ l_{10}^T & \lambda_{11} & l_{12}^T \\ L_{20} & l_{21} & L_{22} \end{pmatrix}, \begin{pmatrix} Q_T \\ Q_B \end{pmatrix} \leftarrow \begin{pmatrix} Q_0 \\ q_1^T \\ Q_2 \end{pmatrix}$$

**endwhile**

Figure 7.1: Algorithm skeleton for CGS-like LQ factorization.

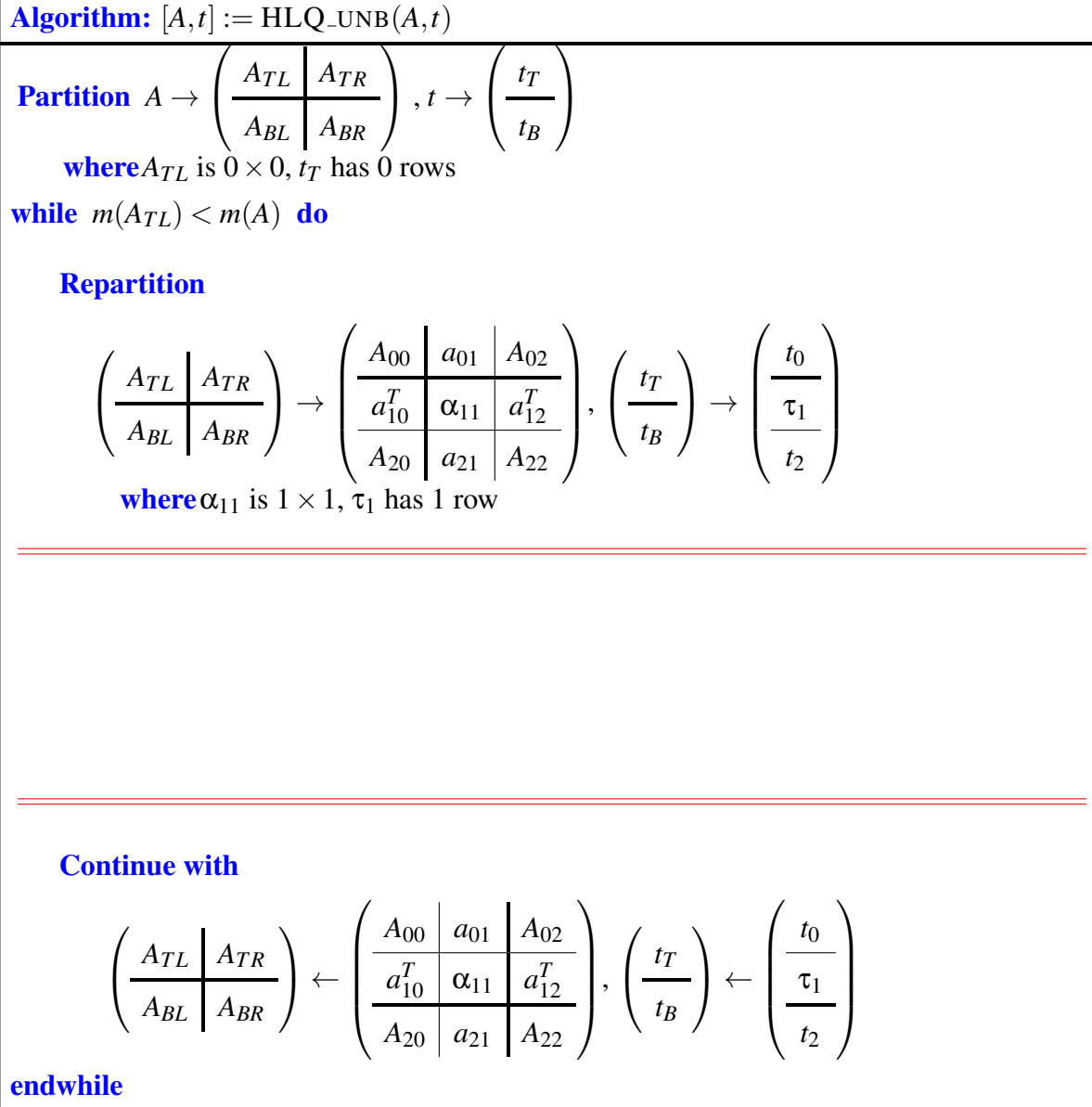


Figure 7.2: Algorithm skeleton for Householder QR factorization inspired LQ factorization.



## Notes on the Condition of a Problem

Correctness in the presence of error (e.g., when floating point computations are performed) takes on a different meaning. For many problems for which computers are used, there is one correct answer, and we expect that answer to be computed by our program. The problem is that, as we will see later, most real numbers cannot be stored exactly in a computer memory. They are stored as approximations, floating point numbers, instead. Hence storing them and/or computing with them inherently incurs error.

Naively, we would like to be able to define a program that computes with floating point numbers as being “correct” if it computes an answer that is close to the exact answer. Unfortunately, some problems that are computed this way have the property that a small change in the input yields a large change in the output. Surely we can’t blame the program for not computing an answer close to the exact answer in this case. The mere act of storing the input data as a floating point number may cause a completely different output, even if all computation is exact. We will later define *stability* to be a property of a program. It is what takes the place of correctness. In this note, we instead will focus on when a problem is a “good” problem, meaning that in exact arithmetic a “small” change in the input will always cause at most a “small” change in the output, or a “bad” problem if a “small” change may yield a “large” change. A good problem will be called *well-conditioned*. A bad problem will be called *ill-conditioned*.

Notice that “small” and “large” are vague. To some degree, norms help us measure size. To some degree, “small” and “large” will be in the eyes of the beholder (in other words, situation dependent).

### Video

Read disclaimer regarding the videos in the preface!

Video did not turn out...

## Outline

<b>Video</b> . . . . .	127
<b>Outline</b> . . . . .	128
<b>8.1. Notation</b> . . . . .	129
<b>8.2. The Prototypical Example: Solving a Linear System</b> . . . . .	129
<b>8.3. Condition Number of a Rectangular Matrix</b> . . . . .	133
<b>8.4. Why Using the Method of Normal Equations Could be Bad</b> . . . . .	134
<b>8.5. Why Multiplication with Unitary Matrices is a Good Thing</b> . . . . .	135



## 8.1 Notation

Throughout this note, we will talk about small changes (perturbations) to scalars, vectors, and matrices. To denote these, we attach a “delta” to the symbol for a scalar, vector, or matrix.

- A small change to scalar  $\alpha \in \mathbb{C}$  will be denoted by  $\delta\alpha \in \mathbb{C}$ ;
- A small change to vector  $x \in \mathbb{C}^n$  will be denoted by  $\delta x \in \mathbb{C}^n$ ; and
- A small change to matrix  $A \in \mathbb{C}^{m \times n}$  will be denoted by  $\delta A \in \mathbb{C}^{m \times n}$ .

Notice that the “delta” touches the  $\alpha$ ,  $x$ , and  $A$ , so that, for example,  $\delta x$  is not mistaken for  $\delta \cdot x$ .

## 8.2 The Prototypical Example: Solving a Linear System

Assume that  $A \in \mathbb{R}^{n \times n}$  is nonsingular and  $x, y \in \mathbb{R}^n$  with  $Ax = y$ . The problem here is the function that computes  $x$  from  $y$  and  $A$ . Let us assume that no error is introduced in the matrix  $A$  when it is stored, but that in the process of storing  $y$  a small error is introduced:  $\delta y \in \mathbb{R}^n$  so that now  $y + \delta y$  is stored. The question becomes by how much the solution  $x$  changes as a function of  $\delta y$ . In particular, we would like to quantify how a relative change in the right-hand side  $y$  ( $\|\delta y\|/\|y\|$  in some norm) translates to a relative change in the solution  $x$  ( $\|\delta x\|/\|x\|$ ). It turns out that we will need to compute norms of matrices, using the norm induced by the vector norm that we use.

Since  $Ax = y$ , if we use a consistent (induced) matrix norm,

$$\|y\| = \|Ax\| \leq \|A\|\|x\| \text{ or, equivalently, } \frac{1}{\|x\|} \leq \|A\| \frac{1}{\|y\|}. \quad (8.1)$$

Also,

$$\left. \begin{array}{rcl} A(x + \delta x) & = & y + \delta y \\ Ax & = & y \end{array} \right\} \text{ implies that } A\delta x = \delta y \text{ so that } \delta x = A^{-1}\delta y.$$

Hence

$$\|\delta x\| = \|A^{-1}\delta y\| \leq \|A^{-1}\|\|\delta y\|. \quad (8.2)$$

Combining (8.1) and (8.2) we conclude that

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\| \frac{\|\delta y\|}{\|y\|}.$$

**What does this mean?** It means that the relative error in the solution is at worst the relative error in the right-hand side, amplified by  $\|A\|\|A^{-1}\|$ . So, if that quantity is “small” *and* the relative error in the right-hand side is “small” *and* exact arithmetic is used, then one is guaranteed a solution with a relatively “small” error.

The quantity  $\kappa_{\|\cdot\|}(A) = \|A\|\|A^{-1}\|$  is called the *condition number* of nonsingular matrix  $A$  (associated with norm  $\|\cdot\|$ ).

**Are we overestimating by how much the relative error can be amplified?** The answer to this is **no**. For every nonsingular matrix  $A$ , there exists a right-hand side  $y$  and perturbation  $\delta y$  such that, if  $A(x + \delta x) = y + \delta y$ ,

$$\frac{\|\delta x\|}{\|x\|} = \|A\| \|A^{-1}\| \frac{\|\delta y\|}{\|y\|}.$$

In order for this equality to hold, we need to find  $y$  and  $\delta y$  such that

$$\|y\| = \|Ax\| = \|A\| \|x\| \text{ or, equivalently, } \|A\| = \frac{\|Ax\|}{\|x\|}$$

and

$$\|\delta x\| = \|A^{-1}\delta y\| = \|A^{-1}\| \|\delta y\|. \text{ or, equivalently, } \|A^{-1}\| = \frac{\|A^{-1}\delta y\|}{\|\delta y\|}.$$

In other words,  $x$  can be chosen as a vector that maximizes  $\|Ax\|/\|x\|$  and  $\delta y$  should maximize  $\|A^{-1}\delta y\|/\|\delta y\|$ . The vector  $y$  is then chosen as  $y = Ax$ .

**What if we use the 2-norm?** For this norm,  $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \sigma_0/\sigma_{n-1}$ . So, the ratio between the largest and smallest singular value determines whether a matrix is well-conditioned or ill-conditioned.

To show for what vectors the maximal magnification is attained, consider the SVD

$$A = U \Sigma V^T = \left( u_0 \mid u_1 \mid \cdots \mid u_{n-1} \right) \begin{pmatrix} \sigma_0 & & & \\ & \sigma_1 & & \\ & & \ddots & \\ & & & \sigma_{n-1} \end{pmatrix} \left( v_0 \mid v_1 \mid \cdots \mid v_{n-1} \right)^H.$$

Recall that

- $\|A\|_2 = \sigma_0$ ,  $v_0$  is the vector that maximizes  $\max_{\|z\|_2=1} \|Az\|_2$ , and  $Av_0 = \sigma_0 u_0$ ;
- $\|A^{-1}\|_2 = 1/\sigma_{n-1}$ ,  $u_{n-1}$  is the vector that maximizes  $\max_{\|z\|_2=1} \|A^{-1}z\|_2$ , and  $Av_{n-1} = \sigma_{n-1} u_{n-1}$ .

Now, take  $y = \sigma_0 u_0$ . Then  $Ax = y$  is solved by  $x = v_0$ . Take  $\delta y = \beta \sigma_1 u_1$ . Then  $A\delta x = \delta y$  is solved by  $x = \beta v_1$ . Now,

$$\frac{\|\delta y\|_2}{\|y\|_2} = \frac{|\beta| \sigma_1}{\sigma_0} \text{ and } \frac{\|\delta x\|_2}{\|x\|_2} = |\beta|.$$

Hence

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\sigma_0}{\sigma_{n-1}} \frac{\|\delta y\|_2}{\|y\|_2}$$

This is depicted in Figure 8.1 for  $n = 2$ .

The SVD can be used to show that  $A$  maps the unit ball to an ellipsoid. The singular values are the lengths of the various axes of the ellipsoid. The condition number thus captures the eccentricity of the ellipsoid: the ratio between the lengths of the largest and smallest axes. This is also illustrated in Figure 8.1.

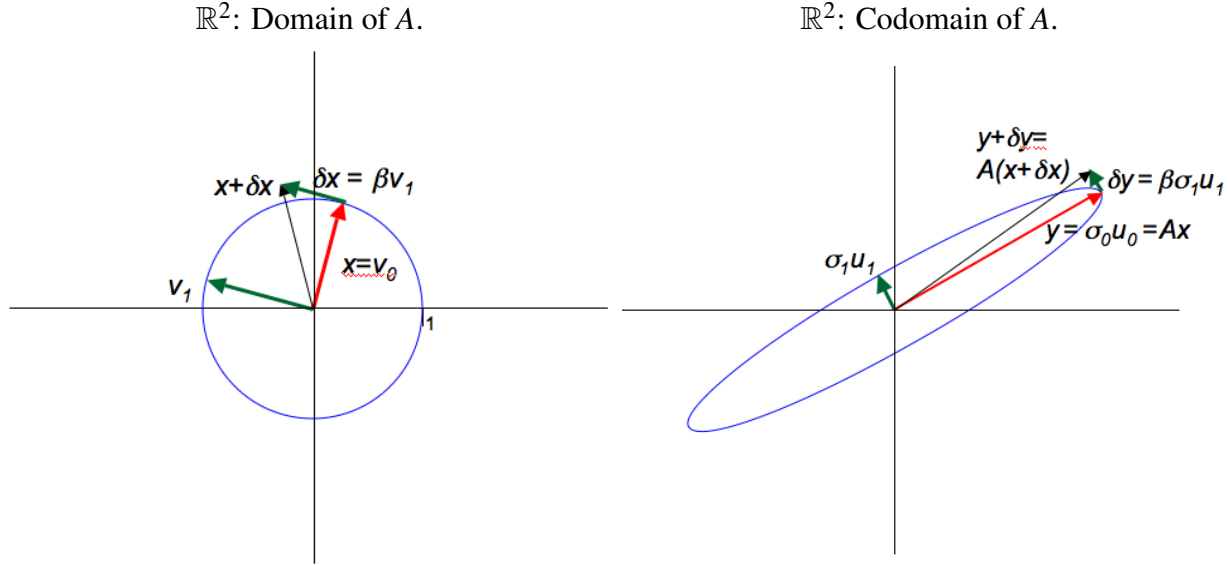


Figure 8.1: Illustration for choices of vectors  $y$  and  $\delta y$  that result in  $\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\sigma_0}{\sigma_{n-1}} \frac{\|\delta y\|_2}{\|y\|_2}$ . Because of the eccentricity of the ellipse, the relatively small change  $\delta y$  relative to  $y$  is amplified into a relatively large change  $\delta x$  relative to  $x$ . On the right, we see that  $\|\delta y\|_2 / \|y\|_2 = \beta \sigma_1 / \sigma_0$  (since  $\|u_0\|_2 = \|u_1\|_2 = 1$ ). On the left, we see that  $\|\delta x\|_2 / \|x\|_2 = \beta$  (since  $\|v_0\|_2 = \|v_1\|_2 = 1$ ).

**Number of accurate digits** Notice that for scalars  $\delta\psi$  and  $\psi$ ,  $\log_{10}(\frac{\delta\psi}{\psi}) = \log_{10}(\delta\psi) - \log_{10}(\psi)$  roughly equals the number leading decimal digits of  $\psi + \delta\psi$  that are accurate, relative to  $\psi$ . For example, if  $\psi = 32.512$  and  $\delta\psi = 0.02$ , then  $\psi + \delta\psi = 32.532$  which has three accurate digits (highlighted in red). Now,  $\log_{10}(32.512) - \log_{10}(0.02) \approx 1.5 - (-1.7) = 3.2$ .

Now, if

$$\frac{\|\delta x\|}{\|x\|} = \kappa(A) \frac{\|\delta y\|}{\|y\|}.$$

then

$$\log_{10}(\|\delta x\|) - \log_{10}(\|x\|) = \log_{10}(\kappa(A)) + \log_{10}(\|\delta y\|) - \log_{10}(\|y\|)$$

so that

$$\log_{10}(\|x\|) - \log_{10}(\|\delta x\|) = [\log_{10}(\|y\|) - \log_{10}(\|\delta y\|)] - \log_{10}(\kappa(A)).$$

In other words, if there were  $k$  digits of accuracy in the right-hand side, then it is possible that (due only to the condition number of  $A$ ) there are only  $k - \log_{10}(\kappa(A))$  digits of accuracy in the solution. If we start with only 8 digits of accuracy and  $\kappa(A) = 10^5$ , we may only get 3 digits of accuracy. If  $\kappa(A) \geq 10^8$ , we may not get *any* digits of accuracy...

**Homework 8.1** Show that, if  $A$  is a nonsingular matrix, for a consistent matrix norm,  $\kappa(A) \geq 1$ .

SEE ANSWER

We conclude from this that we can generally only expect as much relative accuracy in the solution as we had in the right-hand side.

**Alternative exposition** Note: the below links conditioning of matrices to the relative condition number of a more general function. For a more thorough treatment, you may want to read Lecture 12 of “Trefethen and Bau”. That book discusses the subject in much more generality than is needed for our discussion of linear algebra. Thus, if this alternative exposition baffles you, just skip it!

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a continuous function such that  $f(y) = x$ . Let  $\|\cdot\|$  be a vector norm. Consider for  $y \neq 0$

$$\kappa^f(y) = \lim_{\delta \rightarrow 0} \sup_{\|\delta y\| \leq \delta} \left( \frac{\|f(y + \delta y) - f(y)\|}{\|f(y)\|} \right) / \left( \frac{\|\delta y\|}{\|y\|} \right).$$

Letting  $f(y + \delta y) = x + \delta x$ , we find that

$$\kappa^f(y) = \lim_{\delta \rightarrow 0} \sup_{\|\delta y\| \leq \delta} \left( \frac{\|x + \delta x\|}{\|x\|} \right) / \left( \frac{\|\delta y\|}{\|y\|} \right).$$

(Obviously, if  $\delta y = 0$  or  $y = 0$  or  $f(y) = 0$ , things get a bit hairy, so let's not allow that.)

Roughly speaking,  $\kappa^f(y)$  equals the maximum that a(n infinitesimally) small relative error in  $y$  is magnified into a relative error in  $f(y)$ . This can be considered the **relative condition number** of function  $f$ . A large relative condition number means a small relative error in the input ( $y$ ) can be magnified into a large relative error in the output ( $x = f(y)$ ). This is bad, since small errors will invariable occur.

Now, if  $f(y) = x$  is the function that returns  $x$  where  $Ax = y$  for a nonsingular matrix  $A \in \mathbb{C}^{n \times n}$ , then via an argument similar to what we did earlier in this section we find that  $\kappa^f(y) \leq \kappa(A) = \|A\| \|A^{-1}\|$ , the condition number of matrix  $A$ :

$$\begin{aligned} & \lim_{\delta \rightarrow 0} \sup_{\|\delta y\| \leq \delta} \left( \frac{\|f(y + \delta y) - f(y)\|}{\|f(y)\|} \right) / \left( \frac{\|\delta y\|}{\|y\|} \right) \\ &= \lim_{\delta \rightarrow 0} \sup_{\|\delta y\| \leq \delta} \left( \frac{\|A^{-1}(y + \delta y) - A^{-1}(y)\|}{\|A^{-1}y\|} \right) / \left( \frac{\|\delta y\|}{\|y\|} \right) \\ &= \lim_{\delta \rightarrow 0} \max_{\|z\| = 1} \left( \frac{\|A^{-1}(y + \delta y) - A^{-1}(y)\|}{\|A^{-1}y\|} \right) / \left( \frac{\|\delta y\|}{\|y\|} \right) \\ & \quad \delta y = \delta \cdot z \\ &= \lim_{\delta \rightarrow 0} \max_{\|z\| = 1} \left( \frac{\|A^{-1}\delta y\|}{\|\delta y\|} \right) / \left( \frac{\|A^{-1}y\|}{\|y\|} \right) \\ & \quad \delta y = \delta \cdot z \\ &= \lim_{\delta \rightarrow 0} \max_{\|z\| = 1} \left( \frac{\|A^{-1}\delta y\|}{\|\delta y\|} \right) \left( \frac{\|y\|}{\|A^{-1}y\|} \right) \\ & \quad \delta y = \delta \cdot z \end{aligned}$$

$$\begin{aligned}
&= \left[ \lim_{\delta \rightarrow 0} \max_{\substack{\|z\|=1 \\ \delta y = \delta \cdot z}} \left( \frac{\|A^{-1} \delta y\|}{\|\delta y\|} \right) \right] \left[ \left( \frac{\|y\|}{\|A^{-1} y\|} \right) \right] \\
&= \lim_{\delta \rightarrow 0} \max_{\substack{\|z\|=1 \\ \delta y = \delta \cdot z}} \left( \frac{\|A^{-1}(\delta \cdot z)\|}{\|\delta \cdot z\|} \right) \left( \frac{\|y\|}{\|A^{-1} y\|} \right) \\
&= \max_{\|z\|=1} \left( \frac{\|A^{-1} z\|}{\|z\|} \right) \left( \frac{\|y\|}{\|A^{-1} y\|} \right) \\
&= \max_{\|z\|=1} \left( \frac{\|A^{-1} z\|}{\|z\|} \right) \left( \frac{\|Ax\|}{\|x\|} \right) \\
&\leq \left[ \max_{\|z\|=1} \left( \frac{\|A^{-1} z\|}{\|z\|} \right) \right] \left[ \max_{x \neq 0} \left( \frac{\|Ax\|}{\|x\|} \right) \right] \\
&= \|A\| \|A^{-1}\|,
\end{aligned}$$

where  $\|\cdot\|$  is the matrix norm induced by vector norm  $\|\cdot\|$ .

### 8.3 Condition Number of a Rectangular Matrix

Given  $A \in \mathbb{C}^{m \times n}$  with linearly independent columns and  $y \in \mathbb{C}^m$ , consider the linear least-squares (LLS) problem

$$\|Ax - y\|_2 = \min_w \|Aw - y\|_2 \quad (8.3)$$

and the perturbed problem

$$\|A(x + \delta x) - y\|_2 = \min_{w + \delta w} \|A(w + \delta w) - (y + \delta y)\|_2. \quad (8.4)$$

We will again bound by how much the relative error in  $y$  is amplified.

Notice that the solutions to (8.3) and (8.4) respectively satisfy

$$\begin{aligned}
A^H A x &= A^H y \\
A^H A (x + \delta x) &= A^H (y + \delta y)
\end{aligned}$$

so that  $A^H A \delta x = A^H \delta y$  (subtracting the first equation from the second) and hence

$$\|\delta x\|_2 = \|(A^H A)^{-1} A^H \delta y\|_2 \leq \|(A^H A)^{-1} A^H\|_2 \|\delta y\|_2.$$

Now, let  $z = A(A^H A)^{-1} A^H y$  be the projection of  $y$  onto  $C(A)$  and let  $\theta$  be the angle between  $z$  and  $y$ . Let us assume that  $y$  is not orthogonal to  $C(A)$  so that  $z \neq 0$ . Then  $\cos \theta = \|z\|_2 / \|y\|_2$  so that

$$\cos \theta \|y\|_2 = \|z\|_2 = \|Ax\|_2 \leq \|A\|_2 \|x\|_2$$

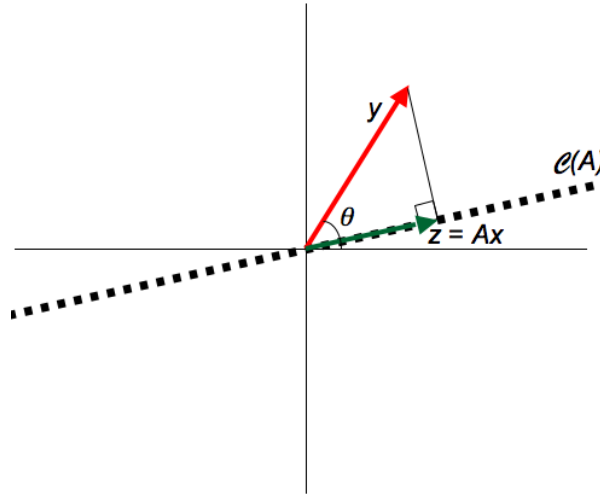


Figure 8.2: Linear least-squares problem  $\|Ax - y\|_2 = \min_v \|Av - y\|_2$ . Vector  $z$  is the projection of  $y$  onto  $C(A)$ .

and hence

$$\frac{1}{\|x\|_2} \leq \frac{\|A\|_2}{\cos \theta \|y\|_2}$$

We conclude that

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\|A\|_2 \|(A^H A)^{-1} A^H\|_2}{\cos \theta} \frac{\|\delta y\|_2}{\|y\|_2} = \frac{1}{\cos \theta} \frac{\sigma_0}{\sigma_{n-1}} \frac{\|\delta y\|_2}{\|y\|_2}$$

where  $\sigma_0$  and  $\sigma_{n-1}$  are (respectively) the largest and smallest singular values of  $A$ , because of the following result:

**Homework 8.2** If  $A$  has linearly independent columns, show that  $\|(A^H A)^{-1} A^H\|_2 = 1/\sigma_{n-1}$ , where  $\sigma_{n-1}$  equals the smallest singular value of  $A$ . Hint: Use the SVD of  $A$ .

🔗 [SEE ANSWER](#)

The condition number of  $A \in \mathbb{C}^{m \times n}$  with linearly independent columns is  $\kappa_2(A) = \sigma_0/\sigma_{n-1}$ .

Notice the effect of the  $\cos \theta$ . When  $y$  is almost perpendicular to  $C(A)$ , then its projection  $z$  is small and  $\cos \theta$  is small. Hence a small relative change in  $y$  can be greatly amplified. This makes sense: if  $y$  is almost perpendicular to  $C(A)$ , then  $x \approx 0$ , and any small  $\delta y \in C(A)$  can yield a *relatively* large change  $\delta x$ .

## 8.4 Why Using the Method of Normal Equations Could be Bad

**Homework 8.3** Let  $A$  have linearly independent columns. Show that  $\kappa_2(A^H A) = \kappa_2(A)^2$ .

🔗 [SEE ANSWER](#)

**Homework 8.4** Let  $A \in \mathbb{C}^{n \times n}$  have linearly independent columns.

- Show that  $Ax = y$  if and only if  $A^H Ax = A^H y$ .
- Reason that using the method of normal equations to solve  $Ax = y$  has a condition number of  $\kappa_2(A)^2$ .

🔗 [SEE ANSWER](#)

Let  $A \in \mathbb{C}^{m \times n}$  have linearly independent columns. If one uses the Method of Normal Equations to solve the linear least-squares problem  $\min_x \|Ax - y\|_2$ , one ends up solving the square linear system  $A^H A x = A^H y$ . Now,  $\kappa_2(A^H A) = \kappa_2(A)^2$ . Hence, using this method squares the condition number of the matrix being used.

## 8.5 Why Multiplication with Unitary Matrices is a Good Thing

Next, consider the computation  $C = AB$  where  $A \in \mathbb{C}^{m \times m}$  is nonsingular and  $B, \Delta B, C, \Delta C \in \mathbb{C}^{m \times n}$ . Then

$$\begin{aligned}(C + \Delta C) &= A(B + \Delta B) \\ C &= AB \\ \Delta C &= A\Delta B\end{aligned}$$

Thus,

$$\|\Delta C\|_2 = \|A\Delta B\|_2 \leq \|A\|_2 \|\Delta B\|_2.$$

Also,  $B = A^{-1}C$  so that

$$\|B\|_2 = \|A^{-1}C\|_2 \leq \|A^{-1}\|_2 \|C\|_2$$

and hence

$$\frac{1}{\|C\|_2} \leq \|A^{-1}\|_2 \frac{1}{\|B\|_2}.$$

Thus,

$$\frac{\|\Delta C\|_2}{\|C\|_2} \leq \|A\|_2 \|A^{-1}\|_2 \frac{\|\Delta B\|_2}{\|B\|_2} = \kappa_2(A) \frac{\|\Delta B\|_2}{\|B\|_2}.$$

This means that the relative error in matrix  $C = AB$  is at most  $\kappa_2(A)$  greater than the relative error in  $B$ .

The following exercise gives us a hint as to why algorithms that cast computation in terms of multiplication by unitary matrices avoid the buildup of error:

**Homework 8.5** Let  $U \in \mathbb{C}^{n \times n}$  be unitary. Show that  $\kappa_2(U) = 1$ .

🔗 [SEE ANSWER](#)

This means is that the relative error in matrix  $C = UB$  is no greater than the relative error in  $B$  when  $U$  is unitary.

**Homework 8.6** Characterize the set of all square matrices  $A$  with  $\kappa_2(A) = 1$ .

🔗 [SEE ANSWER](#)





## Notes on the Stability of an Algorithm

Based on “Goal-Oriented and Modular Stability Analysis” [6, 7] by Paolo Bientinesi and Robert van de Geijn.

### Video

Read disclaimer regarding the videos in the preface!

👉 [Lecture on the Stability of an Algorithm](#)

👉 [YouTube](#)

👉 [Download from UT Box](#)

👉 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

## Outline

<b>Video</b> . . . . .	<b>137</b>
<b>Outline</b> . . . . .	<b>138</b>
<b>9.1. Motivation</b> . . . . .	<b>139</b>
<b>9.2. Floating Point Numbers</b> . . . . .	<b>140</b>
<b>9.3. Notation</b> . . . . .	<b>142</b>
<b>9.4. Floating Point Computation</b> . . . . .	<b>142</b>
9.4.1. Model of floating point computation . . . . .	142
9.4.2. Stability of a numerical algorithm . . . . .	143
9.4.3. Absolute value of vectors and matrices . . . . .	143
<b>9.5. Stability of the Dot Product Operation</b> . . . . .	<b>144</b>
9.5.1. An algorithm for computing DOT . . . . .	144
9.5.2. A simple start . . . . .	144
9.5.3. Preparation . . . . .	146
9.5.4. Target result . . . . .	148
9.5.5. A proof in traditional format . . . . .	149
9.5.6. A weapon of math induction for the war on error (optional) . . . . .	149
9.5.7. Results . . . . .	152
<b>9.6. Stability of a Matrix-Vector Multiplication Algorithm</b> . . . . .	<b>152</b>
9.6.1. An algorithm for computing GEMV . . . . .	152
9.6.2. Analysis . . . . .	152
<b>9.7. Stability of a Matrix-Matrix Multiplication Algorithm</b> . . . . .	<b>154</b>
9.7.1. An algorithm for computing GEMM . . . . .	154
9.7.2. Analysis . . . . .	154
9.7.3. An application . . . . .	155

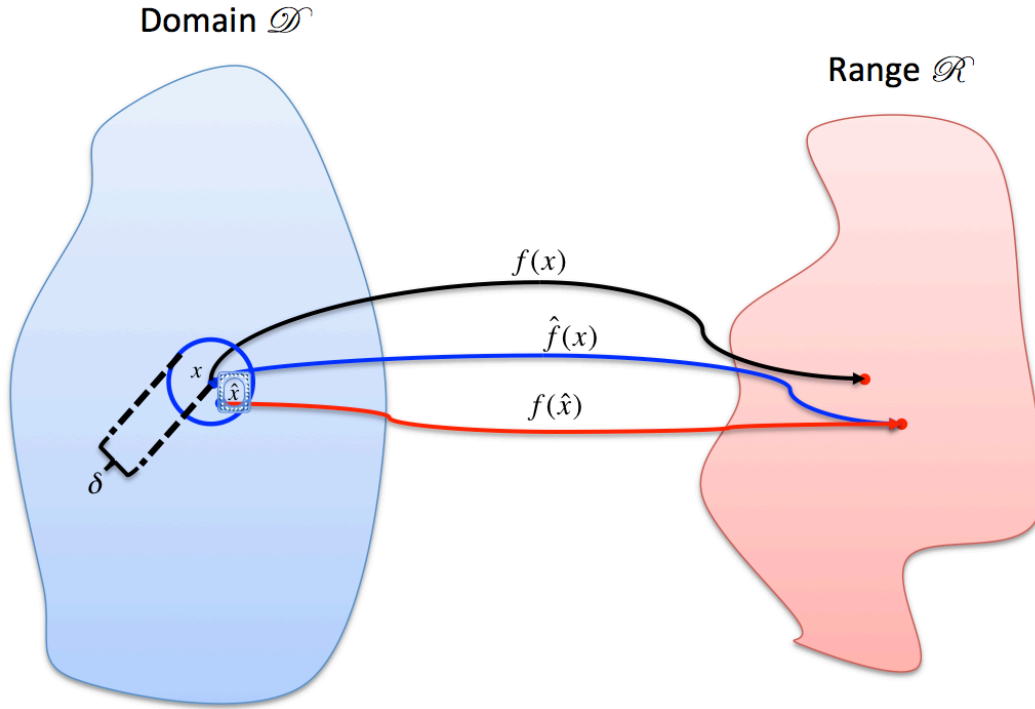


Figure 9.1: In this illustration,  $f : \mathcal{D} \rightarrow \mathcal{R}$  is a function to be evaluated. The function  $\hat{f}$  represents the implementation of the function that uses floating point arithmetic, thus incurring errors. The fact that for a nearby value  $\hat{x}$  the computed value equals the exact function applied to the slightly perturbed  $x$ ,  $f(\hat{x}) = \hat{f}(x)$ , means that the error in the computation can be attributed to a small change in the input. If this is true, then  $\hat{f}$  is said to be a (numerically) stable implementation of  $f$  for input  $x$ .

## 9.1 Motivation

Correctness in the presence of error (e.g., when floating point computations are performed) takes on a different meaning. For many problems for which computers are used, there is one correct answer and we expect that answer to be computed by our program. The problem is that most real numbers cannot be stored exactly in a computer memory. They are stored as approximations, floating point numbers, instead. Hence storing them and/or computing with them inherently incurs error. The question thus becomes “When is a program correct in the presense of such errors?”

Let us assume that we wish to evaluate the mapping  $f : \mathcal{D} \rightarrow \mathcal{R}$  where  $\mathcal{D} \subset \mathbb{R}^n$  is the domain and  $\mathcal{R} \subset \mathbb{R}^m$  is the range (codomain). Now, we will let  $\hat{f} : \mathcal{D} \rightarrow \mathcal{R}$  denote a computer implementation of this function. Generally, for  $x \in \mathcal{D}$  it is the case that  $f(x) \neq \hat{f}(x)$ . Thus, the computed value is not “correct”. From the Notes on Conditioning, we know that it may not be the case that  $\hat{f}(x)$  is “close to”  $f(x)$ . After all, even if  $\hat{f}$  is an exact implementation of  $f$ , the mere act of storing  $x$  may introduce a small error  $\delta x$  and  $f(x + \delta x)$  may be far from  $f(x)$  if  $f$  is ill-conditioned.

The following defines a property that captures correctness in the presense of the kinds of errors that are introduced by computer arithmetic:

**Definition 9.1** Let given the mapping  $f : \mathcal{D} \rightarrow \mathcal{R}$ , where  $\mathcal{D} \subset \mathbb{R}^n$  is the domain and  $\mathcal{R} \subset \mathbb{R}^m$  is the range (codomain), let  $\hat{f} : \mathcal{D} \rightarrow \mathcal{R}$  be a computer implementation of this function. We will call  $\hat{f}$  a (numerically) **stable** implementation of  $f$  on domain  $\mathcal{D}$  if for all  $x \in \mathcal{D}$  there exists a  $\hat{x}$  “close” to  $x$  such that  $\hat{f}(x) = f(\hat{x})$ .

In other words,  $\hat{f}$  is a stable implementation if the error that is introduced is similar to that introduced when  $f$  is evaluated with a slightly changed input. This is illustrated in Figure 9.1 for a specific input  $x$ . If an implementation is not stable, it is numerically unstable.

## 9.2 Floating Point Numbers

Only a finite number of (binary) digits can be used to store a real number. For so-called single-precision and double-precision floating point numbers 32 bits and 64 bits are typically employed, respectively. Let us focus on double precision numbers.

Recall that any real number can be written as  $\mu \times \beta^e$ , where  $\beta$  is the base (an integer greater than one),  $\mu \in (-1, 1)$  is the mantissa, and  $e$  is the exponent (an integer). For our discussion, we will define  $F$  as the set of all numbers  $\chi = \mu\beta^e$  such that  $\beta = 2$ ,  $\mu = \pm.\delta_0\delta_1\cdots\delta_{t-1}$  has only  $t$  (binary) digits ( $\delta_j \in \{0, 1\}$ ),  $\delta_0 = 0$  iff  $\mu = 0$  (the mantissa is normalized), and  $-L \leq e \leq U$ . Elements in  $F$  can be stored with a finite number of (binary) digits.

- There is a largest number (in absolute value) that can be stored. Any number that is larger “overflows”. Typically, this causes a value that denotes a NaN (Not-a-Number) to be stored.
- There is a smallest number (in absolute value) that can be stored. Any number that is smaller “underflows”. Typically, this causes a zero to be stored.

**Example 9.2** For  $x \in \mathbb{R}^n$ , consider computing

$$\|x\|_2 = \sqrt{\sum_{i=0}^{n-1} \chi_i^2}. \quad (9.1)$$

Notice that

$$\|x\|_2 \leq \sqrt{n} \max_{i=0}^{n-1} |\chi_i|$$

and hence unless some  $\chi_i$  is close to overflowing, the result will not overflow. The problem is that if some element  $\chi_i$  has the property that  $\chi_i^2$  overflows, intermediate results in the computation in (9.1) will overflow. The solution is to determine  $k$  such that

$$|\chi_k| = \max_{i=0}^{n-1} |\chi_i|$$

and to then instead compute

$$\|x\|_2 = |\chi_k| \sqrt{\sum_{i=0}^{n-1} \left(\frac{\chi_i}{\chi_k}\right)^2}.$$

It can be argued that the same approach also avoids underflow if underflow can be avoided..

In our further discussions, we will ignore overflow and underflow issues.

What is important is that any time a real number is stored in our computer, it is stored as the nearest floating point number (element in  $F$ ). We first assume that it is truncated (which makes our explanation slightly simpler).

Let positive  $\chi$  be represented by

$$\chi = .\delta_0\delta_1 \cdots \times 2^e,$$

where  $\delta_0 = 1$  (the mantissa is normalized). If  $t$  digits are stored by our floating point system, then  $\hat{\chi} = .\delta_0\delta_1 \cdots \delta_{t-1} \times 2^e$  is stored. Let  $\delta\chi = \chi - \hat{\chi}$ . Then

$$\delta\chi = \underbrace{.\delta_0\delta_1 \cdots \delta_{t-1}\delta_t \cdots \times 2^e}_{\chi} - \underbrace{.\delta_0\delta_1 \cdots \delta_{t-1} \times 2^e}_{\hat{\chi}} = \underbrace{.0 \cdots 0}_{t} \delta_t \cdots \times 2^e < \underbrace{.0 \cdots 01}_{t} \times 2^e = 2^{e-t}.$$

Also, since  $\chi$  is positive,

$$\chi = .\delta_0\delta_1 \cdots \times 2^e \geq .1 \times 2^e \geq 2^{e-1}.$$

Thus,

$$\frac{\delta\chi}{\chi} \leq \frac{2^{e-t}}{2^{e-1}} = 2^{1-t}$$

which can also be written as

$$\delta\chi \leq 2^{1-t}\chi.$$

A careful analysis of what happens when  $\chi$  might equal zero or be negative yields

$$|\delta\chi| \leq 2^{1-t}|\chi|.$$

Now, in practice any base  $\beta$  can be used and floating point computation uses rounding rather than truncating. A similar analysis can be used to show that then

$$|\delta\chi| \leq \mathbf{u}|\chi|$$

where  $\mathbf{u} = \frac{1}{2}\beta^{1-t}$  is known as the **machine epsilon** or **unit roundoff**. When using single precision or double precision real arithmetic,  $\mathbf{u} \approx 10^{-8}$  or  $10^{-16}$ , respectively. The quantity  $\mathbf{u}$  is machine dependent; it is a function of the parameters characterizing the machine arithmetic. The unit roundoff is often alternatively defined as the maximum positive floating point number which can be added to the number stored as 1 without changing the number stored as 1. In the notation introduced below,  $\text{fl}(1 + \mathbf{u}) = 1$ .

**Homework 9.3** Assume a floating point number system with  $\beta = 2$  and a mantissa with  $t$  digits so that a typical positive number is written as  $.d_0d_1 \cdots d_{t-1} \times 2^e$ , with  $d_i \in \{0, 1\}$ .

- Write the number 1 as a floating point number.
- What is the largest positive real number  $\mathbf{u}$  (represented as a binary fraction) such that the floating point representation of  $1 + \mathbf{u}$  equals the floating point representation of 1? (Assume rounded arithmetic.)
- Show that  $\mathbf{u} = \frac{1}{2}2^{1-t}$ .

## 9.3 Notation

When discussing error analyses, we will distinguish between exact and computed quantities. The function  $\text{fl}(\text{expression})$  returns the result of the evaluation of *expression*, where every operation is executed in floating point arithmetic. For example, assuming that the expressions are evaluated from left to right,  $\text{fl}(\chi + \psi + \zeta/\omega)$  is equivalent to  $\text{fl}(\text{fl}(\text{fl}(\chi) + \text{fl}(\psi)) + \text{fl}(\text{fl}(\zeta)/\text{fl}(\omega)))$ . Equality between the quantities *lhs* and *rhs* is denoted by  $\text{lhs} = \text{rhs}$ . Assignment of *rhs* to *lhs* is denoted by  $\text{lhs} := \text{rhs}$  (*lhs* becomes *rhs*). In the context of a program, the statements  $\text{lhs} := \text{rhs}$  and  $\text{lhs} := \text{fl}(\text{rhs})$  are equivalent. Given an assignment  $\kappa := \text{expression}$ , we use the notation  $\check{\kappa}$  (pronounced “check kappa”) to denote the quantity resulting from  $\text{fl}(\text{expression})$ , which is actually stored in the variable  $\kappa$ .

## 9.4 Floating Point Computation

We introduce definitions and results regarding floating point arithmetic. In this note, we focus on real valued arithmetic only. Extensions to complex arithmetic are straightforward.

### 9.4.1 Model of floating point computation

The **Standard Computational Model (SCM)** assumes that, for any two floating point numbers  $\chi$  and  $\psi$ , the basic arithmetic operations satisfy the equality

$$\text{fl}(\chi \text{ op } \psi) = (\chi \text{ op } \psi)(1 + \varepsilon), \quad |\varepsilon| \leq \mathbf{u}, \text{ and } \text{op} \in \{+, -, *, /\}.$$

The quantity  $\varepsilon$  is a function of  $\chi, \psi$  and  $\text{op}$ . Sometimes we add a subscript ( $\varepsilon_+, \varepsilon_*, \dots$ ) to indicate what operation generated the  $(1 + \varepsilon)$  error factor. We always assume that all the input variables to an operation are floating point numbers. **We can interpret the SCM as follows: These operations are performed exactly and it is only in storing the result that a roundoff error occurs (equal to that introduced when a real number is stored as a floating point number).**

**Remark 9.4** Given  $\chi, \psi \in F$ , performing any operation  $\text{op} \in \{+, -, *, /\}$  with  $\chi$  and  $\psi$  in floating point arithmetic,  $[\chi \text{ op } \psi]$ , is a stable operation: Let  $\zeta = \chi \text{ op } \psi$  and  $\hat{\zeta} = \zeta + \delta\zeta = [\chi \text{ (op) } \psi]$ . Then  $|\delta\zeta| \leq \mathbf{u}|\zeta|$  and hence  $\hat{\zeta}$  is close to  $\zeta$  (it has  $k$  correct binary digits).

For certain problems it is convenient to use the **Alternative Computational Model (ACM)** [25], which also assumes for the basic arithmetic operations that

$$\text{fl}(\chi \text{ op } \psi) = \frac{\chi \text{ op } \psi}{1 + \varepsilon}, \quad |\varepsilon| \leq \mathbf{u}, \text{ and } \text{op} \in \{+, -, *, /\}.$$

As for the standard computation model, the quantity  $\varepsilon$  is a function of  $\chi, \psi$  and  $\text{op}$ . Note that the  $\varepsilon$ 's produced using the standard and alternative models are generally not equal.

**Remark 9.5** Notice that the Taylor series expansion of  $1/(1 + \varepsilon)$  is given by

$$\frac{1}{1 + \varepsilon} = 1 + (-\varepsilon) + O(\varepsilon^2),$$

which explains how the SCM and ACM are related.

**Remark 9.6** Sometimes it is more convenient to use the SCM and sometimes the ACM. Trial and error, and eventually experience, will determine which one to use.

### 9.4.2 Stability of a numerical algorithm

In the presence of round-off error, an algorithm involving numerical computations cannot be expected to yield the exact result. Thus, the notion of “correctness” applies only to the execution of algorithms in exact arithmetic. Here we briefly introduce the notion of “stability” of algorithms.

Let  $f : \mathcal{D} \rightarrow \mathcal{R}$  be a mapping from the domain  $\mathcal{D}$  to the range  $\mathcal{R}$  and let  $\check{f} : \mathcal{D} \rightarrow \mathcal{R}$  represent the mapping that captures the execution in floating point arithmetic of a given algorithm which computes  $f$ .

The algorithm is said to be **backward stable** if for all  $x \in \mathcal{D}$  there exists a perturbed input  $\check{x} \in \mathcal{D}$ , close to  $x$ , such that  $\check{f}(x) = f(\check{x})$ . In other words, the computed result equals the result obtained when the exact function is applied to slightly perturbed data. The difference between  $\check{x}$  and  $x$ ,  $\delta x = \check{x} - x$ , is the perturbation to the original input  $x$ .

The reasoning behind backward stability is as follows. The input to a function typically has some errors associated with it. Uncertainty may be due to measurement errors when obtaining the input and/or may be the result of converting real numbers to floating point numbers when storing the input on a computer. If it can be shown that an implementation is backward stable, then it has been proved that the result could have been obtained through exact computations performed on slightly corrupted input. Thus, one can think of the error introduced by the implementation as being comparable to the error introduced when obtaining the input data in the first place.

When discussing error analyses,  $\delta x$ , the difference between  $x$  and  $\check{x}$ , is the backward error and the difference  $\check{f}(x) - f(x)$  is the forward error. Throughout the remainder of this note we will be concerned with bounding the backward and/or forward errors introduced by the algorithms executed with floating point arithmetic.

The algorithm is said to be **forward stable** on domain  $\mathcal{D}$  if for all  $x \in \mathcal{D}$  it is that case that  $\check{f}(x) \approx f(x)$ . In other words, the computed result equals a slight perturbation of the exact result.

### 9.4.3 Absolute value of vectors and matrices

In the above discussion of error, the vague notions of “near” and “slightly perturbed” are used. Making these notions exact usually requires the introduction of measures of size for vectors and matrices, i.e., norms. Instead, for the operations analyzed in this note, all bounds are given in terms of the absolute values of the individual elements of the vectors and/or matrices. While it is easy to convert such bounds to bounds involving norms, the converse is not true.

**Definition 9.7** Given  $x \in \mathbb{R}^n$  and  $A \in \mathbb{R}^{m \times n}$ ,

$$|x| = \begin{pmatrix} |\chi_0| \\ |\chi_1| \\ \vdots \\ |\chi_{n-1}| \end{pmatrix} \quad \text{and} \quad |A| = \begin{pmatrix} |\alpha_{0,0}| & |\alpha_{0,1}| & \dots & |\alpha_{0,n-1}| \\ |\alpha_{1,0}| & |\alpha_{1,1}| & \dots & |\alpha_{1,n-1}| \\ \vdots & \vdots & \ddots & \vdots \\ |\alpha_{m-1,0}| & |\alpha_{m-1,1}| & \dots & |\alpha_{m-1,n-1}| \end{pmatrix}.$$

**Definition 9.8** Let  $\Delta \in \{<, \leq, =, \geq, >\}$  and  $x, y \in \mathbb{R}^n$ . Then

$$|x| \Delta |y| \quad \text{iff} \quad |\chi_i| \Delta |\psi_i|,$$

with  $i = 0, \dots, n-1$ . Similarly, given  $A$  and  $B \in \mathbb{R}^{m \times n}$ ,

$$|A| \Delta |B| \quad \text{iff} \quad |\alpha_{ij}| \Delta |\beta_{ij}|,$$

with  $i = 0, \dots, m-1$  and  $j = 0, \dots, n-1$ .

The next Lemma is exploited in later sections:

**Lemma 9.9** *Let  $A \in \mathbb{R}^{m \times k}$  and  $B \in \mathbb{R}^{k \times n}$ . Then  $|AB| \leq |A||B|$ .*

**Homework 9.10** *Prove Lemma 9.9.*

➡ [SEE ANSWER](#)

The fact that the bounds that we establish can be easily converted into bounds involving norms is a consequence of the following theorem, where  $\|\cdot\|_F$  indicates the Frobenius matrix norm.

**Theorem 9.11** *Let  $A, B \in \mathbb{R}^{m \times n}$ . If  $|A| \leq |B|$  then  $\|A\|_1 \leq \|B\|_1$ ,  $\|A\|_\infty \leq \|B\|_\infty$ , and  $\|A\|_F \leq \|B\|_F$ .*

**Homework 9.12** *Prove Theorem 9.11.*

➡ [SEE ANSWER](#)

## 9.5 Stability of the Dot Product Operation

The matrix-vector multiplication algorithm discussed in the next section requires the computation of the dot (inner) product (DOT) of vectors  $x, y \in \mathbb{R}^n$ :  $\kappa := x^T y$ . In this section, we give an algorithm for this operation and the related error results.

### 9.5.1 An algorithm for computing DOT

We will consider the algorithm given in Figure 9.2. It uses the FLAME notation [24, 4] to express the computation

$$\kappa := \left( ((\chi_0 \psi_0 + \chi_1 \psi_1) + \cdots) + \chi_{n-2} \psi_{n-2} \right) + \chi_{n-1} \psi_{n-1} \quad (9.2)$$

in the indicated order.

### 9.5.2 A simple start

Before giving a general result, let us focus on the case where  $n = 2$ :

$$\kappa := \chi_0 \psi_0 + \chi_1 \psi_1.$$

Then, under the computational model given in Section 9.4, if  $\kappa := \chi_0 \psi_0 + \chi_1 \psi_1$  is executed, the computed result,  $\check{\kappa}$ , satisfies

$$\begin{aligned} \check{\kappa} &= \left[ \begin{pmatrix} \chi_0 \\ \chi_1 \end{pmatrix}^T \begin{pmatrix} \psi_0 \\ \psi_1 \end{pmatrix} \right] \\ &= [\chi_0 \psi_0 + \chi_1 \psi_1] \\ &= [[\chi_0 \psi_0] + [\chi_1 \psi_1]] \\ &= [\chi_0 \psi_0 (1 + \epsilon_*^{(0)}) + \chi_1 \psi_1 (1 + \epsilon_*^{(1)})] \\ &= (\chi_0 \psi_0 (1 + \epsilon_*^{(0)}) + \chi_1 \psi_1 (1 + \epsilon_*^{(1)})) (1 + \epsilon_+^{(1)}) \\ &= \chi_0 \psi_0 (1 + \epsilon_*^{(0)}) (1 + \epsilon_+^{(1)}) + \chi_1 \psi_1 (1 + \epsilon_*^{(1)}) (1 + \epsilon_+^{(1)}) \end{aligned}$$



**Algorithm:**  $[A, t] := \text{HOUSEQR\_BLK\_VAR1}(A, t)$

$$x \rightarrow \begin{pmatrix} \frac{x_T}{x_B} \end{pmatrix}, y \rightarrow \begin{pmatrix} \frac{y_T}{y_B} \end{pmatrix}$$

**while**  $m(x_T) < m(x)$  **do**

$$\begin{pmatrix} \frac{x_T}{x_B} \end{pmatrix} \rightarrow \begin{pmatrix} \frac{x_0}{\chi_1} \\ \frac{x_2}{x_2} \end{pmatrix}, \begin{pmatrix} \frac{y_T}{y_B} \end{pmatrix} \rightarrow \begin{pmatrix} \frac{y_0}{\psi_1} \\ \frac{y_2}{y_2} \end{pmatrix}$$

---


$$\kappa := \kappa + \chi_1 \psi_1$$


---

$$\begin{pmatrix} \frac{x_T}{x_B} \end{pmatrix} \leftarrow \begin{pmatrix} \frac{x_0}{\chi_1} \\ \frac{x_2}{x_2} \end{pmatrix}, \begin{pmatrix} \frac{y_T}{y_B} \end{pmatrix} \leftarrow \begin{pmatrix} \frac{y_0}{\psi_1} \\ \frac{y_2}{y_2} \end{pmatrix}$$

**endwhile**

Figure 9.2: Algorithm for computing  $\kappa := x^T y$ .

$$\begin{aligned} &= \begin{pmatrix} \chi_0 \\ \chi_1 \end{pmatrix}^T \begin{pmatrix} \psi_0(1 + \epsilon_*^{(0)})(1 + \epsilon_+^{(1)}) \\ \psi_1(1 + \epsilon_*^{(1)})(1 + \epsilon_+^{(1)}) \end{pmatrix} \\ &= \begin{pmatrix} \chi_0 \\ \chi_1 \end{pmatrix}^T \begin{pmatrix} (1 + \epsilon_*^{(0)})(1 + \epsilon_+^{(1)}) & 0 \\ 0 & (1 + \epsilon_*^{(1)})(1 + \epsilon_+^{(1)}) \end{pmatrix} \begin{pmatrix} \psi_0 \\ \psi_1 \end{pmatrix} \\ &= \begin{pmatrix} \chi_0(1 + \epsilon_*^{(0)})(1 + \epsilon_+^{(1)}) \\ \chi_1(1 + \epsilon_*^{(1)})(1 + \epsilon_+^{(1)}) \end{pmatrix}^T \begin{pmatrix} \psi_0 \\ \psi_1 \end{pmatrix}, \end{aligned}$$

where  $|\epsilon_*^{(0)}|, |\epsilon_*^{(1)}|, |\epsilon_+^{(1)}| \leq \mathbf{u}$ .

**Homework 9.13** Repeat the above steps for the computation

$$\kappa := ((\chi_0 \psi_0 + \chi_1 \psi_1) + \chi_2 \psi_2),$$

computing in the indicated order.

 [SEE ANSWER](#)

### 9.5.3 Preparation

Under the computational model given in Section 9.4, the computed result of (9.2),  $\check{\mathbf{x}}$ , satisfies

$$\begin{aligned}\check{\mathbf{x}} &= \left( \left( (\chi_0 \psi_0 (1 + \epsilon_*^{(0)}) + \chi_1 \psi_1 (1 + \epsilon_*^{(1)})) (1 + \epsilon_+^{(1)}) + \dots \right) (1 + \epsilon_+^{(n-2)}) \right. \\ &\quad \left. + \chi_{n-1} \psi_{n-1} (1 + \epsilon_*^{(n-1)}) \right) (1 + \epsilon_+^{(n-1)}) \\ &= \sum_{i=0}^{n-1} \left( \chi_i \psi_i (1 + \epsilon_*^{(i)}) \prod_{j=i}^{n-1} (1 + \epsilon_+^{(j)}) \right),\end{aligned}\tag{9.3}$$

where  $\epsilon_+^{(0)} = 0$  and  $|\epsilon_*^{(0)}|, |\epsilon_*^{(j)}|, |\epsilon_+^{(j)}| \leq \mathbf{u}$  for  $j = 1, \dots, n-1$ .

Clearly, a notation to keep expressions from becoming unreadable is desirable. For this reason we introduce the symbol  $\theta_j$ :

**Lemma 9.14** *Let  $\epsilon_i \in \mathbb{R}$ ,  $0 \leq i \leq n-1$ ,  $n\mathbf{u} < 1$ , and  $|\epsilon_i| \leq \mathbf{u}$ . Then  $\exists \theta_n \in \mathbb{R}$  such that*

$$\prod_{i=0}^{n-1} (1 + \epsilon_i)^{\pm 1} = 1 + \theta_n,$$

with  $|\theta_n| \leq n\mathbf{u}/(1 - n\mathbf{u})$ .

**Proof:** By Mathematical Induction.

**Base case.**  $n = 1$ . Trivial.

**Inductive Step.** The Inductive Hypothesis (I.H.) tells us that for all  $\epsilon_i \in \mathbb{R}$ ,  $0 \leq i \leq n-1$ ,  $n\mathbf{u} < 1$ , and  $|\epsilon_i| \leq \mathbf{u}$ , there exists a  $\theta_n \in \mathbb{R}$  such that

$$\prod_{i=0}^{n-1} (1 + \epsilon_i)^{\pm 1} = 1 + \theta_n, \text{ with } |\theta_n| \leq n\mathbf{u}/(1 - n\mathbf{u}).$$

We will show that if  $\epsilon_i \in \mathbb{R}$ ,  $0 \leq i \leq n$ ,  $(n+1)\mathbf{u} < 1$ , and  $|\epsilon_i| \leq \mathbf{u}$ , then there exists a  $\theta_{n+1} \in \mathbb{R}$  such that

$$\prod_{i=0}^n (1 + \epsilon_i)^{\pm 1} = 1 + \theta_{n+1}, \text{ with } |\theta_{n+1}| \leq (n+1)\mathbf{u}/(1 - (n+1)\mathbf{u}).$$

**Case 1:**  $\prod_{i=0}^n (1 + \epsilon_i)^{\pm 1} = \prod_{i=0}^{n-1} (1 + \epsilon_i)^{\pm 1} (1 + \epsilon_n)$ . See Exercise 9.15.

**Case 2:**  $\prod_{i=0}^n (1 + \epsilon_i)^{\pm 1} = (\prod_{i=0}^{n-1} (1 + \epsilon_i)^{\pm 1}) / (1 + \epsilon_n)$ . By the I.H. there exists a  $\theta_n$  such that  $(1 + \theta_n) = \prod_{i=0}^{n-1} (1 + \epsilon_i)^{\pm 1}$  and  $|\theta_n| \leq n\mathbf{u}/(1 - n\mathbf{u})$ . Then

$$\frac{\prod_{i=0}^{n-1} (1 + \epsilon_i)^{\pm 1}}{1 + \epsilon_n} = \frac{1 + \theta_n}{1 + \epsilon_n} = 1 + \underbrace{\frac{\theta_n - \epsilon_n}{1 + \epsilon_n}}_{\theta_{n+1}},$$

which tells us how to pick  $\theta_{n+1}$ . Now

$$\begin{aligned}|\theta_{n+1}| &= \left| \frac{\theta_n - \epsilon_n}{1 + \epsilon_n} \right| \leq \frac{|\theta_n| + \mathbf{u}}{1 - \mathbf{u}} \leq \frac{\frac{n\mathbf{u}}{1 - n\mathbf{u}} + \mathbf{u}}{1 - \mathbf{u}} = \frac{n\mathbf{u} + (1 - n\mathbf{u})\mathbf{u}}{(1 - n\mathbf{u})(1 - \mathbf{u})} \\ &= \frac{(n+1)\mathbf{u} - n\mathbf{u}^2}{1 - (n+1)\mathbf{u} + n\mathbf{u}^2} \leq \frac{(n+1)\mathbf{u}}{1 - (n+1)\mathbf{u}}.\end{aligned}$$

By the Principle of Mathematical Induction, the result holds.

**Homework 9.15** Complete the proof of Lemma 9.14.

➡ SEE ANSWER

The quantity  $\theta_n$  will be used throughout this note. **It is not intended to be a specific number.** Instead, it is an order of magnitude identified by the subscript  $n$ , which indicates the number of error factors of the form  $(1 + \varepsilon_i)$  and/or  $(1 + \varepsilon_i)^{-1}$  that are grouped together to form  $(1 + \theta_n)$ . Since the bound on  $|\theta_n|$  occurs often, we assign it a symbol as follows:

**Definition 9.16** For all  $n \geq 1$  and  $n\mathbf{u} < 1$ , define  $\gamma_n := n\mathbf{u}/(1 - n\mathbf{u})$ .

With this notation, (9.3) simplifies to

$$\check{\mathbf{x}} = \chi_0 \psi_0 (1 + \theta_n) + \chi_1 \psi_1 (1 + \theta_n) + \cdots + \chi_{n-1} \psi_{n-1} (1 + \theta_2) \quad (9.4)$$

$$= \begin{pmatrix} \chi_0 \\ \chi_1 \\ \chi_2 \\ \vdots \\ \chi_{n-1} \end{pmatrix}^T \begin{pmatrix} (1 + \theta_n) & 0 & 0 & \cdots & 0 \\ 0 & (1 + \theta_n) & 0 & \cdots & 0 \\ 0 & 0 & (1 + \theta_{n-1}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (1 + \theta_2) \end{pmatrix} \begin{pmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{n-1} \end{pmatrix} \quad (9.5)$$

$$= \begin{pmatrix} \chi_0 \\ \chi_1 \\ \chi_2 \\ \vdots \\ \chi_{n-1} \end{pmatrix}^T \left( I + \begin{pmatrix} \theta_n & 0 & 0 & \cdots & 0 \\ 0 & \theta_n & 0 & \cdots & 0 \\ 0 & 0 & \theta_{n-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \theta_2 \end{pmatrix} \right) \begin{pmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{n-1} \end{pmatrix},$$

where  $|\theta_j| \leq \gamma_j$ ,  $j = 2, \dots, n$ .

Two instances of the symbol  $\theta_n$ , appearing even in the same expression, typically do not represent the same number. For example, in (9.4) a  $(1 + \theta_n)$  multiplies each of the terms  $\chi_0 \psi_0$  and  $\chi_1 \psi_1$ , but these two instances of  $\theta_n$ , as a rule, do not denote the same quantity. In particular, One should be careful when factoring out such quantities.

As part of the analyses the following bounds will be useful to bound error that accumulates:

**Lemma 9.17** If  $n, b \geq 1$  then  $\gamma_n \leq \gamma_{n+b}$  and  $\gamma_n + \gamma_b + \gamma_n \gamma_b \leq \gamma_{n+b}$ .

**Homework 9.18** Prove Lemma 9.17.

➡ SEE ANSWER

### 9.5.4 Target result

It is of interest to accumulate the roundoff error encountered during computation as a perturbation of input and/or output parameters:

- $\check{\kappa} = (x + \delta x)^T y$ ; (  $\check{\kappa}$  is the exact output for a slightly perturbed  $x$  )
- $\check{\kappa} = x^T (y + \delta y)$ ; (  $\check{\kappa}$  is the exact output for a slightly perturbed  $y$  )
- $\check{\kappa} = x^T y + \delta \kappa$ . (  $\check{\kappa}$  equals the exact result plus an error )

The first two are backward error results (error is accumulated onto input parameters, showing that the algorithm is numerically stable since it yields the exact output for a slightly perturbed input) while the last one is a forward error result (error is accumulated onto the answer). We will see that in different situations, a different error result may be needed by analyses of operations that require a dot product.

Let us focus on the second result. Ideally one would show that each of the entries of  $y$  is slightly perturbed relative to that entry:

$$\delta y = \begin{pmatrix} \sigma_0 \psi_0 \\ \vdots \\ \sigma_{n-1} \psi_{n-1} \end{pmatrix} = \begin{pmatrix} \sigma_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{n-1} \end{pmatrix} \begin{pmatrix} \psi_0 \\ \vdots \\ \psi_{n-1} \end{pmatrix} = \Sigma y,$$

where each  $\sigma_i$  is “small” and  $\Sigma = \text{diag}((\sigma_0, \dots, \sigma_{n-1}))$ . The following special structure of  $\Sigma$ , inspired by (9.5) will be used in the remainder of this note:

$$\Sigma^{(n)} = \begin{cases} 0 \times 0 \text{ matrix} & \text{if } n = 0 \\ \theta_1 & \text{if } n = 1 \\ \text{diag}((\theta_n, \theta_n, \theta_{n-1}, \dots, \theta_2)) & \text{otherwise.} \end{cases} \quad (9.6)$$

Recall that  $\theta_j$  is an order of magnitude variable with  $|\theta_j| \leq \gamma_j$ .

**Homework 9.19** Let  $k \geq 0$  and assume that  $|\varepsilon_1|, |\varepsilon_2| \leq \mathbf{u}$ , with  $\varepsilon_1 = 0$  if  $k = 0$ . Show that

$$\left( \begin{array}{c|c} I + \Sigma^{(k)} & 0 \\ \hline 0 & (1 + \varepsilon_1) \end{array} \right) (1 + \varepsilon_2) = (I + \Sigma^{(k+1)}).$$

*Hint: reason the case where  $k = 0$  separately from the case where  $k > 0$ .*

 [SEE ANSWER](#)

We state a theorem that captures how error is accumulated by the algorithm.

**Theorem 9.20** Let  $x, y \in \mathbb{R}^n$  and let  $\kappa := x^T y$  be computed by executing the algorithm in Figure 9.2. Then

$$\check{\kappa} = [x^T y] = x^T (I + \Sigma^{(n)}) y.$$

### 9.5.5 A proof in traditional format

In the below proof, we will pick symbols to denote vectors so that the proof can be easily related to the alternative framework to be presented in Section 9.5.6.

**Proof:** By Mathematical Induction on  $n$ , the length of vectors  $x$  and  $y$ .

**Base case.**  $m(x) = m(y) = 0$ . Trivial.

**Inductive Step.** I.H.: Assume that if  $x_T, y_T \in \mathbb{R}^k$ ,  $k > 0$ , then

$$\text{fl}(x_T^T y_T) = x_T^T (I + \Sigma_T) y_T, \text{ where } \Sigma_T = \Sigma^{(k)}.$$

We will show that when  $x_T, y_T \in \mathbb{R}^{k+1}$ , the equality  $\text{fl}(x_T^T y_T) = x_T^T (I + \Sigma_T) y_T$  holds *true* again.

Assume that  $x_T, y_T \in \mathbb{R}^{k+1}$ , and partition  $x_T \rightarrow \begin{pmatrix} x_0 \\ \chi_1 \end{pmatrix}$  and  $y_T \rightarrow \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix}$ . Then

$$\begin{aligned} \text{fl}\left(\begin{pmatrix} x_0 \\ \chi_1 \end{pmatrix}^T \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix}\right) &= \text{fl}(\text{fl}(x_0^T y_0) + \text{fl}(\chi_1^T \psi_1)) && \text{(definition)} \\ &= \text{fl}(x_0^T (I + \Sigma_0) y_0 + \text{fl}(\chi_1^T \psi_1)) && \text{(I.H. with } x_T = x_0, \\ &&& y_T = y_0, \text{ and } \Sigma_0 = \Sigma^{(k)}) \\ &= (x_0^T (I + \Sigma_0) y_0 + \chi_1^T \psi_1 (1 + \varepsilon_*)) (1 + \varepsilon_+) && \text{(SCM, twice)} \\ &= \begin{pmatrix} x_0 \\ \chi_1 \end{pmatrix}^T \left( \begin{array}{c|c} (I + \Sigma_0) & 0 \\ \hline 0 & (1 + \varepsilon_*) \end{array} \right) (1 + \varepsilon_+) \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix} && \text{(rearrangement)} \\ &= x_T^T (I + \Sigma_T) y_T && \text{(renaming),} \end{aligned}$$

where  $|\varepsilon_*|, |\varepsilon_+| \leq \mathbf{u}$ ,  $\varepsilon_+ = 0$  if  $k = 0$ , and  $(I + \Sigma_T) = \begin{pmatrix} (I + \Sigma_0) & 0 \\ \hline 0 & (1 + \varepsilon_*) \end{pmatrix} (1 + \varepsilon_+)$  so that  $\Sigma_T = \Sigma^{(k+1)}$ .

**By the Principle of Mathematical Induction**, the result holds.

### 9.5.6 A weapon of math induction for the war on error (optional)

We focus the reader's attention on Figure 9.3 in which we present a framework, which we will call the **error worksheet**, for presenting the inductive proof of Theorem 9.20 side-by-side with the algorithm for DOT. This framework, in a slightly different form, was first introduced in [3]. The expressions enclosed by  $\{ \}$  (in the grey boxes) are predicates describing the state of the variables used in the algorithms and in their analysis. In the worksheet, we use superscripts to indicate the iteration number, thus, the symbols  $v^i$  and  $v^{i+1}$  do not denote two different variables, but two different states of variable  $v$ .

		Error side	Step
$\kappa := 0$		$\{ \Sigma = 0 \}$	1a
<b>Partition</b> $x \rightarrow \left( \frac{x_T}{x_B} \right), y \rightarrow \left( \frac{y_T}{y_B} \right),$ <b>where</b> $x_T$ and $y_T$ are empty, and $\Sigma_T$ is $0 \times 0$		$\Sigma \rightarrow \left( \frac{\Sigma_T \mid 0}{0 \mid \Sigma_B} \right)$	4
$\{ \check{\kappa} = x_T^T (I + \Sigma_T) y_T \wedge \Sigma_T = \Sigma^{(k)} \wedge m(x_T) = k \}$			2a
<b>while</b> $m(x_T) < m(x)$ <b>do</b>			3
$\{ \check{\kappa} = x_T^T (I + \Sigma_T) y_T \wedge \Sigma_T = \Sigma^{(k)} \wedge m(x_T) = k \}$			2b
<b>Repartition</b> $\left( \frac{x_T}{x_B} \right) \rightarrow \left( \frac{x_0}{\chi_1} \right), \left( \frac{y_T}{y_B} \right) \rightarrow \left( \frac{y_0}{\psi_1} \right),$ <b>where</b> $\chi_1, \psi_1$ , and $\sigma_1^i$ are scalars		$\left( \frac{\Sigma_T \mid 0}{0 \mid \Sigma_B} \right) \rightarrow \left( \frac{\Sigma_0^i \mid 0 \mid 0}{0 \mid \sigma_1^i \mid 0} \right)$	5a
$\{ \check{\kappa}^i = x_0^T (I + \Sigma_0^i) y_0 \wedge \Sigma_0^i = \Sigma^{(k)} \wedge m(x_0) = k \}$			6
$\kappa := \kappa + \chi_1 \psi_1$	$\check{\kappa}^{i+1} = (\check{\kappa}^i + \chi_1 \psi_1 (1 + \epsilon_*)) (1 + \epsilon_+)$ $= (x_0^T (I + \Sigma_0^{(k)}) y_0 + \chi_1 \psi_1 (1 + \epsilon_*)) (1 + \epsilon_+)$ $= \left( \frac{x_0}{\chi_1} \right)^T \left( \frac{I + \Sigma_0^{(k)} \mid 0}{0 \mid 1 + \epsilon_*} \right) (1 + \epsilon_+) \left( \frac{y_0}{\psi_1} \right)$ $= \left( \frac{x_0}{\chi_1} \right)^T (I + \Sigma^{(k+1)}) \left( \frac{y_0}{\psi_1} \right)$	SCM, twice $(\epsilon_+ = 0 \text{ if } k = 0)$ Step 6: I.H. Rearrange Exercise 9.19	8
$\left\{ \check{\kappa}^{i+1} = \left( \frac{x_0}{\chi_1} \right)^T \left( I + \left( \frac{\Sigma_0^{i+1} \mid 0}{0 \mid \sigma_1^{i+1}} \right) \right) \left( \frac{y_0}{\psi_1} \right) \right\}$ $\wedge \left( \frac{\Sigma_0^{i+1} \mid 0}{0 \mid \sigma_1^{i+1}} \right) = \Sigma^{(k+1)} \wedge m \left( \frac{x_0}{\chi_1} \right) = (k+1)$			7
<b>Continue with</b> $\left( \frac{x_T}{x_B} \right) \leftarrow \left( \frac{x_0}{\chi_1} \right), \left( \frac{y_T}{y_B} \right) \leftarrow \left( \frac{y_0}{\psi_1} \right),$ $\left( \frac{\Sigma_T \mid 0}{0 \mid \Sigma_B} \right) \leftarrow \left( \frac{\Sigma_0^{i+1} \mid 0 \mid 0}{0 \mid \sigma_1^{i+1} \mid 0} \right)$			5b
$\{ \check{\kappa} = x_T^T (I + \Sigma_T) y_T \wedge \Sigma_T = \Sigma^{(k)} \wedge m(x_T) = k \}$			2c
<b>endwhile</b>			
$\{ \check{\kappa} = x_T^T (I + \Sigma_T) y_T \wedge \Sigma_T = \Sigma^{(k)} \wedge m(x_T) = k \wedge m(x_T) = m(x) \}$			2d
$\{ \check{\kappa} = x^T (I + \Sigma^{(n)}) y \wedge m(x) = n \}$			1b

Figure 9.3: Error worksheet completed to establish the backward error result for the given algorithm that computes the DOT operation.

The proof presented in Figure 9.3 goes hand in hand with the algorithm, as it shows that before and after each iteration of the loop that computes  $\kappa := x^T y$ , the variables  $\check{\kappa}, x_T, y_T, \Sigma_T$  are such that the predicate

$$\{\check{\kappa} = x_T^T (I + \Sigma_T) y_T \wedge k = m(x_T) \wedge \Sigma_T = \Sigma^{(k)}\} \quad (9.7)$$

holds *true*. This relation is satisfied at each iteration of the loop, so it is also satisfied when the loop completes. Upon completion, the loop guard is  $m(x_T) = m(x) = n$ , which implies that  $\check{\kappa} = x^T (I + \Sigma^{(n)}) y$ , i.e., the thesis of the theorem, is satisfied too.

In details, the inductive proof of Theorem 9.20 is captured by the error worksheet as follows:

**Base case.** In Step 2a, i.e. before the execution of the loop, predicate (9.7) is satisfied, as  $k = m(x_T) = 0$ .

**Inductive step.** Assume that the predicate (9.7) holds *true* at Step 2b, i.e., at the top of the loop. Then Steps 6, 7, and 8 in Figure 9.3 prove that the predicate is satisfied again at Step 2c, i.e., the bottom of the loop. Specifically,

- Step 6 holds by virtue of the equalities  $x_0 = x_T, y_0 = y_T$ , and  $\Sigma_0^i = \Sigma_T$ .
- The update in Step 8-left introduces the error indicated in Step 8-right (SCM, twice), yielding the results for  $\Sigma_0^{i+1}$  and  $\sigma_1^{i+1}$ , leaving the variables in the state indicated in Step 7.
- Finally, the redefinition of  $\Sigma_T$  in Step 5b transforms the predicate in Step 7 into that of Step 2c, completing the inductive step.

**By the Principle of Mathematical Induction**, the predicate (9.7) holds for all iterations. In particular, when the loop terminates, the predicate becomes

$$\check{\kappa} = x^T (I + \Sigma^{(n)}) y \wedge n = m(x_T).$$

This completes the discussion of the proof as captured by Figure 9.3.

In the derivation of algorithms, the concept of **loop-invariant** plays a central role. Let  $\mathcal{L}$  be a loop and  $\mathcal{P}$  a predicate. If  $\mathcal{P}$  is *true* before the execution of  $\mathcal{L}$ , at the beginning and at the end of each iteration of  $\mathcal{L}$ , and after the completion of  $\mathcal{L}$ , then predicate  $\mathcal{P}$  is a *loop-invariant* with respect to  $\mathcal{L}$ . Similarly, we give the definition of **error-invariant**.

**Definition 9.21** *We call the predicate involving the operands and error operands in Steps 2a–d the **error-invariant** for the analysis. This predicate is true before and after each iteration of the loop.*

For any algorithm, the loop-invariant and the error-invariant are related in that the former describes the status of the computation at the beginning and the end of each iteration, while the latter captures an error result for the computation indicated by the loop-invariant.

The reader will likely think that the error worksheet is an overkill when proving the error result for the dot product. We agree. However, it links a proof by induction to the execution of a loop, which we believe is useful. Elsewhere, as more complex operations are analyzed, the benefits of the structure that the error worksheet provides will become more obvious. (We will analyze more complex algorithms as the course proceeds.)

### 9.5.7 Results

A number of useful consequences of Theorem 9.20 follow. These will be used later as an inventory (library) of error results from which to draw when analyzing operations and algorithms that utilize DOT.

**Corollary 9.22** *Under the assumptions of Theorem 9.20 the following relations hold:*

*R1-B: (Backward analysis)  $\check{x} = (x + \delta x)^T y$ , where  $|\delta x| \leq \gamma_n |x|$ , and  $\check{y} = x^T (y + \delta y)$ , where  $|\delta y| \leq \gamma_n |y|$ ;*

*R1-F: (Forward analysis)  $\check{y} = x^T y + \delta \kappa$ , where  $|\delta \kappa| \leq \gamma_n |x|^T |y|$ .*

**Proof:** We leave the proof of R1-B as an exercise. For R1-F, let  $\delta \kappa = x^T \Sigma^{(n)} y$ , where  $\Sigma^{(n)}$  is as in Theorem 9.20. Then

$$\begin{aligned} |\delta \kappa| &= |x^T \Sigma^{(n)} y| \\ &\leq |\chi_0| |\theta_n| |\psi_0| + |\chi_1| |\theta_n| |\psi_1| + \cdots + |\chi_{n-1}| |\theta_2| |\psi_{n-1}| \\ &\leq \gamma_n |\chi_0| |\psi_0| + \gamma_n |\chi_1| |\psi_1| + \cdots + \gamma_n |\chi_{n-1}| |\psi_{n-1}| \\ &\leq \gamma_n |x|^T |y|. \end{aligned}$$

**Homework 9.23** *Prove R1-B.*

 [SEE ANSWER](#)

## 9.6 Stability of a Matrix-Vector Multiplication Algorithm

In this section, we discuss the numerical stability of the specific matrix-vector multiplication algorithm that computes  $y := Ax$  via dot products. This allows us to show how results for the dot product can be used in the setting of a more complicated algorithm.

### 9.6.1 An algorithm for computing GEMV

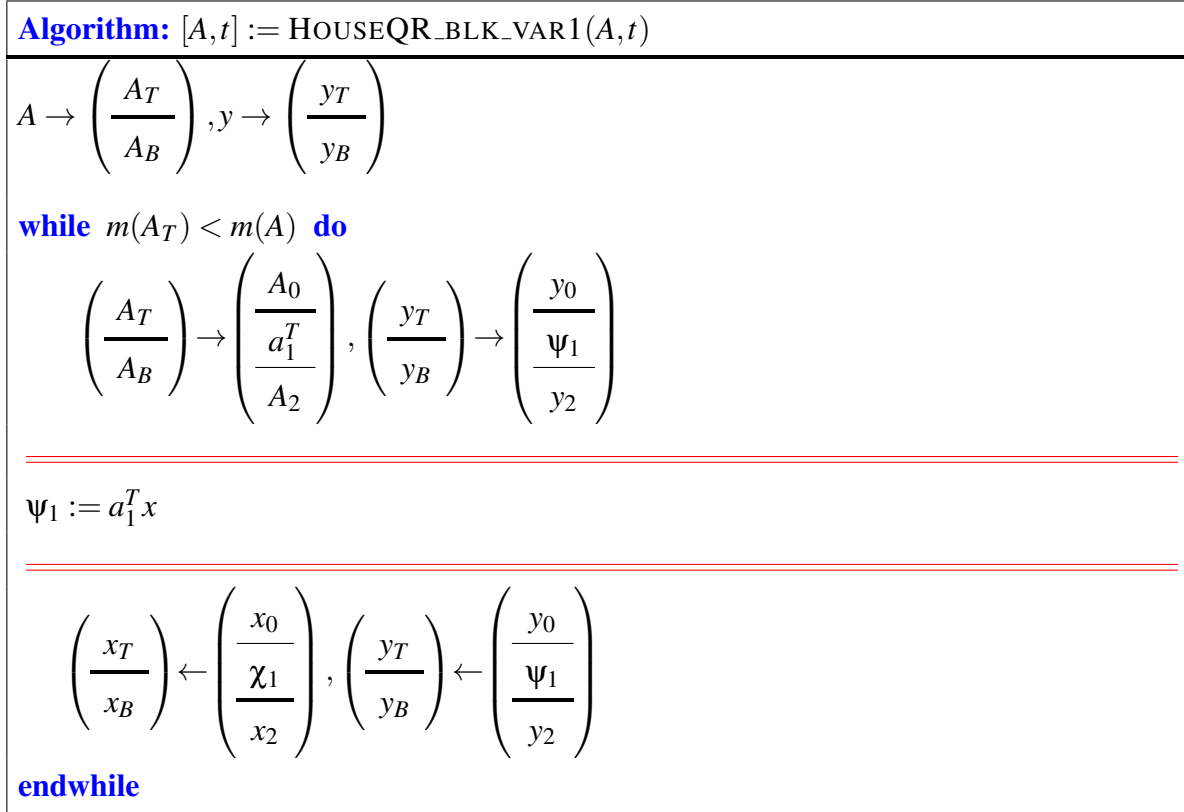
We will consider the algorithm given in Figure 9.4 for computing  $y := Ax$ , which computes  $y$  via dot products.

### 9.6.2 Analysis

Assume  $A \in \mathbb{R}^{m \times n}$  and partition

$$A = \begin{pmatrix} a_0^T \\ a_1^T \\ \vdots \\ a_{m-1}^T \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix}.$$



Figure 9.4: Algorithm for computing  $y := Ax$ .

Then

$$\begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix} := \begin{pmatrix} a_0^T x \\ a_1^T x \\ \vdots \\ a_{m-1}^T x \end{pmatrix}.$$

From Corollary 9.22 R1-B regarding the dot product we know that

$$\check{y} = \begin{pmatrix} \check{\psi}_0 \\ \check{\psi}_1 \\ \vdots \\ \check{\psi}_{m-1} \end{pmatrix} = \begin{pmatrix} (a_0 + \delta a_0)^T x \\ (a_1 + \delta a_1)^T x \\ \vdots \\ (a_{m-1} + \delta a_{m-1})^T x \end{pmatrix} = \left( \begin{pmatrix} a_0^T \\ a_1^T \\ \vdots \\ a_{m-1}^T \end{pmatrix} + \begin{pmatrix} \delta a_0^T \\ \delta a_1^T \\ \vdots \\ \delta a_{m-1}^T \end{pmatrix} \right) x = (A + \Delta A)x,$$

where  $|\delta a_i| \leq \gamma_n |a_i|$ ,  $i = 0, \dots, m-1$ , and hence  $|\Delta A| \leq \gamma_n |A|$ .

Also, from Corollary 9.22 R1-F regarding the dot product we know that

$$\tilde{y} = \begin{pmatrix} \tilde{\psi}_0 \\ \tilde{\psi}_1 \\ \vdots \\ \tilde{\psi}_{m-1} \end{pmatrix} = \begin{pmatrix} a_0^T x + \delta\psi_0 \\ a_1^T x + \delta\psi_1 \\ \vdots \\ a_{m-1}^T x + \delta\psi_{m-1} \end{pmatrix} = \begin{pmatrix} a_0^T \\ a_1^T \\ \vdots \\ a_{m-1}^T \end{pmatrix} x + \begin{pmatrix} \delta\psi_0 \\ \delta\psi_1 \\ \vdots \\ \delta\psi_{m-1} \end{pmatrix} = Ax + \delta y.$$

where  $|\delta\psi_i| \leq \gamma_n |a_i|^T |x|$  and hence  $|\delta y| \leq \gamma_n |A| |x|$ .

The above observations can be summarized in the following theorem:

**Theorem 9.24 Error results for matrix-vector multiplication.** *Let  $A \in \mathbb{R}^{m \times n}$ ,  $x \in \mathbb{R}^n$ ,  $y \in \mathbb{R}^m$  and consider the assignment  $y := Ax$  implemented via the algorithm in Figure 9.4. Then these equalities hold:*

R1-B:  $\tilde{y} = (A + \Delta A)x$ , where  $|\Delta A| \leq \gamma_n |A|$ .

R2-F:  $\tilde{y} = Ax + \delta y$ , where  $|\delta y| \leq \gamma_n |A| |x|$ .

**Homework 9.25** *In the above theorem, could one instead prove the result*

$$\tilde{y} = A(x + \delta x),$$

where  $\delta x$  is “small”?

 [SEE ANSWER](#)

## 9.7 Stability of a Matrix-Matrix Multiplication Algorithm

In this section, we discuss the numerical stability of the specific matrix-matrix multiplication algorithm that computes  $C := AB$  via the matrix-vector multiplication algorithm from the last section, where  $C \in \mathbb{R}^{m \times n}$ ,  $A \in \mathbb{R}^{m \times k}$ , and  $B \in \mathbb{R}^{k \times n}$ .

### 9.7.1 An algorithm for computing GEMM

We will consider the algorithm given in Figure 9.5 for computing  $C := AB$ , which computes one column at a time so that the matrix-vector multiplication algorithm from the last section can be used.

### 9.7.2 Analysis

Partition

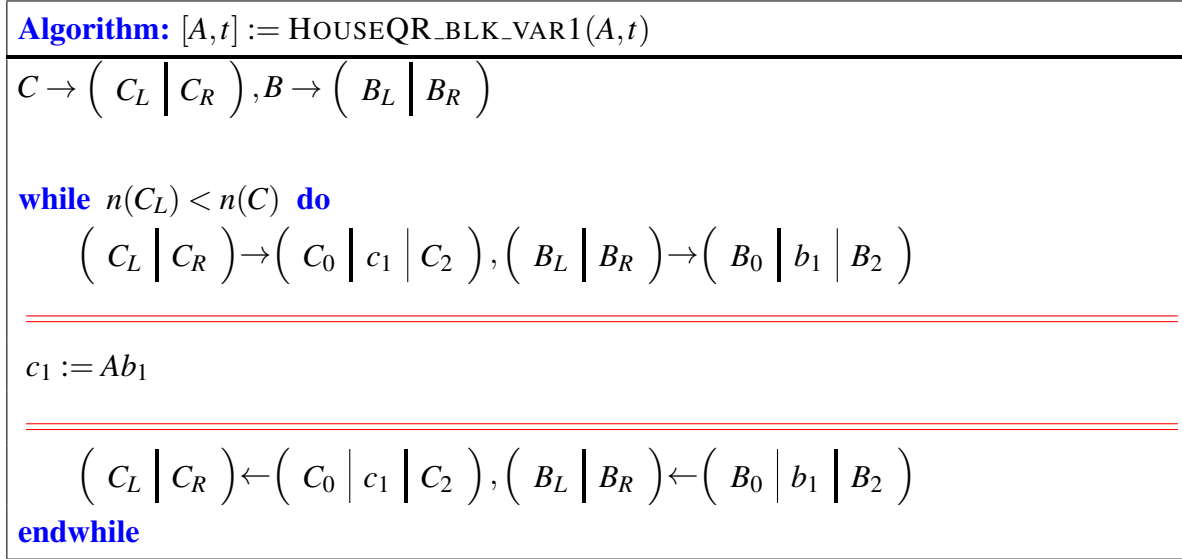
$$C = \left( c_0 \mid c_1 \mid \cdots \mid c_{n-1} \right) \quad \text{and} \quad B = \left( b_0 \mid b_1 \mid \cdots \mid b_{n-1} \right).$$

Then

$$\left( c_0 \mid c_1 \mid \cdots \mid c_{n-1} \right) := \left( Ab_0 \mid Ab_1 \mid \cdots \mid Ab_{n-1} \right).$$

From Corollary 9.22 R1-B regarding the dot product we know that

$$\begin{aligned} \left( \check{c}_0 \mid \check{c}_1 \mid \cdots \mid \check{c}_{n-1} \right) &= \left( Ab_0 + \delta c_0 \mid Ab_1 + \delta c_1 \mid \cdots \mid Ab_{n-1} + \delta c_{n-1} \right) \\ &= \left( Ab_0 \mid Ab_1 \mid \cdots \mid Ab_{n-1} \right) + \left( \delta c_0 \mid \delta c_1 \mid \cdots \mid \delta c_{n-1} \right) \\ &= AB + \Delta C. \end{aligned}$$

Figure 9.5: Algorithm for computing  $C := AB$  one column at a time.

where  $|\delta c_j| \leq \gamma_k |A| |b_j|$ ,  $j = 0, \dots, n-1$ , and hence  $|\Delta C| \leq \gamma_k |A| |B|$ .

The above observations can be summarized in the following theorem:

**Theorem 9.26 (Forward) error results for matrix-matrix multiplication.** *Let  $C \in \mathbb{R}^{m \times n}$ ,  $A \in \mathbb{R}^{m \times k}$ , and  $B \in \mathbb{R}^{k \times n}$  and consider the assignment  $C := AB$  implemented via the algorithm in Figure 9.5. Then the following equality holds:*

*R1-F:  $\check{C} = AB + \Delta C$ , where  $|\Delta C| \leq \gamma_k |A| |B|$ .*

**Homework 9.27** *In the above theorem, could one instead prove the result*

$$\check{C} = (A + \Delta A)(B + \Delta B),$$

*where  $\Delta A$  and  $\Delta B$  are “small”?*

🔗 [SEE ANSWER](#)

### 9.7.3 An application

A collaborator of ours recently implemented a matrix-matrix multiplication algorithm and wanted to check if it gave the correct answer. To do so, he followed the following steps:

- He created random matrices  $A \in \mathbb{R}^{m \times k}$ , and  $C \in \mathbb{R}^{m \times n}$ , with positive entries in the range  $(0, 1)$ .
- He computed  $C = AB$  with an implementation that was known to be “correct” and assumed it yields the exact solution. (Of course, it has error in it as well. We discuss how he compensated for that, below.)
- He computed  $\check{C} = AB$  with his new implementation.
- He computed  $\Delta C = \check{C} - C$  and checked that each of its entries satisfied  $\delta \gamma_{i,j} \leq 2k u \gamma_{i,j}$ .

- In the above, he took advantage of the fact that  $A$  and  $B$  had positive entries so that  $|A||B| = AB = C$ . He also approximated  $\gamma_k = \frac{k\mathbf{u}}{1-k\mathbf{u}}$  with  $k\mathbf{u}$ , and introduced the factor 2 to compensate for the fact that  $C$  itself was inexactly computed.
-

## Notes on Performance

How to attain high performance on modern architectures is of importance: linear algebra is fundamental to scientific computing. Scientific computing often involves very large problems that require the fastest computers in the world to be employed. One wants to use such computers efficiently.

For now, we suggest the reader become familiar with the following resources:

- Week 5 of [Linear Algebra: Foundations to Frontiers - Notes to LAFF With \[29\]](#).  
Focus on Section 5.4 Enrichment.
- Kazushige Goto and Robert van de Geijn.  
Anatomy of high-performance matrix multiplication [22].  
ACM Transactions on Mathematical Software, 34 (3), 2008.
- Field G. Van Zee and Robert van de Geijn.  
BLIS: A Framework for Rapid Instantiation of BLAS Functionality [41].  
ACM Transactions on Mathematical Software, to appear.
- Robert van de Geijn.  
How to Optimize Gemm.  
[wiki.cs.utexas.edu/rvdg/HowToOptimizeGemm](http://wiki.cs.utexas.edu/rvdg/HowToOptimizeGemm).  
(An exercise on how to write a high-performance matrix-matrix multiplication in C.)

A similar exercise:

Michael Lehn

GEMM: From Pure C to SSE Optimized Micro Kernels

<http://apfel.mathematik.uni-ulm.de/lehn/sghpc/gemm/index.html>.



## Notes on Gaussian Elimination and LU Factorization

The LU factorization is also known as the LU decomposition and the operations it performs are equivalent to those performed by Gaussian elimination. For details, we recommend that the reader consult Weeks 6 and 7 of

“[Linear Algebra: Foundations to Frontiers - Notes to LAFF With](#)” [29].

### Video

Read disclaimer regarding the videos in the preface!

Lecture on Gaussian Elimination and LU factorization:

- [YouTube](#)
- [Download from UT Box](#)
- [View After Local Download](#)

Lecture on deriving dense linear algebra algorithms:

- [YouTube](#)
- [Download from UT Box](#)
- [View After Local Download](#)
- [Slides](#)

(For help on viewing, see [Appendix A](#).)

## Outline

<b>Video</b> . . . . .	<b>159</b>
<b>Outline</b> . . . . .	<b>160</b>
<b>11.1. Definition and Existence</b> . . . . .	<b>161</b>
<b>11.2. LU Factorization</b> . . . . .	<b>161</b>
11.2.1. First derivation . . . . .	161
11.2.2. Gauss transforms . . . . .	162
11.2.3. Cost of LU factorization . . . . .	164
<b>11.3. LU Factorization with Partial Pivoting</b> . . . . .	<b>165</b>
11.3.1. Permutation matrices . . . . .	165
11.3.2. The algorithm . . . . .	167
<b>11.4. Proof of Theorem 11.3</b> . . . . .	<b>173</b>
<b>11.5. LU with Complete Pivoting</b> . . . . .	<b>174</b>
<b>11.6. Solving <math>Ax = y</math> Via the LU Factorization with Pivoting</b> . . . . .	<b>175</b>
<b>11.7. Solving Triangular Systems of Equations</b> . . . . .	<b>175</b>
11.7.1. $Lz = y$ . . . . .	175
11.7.2. $Ux = z$ . . . . .	178
<b>11.8. Other LU Factorization Algorithms</b> . . . . .	<b>178</b>
11.8.1. Variant 1: Bordered algorithm . . . . .	183
11.8.2. Variant 2: Left-looking algorithm . . . . .	183
11.8.3. Variant 3: Up-looking variant . . . . .	184
11.8.4. Variant 4: Crout variant . . . . .	184
11.8.5. Variant 5: Classical LU factorization . . . . .	185
11.8.6. All algorithms . . . . .	185
11.8.7. Formal derivation of algorithms . . . . .	185
<b>11.9. Numerical Stability Results</b> . . . . .	<b>187</b>
<b>11.10 Is LU with Partial Pivoting Stable?</b> . . . . .	<b>188</b>
<b>11.11 Blocked Algorithms</b> . . . . .	<b>188</b>
11.11.1. Blocked classical LU factorization (Variant 5) . . . . .	188
11.11.2. Blocked classical LU factorization with pivoting (Variant 5) . . . . .	191
<b>11.12 Variations on a Triple-Nested Loop</b> . . . . .	<b>192</b>



## 11.1 Definition and Existence

**Definition 11.1 LU factorization (decomposition)** Given a matrix  $A \in \mathbb{C}^{m \times n}$  with  $m \leq n$  its LU factorization is given by  $A = LU$  where  $L \in \mathbb{C}^{m \times n}$  is unit lower trapezoidal and  $U \in \mathbb{C}^{n \times n}$  is upper triangular.

The first question we will ask is when the LU factorization exists. For this, we need a definition.

**Definition 11.2** The  $k \times k$  principle leading submatrix of a matrix  $A$  is defined to be the square matrix  $A_{TL} \in \mathbb{C}^{k \times k}$  such that  $A = \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)$ .

This definition allows us to indicate when a matrix has an LU factorization:

**Theorem 11.3 Existence** Let  $A \in \mathbb{C}^{m \times n}$  and  $m \leq n$  have linearly independent columns. Then  $A$  has a unique LU factorization if and only if all its principle leading submatrices are nonsingular.

The proof of this theorem is a bit involved and can be found in Section 11.4.

## 11.2 LU Factorization

We are going to present two different ways of deriving the most commonly known algorithm. The first is a straight forward derivation. The second presents the operation as the application of a sequence of Gauss transforms.

### 11.2.1 First derivation

Partition  $A$ ,  $L$ , and  $U$  as follows:

$$A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right), \quad L \rightarrow \left( \begin{array}{c|c} 1 & 0 \\ \hline l_{21} & L_{22} \end{array} \right), \quad \text{and} \quad U \rightarrow \left( \begin{array}{c|c} u_{11} & u_{12}^T \\ \hline 0 & U_{22} \end{array} \right).$$

Then  $A = LU$  means that

$$\left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c} 1 & 0 \\ \hline l_{21} & L_{22} \end{array} \right) \left( \begin{array}{c|c} u_{11} & u_{12}^T \\ \hline 0 & U_{22} \end{array} \right) = \left( \begin{array}{c|c} u_{11} & u_{12}^T \\ \hline l_{21}u_{11} & l_{21}u_{12}^T + L_{22}U_{22} \end{array} \right).$$

This means that

$$\begin{array}{c|c} \alpha_{11} = u_{11} & a_{12}^T = u_{12}^T \\ \hline a_{21} = u_{11}l_{21} & A_{22} = l_{21}u_{12}^T + L_{22}U_{22} \end{array}$$

or, equivalently,

$$\begin{array}{c|c} \alpha_{11} = u_{11} & a_{12}^T = u_{12}^T \\ \hline a_{21} = u_{11}l_{21} & A_{22} - l_{21}u_{12}^T = L_{22}U_{22} \end{array}.$$

If we let  $U$  overwrite the original matrix  $A$  this suggests the algorithm

- $l_{21} = a_{21}/\alpha_{11}$ .
- $a_{21} = 0$ .
- $A_{22} := A_{22} - l_{21}a_{12}^T$ .
- Continue by overwriting the updated  $A_{22}$  with its LU factorization.

This is captured in the algorithm in Figure 11.1.

### 11.2.2 Gauss transforms

**Definition 11.4** A matrix  $L_k$  of the form  $L_k = \left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & 0 \end{array} \right)$  where  $I_k$  is  $k \times k$  is called a Gauss transform.

**Example 11.5** Gauss transforms can be used to take multiples of a row and subtract these multiples from other rows:

$$\left( \begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & -\lambda_{21} & 1 & 0 \\ \hline 0 & -\lambda_{31} & 0 & 1 \end{array} \right) \left( \begin{array}{c} \hat{a}_0^T \\ \hat{a}_1^T \\ \hat{a}_2^T \\ \hat{a}_3^T \end{array} \right) = \left( \begin{array}{c} \hat{a}_0^T \\ \hline \hat{a}_1^T \\ \hline \left( \begin{array}{c} \hat{a}_2^T \\ \hat{a}_3^T \end{array} \right) - \left( \begin{array}{c} \lambda_{21} \\ \lambda_{31} \end{array} \right) \hat{a}_1^T \end{array} \right) = \left( \begin{array}{c} \hat{a}_0^T \\ \hline \hat{a}_1^T \\ \hline \hat{a}_2^T - \lambda_{21}\hat{a}_1^T \\ \hat{a}_3^T - \lambda_{31}\hat{a}_1^T \end{array} \right).$$

Notice the similarity with what one does in Gaussian Elimination: take a multiples of one row and subtract these from other rows.

Now assume that the LU factorization in the previous subsection has proceeded to where  $A$  contains

$$\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right)$$

where  $A_{00}$  is upper triangular (recall: it is being overwritten by  $U$ !). What we would like to do is eliminate the elements in  $a_{21}$  by taking multiples of the “current row”  $\left( \alpha_{11} \mid a_{12}^T \right)$  and subtract these from the rest of the rows:  $\left( a_{21} \mid A_{22} \right)$ . The vehicle is a Gauss transform: we must determine  $l_{21}$  so that

$$\left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right).$$

**Algorithm:** Compute LU factorization of  $A$ , overwriting  $L$  with factor  $L$  and  $A$  with factor  $U$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), L \rightarrow \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right)$

**where**  $A_{TL}$  and  $L_{TL}$  are  $0 \times 0$

**while**  $n(A_{TL}) < n(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & \lambda_{11} & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right)$$

**where**  $\alpha_{11}, \lambda_{11}$  are  $1 \times 1$

$$\left\{ \begin{array}{l} l_{21} := a_{21}/\alpha_{11} \\ A_{22} := A_{22} - l_{21}a_{12}^T \\ (a_{21} := 0) \end{array} \right.$$

or, alternatively,

$$\left\{ \begin{array}{l} l_{21} := a_{21}/\alpha_{11} \\ \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & 0 \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) \\ \\ = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right) \end{array} \right.$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & \lambda_{11} & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right)$$

**endwhile**

Figure 11.1: Most commonly known algorithm for overwriting a matrix with its LU factorization.

This means we must pick  $l_{21} = a_{21}/\alpha_{11}$  since

$$\left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} - \alpha_{11}l_{21} & A_{22} - l_{21}a_{12}^T \end{array} \right).$$

The resulting algorithm is summarized in Figure 11.1 under “or, alternatively,”. Notice that this algorithm is identical to the algorithm for computing LU factorization discussed before!

How can this be? The following set of exercises explains it.

**Homework 11.6** Show that

$$\left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right)^{-1} = \left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right)$$

SEE ANSWER

Now, clearly, what the algorithm does is to compute a sequence of  $n$  Gauss transforms  $\hat{L}_0, \dots, \hat{L}_{n-1}$  such that  $\hat{L}_{n-1}\hat{L}_{n-2}\cdots\hat{L}_1\hat{L}_0A = U$ . Or, equivalently,  $A = L_0L_1\cdots L_{n-2}L_{n-1}U$ , where  $L_k = \hat{L}_k^{-1}$ . What will show next is that  $L = L_0L_1\cdots L_{n-2}L_{n-1}$  is the unit lower triangular matrix computed by the LU factorization.

**Homework 11.7** Let  $\tilde{L}_k = L_0L_1\cdots L_k$ . Assume that  $\tilde{L}_k$  has the form  $\tilde{L}_{k-1} = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 1 & 0 \\ \hline L_{20} & 0 & I \end{array} \right)$ , where  $\tilde{L}_{00}$

is  $k \times k$ . Show that  $\tilde{L}_k$  is given by  $\tilde{L}_k = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 1 & 0 \\ \hline L_{20} & l_{21} & I \end{array} \right)$  .. (Recall:  $\hat{L}_k = \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right)$ .)

SEE ANSWER

What this exercise shows is that  $L = L_0L_1\cdots L_{n-2}L_{n-1}$  is the triangular matrix that is created by simply placing the computed vectors  $l_{21}$  below the diagonal of a unit lower triangular matrix.

### 11.2.3 Cost of LU factorization

The cost of the LU factorization algorithm given in Figure 11.1 can be analyzed as follows:

- Assume  $A$  is  $n \times n$ .
- During the  $k$ th iteration,  $A_{TL}$  is initially  $k \times k$ .

- Computing  $l_{21} := a_{21}/\alpha_{11}$  is typically implemented as  $\beta := 1/\alpha_{11}$  and then the scaling  $l_{21} := \beta a_{21}$ . The reason is that divisions are expensive relative to multiplications. We will ignore the cost of the division (which will be insignificant if  $n$  is large). Thus, we count this as  $n - k - 1$  multiplies.
- The rank-1 update of  $A_{22}$  requires  $(n - k - 1)^2$  multiplications and  $(n - k - 1)^2$  additions.
- Thus, the total cost (in flops) can be approximated by

$$\begin{aligned}
 \sum_{k=0}^{n-1} [(n - k - 1) + 2(n - k - 1)^2] &= \sum_{j=0}^{n-1} [j + 2j^2] \quad (\text{Change of variable: } j = n - k - 1) \\
 &= \sum_{j=0}^{n-1} j + 2 \sum_{j=0}^{n-1} j^2 \\
 &\approx \frac{n(n-1)}{2} + 2 \int_0^n x^2 dx \\
 &= \frac{n(n-1)}{2} + \frac{2}{3}n^3 \\
 &\approx \frac{2}{3}n^3
 \end{aligned}$$

Notice that this involves roughly half the number of floating point operations as are required for a Householder transformation based QR factorization.

## 11.3 LU Factorization with Partial Pivoting

It is well-known that the LU factorization is numerically unstable under general circumstances. In particular, a backward stability analysis, given for example in [3, 8, 6] and summarized in Section 11.9, shows that the computed matrices  $\check{L}$  and  $\check{U}$  satisfy

$$(A + \Delta A) = \check{L}\check{U} \quad \text{where } |\Delta A| \leq \gamma_n |\check{L}||\check{U}|.$$

(This is the backward error result for the Crout variant for LU factorization, discussed later in this note. Some of the other variants have an error result of  $(A + \Delta A) = \check{L}\check{U}$  where  $|\Delta A| \leq \gamma_n(|A| + |\check{L}||\check{U}|)$ .) Now, if  $\alpha$  is small in magnitude compared to the entries of  $a_{21}$  then not only will  $l_{21}$  have large entries, but the update  $A_{22} - l_{21}a_{12}^T$  will potentially introduce large entries in the updated  $A_{22}$  (in other words, the part of matrix  $A$  from which the future matrix  $U$  will be computed), a phenomenon referred to as *element growth*. To overcome this, we take will swap rows in  $A$  as the factorization proceeds, resulting in an algorithm known as LU factorization with partial pivoting.

### 11.3.1 Permutation matrices

**Definition 11.8** An  $n \times n$  matrix  $P$  is said to be a permutation matrix, or permutation, if, when applied to a vector  $x = (\chi_0, \chi_1, \dots, \chi_{n-1})^T$ , it merely rearranges the order of the elements in that vector. Such a permutation can be represented by the vector of integers,  $(\pi_0, \pi_1, \dots, \pi_{n-1})^T$ , where  $\{\pi_0, \pi_1, \dots, \pi_{n-1}\}$  is a permutation of the integers  $\{0, 1, \dots, n-1\}$  and the permuted vector  $Px$  is given by  $(\chi_{\pi_0}, \chi_{\pi_1}, \dots, \chi_{\pi_{n-1}})^T$ .

**Algorithm:** Compute LU factorization with partial pivoting of  $A$ , overwriting  $L$  with factor  $L$  and  $A$  with factor  $U$ . The pivot vector is returned in  $p$ .

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right),$

$L \rightarrow \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right), p \rightarrow \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right).$

**where**  $A_{TL}$  and  $L_{TL}$  are  $0 \times 0$  and  $p_T$  is  $0 \times 1$

**while**  $n(A_{TL}) < n(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & \lambda_{11} & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \rightarrow \left( \begin{array}{c} p_0 \\ \hline \pi_1 \\ \hline p_2 \end{array} \right)$$

**where**  $\alpha_{11}, \lambda_{11}, \pi_1$  are  $1 \times 1$

$$\pi_1 = \max_i \left( \frac{\alpha_{11}}{a_{21}} \right)$$

$$\left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) := P(\pi_1) \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right)$$

$$l_{21} := a_{21} / \alpha_{11}$$

$$A_{22} := A_{22} - l_{21} a_{12}^T$$

$$(a_{21} := 0)$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & \lambda_{11} & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \leftarrow \left( \begin{array}{c} p_0 \\ \hline \pi_1 \\ \hline p_2 \end{array} \right)$$

**endwhile**

Figure 11.2: LU factorization with partial pivoting.

If  $P$  is a permutation matrix then  $PA$  rearranges the rows of  $A$  exactly as the elements of  $x$  are rearranged by  $Px$ .

We will see that when discussing the LU factorization with partial pivoting, a permutation matrix that swaps the first element of a vector with the  $\pi$ -th element of that vector is a fundamental tool. We will denote that matrix by

$$P(\pi) = \begin{cases} I_n & \text{if } \pi = 0 \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & I_{\pi-1} & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{n-\pi-1} \end{pmatrix} & \text{otherwise,} \end{cases}$$

where  $n$  is the dimension of the permutation matrix. In the following we will use the notation  $P_n$  to indicate that the matrix  $P$  is of size  $n$ . Let  $p$  be a vector of integers satisfying the conditions

$$p = (\pi_0, \dots, \pi_{k-1})^T, \text{ where } 1 \leq k \leq n \text{ and } 0 \leq \pi_i < n-i, \quad (11.1)$$

then  $P_n(p)$  will denote the permutation:

$$P_n(p) = \begin{pmatrix} I_{k-1} & 0 \\ 0 & P_{n-k+1}(\pi_{k-1}) \end{pmatrix} \begin{pmatrix} I_{k-2} & 0 \\ 0 & P_{n-k+2}(\pi_{k-2}) \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}(\pi_1) \end{pmatrix} P_n(\pi_0).$$

**Remark 11.9** In the algorithms, the subscript that indicates the matrix dimensions is omitted.

**Example 11.10** Let  $a_0^T, a_1^T, \dots, a_{n-1}^T$  be the rows of a matrix  $A$ . The application of  $P(p)$  to  $A$  yields a matrix that results from swapping row  $a_0^T$  with  $a_{\pi_0}^T$ , then swapping  $a_1^T$  with  $a_{\pi_1+1}^T$ ,  $a_2^T$  with  $a_{\pi_2+2}^T$ , until finally  $a_{k-1}^T$  is swapped with  $a_{\pi_{k-1}+k-1}^T$ .

**Remark 11.11** For those familiar with how pivot information is stored in LINPACK and LAPACK, notice that those packages store the vector of pivot information  $(\pi_0 + 1, \pi_1 + 2, \dots, \pi_{k-1} + k)^T$ .

### 11.3.2 The algorithm

Having introduced our notation for permutation matrices, we can now define the LU factorization with partial pivoting: Given an  $n \times n$  matrix  $A$ , we wish to compute a) a vector  $p$  of  $n$  integers which satisfies the conditions (11.1), b) a unit lower trapezoidal matrix  $L$ , and c) an upper triangular matrix  $U$  so that  $P(p)A = LU$ . An algorithm for computing this operation is typically represented by

$$[A, p] := \text{LU piv } A,$$

where upon completion  $A$  has been overwritten by  $\{L \setminus U\}$ .

Let us start with revisiting the first derivation of the LU factorization. The first step is to find a first permutation matrix  $P(\pi_1)$  such that the element on the diagonal in the first column is maximal in value. For this, we will introduce the function  $\text{maxi}(x)$  which, given a vector  $x$ , returns the index of the element in  $x$  with maximal magnitude (absolute value). The algorithm then proceeds as follows:

- Partition  $A$ ,  $L$  as follows:

$$A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right), \quad \text{and} \quad L \rightarrow \left( \begin{array}{c|c} 1 & 0 \\ \hline l_{21} & L_{22} \end{array} \right).$$

- Compute  $\pi_1 = \max_i \left( \frac{\alpha_{11}}{a_{21}} \right)$ .
- Permute the rows:  $\left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) := P(\pi_1) \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right)$ .
- Compute  $l_{21} := a_{21}/\alpha_{11}$ .
- Update  $A_{22} := A_{22} - l_{21}a_{12}^T$ .

Now, in general, assume that the computation has proceeded to the point where matrix  $A$  has been overwritten by

$$\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right)$$

where  $A_{00}$  is upper triangular. If no pivoting was added one would compute  $l_{21} := a_{21}/\alpha_{11}$  followed by the update

$$\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right).$$

Now, instead one performs the steps

- Compute  $\pi_1 = \max_i \left( \frac{\alpha_{11}}{a_{21}} \right)$ .
- Permute the rows:  $\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right)$
- Compute  $l_{21} := a_{21}/\alpha_{11}$ .



- Update

$$\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right).$$

This algorithm is summarized in Figure 11.2.

Now, what this algorithm computes is a sequence of Gauss transforms  $\hat{L}_0, \dots, \hat{L}_{n-1}$  and permutations  $P_0, \dots, P_{n-1}$  such that

$$\hat{L}_{n-1}P_{n-1} \cdots \hat{L}_0P_0A = U$$

or, equivalently,

$$A = P_0^T L_0 \cdots \hat{P}_{n-1}^T L_{n-1} U,$$

where  $L_k = \hat{L}_k^{-1}$ . What we will finally show is that there are Gauss transforms  $\bar{L}_0, \dots, \bar{L}_{n-1}$  (here the “bar” does NOT mean conjugation. It is just a symbol) such that

$$A = P_0^T \cdots P_{n-1}^T \underbrace{\bar{L}_0 \cdots \bar{L}_{n-1}}_L U$$

or, equivalently,

$$P(p)A = P_{n-1} \cdots P_0 A = \underbrace{\bar{L}_0 \cdots \bar{L}_{n-1}}_L U,$$

which is what we set out to compute.

Here is the insight. Assume that after  $k$  steps of LU factorization we have computed  $p_T, L_{TL}, L_{BL}$ , etc. so that

$$P(p_T)A = \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & I \end{array} \right) \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline 0 & A_{BR} \end{array} \right),$$

where  $A_{TL}$  is upper triangular and  $k \times k$ .

Now compute the next step of LU factorization with partial pivoting with  $\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline 0 & A_{BR} \end{array} \right)$ :

- Partition

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline 0 & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{01} & A_{02} \end{array} \right)$$

- Compute  $\pi_1 = \max_i \left( \frac{\alpha_{11}}{a_{21}} \right)$

**Algorithm:** Compute LU factorization with partial pivoting of  $A$ , overwriting  $L$  with factor  $L$  and  $A$  with factor  $U$ . The pivot vector is returned in  $p$ .

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), L \rightarrow \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right), p \rightarrow \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right).$

**where**  $A_{TL}$  and  $L_{TL}$  are  $0 \times 0$  and  $p_T$  is  $0 \times 1$

**while**  $n(A_{TL}) < n(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & \lambda_{11} & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \rightarrow \left( \begin{array}{c} p_0 \\ \hline \pi_1 \\ \hline p_2 \end{array} \right)$$

**where**  $\alpha_{11}, \lambda_{11}, \pi_1$  are  $1 \times 1$

$$\pi_1 = \max_i \left( \frac{\alpha_{11}}{a_{21}} \right)$$

$$\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline l_{10}^T & \alpha_{11} & a_{12}^T \\ \hline L_{20} & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline l_{10}^T & \alpha_{11} & a_{12}^T \\ \hline L_{20} & a_{21} & A_{22} \end{array} \right)$$

$$l_{21} := a_{21} / \alpha_{11}$$

$$\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & 0 \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right)$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & \lambda_{11} & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \leftarrow \left( \begin{array}{c} p_0 \\ \hline \pi_1 \\ \hline p_2 \end{array} \right)$$

**endwhile**

Figure 11.3: LU factorization with partial pivoting.

**Algorithm:** Compute LU factorization with partial pivoting of  $A$ , overwriting  $A$  with factors  $L$  and  $U$ . The pivot vector is returned in  $p$ .

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), p \rightarrow \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right).$

**where**  $A_{TL}$  is  $0 \times 0$  and  $p_T$  is  $0 \times 1$

**while**  $n(A_{TL}) < n(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \rightarrow \left( \begin{array}{c} p_0 \\ \hline \pi_1 \\ \hline p_2 \end{array} \right)$$

**where**  $\alpha_{11}, \lambda_{11}, \pi_1$  are  $1 \times 1$

$$\pi_1 = \max_i \left( \frac{\alpha_{11}}{a_{21}} \right)$$

$$\left( \begin{array}{c|c|c} a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right) := P(\pi_1) \left( \begin{array}{c|c|c} a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$

$$a_{21} := a_{21}/\alpha_{11}$$

$$A_{22} := A_{22} - a_{21}a_{12}^T$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \leftarrow \left( \begin{array}{c} p_0 \\ \hline \pi_1 \\ \hline p_2 \end{array} \right)$$

**endwhile**

Figure 11.4: LU factorization with partial pivoting, overwriting  $A$  with the factors.

- Permute  $\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right)$
- Compute  $l_{21} := a_{21}/\alpha_{11}$ .

- Update

$$\left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right)$$

After this,

$$P(p_T)A = \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & I \end{array} \right) \left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right) \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right) \quad (11.2)$$

But

$$\begin{aligned} & \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & I \end{array} \right) \left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right) \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right) \\ &= \left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right) \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline P(\pi_1)L_{BL} & I \end{array} \right) \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right) \\ &= \left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right) \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline \bar{L}_{BL} & I \end{array} \right) \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right), \end{aligned}$$

**NOTE:** There is a “bar” above some of  $L$ ’s and  $l$ ’s. Very hard to see!!! For this reason, these show up in red as well. Here we use the fact that  $P(\pi_1) = P(\pi_1)^T$  because of its very special structure.

Bringing the permutation to the left of (11.2) and “repartitioning” we get

$$\underbrace{\left( \begin{array}{c|c} I & 0 \\ \hline 0 & P(\pi_1) \end{array} \right)}_{P \begin{pmatrix} p_0 \\ \pi_1 \end{pmatrix}} P(p_0) A = \underbrace{\left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \bar{l}_{10}^T & 1 & 0 \\ \hline \bar{L}_{20} & 0 & I \end{array} \right)}_{\left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \bar{l}_{10}^T & 1 & 0 \\ \hline \bar{L}_{20} & l_{21} & I \end{array} \right)} \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline 0 & \alpha_{11} & a_{12}^T \\ \hline 0 & 0 & A_{22} - l_{21}a_{12}^T \end{array} \right).$$

This explains how the algorithm in Figure 11.3 compute  $p$ ,  $L$ , and  $U$  (overwriting  $A$  with  $U$ ) so that  $P(p)A = LU$ .

Finally, we recognize that  $L$  can overwrite the entries of  $A$  below its diagonal, yielding the algorithm in Figure 11.4.

## 11.4 Proof of Theorem 11.3

### Proof:

( $\Rightarrow$ ) Let nonsingular  $A$  have a (unique) LU factorization. We will show that its principle leading submatrices are nonsingular. Let

$$\underbrace{\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)}_A = \underbrace{\left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right)}_L \underbrace{\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline 0 & U_{BR} \end{array} \right)}_U$$

be the LU factorization of  $A$  where  $A_{TL}$ ,  $L_{TL}$ , and  $U_{TL}$  are  $k \times k$ . Notice that  $U$  cannot have a zero on the diagonal since then  $A$  would not have linearly independent columns. Now, the  $k \times k$  principle leading submatrix  $A_{TL}$  equals  $A_{TL} = L_{TL}U_{TL}$  which is nonsingular since  $L_{TL}$  has a unit diagonal and  $U_{TL}$  has no zeroes on the diagonal. Since  $k$  was chosen arbitrarily, this means that all principle leading submatrices are nonsingular.

( $\Leftarrow$ ) We will do a proof by induction on  $n$ .

**Base Case:**  $n = 1$ . Then  $A$  has the form  $A = \begin{pmatrix} \alpha_{11} \\ a_{21} \end{pmatrix}$  where  $\alpha_{11}$  is a scalar. Since the principle leading submatrices are nonsingular  $\alpha_{11} \neq 0$ . Hence  $A = \underbrace{\begin{pmatrix} 1 \\ a_{21}/\alpha_{11} \end{pmatrix}}_L \underbrace{\alpha_{11}}_U$  is the LU factorization of

$A$ . This LU factorization is unique because the first element of  $L$  must be 1.

**Inductive Step:** Assume the result is true for all matrices with  $n = k$ . Show it is true for matrices with  $n = k + 1$ .

Let  $A$  of size  $n = k + 1$  have nonsingular principle leading submatrices. Now, if an LU factorization of  $A$  exists,  $A = LU$ , then it would have to form

$$\underbrace{\left( \begin{array}{c|c} A_{00} & a_{01} \\ \hline a_{10}^T & \alpha_{11} \\ \hline A_{20} & a_{21} \end{array} \right)}_A = \underbrace{\left( \begin{array}{c|c} L_{00} & 0 \\ \hline l_{10}^T & 1 \\ \hline L_{20} & l_{21} \end{array} \right)}_L \underbrace{\left( \begin{array}{c|c} U_{00} & u_{01} \\ \hline 0 & \mathfrak{u}_{11} \end{array} \right)}_U. \quad (11.3)$$

If we can show that the different parts of  $L$  and  $U$  exist and are unique, we are done. Equation (11.3) can be rewritten as

$$\begin{pmatrix} A_{00} \\ a_{10}^T \\ A_{20} \end{pmatrix} = \begin{pmatrix} L_{00} \\ l_{10}^T \\ L_{20} \end{pmatrix} U_{00} \quad \text{and} \quad \begin{pmatrix} a_{01} \\ \alpha_{11} \\ a_{21} \end{pmatrix} = \begin{pmatrix} L_{00}u_{01} \\ l_{10}^T u_{01} + \mathfrak{u}_{11} \\ L_{20}u_{01} + l_{21}\mathfrak{u}_{11} \end{pmatrix}.$$

Now, by the Induction Hypothesis  $L_{11}$ ,  $l_{10}^T$ , and  $L_{20}$  exist and are unique. So the question is whether  $u_{01}$ ,  $v_{11}$ , and  $l_{21}$  exist and are unique:

- **$u_{01}$  exists and is unique.** Since  $L_{00}$  is nonsingular (it has ones on its diagonal)  $L_{00}u_{01} = a_{01}$  has a solution that is unique.
- **$v_{11}$  exists, is unique, and is nonzero.** Since  $l_{10}^T$  and  $u_{01}$  exist and are unique,  $v_{11} = \alpha_{11} - l_{10}^T u_{01}$  exists and is unique. It is also nonzero since the principle leading submatrix of  $A$  given by

$$\left( \begin{array}{c|c} A_{00} & a_{01} \\ \hline a_{10}^T & \alpha_{11} \end{array} \right) = \left( \begin{array}{c|c} L_{00} & 0 \\ \hline l_{10}^T & 1 \end{array} \right) \left( \begin{array}{c|c} U_{00} & u_{01} \\ \hline 0 & v_{11} \end{array} \right),$$

is nonsingular by assumption and therefore  $v_{11}$  must be nonzero.

- **$l_{21}$  exists and is unique.** Since  $v_{11}$  exists and is nonzero,  $l_{21} = a_{21}/v_{11}$  exists and is uniquely determined.

Thus the  $m \times (k+1)$  matrix  $A$  has a unique LU factorization.

**By the Principle of Mathematical Induction** the result holds.

**Homework 11.12** Implement LU factorization with partial pivoting with the FLAME@lab API, in M-script.

🔗 [SEE ANSWER](#)

## 11.5 LU with Complete Pivoting

LU factorization with partial pivoting builds on the insight that pivoting (rearranging) rows in a linear system does not change the solution: if  $Ax = b$  then  $P(p)Ax = P(p)b$ , where  $p$  is a pivot vector. Now, if  $r$  is another pivot vector, then notice that  $P(r)^T P(r) = I$  (a simple property of pivot matrices) and  $AP(r)^T$  permutes the columns of  $A$  in exactly the same order as  $P(r)A$  permutes the rows of  $A$ .

What this means is that if  $Ax = b$  then  $P(p)AP(r)^T[P(r)x] = P(p)b$ . This supports the idea that one might want to not only permute rows of  $A$ , as in partial pivoting, but also columns of  $A$ . This is done in a variation on LU factorization that is known as LU factorization with *complete pivoting*.

The idea is as follows: Given matrix  $A$ , partition

$$A = \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right).$$

Now, instead of finding the largest element in magnitude in the first column, find the largest element in magnitude in the entire matrix. Let's say it is element  $(\pi_0, \rho_0)$ . Then, one permutes

$$\left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) := P(\pi_0) \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) P(\rho_0)^T,$$

making  $\alpha_{11}$  the largest element in magnitude. This then reduces the magnitude of multipliers and element growth.

It can be shown that the maximal element growth experienced when employing LU with complete pivoting indeed reduces element growth. The problem is that it requires  $O(n^2)$  comparisons per iteration. Worse, it completely destroys the ability to utilize blocked algorithms, which attain much greater performance.

In practice LU with complete pivoting is not used.

## 11.6 Solving $Ax = y$ Via the LU Factorization with Pivoting

Given nonsingular matrix  $A \in \mathbb{C}^{m \times m}$ , the above discussions have yielded an algorithm for computing permutation matrix  $P$ , unit lower triangular matrix  $L$  and upper triangular matrix  $U$  such that  $PA = LU$ . We now discuss how these can be used to solve the system of linear equations  $Ax = y$ .

Starting with

$$Ax = y$$

we multiply both sides of the equation by permutation matrix  $P$

$$PAx = \underbrace{Py}_{\hat{y}}$$

and substitute  $LU$  for  $PA$

$$L \underbrace{Ux}_z \hat{y}.$$

We now notice that we can solve the lower triangular system

$$Lz = \hat{y}$$

after which  $x$  can be computed by solving the upper triangular system

$$Ux = z.$$

## 11.7 Solving Triangular Systems of Equations

### 11.7.1 $Lz = y$

First, we discuss solving  $Lz = y$  where  $L$  is a unit lower triangular matrix.

#### Variant 1

Consider  $Lz = y$  where  $L$  is unit lower triangular. Partition

$$L \rightarrow \left( \begin{array}{c|c} 1 & 0 \\ \hline l_{21} & L_{22} \end{array} \right), \quad z \rightarrow \begin{pmatrix} \zeta_1 \\ z_2 \end{pmatrix} \quad \text{and} \quad y \rightarrow \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix}.$$

**Algorithm:** Solve  $Lz = y$ , overwriting  $y$  (Var. 1)

**Partition**  $L \rightarrow \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right), y \rightarrow \left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right)$

**where**  $L_{TL}$  is  $0 \times 0$ ,  $y_T$  has 0 rows

**while**  $m(L_{TL}) < m(L)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} L_{00} & l_{01} & L_{02} \\ \hline l_{10}^T & \lambda_{11} & l_{12}^T \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \rightarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

---


$$y_2 := y_2 - \psi_1 l_{21}$$


---

**Continue with**

$$\left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} L_{00} & l_{01} & L_{02} \\ \hline l_{10}^T & \lambda_{11} & l_{12}^T \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \leftarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

**endwhile**

**Algorithm:** Solve  $Lz = y$ , overwriting  $y$  (Var. 2)

**Partition**  $L \rightarrow \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right), y \rightarrow \left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right)$

**where**  $L_{TL}$  is  $0 \times 0$ ,  $y_T$  has 0 rows

**while**  $m(L_{TL}) < m(L)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} L_{00} & l_{01} & L_{02} \\ \hline l_{10}^T & \lambda_{11} & l_{12}^T \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \rightarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

---


$$\psi_1 := \psi_1 - l_{10}^T y_0$$


---

**Continue with**

$$\left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} L_{00} & l_{01} & L_{02} \\ \hline l_{10}^T & \lambda_{11} & l_{12}^T \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \leftarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

**endwhile**

Figure 11.5: Algorithms for the solution of a unit lower triangular system  $Lz = y$  that overwrite  $y$  with  $z$ .



Then

$$\underbrace{\left( \begin{array}{c|c} 1 & 0 \\ l_{21} & L_{22} \end{array} \right)}_L \underbrace{\left( \begin{array}{c} \zeta_1 \\ z_2 \end{array} \right)}_z = \underbrace{\left( \begin{array}{c} \psi_1 \\ y_2 \end{array} \right)}_y.$$

Multiplying out the left-hand side yields

$$\left( \begin{array}{c} \zeta_1 \\ \zeta_1 l_{21} + L_{22} z_2 \end{array} \right) = \left( \begin{array}{c} \psi_1 \\ y_2 \end{array} \right)$$

and the equalities

$$\begin{aligned} \zeta_1 &= \psi_1 \\ \zeta_1 l_{21} + L_{22} z_2 &= y_2, \end{aligned}$$

which can be rearranged as

$$\begin{aligned} \zeta_1 &= \psi_1 \\ L_{22} z_2 &= y_2 - \zeta_1 l_{21}. \end{aligned}$$

These insights justify the algorithm in Figure 11.5 (left), which overwrites  $y$  with the solution to  $Lz = y$ .

### Variant 2

An alternative algorithm can be derived as follows: Partition

$$L \rightarrow \left( \begin{array}{c|c} L_{00} & 0 \\ l_{10}^T & 1 \end{array} \right), \quad z \rightarrow \left( \begin{array}{c} z_0 \\ \zeta_1 \end{array} \right) \quad \text{and} \quad y \rightarrow \left( \begin{array}{c} y_0 \\ \psi_1 \end{array} \right).$$

Then

$$\underbrace{\left( \begin{array}{c|c} L_{00} & 0 \\ l_{10}^T & 1 \end{array} \right)}_L \underbrace{\left( \begin{array}{c} z_0 \\ \zeta_1 \end{array} \right)}_z = \underbrace{\left( \begin{array}{c} y_0 \\ \psi_1 \end{array} \right)}_y.$$

Multiplying out the left-hand side yields

$$\left( \begin{array}{c} L_{00} z_0 \\ l_{10}^T z_0 + \zeta_1 \end{array} \right) = \left( \begin{array}{c} y_0 \\ \psi_1 \end{array} \right)$$

and the equalities

$$\begin{aligned} L_{00} z_0 &= y_0 \\ l_{10}^T z_0 + \zeta_1 &= \psi_1. \end{aligned}$$

The idea now is as follows: Assume that the elements of  $z_0$  were computed in previous iterations in the algorithm in Figure 11.5 (left), overwriting  $y_0$ . Then in the current iteration we can compute  $\zeta_1 := \psi_0 - l_{10}^T z_0$ , overwriting  $\psi_1$ .

### Discussion

Notice that Variant 1 casts the computation in terms of an AXPY operation while Variant 2 casts it in terms of DOT products.

#### 11.7.2 $Ux = z$

Next, we discuss solving  $Ux = y$  where  $U$  is an upper triangular matrix (with no assumptions about its diagonal entries).

**Homework 11.13** *Derive an algorithm for solving  $Ux = y$ , overwriting  $y$  with the solution, that casts most computation in terms of DOT products. Hint: Partition*

$$U \rightarrow \left( \begin{array}{c|c} u_{11} & u_{12}^T \\ \hline 0 & U_{22} \end{array} \right).$$

Call this Variant 1 and use Figure 11.6 to state the algorithm.

🔗 [SEE ANSWER](#)

**Homework 11.14** *Derive an algorithm for solving  $Ux = y$ , overwriting  $y$  with the solution, that casts most computation in terms of AXPY operations. Call this Variant 2 and use Figure 11.6 to state the algorithm.*

🔗 [SEE ANSWER](#)

## 11.8 Other LU Factorization Algorithms

There are actually five different (unblocked) algorithms for computing the LU factorization that were discovered over the course of the centuries<sup>1</sup>. The LU factorization in Figure 11.1 is sometimes called *classical LU factorization* or the *right-looking* algorithm. We now briefly describe how to derive the other algorithms.

Finding the algorithms starts with the following observations.

- Our algorithms will overwrite the matrix  $A$ , and hence we introduce  $\hat{A}$  to denote the original contents of  $A$ . We will say that the *precondition* for the algorithm is that

$$A = \hat{A}$$

( $A$  starts by containing the original contents of  $A$ .)

- We wish to overwrite  $A$  with  $L$  and  $U$ . Thus, the *postcondition* for the algorithm (the state in which we wish to exit the algorithm) is that

$$A = L \setminus U \wedge LU = \hat{A}$$

( $A$  is overwritten by  $L$  below the diagonal and  $U$  on and above the diagonal, where multiplying  $L$  and  $U$  yields the original matrix  $A$ .)

---

<sup>1</sup>For a thorough discussion of the different LU factorization algorithms that also gives a historic perspective, we recommend “Matrix Algorithms Volume 1” by G.W. Stewart [34]

**Algorithm:** Solve  $Uz = y$ , overwriting  $y$  (Variant 1)**Partition**  $U \rightarrow \left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right), y \rightarrow \left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right)$ **where**  $U_{BR}$  is  $0 \times 0$ ,  $y_B$  has 0 rows**while**  $m(U_{BR}) < m(U)$  **do****Repartition**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \rightarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$


---

**Continue with**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \leftarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

**endwhile****Algorithm:** Solve  $Uz = y$ , overwriting  $y$  (Variant 2)**Partition**  $U \rightarrow \left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right), y \rightarrow \left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right)$ **where**  $U_{BR}$  is  $0 \times 0$ ,  $y_B$  has 0 rows**while**  $m(U_{BR}) < m(U)$  **do****Repartition**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \rightarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$


---

**Continue with**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ \hline y_B \end{array} \right) \leftarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ \hline y_2 \end{array} \right)$$

**endwhile**Figure 11.6: Algorithms for the solution of an upper triangular system  $Ux = y$  that overwrite  $y$  with  $x$ .

- All the algorithms will march through the matrices from top-left to bottom-right. Thus, at a representative point in the algorithm, the matrices are viewed as quadrants:

$$A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), \quad L \rightarrow \left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right), \quad \text{and} \quad U \rightarrow \left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline 0 & U_{BR} \end{array} \right).$$

where  $A_{TL}$ ,  $L_{TL}$ , and  $U_{TL}$  are all square and equally sized.

- In terms of these exposed quadrants, in the end we wish for matrix  $A$  to contain

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L \backslash U_{TL} & U_{TR} \\ \hline L_{BL} & L \backslash U_{BR} \end{array} \right)$$

where  $\left( \begin{array}{c|c} L_{TL} & 0 \\ \hline L_{BL} & L_{BR} \end{array} \right) \left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline 0 & U_{BR} \end{array} \right) = \left( \begin{array}{c|c} \hat{A}_{TL} & \hat{A}_{TR} \\ \hline \hat{A}_{BL} & \hat{A}_{BR} \end{array} \right)$

- Manipulating this yields what we call the Partitioned Matrix Expression (PME), which can be viewed as a recursive definition of the LU factorization:

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L \backslash U_{TL} & U_{TR} \\ \hline L_{BL} & L \backslash U_{BR} \end{array} \right)$$

$$\wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad L_{TL}U_{TR} = \hat{A}_{TR}}{L_{BL}U_{TL} = \hat{A}_{BL} \quad L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}}$$

Now, consider the code skeleton for the LU factorization in Figure 11.7. At the top of the loop (right after the **while**), we want to maintain certain contents in matrix  $A$ . Since we are in a loop, we haven't yet overwritten  $A$  with the final result. Instead, some progress toward this final result have been made. The way we can find what the state of  $A$  is that we would like to maintain is to take the PME and delete subexpression. For example, consider the following condition on the contents of  $A$ :

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L \backslash U_{TL} & U_{TR} \\ \hline L_{BL} & \hat{A}_{BR} - L_{BL}U_{TR} \end{array} \right)$$

$$\wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad L_{TL}U_{TR} = \hat{A}_{TR}}{L_{BL}U_{TL} = \hat{A}_{BL} \quad \cancel{L_{BL}U_{TR}} + \cancel{L_{BR}U_{BR}} = \hat{A}_{BR}}$$

What we are saying is that  $A_{TL}$ ,  $A_{TR}$ , and  $A_{BL}$  have been completely updated with the corresponding parts of  $L$  and  $U$ , and  $A_{BR}$  has been partially updated. **This is exactly the state that the algorithm that we discussed previously in this document maintains!** What is left is to factor  $A_{BR}$ , since it contains  $\hat{A}_{BR} - L_{BL}U_{TR}$ , and  $\hat{A}_{BR} - L_{BL}U_{TR} = L_{BR}U_{BR}$ .

- By carefully analyzing the order in which computation must occur (in compiler lingo: by performing a dependence analysis), we can identify five states that can be maintained at the top of the loop, by deleting subexpressions from the PME. These are called *loop invariants* and are listed in Figure 11.8.
- Key to figuring out what updates must occur in the loop for each of the variants is to look at how the matrices are repartitioned at the top and bottom of the loop body.

**Algorithm:**  $A := \text{LU}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), L \rightarrow \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right), U \rightarrow \left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$ ,  $L_{TL}$  is  $0 \times 0$ ,  $U_{TL}$  is  $0 \times 0$

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} L_{00} & l_{01} & L_{02} \\ \hline l_{10}^T & \lambda_{11} & l_{12}^T \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right),$$

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right)$$

**where**  $\alpha_{11}$  is  $1 \times 1$ ,  $\lambda_{11}$  is  $1 \times 1$ ,  $v_{11}$  is  $1 \times 1$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c|c} L_{TL} & L_{TR} \\ \hline L_{BL} & L_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} L_{00} & l_{01} & L_{02} \\ \hline l_{10}^T & \lambda_{11} & l_{12}^T \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right),$$

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right)$$

**endwhile**

Figure 11.7: Code skeleton for LU factorization.

Variant	Algorithm	State (loop invariant)
1	Bordered	$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} L \backslash U_{TL} & \hat{A}_{TR} \\ \hline \hat{A}_{BL} & \hat{A}_{BR} \end{array} \right)$ $\wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad \color{red}{L_{TL}U_{TR} = \hat{A}_{TR}}}{\color{red}{L_{BL}U_{TL} = \hat{A}_{BL}} \quad \color{red}{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}}}$
2	Left-looking	$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} L \backslash U_{TL} & \hat{A}_{TR} \\ \hline L_{BL} & \hat{A}_{BR} \end{array} \right)$ $\wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad \color{red}{L_{TL}U_{TR} = \hat{A}_{TR}}}{L_{BL}U_{TL} = \hat{A}_{BL} \quad \color{red}{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}}}$
3	Up-looking	$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} L \backslash U_{TL} & U_{TR} \\ \hline \hat{A}_{BL} & \hat{A}_{BR} \end{array} \right)$ $\wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad L_{TL}U_{TR} = \hat{A}_{TR}}{\color{red}{L_{BL}U_{TL} = \hat{A}_{BL}} \quad \color{red}{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}}}$
4	Crout variant	$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} L \backslash U_{TL} & U_{TR} \\ \hline L_{BL} & \hat{A}_{BR} \end{array} \right)$ $\wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad L_{TL}U_{TR} = \hat{A}_{TR}}{L_{BL}U_{TL} = \hat{A}_{BL} \quad \color{red}{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}}}$
5	Classical LU	$\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c c} L \backslash U_{TL} & U_{TR} \\ \hline L_{BL} & \hat{A}_{BR} - L_{BL}U_{TR} \end{array} \right)$ $\wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad L_{TL}U_{TR} = \hat{A}_{TR}}{L_{BL}U_{TL} = \hat{A}_{BL} \quad \color{red}{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}}}$

Figure 11.8: Loop invariants for various LU factorization algorithms.

### 11.8.1 Variant 1: Bordered algorithm

Consider the loop invariant:

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L \backslash U_{TL} & \hat{A}_{TR} \\ \hline \hat{A}_{BL} & \hat{A}_{BR} \end{array} \right)$$

$$\wedge \begin{array}{c|c} L_{TL}U_{TL} = \hat{A}_{TL} & \textcolor{red}{L_{TL}U_{TR} = \hat{A}_{TR}} \\ \hline \textcolor{red}{L_{BL}U_{TL} = \hat{A}_{BL}} & \textcolor{red}{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}} \end{array}$$

At the top of the loop, after repartitioning,  $A$  contains

$$\begin{array}{c|c|c} L \backslash U_{00} & \hat{a}_{01} & \hat{A}_{02} \\ \hline \hat{a}_{10}^T & \hat{\alpha}_{11} & \hat{a}_{12}^T \\ \hline \hat{A}_{20} & \hat{a}_{21} & \hat{A}_{22} \end{array}$$

while at the bottom it must contain

$$\begin{array}{c|c|c} L \backslash U_{00} & \textcolor{blue}{u}_{01} & \hat{A}_{02} \\ \hline \textcolor{blue}{l}_{10}^T & \textcolor{blue}{v}_{11} & \hat{a}_{12}^T \\ \hline \hat{A}_{20} & \hat{a}_{21} & \hat{A}_{22} \end{array}$$

where the entries in blue are to be computed. Now, considering  $LU = \hat{A}$  we notice that

$$\begin{array}{c|c|c} \textcolor{red}{L_{00}}U_{00} = \hat{A}_{00} & \textcolor{red}{L_{00}u_{01} = \hat{a}_{01}} & \textcolor{red}{L_{00}U_{02} = \hat{A}_{02}} \\ \hline \textcolor{red}{l_{10}^T U_{00} = \hat{a}_{10}^T} & \textcolor{red}{l_{10}^T u_{01} + v_{11} = \hat{\alpha}_{11}} & \textcolor{red}{l_{10}^T U_{02} + u_{12}^T = \hat{a}_{12}^T} \\ \hline L_{20}U_{00} = \hat{A}_{20} & L_{20}u_{01} + v_{11}l_{21} = \hat{a}_{21} & L_{20}U_{02} + l_{21}u_{12}^T + L_{22}U_{22} = \hat{A}_{22} \end{array}$$

where the entries in red are already known. The equalities in yellow can be used to compute the desired parts of  $L$  and  $U$ :

- Solve  $L_{00}u_{01} = \hat{a}_{01}$  for  $u_{01}$ , overwriting  $a_{01}$  with the result.
- Solve  $l_{10}^T U_{00} = \hat{a}_{10}^T$  (or, equivalently,  $U_{00}^T (l_{10}^T)^T = (\hat{a}_{10}^T)^T$  for  $l_{10}^T$ ), overwriting  $a_{10}^T$  with the result.
- Compute  $v_{11} = \hat{\alpha}_{11} - l_{10}^T u_{01}$ , overwriting  $\alpha_{11}$  with the result.

**Homework 11.15** If  $A$  is an  $n \times n$  matrix, show that the cost of Variant 1 is approximately  $\frac{2}{3}n^3$  flops.

• SEE ANSWER

### 11.8.2 Variant 2: Left-looking algorithm

Consider the loop invariant:

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L \backslash U_{TL} & \hat{A}_{TR} \\ \hline L_{BL} & \hat{A}_{BR} \end{array} \right)$$

$$\wedge \begin{array}{c|c} L_{TL}U_{TL} = \hat{A}_{TL} & \textcolor{red}{L_{TL}U_{TR} = \hat{A}_{TR}} \\ \hline L_{BL}U_{TL} = \hat{A}_{BL} & \textcolor{red}{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}} \end{array}$$

At the top of the loop, after repartitioning,  $A$  contains

$$\begin{array}{c|c|c} L \setminus U_{00} & \hat{a}_{01} & \hat{A}_{02} \\ \hline l_{10}^T & \hat{\alpha}_{11} & \hat{a}_{12}^T \\ \hline L_{20} & \hat{a}_{21} & \hat{A}_{22} \end{array}$$

while at the bottom it must contain

$$\begin{array}{c|c|c} L \setminus U_{00} & u_{01} & \hat{A}_{02} \\ \hline l_{10}^T & v_{11} & \hat{a}_{12}^T \\ \hline L_{20} & l_{21} & \hat{A}_{22} \end{array}$$

where the entries in blue are to be computed. Now, considering  $LU = \hat{A}$  we notice that

$$\begin{array}{c|c|c} L_{00}U_{00} = \hat{A}_{00} & L_{00}u_{01} = \hat{a}_{01} & L_{00}U_{02} = \hat{A}_{02} \\ \hline l_{10}^T U_{00} = \hat{a}_{10}^T & l_{10}^T u_{01} + v_{11} = \hat{\alpha}_{11} & l_{10}^T U_{02} + u_{12}^T = \hat{a}_{12}^T \\ \hline L_{20}U_{00} = \hat{A}_{20} & L_{20}u_{01} + v_{11}l_{21} = \hat{a}_{21} & L_{20}U_{02} + l_{21}u_{12}^T + L_{22}U_{22} = \hat{A}_{22} \end{array}$$

The equalities in yellow can be used to compute the desired parts of  $L$  and  $U$ :

- Solve  $L_{00}u_{01} = a_{01}$  for  $u_{01}$ , overwriting  $a_{01}$  with the result.
- Compute  $v_{11} = \alpha_{11} - l_{10}^T u_{01}$ , overwriting  $\alpha_{11}$  with the result.
- Compute  $l_{21} := (a_{21} - L_{20}u_{01})/v_{11}$ , overwriting  $a_{21}$  with the result.

### 11.8.3 Variant 3: Up-looking variant

**Homework 11.16** Derive the up-looking variant for computing the LU factorization.

🔗 [SEE ANSWER](#)

### 11.8.4 Variant 4: Crout variant

Consider the loop invariant:

$$\begin{array}{c} \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L \setminus U_{TL} & U_{TR} \\ \hline L_{BL} & \hat{A}_{BR} \end{array} \right) \\ \wedge \frac{L_{TL}U_{TL} = \hat{A}_{TL} \quad L_{TL}U_{TR} = \hat{A}_{TR}}{L_{BL}U_{TL} = \hat{A}_{BL} \quad \cancel{L_{BL}U_{TR}} + \cancel{L_{BR}U_{BR}} = \hat{A}_{BR}} \end{array}$$

At the top of the loop, after repartitioning,  $A$  contains

$$\begin{array}{c|c|c} L \setminus U_{00} & u_{01} & U_{02} \\ \hline l_{10}^T & \hat{\alpha}_{11} & \hat{a}_{12}^T \\ \hline L_{20} & \hat{a}_{21} & \hat{A}_{22} \end{array}$$



while at the bottom it must contain

$$\begin{array}{c|c|c} L \backslash U_{00} & u_{01} & U_{02} \\ \hline l_{10}^T & \mathfrak{v}_{11} & u_{12}^T \\ \hline L_{20} & \mathfrak{l}_{21} & \hat{A}_{22} \end{array}$$

where the entries in blue are to be computed. Now, considering  $LU = \hat{A}$  we notice that

$$\begin{array}{c|c|c} L_{00}U_{00} = \hat{A}_{00} & L_{00}u_{01} = \hat{a}_{01} & L_{00}U_{02} = \hat{A}_{02} \\ \hline l_{10}^T U_{00} = \hat{a}_{10}^T & l_{10}^T u_{01} + \mathfrak{v}_{11} = \hat{\alpha}_{11} & l_{10}^T U_{02} + u_{12}^T = \hat{a}_{12}^T \\ \hline L_{20}U_{00} = \hat{A}_{20} & L_{20}u_{01} + \mathfrak{v}_{11}l_{21} = \hat{a}_{21} & L_{20}U_{02} + l_{21}u_{12}^T + L_{22}U_{22} = \hat{A}_{22} \end{array}$$

The equalities in yellow can be used to compute the desired parts of  $L$  and  $U$ :

- Compute  $\mathfrak{v}_{11} = \alpha_{11} - l_{10}^T u_{01}$ , overwriting  $\alpha_{11}$  with the result.
- Compute  $\mathfrak{l}_{21} := (\alpha_{21} - L_{20}u_{01})/\mathfrak{v}_{11}$ , overwriting  $a_{21}$  with the result.
- Compute  $u_{12}^T := a_{12}^T - l_{10}^T U_{02}$ , overwriting  $a_{12}^T$  with the result.

### 11.8.5 Variant 5: Classical LU factorization

We have already derived this algorithm. You may want to try rederiving it using the techniques discussed in this section.

### 11.8.6 All algorithms

All five algorithms for LU factorization are summarized in Figure 11.9.

**Homework 11.17** Implement all five LU factorization algorithms with the FLAME@lab API, in M-script.

🔗 [SEE ANSWER](#)

**Homework 11.18** Which of the five variants can be modified to incorporate partial pivoting?

🔗 [SEE ANSWER](#)

### 11.8.7 Formal derivation of algorithms

The described approach to deriving algorithms, linking the process to the *a priori* identification of loop invariants, was first proposed in [24]. It was refined into what we call the “worksheet” for deriving algorithms hand-in-hand with their proofs of correctness, in [4]. A book that describes the process at a level also appropriate for the novice is “The Science of Programming Matrix Computations” [38].

**Algorithm:**  $A := L \setminus U = \text{LUA}$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$

**while**  $n(A_{TL}) < n(A)$  **do**

**Repartition**

$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$

**where**  $\alpha_{11}$  is  $1 \times 1$

$A_{00}$  contains  $L_{00}$  and  $U_{00}$  in its strictly lower and upper triangular part, respectively.

Variant 1	Variant 2	Variant 3	Variant 4	Variant 5
Bordered	Left-looking	Up-looking	Crout variant	Classical LU
$a_{01} := L_{00}^{-1} a_{01}$ $a_{10}^T := a_{10}^T U_{00}^{-1}$ $\alpha_{11} := \alpha_{11} - a_{21}^T a_{01}$	$a_{01} := L_{00}^{-1} a_{01}$ $\alpha_{11} := \alpha_{11} - a_{21}^T a_{01}$  $a_{21} := a_{21} - A_{20} a_{01}$ $a_{21} := a_{21} / \alpha_{11}$	Exercise in 11.16	$\alpha_{11} := \alpha_{11} - a_{21}^T a_{01}$ $a_{12}^T := a_{12}^T - a_{10}^T A_{02}$ $a_{21} := a_{21} - A_{20} a_{01}$ $a_{21} := a_{21} / \alpha_{11}$	$a_{21} := a_{21} / \alpha_{11}$ $A_{22} := A_{22} - a_{21} a_{12}^T$

**Continue with**

$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$

**endwhile**

Figure 11.9: All five LU factorization algorithms.

## 11.9 Numerical Stability Results

The numerical stability of various LU factorization algorithms as well as the triangular solve algorithms can be found in standard graduate level numerical linear algebra texts and references [21, 25, 34]. Of particular interest may be the analysis of the Crout variant (Variant 4) in [8], since it uses our notation as well as the results in “Notes on Numerical Stability”. (We recommend the technical report version [6] of the paper, since it has more details as well as exercises to help the reader understand.) In that paper, a systematic approach towards the derivation of backward error results is given that mirrors the systematic approach to deriving the algorithms given in [?, 4, 38].

Here are pertinent results from that paper, assuming floating point arithmetic obeys the model of computation given in “Notes on Numerical Stability” (as well as [8, 6, 25]). It is assumed that the reader is familiar with those notes.

**Theorem 11.19** *Let  $A \in \mathbb{R}^{n \times n}$  and let the LU factorization of  $A$  be computed via the Crout variant (Variant 4), yielding approximate factors  $\check{L}$  and  $\check{U}$ . Then*

$$(A + \Delta A) = \check{L}\check{U} \quad \text{with} \quad |\Delta A| \leq \gamma_n |\check{L}| |\check{U}|.$$

**Theorem 11.20** *Let  $L \in \mathbb{R}^{n \times n}$  be lower triangular and  $y, z \in \mathbb{R}^n$  with  $Lz = y$ . Let  $\check{z}$  be the approximate solution that is computed. Then*

$$(L + \Delta L)\check{z} = y \quad \text{with} \quad |\Delta L| \leq \gamma_n |L|.$$

**Theorem 11.21** *Let  $U \in \mathbb{R}^{n \times n}$  be upper triangular and  $x, z \in \mathbb{R}^n$  with  $Ux = z$ . Let  $\check{x}$  be the approximate solution that is computed. Then*

$$(U + \Delta U)\check{x} = z \quad \text{with} \quad |\Delta U| \leq \gamma_n |U|.$$

**Theorem 11.22** *Let  $A \in \mathbb{R}^{n \times n}$  and  $x, y \in \mathbb{R}^n$  with  $Ax = y$ . Let  $\check{x}$  be the approximate solution computed via the following steps:*

- *Compute the LU factorization, yielding approximate factors  $\check{L}$  and  $\check{U}$ .*
- *Solve  $\check{L}z = y$ , yielding approximate solution  $\check{z}$ .*
- *Solve  $\check{U}x = \check{z}$ , yielding approximate solution  $\check{x}$ .*

*Then  $(A + \Delta A)\check{x} = y$  with  $|\Delta A| \leq (3\gamma_n + \gamma_n^2)|\check{L}||\check{U}|$ .*

The analysis of LU factorization without partial pivoting is related that of LU factorization with partial pivoting. We have shown that LU with partial pivoting is equivalent to the LU factorization without partial pivoting on a pre-permuted matrix:  $PA = LU$ , where  $P$  is a permutation matrix. The permutation doesn't involve any floating point operations and therefore does not generate error. It can therefore be argued that, as a result, the error that is accumulated is equivalent with or without partial pivoting

## 11.10 Is LU with Partial Pivoting Stable?

**Homework 11.23** Apply LU with partial pivoting to

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ -1 & -1 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & & \cdots & 1 & 1 \\ -1 & -1 & & \cdots & -1 & 1 \end{pmatrix}.$$

Pivot only when necessary.

 [SEE ANSWER](#)

From this exercise we conclude that even LU factorization with partial pivoting can yield large (exponential) element growth in  $U$ . You may enjoy the collection of problems for which Gaussian elimination with partial pivoting is unstable by Stephen Wright [46].

In practice, this does not seem to happen and LU factorization is considered to be stable.

## 11.11 Blocked Algorithms

It is well-known that matrix-matrix multiplication can achieve high performance on most computer architectures [2, 22, 19]. As a result, many dense matrix algorithms are reformulated to be rich in matrix-matrix multiplication operations. An interface to a library of such operations is known as the level-3 Basic Linear Algebra Subprograms (BLAS) [17]. In this section, we show how LU factorization can be rearranged so that most computation is in matrix-matrix multiplications.

### 11.11.1 Blocked classical LU factorization (Variant 5)

Partition  $A$ ,  $L$ , and  $U$  as follows:

$$A \rightarrow \left( \begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right), \quad L \rightarrow \left( \begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & L_{22} \end{array} \right), \quad \text{and} \quad U \rightarrow \left( \begin{array}{c|c} U_{11} & U_{12} \\ \hline 0 & U_{22} \end{array} \right),$$

where  $A_{11}$ ,  $L_{11}$ , and  $U_{11}$  are  $b \times b$ . Then  $A = LU$  means that

$$\left( \begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & L_{22} \end{array} \right) \left( \begin{array}{c|c} U_{11} & U_{12} \\ \hline 0 & U_{22} \end{array} \right) = \left( \begin{array}{c|c} L_{11}U_{11} & L_{11}U_{12} \\ \hline L_{21}U_{11} & L_{21}U_{12} + L_{22}U_{22} \end{array} \right).$$

This means that

$$\begin{array}{c|c} A_{11} = L_{11}U_{11} & A_{12} = L_{11}U_{12} \\ \hline A_{21} = L_{21}U_{11} & A_{22} = L_{21}U_{12} + L_{22}U_{22} \end{array}$$

**Algorithm:**  $A := LU(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$

**while**  $m(A_{TL}) < m(A)$  **do**

**Determine block size**  $b$

**Repartition**

$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right)$

**where**  $A_{11}$  is  $b \times b$

$A_{00}$  contains  $L_{00}$  and  $U_{00}$  in its strictly lower and upper triangular part, respectively.

Variant 1:

$$A_{01} := L_{00}^{-1} A_{01}$$

$$A_{10} := A_{10} U_{00}^{-1}$$

$$A_{11} := A_{11} - A_{10} A_{01}$$

$$A_{11} := L U A_{11}$$

Variant 2:

$$A_{01} := L_{00}^{-1} A_{01}$$

$$A_{11} := A_{11} - A_{10} A_{01}$$

$$A_{11} := L U A_{11}$$

$$A_{21} := (A_{21} - A_{20} A_{01}) U_{11}^{-1}$$

Variant 3:

$$A_{10} := A_{10} U_{00}^{-1}$$

$$A_{11} := A_{11} - A_{10} A_{01}$$

$$A_{11} := L U A_{11}$$

$$A_{12} := A_{12} - A_{10} A_{02}$$

Variant 4:

$$A_{11} := A_{11} - A_{10} A_{01}$$

$$A_{11} := L U A_{11}$$

$$A_{21} := (A_{21} - A_{20} A_{01}) U_{11}^{-1}$$

$$A_{12} := L_{11}^{-1} (A_{12} - A_{10} A_{02})$$

Variant 5:

$$A_{11} := L U A_{11}$$

$$A_{21} := A_{21} U_{11}^{-1}$$

$$A_{12} := L_{11}^{-1} A_{12}$$

$$A_{22} := A_{22} - A_{21} A_{12}$$

**Continue with**

$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right)$

**endwhile**

Figure 11.10: Blocked algorithms for computing the LU factorization.

**Algorithm:**  $[A, p] := \text{LUPIV\_BLK}(A, p)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), p \rightarrow \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$ ,  $p_T$  has 0 elements.

**while**  $n(A_{TL}) < n(A)$  **do**

**Determine block size**  $b$

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \rightarrow \left( \begin{array}{c} p_0 \\ \hline p_1 \\ \hline p_2 \end{array} \right)$$

**where**  $A_{11}$  is  $b \times b$ ,  $p_1$  has  $b$  elements

Variant 2:

$$A_{01} := L_{00}^{-1} A_{01}$$

$$A_{11} := A_{11} - A_{10} A_{01}$$

$$A_{21} := A_{21} - A_{20} A_{01}$$

$$\left[ \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right), p_1 \right] :=$$

$$\text{LUpiv} \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right), p_1$$

$$\left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right) :=$$

$$P(p_1) \left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right)$$

Variant 4:

$$A_{11} := A_{11} - A_{10} A_{01}$$

$$A_{21} := A_{21} - A_{20} A_{01}$$

$$\left[ \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right), p_1 \right] :=$$

$$\text{LUpiv} \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right), p_1$$

$$\left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right) :=$$

$$P(p_1) \left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right)$$

$$A_{12} := A_{12} - A_{10} A_{02}$$

$$A_{12} := L_{11}^{-1} A_{12}$$

Variant 5:

$$\left[ \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right), p_1 \right] :=$$

$$\text{LUpiv} \left( \begin{array}{c} A_{11} \\ \hline A_{21} \end{array} \right), p_1$$

$$\left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right) :=$$

$$P(p_1) \left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right)$$

$$A_{12} := L_{11}^{-1} A_{12}$$

$$A_{22} := A_{22} - A_{21} A_{12}$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} p_T \\ \hline p_B \end{array} \right) \leftarrow \left( \begin{array}{c} p_0 \\ \hline p_1 \\ \hline p_2 \end{array} \right)$$

**endwhile**

Figure 11.11: Blocked algorithms for computing the LU factorization with partial pivoting.

or, equivalently,

$$\left( \begin{array}{c|c} A_{11} = L_{11}U_{11} & A_{12} = L_{11}U_{12} \\ \hline A_{21} = L_{21}U_{11} & A_{22} - L_{21}U_{12} = L_{22}U_{22} \end{array} \right).$$

If we let  $L$  and  $U$  overwrite the original matrix  $A$  this suggests the algorithm

- Compute the LU factorization  $A_{11} = L_{11}U_{11}$ , overwriting  $A_{11}$  with  $L_{11}$ . Notice that any of the “unblocked” algorithms previously discussed in this note can be used for this factorization.
- Solve  $L_{11}U_{12} = A_{12}$ , overwriting  $A_{12}$  with  $U_{12}$ . (This can also be expressed as  $A_{12} := L_{11}^{-1}A_{12}$ .)
- Solve  $L_{21}U_{11} = A_{21}$ , overwriting  $A_{21}$  with  $U_{21}$ . (This can also be expressed as  $A_{21} := A_{21}U_{11}^{-1}$ .)
- Update  $A_{22} := A_{22} - A_{21}A_{12}$ .
- Continue by overwriting the updated  $A_{22}$  with its LU factorization.

If  $b$  is small relative to  $n$ , then most computation is in the last step, which is a matrix-matrix multiplication. Similarly, blocked algorithms for the other variants can be derived. All are given in Figure 11.10.

### 11.11.2 Blocked classical LU factorization with pivoting (Variant 5)

Pivoting can be added to some of the blocked algorithms. Let us focus once again on Variant 5.

Partition  $A$ ,  $L$ , and  $U$  as follows:

$$A \rightarrow \left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right), \quad L \rightarrow \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline L_{10} & L_{11} & 0 \\ \hline L_{20} & L_{21} & L_{22} \end{array} \right), \quad \text{and} \quad U \rightarrow \left( \begin{array}{c|c|c} U_{00} & U_{01} & U_{02} \\ \hline 0 & U_{11} & U_{12} \\ \hline 0 & 0 & U_{22} \end{array} \right),$$

where  $A_{00}$ ,  $L_{00}$ , and  $U_{00}$  are  $k \times k$ , and  $A_{11}$ ,  $L_{11}$ , and  $U_{11}$  are  $b \times b$ .

Assume that the computation has proceeded to the point where  $A$  contains

$$\left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c|c} L \backslash U_{00} & U_{01} & U_{02} \\ \hline L_{10} & \hat{A}_{11} - L_{10}U_{01} & A_{12} - L_{10}U_{02} \\ \hline L_{20} & \hat{A}_{21} - L_{20}U_{01} & A_{22} - L_{20}U_{02} \end{array} \right),$$

where, as before,  $\hat{A}$  denotes the original contents of  $A$  and

$$P(p_0) \left( \begin{array}{c|c|c} \hat{A}_{00} & \hat{A}_{01} & \hat{A}_{02} \\ \hline \hat{A}_{10} & \hat{A}_{11} & \hat{A}_{12} \\ \hline \hat{A}_{20} & \hat{A}_{21} & \hat{A}_{22} \end{array} \right) = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline L_{10} & I & 0 \\ \hline L_{20} & 0 & I \end{array} \right) \left( \begin{array}{c|c|c} U_{00} & U_{01} & U_{02} \\ \hline 0 & A_{11} & A_{12} \\ \hline 0 & A_{21} & A_{22} \end{array} \right).$$

In the current blocked step, we now perform the following computations

- Compute the LU factorization with pivoting of the “current panel”  $\begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix}$ :

$$P(p_1) \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix} = \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} U_{11},$$

overwriting  $A_{11}$  with  $L \backslash U_{11}$  and  $A_{21}$  with  $L_{21}$ .

- Correspondingly, swap rows in the remainder of the matrix

$$\left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right) := P(p_1) \left( \begin{array}{c|c} A_{10} & A_{12} \\ \hline A_{20} & A_{22} \end{array} \right).$$

- Solve  $L_{11}U_{12} = A_{12}$ , overwriting  $A_{12}$  with  $U_{12}$ . (This can also be more concisely written as  $A_{12} := L_{11}^{-1}A_{12}$ .)
- Update  $A_{22} := A_{22} - A_{21}A_{12}$ .

Careful consideration shows that this puts the matrix  $A$  in the state

$$\left( \begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c|c} L \backslash U_{00} & U_{01} & U_{02} \\ \hline L_{10} & L \backslash U_{11} & U_{12} \\ \hline L_{20} & L_{21} & \hat{A}_{22} - L_{20}U_{02} - L_{21}U_{12} \end{array} \right),$$

where

$$P \left( \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} \right) \left( \begin{array}{c|c|c} \hat{A}_{00} & \hat{A}_{01} & \hat{A}_{02} \\ \hline \hat{A}_{10} & \hat{A}_{11} & \hat{A}_{12} \\ \hline \hat{A}_{20} & \hat{A}_{21} & \hat{A}_{22} \end{array} \right) = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline L_{10} & L_{11} & 0 \\ \hline L_{20} & L_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} U_{00} & U_{01} & U_{02} \\ \hline 0 & U_{11} & U_{12} \\ \hline 0 & 0 & A_{22} \end{array} \right).$$

Similarly, blocked algorithms with pivoting for some of the other variants can be derived. All are given in Figure 11.10.

## 11.12 Variations on a Triple-Nested Loop

All LU factorization algorithms presented in this note perform exactly the same floating point operations (with some rearrangement of data thrown in for the algorithms that perform pivoting) as does the triple-nested loop that implements Gaussian elimination:



---

<b>for</b> $j = 0, \dots, n - 1$	(zero the elements below $(j, j)$ element)
<b>for</b> $i = j + 1, \dots, n - 1$	
$\alpha_{i,j} := \alpha_{i,j} / \alpha_{j,j}$	(compute multiplier $\lambda_{i,j}$ , overwriting $\alpha_{i,j}$ )
<b>for</b> $k = j + 1, \dots, n - 1$	(subtract $\lambda_{i,j}$ times the $j$ th row from $i$ th row)
$\alpha_{i,k} := \alpha_{i,k} - \alpha_{i,j} \alpha_{j,k}$	
<b>endfor</b>	
<b>endfor</b>	
<b>endfor</b>	

---



# Chapter 12

## Notes on Cholesky Factorization

### Video

Read disclaimer regarding the videos in the preface!

 [YouTube](#)

 [Download from UT Box](#)

 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

## Outline

<b>Video</b>	<b>195</b>
<b>Outline</b>	<b>196</b>
<b>12.1. Definition and Existence</b>	<b>197</b>
<b>12.2. Application</b>	<b>197</b>
<b>12.3. An Algorithm</b>	<b>198</b>
<b>12.4. Proof of the Cholesky Factorization Theorem</b>	<b>199</b>
<b>12.5. Blocked Algorithm</b>	<b>200</b>
<b>12.6. Alternative Representation</b>	<b>200</b>
<b>12.7. Cost</b>	<b>203</b>
<b>12.8. Solving the Linear Least-Squares Problem via the Cholesky Factorization</b>	<b>204</b>
<b>12.9. Other Cholesky Factorization Algorithms</b>	<b>204</b>
<b>12.10 Implementing the Cholesky Factorization with the (Traditional) BLAS</b>	<b>206</b>
12.10.1. What are the BLAS?	206
12.10.2. A simple implementation in Fortran	209
12.10.3. Implementation with calls to level-1 BLAS	209
12.10.4. Matrix-vector operations (level-2 BLAS)	209
12.10.5. Matrix-matrix operations (level-3 BLAS)	213
12.10.6. Impact on performance	213
<b>12.11 Alternatives to the BLAS</b>	<b>214</b>
12.11.1. The FLAME/C API	214
12.11.2. BLIS	214

## 12.1 Definition and Existence

This operation is only defined for Hermitian positive definite matrices:

**Definition 12.1** A matrix  $A \in \mathbb{C}^{m \times m}$  is Hermitian positive definite (HPD) if and only if it is Hermitian ( $A^H = A$ ) and for all nonzero vectors  $x \in \mathbb{C}^m$  it is the case that  $x^H A x > 0$ . If in addition  $A \in \mathbb{R}^{m \times m}$  then  $A$  is said to be symmetric positive definite (SPD).

(If you feel uncomfortable with complex arithmetic, just replace the word “Hermitian” with “Symmetric” in this document and the Hermitian transpose operation,  $^H$ , with the transpose operation,  $^T$ .)

**Example 12.2** Consider the case where  $m = 1$  so that  $A$  is a real scalar,  $\alpha$ . Notice that then  $A$  is SPD if and only if  $\alpha > 0$ . This is because then for all nonzero  $\chi \in \mathbb{R}$  it is the case that  $\alpha \chi^2 > 0$ .

First some exercises:

**Homework 12.3** Let  $B \in \mathbb{C}^{m \times n}$  have linearly independent columns. Prove that  $A = B^H B$  is HPD.

🔗 [SEE ANSWER](#)

**Homework 12.4** Let  $A \in \mathbb{C}^{m \times m}$  be HPD. Show that its diagonal elements are real and positive.

🔗 [SEE ANSWER](#)

We will prove the following theorem in Section 12.4:

**Theorem 12.5 (Cholesky Factorization Theorem)** Given a HPD matrix  $A$  there exists a lower triangular matrix  $L$  such that  $A = LL^H$ .

Obviously, there similarly exists an upper triangular matrix  $U$  such that  $A = U^H U$  since we can choose  $U^H = L$ .

The lower triangular matrix  $L$  is known as the Cholesky factor and  $LL^H$  is known as the Cholesky factorization of  $A$ . It is unique if the diagonal elements of  $L$  are restricted to be positive.

The operation that overwrites the lower triangular part of matrix  $A$  with its Cholesky factor will be denoted by  $A := \text{Chol}A$ , which should be read as “ $A$  becomes its Cholesky factor.” Typically, only the lower (or upper) triangular part of  $A$  is stored, and it is that part that is then overwritten with the result. In this discussion, we will assume that the lower triangular part of  $A$  is stored and overwritten.

## 12.2 Application

The Cholesky factorization is used to solve the linear system  $Ax = y$  when  $A$  is HPD: Substituting the factors into the equation yields  $LL^H x = y$ . Letting  $z = L^H x$ ,

$$Ax = L \underbrace{(L^H x)}_z = Lz = y.$$

Thus,  $z$  can be computed by solving the triangular system of equations  $Lz = y$  and subsequently the desired solution  $x$  can be computed by solving the triangular linear system  $L^H x = z$ .

## 12.3 An Algorithm

The most common algorithm for computing  $A := \text{Chol}A$  can be derived as follows: Consider  $A = LL^H$ . Partition

$$A = \left( \begin{array}{c|c} \alpha_{11} & \star \\ \hline a_{21} & A_{22} \end{array} \right) \quad \text{and} \quad L = \left( \begin{array}{c|c} \lambda_{11} & 0 \\ \hline l_{21} & L_{22} \end{array} \right). \quad (12.1)$$

**Remark 12.6** We adopt the commonly used notation where Greek lower case letters refer to scalars, lower case letters refer to (column) vectors, and upper case letters refer to matrices. (This is convention is often attributed to Alston Householder.) The  $\star$  refers to a part of  $A$  that is neither stored nor updated.

By substituting these partitioned matrices into  $A = LL^H$  we find that

$$\begin{aligned} \left( \begin{array}{c|c} \alpha_{11} & \star \\ \hline a_{21} & A_{22} \end{array} \right) &= \left( \begin{array}{c|c} \lambda_{11} & 0 \\ \hline l_{21} & L_{22} \end{array} \right) \left( \begin{array}{c|c} \lambda_{11} & 0 \\ \hline l_{21} & L_{22} \end{array} \right)^H = \left( \begin{array}{c|c} \lambda_{11} & 0 \\ \hline l_{21} & L_{22} \end{array} \right) \left( \begin{array}{c|c} \bar{\lambda}_{11} & l_{21}^H \\ \hline 0 & L_{22}^H \end{array} \right) \\ &= \left( \begin{array}{c|c} |\lambda_{11}|^2 & \star \\ \hline \bar{\lambda}_{11} l_{21} & l_{21} l_{21}^H + L_{22} L_{22}^H \end{array} \right), \end{aligned}$$

from which we conclude that

$$\frac{|\lambda_{11}| = \sqrt{\alpha_{11}}}{l_{21} = a_{21}/\bar{\lambda}_{11}} \left| \begin{array}{c|c} \star \\ \hline L_{22} = \text{Chol}A_{22} - l_{21} l_{21}^H \end{array} \right|.$$

These equalities motivate the algorithm

1. Partition  $A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & \star \\ \hline a_{21} & A_{22} \end{array} \right)$ .
2. Overwrite  $\alpha_{11} := \lambda_{11} = \sqrt{\alpha_{11}}$ . (Picking  $\lambda_{11} = \sqrt{\alpha_{11}}$  makes it positive and real, and ensures uniqueness.)
3. Overwrite  $a_{21} := l_{21} = a_{21}/\lambda_{11}$ .
4. Overwrite  $A_{22} := A_{22} - l_{21} l_{21}^H$  (updating only the lower triangular part of  $A_{22}$ ). This operation is called a *symmetric rank-1 update*.
5. Continue with  $A = A_{22}$ . (Back to Step 1.)

**Remark 12.7** Similar to the `tril` function in Matlab, we use  $\text{tril}(B)$  to denote the lower triangular part of matrix  $B$ .

## 12.4 Proof of the Cholesky Factorization Theorem

In this section, we partition  $A$  as in (12.1):

$$A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{21}^H \\ \hline a_{21} & A_{22} \end{array} \right).$$

The following lemmas are key to the proof:

**Lemma 12.8** *Let  $A \in \mathbb{C}^{n \times n}$  be HPD. Then  $\alpha_{11}$  is real and positive.*

**Proof:** This is special case of Exercise 12.4.

**Lemma 12.9** *Let  $A \in \mathbb{C}^{m \times m}$  be HPD and  $l_{21} = a_{21}/\sqrt{\alpha_{11}}$ . Then  $A_{22} - l_{21}l_{21}^H$  is HPD.*

**Proof:** Since  $A$  is Hermitian so are  $A_{22}$  and  $A_{22} - l_{21}l_{21}^H$ .

Let  $x_2 \in \mathbb{C}^{(n-1) \times (n-1)}$  be an arbitrary nonzero vector. Define  $x = \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix}$  where  $\chi_1 = -a_{21}^H x_2 / \alpha_{11}$ .

Then, since  $x \neq 0$ ,

$$\begin{aligned} 0 < x^H A x &= \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix}^H \begin{pmatrix} \alpha_{11} & a_{21}^H \\ \hline a_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix}^H \begin{pmatrix} \alpha_{11}\chi_1 + a_{21}^H x_2 \\ a_{21}\chi_1 + A_{22}x_2 \end{pmatrix} \\ &= \alpha_{11}|\chi_1|^2 + \bar{\chi}_1 a_{21}^H x_2 + x_2^H a_{21} \chi_1 + x_2^H A_{22} x_2 \\ &= \alpha_{11} \frac{a_{21}^H x_2}{\alpha_{11}} \frac{x_2^H a_{21}}{\alpha_{11}} - \frac{x_2^H a_{21}}{\alpha_{11}} a_{21}^H x_2 - x_2^H a_{21} \frac{a_{21}^H x_2}{\alpha_{11}} + x_2^H A_{22} x_2 \\ &= x_2^H \left( A_{22} - \frac{a_{21} a_{21}^H}{\alpha_{11}} \right) x_2 \quad (\text{since } x_2^H a_{21} a_{21}^H x_2 \text{ is real and hence equals } a_{21}^H x_2 x_2^H a_{21}) \\ &= x_2^H (A_{22} - l_{21} l_{21}^H) x_2. \end{aligned}$$

We conclude that  $A_{22} - l_{21}l_{21}^H$  is HPD.

**Proof:** of the **Cholesky Factorization Theorem**

Proof by induction.

**Base case:**  $n = 1$ . Clearly the result is true for a  $1 \times 1$  matrix  $A = \alpha_{11}$ : In this case, the fact that  $A$  is HPD means that  $\alpha_{11}$  is real and positive and a Cholesky factor is then given by  $\lambda_{11} = \sqrt{\alpha_{11}}$ , with uniqueness if we insist that  $\lambda_{11}$  is positive.

**Inductive step:** Assume the result is true for HPD matrix  $A \in \mathbb{C}^{(n-1) \times (n-1)}$ . We will show that it holds for  $A \in \mathbb{C}^{n \times n}$ . Let  $A \in \mathbb{C}^{n \times n}$  be HPD. Partition  $A$  and  $L$  as in (12.1) and let  $\lambda_{11} = \sqrt{\alpha_{11}}$  (which is well-defined by Lemma 12.8),  $l_{21} = a_{21}/\lambda_{11}$ , and  $L_{22} = \text{Chol}A_{22} - l_{21}l_{21}^H$  (which exists as a consequence of the Inductive Hypothesis and Lemma 12.9). Then  $L$  is the desired Cholesky factor of  $A$ .

**By the principle of mathematical induction**, the theorem holds.

## 12.5 Blocked Algorithm

In order to attain high performance, the computation is cast in terms of matrix-matrix multiplication by so-called blocked algorithms. For the Cholesky factorization a blocked version of the algorithm can be derived by partitioning

$$A \rightarrow \left( \begin{array}{c|c} A_{11} & \star \\ \hline A_{21} & A_{22} \end{array} \right) \quad \text{and} \quad L \rightarrow \left( \begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & L_{22} \end{array} \right),$$

where  $A_{11}$  and  $L_{11}$  are  $b \times b$ . By substituting these partitioned matrices into  $A = LL^H$  we find that

$$\left( \begin{array}{c|c} A_{11} & \star \\ \hline A_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & L_{22} \end{array} \right) \left( \begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & L_{22} \end{array} \right)^H = \left( \begin{array}{c|c} L_{11}L_{11}^H & \star \\ \hline L_{21}L_{11}^H & L_{21}L_{21}^H + L_{22}L_{22}^H \end{array} \right).$$

From this we conclude that

$$\begin{array}{c|c} L_{11} = \text{Chol}A_{11} & \star \\ \hline L_{21} = A_{21}L_{11}^{-H} & L_{22} = \text{Chol}A_{22} - L_{21}L_{21}^H \end{array}.$$

An algorithm is then described by the steps

1. Partition  $A \rightarrow \left( \begin{array}{c|c} A_{11} & \star \\ \hline A_{21} & A_{22} \end{array} \right)$ , where  $A_{11}$  is  $b \times b$ .
2. Overwrite  $A_{11} := L_{11} = \text{Chol}A_{11}$ .
3. Overwrite  $A_{21} := L_{21} = A_{21}L_{11}^{-H}$ .
4. Overwrite  $A_{22} := A_{22} - L_{21}L_{21}^H$  (updating only the lower triangular part).
5. Continue with  $A = A_{22}$ . (Back to Step 1.)

**Remark 12.10** The Cholesky factorization  $A_{11} := L_{11} = \text{Chol}A_{11}$  can be computed with the unblocked algorithm or by calling the blocked Cholesky factorization algorithm recursively.

**Remark 12.11** Operations like  $L_{21} = A_{21}L_{11}^{-H}$  are computed by solving the equivalent linear system with multiple right-hand sides  $L_{11}L_{21}^H = A_{21}^H$ .

## 12.6 Alternative Representation

When explaining the above algorithm in a classroom setting, invariably it is accompanied by a picture sequence like the one in Figure 12.1(left)<sup>1</sup> and the (verbal) explanation:

<sup>1</sup> Picture modified from a similar one in [24].



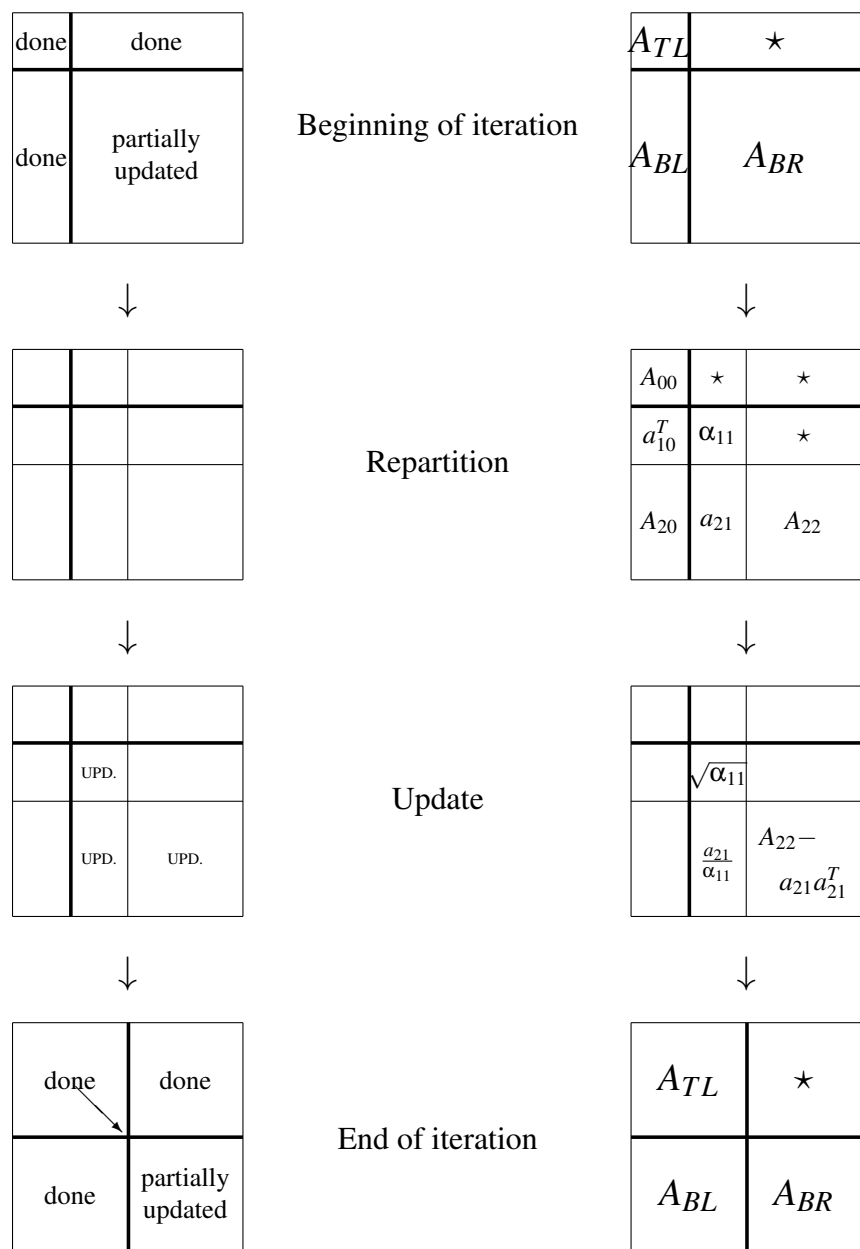


Figure 12.1: Left: Progression of pictures that explain Cholesky factorization algorithm. Right: Same pictures, annotated with labels and updates.

**Algorithm:**  $A := \text{CHOL\_UNB}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$

$$\alpha_{11} := \sqrt{\alpha_{11}}$$

$$a_{21} := a_{21} / \alpha_{11}$$

$$A_{22} := A_{22} - \text{tril}(a_{21} a_{21}^H)$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$

**endwhile**

**Algorithm:**  $A := \text{CHOL\_BLK}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline A_{10} & A_{11} & \star \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right)$$

$$A_{11} := \text{Chol}A_{11}$$

$$A_{21} := A_{21} \text{tril}(A_{11})^{-H}$$

$$A_{22} := A_{22} - \text{tril}(A_{21} A_{21}^H)$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline A_{10} & A_{11} & \star \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right)$$

**endwhile**

Figure 12.2: Unblocked and blocked algorithms for computing the Cholesky factorization in FLAME notation.

**Beginning of iteration:** At some stage of the algorithm (Top of the loop), the computation has moved through the matrix to the point indicated by the thick lines. Notice that we have finished with the parts of the matrix that are in the top-left, top-right (which is not to be touched), and bottom-left quadrants. The bottom-right quadrant has been updated to the point where we only need to perform a Cholesky factorization of it.

**Repartition:** We now repartition the bottom-right submatrix to expose  $\alpha_{11}$ ,  $a_{21}$ , and  $A_{22}$ .

**Update:**  $\alpha_{11}$ ,  $a_{21}$ , and  $A_{22}$  are updated as discussed before.

**End of iteration:** The thick lines are moved, since we now have completed more of the computation, and only a factorization of  $A_{22}$  (which becomes the new bottom-right quadrant) remains to be performed.

**Continue:** The above steps are repeated until the submatrix  $A_{BR}$  is empty.

To motivate our notation, we annotate this progression of pictures as in Figure 12.1 (right). In those pictures, “T”, “B”, “L”, and “R” stand for “Top”, “Bottom”, “Left”, and “Right”, respectively. This then motivates the format of the algorithm in Figure 12.2 (left). It uses what we call the FLAME notation for representing algorithms [24, 23, 38]. A similar explanation can be given for the blocked algorithm, which is given in Figure 12.2 (right). In the algorithms,  $m(A)$  indicates the number of rows of matrix  $A$ .

**Remark 12.12** *The indices in our more stylized presentation of the algorithms are subscripts rather than indices in the conventional sense.*

**Remark 12.13** *The notation in Figs. 12.1 and 12.2 allows the contents of matrix  $A$  at the beginning of the iteration to be formally stated:*

$$A = \left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L_{TL} & \star \\ \hline L_{BL} & \hat{A}_{BR} - \text{tril}(L_{BL}L_{BL}^H) \end{array} \right),$$

where  $L_{TL} = \text{Chol} \hat{A}_{TL}$ ,  $L_{BL} = \hat{A}_{BL}L_{TL}^{-H}$ , and  $\hat{A}_{TL}$ ,  $\hat{A}_{BL}$  and  $\hat{A}_{BR}$  denote the original contents of the quadrants  $A_{TL}$ ,  $A_{BL}$  and  $A_{BR}$ , respectively.

**Homework 12.14** *Implement the Cholesky factorization with M-script.*

🔗 [SEE ANSWER](#)

## 12.7 Cost

The cost of the Cholesky factorization of  $A \in \mathbb{C}^{m \times m}$  can be analyzed as follows: In Figure 12.2 (left) during the  $k$ th iteration (starting  $k$  at zero)  $A_{00}$  is  $k \times k$ . Thus, the operations in that iteration cost

- $\alpha_{11} := \sqrt{\alpha_{11}}$ : negligible when  $k$  is large.
- $a_{21} := a_{21}/\alpha_{11}$ : approximately  $(m - k - 1)$  flops.
- $A_{22} := A_{22} - \text{tril}(a_{21}a_{21}^H)$ : approximately  $(m - k - 1)^2$  flops. (A rank-1 update of all of  $A_{22}$  would have cost  $2(m - k - 1)^2$  flops. Approximately half the entries of  $A_{22}$  are updated.)

Thus, the total cost in flops is given by

$$\begin{aligned}
 C_{\text{Chol}}(m) &\approx \underbrace{\sum_{k=0}^{m-1} (m-k-1)^2}_{\text{(Due to update of } A_{22})} + \underbrace{\sum_{k=0}^{m-1} (m-k-1)}_{\text{(Due to update of } a_{21})} \\
 &= \sum_{j=0}^{m-1} j^2 + \sum_{j=0}^{m-1} j \approx \frac{1}{3}m^3 + \frac{1}{2}m^2 \approx \frac{1}{3}m^3
 \end{aligned}$$

which allows us to state that (obvious) most computation is in the update of  $A_{22}$ . It can be shown that the blocked Cholesky factorization algorithm performs exactly the same number of floating point operations.

Comparing the cost of the Cholesky factorization to that of the LU factorization from “Notes on LU Factorization” we see that taking advantage of symmetry cuts the cost approximately in half.

## 12.8 Solving the Linear Least-Squares Problem via the Cholesky Factorization

Recall that if  $B \in \mathbb{C}^{m \times n}$  has linearly independent columns, then  $A = B^H B$  is HPD. Also, recall from “Notes on Linear Least-Squares” that the solution,  $x \in \mathbb{C}^n$  to the linear least-squares (LLS) problem

$$\|Bx - y\|_2 = \min_{z \in \mathbb{C}^n} \|Bz - y\|_2$$

equals the solution to the normal equations

$$\underbrace{B^H B}_A x = \underbrace{B^H y}_{\hat{y}}.$$

This makes it obvious how the Cholesky factorization can (and often is) used to solve the LLS problem.

**Homework 12.15** Consider  $B \in \mathbb{C}^{m \times n}$  with linearly independent columns. Recall that  $B$  has a QR factorization,  $B = QR$  where  $Q$  has orthonormal columns and  $R$  is an upper triangular matrix with positive diagonal elements. How are the Cholesky factorization of  $B^H B$  and the QR factorization of  $B$  related?

🔗 [SEE ANSWER](#)

## 12.9 Other Cholesky Factorization Algorithms

There are actually three different unblocked and three different blocked algorithms for computing the Cholesky factorization. The algorithms we discussed so far in this note are sometimes called *right-looking* algorithms. Systematic derivation of all these algorithms, as well as their blocked counterparts, are given in Chapter 6 of [38]. In this section, a sequence of exercises leads to what is often referred to as the *bordered* Cholesky factorization algorithm.

**Algorithm:**  $A := \text{CHOL\_UNB}(A)$  Bordered algorithm)

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$

**where**  $\alpha_{11}$  is  $1 \times 1$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$

**endwhile**

Figure 12.3: Unblocked Cholesky factorization, bordered variant, for Homework 12.17.

**Homework 12.16** Let  $A$  be SPD and partition

$$A \rightarrow \left( \begin{array}{c|c} A_{00} & a_{10} \\ \hline a_{10}^T & \alpha_{11} \end{array} \right)$$

(Hint: For this exercise, use techniques similar to those in Section 12.4.)

1. Show that  $A_{00}$  is SPD.
2. Assuming that  $A_{00} = L_{00}L_{00}^T$ , where  $L_{00}$  is lower triangular and nonsingular, argue that the assignment  $l_{10}^T := a_{10}^T L_{00}^{-T}$  is well-defined.
3. Assuming that  $A_{00}$  is SPD,  $A_{00} = L_{00}L_{00}^T$  where  $L_{00}$  is lower triangular and nonsingular, and  $l_{10}^T = a_{10}^T L_{00}^{-T}$ , show that  $\alpha_{11} - l_{10}^T l_{10} > 0$  so that  $\lambda_{11} := \sqrt{\alpha_{11} - l_{10}^T l_{10}}$  is well-defined.

4. Show that

$$\left( \begin{array}{c|c} A_{00} & a_{10} \\ \hline a_{10}^T & \alpha_{11} \end{array} \right) = \left( \begin{array}{c|c} L_{00} & 0 \\ \hline l_{10}^T & \lambda_{11} \end{array} \right) \left( \begin{array}{c|c} L_{00} & 0 \\ \hline l_{10}^T & \lambda_{11} \end{array} \right)^T$$

➡ SEE ANSWER

**Homework 12.17** Use the results in the last exercise to give an alternative proof by induction of the Cholesky Factorization Theorem and to give an algorithm for computing it by filling in Figure 12.3. This algorithm is often referred to as the bordered Cholesky factorization algorithm.

➡ SEE ANSWER

**Homework 12.18** Show that the cost of the bordered algorithm is, approximately,  $\frac{1}{3}n^3$  flops.

➡ SEE ANSWER

## 12.10 Implementing the Cholesky Factorization with the (Traditional) BLAS

The Basic Linear Algebra Subprograms (BLAS) are an interface to commonly used fundamental linear algebra operations. In this section, we illustrate how the unblocked and blocked Cholesky factorization algorithms can be implemented in terms of the BLAS. The explanation draws from the entry we wrote for the BLAS in the Encyclopedia of Parallel Computing [37].

### 12.10.1 What are the BLAS?

The BLAS interface [28, 18, 17] was proposed to support portable high-performance implementation of applications that are matrix and/or vector computation intensive. The idea is that one casts computation in terms of the BLAS interface, leaving the architecture-specific optimization of that interface to an expert.

	Algorithm	Code
Simple	<pre> <b>for</b> <math>j = 1 : n</math>   <math>\alpha_{j,j} := \sqrt{\alpha_{j,j}}</math>    <b>for</b> <math>i = j + 1 : n</math>     <math>\alpha_{i,j} := \alpha_{i,j} / \alpha_{j,j}</math>   <b>endfor</b>    <b>for</b> <math>k = j + 1 : n</math>     <b>for</b> <math>i = k : n</math>       <math>\alpha_{i,k} := \alpha_{i,k} - \alpha_{i,j} \alpha_{k,j}</math>     <b>endfor</b>   <b>endfor</b> <b>endfor</b> </pre>	<pre> do j=1, n   A(j,j) = sqrt(A(j,j))    do i=j+1,n     A(i,j) = A(i,j) / A(j,j)   enddo    do k=j+1,n     do i=k,n       A(i,k) = A(i,k) - A(i,j) * A(k,j)     enddo   enddo enddo </pre>
Vector-vector	<pre> <b>for</b> <math>j = 1 : n</math>   <math>\alpha_{j,j} := \sqrt{\alpha_{j,j}}</math>    <math>\alpha_{j+1:n,j} := \alpha_{j+1:n,j} / \alpha_{j,j}</math>    <b>for</b> <math>k = j + 1 : n</math>     <math>\alpha_{k:n,k} := -\alpha_{k,j} \alpha_{k:n,j} + \alpha_{k:n,k}</math>   <b>endfor</b> <b>endfor</b> </pre>	<pre> do j=1, n   A( j,j ) = sqrt( A( j,j ) )    call dscal( n-j, 1.0d00 / A(j,j), A(j+1,j), 1 )    do k=j+1,n     call daxpy( n-k+1, -A(k,j), A(k,j), 1,                A(k,k), 1 )   enddo enddo </pre>

Figure 12.4: Simple and vector-vector (level-1 BLAS) based representations of the right-looking algorithm.

	Algorithm	Code
Matrix-vector	<b>for</b> $j = 1 : n$ $\alpha_{j,j} := \sqrt{\alpha_{j,j}}$ $\alpha_{j+1:n,j} := \alpha_{j+1:n,j} / \alpha_{j,j}$ $\alpha_{j+1:n,j+1:n} :=$ $\quad -\text{tril}(\alpha_{j+1:n,j} \alpha_{j+1:n,j}^T) + \alpha_{j+1:n,j+1:n}$ <b>endfor</b>	<pre> do j=1, n   A(j,j) = sqrt(A(j,j))    call dscal( n-j, 1.0d00 / A(j,j), A(j+1,j), 1 )    call dsyr( 'lower triangular',             n-j, -1.0, A(j+1,j), 1, A(j+1,j+1), lda ) enddo </pre>
FLAME Notation	<p><b>Partition</b> <math>A \rightarrow \left( \begin{array}{c c} A_{TL} &amp; \star \\ \hline A_{BL} &amp; A_{BR} \end{array} \right)</math></p> <p><b>where</b> <math>A_{TL}</math> is <math>0 \times 0</math></p> <p><b>while</b> <math>m(A_{TL}) &lt; m(A)</math> <b>do</b></p> <p><b>Repartition</b></p> $\left( \begin{array}{c c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c c c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$ <p><b>where</b> <math>\alpha_{11}</math> is <math>1 \times 1</math></p> <hr/> $\alpha_{11} := \sqrt{\alpha_{11}}$ $a_{21} := a_{21} / \alpha_{11}$ $A_{22} := A_{22} - \text{tril}(a_{21} a_{21}^H)$ <hr/> <p><b>Continue with</b></p> $\left( \begin{array}{c c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c c c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$ <p><b>endwhile</b></p>	<pre> int Chol_unb_var3( FLA_Obj A ) {   FLA_Obj ATL,   ATR,     A00, a01,   A02,           ABL,   ABR,     a10t, alpha11, a12t,           A20, a21,     A22;    FLA_Part_2x2( A,      &amp;ATL, &amp;ATR,                 &amp;ABL, &amp;ABR,      0, 0, FLA_TL );    while ( FLA_Obj_length( ATL ) &lt; FLA_Obj_length( A ) ){     FLA_Repart_2x2_to_3x3(       ATL, /**/ ATR,      &amp;A00, /**/ &amp;a01,      &amp;A02,       /* ***** */ /* ***** */       &amp;a10t, /**/ &amp;alpha11, &amp;a12t,       ABL, /**/ ABR,      &amp;A20, /**/ &amp;a21,      &amp;A22,       1, 1, FLA_BR );      /*-----*/     FLA_Sqrt( alpha11 );     FLA_Invscl( alpha11, a21 );     FLA_Syr( FLA_LOWER_TRIANGULAR, FLA_MINUS_ONE,              A21, A22 );     /*-----*/     FLA_Cont_with_3x3_to_2x2(       &amp;ATL, /**/ &amp;ATR,      A00, a01,      /**/ A02,       &amp;a10t, alpha11, /**/ a12t,       /* ***** */ /* ***** */       &amp;ABL, /**/ &amp;ABR,      A20, a21,      /**/ A22,       FLA_TL );   }   return FLA_SUCCESS; } </pre>

Figure 12.5: Matrix-vector (level-2 BLAS) based representations of the right-looking algorithm.



### 12.10.2 A simple implementation in Fortran

We start with a simple implementation in Fortran. A simple algorithm that does not use BLAS and the corresponding code is given in the row labeled “Simple” in Figure 12.4. This sets the stage for our explanation of how the algorithm and code can be represented with vector-vector, matrix-vector, and matrix-matrix operations, and the corresponding calls to BLAS routines.

### 12.10.3 Implementation with calls to level-1 BLAS

The first BLAS interface [28] was proposed in the 1970s when vector supercomputers like the early Cray architectures reigned. On such computers, it sufficed to cast computation in terms of vector operations. As long as memory was accessed mostly contiguously, near-peak performance could be achieved. This interface is now referred to as the Level-1 BLAS. It was used for the implementation of the first widely used dense linear algebra package, LINPACK [16].

Let  $x$  and  $y$  be vectors of appropriate length and  $\alpha$  be scalar. In this and other notes we have *vector-vector operations* such as scaling of a vector ( $x := \alpha x$ ), inner (dot) product ( $\alpha := x^T y$ ), and scaled vector addition ( $y := \alpha x + y$ ). This last operation is known as an `axpy`, which stands for alpha times x plus y.

Our Cholesky factorization algorithm expressed in terms of such vector-vector operations and the corresponding code are given in Figure 12.4 in the row labeled “Vector-vector”. If the operations supported by `dscal` and `daxpy` achieve high performance on a target architecture (as it was in the days of vector supercomputers) then so will the implementation of the Cholesky factorization, since it casts most computation in terms of those operations. Unfortunately, vector-vector operations perform  $O(n)$  computation on  $O(n)$  data, meaning that these days the bandwidth to memory typically limits performance, since retrieving a data item from memory is often more than an order of magnitude more costly than a floating point operation with that data item.

We summarize information about level-1 BLAS in Figure 12.6.

### 12.10.4 Matrix-vector operations (level-2 BLAS)

The next level of BLAS supports operations with matrices and vectors. The simplest example of such an operation is the matrix-vector product:  $y := Ax$  where  $x$  and  $y$  are vectors and  $A$  is a matrix. Another example is the computation  $A_{22} = -a_{21}a_{21}^T + A_{22}$  (symmetric rank-1 update) in the Cholesky factorization. Here only the lower (or upper) triangular part of the matrix is updated, taking advantage of symmetry.

The use of symmetric rank-1 update is illustrated in Figure 12.5, in the row labeled “Matrix-vector”. There `dsyr` is the routine that implements a double symmetric rank-1 update. Readability of the code is improved by casting computation in terms of routines that implement the operations that appear in the algorithm: `dscal` for  $a_{21} = a_{21}/\alpha_{11}$  and `dsyr` for  $A_{22} = -a_{21}a_{21}^T + A_{22}$ .

If the operation supported by `dsyr` achieves high performance on a target architecture (as it was in the days of vector supercomputers) then so will this implementation of the Cholesky factorization, since it casts most computation in terms of that operation. Unfortunately, matrix-vector operations perform  $O(n^2)$  computation on  $O(n^2)$  data, meaning that these days the bandwidth to memory typically limits performance.

We summarize information about level-2 BLAS in Figure 12.7.

A proto-typical calling sequence for a level-1 BLAS routine is

`□axpy( n, alpha, x, incx, y, incy ),`

which implements the scaled vector addition operation  $y = \alpha x + y$ . Here

- The “□” indicates the data type. The choices for this first letter are

s	<u>s</u> ingle precision
d	<u>d</u> ouble precision
c	single precision <u>c</u> omplex
z	double precision complex

- The operation is identified as `axpy`: alpha times x plus y.
- `n` indicates the number of elements in the vectors  $x$  and  $y$ .
- `alpha` is the scalar  $\alpha$ .
- `x` and `y` indicate the memory locations where the first elements of  $x$  and  $y$  are stored, respectively.
- `incx` and `incy` equal the increment by which one has to stride through memory for the elements of vectors  $x$  and  $y$ , respectively

The following are the most frequently used level-1 BLAS:

routine/ function	operation
□swap	$x \leftrightarrow y$
□scal	$x \leftarrow \alpha x$
□copy	$y \leftarrow x$
□axpy	$y \leftarrow \alpha x + y$
□dot	$x^T y$
□nrm2	$\ x\ _2$
□asum	$\ \operatorname{re}(x)\ _1 + \ \operatorname{im}(x)\ _1$
i□max	$\min(k) :  \operatorname{re}(x_k)  +  \operatorname{im}(x_k)  = \max( \operatorname{re}(x_i)  +  \operatorname{im}(x_i) )$

Figure 12.6: Summary of the most commonly used level-1 BLAS.

The naming convention for level-2 BLAS routines is given by

$$\square\text{XXYY},$$

where

- “ $\square$ ” can take on the values s, d, c, z.
- XX indicates the shape of the matrix.
- YY indicates the operation to be performed:

XX	matrix shape	YY	matrix shape
ge	<u>g</u> eneral (rectangular)	mv	<u>m</u> atrix <u>v</u> ector multiplication
sy	<u>s</u> ymmetric	sv	<u>s</u> olve <u>v</u> ector
he	<u>H</u> ermitian	r	<u>r</u> ank-1 update
tr	<u>t</u> riangular	r2	<u>r</u> ank- <u>2</u> update

- In addition, operations with banded matrices are supported, which we do not discuss here.

A representative call to a level-2 BLAS operation is given by

```
dsyr( uplo, n, alpha, x, incx, A, lda )
```

which implements the operation  $A = \alpha x x^T + A$ , updating the lower or upper triangular part of  $A$  by choosing `uplo` as ‘Lower triangular’ or ‘Upper triangular’, respectively. The parameter `lda` (the leading dimension of matrix  $A$ ) indicates the increment by which memory has to be traversed in order to address successive elements in a row of matrix  $A$ .

The following table gives the most commonly used level-2 BLAS operations:

routine/ function	operation
$\square$ gemv	<u>g</u> eneral <u>m</u> atrix- <u>v</u> ector multiplication
$\square$ symv	<u>s</u> ymmetric <u>m</u> atrix- <u>v</u> ector multiplication
$\square$ trmv	<u>t</u> riangular <u>m</u> atrix- <u>v</u> ector multiplication
$\square$ trsv	<u>t</u> riangular <u>s</u> olve <u>v</u> ector
$\square$ ger	<u>g</u> eneral <u>r</u> ank-1 update
$\square$ syr	<u>s</u> ymmetric <u>r</u> ank-1 update
$\square$ syr2	<u>s</u> ymmetric <u>r</u> ank- <u>2</u> update

Figure 12.7: Summary of the most commonly used level-2 BLAS.

	Algorithm	Code
Matrix-matrix	<p><b>for</b> <math>j = 1 : n</math> <b>in steps of</b> <math>n_b</math></p> <p><math>b := \min(n - j + 1, n_b)</math></p> <p><math>A_{j:j+b-1, j:j+b-1} := \text{Chol}A_{j:j+b-1, j:j+b-1}</math></p> <p><math>A_{j+b:n, j:j+b-1} :=</math></p> <p><math>A_{j+b:n, j:j+b-1} A_{j:j+b-1, j:j+b-1}^{-H}</math></p> <p><math>A_{j+b:n, j+b:n} := A_{j+b:n, j+b:n}</math></p> <p><math>- \text{tril} \left( A_{j+b:n, j:j+b-1} A_{j+b:n, j:j+b-1}^H \right)</math></p> <p><b>endfor</b></p>	<pre> do j=1, n, nb   jb = min( nb, n-j+1 )    call chol( jb, A( j, j ), lda )    call dtrsm( 'Right', 'Lower triangular',               'Transpose', 'Nonunit diag',               J-JB+1, JB, 1.0d00, A( j, j ), lda,               A( j+jb, j ), lda )    call dsyrk( 'Lower triangular', 'No transpose',               J-JB+1, JB, -1.0d00, A( j+jb, j ), lda,               1.0d00, A( j+jb, j+jb ), lda )  enddo </pre>
FLAME Notation	<p><b>Partition</b> <math>A \rightarrow \left( \begin{array}{c c} A_{TL} &amp; * \\ \hline A_{BL} &amp; A_{BR} \end{array} \right)</math></p> <p><b>where</b> <math>A_{TL}</math> is <math>0 \times 0</math></p> <p><b>while</b> <math>m(A_{TL}) &lt; m(A)</math> <b>do</b></p> <p><b>Determine block size</b> <math>b</math></p> <p><b>Repartition</b></p> <p><math>\left( \begin{array}{c c} A_{TL} &amp; * \\ \hline A_{BL} &amp; A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c c c} A_{00} &amp; * &amp; * \\ \hline A_{10} &amp; A_{11} &amp; * \\ \hline A_{20} &amp; A_{21} &amp; A_{22} \end{array} \right)</math></p> <p><b>where</b> <math>A_{11}</math> is <math>b \times b</math></p> <hr/> <p><math>A_{11} := \text{Chol}A_{11}</math></p> <p><math>A_{21} := A_{21} \text{tril}(A_{11})^{-H}</math></p> <p><math>A_{22} := A_{22} - \text{tril}(A_{21}A_{21}^H)</math></p> <hr/> <p><b>Continue with</b></p> <p><math>\left( \begin{array}{c c} A_{TL} &amp; * \\ \hline A_{BL} &amp; A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c c c} A_{00} &amp; * &amp; * \\ \hline A_{10} &amp; A_{11} &amp; * \\ \hline A_{20} &amp; A_{21} &amp; A_{22} \end{array} \right)</math></p> <p><b>endwhile</b></p>	<pre> int Chol_unb_var3( FLA_Obj A ) {   FLA_Obj ATL,   ATR,   A00, a01,   A02,           ABL,   ABR,   a10t, alpha11, a12t,           A20, a21,   A22;    FLA_Part_2x2( A,      &amp;ATL, &amp;ATR,                 &amp;ABL, &amp;ABR,      0, 0, FLA_TL );    while ( FLA_Obj_length( ATL ) &lt; FLA_Obj_length( A ) ){     FLA_Repart_2x2_to_3x3(       ATL, /**/ ATR,      &amp;A00, /**/ &amp;a01,      &amp;A02,       /* ***** */ /* ***** */       &amp;a10t, /**/ &amp;alpha11, &amp;a12t,       ABL, /**/ ABR,      &amp;A20, /**/ &amp;a21,      &amp;A22,       1, 1, FLA_BR );      /*-----*/     FLA_Sqrt( alpha11 );     FLA_Invscal( alpha11, a21 );     FLA_Syr( FLA_LOWER_TRIANGULAR, FLA_MINUS_ONE,              A21, A22 );      /*-----*/     FLA_Cont_with_3x3_to_2x2(       &amp;ATL, /**/ &amp;ATR,      A00, a01,      /**/ A02,       &amp;a10t, alpha11, /**/ a12t,       /* ***** */ /* ***** */       &amp;ABL, /**/ &amp;ABR,      A20, a21,      /**/ A22,       FLA_TL );   }   return FLA_SUCCESS; } </pre>

Figure 12.8: Blocked algorithm and implementation with level-3 BLAS.

The naming convention for level-3 BLAS routines are similar to those for the level-2 BLAS. A representative call to a level-3 BLAS operation is given by

```
dsyrk( uplo, trans, n, k, alpha, A, lda, beta, C, ldc )
```

which implements the operation  $C := \alpha AA^T + \beta C$  or  $C := \alpha A^T A + \beta C$  depending on whether `trans` is chosen as 'No transpose' or 'Transpose', respectively. It updates the lower or upper triangular part of  $C$  depending on whether `uplo` equal 'Lower triangular' or 'Upper triangular', respectively. The parameters `lda` and `ldc` are the leading dimensions of arrays  $A$  and  $C$ , respectively.

The following table gives the most commonly used Level-3 BLAS operations

routine/ function	operation
<input type="checkbox"/> <code>gemm</code>	<u>g</u> eneral <u>m</u> atrix- <u>m</u> atrix multiplication
<input type="checkbox"/> <code>symm</code>	<u>s</u> ymmetric <u>m</u> atrix- <u>m</u> atrix multiplication
<input type="checkbox"/> <code>trmm</code>	<u>t</u> riangular <u>m</u> atrix- <u>m</u> atrix multiplication
<input type="checkbox"/> <code>trsm</code>	<u>t</u> riangular <u>s</u> olve with <u>m</u> ultiple right-hand sides
<input type="checkbox"/> <code>syrk</code>	<u>s</u> ymmetric <u>r</u> ank- <u>k</u> update
<input type="checkbox"/> <code>syr2k</code>	<u>s</u> ymmetric rank- <u>2k</u> update

Figure 12.9: Summary of the most commonly used level-3 BLAS.

### 12.10.5 Matrix-matrix operations (level-3 BLAS)

Finally, we turn to the implementation of the blocked Cholesky factorization algorithm from Section 12.5. The algorithm is expressed with FLAME notation and Matlab-like notation in Figure 12.8.

The routines `dtrsm` and `dsyrk` are level-3 BLAS routines:

- The call to `dtrsm` implements  $A_{21} := L_{21}$  where  $L_{21}L_{11}^T = A_{21}$ .
- The call to `dsyrk` implements  $A_{22} := -L_{21}L_{21}^T + A_{22}$ .

The bulk of the computation is now cast in terms of matrix-matrix operations which can achieve high performance.

We summarize information about level-3 BLAS in Figure 12.9.

### 12.10.6 Impact on performance

Figure 12.10 illustrates the performance benefits that come from using the different levels of BLAS on a typical architecture.

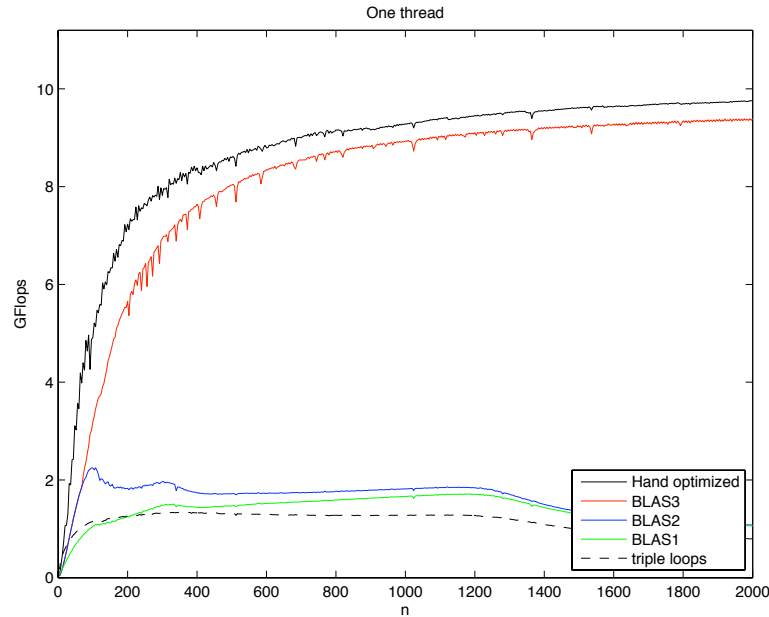


Figure 12.10: Performance of the different implementations of Cholesky factorization that use different levels of BLAS. The target processor has a peak of 11.2 Gflops (billions of floating point operations per second). BLAS1, BLAS2, and BLAS3 indicate that the bulk of computation was cast in terms of level-1, -2, or -3 BLAS, respectively.

## 12.11 Alternatives to the BLAS

### 12.11.1 The FLAME/C API

In a number of places in these notes we presented algorithms in FLAME notation. Clearly, there is a disconnect between this notation and how the algorithms are then encoded with the BLAS interface. In Figures 12.4, 12.5, and 12.8 we also show how the FLAME API for the C programming language [5] allows the algorithms to be more naturally translated into code. While the traditional BLAS interface underlies the implementation of Cholesky factorization and other algorithms in the widely used LAPACK library [1], the FLAME/C API is used in our `libflame` library [24, 39, 40].

### 12.11.2 BLIS

The implementations that call BLAS in this paper are coded in Fortran. More recently, the languages of choice for scientific computing have become C and C++. While there is a C interface to the traditional BLAS called the CBLAS [10], we believe a more elegant such interface is the BLAS-like Library Instantiation Software (BLIS) interface [41]. BLIS is not only a framework for rapid implementation of the traditional BLAS, but also presents an alternative interface for C and C++ users.

# Chapter 13

## Notes on Eigenvalues and Eigenvectors

If you have forgotten how to find the eigenvalues and eigenvectors of  $2 \times 2$  and  $3 \times 3$  matrices, you may want to review

[Linear Algebra: Foundations to Frontiers - Notes to LAFF With \[29\]](#).

### Video

Read disclaimer regarding the videos in the preface!

 [YouTube](#)

 [Download from UT Box](#)

 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

## Outline

<a href="#">Video</a> . . . . .	215
<a href="#">Outline</a> . . . . .	216
<a href="#">13.1. Definition</a> . . . . .	217
<a href="#">13.2. The Schur and Spectral Factorizations</a> . . . . .	220
<a href="#">13.3. Relation Between the SVD and the Spectral Decomposition</a> . . . . .	222

---



## 13.1 Definition

**Definition 13.1** Let  $A \in \mathbb{C}^{m \times m}$ . Then  $\lambda \in \mathbb{C}$  and nonzero  $x \in \mathbb{C}^m$  are said to be an eigenvalue and corresponding eigenvector if  $Ax = \lambda x$ . The tuple  $(\lambda, x)$  is said to be an eigenpair. The set of all eigenvalues of  $A$  is denoted by  $\Lambda(A)$  and is called the spectrum of  $A$ . The spectral radius of  $A$ ,  $\rho(A)$  equals the magnitude of the largest eigenvalue, in magnitude:

$$\rho(A) = \max_{\lambda \in \Lambda(A)} |\lambda|.$$

The action of  $A$  on an eigenvector  $x$  is as if it were multiplied by a scalar. The direction does not change, only its length is scaled:

$$Ax = \lambda x.$$

**Theorem 13.2** Scalar  $\lambda$  is an eigenvalue of  $A$  if and only if

$$(\lambda I - A) \left\{ \begin{array}{l} \text{is singular} \\ \text{has a nontrivial null-space} \\ \text{has linearly dependent columns} \\ \det(\lambda I - A) = 0 \\ (\lambda I - A)x = 0 \text{ has a nontrivial solution} \\ \text{etc.} \end{array} \right.$$

The following exercises expose some other basic properties of eigenvalues and eigenvectors:

**Homework 13.3** Eigenvectors are not unique.

 [SEE ANSWER](#)

**Homework 13.4** Let  $\lambda$  be an eigenvalue of  $A$  and let  $\mathcal{E}_\lambda(A) = \{x \in \mathbb{C}^m | Ax = \lambda x\}$  denote the set of all eigenvectors of  $A$  associated with  $\lambda$  (including the zero vector, which is not really considered an eigenvector). Show that this set is a (nontrivial) subspace of  $\mathbb{C}^m$ .

 [SEE ANSWER](#)

**Definition 13.5** Given  $A \in \mathbb{C}^{m \times m}$ , the function  $p_m(\lambda) = \det(\lambda I - A)$  is a polynomial of degree at most  $m$ . This polynomial is called the characteristic polynomial of  $A$ .

The definition of  $p_m(\lambda)$  and the fact that is a polynomial of degree at most  $m$  is a consequence of the definition of the determinant of an arbitrary square matrix. This definition is not particularly enlightening other than that it allows one to succinctly related eigenvalues to the roots of the characteristic polynomial.

**Remark 13.6** The relation between eigenvalues and the roots of the characteristic polynomial yield a disconcerting insight: A general formula for the eigenvalues of a  $m \times m$  matrix with  $m > 4$  does not exist.

The reason is that there is no general formula for the roots of a polynomial of degree  $m > 4$ . Given any polynomial  $p_m(\chi)$  of degree  $m$ , an  $m \times m$  matrix can be constructed such that its characteristic polynomial is  $p_m(\lambda)$ . If

$$p_m(\chi) = \alpha_0 + \alpha_1\chi + \cdots + \alpha_{m-1}\chi^{m-1} + \chi^m$$

and

$$A = \begin{pmatrix} -\alpha_{n-1} & -\alpha_{n-2} & -\alpha_{n-3} & \cdots & -\alpha_1 & -\alpha_0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

then

$$p_m(\lambda) = \det(\lambda I - A)$$

Hence, we conclude that no general formula can be found for the eigenvalues for  $m \times m$  matrices when  $m > 4$ . What we will see in future “Notes on ...” is that we will instead create algorithms that *converge* to the eigenvalues and/or eigenvalues of matrices.

**Theorem 13.7** Let  $A \in \mathbb{C}^{m \times m}$  and  $p_m(\lambda)$  be its characteristic polynomial. Then  $\lambda \in \Lambda(A)$  if and only if  $p_m(\lambda) = 0$ .

**Proof:** This is an immediate consequence of Theorem 13.2.

In other words,  $\lambda$  is an eigenvalue of  $A$  if and only if it is a root of  $p_m(\lambda)$ . This has the immediate consequence that  $A$  has at most  $m$  eigenvalues and, if one counts multiple roots by their multiplicity, it has exactly  $m$  eigenvalues. (One says “Matrix  $A \in \mathbb{C}^{m \times m}$  has  $m$  eigenvalues, multiplicity counted.”)

**Homework 13.8** The eigenvalues of a diagonal matrix equal the values on its diagonal. The eigenvalues of a triangular matrix equal the values on its diagonal.

➡ SEE ANSWER

**Corollary 13.9** If  $A \in \mathbb{R}^{m \times m}$  is real valued then some or all of its eigenvalues may be complex valued. In this case, if  $\lambda \in \Lambda(A)$  then so is its conjugate,  $\bar{\lambda}$ .

**Proof:** It can be shown that if  $A$  is real valued, then the coefficients of its characteristic polynomial are all real valued. Complex roots of a polynomial with real coefficients come in conjugate pairs.

It is not hard to see that an eigenvalue that is a root of multiplicity  $k$  has at most  $k$  eigenvectors. It is, however, not necessarily the case that an eigenvalue that is a root of multiplicity  $k$  also has  $k$  linearly independent eigenvectors. In other words, the null space of  $\lambda I - A$  may have dimension less than the algebraic multiplicity of  $\lambda$ . The prototypical counter example is the  $k \times k$  matrix

$$J(\mu) = \begin{pmatrix} \mu & 1 & 0 & \cdots & 0 & 0 \\ 0 & \mu & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & \mu & 1 \\ 0 & 0 & 0 & \cdots & 0 & \mu \end{pmatrix}$$

where  $k > 1$ . Observe that  $\lambda I - J(\mu)$  is singular if and only if  $\lambda = \mu$ . Since  $\mu I - J(\mu)$  has  $k - 1$  linearly independent columns its null-space has dimension one: all eigenvectors are scalar multiples of each other. This matrix is known as a *Jordan block*.

**Definition 13.10** A matrix  $A \in \mathbb{C}^{m \times m}$  that has fewer than  $m$  linearly independent eigenvectors is said to be defective. A matrix that does have  $m$  linearly independent eigenvectors is said to be nondefective.

**Theorem 13.11** Let  $A \in \mathbb{C}^{m \times m}$ . There exist nonsingular matrix  $X$  and diagonal matrix  $\Lambda$  such that  $A = X\Lambda X^{-1}$  if and only if  $A$  is nondefective.

**Proof:**

( $\Rightarrow$ ). Assume there exist nonsingular matrix  $X$  and diagonal matrix  $\Lambda$  so that  $A = X\Lambda X^{-1}$ . Then, equivalently,  $AX = X\Lambda$ . Partition  $X$  by columns so that

$$\begin{aligned} A \left( \begin{array}{c|c|c|c} x_0 & x_1 & \cdots & x_{m-1} \end{array} \right) &= \left( \begin{array}{c|c|c|c} x_0 & x_1 & \cdots & x_{m-1} \end{array} \right) \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{m-1} \end{pmatrix} \\ &= \left( \begin{array}{c|c|c|c} \lambda_0 x_0 & \lambda_1 x_1 & \cdots & \lambda_{m-1} x_{m-1} \end{array} \right). \end{aligned}$$

Then, clearly,  $Ax_j = \lambda_j x_j$  so that  $A$  has  $m$  linearly independent eigenvectors and is thus nondefective.

( $\Leftarrow$ ). Assume that  $A$  is nondefective. Let  $\{x_0, \dots, x_{m-1}\}$  equal  $m$  linearly independent eigenvectors corresponding to eigenvalues  $\{\lambda_0, \dots, \lambda_{m-1}\}$ . If  $X = \left( \begin{array}{c|c|c|c} x_0 & x_1 & \cdots & x_{m-1} \end{array} \right)$  then  $AX = X\Lambda$  where  $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{m-1})$ . Hence  $A = X\Lambda X^{-1}$ .

**Definition 13.12** Let  $\mu \in \Lambda(A)$  and  $p_m(\lambda)$  be the characteristic polynomial of  $A$ . Then the algebraic multiplicity of  $\mu$  is defined as the multiplicity of  $\mu$  as a root of  $p_m(\lambda)$ .

**Definition 13.13** Let  $\mu \in \Lambda(A)$ . Then the geometric multiplicity of  $\mu$  is defined to be the dimension of  $\mathcal{E}_\mu(A)$ . In other words, the geometric multiplicity of  $\mu$  equals the number of linearly independent eigenvectors that are associated with  $\mu$ .

**Theorem 13.14** Let  $A \in \mathbb{C}^{m \times m}$ . Let the eigenvalues of  $A$  be given by  $\lambda_0, \lambda_1, \dots, \lambda_{k-1}$ , where an eigenvalue is listed exactly  $n$  times if it has geometric multiplicity  $n$ . There exists a nonsingular matrix  $X$  such that

$$A = X \begin{pmatrix} J(\lambda_0) & 0 & \cdots & 0 \\ 0 & J(\lambda_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J(\lambda_{k-1}) \end{pmatrix}.$$

For our discussion, the sizes of the Jordan blocks  $J(\lambda_i)$  are not particularly important. Indeed, this decomposition, known as the Jordan Canonical Form of matrix  $A$ , is not particularly interesting in practice. For this reason, we don't discuss it further and do not give its proof.

## 13.2 The Schur and Spectral Factorizations

**Theorem 13.15** Let  $A, Y, B \in \mathbb{C}^{m \times m}$ , assume  $Y$  is nonsingular, and let  $B = Y^{-1}AY$ . Then  $\Lambda(A) = \Lambda(B)$ .

**Proof:** Let  $\lambda \in \Lambda(A)$  and  $x$  be an associated eigenvector. Then  $Ax = \lambda x$  if and only if  $Y^{-1}AYY^{-1}x = Y^{-1}\lambda x$  if and only if  $B(Y^{-1}x) = \lambda(Y^{-1}x)$ .

**Definition 13.16** Matrices  $A$  and  $B$  are said to be similar if there exists a nonsingular matrix  $Y$  such that  $B = Y^{-1}AY$ .

Given a nonsingular matrix  $Y$  the transformation  $Y^{-1}AY$  is called a similarity transformation of  $A$ .

It is not hard to expand the last proof to show that if  $A$  is similar to  $B$  and  $\lambda \in \Lambda(A)$  has algebraic/geometric multiplicity  $k$  then  $\lambda \in \Lambda(B)$  has algebraic/geometric multiplicity  $k$ .

The following is the fundamental theorem for the algebraic eigenvalue problem:

**Theorem 13.17 Schur Decomposition Theorem** Let  $A \in \mathbb{C}^{m \times m}$ . Then there exist a unitary matrix  $Q$  and upper triangular matrix  $U$  such that  $A = QUQ^H$ . This decomposition is called the Schur decomposition of matrix  $A$ .

In the above theorem,  $\Lambda(A) = \Lambda(U)$  and hence the eigenvalues of  $A$  can be found on the diagonal of  $U$ .

**Proof:** We will outline how to construct  $Q$  so that  $Q^HAQ = U$ , an upper triangular matrix.

Since a polynomial of degree  $m$  has at least one root, matrix  $A$  has at least one eigenvalue,  $\lambda_1$ , and corresponding eigenvector  $q_1$ , where we normalize this eigenvector to have length one. Thus  $Aq_1 = \lambda_1 q_1$ . Choose  $Q_2$  so that  $Q = \begin{pmatrix} q_1 & Q_2 \end{pmatrix}$  is unitary. Then

$$\begin{aligned} Q^HAQ &= \begin{pmatrix} q_1 & Q_2 \end{pmatrix}^H A \begin{pmatrix} q_1 & Q_2 \end{pmatrix} \\ &= \begin{pmatrix} q_1^H A q_1 & q_1^H A Q_2 \\ Q_2^H A q_1 & Q_2^H A Q_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & q_1^H A Q_2 \\ \lambda Q_2^H q_1 & Q_2^H A Q_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & w^T \\ 0 & B \end{pmatrix}, \end{aligned}$$

where  $w^T = q_1^H A Q_2$  and  $B = Q_2^H A Q_2$ . This insight can be used to construct an inductive proof.

One should not mistake the above theorem and its proof as a constructive way to compute the Schur decomposition: finding an eigenvalue and/or the eigenvalue associated with it is difficult.

**Lemma 13.18** Let  $A \in \mathbb{C}^{m \times m}$  be of form  $A = \begin{pmatrix} A_{TL} & A_{TR} \\ 0 & A_{BR} \end{pmatrix}$ . Assume that  $Q_{TL}$  and  $Q_{BR}$  are unitary “of appropriate size”. Then

$$A = \begin{pmatrix} Q_{TL} & 0 \\ 0 & Q_{BR} \end{pmatrix}^H \begin{pmatrix} Q_{TL} A_{TL} Q_{TL}^H & Q_{TL} A_{TR} Q_{BR}^H \\ 0 & Q_{BR} A_{BR} Q_{BR}^H \end{pmatrix} \begin{pmatrix} Q_{TL} & 0 \\ 0 & Q_{BR} \end{pmatrix}.$$

**Homework 13.19** Prove Lemma 13.18. Then generalize it to a result for block upper triangular matrices:

$$A = \left( \begin{array}{c|c|c|c} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ \hline 0 & A_{1,1} & \cdots & A_{1,N-1} \\ \hline 0 & 0 & \ddots & \vdots \\ \hline 0 & 0 & \cdots & A_{N-1,N-1} \end{array} \right).$$

➡ SEE ANSWER

**Corollary 13.20** Let  $A \in \mathbb{C}^{m \times m}$  be of the form  $A = \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline 0 & A_{BR} \end{array} \right)$ . Then  $\Lambda(A) = \Lambda(A_{TL}) \cup \Lambda(A_{BR})$ .

**Homework 13.21** Prove Corollary 13.20. Then generalize it to a result for block upper triangular matrices.

➡ SEE ANSWER

A theorem that will later allow the eigenvalues and vectors of a real matrix to be computed (mostly) without requiring complex arithmetic is given by

**Theorem 13.22** Let  $A \in \mathbb{R}^{m \times m}$ . Then there exist a unitary matrix  $Q \in \mathbb{R}^{m \times m}$  and quasi upper triangular matrix  $U \in \mathbb{R}^{m \times m}$  such that  $A = QUQ^T$ .

A quasi upper triangular matrix is a block upper triangular matrix where the blocks on the diagonal are  $1 \times 1$  or  $2 \times 2$ . Complex eigenvalues of  $A$  are found as the complex eigenvalues of those  $2 \times 2$  blocks on the diagonal.

**Theorem 13.23 Spectral Decomposition Theorem** Let  $A \in \mathbb{C}^{m \times m}$  be Hermitian. Then there exist a unitary matrix  $Q$  and diagonal matrix  $\Lambda \in \mathbb{R}^{m \times m}$  such that  $A = Q\Lambda Q^H$ . This decomposition is called the Spectral decomposition of matrix  $A$ .

**Proof:** From the Schur Decomposition Theorem we know that there exist a matrix  $Q$  and upper triangular matrix  $U$  such that  $A = QUQ^H$ . Since  $A = A^H$  we know that  $QUQ^H = QU^H Q^H$  and hence  $U = U^H$ . But a Hermitian triangular matrix is diagonal with real valued diagonal entries.

What we conclude is that a Hermitian matrix is nondefective and its eigenvectors can be chosen to form an orthogonal basis.

**Homework 13.24** Let  $A$  be Hermitian and  $\lambda$  and  $\mu$  be distinct eigenvalues with eigenvectors  $x_\lambda$  and  $x_\mu$ , respectively. Then  $x_\lambda^H x_\mu = 0$ . (In other words, the eigenvectors of a Hermitian matrix corresponding to distinct eigenvalues are orthogonal.)

➡ SEE ANSWER

### 13.3 Relation Between the SVD and the Spectral Decomposition

**Homework 13.25** Let  $A \in \mathbb{C}^{m \times m}$  be a Hermitian matrix,  $A = Q\Lambda Q^H$  its Spectral Decomposition, and  $A = U\Sigma V^H$  its SVD. Relate  $Q$ ,  $U$ ,  $V$ ,  $\Lambda$ , and  $\Sigma$ .

👉 [SEE ANSWER](#)

**Homework 13.26** Let  $A \in \mathbb{C}^{m \times m}$  and  $A = U\Sigma V^H$  its SVD. Relate the Spectral decompositions of  $A^H A$  and  $AA^H$  to  $U$ ,  $V$ , and  $\Sigma$ .

👉 [SEE ANSWER](#)

# Chapter 14

## Notes on the Power Method and Related Methods

You may want to review Chapter 12 of

[Linear Algebra: Foundations to Frontiers - Notes to LAFF With \[29\]](#)

in which the Power Method and Inverse Power Methods are discussed at a more rudimentary level.

### Video

Read disclaimer regarding the videos in the preface!

Tragically, I forgot to turn on the camera... This was a great lecture!

## Outline

<b>Video</b> . . . . .	<b>223</b>
<b>Outline</b> . . . . .	<b>224</b>
<b>14.1. The Power Method</b> . . . . .	<b>225</b>
14.1.1. First attempt . . . . .	225
14.1.2. Second attempt . . . . .	226
14.1.3. Convergence . . . . .	227
14.1.4. Practical Power Method . . . . .	230
14.1.5. The Rayleigh quotient . . . . .	230
14.1.6. What if $ \lambda_0  \geq  \lambda_1 $ ? . . . .	231
<b>14.2. The Inverse Power Method</b> . . . . .	<b>231</b>
<b>14.3. Rayleigh-quotient Iteration</b> . . . . .	<b>232</b>



## 14.1 The Power Method

The Power Method is a simple method that under mild conditions yields a vector corresponding to the eigenvalue that is largest in magnitude.

Throughout this section we will assume that a given matrix  $A \in \mathbb{C}^{m \times m}$  is *nondeficient*: there exists a nonsingular matrix  $X$  and diagonal matrix  $\Lambda$  such that  $A = X\Lambda X^{-1}$ . (Sometimes this is called a diagonalizable matrix since there exists a matrix  $X$  so that

$$X^{-1}AX = \Lambda \text{ or, equivalently, } A = X\Lambda X^{-1}.$$

From “Notes on Eigenvalues and Eigenvectors” we know then that the columns of  $X$  equal eigenvectors of  $A$  and the elements on the diagonal of  $\Lambda$  equal the eigenvalues::

$$X = \left( x_0 \mid x_1 \mid \cdots \mid x_{m-1} \right) \quad \text{and} \quad \Lambda = \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{m-1} \end{pmatrix}$$

so that

$$Ax_i = \lambda_i x_i \quad \text{for } i = 0, \dots, m-1.$$

For most of this section we will assume that

$$|\lambda_0| > |\lambda_1| \geq \cdots \geq |\lambda_{m-1}|.$$

In particular,  $\lambda_0$  is the eigenvalue with maximal absolute value.

### 14.1.1 First attempt

Now, let  $v^{(0)} \in \mathbb{C}^{m \times m}$  be an “initial guess”. Our (first attempt at the) Power Method iterates as follows:

```
for  $k = 0, \dots$ 
   $v^{(k+1)} = Av^{(k)}$ 
endfor
```

Clearly  $v^{(k)} = A^k v^{(0)}$ . Let

$$v^{(0)} = Xy = \psi_0 x_0 + \psi_1 x_1 + \cdots + \psi_{m-1} x_{m-1}.$$

What does this mean? We view the columns of  $X$  as forming a basis for  $\mathbb{C}^m$  and then the elements in vector  $y = X^{-1}v^{(0)}$  equal the coefficients for describing  $v^{(0)}$  in that basis. Then

$$\begin{aligned} v^{(1)} = Av^{(0)} &= A(\psi_0 x_0 + \psi_1 x_1 + \cdots + \psi_{m-1} x_{m-1}) \\ &= \psi_0 \lambda_0 x_0 + \psi_1 \lambda_1 x_1 + \cdots + \psi_{m-1} \lambda_{m-1} x_{m-1}, \\ v^{(2)} = Av^{(1)} &= \psi_0 \lambda_0^2 x_0 + \psi_1 \lambda_1^2 x_1 + \cdots + \psi_{m-1} \lambda_{m-1}^2 x_{m-1}, \\ &\vdots \\ v^{(k)} = Av^{(k-1)} &= \psi_0 \lambda_0^k x_0 + \psi_1 \lambda_1^k x_1 + \cdots + \psi_{m-1} \lambda_{m-1}^k x_{m-1}. \end{aligned}$$

Now, as long as  $\psi_0 \neq 0$  clearly  $\psi_0 \lambda_0^k x_0$  will eventually dominate which means that  $v^{(k)}$  will start pointing in the direction of  $x_0$ . In other words, it will start pointing in the direction of an eigenvector corresponding to  $\lambda_0$ . The problem is that it will become infinitely long if  $|\lambda_0| > 1$  or infinitesimally short if  $|\lambda_0| < 1$ . All is good if  $|\lambda_0| = 1$ .

### 14.1.2 Second attempt

Again, let  $v^{(0)} \in \mathbb{C}^{m \times m}$  be an “initial guess”. The second attempt at the Power Method iterates as follows:

```
for  $k = 0, \dots$ 
     $v^{(k+1)} = Av^{(k)} / \lambda_0$ 
endfor
```

It is not hard to see that then

$$\begin{aligned} v^{(k)} &= Av^{(k-1)} / \lambda_0 = A^k v^{(0)} / \lambda_0^k \\ &= \psi_0 \left( \frac{\lambda_0}{\lambda_0} \right)^k x_0 + \psi_1 \left( \frac{\lambda_1}{\lambda_0} \right)^k x_1 + \dots + \psi_{m-1} \left( \frac{\lambda_{m-1}}{\lambda_0} \right)^k x_{m-1} \\ &= \psi_0 x_0 + \psi_1 \left( \frac{\lambda_1}{\lambda_0} \right)^k x_1 + \dots + \psi_{m-1} \left( \frac{\lambda_{m-1}}{\lambda_0} \right)^k x_{m-1}. \end{aligned}$$

Clearly  $\lim_{k \rightarrow \infty} v^{(k)} = \psi_0 x_0$ , as long as  $\psi_0 \neq 0$ , since  $\left| \frac{\lambda_k}{\lambda_0} \right| < 1$  for  $k > 0$ .

Another way of stating this is to notice that

$$A^k = \underbrace{(AA \cdots A)}_{k \text{ times}} = \underbrace{(X\Lambda X^{-1})(X\Lambda X^{-1}) \cdots (X\Lambda X^{-1})}_{\Lambda^k} = X\Lambda^k X^{-1}.$$

so that

$$\begin{aligned} v^{(k)} &= A^k v^{(0)} / \lambda_0^k \\ &= A^k X y / \lambda_0^k \\ &= X \Lambda^k X^{-1} X y / \lambda_0^k \\ &= X \Lambda^k y / \lambda_0^k \\ &= X \left( \Lambda^k / \lambda_0^k \right) y = X \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \left( \frac{\lambda_1}{\lambda_0} \right)^k & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \left( \frac{\lambda_{m-1}}{\lambda_0} \right)^k \end{pmatrix} y. \end{aligned}$$

Now, since  $\left| \frac{\lambda_k}{\lambda_0} \right| < 1$  for  $k > 1$  we can argue that

$$\begin{aligned} \lim_{k \rightarrow \infty} v^{(k)} &= \lim_{k \rightarrow \infty} X \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \left(\frac{\lambda_1}{\lambda_0}\right)^k & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\lambda_{m-1}}{\lambda_0}\right)^k \end{pmatrix} y = X \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} y \\ &= X\psi_0 e_0 = \psi_0 X e_0 = \psi_0 x_0. \end{aligned}$$

Thus, as long as  $\psi_0 \neq 0$  (which means  $v$  must have a component in the direction of  $x_0$ ) this method will eventually yield a vector in the direction of  $x_0$ . However, this time the problem is that we don't know  $\lambda_0$  when we start.

### 14.1.3 Convergence

Before we make the algorithm practical, let us examine how fast the iteration converges. This requires a few definitions regarding rates of convergence.

**Definition 14.1** Let  $\alpha_0, \alpha_1, \alpha_2, \dots \in \mathbb{C}$  be an infinite sequence of scalars. Then  $\alpha_k$  is said to converge to  $\alpha$  if

$$\lim_{k \rightarrow \infty} |\alpha_k - \alpha| = 0.$$

Let  $x_0, x_1, x_2, \dots \in \mathbb{C}^m$  be an infinite sequence of vectors. Then  $x_k$  is said to converge to  $x$  in the  $\|\cdot\|$  norm if

$$\lim_{k \rightarrow \infty} \|x_k - x\| = 0.$$

Notice that because of the equivalence of norms, if the sequence converges in one norm, it converges in all norms.

**Definition 14.2** Let  $\alpha_0, \alpha_1, \alpha_2, \dots \in \mathbb{C}$  be an infinite sequence of scalars that converges to  $\alpha$ . Then

- $\alpha_k$  is said to converge linearly to  $\alpha$  if for large enough  $k$

$$|\alpha_{k+1} - \alpha| \leq C|\alpha_k - \alpha|$$

for some constant  $C < 1$ .

- $\alpha_k$  is said to converge super-linearly to  $\alpha$  if

$$|\alpha_{k+1} - \alpha| \leq C_k |\alpha_k - \alpha|$$

with  $C_k \rightarrow 0$ .

- $\alpha_k$  is said to converge quadratically to  $\alpha$  if for large enough  $k$

$$|\alpha_{k+1} - \alpha| \leq C |\alpha_k - \alpha|^2$$

for some constant  $C$ .

- $\alpha_k$  is said to converge super-quadratically to  $\alpha$  if

$$|\alpha_{k+1} - \alpha| \leq C_k |\alpha_k - \alpha|^2$$

with  $C_k \rightarrow 0$ .

- $\alpha_k$  is said to converge cubically to  $\alpha$  if for large enough  $k$

$$|\alpha_{k+1} - \alpha| \leq C |\alpha_k - \alpha|^3$$

for some constant  $C$ .

Linear convergence can be slow. Let's say that for  $k \geq K$  we observe that

$$|\alpha_{k+1} - \alpha| \leq C |\alpha_k - \alpha|.$$

Then, clearly,  $|\alpha_{k+n} - \alpha| \leq C^n |\alpha_k - \alpha|$ . If  $C = 0.99$ , progress may be very, very slow. If  $|\alpha_k - \alpha| = 1$ , then

$$\begin{aligned} |\alpha_{k+1} - \alpha| &\leq 0.99000 \\ |\alpha_{k+2} - \alpha| &\leq 0.98010 \\ |\alpha_{k+3} - \alpha| &\leq 0.97030 \\ |\alpha_{k+4} - \alpha| &\leq 0.96060 \\ |\alpha_{k+5} - \alpha| &\leq 0.95099 \\ |\alpha_{k+6} - \alpha| &\leq 0.94148 \\ |\alpha_{k+7} - \alpha| &\leq 0.93206 \\ |\alpha_{k+8} - \alpha| &\leq 0.92274 \\ |\alpha_{k+9} - \alpha| &\leq 0.91351 \end{aligned}$$

Quadratic convergence is fast. Now

$$\begin{aligned} |\alpha_{k+1} - \alpha| &\leq C |\alpha_k - \alpha|^2 \\ |\alpha_{k+2} - \alpha| &\leq C |\alpha_{k+1} - \alpha|^2 \leq C (C |\alpha_k - \alpha|^2)^2 = C^3 |\alpha_k - \alpha|^4 \\ |\alpha_{k+3} - \alpha| &\leq C |\alpha_{k+2} - \alpha|^2 \leq C (C^3 |\alpha_k - \alpha|^4)^2 = C^7 |\alpha_k - \alpha|^8 \\ &\vdots \\ |\alpha_{k+n} - \alpha| &\leq C^{2^n - 1} |\alpha_k - \alpha|^{2^n} \end{aligned}$$

Even  $C = 0.99$  and  $|\alpha_k - \alpha| = 1$ , then

$$\begin{aligned} |\alpha_{k+1} - \alpha| &\leq 0.99000 \\ |\alpha_{k+2} - \alpha| &\leq 0.970299 \\ |\alpha_{k+3} - \alpha| &\leq 0.932065 \\ |\alpha_{k+4} - \alpha| &\leq 0.860058 \\ |\alpha_{k+5} - \alpha| &\leq 0.732303 \\ |\alpha_{k+6} - \alpha| &\leq 0.530905 \\ |\alpha_{k+7} - \alpha| &\leq 0.279042 \\ |\alpha_{k+8} - \alpha| &\leq 0.077085 \\ |\alpha_{k+9} - \alpha| &\leq 0.005882 \\ |\alpha_{k+10} - \alpha| &\leq 0.000034 \end{aligned}$$

If we consider  $\alpha$  the correct result then, eventually, the number of correct digits roughly doubles in each iteration. This can be explained as follows: If  $|\alpha_k - \alpha| < 1$ , then the number of correct decimal digits is given by

$$-\log_{10} |\alpha_k - \alpha|.$$

Since  $\log_{10}$  is a monotonically increasing function,

$$\log_{10} |\alpha_{k+1} - \alpha| \leq \log_{10} C |\alpha_k - \alpha|^2 = \log_{10}(C) + 2 \log_{10} |\alpha_k - \alpha| \leq 2 \log_{10} (|\alpha_k - \alpha|)$$

and hence

$$\underbrace{-\log_{10} |\alpha_{k+1} - \alpha|}_{\substack{\text{number of correct} \\ \text{digits in } \alpha_{k+1}}} \geq 2 \left( \underbrace{-\log_{10} (|\alpha_k - \alpha|)}_{\substack{\text{number of correct} \\ \text{digits in } \alpha_k}} \right).$$

Cubic convergence is dizzyingly fast: Eventually the number of correct digits triples from one iteration to the next.

We now define a convenient norm.

**Lemma 14.3** *Let  $X \in \mathbb{C}^{m \times m}$  be nonsingular. Define  $\|\cdot\|_X : \mathbb{C}^m \rightarrow \mathbb{R}$  by  $\|y\|_X = \|Xy\|$  for some given norm  $\|\cdot\| : \mathbb{C}^m \rightarrow \mathbb{R}$ . Then  $\|\cdot\|_X$  is a norm.*

**Homework 14.4** *Prove Lemma 14.3.*

 [SEE ANSWER](#)

With this new norm, we can do our convergence analysis:

$$\begin{aligned} v^{(k)} - \psi_0 x_0 &= A^k v^{(0)} / \lambda_0^k - \psi_0 x_0 = X \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \left(\frac{\lambda_1}{\lambda_0}\right)^k & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\lambda_{m-1}}{\lambda_0}\right)^k \end{pmatrix} X^{-1} v^{(0)} - \psi_0 x_0 \\ &= X \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \left(\frac{\lambda_1}{\lambda_0}\right)^k & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\lambda_{m-1}}{\lambda_0}\right)^k \end{pmatrix} y - \psi_0 x_0 = X \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \left(\frac{\lambda_1}{\lambda_0}\right)^k & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\lambda_{m-1}}{\lambda_0}\right)^k \end{pmatrix} y \end{aligned}$$

Hence

$$X^{-1}(v^{(k)} - \psi_0 x_0) = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \left(\frac{\lambda_1}{\lambda_0}\right)^k & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\lambda_{m-1}}{\lambda_0}\right)^k \end{pmatrix} y$$

and

$$X^{-1}(v^{(k+1)} - \psi_0 x_0) = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \frac{\lambda_1}{\lambda_0} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\lambda_{m-1}}{\lambda_0} \end{pmatrix} X^{-1}(v^{(k)} - \psi_0 x_0).$$

Now, let  $\|\cdot\|$  be a p-norm<sup>1</sup> and its induced matrix norm and  $\|\cdot\|_{X^{-1}}$  as defined in Lemma 14.3. Then

$$\begin{aligned} \|v^{(k+1)} - \psi_0 x_0\|_{X^{-1}} &= \|X^{-1}(v^{(k+1)} - \psi_0 x_0)\| \\ &= \left\| \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \frac{\lambda_1}{\lambda_0} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\lambda_{m-1}}{\lambda_0} \end{pmatrix} X^{-1}(v^{(k)} - \psi_0 x_0) \right\| \\ &\leq \left| \frac{\lambda_1}{\lambda_0} \right| \|X^{-1}(v^{(k)} - \psi_0 x_0)\| = \left| \frac{\lambda_1}{\lambda_0} \right| \|v^{(k)} - \psi_0 x_0\|_{X^{-1}}. \end{aligned}$$

This shows that, in this norm, the convergence of  $v^{(k)}$  to  $\psi_0 x_0$  is linear: The difference between current approximation,  $v^{(k)}$ , and the solution,  $\psi_0 x_0$ , is reduced by at least a constant factor in each iteration.

#### 14.1.4 Practical Power Method

The following algorithm, known as the Power Method, avoids the problem of  $v^{(k)}$  growing or shrinking in length, without requiring  $\lambda_0$  to be known, by scaling it to be of unit length at each step:

```
for  $k = 0, \dots$ 
   $v^{(k+1)} = Av^{(k)}$ 
   $v^{(k+1)} = v^{(k+1)} / \|v^{(k+1)}\|$ 
endfor
```

#### 14.1.5 The Rayleigh quotient

A question is how to extract an approximation of  $\lambda_0$  given an approximation of  $x_0$ . The following theorem provides the answer:

**Theorem 14.5** *If  $x$  is an eigenvector of  $A$  then  $\lambda = x^H A x / (x^H x)$  is the associated eigenvalue of  $A$ . This ratio is known as the Rayleigh quotient.*

**Proof:** Let  $x$  be an eigenvector of  $A$  and  $\lambda$  the associated eigenvalue. Then  $Ax = \lambda x$ . Multiplying on the left by  $x^H$  yields  $x^H A x = \lambda x^H x$  which, since  $x \neq 0$  means that  $\lambda = x^H A x / (x^H x)$ .

Clearly this ratio as a function of  $x$  is continuous and hence an approximation to  $x_0$  when plugged into this formula would yield an approximation to  $\lambda_0$ .

<sup>1</sup>We choose a p-norm to make sure that the norm of a diagonal matrix equals the absolute value of the largest element (in magnitude) on its diagonal.

### 14.1.6 What if $|\lambda_0| \geq |\lambda_1|$ ?

Now, what if

$$|\lambda_0| = \dots = |\lambda_{k-1}| > |\lambda_k| \geq \dots \geq |\lambda_{m-1}|?$$

By extending the above analysis one can easily show that  $v^{(k)}$  will converge to a vector in the subspace spanned by the eigenvectors associated with  $\lambda_0, \dots, \lambda_{k-1}$ .

An important special case is when  $k = 2$ : if  $A$  is real valued then  $\lambda_0$  still may be complex valued in which case  $\bar{\lambda}_0$  is also an eigenvalue and it has the same magnitude as  $\lambda_0$ . We deduce that  $v^{(k)}$  will always be in the space spanned by the eigenvectors corresponding to  $\lambda_0$  and  $\bar{\lambda}_0$ .

## 14.2 The Inverse Power Method

The Power Method homes in on an eigenvector associated with the largest (in magnitude) eigenvalue. The Inverse Power Method homes in on an eigenvector associated with the smallest eigenvalue (in magnitude).

Throughout this section we will assume that a given matrix  $A \in \mathbb{C}^{m \times m}$  is *nondeficient* and nonsingular so that there exist matrix  $X$  and diagonal matrix  $\Lambda$  such that  $A = X\Lambda X^{-1}$ . We further assume that  $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{m-1})$  and

$$|\lambda_0| \geq |\lambda_1| \geq \dots \geq |\lambda_{m-2}| > |\lambda_{m-1}|.$$

**Theorem 14.6** *Let  $A \in \mathbb{C}^{m \times m}$  be nonsingular. Then  $\lambda$  and  $x$  are an eigenvalue and associated eigenvector of  $A$  if and only if  $1/\lambda$  and  $x$  are an eigenvalue and associated eigenvector of  $A^{-1}$ .*

**Homework 14.7** Assume that

$$|\lambda_0| \geq |\lambda_1| \geq \dots \geq |\lambda_{m-2}| > |\lambda_{m-1}| > 0.$$

Show that

$$\left| \frac{1}{\lambda_{m-1}} \right| > \left| \frac{1}{\lambda_{m-2}} \right| \geq \left| \frac{1}{\lambda_{m-3}} \right| \geq \dots \geq \left| \frac{1}{\lambda_0} \right|.$$

 [SEE ANSWER](#)

Thus, an eigenvector associated with the smallest (in magnitude) eigenvalue of  $A$  is an eigenvector associated with the largest (in magnitude) eigenvalue of  $A^{-1}$ . This suggests the following naive iteration:

```
for  $k = 0, \dots$ 
   $v^{(k+1)} = A^{-1}v^{(k)}$ 
   $\lambda_{m-1}^{(k+1)} = \lambda_{m-1}v^{(k+1)}$ 
endfor
```

Of course, we would want to factor  $A = LU$  once and solve  $L(Uv^{(k+1)}) = v^{(k)}$  rather than multiplying with  $A^{-1}$ . From the analysis of the convergence of the “second attempt” for a Power Method algorithm we conclude that now

$$\|v^{(k+1)} - \Psi_{m-1}x_{m-1}\|_{X^{-1}} \leq \left| \frac{\lambda_{m-1}}{\lambda_{m-2}} \right| \|v^{(k)} - \Psi_{m-1}x_{m-1}\|_{X^{-1}}.$$

A practical Inverse Power Method algorithm is given by

```

for  $k = 0, \dots$ 
   $v^{(k+1)} = A^{-1}v^{(k)}$ 
   $v^{(k+1)} = v^{(k+1)} / \|v^{(k+1)}\|$ 
endfor

```

Often, we would expect the Invert Power Method to converge faster than the Power Method. For example, take the case where  $|\lambda_k|$  are equally spaced between 0 and  $m$ :  $|\lambda_k| = (k+1)$ . Then

$$\left| \frac{\lambda_1}{\lambda_0} \right| = \frac{m-1}{m} \quad \text{and} \quad \left| \frac{\lambda_{m-1}}{\lambda_{m-2}} \right| = \frac{1}{2}.$$

which means that the Power Method converges much more slowly than the Inverse Power Method.

### 14.3 Rayleigh-quotient Iteration

The next observation is captured in the following lemma:

**Lemma 14.8** *Let  $A \in \mathbb{C}^{m \times m}$  and  $\mu \in \mathbb{C}$ . Then  $(\lambda, x)$  is an eigenpair of  $A$  if and only if  $(\lambda - \mu, x)$  is an eigenpair of  $(A - \mu I)$ .*

**Homework 14.9** *Prove Lemma 14.8.*

 [SEE ANSWER](#)

The matrix  $A - \mu I$  is referred to as the matrix  $A$  that has been “shifted” by  $\mu$ . What the lemma says is that shifting  $A$  by  $\mu$  shifts the spectrum of  $A$  by  $\mu$ :

**Lemma 14.10** *Let  $A \in \mathbb{C}^{m \times m}$ ,  $A = X\Lambda X^{-1}$  and  $\mu \in \mathbb{C}$ . Then  $A - \mu I = X(\Lambda - \mu I)X^{-1}$ .*

**Homework 14.11** *Prove Lemma 14.10.*

 [SEE ANSWER](#)

This suggests the following (naive) iteration: Pick a value  $\mu$  close to  $\lambda_{m-1}$ . Iterate

```

for  $k = 0, \dots$ 
   $v^{(k+1)} = (A - \mu I)^{-1}v^{(k)}$ 
   $v^{(k+1)} = (\lambda_{m-1} - \mu)v^{(k+1)}$ 
endfor

```

Of course one would solve  $(A - \mu I)v^{(k+1)} = v^{(k)}$  rather than computing and applying the inverse of  $A$ .

If we index the eigenvalues so that  $|\lambda_0 - \mu| \leq \dots \leq |\lambda_{m-2} - \mu| < |\lambda_{m-1} - \mu|$  then

$$\|v^{(k+1)} - \Psi_{m-1}x_{m-1}\|_{X^{-1}} \leq \left| \frac{\lambda_{m-1} - \mu}{\lambda_{m-2} - \mu} \right| \|v^{(k)} - \Psi_{m-1}x_{m-1}\|_{X^{-1}}.$$

The closer to  $\lambda_{m-1}$  the “shift” (so named because it shifts the spectrum of  $A$ ) is chosen, the more favorable the ratio that dictates convergence.

A more practical algorithm is given by



```

for  $k = 0, \dots$ 
   $v^{(k+1)} = (A - \mu I)^{-1} v^{(k)}$ 
   $v^{(k+1)} = v^{(k+1)} / \|v^{(k+1)}\|$ 
endfor

```

The question now becomes how to choose  $\mu$  so that it is a good guess for  $\lambda_{m-1}$ . Often an application inherently supplies a reasonable approximation for the smallest eigenvalue or an eigenvalue of particular interest. However, we know that eventually  $v^{(k)}$  becomes a good approximation for  $x_{m-1}$  and therefore the Rayleigh quotient gives us a way to find a good approximation for  $\lambda_{m-1}$ . This suggests the (naive) Rayleigh-quotient iteration:

```

for  $k = 0, \dots$ 
   $\mu_k = v^{(k)H} A v^{(k)} / (v^{(k)H} v^{(k)})$ 
   $v^{(k+1)} = (A - \mu_k I)^{-1} v^{(k)}$ 
   $v^{(k+1)} = (\lambda_{m-1} - \mu_k) v^{(k+1)}$ 
endfor

```

Now<sup>2</sup>

$$\|v^{(k+1)} - \Psi_{m-1} x_{m-1}\|_{X^{-1}} \leq \left| \frac{\lambda_{m-1} - \mu_k}{\lambda_{m-2} - \mu_k} \right| \|v^{(k)} - \Psi_{m-1} x_{m-1}\|_{X^{-1}}$$

with

$$\lim_{k \rightarrow \infty} (\lambda_{m-1} - \mu_k) = 0$$

which means *super linear* convergence is observed. In fact, it can be shown that once  $k$  is large enough

$$\|v^{(k+1)} - \Psi_{m-1} x_{m-1}\|_{X^{-1}} \leq C \|v^{(k)} - \Psi_{m-1} x_{m-1}\|_{X^{-1}}^2,$$

which is known as quadratic convergence. Roughly speaking this means that every iteration doubles the number of correct digits in the current approximation. To prove this, one shows that  $|\lambda_{m-1} - \mu_k| \leq C \|v^{(k)} - \Psi_{m-1} x_{m-1}\|_{X^{-1}}$ .

Better yet, it can be shown that if  $A$  is Hermitian, then, once  $k$  is large enough,

$$\|v^{(k+1)} - \Psi_{m-1} x_{m-1}\|_{X^{-1}} \leq C \|v^{(k)} - \Psi_{m-1} x_{m-1}\|_{X^{-1}}^3,$$

which is known as cubic convergence. Roughly speaking this means that every iteration triples the number of correct digits in the current approximation. This is mind-boggling fast convergence!

A practical Rayleigh quotient iteration is given by

```

 $v^{(0)} = v^{(0)} / \|v^{(0)}\|_2$ 
for  $k = 0, \dots$ 
   $\mu_k = v^{(k)H} A v^{(k)}$  (Now  $\|v^{(k)}\|_2 = 1$ )
   $v^{(k+1)} = (A - \mu_k I)^{-1} v^{(k)}$ 
   $v^{(k+1)} = v^{(k+1)} / \|v^{(k+1)}\|$ 
endfor

```

---

<sup>2</sup> I think... I have not checked this thoroughly. But the general idea holds.  $\lambda_{m-1}$  has to be defined as the eigenvalue to which the method eventually converges.



# Chapter 15

## Notes on the QR Algorithm and other Dense Eigensolvers

### Video

Read disclaimer regarding the videos in the preface!

Tragically, the camera ran out of memory for the first lecture... Here is the second lecture, which discusses the implicit QR algorithm

 [YouTube](#)

 [Download from UT Box](#)

 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

In most of this ote, we focus on the case where  $A$  is symmetric and real valued. The reason for this is that many of the techniques can be more easily understood in that setting.

## Outline

<b>Video</b>	<b>235</b>
<b>Outline</b>	<b>236</b>
<b>15.1. Preliminaries</b>	<b>237</b>
<b>15.2. Subspace Iteration</b>	<b>237</b>
<b>15.3. The QR Algorithm</b>	<b>242</b>
15.3.1. A basic (unshifted) QR algorithm	242
15.3.2. A basic shifted QR algorithm	242
<b>15.4. Reduction to Tridiagonal Form</b>	<b>244</b>
15.4.1. Householder transformations (reflectors)	244
15.4.2. Algorithm	245
<b>15.5. The QR algorithm with a Tridiagonal Matrix</b>	<b>247</b>
15.5.1. Givens' rotations	247
<b>15.6. QR Factorization of a Tridiagonal Matrix</b>	<b>248</b>
<b>15.7. The Implicitly Shifted QR Algorithm</b>	<b>250</b>
15.7.1. Upper Hessenberg and tridiagonal matrices	250
15.7.2. The Implicit Q Theorem	251
15.7.3. The Francis QR Step	252
15.7.4. A complete algorithm	254
<b>15.8. Further Reading</b>	<b>258</b>
15.8.1. More on reduction to tridiagonal form	258
15.8.2. Optimizing the tridiagonal QR algorithm	258
<b>15.9. Other Algorithms</b>	<b>258</b>
15.9.1. Jacobi's method for the symmetric eigenvalue problem	258
15.9.2. Cuppen's Algorithm	261
15.9.3. The Method of Multiple Relatively Robust Representations (MRRR)	261
<b>15.10. The Nonsymmetric QR Algorithm</b>	<b>261</b>
15.10.1. A variant of the Schur decomposition	261
15.10.2. Reduction to upper Hessenberg form	262
15.10.3. The implicitly double-shifted QR algorithm	265

## 15.1 Preliminaries

The QR algorithm is a standard method for computing all eigenvalues and eigenvectors of a matrix. In this note, we focus on the real valued symmetric eigenvalue problem (the case where  $A \in \mathbb{R}^{n \times n}$ ). For this case, recall the Spectral Decomposition Theorem:

**Theorem 15.1** *If  $A \in \mathbb{R}^{n \times n}$  then there exists unitary matrix  $Q$  and diagonal matrix  $\Lambda$  such that  $A = Q\Lambda Q^T$ .*

We will partition  $Q = \left( q_0 \mid \cdots \mid q_{n-1} \right)$  and assume that  $\Lambda = \text{diag}((\lambda_0, \dots, \lambda_{n-1}))$  so that throughout this note,  $q_i$  and  $\lambda_i$  refer to the  $i$ th column of  $Q$  and the  $i$ th diagonal element of  $\Lambda$ , which means that each tuple  $(\lambda, q_i)$  is an eigenpair.

## 15.2 Subspace Iteration

We start with a matrix  $V \in \mathbb{R}^{n \times r}$  with normalized columns and iterate something like

```

 $V^{(0)} = V$ 
for  $k = 0, \dots$  convergence
     $V^{(k+1)} = AV^{(k)}$ 
    Normalize the columns to be of unit length.
end for

```

The problem with this approach is that all columns will (likely) converge to an eigenvector associated with the dominant eigenvalue, since the Power Method is being applied to all columns simultaneously. We will now lead the reader through a succession of insights towards a practical algorithm.

Let us examine what  $\widehat{V} = AV$  looks like, for the simple case where  $V = \left( v_0 \mid v_1 \mid v_2 \right)$  (three columns). We know that

$$v_j = Q \underbrace{Q^T v_j}_{y_j}.$$

Hence

$$\begin{aligned} v_0 &= \sum_{j=0}^{n-1} \psi_{0,j} q_j, \\ v_1 &= \sum_{j=0}^{n-1} \psi_{1,j} q_j, \text{ and} \\ v_2 &= \sum_{j=0}^{n-1} \psi_{2,j} q_j, \end{aligned}$$

where  $\psi_{i,j}$  equals the  $i$ th element of  $y_j$ . Then

$$\begin{aligned}
 AV &= A \left( v_0 \mid v_1 \mid v_2 \right) \\
 &= A \left( \sum_{j=0}^{n-1} \psi_{0,j} q_j \mid \sum_{j=0}^{n-1} \psi_{1,j} q_j \mid \sum_{j=0}^{n-1} \psi_{2,j} q_j \right) \\
 &= \left( \sum_{j=0}^{n-1} \psi_{0,j} A q_j \mid \sum_{j=0}^{n-1} \psi_{1,j} A q_j \mid \sum_{j=0}^{n-1} \psi_{2,j} A q_j \right) \\
 &= \left( \sum_{j=0}^{n-1} \psi_{0,j} \lambda_j q_j \mid \sum_{j=0}^{n-1} \psi_{1,j} \lambda_j q_j \mid \sum_{j=0}^{n-1} \psi_{2,j} \lambda_j q_j \right)
 \end{aligned}$$

- If we happened to know  $\lambda_0$ ,  $\lambda_1$ , and  $\lambda_2$  then we could divide the columns by these, respectively, and get new vectors

$$\begin{aligned}
 \left( \hat{v}_0 \mid \hat{v}_1 \mid \hat{v}_2 \right) &= \left( \sum_{j=0}^{n-1} \psi_{0,j} \left( \frac{\lambda_j}{\lambda_0} \right) q_j \mid \sum_{j=0}^{n-1} \psi_{1,j} \left( \frac{\lambda_j}{\lambda_1} \right) q_j \mid \sum_{j=0}^{n-1} \psi_{2,j} \left( \frac{\lambda_j}{\lambda_2} \right) q_j \right) \\
 &= \left( \begin{array}{c|c|c} \psi_{0,0} q_0 + & \psi_{1,0} \left( \frac{\lambda_0}{\lambda_1} \right) q_0 + & \psi_{2,0} \left( \frac{\lambda_0}{\lambda_2} \right) q_0 + \psi_{2,1} \left( \frac{\lambda_1}{\lambda_2} \right) q_1 + \\ \sum_{j=1}^{n-1} \psi_{0,j} \left( \frac{\lambda_j}{\lambda_0} \right) q_j & \sum_{j=2}^{n-1} \psi_{1,j} \left( \frac{\lambda_j}{\lambda_1} \right) q_j & \sum_{j=3}^{n-1} \psi_{2,j} \left( \frac{\lambda_j}{\lambda_2} \right) q_j \end{array} \right) \quad (15.1)
 \end{aligned}$$

- Assume that  $|\lambda_0| > |\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_{n-1}|$ . Then, similar as for the power method,
  - The first column will see components in the direction of  $\{q_1, \dots, q_{n-1}\}$  shrink relative to the component in the direction of  $q_0$ .
  - The second column will see components in the direction of  $\{q_2, \dots, q_{n-1}\}$  shrink relative to the component in the direction of  $q_1$ , but the component in the direction of  $q_0$  increases, relatively, since  $|\lambda_0/\lambda_1| > 1$ .
  - The third column will see components in the direction of  $\{q_3, \dots, q_{n-1}\}$  shrink relative to the component in the direction of  $q_2$ , but the components in the directions of  $q_0$  and  $q_1$  increase, relatively, since  $|\lambda_0/\lambda_2| > 1$  and  $|\lambda_1/\lambda_2| > 1$ .

How can we make it so that  $v_j$  converges to a vector in the direction of  $q_j$ ?

- If we happen to know  $q_0$ , then we can subtract out the component of

$$\hat{v}_1 = \psi_{1,0} \frac{\lambda_0}{\lambda_1} q_0 + \psi_{1,1} q_1 + \sum_{j=2}^{n-1} \psi_{1,j} \frac{\lambda_j}{\lambda_1} q_j$$

in the direction of  $q_0$ :

$$\hat{v}_1 - q_0^T \hat{v}_1 q_0 = \psi_{1,1} q_1 + \sum_{j=2}^{n-1} \psi_{1,j} \frac{\lambda_j}{\lambda_1} q_j$$

so that we are left with the component in the direction of  $q_1$  and components in directions of  $q_2, \dots, q_{n-1}$  that are suppressed every time through the loop.

- Similarly, if we also know  $q_1$ , the components of  $\hat{v}_2$  in the direction of  $q_0$  and  $q_1$  can be subtracted from that vector.

- We do not know  $\lambda_0$ ,  $\lambda_1$ , and  $\lambda_2$  but from the discussion about the Power Method we remember that we can just normalize the so updated  $\hat{v}_0$ ,  $\hat{v}_1$ , and  $\hat{v}_2$  to have unit length.

How can we make these insights practical?

- We do not know  $q_0$ ,  $q_1$ , and  $q_2$ , but we can informally argue that if we keep iterating,
  - The vector  $\hat{v}_0$ , normalized in each step, will eventually point in the direction of  $q_0$ .
  - $\mathcal{S}(\hat{v}_0, \hat{v}_1)$  will eventually equal  $\mathcal{S}(q_0, q_1)$ .
  - In each iteration, we can subtract the component of  $\hat{v}_1$  in the direction of  $\hat{v}_0$  from  $\hat{v}_1$ , and then normalize  $\hat{v}_1$  so that eventually result in a the vector that points in the direction of  $q_1$ .
  - $\mathcal{S}(\hat{v}_0, \hat{v}_1, \hat{v}_2)$  will eventually equal  $\mathcal{S}(q_0, q_1, q_2)$ .
  - In each iteration, we can subtract the component of  $\hat{v}_2$  in the directions of  $\hat{v}_0$  and  $\hat{v}_1$  from  $\hat{v}_2$ , and then normalize the result, to make  $\hat{v}_2$  eventually point in the direction of  $q_2$ .

What we recognize is that normalizing  $\hat{v}_0$ , subtracting out the component of  $\hat{v}_1$  in the direction of  $\hat{v}_0$ , and then normalizing  $\hat{v}_1$ , etc., is exactly what the Gram-Schmidt process does. And thus, we can use any convenient (and stable) QR factorization method. This also shows how the method can be generalized to work with more than three columns and even all columns simultaneously.

The algorithm now becomes:

$$\begin{aligned}
 V^{(0)} &= I^{n \times p} \quad (I^{n \times p} \text{ represents the first } p \text{ columns of } I) \\
 &\text{for } k = 0, \dots \text{ convergence} \\
 &\quad AV^{(k)} \rightarrow V^{(k+1)}R^{(k+1)} \quad (\text{QR factorization with } R^{(k+1)} \in \mathbb{R}^{p \times p}) \\
 &\text{end for}
 \end{aligned}$$

Now consider again (15.1), focusing on the third column:

$$\begin{pmatrix} \Psi_{2,0} \left( \frac{\lambda_0}{\lambda_2} \right) q_0 + \Psi_{2,1} \left( \frac{\lambda_1}{\lambda_2} \right) q_1 + \\ \Psi_{2,2} q_2 + \\ \sum_{j=3}^{n-1} \Psi_j \left( \frac{\lambda_j}{\lambda_2} \right) q_j \end{pmatrix} = \begin{pmatrix} \Psi_{2,0} \left( \frac{\lambda_0}{\lambda_2} \right) q_0 + \Psi_{2,1} \left( \frac{\lambda_1}{\lambda_2} \right) q_1 + \\ \Psi_{2,2} q_2 + \\ \Psi_j \left( \frac{\lambda_3}{\lambda_2} \right) q_3 + \sum_{j=4}^{n-1} \Psi_{2,j} \left( \frac{\lambda_j}{\lambda_2} \right) q_j \end{pmatrix}.$$

This shows that, if the components in the direction of  $q_0$  and  $q_1$  are subtracted out, it is the component in the direction of  $q_3$  that is diminished in length the most slowly, dictated by the ratio  $\left| \frac{\lambda_3}{\lambda_2} \right|$ . This, of course, generalizes: the  $j$ th column of  $V^{(k)}$ ,  $v_j^{(k)}$  will have a component in the direction of  $q_{j+1}$ , of length  $|q_{j+1}^T v_j^{(k)}|$ , that can be expected to shrink most slowly.

We demonstrate this in Figure 15.1, which shows the execution of the algorithm with  $p = n$  for a  $5 \times 5$  matrix, and shows how  $|q_{j+1}^T v_j^{(k)}|$  converge to zero as as function of  $k$ .

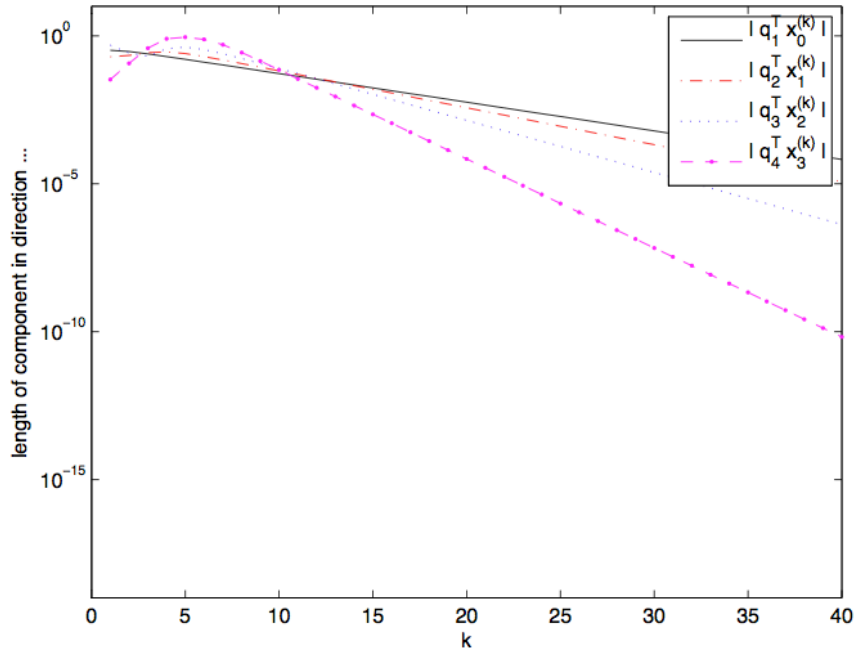


Figure 15.1: Convergence of the subspace iteration for a  $5 \times 5$  matrix. This graph is mislabeled:  $x$  should be labeled with  $v$ . The (linear) convergence of  $v_j$  to a vector in the direction of  $q_j$  is dictated by how quickly the component in the direction  $q_{j+1}$  converges to zero. The line labeled  $|q_{j+1}^T x_j|$  plots the length of the component in the direction  $q_{j+1}$  as a function of the iteration number.

Next, we observe that if  $V \in \mathbb{R}^{n \times n}$  in the above iteration (which means we are iterating with  $n$  vectors at a time), then  $AV$  yields a next-to last column of the form

$$\begin{pmatrix} \sum_{j=0}^{n-3} \psi_{n-2,j} \left( \frac{\lambda_j}{\lambda_{n-2}} \right) q_j + \\ \psi_{n-2,n-2} q_{n-2} + \\ \psi_{n-2,n-1} \left( \frac{\lambda_{n-1}}{\lambda_{n-2}} \right) q_{n-1} \end{pmatrix},$$

where  $\psi_{i,j} = q_j^T v_i$ . Thus, given that the components in the direction of  $q_j$ ,  $j = 0, \dots, n-2$  can be expected in later iterations to be greatly reduced by the QR factorization that subsequently happens with  $AV$ , we notice that it is  $\left| \frac{\lambda_{n-1}}{\lambda_{n-2}} \right|$  that dictates how fast the component in the direction of  $q_{n-1}$  disappears from  $v_{n-2}^{(k)}$ . This is a ratio we also saw in the Inverse Power Method and that we noticed we could accelerate in the Rayleigh Quotient Iteration: At each iteration we should shift the matrix to  $(A - \mu_k I)$  where  $\mu_k \approx \lambda_{n-1}$ . Since the last column of  $V^{(k)}$  is supposed to be converging to  $q_{n-1}$ , it seems reasonable to use  $\mu_k = v_{n-1}^{(k)T} A v_{n-1}^{(k)}$  (recall that  $v_{n-1}^{(k)}$  has unit length, so this is the Rayleigh quotient.)

The above discussion motivates the iteration



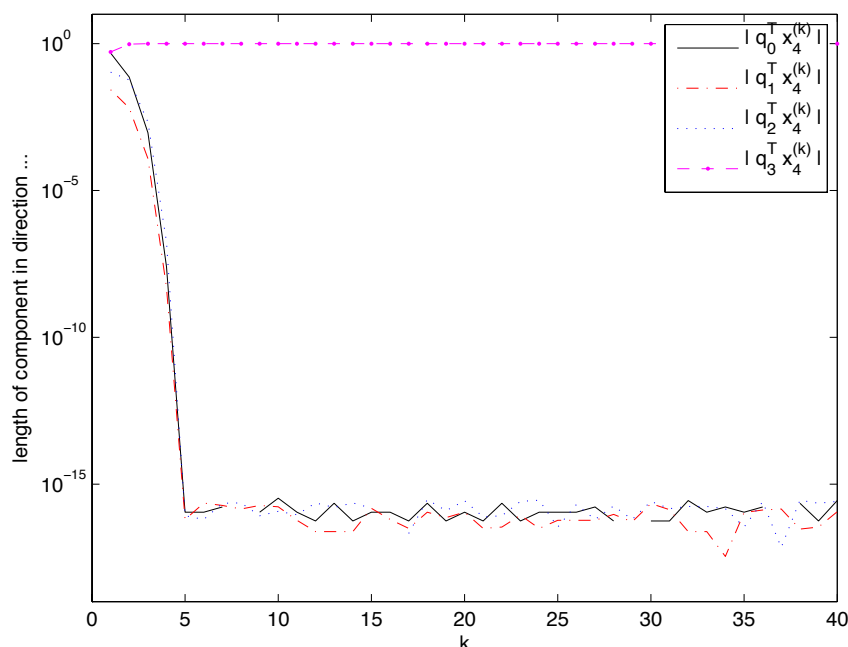


Figure 15.2: Convergence of the shifted subspace iteration for a  $5 \times 5$  matrix. This graph is mislabeled:  $x$  should be labeled with  $v$ . What this graph shows is that the components of  $v_4$  in the directions  $q_0$  through  $q_3$  disappear very quickly. The vector  $v_4$  quickly points in the direction of the eigenvector associated with the smallest (in magnitude) eigenvalue. Just like the Rayleigh-quotient iteration is not guaranteed to converge to the eigenvector associated with the smallest (in magnitude) eigenvalue, the shifted subspace iteration may home in on a different eigenvector than the one associated with the smallest (in magnitude) eigenvalue. **Something is wrong in this graph: All curves should quickly drop to (near) zero!**

```

 $V^{(0)} := I$  ( $V^{(0)} \in \mathbb{R}^{n \times n}$ !)
for  $k := 0, \dots$  convergence
     $\mu_k := v_{n-1}^{(k)T} A v_{n-1}^{(k)}$  (Rayleigh quotient)
     $(A - \mu_k I) V^{(k)} \rightarrow V^{(k+1)} R^{(k+1)}$  (QR factorization)
end for

```

Notice that this does *not* require one to solve with  $(A - \mu_k I)$ , unlike in the Rayleigh Quotient Iteration. However, it does require a QR factorization, which requires more computation than the LU factorization (approximately  $\frac{4}{3}n^3$  flops).

We demonstrate the convergence in Figure 15.2, which shows the execution of the algorithm with a  $5 \times 5$  matrix and illustrates how  $|q_j^T v_{n-1}^{(k)}|$  converge to zero as a function of  $k$ .

Subspace iteration	QR algorithm
$\hat{A}^{(0)} := A$	$A^{(0)} := A$
$\hat{V}^{(0)} := I$	$V^{(0)} := I$
for $k := 0, \dots$ until convergence	for $k := 0, \dots$ until convergence
$A\hat{V}^{(k)} \rightarrow \hat{V}^{(k+1)}\hat{R}^{(k+1)}$ (QR factorization)	$A^{(k)} \rightarrow Q^{(k+1)}R^{(k+1)}$ (QR factorization)
$\hat{A}^{(k+1)} := \hat{V}^{(k+1)T}A\hat{V}^{(k+1)}$	$A^{(k+1)} := R^{(k+1)}Q^{(k+1)}$
	$V^{(k+1)} := V^{(k)}Q^{(k+1)}$
end for	end for

Figure 15.3: Basic subspace iteration and basic QR algorithm.

## 15.3 The QR Algorithm

The QR algorithm is a classic algorithm for computing all eigenvalues and eigenvectors of a matrix. While we explain it for the symmetric eigenvalue problem, it generalizes to the nonsymmetric eigenvalue problem as well.

### 15.3.1 A basic (unshifted) QR algorithm

We have informally argued that the columns of the orthogonal matrices  $V^{(k)} \in \mathbb{R}^{n \times n}$  generated by the (unshifted) subspace iteration converge to eigenvectors of matrix  $A$ . (The exact conditions under which this happens have not been fully discussed.) In Figure 15.3 (left), we restate the subspace iteration. In it, we denote matrices  $V^{(k)}$  and  $R^{(k)}$  from the subspace iteration by  $\hat{V}^{(k)}$  and  $\hat{R}$  to distinguish them from the ones computed by the algorithm on the right. The algorithm on the left also computes the matrix  $\hat{A}^{(k)} = V^{(k)T}AV^{(k)}$ , a matrix that hopefully converges to  $\Lambda$ , the diagonal matrix with the eigenvalues of  $A$  on its diagonal. To the right is the QR algorithm. The claim is that the two algorithms compute the same quantities.

**Homework 15.2** Prove that in Figure 15.3,  $\hat{V}^{(k)} = V^{(k)}$ , and  $\hat{A}^{(k)} = A^{(k)}$ ,  $k = 0, \dots$

🔗 [SEE ANSWER](#)

We conclude that if  $\hat{V}^{(k)}$  converges to the matrix of orthonormal eigenvectors when the subspace iteration is applied to  $V^{(0)} = I$ , then  $A^{(k)}$  converges to the diagonal matrix with eigenvalues along the diagonal.

### 15.3.2 A basic shifted QR algorithm

In Figure 15.4 (left), we restate the subspace iteration with shifting. In it, we denote matrices  $V^{(k)}$  and  $R^{(k)}$  from the subspace iteration by  $\hat{V}^{(k)}$  and  $\hat{R}$  to distinguish them from the ones computed by the algorithm on the right. The algorithm on the left also computes the matrix  $\hat{A}^{(k)} = V^{(k)T}AV^{(k)}$ , a matrix that hopefully converges to  $\Lambda$ , the diagonal matrix with the eigenvalues of  $A$  on its diagonal. To the right is the shifted QR algorithm. The claim is that the two algorithms compute the same quantities.

Subspace iteration	QR algorithm
$\hat{A}^{(0)} := A$	$A^{(0)} := A$
$\hat{V}^{(0)} := I$	$V^{(0)} := I$
for $k := 0, \dots$ until convergence	for $k := 0, \dots$ until convergence
$\hat{\mu}_k := \hat{v}_{n-1}^{(k)T} A \hat{v}_{n-1}^{(k)}$	$\mu_k = \alpha_{n-1,n-1}^{(k)}$
$(A - \hat{\mu}_k I) \hat{V}^{(k)} \rightarrow \hat{V}^{(k+1)} \hat{R}^{(k+1)}$ (QR factorization)	$A^{(k)} - \mu_k I \rightarrow Q^{(k+1)} R^{(k+1)}$ (QR factorization)
$\hat{A}^{(k+1)} := \hat{V}^{(k+1)T} A \hat{V}^{(k+1)}$	$A^{(k+1)} := R^{(k+1)} Q^{(k+1)} + \mu_k I$
	$V^{(k+1)} := V^{(k)} Q^{(k+1)}$
end for	end for

Figure 15.4: Basic shifted subspace iteration and basic shifted QR algorithm.

**Homework 15.3** Prove that in Figure 15.4,  $\hat{V}^{(k)} = V^{(k)}$ , and  $\hat{A}^{(k)} = A^{(k)}$ ,  $k = 1, \dots$

SEE ANSWER

We conclude that if  $\hat{V}^{(k)}$  converges to the matrix of orthonormal eigenvectors when the shifted subspace iteration is applied to  $V^{(0)} = I$ , then  $A^{(k)}$  converges to the diagonal matrix with eigenvalues along the diagonal.

The convergence of the basic shifted QR algorithm is illustrated below. Pay particular attention to the convergence of the last row and column.

$$\begin{aligned}
 A^{(0)} &= \begin{pmatrix} 2.01131953448 & 0.05992695085 & 0.14820940917 \\ 0.05992695085 & 2.30708673171 & 0.93623515213 \\ 0.14820940917 & 0.93623515213 & 1.68159373379 \end{pmatrix} & A^{(1)} &= \begin{pmatrix} 2.21466116574 & 0.34213192482 & 0.31816754245 \\ 0.34213192482 & 2.54202325042 & 0.57052186467 \\ 0.31816754245 & 0.57052186467 & 1.24331558383 \end{pmatrix} \\
 A^{(2)} &= \begin{pmatrix} 2.63492207667 & 0.47798481637 & 0.07654607908 \\ 0.47798481637 & 2.35970859985 & 0.06905042811 \\ 0.07654607908 & 0.06905042811 & 1.00536932347 \end{pmatrix} & A^{(3)} &= \begin{pmatrix} 2.87588550968 & 0.32971207176 & 0.00024210487 \\ 0.32971207176 & 2.12411444949 & 0.00014361630 \\ 0.00024210487 & 0.00014361630 & 1.00000004082 \end{pmatrix} \\
 A^{(4)} &= \begin{pmatrix} 2.96578660126 & 0.18177690194 & 0.00000000000 \\ 0.18177690194 & 2.03421339873 & 0.00000000000 \\ 0.00000000000 & 0.00000000000 & 1.00000000000 \end{pmatrix} & A^{(5)} &= \begin{pmatrix} 2.9912213907 & 0.093282073553 & 0.00000000000 \\ 0.0932820735 & 2.008778609226 & 0.00000000000 \\ 0.00000000000 & 0.00000000000 & 1.00000000000 \end{pmatrix}
 \end{aligned}$$

Once the off-diagonal elements of the last row and column have converged (are sufficiently small), the problem can be *deflated* by applying the following theorem:

**Theorem 15.4** Let

$$A = \left( \begin{array}{c|c|c|c} A_{0,0} & A_{01} & \cdots & A_{0,N-1} \\ \hline 0 & A_{1,1} & \cdots & A_{1,N-1} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & A_{N-1,N-1} \end{array} \right),$$

where  $A_{k,k}$  are all square. Then  $\Lambda(A) = \cup_{k=0}^{N-1} \Lambda(A_{k,k})$ .

**Homework 15.5** *Prove the above theorem.*

 [SEE ANSWER](#)

In other words, once the last row and column have converged, the algorithm can continue with the submatrix that consists of the first  $n - 1$  rows and columns.

The problem with the QR algorithm, as stated, is that each iteration requires  $O(n^3)$  operations, which is too expensive given that many iterations are required to find all eigenvalues and eigenvectors.

## 15.4 Reduction to Tridiagonal Form

In the next section, we will see that if  $A^{(0)}$  is a tridiagonal matrix, then so are all  $A^{(k)}$ . This reduces the cost of each iteration from  $O(n^3)$  to  $O(n)$ . We first show how unitary similarity transformations can be used to reduce a matrix to tridiagonal form.

### 15.4.1 Householder transformations (reflectors)

We briefly review the main tool employed to reduce a matrix to tridiagonal form: the Householder transform, also known as a reflector. Full details were given in Chapter 6.

**Definition 15.6** Let  $u \in \mathbb{R}^n$ ,  $\tau \in \mathbb{R}$ . Then  $H = H(u) = I - uu^T / \tau$ , where  $\tau = \frac{1}{2}u^T u$ , is said to be a reflector or Householder transformation.

We observe:

- Let  $z$  be any vector that is perpendicular to  $u$ . Applying a Householder transform  $H(u)$  to  $z$  leaves the vector unchanged:  $H(u)z = z$ .
- Let any vector  $x$  be written as  $x = z + u^T x u$ , where  $z$  is perpendicular to  $u$  and  $u^T x u$  is the component of  $x$  in the direction of  $u$ . Then  $H(u)x = z - u^T x u$ .

This can be interpreted as follows: The space perpendicular to  $u$  acts as a “mirror”: any vector in that space (along the mirror) is not reflected, while any other vector has the component that is orthogonal to the space (the component outside and orthogonal to the mirror) reversed in direction. Notice that a reflection preserves the length of the vector. Also, it is easy to verify that:

1.  $HH = I$  (reflecting the reflection of a vector results in the original vector);
2.  $H = H^T$ , and so  $H^T H = HH^T = I$  (a reflection is an orthogonal matrix and thus preserves the norm); and
3. if  $H_0, \dots, H_{k-1}$  are Householder transformations and  $Q = H_0 H_1 \cdots H_{k-1}$ , then  $Q^T Q = QQ^T = I$  (an accumulation of reflectors is an orthogonal matrix).

As part of the reduction to condensed form operations, given a vector  $x$  we will wish to find a Householder transformation,  $H(u)$ , such that  $H(u)x$  equals a vector with zeroes below the first element:  $H(u)x = \mp \|x\|_2 e_0$  where  $e_0$  equals the first column of the identity matrix. It can be easily checked that choosing  $u = x \pm \|x\|_2 e_0$  yields the desired  $H(u)$ . Notice that any nonzero scaling of  $u$  has the same property, and the convention is to scale  $u$  so that the first element equals one. Let us define  $[u, \tau, h] = \text{HouseV}(x)$  to be the function that returns  $u$  with first element equal to one,  $\tau = \frac{1}{2}u^T u$ , and  $h = H(u)x$ .

### 15.4.2 Algorithm

The first step towards computing the eigenvalue decomposition of a symmetric matrix is to reduce the matrix to tridiagonal form.

The basic algorithm for reducing a symmetric matrix to tridiagonal form, overwriting the original matrix with the result, can be explained as follows. We assume that symmetric  $A$  is stored only in the lower triangular part of the matrix and that only the diagonal and subdiagonal of the symmetric tridiagonal matrix is computed, overwriting those parts of  $A$ . Finally, the Householder vectors used to zero out parts of  $A$  overwrite the entries that they annihilate (set to zero).

- Partition  $A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{21}^T \\ \hline a_{21} & A_{22} \end{array} \right).$

- Let  $[u_{21}, \tau, a_{21}] := \text{HouseV}(a_{21}).$ <sup>1</sup>

- Update

$$\left( \begin{array}{c|c} \alpha_{11} & a_{21}^T \\ \hline a_{21} & A_{22} \end{array} \right) := \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & H \end{array} \right) \left( \begin{array}{c|c} \alpha_{11} & a_{21}^T \\ \hline a_{21} & A_{22} \end{array} \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & H \end{array} \right) = \left( \begin{array}{c|c} \alpha_{11} & a_{21}^T H \\ \hline H a_{21} & H A_{22} H \end{array} \right)$$

where  $H = H(u_{21})$ . Note that  $a_{21} := H a_{21}$  need not be executed since this update was performed by the instance of `HouseV` above.<sup>2</sup> Also,  $a_{12}^T$  is not stored nor updated due to symmetry. Finally, only the lower triangular part of  $H A_{22} H$  is computed, overwriting  $A_{22}$ . The update of  $A_{22}$  warrants closer scrutiny:

$$\begin{aligned} A_{22} &:= (I - \frac{1}{\tau} u_{21} u_{21}^T) A_{22} (I - \frac{1}{\tau} u_{21} u_{21}^T) \\ &= (A_{22} - \frac{1}{\tau} u_{21} \underbrace{u_{21}^T A_{22}}_{y_{21}^T}) (I - \frac{1}{\tau} u_{21} u_{21}^T) \\ &= A_{22} - \frac{1}{\tau} u_{21} y_{21}^T - \frac{1}{\tau} \underbrace{A u_{21}}_{y_{21}} u_{21}^T + \frac{1}{\tau^2} u_{21} \underbrace{y_{21}^T u_{21}}_{2\beta} u_{21}^T \\ &= A_{22} - \left( \frac{1}{\tau} u_{21} y_{21}^T - \frac{\beta}{\tau^2} u_{21} u_{21}^T \right) - \left( \frac{1}{\tau} y_{21} u_{21}^T - \frac{\beta}{\tau^2} u_{21} u_{21}^T \right) \\ &= A_{22} - u_{21} \underbrace{\frac{1}{\tau} \left( y_{21}^T - \frac{\beta}{\tau} u_{21}^T \right)}_{w_{21}^T} - \underbrace{\frac{1}{\tau} \left( y_{21} - \frac{\beta}{\tau} u_{21} \right)}_{w_{21}} u_{21}^T \\ &= \underbrace{A_{22} - u_{21} w_{21}^T - w_{21} u_{21}^T}_{\text{symmetric}} \\ &\quad \text{rank-2 update} \end{aligned}$$

<sup>1</sup> Note that the semantics here indicate that  $a_{21}$  is overwritten by  $H a_{21}$ .

<sup>2</sup> In practice, the zeros below the first element of  $H a_{21}$  are not actually written. Instead, the implementation overwrites these elements with the corresponding elements of the vector  $u_{21}$ .

**Algorithm:**  $[A, t] := \text{TRIRED\_UNB}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), t \rightarrow \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$  and  $t_b$  has 0 elements

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \rightarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ t_2 \end{array} \right)$$

**where**  $\alpha_{11}$  and  $\tau_1$  are scalars

---

---


$$[u_{21}, \tau_1, a_{21}] := \text{HouseV}(a_{21})$$

$$y_{21} := A_{22}u_{21}$$

$$\beta := u_{21}^T y_{21} / 2$$

$$w_{21} := (y_{21} - \beta u_{21} / \tau_1) / \tau_1$$

$$A_{22} := A_{22} - \text{tril}(u_{21} w_{21}^T + w_{21} u_{21}^T) \quad (\text{symmetric rank-2 update})$$


---

---

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \leftarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ t_2 \end{array} \right)$$

**endwhile**

Figure 15.5: Basic algorithm for reduction of a symmetric matrix to tridiagonal form.

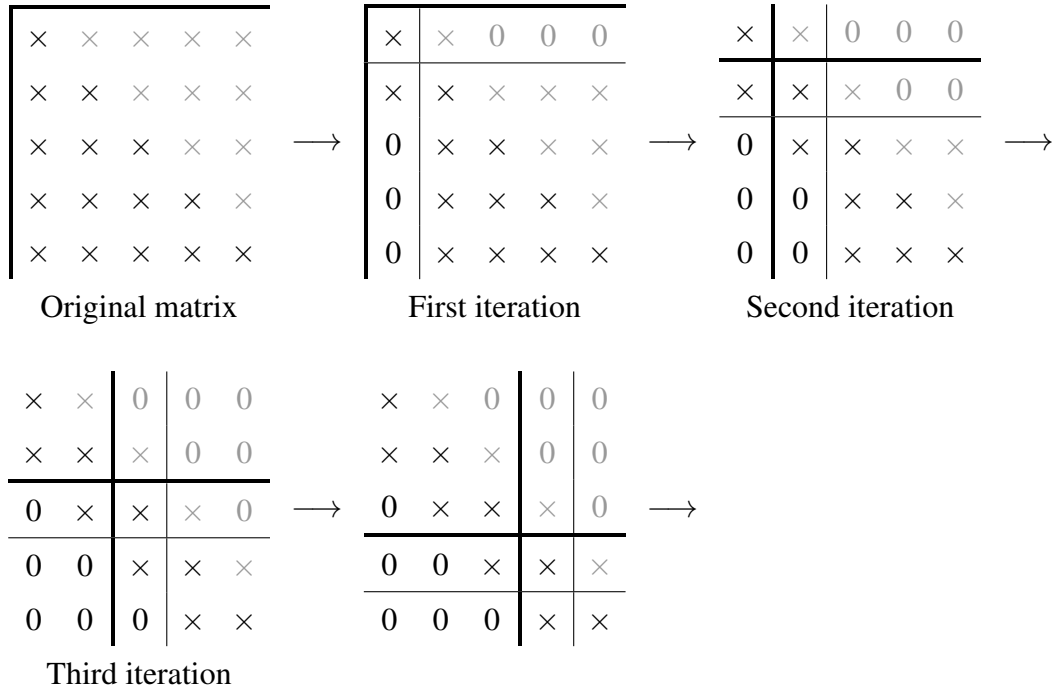


Figure 15.6: Illustration of reduction of a symmetric matrix to tridiagonal form. The  $\times$ s denote nonzero elements in the matrix. The gray entries above the diagonal are not actually updated.

- Continue this process with the updated  $A_{22}$ .

This is captured in the algorithm in Figure 15.5. It is also illustrated in Figure 15.6.

The total cost for reducing  $A \in \mathbb{R}^{n \times n}$  is approximately

$$\sum_{k=0}^{n-1} (4(n-k-1)^2) \text{ flops} \approx \frac{4}{3}n^3 \text{ flops.}$$

This equals, approximately, the cost of one QR factorization of matrix  $A$ .

## 15.5 The QR algorithm with a Tridiagonal Matrix

We are now ready to describe an algorithm for the QR algorithm with a tridiagonal matrix.

### 15.5.1 Givens' rotations

First, we introduce another important class of unitary matrices known as Givens' rotations. Given a vector  $x = \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} \in \mathbb{R}^2$ , there exists an orthogonal matrix  $G$  such that  $G^T x = \begin{pmatrix} \pm \|x\|_2 \\ 0 \end{pmatrix}$ . The Householder

transformation is one example of such a matrix  $G$ . An alternative is the Givens' rotation:  $G = \begin{pmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{pmatrix}$

where  $\gamma^2 + \sigma^2 = 1$ . (Notice that  $\gamma$  and  $\sigma$  can be thought of as the cosine and sine of an angle.) Then

$$\begin{aligned} G^T G &= \begin{pmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{pmatrix}^T \begin{pmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{pmatrix} = \begin{pmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{pmatrix} \begin{pmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{pmatrix} \\ &= \begin{pmatrix} \gamma^2 + \sigma^2 & -\gamma\sigma + \gamma\sigma \\ \gamma\sigma - \gamma\sigma & \gamma^2 + \sigma^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

which means that a Givens' rotation is a unitary matrix.

Now, if  $\gamma = \chi_1 / \|x\|_2$  and  $\sigma = \chi_2 / \|x\|_2$ , then  $\gamma^2 + \sigma^2 = (\chi_1^2 + \chi_2^2) / \|x\|_2^2 = 1$  and

$$\begin{pmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{pmatrix}^T \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} = \begin{pmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{pmatrix} \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} = \begin{pmatrix} (\chi_1^2 + \chi_2^2) / \|x\|_2 \\ (\chi_1\chi_2 - \chi_1\chi_2) / \|x\|_2 \end{pmatrix} = \begin{pmatrix} \|x\|_2 \\ 0 \end{pmatrix}.$$

## 15.6 QR Factorization of a Tridiagonal Matrix

Now, consider the  $4 \times 4$  tridiagonal matrix

$$\begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & 0 & 0 \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & 0 \\ 0 & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{pmatrix}$$

From  $\begin{pmatrix} \alpha_{0,0} \\ \alpha_{1,0} \end{pmatrix}$  one can compute  $\gamma_{1,0}$  and  $\sigma_{1,0}$  so that

$$\begin{pmatrix} \gamma_{1,0} & -\sigma_{1,0} \\ \sigma_{1,0} & \gamma_{1,0} \end{pmatrix}^T \begin{pmatrix} \alpha_{0,0} \\ \alpha_{1,0} \end{pmatrix} = \begin{pmatrix} \hat{\alpha}_{0,0} \\ 0 \end{pmatrix}.$$

Then

$$\begin{pmatrix} \hat{\alpha}_{0,0} & \hat{\alpha}_{0,1} & \hat{\alpha}_{0,2} & 0 \\ 0 & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & 0 \\ \hline 0 & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ \hline 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{pmatrix} = \begin{pmatrix} \gamma_{1,0} & \sigma_{1,0} & 0 & 0 \\ -\sigma_{1,0} & \gamma_{1,0} & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & 0 & 0 \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & 0 \\ \hline 0 & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ \hline 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{pmatrix}$$

Next, from  $\begin{pmatrix} \hat{\alpha}_{1,1} \\ \alpha_{2,1} \end{pmatrix}$  one can compute  $\gamma_{2,1}$  and  $\sigma_{2,1}$  so that

$$\begin{pmatrix} \gamma_{2,1} & -\sigma_{2,1} \\ \sigma_{2,1} & \gamma_{2,1} \end{pmatrix}^T \begin{pmatrix} \hat{\alpha}_{1,1} \\ \alpha_{2,1} \end{pmatrix} = \begin{pmatrix} \hat{\hat{\alpha}}_{1,1} \\ 0 \end{pmatrix}.$$



Then

$$\left( \begin{array}{c|c|c|c} \hat{\alpha}_{0,0} & \hat{\alpha}_{0,1} & \hat{\alpha}_{0,2} & 0 \\ \hline 0 & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & \hat{\alpha}_{1,3} \\ \hline 0 & 0 & \hat{\alpha}_{2,2} & \hat{\alpha}_{2,3} \\ \hline 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{array} \right) = \left( \begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & \gamma_{2,1} & \sigma_{2,1} & 0 \\ \hline 0 & -\sigma_{2,1} & \gamma_{2,1} & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right) \left( \begin{array}{c|c|c|c} \hat{\alpha}_{0,0} & \hat{\alpha}_{0,1} & \hat{\alpha}_{0,2} & 0 \\ \hline 0 & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & 0 \\ \hline 0 & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ \hline 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{array} \right)$$

Finally, from  $\begin{pmatrix} \hat{\alpha}_{2,2} \\ \alpha_{3,2} \end{pmatrix}$  one can compute  $\gamma_{3,2}$  and  $\sigma_{3,2}$  so that  $\begin{pmatrix} \gamma_{3,2} & -\sigma_{3,2} \\ \sigma_{3,2} & \gamma_{3,2} \end{pmatrix}^T \begin{pmatrix} \hat{\alpha}_{2,2} \\ \alpha_{3,2} \end{pmatrix} = \begin{pmatrix} \hat{\alpha}_{2,2} \\ 0 \end{pmatrix}$ .

Then

$$\left( \begin{array}{c|c|c|c} \hat{\alpha}_{0,0} & \hat{\alpha}_{0,1} & \hat{\alpha}_{0,2} & 0 \\ \hline 0 & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & \hat{\alpha}_{1,3} \\ \hline 0 & 0 & \hat{\alpha}_{2,2} & \hat{\alpha}_{2,3} \\ \hline 0 & 0 & 0 & \hat{\alpha}_{3,3} \end{array} \right) = \left( \begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 1 & 0 & \gamma_{3,2} & \sigma_{3,2} \\ \hline 0 & 1 & -\sigma_{3,2} & \gamma_{3,2} \end{array} \right) \left( \begin{array}{c|c|c|c} \hat{\alpha}_{0,0} & \hat{\alpha}_{0,1} & \hat{\alpha}_{0,2} & 0 \\ \hline 0 & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & \hat{\alpha}_{1,3} \\ \hline 0 & 0 & \hat{\alpha}_{2,2} & \hat{\alpha}_{2,3} \\ \hline 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{array} \right)$$

The matrix  $Q$  is the orthogonal matrix that results from multiplying the different Givens' rotations together:

$$Q = \left( \begin{array}{c|c|c|c} \gamma_{1,0} & -\sigma_{1,0} & 0 & 0 \\ \hline \sigma_{1,0} & \gamma_{1,0} & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right) \left( \begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & \gamma_{2,1} & -\sigma_{2,1} & 0 \\ \hline 0 & \sigma_{2,1} & \gamma_{2,1} & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right) \left( \begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & \gamma_{3,2} & -\sigma_{3,2} \\ \hline 0 & 0 & \sigma_{3,2} & \gamma_{3,2} \end{array} \right). \quad (15.2)$$

However, it is typically not explicitly formed.

The next question is how to compute  $RQ$  given the QR factorization of the tridiagonal matrix:

$$\begin{pmatrix} \hat{\alpha}_{0,0} & \hat{\alpha}_{0,1} & \hat{\alpha}_{0,2} & 0 \\ 0 & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & \hat{\alpha}_{1,3} \\ 0 & 0 & \hat{\alpha}_{2,2} & \hat{\alpha}_{2,3} \\ 0 & 0 & 0 & \hat{\alpha}_{3,3} \end{pmatrix} \begin{pmatrix} \gamma_{1,0} & -\sigma_{1,0} & 0 & 0 \\ \sigma_{1,0} & \gamma_{1,0} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \gamma_{2,1} & -\sigma_{2,1} & 0 \\ 0 & \sigma_{2,1} & \gamma_{2,1} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma_{3,2} & -\sigma_{3,2} \\ 0 & 0 & \sigma_{3,2} & \gamma_{3,2} \end{pmatrix} \\
 \underbrace{\begin{pmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{0,1} & \tilde{\alpha}_{0,2} & 0 \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \tilde{\alpha}_{1,2} & \tilde{\alpha}_{1,3} \\ 0 & 0 & \tilde{\alpha}_{2,2} & \tilde{\alpha}_{2,3} \\ 0 & 0 & 0 & \tilde{\alpha}_{3,3} \end{pmatrix}}_{\begin{pmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{0,1} & \tilde{\alpha}_{0,2} & 0 \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \tilde{\alpha}_{1,2} & \tilde{\alpha}_{1,3} \\ 0 & \tilde{\alpha}_{2,1} & \tilde{\alpha}_{2,2} & \tilde{\alpha}_{2,3} \\ 0 & 0 & 0 & \tilde{\alpha}_{3,3} \end{pmatrix}} \underbrace{\begin{pmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{0,1} & \tilde{\alpha}_{0,2} & 0 \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \tilde{\alpha}_{1,2} & \tilde{\alpha}_{1,3} \\ 0 & \tilde{\alpha}_{2,1} & \tilde{\alpha}_{2,2} & \tilde{\alpha}_{2,3} \\ 0 & 0 & 0 & \tilde{\alpha}_{3,3} \end{pmatrix}}_{\begin{pmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{0,1} & \tilde{\alpha}_{0,2} & 0 \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \tilde{\alpha}_{1,2} & \tilde{\alpha}_{1,3} \\ 0 & \tilde{\alpha}_{2,1} & \tilde{\alpha}_{2,2} & \tilde{\alpha}_{2,3} \\ 0 & 0 & \tilde{\alpha}_{3,2} & \tilde{\alpha}_{3,3} \end{pmatrix}}.$$

A symmetry argument can be used to motivate that  $\tilde{\alpha}_{0,2} = \tilde{\alpha}_{1,3} = 0$ .

## 15.7 The Implicitly Shifted QR Algorithm

### 15.7.1 Upper Hessenberg and tridiagonal matrices

**Definition 15.7** A matrix is said to be upper Hessenberg if all entries below its first subdiagonal equal zero.

In other words, if matrix  $A \in \mathbb{R}^{n \times n}$  is upper Hessenberg, it looks like

$$A = \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} & \cdots & \alpha_{0,n-1} & \alpha_{0,n-1} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,n-1} & \alpha_{1,n-1} \\ 0 & \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,n-1} & \alpha_{2,n-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & \alpha_{n-2,n-2} & \alpha_{n-2,n-2} \\ 0 & 0 & 0 & \cdots & \alpha_{n-1,n-2} & \alpha_{n-1,n-2} \end{pmatrix}.$$

Obviously, a tridiagonal matrix is a special case of an upper Hessenberg matrix.

### 15.7.2 The Implicit Q Theorem

The following theorem sets up one of the most remarkable algorithms in numerical linear algebra, which allows us to greatly simplify the implementation of the shifted QR algorithm when  $A$  is tridiagonal.

**Theorem 15.8 (Implicit Q Theorem)** *Let  $A, B \in \mathbb{R}^{n \times n}$  where  $B$  is upper Hessenberg and has only positive elements on its first subdiagonal and assume there exists a unitary matrix  $Q$  such that  $Q^T A Q = B$ . Then  $Q$  and  $B$  are uniquely determined by  $A$  and the first column of  $Q$ .*

**Proof:** Partition

$$Q = \left( \begin{array}{c|c|c|c|c|c} q_0 & q_1 & q_2 & \cdots & q_{n-2} & q_{n-1} \end{array} \right) \text{ and } B = \left( \begin{array}{c|c|c|c|c|c} \beta_{0,0} & \beta_{0,1} & \beta_{0,2} & \cdots & \beta_{0,n-2} & \beta_{0,n-1} \\ \beta_{1,0} & \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,n-2} & \beta_{1,n-1} \\ \hline 0 & \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,n-2} & \beta_{2,n-1} \\ 0 & 0 & \beta_{3,2} & \cdots & \beta_{3,n-2} & \beta_{3,n-1} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline 0 & 0 & 0 & \cdots & \beta_{n-1,n-2} & \beta_{n-1,n-1} \end{array} \right).$$

Notice that  $AQ = QB$  and hence

$$\begin{aligned} & A \left( \begin{array}{c|c|c|c|c|c} q_0 & q_1 & q_2 & \cdots & q_{n-2} & q_{n-1} \end{array} \right) \\ &= \left( \begin{array}{c|c|c|c|c|c} q_0 & q_1 & q_2 & \cdots & q_{n-2} & q_{n-1} \end{array} \right) \left( \begin{array}{c|c|c|c|c|c} \beta_{0,0} & \beta_{0,1} & \beta_{0,2} & \cdots & \beta_{0,n-2} & \beta_{0,n-1} \\ \beta_{1,0} & \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,n-2} & \beta_{1,n-1} \\ \hline 0 & \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,n-1} & \\ 0 & 0 & \beta_{3,2} & \cdots & \beta_{3,n-2} & \beta_{3,n-1} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline 0 & 0 & 0 & \cdots & \beta_{n-1,n-2} & \beta_{n-1,n-1} \end{array} \right). \end{aligned}$$

Equating the first column on the left and right, we notice that

$$Aq_0 = \beta_{0,0}q_0 + \beta_{1,0}q_1.$$

Now,  $q_0$  is given and  $\|q_0\|_2$  since  $Q$  is unitary. Hence

$$q_0^T A q_0 = \beta_{0,0} q_0^T q_0 + \beta_{1,0} q_0^T q_1 = \beta_{0,0}.$$

Next,

$$\beta_{1,0}q_1 = Aq_0 - \beta_{0,0}q_0 = \tilde{q}_1.$$

Since  $\|q_1\|_2 = 1$  (it is a column of a unitary matrix) and  $\beta_{1,0}$  is assumed to be positive, then we know that

$$\beta_{1,0} = \|\tilde{q}_1\|_2.$$

Finally,

$$q_1 = \tilde{q}_1 / \beta_{1,0}.$$

The point is that the first column of  $B$  and second column of  $Q$  are prescribed by the first column of  $Q$  and the fact that  $B$  has positive elements on the first subdiagonal.

In this way, each column of  $Q$  and each column of  $B$  can be determined, one by one.

**Homework 15.9** Give all the details of the above proof.

SEE ANSWER

Notice the similarity between the above proof and the proof of the existence and uniqueness of the QR factorization!

To take advantage of the special structure of  $A$  being symmetric, the theorem can be expanded to

**Theorem 15.10 (Implicit Q Theorem)** Let  $A, B \in \mathbb{R}^{n \times n}$  where  $B$  is upper Hessenberg and has only positive elements on its first subdiagonal and assume there exists a unitary matrix  $Q$  such that  $Q^T A Q = B$ . Then  $Q$  and  $B$  are uniquely determined by  $A$  and the first column of  $Q$ . If  $A$  is symmetric, then  $B$  is also symmetric and hence tridiagonal.

### 15.7.3 The Francis QR Step

The Francis QR Step combines the steps  $(A^{(k-1)} - \mu_k I) \rightarrow Q^{(k)} R^{(k)}$  and  $A^{(k+1)} := R^{(k)} Q^{(k)} + \mu_k I$  into a single step.

Now, consider the  $4 \times 4$  tridiagonal matrix

$$\begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & 0 & 0 \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & 0 \\ 0 & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{pmatrix} - \mu I$$

The first Givens' rotation is computed from  $\begin{pmatrix} \alpha_{0,0} - \mu \\ \alpha_{1,0} \end{pmatrix}$ , yielding  $\gamma_{1,0}$  and  $\sigma_{1,0}$  so that

$$\begin{pmatrix} \gamma_{1,0} & -\sigma_{1,0} \\ \sigma_{1,0} & \gamma_{1,0} \end{pmatrix}^T \begin{pmatrix} \alpha_{0,0} - \mu I \\ \alpha_{1,0} \end{pmatrix}$$

has a zero second entry. Now, to preserve eigenvalues, any orthogonal matrix that is applied from the left must also have its transpose applied from the right. Let us compute

$$\begin{pmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{1,0} & \hat{\alpha}_{2,0} & 0 \\ \hat{\alpha}_{1,0} & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & 0 \\ \hat{\alpha}_{2,0} & \hat{\alpha}_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{pmatrix} = \begin{pmatrix} \gamma_{1,0} & \sigma_{1,0} & 0 & 0 \\ -\sigma_{1,0} & \gamma_{1,0} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{0,0} & \alpha_{0,1} & 0 & 0 \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & 0 \\ 0 & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{pmatrix} \begin{pmatrix} \gamma_{1,0} & -\sigma_{1,0} & 0 & 0 \\ \sigma_{1,0} & \gamma_{1,0} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next, from  $\begin{pmatrix} \hat{\alpha}_{1,0} \\ \hat{\alpha}_{2,0} \end{pmatrix}$  one can compute  $\gamma_{2,0}$  and  $\sigma_{2,0}$  so that  $\begin{pmatrix} \gamma_{2,0} & -\sigma_{2,0} \\ \sigma_{2,0} & \gamma_{2,0} \end{pmatrix}^T \begin{pmatrix} \hat{\alpha}_{1,0} \\ \hat{\alpha}_{2,0} \end{pmatrix} = \begin{pmatrix} \tilde{\alpha}_{1,0} \\ 0 \end{pmatrix}$ .

Then

$$\begin{pmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{1,0} & 0 & 0 \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \hat{\alpha}_{2,1} & \hat{\alpha}_{3,1} \\ 0 & \hat{\alpha}_{2,1} & \hat{\alpha}_{2,2} & \hat{\alpha}_{2,3} \\ 0 & \hat{\alpha}_{3,1} & \hat{\alpha}_{3,2} & \alpha_{3,3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \gamma_{2,0} & \sigma_{2,0} & 0 \\ 0 & -\sigma_{2,0} & \gamma_{2,0} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{1,0} & \hat{\alpha}_{2,0} & 0 \\ \hat{\alpha}_{1,0} & \hat{\alpha}_{1,1} & \hat{\alpha}_{1,2} & 0 \\ \hat{\alpha}_{2,0} & \hat{\alpha}_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ 0 & 0 & \alpha_{3,2} & \alpha_{3,3} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \gamma_{2,0} & -\sigma_{2,0} & 0 \\ 0 & \sigma_{2,0} & \gamma_{2,0} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

From: Gene H Golub <golub@stanford.edu>  
Date: Sun, 19 Aug 2007 13:54:47 -0700 (PDT)  
Subject: John Francis, Co-Inventor of QR

Dear Colleagues,

For many years, I have been interested in meeting J G F Francis, one of the co-inventors of the QR algorithm for computing eigenvalues of general matrices. Through a lead provided by the late Erin Brent and with the aid of Google, I finally made contact with him.

John Francis was born in 1934 in London and currently lives in Hove, near Brighton. His residence is about a quarter mile from the sea; he is a widower. In 1954, he worked at the National Research Development Corp (NRDC) and attended some lectures given by Christopher Strachey. In 1955,'56 he was a student at Cambridge but did not complete a degree. He then went back to NRDC as an assistant to Strachey where he got involved in flutter computations and this led to his work on QR.

After leaving NRDC in 1961, he worked at the Ferranti Corp and then at the University of Sussex. Subsequently, he had positions with various industrial organizations and consultancies. He is now retired. His interests were quite general and included Artificial Intelligence, computer languages, systems engineering. He has not returned to numerical computation.

He was surprised to learn there are many references to his work and that the QR method is considered one of the ten most important algorithms of the 20th century. He was unaware of such developments as TeX and Math Lab. Currently he is working on a degree at the Open University.

John Francis did remarkable work and we are all in his debt. Along with the conjugate gradient method, it provided us with one of the basic tools of numerical analysis.

Gene Golub

Figure 15.7: Posting by the late Gene Golub in NA Digest Sunday, August 19, 2007 Volume 07 : Issue 34. An article on the ten most important algorithms of the 20th century, published in SIAM News, can be found at <http://www.uta.edu/faculty/rcli/TopTen/topten.pdf>.

again preserves eigenvalues. Finally, from  $\begin{pmatrix} \hat{\alpha}_{2,1} \\ \hat{\alpha}_{3,1} \end{pmatrix}$  one can compute  $\gamma_{3,1}$  and  $\sigma_{3,1}$  so that

$$\begin{pmatrix} \gamma_{3,1} & -\sigma_{3,1} \\ \sigma_{3,1} & \gamma_{3,1} \end{pmatrix}^T \begin{pmatrix} \hat{\alpha}_{2,1} \\ \hat{\alpha}_{3,1} \end{pmatrix} = \begin{pmatrix} \tilde{\alpha}_{2,1} \\ 0 \end{pmatrix}.$$

Then

$$\left( \begin{array}{cc|cc} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{1,0} & 0 & 0 \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \tilde{\alpha}_{2,1} & 0 \\ \hline 0 & \tilde{\alpha}_{2,1} & \tilde{\alpha}_{2,2} & \tilde{\alpha}_{2,3} \\ 0 & 0 & \tilde{\alpha}_{3,2} & \tilde{\alpha}_{3,3} \end{array} \right) = \left( \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 1 & 0 & \gamma_{3,2} & \sigma_{3,2} \\ 0 & 1 & -\sigma_{3,2} & \gamma_{3,2} \end{array} \right) \left( \begin{array}{cc|cc} \tilde{\alpha}_{0,0} & \tilde{\alpha}_{1,0} & 0 & 0 \\ \tilde{\alpha}_{1,0} & \tilde{\alpha}_{1,1} & \hat{\alpha}_{2,1} & \hat{\alpha}_{3,1} \\ \hline 0 & \hat{\alpha}_{2,1} & \hat{\alpha}_{2,2} & \hat{\alpha}_{2,3} \\ 0 & \hat{\alpha}_{3,1} & \hat{\alpha}_{3,2} & \alpha_{3,3} \end{array} \right) \left( \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 1 & 0 & \gamma_{3,1} & -\sigma_{3,1} \\ 0 & 1 & \sigma_{3,1} & \gamma_{3,1} \end{array} \right)$$

The matrix  $Q$  is the orthogonal matrix that results from multiplying the different Givens' rotations together:

$$Q = \left( \begin{array}{cc|cc} \gamma_{1,0} & -\sigma_{1,0} & 0 & 0 \\ \sigma_{1,0} & \gamma_{1,0} & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right) \left( \begin{array}{c|ccc} 1 & 0 & 0 & 0 \\ \hline 0 & \gamma_{2,0} & -\sigma_{2,0} & 0 \\ 0 & \sigma_{2,0} & \gamma_{2,0} & 0 \\ \hline 0 & 0 & 0 & 1 \end{array} \right) \left( \begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & \gamma_{3,1} & -\sigma_{3,1} \\ 0 & 0 & \sigma_{3,1} & \gamma_{3,1} \end{array} \right).$$

It is important to note that the first columns of  $Q$  is given by  $\begin{pmatrix} \gamma_{1,0} \\ \sigma_{1,0} \\ 0 \\ 0 \end{pmatrix}$ , which is exactly the same first

column had  $Q$  been computed as in Section 15.6 (Equation 15.2). Thus, by the Implicit Q Theorem, the tridiagonal matrix that results from this approach is equal to the tridiagonal matrix that would be computed by applying the QR factorization from Section 15.6 with  $A - \mu I$ ,  $A - \mu I \rightarrow QR$  followed by the formation of  $RQ + \mu I$  using the algorithm for computing  $RQ$  in Section 15.6.

The successive elimination of elements  $\tilde{\alpha}_{i+1,i}$  is often referred to as *chasing the bulge* while the entire process that introduces the bulge and then chases it is known as a Francis Implicit QR Step. Obviously, the method generalizes to matrices of arbitrary size, as illustrated in Figure 15.8. An algorithm for the chasing of the bulge is given in Figure 17.4. (Note that in those figures  $T$  is used for  $A$ , something that needs to be made consistent in these notes, eventually.) In practice, the tridiagonal matrix is not stored as a matrix. Instead, its diagonal and subdiagonal are stored as vectors.

### 15.7.4 A complete algorithm

This last section shows how one iteration of the QR algorithm can be performed on a tridiagonal matrix by implicitly shifting and then “chasing the bulge”. All that is left to complete the algorithm is to note that

- The shift  $\mu_k$  can be chosen to equal  $\alpha_{n-1,n-1}$  (the last element on the diagonal, which tends to converge to the eigenvalue smallest in magnitude). In practice, choosing the shift to be an eigenvalue of the bottom-right  $2 \times 2$  matrix works better. This is known as the *Wilkinson Shift*.

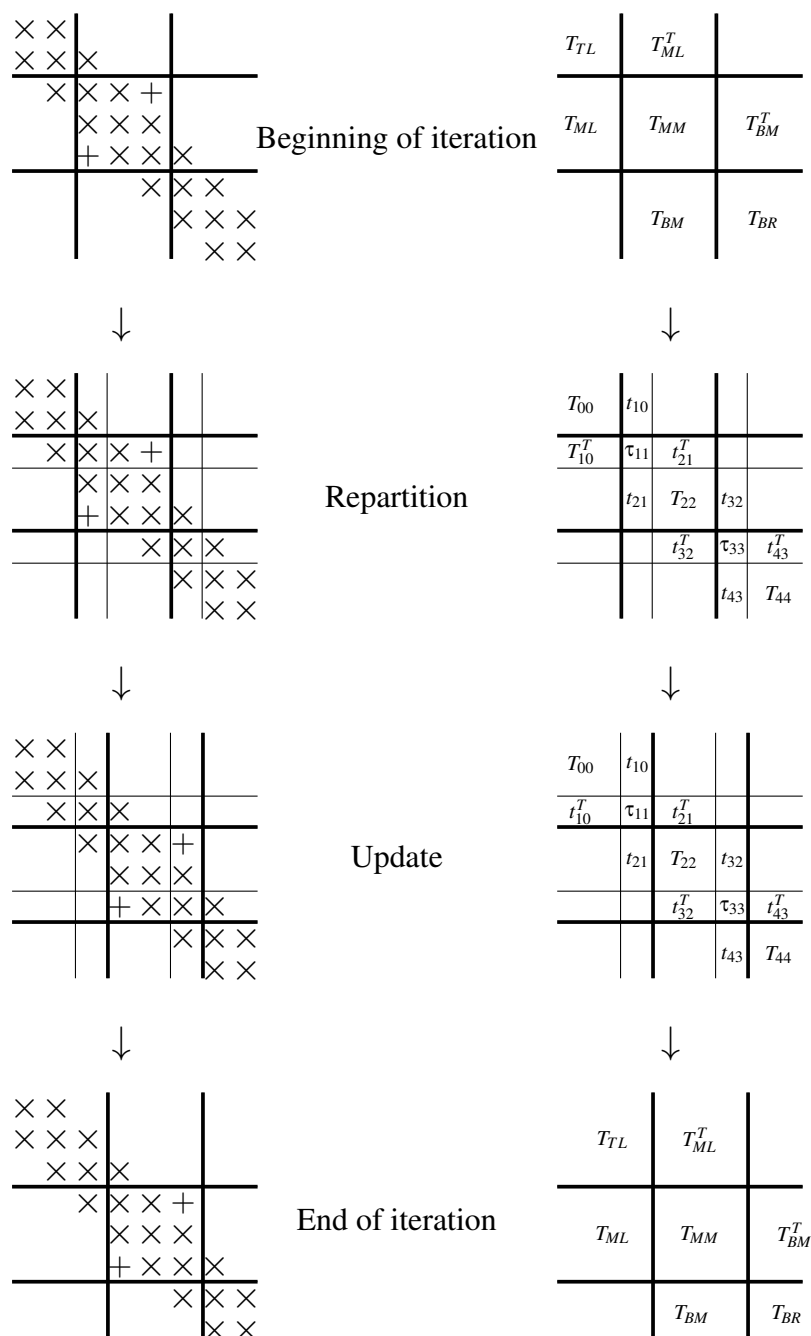


Figure 15.8: One step of “chasing the bulge” in the implicitly shifted symmetric QR algorithm.

- If an element of the subdiagonal (and corresponding element on the superdiagonal) becomes small enough, it can be considered to be zero and the problem deflates (decouples) into two smaller tridiagonal matrices. Small is often taken to mean that  $|\alpha_{i+1,i}| \leq \epsilon(|\alpha_{i,i}| + |\alpha_{i+1,i+1}|)$  where  $\epsilon$  is some quantity close to the machine epsilon (unit roundoff).
- If  $A = QTQ^T$  reduced  $A$  to the tridiagonal matrix  $T$  before the QR algorithm commenced, then the Givens' rotations encountered as part of the implicitly shifted QR algorithm can be applied from the

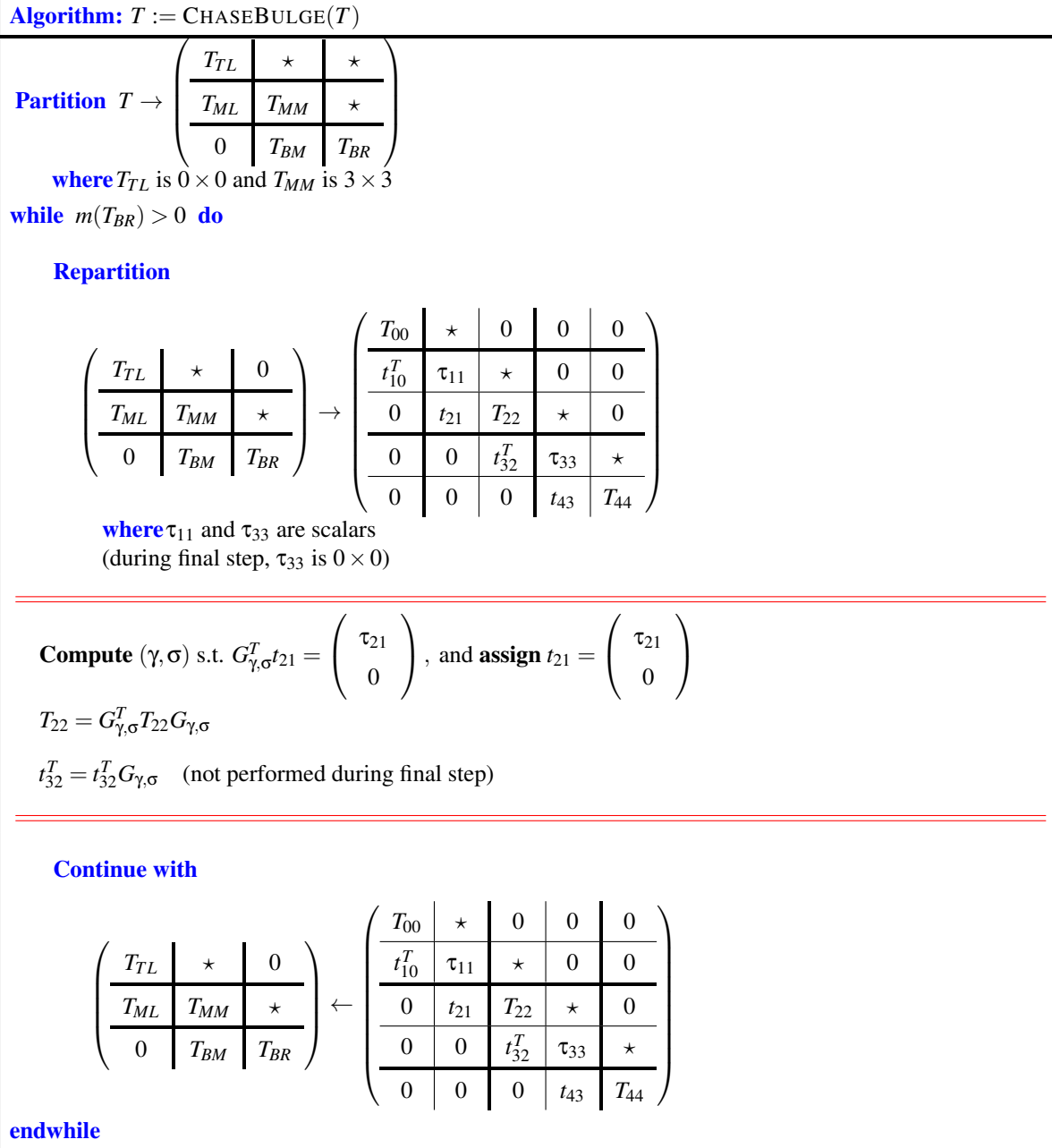


Figure 15.9: Chasing the bulge.

right to the appropriate columns of  $Q$  so that upon completion  $Q$  is overwritten with the eigenvectors of  $A$ . Notice that applying a Givens' rotation to a pair of columns of  $Q$  requires  $O(n)$  computation per Givens rotation. For each Francis implicit QR step  $O(n)$  Givens' rotations are computed, making the application of Givens' rotations to  $Q$  of cost  $O(n^2)$  per iteration of the implicitly shifted QR algorithm. Typically a few (2-3) iterations are needed per eigenvalue that is uncovered (by deflation) meaning that  $O(n)$  iterations are needed. Thus, the QR algorithm is roughly of cost  $O(n^3)$  if the eigenvalues are accumulated (in addition to the cost of forming the  $Q$  from the reduction to tridiagonal form, which takes another  $O(n^3)$  operations.)



- If an element on the subdiagonal becomes zero (or very small), and hence the corresponding element of the superdiagonal, then the problem can be *deflated*: If

$$T = \left( \begin{array}{c|c} T_{00} & 0 \\ \hline 0 & T_{11} \end{array} \right)$$

$$\begin{array}{cccc|cccc} \times & \times & & & & & & \\ \times & \times & \times & & & & & \\ & \times & \times & \times & & & & \\ & & \times & \times & \times & & & \\ & & & \times & \times & \times & & \\ & & & & \times & \times & 0 & \\ \hline & & & & & 0 & \times & \times \\ & & & & & & \times & \times & \times \\ & & & & & & & \times & \times \end{array}$$

then

- The computation can continue separately with  $T_{00}$  and  $T_{11}$ .
- One can pick the shift from the bottom-right of  $T_{00}$  as one continues finding the eigenvalues of  $T_{00}$ , thus accelerating the computation.
- One can pick the shift from the bottom-right of  $T_{11}$  as one continues finding the eigenvalues of  $T_{11}$ , thus accelerating the computation.
- One must continue to accumulate the eigenvectors by applying the rotations to the appropriate columns of  $Q$ .

Because of the connection between the QR algorithm and the Inverse Power Method, subdiagonal entries near the bottom-right of  $T$  are more likely to converge to a zero, so most deflation will happen there.

- A question becomes when an element on the subdiagonal,  $\tau_{i+1,i}$  can be considered to be zero. The answer is when  $|\tau_{i+1,i}|$  is small relative to  $|\tau_i|$  and  $|\tau_{i+1,i+1}|$ . A typical condition that is used is

$$|\tau_{i+1,i}| \leq \mathbf{u} \sqrt{|\tau_{i,i}| \tau_{i+1,i+1}|}.$$

- If  $A \in \mathbb{C}^{n \times n}$  is Hermitian, then its Spectral Decomposition is also computed via the following steps, which mirror those for the real symmetric case:
  - Reduce to tridiagonal form. Householder transformation based similarity transformations can again be used for this. This leaves one with a tridiagonal matrix,  $T$ , with real values along the diagonal (because the matrix is Hermitian) and values on the subdiagonal and superdiagonal that may be complex valued.
  - The matrix  $Q_T$  such  $A = Q_T T Q_T^H$  can then be formed from the Householder transformations.
  - A simple step can be used to then change this tridiagonal form to have real values even on the subdiagonal and superdiagonal. The matrix  $Q_T$  can be updated accordingly.
  - The tridiagonal QR algorithm that we described can then be used to diagonalize the matrix, accumulating the eigenvectors by applying the encountered Givens' rotations to  $Q_T$ . This is where the real expense is: Apply the Givens' rotations to matrix  $T$  requires  $O(n)$  per sweep. Applying the Givens' rotation to  $Q_T$  requires  $O(n^2)$  per sweep.

For details, see some of our papers mentioned in the next section.

## 15.8 Further Reading

### 15.8.1 More on reduction to tridiagonal form

The reduction to tridiagonal form can only be partially cast in terms of matrix-matrix multiplication [20]. This is a severe hindrance to high performance for that first step towards computing all eigenvalues and eigenvector of a symmetric matrix. Worse, a considerable fraction of the total cost of the computation is in that first step.

For a detailed discussion on the blocked algorithm that uses FLAME notation, we recommend [43]

Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí, G. Joseph Elizondo.

**Families of Algorithms for Reducing a Matrix to Condensed Form.**

ACM Transactions on Mathematical Software (TOMS) , Vol. 39, No. 1, 2012

(Reduction to tridiagonal form is one case of what is more generally referred to as “condensed form”).)

### 15.8.2 Optimizing the tridiagonal QR algorithm

As the Givens’ rotations are applied to the tridiagonal matrix, they are also applied to a matrix in which eigenvectors are accumulated. While one Implicit Francis Step requires  $O(n)$  computation, this accumulation of the eigenvectors requires  $O(n^2)$  computation with  $O(n^2)$  data. We have learned before that this means the cost of accessing data dominates on current architectures.

In a recent paper, we showed how accumulating the Givens’ rotations for several Francis Steps allows one to attain performance similar to that attained by a matrix-matrix multiplication. Details can be found in [42]:

Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí.

**Restructuring the Tridiagonal and Bidiagonal QR Algorithms for Performance.**

ACM Transactions on Mathematical Software (TOMS), Vol. 40, No. 3, 2014.

## 15.9 Other Algorithms

### 15.9.1 Jacobi’s method for the symmetric eigenvalue problem

(Not to be mistaken for the Jacobi iteration for solving linear systems.)

The oldest algorithm for computing the eigenvalues and eigenvectors of a matrix is due to Jacobi and dates back to 1846 [26]. This is a method that keeps resurfacing, since it parallelizes easily. The operation count tends to be higher (by a constant factor) than that of reduction to tridiagonal form followed by the tridiagonal QR algorithm.

The idea is as follows: Given a symmetric  $2 \times 2$  matrix

$$A_{31} = \begin{pmatrix} \alpha_{11} & \alpha_{13} \\ \alpha_{31} & \alpha_{33} \end{pmatrix}$$

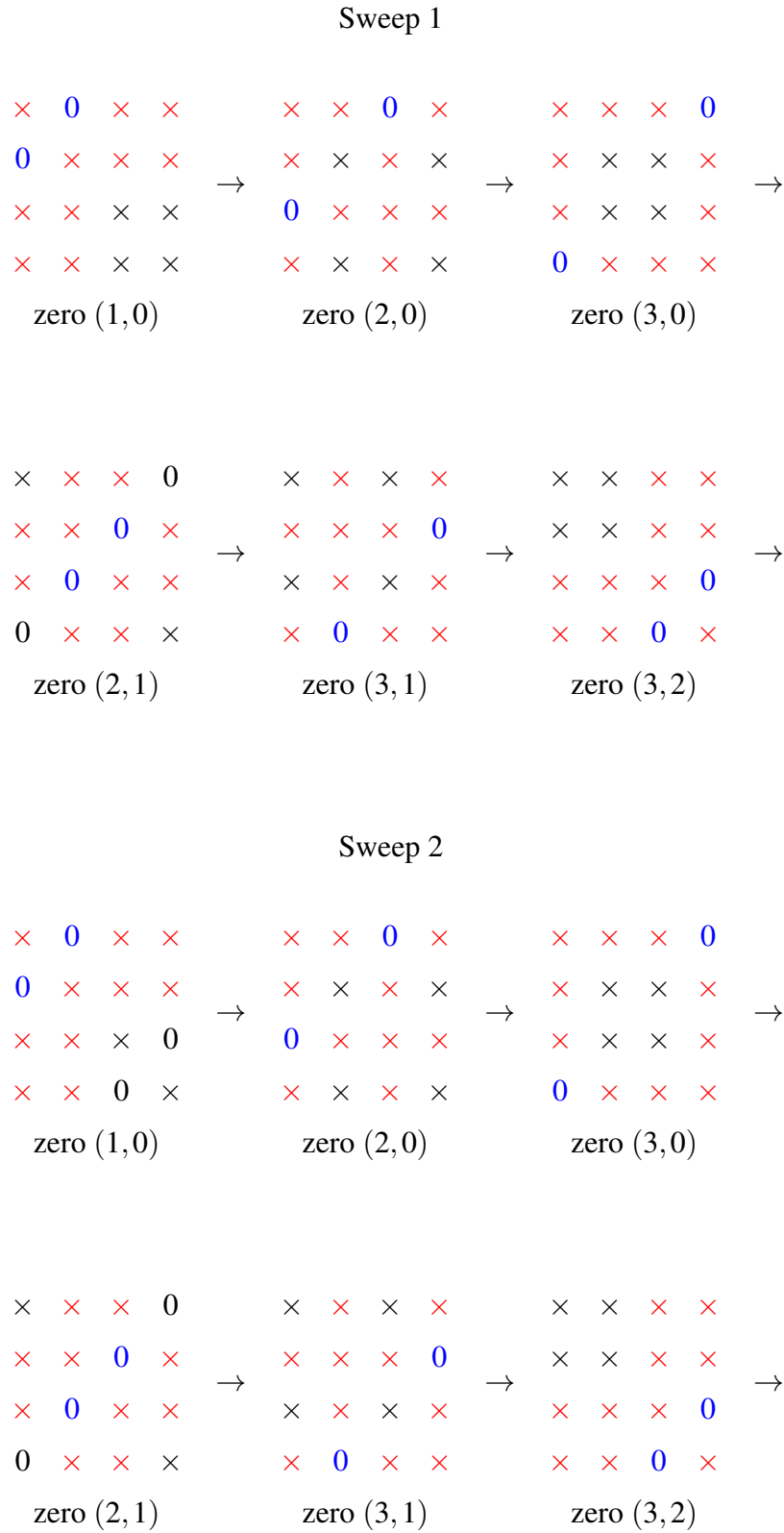


Figure 15.10: Column-cyclic Jacobi algorithm.

There exists a rotation (which is of course unitary)

$$J_{31} = \begin{pmatrix} \gamma_{11} & -\sigma_{13} \\ \sigma_{31} & \gamma_{33} \end{pmatrix}$$

such that

$$J_{31}A_{31}J_{31}^T = \begin{pmatrix} \gamma_{11} & -\sigma_{31} \\ \sigma_{31} & \gamma_{33} \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{31} & \alpha_{33} \end{pmatrix} \begin{pmatrix} \gamma_{11} & -\sigma_{31} \\ \sigma_{31} & \gamma_{33} \end{pmatrix}^T = \begin{pmatrix} \hat{\alpha}_{11} & 0 \\ 0 & \hat{\alpha}_{33} \end{pmatrix}.$$

We know this exists since the Spectral Decomposition of the  $2 \times 2$  matrix exists. Such a rotation is called a Jacobi rotation. (Notice that it is different from a Givens' rotation because it diagonalizes a  $2 \times 2$  matrix when used as a unitary similarity transformation. By contrast, a Givens' rotation zeroes an element when applied from one side of a matrix.)

**Homework 15.11** In the above discussion, show that  $\alpha_{11}^2 + 2\alpha_{31}^2 + \alpha_{33}^2 = \hat{\alpha}_{11}^2 + \hat{\alpha}_{33}^2$ .

SEE ANSWER

Jacobi rotation rotations can be used to selectively zero off-diagonal elements by observing the following:

$$\begin{aligned} JAJ^T &= \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \gamma_{11} & 0 & -\sigma_{31} & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & \sigma_{31} & 0 & \gamma_{33} & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix} \begin{pmatrix} A_{00} & a_{10} & A_{20}^T & a_{30} & A_{40}^T \\ a_{10}^T & \alpha_{11} & a_{21}^T & \alpha_{31} & a_{41}^T \\ A_{20} & a_{21} & A_{22} & a_{32} & A_{42}^T \\ a_{30}^T & \alpha_{31} & a_{32}^T & \alpha_{33} & a_{43}^T \\ A_{40} & a_{41} & A_{42} & a_{43} & A_{44} \end{pmatrix} \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \gamma_{11} & 0 & -\sigma_{13} & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & \sigma_{31} & 0 & \gamma_{33} & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix}^T \\ &= \begin{pmatrix} A_{00} & \hat{a}_{10} & A_{20}^T & \hat{a}_{30} & A_{40}^T \\ \hat{a}_{10}^T & \hat{\alpha}_{11} & \hat{a}_{21}^T & 0 & \hat{a}_{41}^T \\ A_{20} & \hat{a}_{21} & A_{22} & \hat{a}_{32} & A_{42}^T \\ \hat{a}_{30}^T & 0 & \hat{a}_{32}^T & \hat{\alpha}_{33} & \hat{a}_{43}^T \\ A_{40} & \hat{a}_{41} & A_{42} & \hat{a}_{43} & A_{44} \end{pmatrix} = \hat{A}, \end{aligned}$$

where

$$\begin{pmatrix} \gamma_{11} & -\sigma_{31} \\ \sigma_{31} & \gamma_{33} \end{pmatrix} \begin{pmatrix} a_{10}^T & a_{21}^T & a_{41}^T \\ a_{30}^T & a_{32}^T & a_{43}^T \end{pmatrix} = \begin{pmatrix} \hat{a}_{10}^T & \hat{a}_{21}^T & \hat{a}_{41}^T \\ \hat{a}_{30}^T & \hat{a}_{32}^T & \hat{a}_{43}^T \end{pmatrix}.$$

Importantly,

$$\begin{aligned} a_{10}^T a_{10} + a_{30}^T a_{30} &= \hat{a}_{10}^T \hat{a}_{10} + \hat{a}_{30}^T \hat{a}_{30} \\ a_{21}^T a_{21} + a_{32}^T a_{32} &= \hat{a}_{21}^T \hat{a}_{21} + \hat{a}_{32}^T \hat{a}_{32} \\ a_{41}^T a_{41} + a_{43}^T a_{43} &= \hat{a}_{41}^T \hat{a}_{41} + \hat{a}_{43}^T \hat{a}_{43}. \end{aligned}$$

What this means is that if one defines  $\text{off}(A)$  as the square of the Frobenius norm of the off-diagonal elements of  $A$ ,

$$\text{off}(A) = \|A\|_F^2 - \|\text{diag}(A)\|_F^2,$$

then  $\text{off}(\hat{A}) = \text{off}(A) - 2\alpha_{31}^2$ .

- The good news: every time a Jacobi rotation is used to zero an off-diagonal element,  $\text{off}(A)$  decreases by twice the square of that element.
- The bad news: a previously introduced zero may become nonzero in the process.

The original algorithm developed by Jacobi searched for the largest (in absolute value) off-diagonal element and zeroed it, repeating this process until all off-diagonal elements were small. The algorithm was applied by hand by one of his students, Seidel (of Gauss-Seidel fame). The problem with this is that searching for the largest off-diagonal element requires  $O(n^2)$  comparisons. Computing and applying one Jacobi rotation as a similarity transformation requires  $O(n)$  flops. Thus, for large  $n$  this is not practical. Instead, it can be shown that zeroing the off-diagonal elements by columns (or rows) also converges to a diagonal matrix. This is known as the column-cyclic Jacobi algorithm. We illustrate this in Figure 15.10.

### 15.9.2 Cuppen's Algorithm

To be added at a future time.

### 15.9.3 The Method of Multiple Relatively Robust Representations (MRRR)

Even once the problem has been reduced to tridiagonal form, the computation of the eigenvalues and eigenvectors via the QR algorithm requires  $O(n^3)$  computations. A method that reduces this to  $O(n^2)$  time (which can be argued to achieve the lower bound for computation, within a constant, because the  $n$  vectors must be at least written) is achieved by the Method of Multiple Relatively Robust Representations (MRRR) by Dhillon and Partlett [13, 12, 14, 15]. The details of that method go beyond the scope of this note.

## 15.10 The Nonsymmetric QR Algorithm

The QR algorithm that we have described can be modified to compute the Schur decomposition of a nonsymmetric matrix. We briefly describe the high-level ideas.

### 15.10.1 A variant of the Schur decomposition

Let  $A \in \mathbb{R}^{n \times n}$  be nonsymmetric. Recall:

- There exists a unitary matrix  $Q \in \mathbb{C}^{n \times n}$  and upper triangular matrix  $R \in \mathbb{C}^{n \times n}$  such that  $A = QRQ^H$ . Importantly: even if  $A$  is real valued, the eigenvalues and eigenvectors may be complex valued.
- The eigenvalues will come in conjugate pairs.

A variation of the Schur Decomposition theorem is

**Theorem 15.12** Let  $A \in \mathbb{R}^{n \times n}$ . Then there exist unitary  $Q \in \mathbb{R}^{n \times n}$  and quasi upper triangular matrix  $R \in \mathbb{R}^{n \times n}$  such that  $A = QRQ^T$ .

Here quasi upper triangular matrix means that the matrix is block upper triangular with blocks of the diagonal that are  $1 \times 1$  or  $2 \times 2$ .

**Remark 15.13** The important thing is that this alternative to the Schur decomposition can be computed using only real arithmetic.

### 15.10.2 Reduction to upperHessenberg form

The basic algorithm for reducing a real-valued nonsymmetric matrix to upperHessenberg form, overwriting the original matrix with the result, can be explained similar to the explanation of the reduction of a symmetric matrix to tridiagonal form. We assume that the upperHessenberg matrix overwrites the upperHessenbert part of  $A$  and the Householder vectors used to zero out parts of  $A$  overwrite the entries that they annihilate (set to zero).

- Assume that the process has proceeded to where  $A = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right)$  with  $A_{00}$  being  $k \times k$  and upperHessenberg,  $a_{10}^T$  being a row vector with only a last nonzero entry, the rest of the submatrices updated according to the application of previous  $k$  Householder transformations.

- Let  $[u_{21}, \tau, a_{21}] := \text{HouseV}(a_{21})$ .<sup>3</sup>

- Update

$$\begin{aligned} \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) &:= \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & 0 & H \end{array} \right) \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline 0 & a_{21} & A_{22} \end{array} \right) \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & 0 & H \end{array} \right) \\ &= \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02}H \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T H \\ \hline 0 & Ha_{21} & HA_{22}H \end{array} \right) \end{aligned}$$

where  $H = H(u_{21})$ .

- Note that  $a_{21} := Ha_{21}$  need not be executed since this update was performed by the instance of HouseV above.<sup>4</sup>

<sup>3</sup> Note that the semantics here indicate that  $a_{21}$  is overwritten by  $Ha_{21}$ .

<sup>4</sup> In practice, the zeros below the first element of  $Ha_{21}$  are not actually written. Instead, the implementation overwrites these elements with the corresponding elements of the vector  $u_{21}$ .

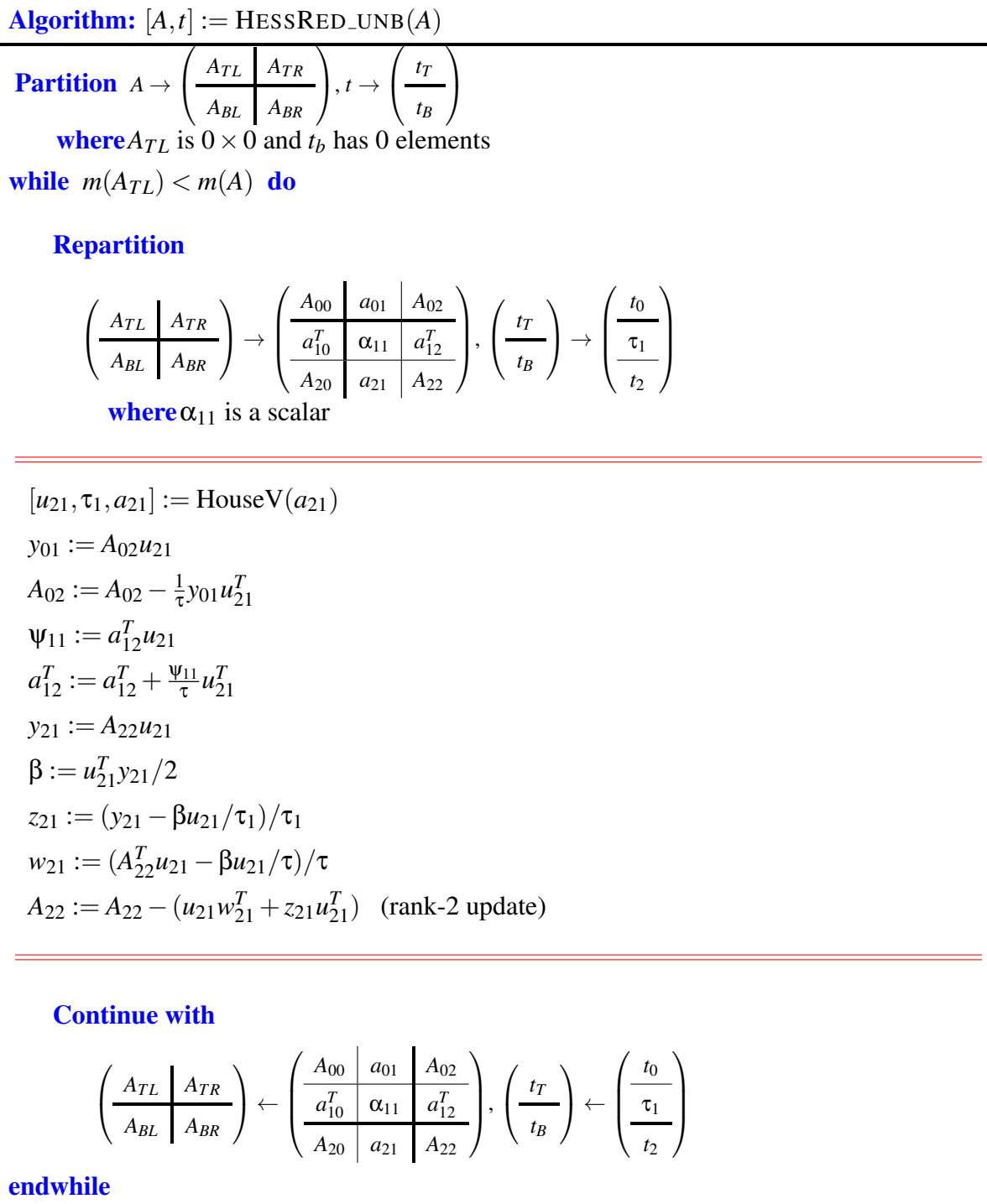


Figure 15.11: Basic algorithm for reduction of a nonsymmetric matrix to upperHessenberg form.

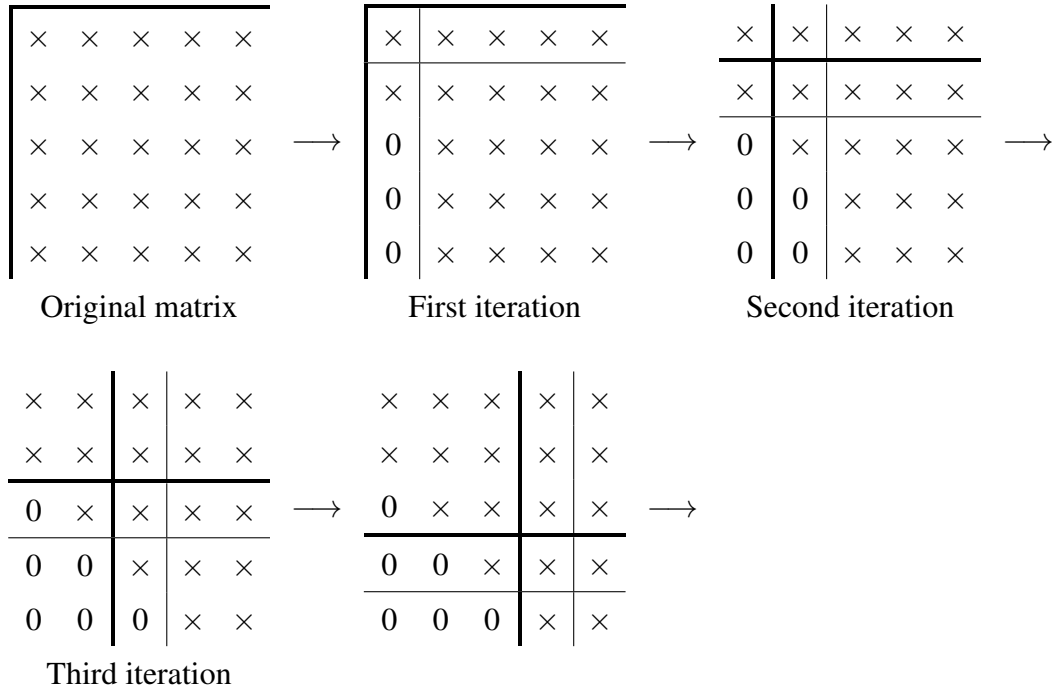


Figure 15.12: Illustration of reduction of a nonsymmetric matrix to upperHessenbert form. The  $\times$ s denote nonzero elements in the matrix.

- The update  $A_{02}H$  requires

$$\begin{aligned}
 A_{02} &:= A_{02}(I - \frac{1}{\tau}u_{21}u_{21}^T) \\
 &= A_{02} - \frac{1}{\tau} \underbrace{A_{02}u_{21}}_{y_{01}} u_{21}^T \\
 &= A_{02} - \frac{1}{\tau} y_{01} u_{21}^T \quad (\text{ger})
 \end{aligned}$$

- The update  $a_{12}^T H$  requires

$$\begin{aligned}
 a_{12}^T &:= a_{12}^T(I - \frac{1}{\tau}u_{21}u_{21}^T) \\
 &= a_{12}^T - \frac{1}{\tau} \underbrace{a_{12}^T u_{21}}_{\psi_{11}} u_{21}^T \\
 &= a_{12}^T - \left(\frac{\psi_{11}}{\tau}\right) u_{21}^T \quad (\text{axpy})
 \end{aligned}$$

- The update of  $A_{22}$  requires

$$A_{22} := (I - \frac{1}{\tau}u_{21}u_{21}^T)A_{22}(I - \frac{1}{\tau}u_{21}u_{21}^T)$$



$$\begin{aligned}
&= (A_{22} - \frac{1}{\tau} u_{21} u_{21}^T A_{22}) (I - \frac{1}{\tau} u_{21} u_{21}^T) \\
&= A_{22} - \frac{1}{\tau} u_{21} u_{21}^T A_{22} - \frac{1}{\tau} A_{22} u_{21} u_{21}^T + \frac{1}{\tau^2} u_{21} \underbrace{u_{21}^T A u_{21}}_{2\beta} u_{21}^T \\
&= A_{22} - \left( \frac{1}{\tau} u_{21} u_{21}^T A_{22} - \frac{\beta}{\tau^2} u_{21} u_{21}^T \right) - \left( \frac{1}{\tau} A u_{21} u_{21}^T - \frac{\beta}{\tau^2} u_{21} u_{21}^T \right) \\
&= A_{22} - \frac{1}{\tau} u_{21} \underbrace{\left( u_{21}^T A_{22} - \frac{\beta}{\tau} u_{21}^T \right)}_{w_{21}^T} - \underbrace{\left( A u_{21} - \frac{\beta}{\tau} u_{21} \right)}_{z_{21}} \frac{1}{\tau} u_{21}^T \\
&= \underbrace{A_{22} - \frac{1}{\tau} u_{21} w_{21}^T - \frac{1}{\tau} z_{21} u_{21}^T}_{\text{rank-2 update}}
\end{aligned}$$

- Continue this process with the updated  $A$ .

It doesn't suffice to only This is captured in the algorithm in Figure 15.11. It is also illustrated in Figure 15.12.

**Homework 15.14** Give the approximate total cost for reducing a nonsymmetric  $A \in \mathbb{R}^{n \times n}$  to upper-Hessenberg form.

👉 SEE ANSWER

For a detailed discussion on the blocked algorithm that uses FLAME notation, we recommend [32]

Gregorio Quintana-Ortí and Robert A. van de Geijn.

**Improving the performance of reduction to Hessenberg form.**

ACM Transactions on Mathematical Software (TOMS) , Vol. 32, No. 2, 2006

and [43]

Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí, G. Joseph Elizondo.

**Families of Algorithms for Reducing a Matrix to Condensed Form.**

ACM Transactions on Mathematical Software (TOMS) , Vol. 39, No. 1, 2012

In those papers, citations to earlier work can be found.

### 15.10.3 The implicitly double-shifted QR algorithm

To be added at a future date!



# Chapter 16

## Notes on the Method of Relatively Robust Representations (MRRR)

The purpose of this note is to give some high level idea of how the Method of Relatively Robust Representations (MRRR) works.

## Outline

<b>Outline</b> . . . . .	<b>268</b>
<b>16.1. MRRR, from 35,000 Feet</b> . . . . .	<b>269</b>
<b>16.2. Cholesky Factorization, Again</b> . . . . .	<b>269</b>
<b>16.3. The <math>LDL^T</math> Factorization</b> . . . . .	<b>272</b>
<b>16.4. The <math>UDU^T</math> Factorization</b> . . . . .	<b>274</b>
<b>16.5. The <math>UDU^T</math> Factorization</b> . . . . .	<b>274</b>
<b>16.6. The Twisted Factorization</b> . . . . .	<b>278</b>
<b>16.7. Computing an Eigenvector from the Twisted Factorization</b> . . . . .	<b>279</b>

## 16.1 MRRR, from 35,000 Feet

The Method of Relatively Robust Representations (MRRR) is an algorithm that, given a tridiagonal matrix, computes eigenvectors associated with that matrix in  $O(n)$  time per eigenvector. This means it computes all eigenvectors in  $O(n^2)$  time, which is much faster than the tridiagonal QR algorithm (which requires  $O(n^3)$  computation). Notice that highly accurate eigenvalues of a tridiagonal matrix can themselves be computed in  $O(n^2)$  time. So, it is legitimate to start by assuming that we have these highly accurate eigenvalues. For our discussion, we only need one.

The MRRR algorithm has at least two benefits of the symmetric QR algorithm for tridiagonal matrices:

- It can compute all eigenvalues and eigenvectors of tridiagonal matrix in  $O(n^2)$  time (versus  $O(n^3)$  time for the symmetric QR algorithm).
- It can efficiently compute eigenvectors corresponding to a subset of eigenvalues.

The benefit when computing all eigenvalues and eigenvectors of a dense matrix is considerably less because transforming the eigenvectors of the tridiagonal matrix back into the eigenvectors of the dense matrix requires an extra  $O(n^3)$  computation, as discussed in [42]. In addition, making the method totally robust has been tricky, since the method does not rely exclusively on unitary similarity transformations.

The fundamental idea is that of a twisted factorization of the tridiagonal matrix. This note builds up to what that factorization is, how to compute it, and how to then compute an eigenvector with it. We start by reminding the reader of what the Cholesky factorization of a symmetric positive definite (SPD) matrix is. Then we discuss the Cholesky factorization of a tridiagonal SPD matrix. This then leads to the  $LDL^T$  factorization of an indefinite and indefinite tridiagonal matrix. Next follows a discussion of the  $UDU^T$  factorization of an indefinite matrix, which then finally yields the twisted factorization. When the matrix is nearly singular, an approximation of the twisted factorization can then be used to compute an approximate eigenvalue.

Notice that the devil is in the details of the MRRR algorithm. We will not tackle those details.

## 16.2 Cholesky Factorization, Again

We have discussed in class the Cholesky factorization,  $A = LL^T$ , which requires a matrix to be symmetric positive definite. (We will restrict our discussion to real matrices.) The following computes the Cholesky factorization:

- Partition  $A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{21}^T \\ \hline a_{21} & A_{22} \end{array} \right)$ .
- Update  $\alpha_{11} := \sqrt{\alpha_{11}}$ .
- Update  $a_{21} := a_{21}/\alpha_{11}$ .
- Update  $A_{22} := A_{22} - a_{21}a_{21}^T$  (updating only the lower triangular part).
- Continue to compute the Cholesky factorization of the updated  $A_{22}$ .

**Algorithm:**  $A := \text{CHOL}(A)$

---

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right)$   
**where**  $A_{TL}$  is  $0 \times 0$   
**while**  $m(A_{TL}) < m(A)$  **do**  
**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$


---

$\alpha_{11} := \sqrt{\alpha_{11}}$   
 $a_{21} := a_{21} / \alpha_{11}$   
 $A_{22} := A_{22} - a_{21} a_{21}^T$

---

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & \star \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & \star & \star \\ \hline a_{10}^T & \alpha_{11} & \star \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right)$$

**endwhile**

Figure 16.1: Unblocked algorithm for computing the Cholesky factorization. Updates to  $A_{22}$  affect only the lower triangular part.

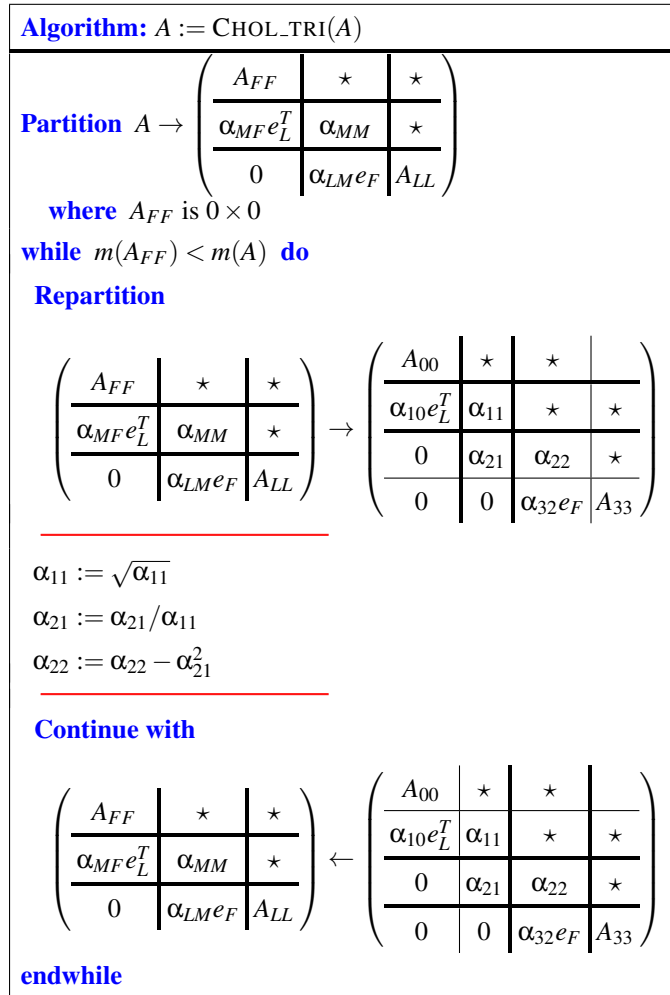


Figure 16.2: Algorithm for computing the the Cholesky factorization of a tridiagonal matrix.

The resulting algorithm is given in Figure 16.1.

In the special case where  $A$  is tridiagonal and SPD, the algorithm needs to be modified so that it can take advantage of zero elements:

- Partition  $A \rightarrow \left( \begin{array}{c|c|c} \alpha_{11} & \star & \star \\ \hline \alpha_{21} & \alpha_{22} & \star \\ \hline 0 & \alpha_{32}e_F & A_{33} \end{array} \right)$ , where  $\star$  indicates the symmetric part that is not stored. Here  $e_F$  indicates the unit basis vector with a “1” as first element.
- Update  $\alpha_{11} := \sqrt{\alpha_{11}}$ .
- Update  $\alpha_{21} := \alpha_{21}/\alpha_{11}$ .
- Update  $\alpha_{22} := \alpha_{22} - \alpha_{21}^2$ .
- Continue to compute the Cholesky factorization of  $\left( \begin{array}{c|c} \alpha_{22} & \star \\ \hline \alpha_{32}e_F & A_{33} \end{array} \right)$

The resulting algorithm is given in Figure 16.2. In that figure, it helps to interpret  $F$ ,  $M$ , and  $L$  as First, Middle, and Last. In that figure,  $e_F$  and  $e_L$  are the unit basis vectors with a “1” as first and last element, respectively. Notice that the Cholesky factor of a tridiagonal matrix is a lower bidiagonal matrix that overwrites the lower triangular part of the tridiagonal matrix  $A$ .

Naturally, the whole matrix needs not be stored. But that detail is not important for our discussion.

### 16.3 The $LDL^T$ Factorization

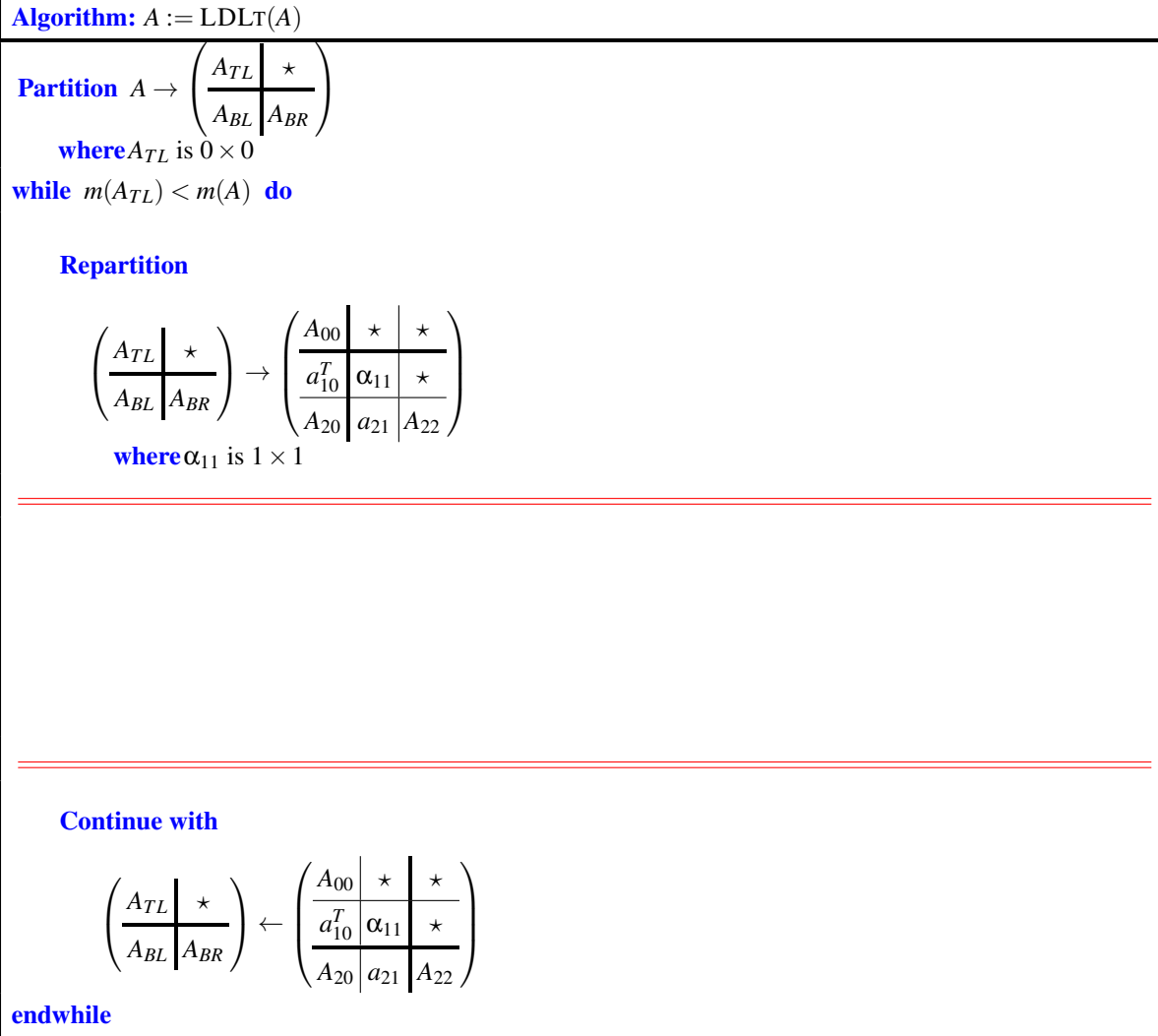
Now, one can alternatively compute  $A = LDL^T$ , where  $L$  is unit lower triangular and  $D$  is diagonal. We will look at how this is done first and will then note that this factorization can be computed for any indefinite (nonsingular) symmetric matrix. (Notice: the so computed  $L$  is *not* the same as the  $L$  computed by the Cholesky factorization.) Partition

$$A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{21}^T \\ \hline a_{21} & A_{22} \end{array} \right), L \rightarrow \left( \begin{array}{c|c} 1 & 0 \\ \hline l_{21} & L_{22} \end{array} \right), \text{ and } D \rightarrow \left( \begin{array}{c|c} \delta_1 & 0 \\ \hline 0 & D_{22} \end{array} \right).$$

Then

$$\begin{aligned} \left( \begin{array}{c|c} \alpha_{11} & a_{21}^T \\ \hline a_{21} & A_{22} \end{array} \right) &= \left( \begin{array}{c|c} 1 & 0 \\ \hline l_{21} & L_{22} \end{array} \right) \left( \begin{array}{c|c} \delta_1 & 0 \\ \hline 0 & D_{22} \end{array} \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline l_{21} & L_{22} \end{array} \right)^T \\ &= \left( \begin{array}{c|c} \delta_{11} & 0 \\ \hline \delta_{11}l_{21} & L_{22}D_{22} \end{array} \right) \left( \begin{array}{c|c} 1 & l_{21}^T \\ \hline 0 & L_{22}^T \end{array} \right) \\ &= \left( \begin{array}{c|c} \delta_{11} & \star \\ \hline \delta_{11}l_{21} & \delta_{11}l_{21}l_{21}^T + L_{22}D_{22}L_{22}^T \end{array} \right) \end{aligned}$$



Figure 16.3: Unblocked algorithm for computing  $A \rightarrow LDL^T$ , overwriting  $A$ .

This suggests the following algorithm for overwriting the strictly lower triangular part of  $A$  with the strictly lower triangular part of  $L$  and the diagonal of  $A$  with  $D$ :

- Partition  $A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{21}^T \\ \hline a_{21} & A_{22} \end{array} \right)$ .
- $\alpha_{11} := \delta_{11} = \alpha_{11}$  (no-op).
- Compute  $l_{21} := a_{21}/\alpha_{11}$ .
- Update  $A_{22} := A_{22} - l_{21}a_{21}^T$  (updating only the lower triangular part).
- $a_{21} := l_{21}$ .
- Continue with computing  $A_{22} \rightarrow L_{22}D_{22}L_{22}^T$ .

This algorithm will complete as long as  $\delta_{11} \neq 0$ , which happens when  $A$  is nonsingular. This is equivalent to  $A$  does not having *zero* as an eigenvalue. Such a matrix is also called *indefinite*.

**Homework 16.1** Modify the algorithm in Figure 16.1 so that it computes the  $LDL^T$  factorization. (Fill in Figure 16.3.)

☛ [SEE ANSWER](#)

**Homework 16.2** Modify the algorithm in Figure 16.2 so that it computes the  $LDL^T$  factorization of a tridiagonal matrix. (Fill in Figure 16.4.) What is the approximate cost, in floating point operations, of computing the  $LDL^T$  factorization of a tridiagonal matrix? Count a divide, multiply, and add/subtract as a floating point operation each. Show how you came up with the algorithm, similar to how we derived the algorithm for the tridiagonal Cholesky factorization.

☛ [SEE ANSWER](#)

Notice that computing the  $LDL^T$  factorization of an indefinite matrix is not a good idea when  $A$  is nearly singular. This would lead to some  $\delta_{11}$  being nearly zero, meaning that dividing by it leads to very large entries in  $l_{21}$  and corresponding element growth in  $A_{22}$ . (In other words, strange things will happen “down stream” from the small  $\delta_{11}$ .) Not a good thing. Unfortunately, we are going to need something like this factorization specifically for the case where  $A$  is nearly singular.

## 16.4 The $UDU^T$ Factorization

The fact that the LU, MRRResky, and  $LDL^T$  factorizations start in the top-left corner of the matrix and work towards the bottom-right is an accident of history. Gaussian elimination works in that direction, hence so does LU factorization, and the rest kind of follow.

One can imagine, given a SPD matrix  $A$ , instead computing  $A = UU^T$ , where  $U$  is upper triangular. Such a computation starts in the lower-right corner of the matrix and works towards the top-left. Similarly, an algorithm can be created for computing  $A = UDU^T$  where  $U$  is unit upper triangular and  $D$  is diagonal.

**Homework 16.3** Derive an algorithm that, given an indefinite matrix  $A$ , computes  $A = UDU^T$ . Overwrite only the upper triangular part of  $A$ . (Fill in Figure 16.5.) Show how you came up with the algorithm, similar to how we derived the algorithm for  $LDL^T$ .

☛ [SEE ANSWER](#)

## 16.5 The $UDU^T$ Factorization

**Homework 16.4** Derive an algorithm that, given an indefinite tridiagonal matrix  $A$ , computes  $A = UDU^T$ . Overwrite only the upper triangular part of  $A$ . (Fill in Figure 16.6.) Show how you came up with the algorithm, similar to how we derived the algorithm for  $LDL^T$ .

☛ [SEE ANSWER](#)

Notice that the  $UDU^T$  factorization has the exact same shortcomings as does  $LDL^T$  when applied to a singular matrix. If  $D$  has a small element on the diagonal, it is again “down stream” that this can cause large elements in the factored matrix. But now “down stream” means towards the top-left since the algorithm moves in the opposite direction.

**Algorithm:**  $A := \text{LDLT\_TRI}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c|c} A_{FF} & \star & \star \\ \hline \alpha_{MF} e_L^T & \alpha_{MM} & \star \\ \hline 0 & \alpha_{LM} e_F & A_{LL} \end{array} \right)$

**where**  $A_{LL}$  is  $0 \times 0$

**while**  $m(A_{FF}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c|c} A_{FF} & \star & \star \\ \hline \alpha_{MF} e_L^T & \alpha_{MM} & \star \\ \hline 0 & \alpha_{LM} e_F & A_{LL} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c|c} A_{00} & \star & \star & \\ \hline \alpha_{10} e_L^T & \alpha_{11} & \star & \star \\ \hline 0 & \alpha_{21} & \alpha_{22} & \star \\ \hline 0 & 0 & \alpha_{32} e_F & A_{33} \end{array} \right)$$

**where**  $\alpha_{22}$  is a scalar

**Continue with**

$$\left( \begin{array}{c|c|c} A_{FF} & \star & \star \\ \hline \alpha_{MF} e_L^T & \alpha_{MM} & \star \\ \hline 0 & \alpha_{LM} e_F & A_{LL} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c|c} A_{00} & \star & \star & \\ \hline \alpha_{10} e_L^T & \alpha_{11} & \star & \star \\ \hline 0 & \alpha_{21} & \alpha_{22} & \star \\ \hline 0 & 0 & \alpha_{32} e_F & A_{33} \end{array} \right)$$

**endwhile**

Figure 16.4: Algorithm for computing the the  $LDL^T$  factorization of a tridiagonal matrix.

**Algorithm:**  $A := \text{UDUT}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline \star & A_{BR} \end{array} \right)$

**where**  $A_{BR}$  is  $0 \times 0$

**while**  $m(A_{BR}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline \star & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline \star & \alpha_{11} & a_{12}^T \\ \hline \star & \star & A_{22} \end{array} \right)$$

**where**  $\alpha_{11}$  is  $1 \times 1$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline \star & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline \star & \alpha_{11} & a_{12}^T \\ \hline \star & \star & A_{22} \end{array} \right)$$

**endwhile**

Figure 16.5: Unblocked algorithm for computing  $A = \text{UDUT}$ . Updates to  $A_{00}$  affect only the upper triangular part.

**Algorithm:**  $A := \text{UDUT\_TRI}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c|c} A_{FF} & \alpha_{FM}e_L^T & 0 \\ \hline * & \alpha_{MM} & \alpha_{ML}e_F^T \\ \hline * & * & A_{LL} \end{array} \right)$

**where**  $A_{FF}$  is  $0 \times 0$

**while**  $m(A_{LL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c|c} A_{FF} & \alpha_{FM}e_L & 0 \\ \hline * & \alpha_{MM} & \alpha_{ML}e_F^T \\ \hline * & * & A_{LL} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c|c} A_{00} & \alpha_{01}e_L & 0 & 0 \\ \hline * & \alpha_{11} & \alpha_{12} & 0 \\ \hline * & * & \alpha_{22} & \alpha_{23}e_F^T \\ \hline * & * & * & A_{33} \end{array} \right)$$

**where**

**Continue with**

$$\left( \begin{array}{c|c|c} A_{FF} & \alpha_{FM}e_L & 0 \\ \hline * & \alpha_{MM} & \alpha_{ML}e_F^T \\ \hline * & * & A_{LL} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c|c} A_{00} & \alpha_{01}e_L & 0 & 0 \\ \hline * & \alpha_{11} & \alpha_{12} & 0 \\ \hline * & * & \alpha_{22} & \alpha_{23}e_F^T \\ \hline * & * & * & A_{33} \end{array} \right)$$

**endwhile**

Figure 16.6: Algorithm for computing the the  $UDU^T$  factorization of a tridiagonal matrix.

## 16.6 The Twisted Factorization

Let us assume that  $A$  is a tridiagonal matrix and that we are *given* one of its eigenvalues (or rather, a very good approximation),  $\lambda$ , and let us assume that this eigenvalue has multiplicity one. Indeed, we are going to assume that it is well-separated meaning there is no other eigenvalue close to it.

Here is a way to compute an associated eigenvector: Find a nonzero vector in the null space of  $B = A - \lambda I$ . Let us think back how one was taught how to do this:

- Reduce  $B$  to row-echelon form. This may will require pivoting.
- Find a column that has no pivot in it. This identifies an independent (free) variable.
- Set the element in vector  $x$  corresponding to the independent variable to *one*.
- Solve for the dependent variables.

There are a number of problems with this approach:

- It does not take advantage of symmetry.
- It is inherently unstable since you are working with a matrix,  $B = A - \lambda I$  that inherently has a bad condition number. (Indeed, it is infinity if  $\lambda$  is an exact eigenvalue.)
- $\lambda$  is not an exact eigenvalue when it is computed by a computer and hence  $B$  is not exactly singular. So, no independent variables will even be found.

These are only some of the problems. To overcome this, one computes something called a twisted factorization.

Again, let  $B$  be a tridiagonal matrix. Assume that  $\lambda$  is an approximate eigenvalue of  $B$  and let  $A = B - \lambda I$ . We are going to compute an approximate eigenvector of  $B$  associated with  $\lambda$  by computing a vector that is in the null space of a singular matrix that is close to  $A$ . We will assume that  $\lambda$  is an eigenvalue of  $B$  that has multiplicity one, and is well-separated from other eigenvalues. Thus, the singular matrix close to  $A$  has a null space with dimension one. Thus, we would expect  $A = LDL^T$  to have one element on the diagonal of  $D$  that is essentially zero. Ditto for  $A = UEU^T$ , where  $E$  is diagonal and we use this letter to be able to distinguish the two diagonal matrices.

Thus, we have

- $\lambda$  is an approximate (but not exact) eigenvalue of  $B$ .
- $A = B - \lambda I$  is indefinite.
- $A = LDL^T$ .  $L$  is bidiagonal, unit lower triangular, and  $D$  is diagonal with one small element on the diagonal.
- $A = UEU^T$ .  $U$  is bidiagonal, unit upper triangular, and  $E$  is diagonal with one small element on the diagonal.

Let

$$A = \left( \begin{array}{c|c|c} A_{00} & \alpha_{10}e_L & 0 \\ \hline \alpha_{10}e_L^T & \alpha_{11} & \alpha_{21}e_F^T \\ \hline 0 & \alpha_{21}e_F & A_{22} \end{array} \right), L = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & 0 \\ \hline 0 & \lambda_{21}e_F & L_{22} \end{array} \right), U = \left( \begin{array}{c|c|c} U_{00} & v_{01}e_L & 0 \\ \hline 0 & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right),$$

$$D = \left( \begin{array}{c|c|c} D_{00} & 0 & 0 \\ \hline 0 & \delta_1 & 0 \\ \hline 0 & 0 & D_{22} \end{array} \right), \text{ and } E = \left( \begin{array}{c|c|c} E_{00} & 0 & 0 \\ \hline 0 & \varepsilon_1 & 0 \\ \hline 0 & 0 & E_{22} \end{array} \right),$$

where all the partitioning is “conformal”.

**Homework 16.5** Show that, provided  $\phi_1$  is chosen appropriately,

$$\left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right) \left( \begin{array}{c|c|c} D_{00} & 0 & 0 \\ \hline 0 & \phi_1 & 0 \\ \hline 0 & 0 & E_{22} \end{array} \right) \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right)^T$$

$$= \left( \begin{array}{c|c|c} A_{00} & \alpha_{01}e_L & 0 \\ \hline \alpha_{01}e_L^T & \alpha_{11} & \alpha_{21}e_F^T \\ \hline 0 & \alpha_{21}e_F & A_{22} \end{array} \right).$$

(Hint: multiply out  $A = LDL^T$  and  $A = UEU^T$  with the partitioned matrices first. Then multiply out the above. Compare and match...) How should  $\phi_1$  be chosen? What is the cost of computing the twisted factorization given that you have already computed the  $LDL^T$  and  $UDU^T$  factorizations? A “Big O” estimate is sufficient. Be sure to take into account what  $e_L^T D_{00} e_L$  and  $e_F^T E_{22} e_F$  equal in your cost estimate.

• SEE ANSWER

## 16.7 Computing an Eigenvector from the Twisted Factorization

The way the method now works for computing the desired approximate eigenvector is to find the  $\phi_1$  that is smallest in value of all possible such values. In other words, you can partition  $A$ ,  $L$ ,  $U$ ,  $D$ , and  $E$  in many ways, singling out any of the diagonal values of these matrices. The partitioning chosen is the one that makes  $\phi_1$  the smallest of all possibilities. It is then set to zero so that

$$\left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right) \left( \begin{array}{c|c|c} D_{00} & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & E_{22} \end{array} \right) \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right)^T \approx \left( \begin{array}{c|c|c} A_{00} & \alpha_{01}e_L & 0 \\ \hline \alpha_{01}e_L^T & \alpha_{11} & \alpha_{21}e_F^T \\ \hline 0 & \alpha_{21}e_F & A_{22} \end{array} \right).$$

Because  $\lambda$  was assumed to have multiplicity one and well-separated, the resulting twisted factorization has “nice” properties. We won’t get into what that exactly means.

**Homework 16.6** Compute  $x_0$ ,  $\chi_1$ , and  $x_2$  so that

$$\left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right) \left( \begin{array}{c|c|c} D_{00} & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & E_{22} \end{array} \right) \underbrace{\left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right)^T \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix}}_{\text{Hint: } \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

where  $x = \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix}$  is not a zero vector. What is the cost of this computation, given that  $L_{00}$  and  $U_{22}$  have special structure?

➡ [SEE ANSWER](#)

The vector  $x$  that is so computed is the desired approximate eigenvector of  $B$ .

Now, if the eigenvalues of  $B$  are well separated, and one follows essentially this procedure to find eigenvectors associated with each eigenvalue, the resulting eigenvectors are quite orthogonal to each other. We know that the eigenvectors of a symmetric matrix with distinct eigenvalues should be orthogonal, so this is a desirable property.

The approach becomes very tricky when eigenvalues are “clustered”, meaning that some of them are very close to each other. A careful scheme that shifts the matrix is then used to make eigenvalues in a cluster relatively well separated. But that goes beyond this note.



# Chapter 17

## Notes on Computing the SVD of a Matrix

For simplicity we will focus on the case where  $A \in \mathbb{R}^{n \times n}$ . We will assume that the reader has read Chapter [15](#).

## Outline

<b>Outline</b> . . . . .	<b>282</b>
<b>17.1. Background</b> . . . . .	<b>283</b>
<b>17.2. Reduction to Bidiagonal Form</b> . . . . .	<b>283</b>
<b>17.3. The QR Algorithm with a Bidiagonal Matrix</b> . . . . .	<b>287</b>
<b>17.4. Putting it all together</b> . . . . .	<b>290</b>

## 17.1 Background

Recall:

**Theorem 17.1** If  $A \in \mathbb{R}^{m \times n}$  then there exists unitary matrices  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$ , and diagonal matrix  $\Sigma \in \mathbb{R}^{m \times n}$  such that  $A = U\Sigma V^T$ . This is known as the Singular Value Decomposition.

There is a relation between the SVD of a matrix  $A$  and the Spectral Decomposition of  $A^T A$ :

**Homework 17.2** If  $A = U\Sigma V^T$  is the SVD of  $A$  then  $A^T A = V\Sigma^2 V^T$  is the Spectral Decomposition of  $A^T A$ .

🔗 [SEE ANSWER](#)

The above exercise suggests steps for computing the Reduced SVD:

- Form  $C = A^T A$ .
- Compute unitary  $V$  and diagonal  $\Lambda$  such that  $C = Q\Lambda Q^T$ , ordering the eigenvalues from largest to smallest on the diagonal of  $\Lambda$ .
- Form  $W = AV$ . Then  $W = U\Sigma$  so that  $U$  and  $\Sigma$  can be computed from  $W$  by choosing the diagonal elements of  $\Sigma$  to equal the lengths of the corresponding columns of  $W$  and normalizing those columns by those lengths.

The problem with this approach is that forming  $A^T A$  squares the condition number of the problem. We will show how to avoid this.

## 17.2 Reduction to Bidiagonal Form

In the last chapter, we saw that it is beneficial to start by reducing a matrix to tridiagonal or upper Hessenberg form when computing the Spectral or Schur decompositions. The corresponding step when computing the SVD of a given matrix is to reduce the matrix to bidiagonal form.

We assume that the bidiagonal matrix overwrites matrix  $A$ . Householder vectors will again play a role, and will again overwrite elements that are annihilated (set to zero).

- Partition  $A \rightarrow \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right)$ . In the first iteration, this can be visualized as

×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	×
×	×	×	×

- We introduce zeroes below the “diagonal” in the first column by computing a Householder transformation and applying this from the left:

**Algorithm:**  $[A, t, r] := \text{BiDRED\_UNB}(A)$

**Partition**  $A \rightarrow \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right), t \rightarrow \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right), r \rightarrow \left( \begin{array}{c} r_T \\ \hline r_B \end{array} \right)$

**where**  $A_{TL}$  is  $0 \times 0$  and  $t_b$  has 0 elements

**while**  $m(A_{TL}) < m(A)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \rightarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right), \left( \begin{array}{c} r_T \\ \hline r_B \end{array} \right) \rightarrow \left( \begin{array}{c} r_0 \\ \hline \rho_1 \\ \hline r_2 \end{array} \right)$$

**where**  $\alpha_{11}$ ,  $\tau_1$ , and  $\rho_1$  are scalars

$$\left[ \left( \begin{array}{c} 1 \\ \hline u_{21} \end{array} \right), \tau_1, \left( \begin{array}{c} \alpha_{11} \\ \hline a_{21} \end{array} \right) \right] := \text{HouseV} \left( \begin{array}{c} \alpha_{11} \\ \hline a_{21} \end{array} \right)$$

$$a_{12}^T := a_{12}^T - w_{12}^T$$

$$z_{21} := (y_{21} - \beta u_{21} / \tau_1) / \tau_1$$

$$A_{22} := A_{22} - u_{21} w_{21}^T \quad (\text{rank-1 update})$$

$$[v_{12}, \rho_1, a_{12}] := \text{HouseV}(a_{12})$$

$$w_{21} := A_{22} v_{12} / \rho_1$$

$$A_{22} := A_{22} - w_{21} v_{12}^T \quad (\text{rank-1 update})$$

**Continue with**

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \left( \begin{array}{c} t_T \\ \hline t_B \end{array} \right) \leftarrow \left( \begin{array}{c} t_0 \\ \hline \tau_1 \\ \hline t_2 \end{array} \right), \left( \begin{array}{c} r_T \\ \hline r_B \end{array} \right) \leftarrow \left( \begin{array}{c} r_0 \\ \hline \rho_1 \\ \hline r_2 \end{array} \right)$$

**endwhile**

Figure 17.1: Basic algorithm for reduction of a nonsymmetric matrix to bidiagonal form.

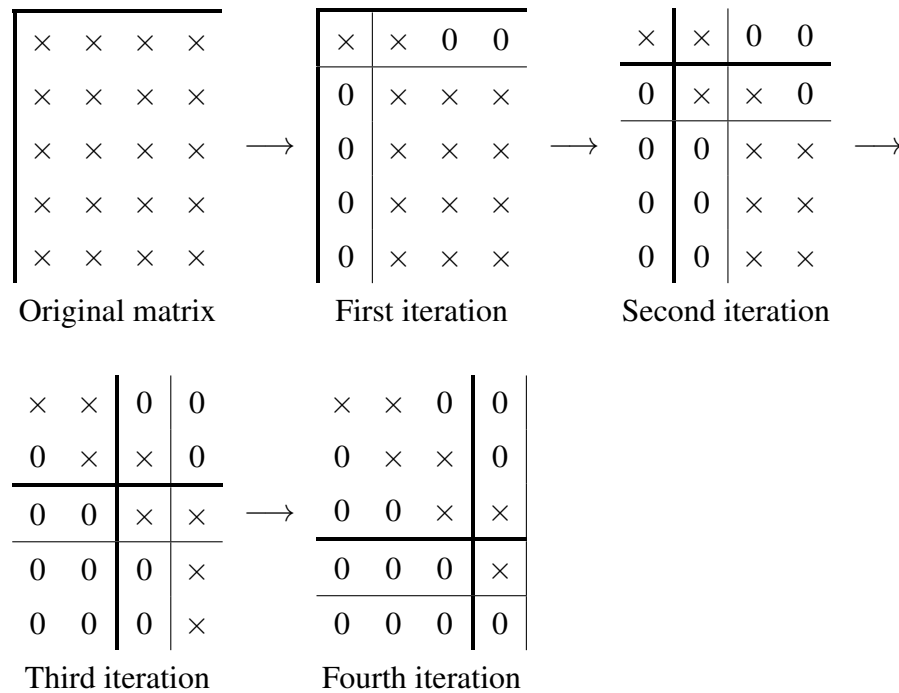


Figure 17.2: Illustration of reduction to bidiagonal form form.

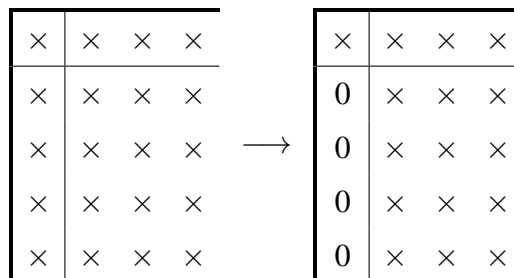
- Let  $\left[ \begin{pmatrix} 1 \\ u_{21} \end{pmatrix}, \tau_1, \begin{pmatrix} \alpha_{11} \\ a_{21} \end{pmatrix} \right] := \text{HouseV} \left( \begin{pmatrix} \alpha_{11} \\ a_{21} \end{pmatrix} \right)$ . This not only computes the Householder vector, but also zeroes the entries in  $a_{21}$  and updates  $\alpha_{11}$ .
- Update

$$\left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) := \left( I - \frac{1}{\tau_1} \begin{pmatrix} 1 \\ u_{21} \end{pmatrix} \begin{pmatrix} 1 \\ u_{21} \end{pmatrix}^T \right) \left( \begin{array}{c|c} \alpha_{11} & a_{12}^T \\ \hline a_{21} & A_{22} \end{array} \right) = \left( \begin{array}{c|c} \tilde{\alpha}_{11} & \tilde{a}_{12}^T \\ \hline 0 & \tilde{A}_{22} \end{array} \right)$$

where

- \*  $\tilde{\alpha}_{11}$  is updated as part of the computation of the Householder vector;
- \*  $w_{12}^T := (a_{12}^T + u_{21}^T A_{22}) / \tau_1$ ;
- \*  $\tilde{a}_{12}^T := a_{12}^T - w_{12}^T$ ; and
- \*  $\tilde{A}_{22} := A_{22} - u_{21} w_{12}^T$ .

This introduces zeroes below the "diagonal" in the first column:



The zeroes below  $\alpha_{11}$  are not actually written. Instead, the implementation overwrites these elements with the corresponding elements of the vector  $u_{21}^T$ .

- Next, we introduce zeroes in the first row to the right of the element on the superdiagonal by computing a Householder transformation from  $\tilde{a}_{12}^T$ :
  - Let  $[v_{12}, \rho_1, a_{12}] := \text{HouseV}(a_{12})$ . This not only computes the Householder vector, but also zeroes all but the first entry in  $\tilde{a}_{12}^T$ , updating that first entry.
  - Update

$$\begin{pmatrix} \frac{a_{12}^T}{A_{22}} \\ \vdots \end{pmatrix} := \begin{pmatrix} \frac{a_{12}^T}{A_{22}} \\ \vdots \end{pmatrix} \left( I - \frac{1}{\rho_1} v_{12} v_{12}^T \right) = \begin{pmatrix} \frac{\tilde{\alpha}_{12} e_0^T}{\tilde{A}_{22}} \\ \vdots \end{pmatrix}$$

where

- \*  $\tilde{\alpha}_{12}$  equals the first element of the updated  $a_{12}^T$ ;
- \*  $w_{21} := A_{22} v_{12} / \rho_1$ ;
- \*  $\tilde{A}_{22} := A_{22} - w_{21} v_{12}^T$ .

This introduces zeroes to the right of the "superdiagonal" in the first row:

$$\begin{array}{c|cccc} \times & \times & \times & \times & \\ \hline 0 & \times & \times & \times & \\ 0 & \times & \times & \times & \\ 0 & \times & \times & \times & \\ 0 & \times & \times & \times & \end{array} \longrightarrow \begin{array}{c|cccc} \times & \times & 0 & 0 & \\ \hline 0 & \times & \times & \times & \\ 0 & \times & \times & \times & \\ 0 & \times & \times & \times & \\ 0 & \times & \times & \times & \end{array}$$

- The zeros to the right of the first element of  $a_{12}^T$  are not actually written. Instead, the implementation overwrites these elements with the corresponding elements of the vector  $v_{12}^T$ .
- Continue this process with the updated  $A_{22}$ .

The above observations are captured in the algorithm in Figure 17.1. It is also illustrated in Figure 17.2.

**Homework 17.3 Homework 17.4** Give the approximate total cost for reducing  $A \in \mathbb{R}^{n \times n}$  to bidiagonal form.

🔗 [SEE ANSWER](#)

For a detailed discussion on the blocked algorithm that uses FLAME notation, we recommend [43]

Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí, G. Joseph Elizondo.

**Families of Algorithms for Reducing a Matrix to Condensed Form.**

ACM Transactions on Mathematical Software (TOMS) , Vol. 39, No. 1, 2012

## 17.3 The QR Algorithm with a Bidiagonal Matrix

Let  $B = U_B^T A V_B$  be bidiagonal where  $U_B$  and  $V_B$  have orthonormal columns.

**Lemma 17.5** *Let  $B \in \mathbb{R}^{m \times n}$  be upper bidiagonal and  $T \in \mathbb{R}^{n \times n}$  with  $T = B^T B$ . Then  $T$  is tridiagonal.*

**Proof:** Partition

$$B = \left( \begin{array}{c|c} B_{00} & \beta_{10} e_l \\ \hline 0 & \beta_{11} \end{array} \right),$$

where  $e_l$  denotes the unit basis vector with a “1” in the last entry. Then

$$\begin{aligned} B^T B &= \left( \begin{array}{c|c} B_{00} & \beta_{10} e_l \\ \hline 0 & \beta_{11} \end{array} \right)^T \left( \begin{array}{c|c} B_{00} & \beta_{10} e_l \\ \hline 0 & \beta_{11} \end{array} \right) = \left( \begin{array}{c|c} B_{00}^T & 0 \\ \hline \beta_{10} e_l^T & \beta_{11} \end{array} \right) \left( \begin{array}{c|c} B_{00} & \beta_{10} e_l \\ \hline 0 & \beta_{11} \end{array} \right) \\ &= \left( \begin{array}{c|c} B_{00}^T B_{00} & \beta_{10} B_{00}^T e_l \\ \hline \beta_{10} e_l^T B_{00} & \beta_{10}^2 + \beta_{11}^2 \end{array} \right), \end{aligned}$$

where  $\beta_{10} B_{00}^T e_l$  is (clearly) a vector with only a nonzero in the last entry.

The above proof also shows that if  $B$  has no zeroes on its superdiagonal, then neither does  $B^T B$ .

This means that one can apply the QR algorithm to matrix  $B^T B$ . The problem, again, is that this squares the condition number of the problem.

The following extension of the Implicit Q Theorem comes to the rescue:

**Theorem 17.6** *Let  $C, B \in \mathbb{R}^{m \times n}$ . If there exist unitary matrices  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$  so that  $B = U^T C V$  is upper bidiagonal and has positive values on its superdiagonal then  $B$  and  $V$  are uniquely determined by  $C$  and the first column of  $V$ .*

The above theorem supports an algorithm that, starting with an upper bidiagonal matrix  $B$ , implicitly performs a shifted QR algorithm with  $T = B^T B$ .

Consider the  $5 \times 5$  bidiagonal matrix

$$B = \begin{pmatrix} \beta_{0,0} & \beta_{0,1} & 0 & 0 & 0 \\ 0 & \beta_{1,1} & \beta_{1,2} & 0 & 0 \\ 0 & 0 & \beta_{2,2} & \beta_{2,3} & 0 \\ 0 & 0 & 0 & \beta_{3,3} & \beta_{3,4} \\ 0 & 0 & 0 & 0 & \beta_{4,4} \end{pmatrix}.$$

Then  $T = B^T B$  equals

$$T = \begin{pmatrix} \beta_{0,0}^2 & \star & 0 & 0 & 0 \\ \beta_{0,1} \beta_{0,0} & \star & \star & 0 & 0 \\ 0 & 0 & \star & \star & 0 \\ 0 & 0 & 0 & \star & \star \\ 0 & 0 & 0 & 0 & \beta_{3,4}^2 + \beta_{4,4}^2 \end{pmatrix},$$

where the  $\star$ s indicate “don’t care” entries. Now, *if* an iteration of the implicitly shifted QR algorithm were executed with this matrix, then the shift would be  $\mu_k = \beta_{3,4}^2 + \beta_{4,4}^2$  (actually, it is usually computed from the bottom-right  $2 \times 2$  matrix, a minor detail) and the first Givens’ rotation would be computed so that

$$\begin{pmatrix} \gamma_{0,1} & \sigma_{0,1} \\ -\sigma_{0,1} & \gamma_{0,1} \end{pmatrix} \begin{pmatrix} \beta_{0,0}^2 - \mu I \\ \beta_{0,1}\beta_{1,1} \end{pmatrix}$$

has a zero second entry. Let us call the  $n \times n$  matrix with this as its top-left submatrix  $G_0$ . Then applying this from the left and right of  $T$  means forming  $G_0 T G_0 = G_0 B^T B G_0^T = (B G_0^T)^T (B G_0^T)$ . What if we only apply it to  $B$ ? Then

$$\begin{pmatrix} \tilde{\beta}_{0,0} & \tilde{\beta}_{0,1} & 0 & 0 & 0 \\ \tilde{\beta}_{1,0} & \tilde{\beta}_{1,1} & \beta_{1,2} & 0 & 0 \\ 0 & 0 & \beta_{2,2} & \beta_{2,3} & 0 \\ 0 & 0 & 0 & \beta_{3,3} & \beta_{3,4} \\ 0 & 0 & 0 & 0 & \beta_{4,4} \end{pmatrix} = \begin{pmatrix} \beta_{0,0} & \beta_{0,1} & 0 & 0 & 0 \\ 0 & \beta_{1,1} & \beta_{1,2} & 0 & 0 \\ 0 & 0 & \beta_{2,2} & \beta_{2,3} & 0 \\ 0 & 0 & 0 & \beta_{3,3} & \beta_{3,4} \\ 0 & 0 & 0 & 0 & \beta_{4,4} \end{pmatrix} \begin{pmatrix} \gamma_{0,1} & -\sigma_{0,1} & 0 & 0 & 0 \\ \sigma_{0,1} & \gamma_{0,1} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The idea now is to apply Givens’ rotation from the left and right to “chase the bulge” except that now the rotations applied from both sides need not be the same. Thus, next, from  $\begin{pmatrix} \tilde{\beta}_{0,0} \\ \tilde{\beta}_{1,0} \end{pmatrix}$  one can compute  $\gamma_{1,0}$

and  $\sigma_{1,0}$  so that  $\begin{pmatrix} \gamma_{1,0} & \sigma_{1,0} \\ -\sigma_{1,0} & \gamma_{1,0} \end{pmatrix} \begin{pmatrix} \tilde{\beta}_{0,0} \\ \tilde{\beta}_{1,0} \end{pmatrix} = \begin{pmatrix} \tilde{\tilde{\beta}}_{0,0} \\ 0 \end{pmatrix}$ . Then

$$\begin{pmatrix} \tilde{\tilde{\beta}}_{0,0} & \tilde{\tilde{\beta}}_{0,1} & \tilde{\tilde{\beta}}_{0,2} & 0 & 0 \\ 0 & \tilde{\tilde{\beta}}_{1,1} & \tilde{\tilde{\beta}}_{1,2} & 0 & 0 \\ 0 & 0 & \beta_{2,2} & \beta_{2,3} & 0 \\ 0 & 0 & 0 & \beta_{3,3} & \beta_{3,4} \\ 0 & 0 & 0 & 0 & \beta_{4,4} \end{pmatrix} = \begin{pmatrix} \gamma_{1,0} & \sigma_{1,0} & 0 & 0 & 0 \\ -\sigma_{1,0} & \gamma_{1,0} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tilde{\beta}_{0,0} & \tilde{\beta}_{0,1} & 0 & 0 & 0 \\ \tilde{\beta}_{1,0} & \tilde{\beta}_{1,1} & \beta_{1,2} & 0 & 0 \\ 0 & 0 & \beta_{2,2} & \beta_{2,3} & 0 \\ 0 & 0 & 0 & \beta_{3,3} & \beta_{3,4} \\ 0 & 0 & 0 & 0 & \beta_{4,4} \end{pmatrix}.$$

Continuing, from  $\begin{pmatrix} \tilde{\tilde{\beta}}_{0,1} \\ \tilde{\tilde{\beta}}_{0,2} \end{pmatrix}$  one can compute  $\gamma_{0,2}$  and  $\sigma_{0,2}$  so that  $\begin{pmatrix} \gamma_{0,2} & \sigma_{0,2} \\ -\sigma_{0,2} & \gamma_{0,2} \end{pmatrix} \begin{pmatrix} \tilde{\tilde{\beta}}_{0,1} \\ \tilde{\tilde{\beta}}_{0,2} \end{pmatrix} = \begin{pmatrix} \tilde{\tilde{\tilde{\beta}}}_{0,0} \\ 0 \end{pmatrix}$ . Then

$$\begin{pmatrix} \tilde{\tilde{\tilde{\beta}}}_{0,0} & \tilde{\tilde{\tilde{\beta}}}_{0,1} & 0 & 0 & 0 \\ 0 & \tilde{\tilde{\tilde{\beta}}}_{1,1} & \tilde{\tilde{\tilde{\beta}}}_{1,2} & 0 & 0 \\ 0 & \tilde{\tilde{\beta}}_{2,1} & \tilde{\tilde{\beta}}_{2,2} & \beta_{2,3} & 0 \\ 0 & 0 & 0 & \beta_{3,3} & \beta_{3,4} \\ 0 & 0 & 0 & 0 & \beta_{4,4} \end{pmatrix} = \begin{pmatrix} \tilde{\tilde{\beta}}_{0,0} & \tilde{\tilde{\beta}}_{0,1} & \tilde{\tilde{\beta}}_{0,2} & 0 & 0 \\ 0 & \tilde{\tilde{\beta}}_{1,1} & \tilde{\tilde{\beta}}_{1,2} & 0 & 0 \\ 0 & 0 & \beta_{2,2} & \beta_{2,3} & 0 \\ 0 & 0 & 0 & \beta_{3,3} & \beta_{3,4} \\ 0 & 0 & 0 & 0 & \beta_{4,4} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \gamma_{1,0} & -\sigma_{1,0} & 0 & 0 \\ 0 & \sigma_{1,0} & \gamma_{1,0} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

And so forth, as illustrated in Figure 17.3.



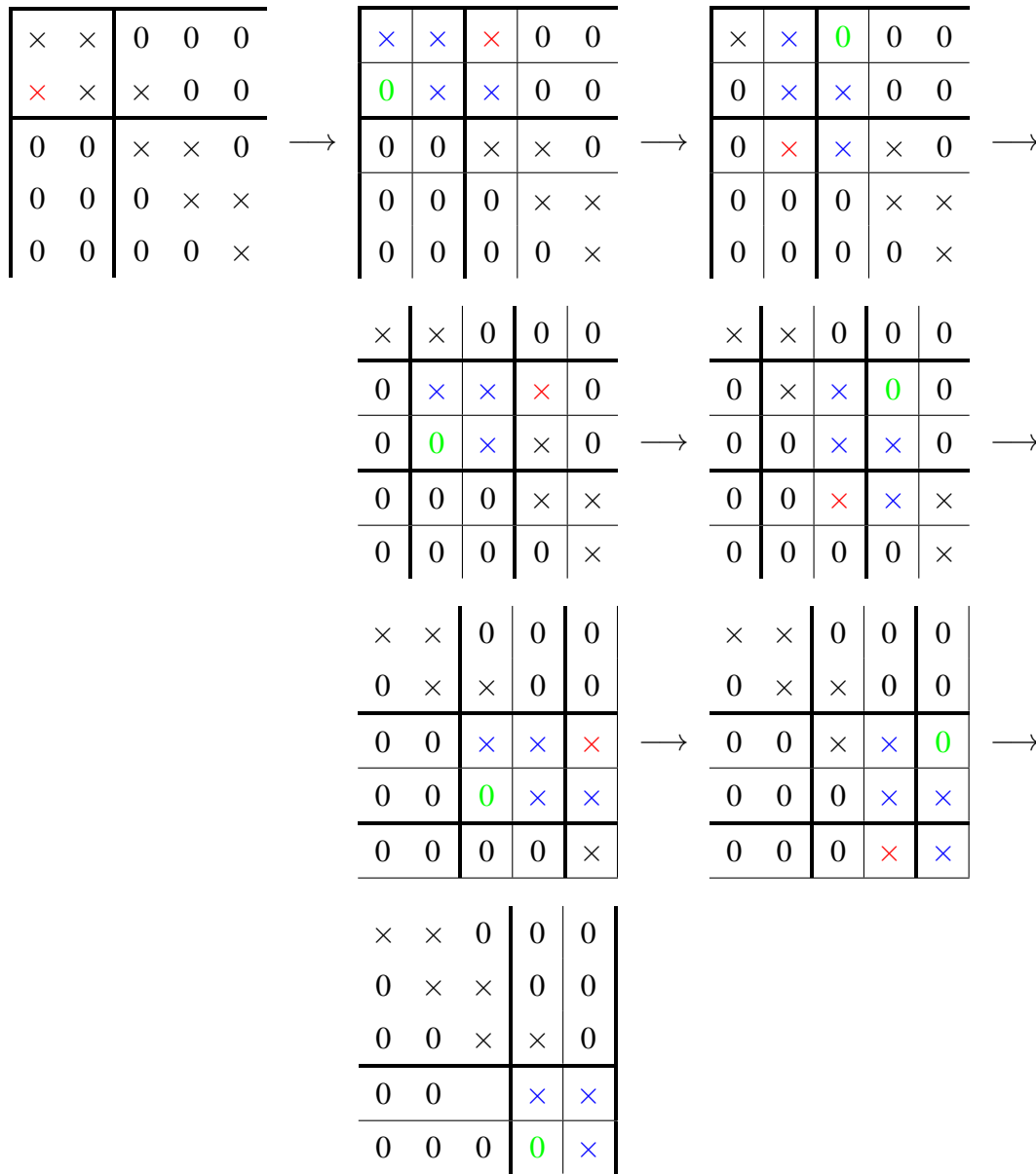


Figure 17.3: Illustration of one sweep of an implicit QR algorithm with a bidiagonal matrix.

## 17.4 Putting it all together

**Algorithm:**  $B := \text{CHASEBULGEID}(B)$

**Partition**  $B \rightarrow$  
$$\left( \begin{array}{c|c|c} B_{TL} & B_{TM} & \star \\ \hline 0 & B_{MM} & B_{MR} \\ \hline 0 & 0 & B_{BR} \end{array} \right)$$

**where**  $B_{TL}$  is  $0 \times 0$  and  $B_{MM}$  is  $2 \times 2$

**while**  $m(B_{BR}) \geq 0$  **do**

**Repartition**

$$\left( \begin{array}{c|c|c} B_{TL} & B_{TM} & \star \\ \hline 0 & B_{MM} & B_{MR} \\ \hline 0 & 0 & B_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c|c|c} B_{00} & \beta_{01}e_l & 0 & 0 & 0 \\ \hline 0 & \beta_{11} & \beta_{12} & 0 & 0 \\ \hline 0 & \beta_{21} & \beta_{22} & \beta_{23} & 0 \\ \hline 0 & 0 & 0 & \beta_{33} & \beta_{34}e_0^T \\ \hline 0 & 0 & 0 & 0 & B_{44} \end{array} \right)$$

**where**  $\tau_{11}$  and  $\tau_{33}$  are scalars  
(during final step,  $\tau_{33}$  is  $0 \times 0$ )

**Compute**  $(\gamma_1^L, \sigma_1^L)$  s.t.  $\left( \begin{array}{c|c} \gamma_1^L & \sigma_1^L \\ \hline -\sigma_1^L & \gamma_1^L \end{array} \right) \left( \begin{array}{c} \beta_{1,1} \\ \beta_{2,1} \end{array} \right) = \left( \begin{array}{c} \tilde{\beta}_{1,1} \\ 0 \end{array} \right)$

**overwriting**  $\beta_{1,1}$  with  $\tilde{\beta}_{1,1}$

$$\left( \begin{array}{c|c} \beta_{1,2} & \beta_{1,3} \\ \hline \beta_{2,2} & \beta_{2,3} \end{array} \right) := \left( \begin{array}{c|c} \gamma_1^L & \sigma_1^L \\ \hline -\sigma_1^L & \gamma_1^L \end{array} \right) \left( \begin{array}{c} \beta_{1,2} \\ \beta_{2,2} \end{array} \right) \left( \begin{array}{c} 0 \\ \beta_{2,3} \end{array} \right)$$

**if**  $m(B_{BR}) \neq 0$

**Compute**  $(\gamma_1^R, \sigma_1^R)$  s.t.  $\left( \begin{array}{c|c} \gamma_1^R & \sigma_1^R \\ \hline -\sigma_1^R & \gamma_1^R \end{array} \right) \left( \begin{array}{c} \beta_{1,2} \\ \beta_{1,3} \end{array} \right) = \left( \begin{array}{c} \tilde{\beta}_{1,2} \\ 0 \end{array} \right)$

**overwriting**  $\beta_{1,2}$  with  $\tilde{\beta}_{1,2}$

$$\left( \begin{array}{c|c} \beta_{1,2} & 0 \\ \hline \beta_{2,2} & \beta_{2,3} \\ \hline \beta_{3,2} & \beta_{3,3} \end{array} \right) := \left( \begin{array}{c|c} \beta_{1,2} & \beta_{1,3} \\ \hline \beta_{2,2} & \beta_{2,3} \\ \hline 0 & \beta_{3,3} \end{array} \right) \left( \begin{array}{c|c} \gamma_1^R & -\sigma_1^R \\ \hline \sigma_1^R & \gamma_1^R \end{array} \right)$$

**Continue with**

$$\left( \begin{array}{c|c|c} B_{TL} & B_{TM} & \star \\ \hline 0 & B_{MM} & B_{MR} \\ \hline 0 & 0 & B_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c|c|c} B_{00} & \beta_{01}e_l & 0 & 0 & 0 \\ \hline 0 & \beta_{11} & \beta_{12} & 0 & 0 \\ \hline 0 & \beta_{21} & \beta_{22} & \beta_{23} & 0 \\ \hline 0 & 0 & 0 & \beta_{33} & \beta_{34}e_0^T \\ \hline 0 & 0 & 0 & 0 & B_{44} \end{array} \right)$$

**endwhile**

Figure 17.4: Chasing the bulge.




# Chapter 18

## Notes on Splitting Methods

### Video

Read disclaimer regarding the videos in the preface!

 YouTube

 [Download from UT Box](#)

 [View After Local Download](#)

(For help on viewing, see Appendix [A](#).)

## Outline

<b>Video</b> . . . . .	<b>293</b>
<b>Outline</b> . . . . .	<b>294</b>
<b>18.1. A Simple Example: One-Dimensional Boundary Value Problem</b> . . . . .	<b>295</b>
<b>18.2. A Two-dimensional Example</b> . . . . .	<b>296</b>
18.2.1. Discretization . . . . .	296
<b>18.3. Direct solution</b> . . . . .	<b>298</b>
<b>18.4. Iterative solution: Jacobi iteration</b> . . . . .	<b>298</b>
18.4.1. Motivating example . . . . .	299
18.4.2. More generally . . . . .	300
<b>18.5. The general case</b> . . . . .	<b>301</b>
<b>18.6. Theory</b> . . . . .	<b>307</b>
18.6.1. Some useful facts . . . . .	307

## 18.1 A Simple Example: One-Dimensional Boundary Value Problem

A computational science application typically starts with a physical problem. This problem is described by a law that governs the physics. Such a law can be expressed as a (continuous) mathematical equation that must be satisfied. Often it is impossible to find an explicit solution for this equation and hence it is approximated by discretizing the problem. At the bottom of the food chain, a linear algebra problem often appears.

Let us illustrate this for a very simple, one-dimensional problem. An example of a one-dimensional Poisson equation with Dirichlet boundary condition on the domain  $[0, 1]$  can be described as

$$u''(x) = f(x) \text{ where } u(0) = u(1) = 0.$$

The Dirichlet boundary condition is just a fancy way of saying that on the boundary (at  $x = 0$  and  $x = 1$ ) function  $u$  is restricted to take on the value 0. Think of this as a string where there is some driving force (e.g., a sound wave hitting the wave) that is given by a continuous function  $f(x)$ , meaning that  $u$  is twice continuously differentiable on the interior (for  $0 < x < 1$ ). The ends of the string are fixed, giving rise to the given boundary condition. The function  $u(x)$  indicates how far from rest (which would happen if  $f(x) = 0$ ) the string is displaced at point  $x$ . There are a few details missing here, such as the fact that  $u$  and  $f$  are probably functions of time as well, but let's ignore that.

Now, we can transform this continuous problem into a discretized problem by partitioning the interval  $[0, 1]$  into  $N + 1$  equal intervals of length  $h$  and looking at the points  $x_0, x_1, \dots, x_{N+1}$  where  $x_i = ih$ , which are the nodes where the intervals meet. We now instead compute  $v_i \approx u(x_i)$ . In our discussion, we will let  $\phi_i = f(x_i)$  (which is exactly available, since  $f(x)$  is given).

Now, we recall that

$$u'(x) \approx \frac{u(x_i + h) - u(x_i)}{h} = \frac{v_{i+1} - v_i}{h}.$$

Also,

$$u''(x) \approx \frac{u'(x_i + h) - u'(x_i)}{h} \approx \frac{\frac{u(x_i + h) - u(x_i)}{h} - \frac{u(x_i) - u(x_{i-1}))}{h}}{h} = \frac{v_{i-1} - 2v_i + v_{i+1}}{h^2}.$$

So, one can approximate our problem with

$$\begin{cases} v_{i-1} - 2v_i + v_{i+1} = h^2 \phi_i & 0 < i \leq N \\ v_0 = v_{N+1} = 0 \end{cases}$$

or, equivalently,

$$\begin{cases} -2v_1 + v_2 = h^2 \phi_1 \\ v_{i-1} - 2v_i + v_{i+1} = h^2 \phi_i & 1 < i < N \\ v_{N-1} - 2v_N = h^2 \phi_N \end{cases}$$

The above compactly expresses the system of linear equations

$$\begin{array}{rcl} -2v_1 + v_2 & & = h^2 \phi_1 \\ v_1 - 2v_2 + v_3 & & = h^2 \phi_2 \\ v_2 - 2v_3 + v_4 & & = h^2 \phi_3 \\ \vdots & \ddots & \vdots \\ v_{N-1} - 2v_N & & = h^2 \phi_N \end{array}$$

or, in matrix notation,

$$\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_N \end{pmatrix} = \begin{pmatrix} h^2 \phi_1 \\ h^2 \phi_2 \\ h^2 \phi_3 \\ \vdots \\ h^2 \phi_N \end{pmatrix}.$$

Thus, one can express this discretized problem as

$$Au = h^2 f,$$

where  $A$  is the indicated tridiagonal matrix.

## 18.2 A Two-dimensional Example

A more typical computational engineering or sciences application starts with a Partial Differential Equation (PDE) that governs the physics of the problem. This is the higher dimensional equivalent of the last example. In this note, we will use one of the simplest, Laplace's equation. Consider

$$-\Delta u = f$$

which in two dimensions is,

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y)$$

with boundary condition  $\partial\Omega = 0$  (meaning that  $u(x, y) = 0$  on the boundary of domain  $\Omega$ ). For example, the domain may be  $0 \leq x, y \leq 1$ ,  $\partial\Omega$  its boundary, and the question may be a membrane with, again,  $f$  being some force from a sound wave.

### 18.2.1 Discretization

To solve the problem computationally with a computer, the problem is again discretized. Relating back to the problem of the membrane on the unit square in the previous section, this means that the continuous domain is viewed as a mesh instead, as illustrated in Figure 18.1. In that figure,  $v_i$  will equal the displacement from rest.

Now, just like before, we will let  $\phi_i$  be the value of  $f(x, y)$  at the mesh point  $i$ . One can approximate

$$\frac{\partial^2 u(x, y)}{\partial x^2} \approx \frac{u(x-h, y) - 2u(x, y) + u(x+h, y)}{h^2}$$

and

$$\frac{\partial^2 u(x, y)}{\partial y^2} \approx \frac{u(x, y-h) - 2u(x, y) + u(x, y+h)}{h^2}$$

so that

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y)$$



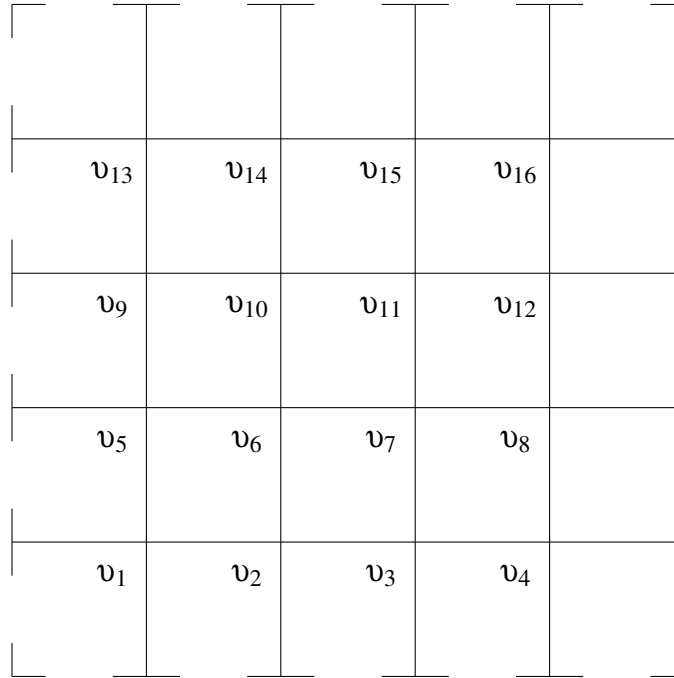


Figure 18.1: Discretized domain, yielding a mesh. The specific ordering of the elements of  $u$ , by rows, is known as the “natural ordering”. If one ordered differently, for example randomly, this would induce a permutation on the rows and columns of matrix  $A$ , since it induces a permutation on the order of the indices.

becomes

$$-\frac{u(x-h, y) + 2u(x, y) - u(x+h, y)}{h^2} - \frac{u(x, y-h) + 2u(x, y) - u(x, y+h)}{h^2} = f(x, y)$$

or, equivalently,

$$\frac{-u(x-h, y) - u(x, y-h) + 4u(x, y) - u(x+h, y) - u(x, y+h)}{h^2} = f(x, y).$$

If  $(x, y)$  corresponds to the point  $i$  in a mesh where the interior points form a  $N \times N$  grid, this translates to the system of linear equations

$$-v_{i-N} - v_{i-1} + 4v_i - v_{i+1} - v_{i+N} = h^2\phi_i.$$

This can be rewritten as

$$v_i = h^2\phi_i - \frac{v_{i-N} + v_{i-1} + v_{i+1} + v_{i+N}}{4}$$

or

$$\begin{array}{ccccccc} 4v_1 & - & v_2 & & & - & v_5 & & & = & h^2\phi_0 \\ -v_1 & + & 4v_2 & - & v_3 & & & - & v_6 & & = & h^2\phi_1 \\ & & - & v_2 & + & 4v_3 & - & v_4 & & & - & v_7 & & = & h^2\phi_2 \\ & & & - & v_3 & + & 4v_4 & & & - & v_8 & & = & h^2\phi_3 \\ -v_1 & & & & + & 4v_5 & - & v_6 & & & - & v_9 & & = & h^2\phi_4 \\ \vdots & & & & & \ddots & \ddots & \ddots & & & \ddots & & = & \vdots \end{array} \quad (18.1)$$

In matrix notation this becomes

$$\left( \begin{array}{ccc|ccc|cc} 4 & -1 & & -1 & & & & \\ -1 & 4 & -1 & & -1 & & & \\ & -1 & 4 & -1 & & -1 & & \\ & & -1 & 4 & & & -1 & \\ \hline -1 & & & 4 & -1 & & -1 & \\ & -1 & & -1 & 4 & -1 & & \ddots \\ & & -1 & & -1 & 4 & -1 & \\ & & & -1 & & -1 & 4 & \\ \hline & & & -1 & & & 4 & \ddots \\ & & & & \ddots & & \ddots & \ddots \end{array} \right) \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \\ \vdots \end{pmatrix} = \begin{pmatrix} h^2\phi_1 \\ h^2\phi_2 \\ h^2\phi_3 \\ h^2\phi_4 \\ h^2\phi_5 \\ h^2\phi_6 \\ h^2\phi_7 \\ h^2\phi_8 \\ h^2\phi_9 \\ \vdots \end{pmatrix}. \quad (18.2)$$

This now shows how solving the discretized Laplace's equation boils down to the solution of a linear system  $Au = h^2f(=b)$ , where  $A$  has a distinct sparsity pattern (pattern of nonzeros).

**Remark 18.1** *The point is that many problems that arise in computational science require the solution to a system of linear equations  $Au = b$  where  $A$  is a (very) sparse matrix.*

**Definition 18.2** *Wilkinson defined a sparse matrix as any matrix with enough zeros that it pays to take advantage of them.*

### 18.3 Direct solution

It is tempting to simply use a dense linear solver to compute the solution to  $Au = b$ . This would require  $O(n^3)$  operations, where  $n$  equals the size of matrix  $A$ . One can take advantage of the fact that beyond the outer band of “ $-1$ ”s the matrix is entire zeroes to reduce the cost to  $O(nB^2)$ , where  $B$  equals the width of the banded matrix. And there are sparse direct methods that try to further reduce fill-in of zeroes to bring the number of operations required to factor to matrix. Details of these go beyond the scope of this note.

### 18.4 Iterative solution: Jacobi iteration

It is more common to solve sparse problems like our model problem via iterative methods which generate a sequence of vectors  $u^{(k)}$ ,  $k = 0, \dots$ . The first vector,  $u^{(0)}$ , represents an initial guess of what the solution is. The hope is that  $u^{(k)}$  converges to the solution  $u$ , in other words, that eventually  $u^{(k)}$  becomes arbitrarily close to the solution. More precisely, we would like for it to be the case that  $\|u^{(k)} - u\| \rightarrow 0$  for some norm  $\|\cdot\|$ , where  $u$  is the true solution.

### 18.4.1 Motivating example

Consider  $Au = b$ . If we guessed perfectly, then  $u^{(0)}$  would solve  $Ax = b$  and it would be the case that

$$\begin{aligned} v_1^{(0)} &= (b_1 + v_2^{(0)} + v_5^{(0)})/4 \\ v_2^{(0)} &= (b_2 + v_1^{(0)} + v_3^{(0)} + v_6^{(0)})/4 \\ v_3^{(0)} &= (b_3 + v_2^{(0)} + v_4^{(0)} + v_7^{(0)})/4 \\ v_4^{(0)} &= (b_4 + v_3^{(0)} + v_8^{(0)})/4 \\ &\vdots \end{aligned}$$

Naturally, that is a bit optimistic. Therefore, a new approximation to the solution is created by computing

$$\begin{aligned} v_1^{(1)} &= (b_1 + v_2^{(0)} + v_5^{(0)})/4 \\ v_2^{(1)} &= (b_2 + v_1^{(0)} + v_3^{(0)} + v_6^{(0)})/4 \\ v_3^{(1)} &= (b_3 + v_2^{(0)} + v_4^{(0)} + v_7^{(0)})/4 \\ v_4^{(1)} &= (b_4 + v_3^{(0)} + v_8^{(0)})/4 \\ &\vdots \end{aligned}$$

In other words, a new guess for the displacement  $u_i$  at point  $i$  is generated by the values at the points around it, plus some contribution from  $\beta_i$  (which itself incorporates contributions from  $\phi_i$  in our example).

This process can be used to compute a sequence of vectors  $u^{(0)}, u^{(1)}, \dots$  by similarly computing  $u^{(k+1)}$  from  $u^{(k)}$ :

$$\begin{aligned} v_1^{(k+1)} &= (b_1 + v_2^{(k)} + v_5^{(k)})/4 \\ v_2^{(k+1)} &= (b_2 + v_1^{(k)} + v_3^{(k)} + v_6^{(k)})/4 \\ v_3^{(k+1)} &= (b_3 + v_2^{(k)} + v_4^{(k)} + v_7^{(k)})/4 \\ v_4^{(k+1)} &= (b_4 + v_3^{(k)} + v_8^{(k)})/4 \\ &\vdots \end{aligned}$$

This can be rewritten as

$$\begin{aligned} 4v_1^{(k+1)} &= v_2^{(k)} + v_5^{(k)} + \beta_1 \\ 4v_2^{(k+1)} &= v_1^{(k)} + v_3^{(k)} + v_6^{(k)} + \beta_2 \\ 4v_3^{(k+1)} &= v_2^{(k)} + v_4^{(k)} + v_7^{(k)} + \beta_3 \\ 4v_4^{(k+1)} &= v_3^{(k)} + v_8^{(k)} + \beta_4 \\ 4v_5^{(k+1)} &= v_1^{(k)} + v_6^{(k)} + v_9^{(k)} + \beta_5 \\ &= \ddots \quad \ddots \quad \ddots \quad \ddots \quad \vdots \end{aligned}$$

or, in matrix notation,

$$4 \begin{pmatrix} v_1^{(k+1)} \\ v_2^{(k+1)} \\ v_3^{(k+1)} \\ v_4^{(k+1)} \\ \hline v_5^{(k+1)} \\ v_6^{(k+1)} \\ v_7^{(k+1)} \\ v_8^{(k+1)} \\ \hline v_9^{(k+1)} \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 & 1 & & & 1 & & & & \\ & 1 & 0 & 1 & & 1 & & & \\ & & 1 & 0 & 1 & & 1 & & \\ & & & 1 & 0 & & & 1 & \\ \hline 1 & & & & 0 & 1 & & & 1 \\ & 1 & & & 1 & 0 & 1 & & \ddots \\ & & 1 & & & 1 & 0 & 1 & \\ & & & 1 & & & 1 & 0 & \\ \hline & & & & 1 & & & 0 & \ddots \\ & & & & & \ddots & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} v_1^{(k)} \\ v_2^{(k)} \\ v_3^{(k)} \\ v_4^{(k)} \\ \hline v_5^{(k)} \\ v_6^{(k)} \\ v_7^{(k)} \\ v_8^{(k)} \\ \hline v_9^{(k)} \\ \vdots \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \hline \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \hline \beta_9 \\ \vdots \end{pmatrix}.$$

Typically, the physics of the problem dictates that this iteration will eventually yield a vector  $u^{(k)}$  that is arbitrarily close to the solution of  $Au = b$ . More on the mathematical reasons for this later. In other words, the vectors  $u^{(k)}$  will *converge* to  $u$ :  $\lim_{k \rightarrow \infty} u^{(k)} = u$  or  $\lim_{k \rightarrow \infty} \|u^{(k)} - u\| = 0$ .

### 18.4.2 More generally

The above discussion can be more concisely described as follows. Split matrix  $A = L + D + U$ , where  $L$  and  $U$  are the strictly lower and upper triangular parts of  $A$ , respectively, and  $D$  its diagonal. Now,

$$Au = b$$

implies that

$$(L + D + U)u = b$$

or

$$u = -D^{-1}[(L + U)u + b].$$

This is an example of a fixed-point equation: Plug  $u$  into  $-D^{-1}[(L + U)u + b]$  and the result is again  $u$ . The iteration is then created by viewing the vector on the left as the next guess given the guess for  $u$  on the right:

$$u^{(k+1)} = -D^{-1}[(L + U)u^{(k)} + b].$$

Now let us see how to extract a more detailed algorithm from this. Partition, conformally,

$$A = \begin{pmatrix} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{pmatrix}, D = \begin{pmatrix} D_{00} & 0 & 0 \\ \hline 0 & \delta_{11} & 0 \\ \hline 0 & 0 & D_{22} \end{pmatrix}, L = \begin{pmatrix} L_{00} & 0 & 0 \\ \hline l_{10}^T & 0 & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{pmatrix}, \quad (18.3)$$

$$U = \begin{pmatrix} U_{00} & u_{01} & U_{02} \\ \hline 0 & 0 & u_{12}^T \\ \hline 0 & 0 & U_{22} \end{pmatrix}, u^{(k)} = \begin{pmatrix} v_0^{(k)} \\ \hline v_1^{(k)} \\ \hline v_2^{(k)} \end{pmatrix}, \text{ and } b^{(k)} = \begin{pmatrix} b_0 \\ \hline \beta_1 \\ \hline b_2 \end{pmatrix}. \quad (18.4)$$

Consider

$$\left( \begin{array}{c|c|c} D_{00} & 0 & 0 \\ \hline 0 & \delta_{11} & 0 \\ \hline 0 & 0 & D_{22} \end{array} \right) \begin{pmatrix} x_0^{(k+1)} \\ v_1^{(k+1)} \\ x_2^{(k+1)} \end{pmatrix} = - \left( \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 0 & 0 \\ \hline L_{20} & l_{21} & L_{22} \end{array} \right) + \left( \begin{array}{c|c|c} U_{00} & v_{01} & U_{02} \\ \hline 0 & 0 & v_{12} \\ \hline 0 & 0 & U_{22} \end{array} \right) \right) \begin{pmatrix} v_0^{(k)} \\ v_1^{(k)} \\ v_2^{(k)} \end{pmatrix} + \begin{pmatrix} b_0 \\ \beta_1 \\ b_2 \end{pmatrix}.$$

Then

$$\delta_{11} v_1^{(k+1)} = \beta_1 - l_{10}^T v_0^{(k)} - v_{12}^T v_2^{(k)}.$$

Now, for the Jacobi iteration  $\delta_{11} = \alpha_{11}$ ,  $l_{10}^T = a_{10}^T$ , and  $v_{12} = a_{12}^T$ , so that

$$v_1^{(k+1)} = (\beta_1 - a_{10}^T v_0^{(k)} - a_{12}^T v_2^{(k)}) / \alpha_{11}.$$

This suggests the algorithm in Figure 18.2. For those who prefer indices (and for these kinds of iterations, indices often clarify rather than obscure), this can be described as

$$v_i^{(k+1)} = \frac{1}{\alpha_{ii}} \left( \beta_i - \sum_{\substack{j \neq i \\ \alpha_{ij} \neq 0}} \alpha_{ij} v_j^{(k)} \right).$$

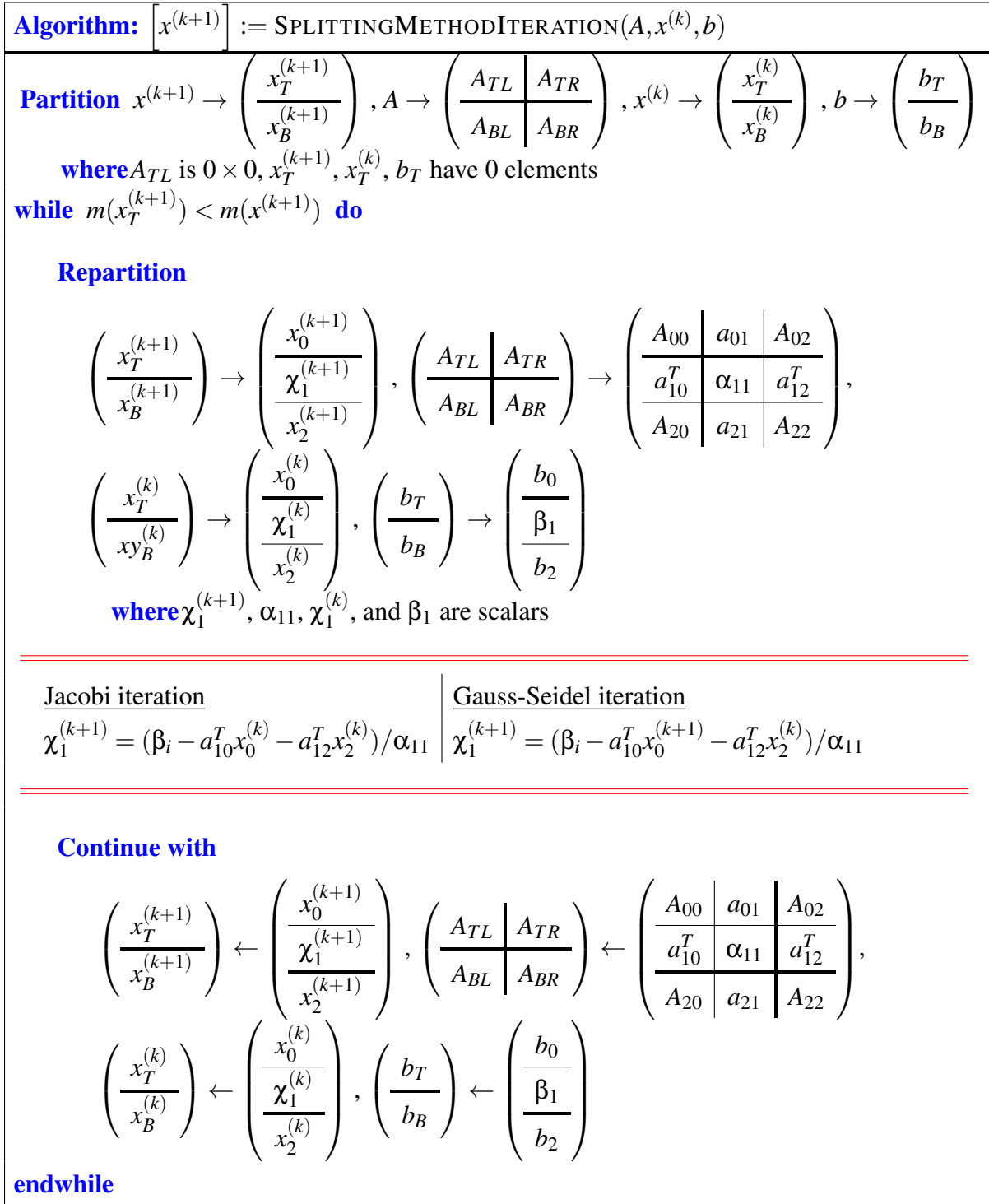
**Remark 18.3** *It is not hard to see that, for the example of the discretized square, this yields exactly the computations that compute  $u^{(k+1)}$  from  $u^{(k)}$  since there*

$$-(L+U) = \left( \begin{array}{ccc|ccc|ccc} 0 & 1 & & 1 & & & & & \\ & 1 & 0 & 1 & & 1 & & & \\ & & 1 & 0 & 1 & & 1 & & \\ & & & 1 & 0 & & & 1 & \\ \hline 1 & & & & & 0 & 1 & & 1 \\ & 1 & & & & 1 & 0 & 1 & \ddots \\ & & 1 & & & & 1 & 0 & 1 \\ & & & 1 & & & & 1 & 0 \\ \hline & & & & & 1 & & & 0 & \ddots \\ & & & & & & \ddots & & \ddots & \ddots \end{array} \right)$$

and  $D = 4I$ .

## 18.5 The general case

The Jacobi iteration is a special case of a family of method known as Splitting Methods. It starts with a *splitting* of a nonsingular matrix  $A$  into  $A = M - N$ . One next observes that  $Au = b$  is equivalent to  $(M - N)u = b$ , which is equivalent to  $Mu = Nu + b$  or  $u = M^{-1}(Nu + b)$ . Notice that  $M^{-1}$  is not explicitly

Figure 18.2: Various splitting methods (one iteration  $x^{(k+1)} = M^{-1}(Nx^{(k)} + b)$ ).

formed: one solves with  $M$  instead. Now, a vector  $u$  is the solution of  $Au = b$  if and only if it satisfies  $u = M^{-1}(Nu + b)$ . The idea is to start with an initial guess (approximation) of  $u$ ,  $u^{(0)}$ , and to then iterate to compute  $u^{(k+1)} = M^{-1}(Nu^{(k)} + b)$ , which requires a multiplication with  $N$  and a solve with  $M$  in each iteration.

**Example 18.4** Let  $A \in \mathbb{C}^{n \times n}$  and split this matrix  $A = L + D + U$ , where  $L$  and  $U$  are the strictly lower and upper triangular parts of  $A$ , respectively, and  $D$  its diagonal. Then choosing  $M = D$  and  $N = -(L + U)$  yields the Jacobi iteration.

The convergence of these methods is summarized by the following theorem:

**Theorem 18.5** Let  $A \in \mathbb{C}^{n \times n}$  be nonsingular and  $x, b \in \mathbb{C}^n$  so that  $Au = b$ . Let  $A = M - N$  be a splitting of  $A$ ,  $u^{(0)}$  be given (an initial guess), and  $u^{(k+1)} = M^{-1}(Nu^{(k)} + b)$ . If  $\|M^{-1}N\| < 1$  for some matrix norm induced by the  $\|\cdot\|$  vector norm, then  $u^{(k)}$  will converge to the solution  $u$ .

**Proof:** Notice  $Au = b$  implies that  $(M - N)u = b$  implies that  $Mu = Nu + b$  and hence  $u = M^{-1}(Nu + b)$ . Hence

$$u^{(k+1)} - u = M^{-1}(Nu^{(k)} + b) - M^{-1}(Nu + b) = M^{-1}N(u^{(k)} - u).$$

Taking norms of both sides, one gets that

$$\|u^{(k+1)} - u\| = \|M^{-1}N(u^{(k)} - u)\| \leq \|M^{-1}N\| \|u^{(k)} - u\|.$$

Since this holds for all  $k$ , once can deduce that

$$\|u^{(k)} - u\| \leq \|M^{-1}N\|^k \|u^{(0)} - u\|.$$

If  $\|M^{-1}N\| < 1$  then the right-hand side of this will converge to zero, meaning that  $u^{(k)}$  converges to  $u$ .

Now, often one checks convergence by computing the residual  $r^{(k)} = b - Au^{(k)}$ . After all, if  $r^{(k)}$  is approximately the zero vector, then  $u^{(k)}$  approximately solves  $Au = b$ . The following corollary links the convergence of  $r^{(k)}$  to the convergence of  $u^{(k)}$ .

**Corollary 18.6** Let  $A \in \mathbb{C}^{n \times n}$  be nonsingular and  $u, b \in \mathbb{C}^n$  so that  $Au = b$ . Let  $u^{(k)} \in \mathbb{C}^n$  and let  $r^{(k)} = b - Au^{(k)}$ . Then  $u^{(k)}$  converges to  $u$  if and only if  $r^{(k)}$  converges to the zero vector.

**Proof:** Assume that  $u^{(k)}$  converges to  $u$ . Note that  $r^{(k)} = b - Au^{(k)} = Au - Au^{(k)} = A(u - u^{(k)})$ . Since  $A$  is nonsingular, clearly  $r^{(k)}$  converges to the zero vector if and only if  $u^{(k)}$  converges to  $u$ .

**Corollary 18.7** Let  $A \in \mathbb{C}^{n \times n}$  be nonsingular and  $u, b \in \mathbb{C}^n$  so that  $Au = b$ . Let  $A = M - N$  be a splitting of  $A$ ,  $u^{(0)}$  be given (an initial guess),  $u^{(k+1)} = M^{-1}(Nu^{(k)} + b)$ , and  $r^{(k)} = b - Au^{(k)}$ . If  $\|M^{-1}N\| < 1$  for some matrix norm induced by the  $\|\cdot\|$  vector norm, then  $r^{(k)}$  will converge to the zero vector.

Now let us see how to extract a more detailed algorithm from this. Partition, conformally,

$$A = \left( \begin{array}{c|c|c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{20} & a_{21} & A_{22} \end{array} \right), \quad M = \left( \begin{array}{c|c|c} M_{00} & m_{01} & M_{02} \\ \hline m_{10}^T & \mu_{11} & m_{12}^T \\ \hline M_{20} & m_{21} & M_{22} \end{array} \right), \quad N = \left( \begin{array}{c|c|c} N_{00} & n_{01} & N_{02} \\ \hline n_{10}^T & \nu_{11} & n_{12}^T \\ \hline N_{20} & n_{21} & N_{22} \end{array} \right),$$

$$u^{(k)} = \begin{pmatrix} x_0^{(k)} \\ \frac{v_1^{(k)}}{x_2^{(k)}} \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} b_0 \\ \frac{\beta_1}{b_2} \end{pmatrix}.$$

Consider

$$\left( \begin{array}{c|c|c} M_{00} & m_{01} & M_{02} \\ \hline m_{10}^T & \mu_{11} & m_{12}^T \\ \hline M_{20} & m_{21} & M_{22} \end{array} \right) \left( \begin{array}{c} x_0^{(k+1)} \\ \hline v_1^{(k+1)} \\ \hline x_2^{(k+1)} \end{array} \right) = \left( \begin{array}{c|c|c} N_{00} & n_{01} & N_{02} \\ \hline n_{10}^T & v_{11} & n_{12}^T \\ \hline N_{20} & n_{21} & N_{22} \end{array} \right) \left( \begin{array}{c} x_0^{(k)} \\ \hline v_1^{(k)} \\ \hline x_2^{(k)} \end{array} \right) + \left( \begin{array}{c} b_0 \\ \hline \beta_1 \\ \hline b_2 \end{array} \right).$$

Then

$$m_{10}^T x_0^{(k+1)} + \mu_{11} v_1^{(k+1)} + m_{12}^T x_2^{(k+1)} = n_{10}^T x_0^{(k)} + v_{11} v_1^{(k)} + n_{12}^T x_2^{(k)} + \beta_1.$$

### Jacobi iteration

**Example 18.8** For the Jacobi iteration  $m_{10}^T = 0$ ,  $\mu_{11} = \alpha_{11}$ ,  $m_{12}^T = 0$ ,  $n_{10}^T = -a_{10}^T$ ,  $v_{11} = 0$ , and  $n_{12}^T = -a_{12}^T$ , so that

$$v_1^{(k+1)} = (\beta_1 - a_{10}^T x_0^{(k)} - a_{12}^T x_2^{(k)}) / \alpha_{11}.$$

This suggests the algorithm in Figure 18.2.

**Gauss-Seidel iteration** A modification of the Jacobi iteration is inspired by the observation that in Eqns. (18.1)–(18.2) when computing  $v_i$  one could use the new (updated) values for  $v_j$ ,  $j < i$ :

$$\begin{aligned} v_1^{(k+1)} &= (\beta_1 + v_2^{(k)} + v_5^{(k)}) / 4 \\ v_2^{(k+1)} &= (\beta_2 + v_1^{(k+1)} + v_3^{(k)} + v_6^{(k)}) / 4 \\ v_3^{(k+1)} &= (\beta_3 + v_2^{(k+1)} + v_4^{(k)} + v_7^{(k)}) / 4 \\ v_4^{(k+1)} &= (\beta_4 + v_3^{(k+1)} + v_8^{(k)}) / 4 \\ v_5^{(k+1)} &= (\beta_5 + v_1^{(k+1)} + v_6^{(k)} + v_9^{(k)}) / 4 \\ &\vdots \end{aligned}$$

This can be rewritten as

$$\begin{aligned} 4v_1^{(k+1)} &= v_2^{(k)} + v_5^{(k)} + \beta_1 \\ 4v_2^{(k+1)} &= v_1^{(k+1)} + v_3^{(k)} + v_6^{(k)} + \beta_2 \\ 4v_3^{(k+1)} &= v_2^{(k+1)} + v_4^{(k)} + v_7^{(k)} + \beta_3 \\ 4v_4^{(k+1)} &= v_3^{(k+1)} + v_8^{(k)} + \beta_4 \\ 4v_5^{(k+1)} &= v_1^{(k+1)} + v_6^{(k)} + v_9^{(k)} + \beta_5 \\ &= \ddots \quad \ddots \quad \ddots \quad \ddots \quad \ddots \end{aligned}$$



or, in matrix notation,

$$\begin{pmatrix}
 4 & & & & & & & & \\
 -1 & 4 & & & & & & & \\
 & -1 & 4 & & & & & & \\
 & & -1 & 4 & & & & & \\
 -1 & & & & 4 & & & & \\
 & -1 & & & -1 & 4 & & & \\
 & & -1 & & & -1 & 4 & & \\
 & & & -1 & & & -1 & 4 & \\
 & & & & -1 & & & & 4
 \end{pmatrix}
 \begin{pmatrix}
 v_1^{(k+1)} \\
 v_2^{(k+1)} \\
 v_3^{(k+1)} \\
 v_4^{(k+1)} \\
 v_5^{(k+1)} \\
 v_6^{(k+1)} \\
 v_7^{(k+1)} \\
 v_8^{(k+1)} \\
 v_9^{(k+1)}
 \end{pmatrix}
 =
 \begin{pmatrix}
 0 & 1 & & & & & & & \\
 0 & 0 & 1 & & & & & & \\
 & 0 & 0 & 1 & & & & & \\
 & & 0 & 0 & 1 & & & & \\
 0 & & & & & 1 & & & \\
 0 & & & & 0 & 1 & & & 1 \\
 & 0 & & & 0 & 0 & 1 & & \ddots \\
 & & 0 & & & 0 & 0 & 1 & \\
 & & & 0 & & & 0 & 0 & \\
 0 & & & & & & & & 0 \\
 & & & & & & & & \ddots
 \end{pmatrix}
 \begin{pmatrix}
 v_1^{(k)} \\
 v_2^{(k)} \\
 v_3^{(k)} \\
 v_4^{(k)} \\
 v_5^{(k)} \\
 v_6^{(k)} \\
 v_7^{(k)} \\
 v_8^{(k)} \\
 v_9^{(k)}
 \end{pmatrix}
 + b.$$

Again, split  $A = L + D + U$  and partition these matrices as in (18.3) and (18.4). Consider

$$\left( \begin{array}{c|c|c} D_{00} + L_{00} & 0 & 0 \\ \hline l_{10}^T & \delta_{11} & 0 \\ \hline L_{20} & l_{21} & D_{22} + L_{22} \end{array} \right) \begin{pmatrix} x_0^{(k+1)} \\ v_1^{(k+1)} \\ x_2^{(k+1)} \end{pmatrix} = - \left( \begin{array}{c|c|c} U_{00} & v_{01} & U_{02} \\ \hline 0 & 0 & v_{12} \\ \hline 0 & 0 & U_{22} \end{array} \right) \begin{pmatrix} x_0^{(k)} \\ v_1^{(k)} \\ x_2^{(k)} \end{pmatrix} + \begin{pmatrix} b_0 \\ \beta_1 \\ b_2 \end{pmatrix}.$$

Then

$$l_{10}^T x_0^{(k+1)} + \delta_{11} v_1^{(k+1)} = \beta_1 - v_{12}^T x_2^{(k)}.$$

Now, for the Gauss-Seidel iteration  $\delta_{11} = \alpha_{11}$ ,  $l_{10}^T = a_{10}^T$ , and  $v_{12} = a_{12}^T$ , so that

$$v_1^{(k+1)} = (\beta_1 - a_{10}^T x_0^{(k+1)} - a_{12}^T x_2^{(k)}) / \alpha_{11}.$$

This suggests the algorithm in Figure 18.2. For those who prefer indices, this can be described as

$$v_i^{(k+1)} = \frac{1}{\alpha_{ii}} \left( \beta_i - \sum_{\substack{j=0 \\ \alpha_{ij} \neq 0}}^{i-1} \alpha_{ij} v_j^{(k+1)} - \sum_{\substack{j=i+1 \\ \alpha_{ij} \neq 0}}^{n-1} \alpha_{ij} v_j^{(k)} \right).$$

Notice that this fits the general framework since now  $M = L + D$  and  $N = -U$ .

**Successive Over Relaxation (SOR)** Consider the update as performed by the Gauss-Seidel iteration

$$v_1^{\text{GS}(k+1)} = (\beta_1 - a_{10}^T x_0^{(k+1)} - a_{12}^T x_2^{(k)}) / \alpha_{11}.$$

Next, one can think of

$$v_1^{\text{GS}(k+1)} - v_1^{(k)}$$

as a direction in which the solution is changing. Now, one can ask the question “What if we go a little further in that direction?” In other words, One can think of this as

$$v_1^{(k+1)} = v_1^{(k)} + \omega [v_1^{\text{GS}(k+1)} - v_1^{(k)}] = \omega v_1^{\text{GS}(k+1)} + (1 - \omega) v_1^{(k)},$$

for some *relaxation* parameter  $\omega = 1 + \alpha$ . Going further would mean picking  $\omega > 1$ . Then

$$\begin{aligned} v_1^{(k+1)} &= \omega [\beta_1 - a_{10}^T x_0^{(k+1)} - a_{12}^T x_2^{(k)}] / \alpha_{11} + (1 - \omega) v_1^{(k)} \\ &= (\omega \beta_1 - \omega a_{10}^T x_0^{(k+1)} - \omega a_{12}^T x_2^{(k)}) / \alpha_{11} + (1 - \omega) v_1^{(k)}, \end{aligned}$$

which can be rewritten as

$$\omega a_{10}^T x_0^{(k+1)} + \alpha_{11} v_1^{(k+1)} = (1 - \omega) \alpha_{11} v_1^{(k)} - \omega a_{12}^T x_2^{(k)} + \omega \beta_1.$$

or

$$a_{10}^T x_0^{(k+1)} + \frac{\alpha_{11}}{\omega} v_1^{(k+1)} = \frac{1 - \omega}{\omega} \alpha_{11} v_1^{(k)} - a_{12}^T x_2^{(k)} + \beta_1.$$

Now, if one once again partitions  $A = L + D + U$  and one takes  $M = (L + \frac{1}{\omega} D)$  and  $N = (\frac{1 - \omega}{\omega} D - U)$  then it can be easily checked that  $M u^{(k+1)} = N u^{(k)} + b$ . Thus,

$$\|u^{(k)} - u\| \leq \|(L + \frac{1}{\omega} D)^{-1} (\frac{1 - \omega}{\omega} D - U)\|^k \|u^{(0)} - u\| = \|(\omega L + D)^{-1} ((1 - \omega) D - \omega U)\|^k \|u^{(0)} - u\|.$$

The idea now is that by choosing  $\omega$  carefully,

$$\|(\omega L + D)^{-1} ((1 - \omega) D - \omega U)\|$$

can be made smaller, meaning that the convergence is faster.

**Symmetric Successive Over Relaxation** Gauss-Seidel and SOR update the unknowns in what is called the *natural ordering* meaning that one updates  $u_0, u_1, \dots$  in that prescribed order, using the most recently updated values and, in the case of SOR, extrapolating the update. The iteration can be made symmetric by updating first in the natural ordering and then updating similarly, but in reverse natural ordering.

Then this is equivalent to choosing  $M_F = (L + \frac{1}{\omega}D)$  and  $N_F = (\frac{1-\omega}{\omega}D - U)$  (the Forward splitting) and  $M_B = (U + \frac{1}{\omega}D)$  and  $N_B = (\frac{1-\omega}{\omega}D - L)$  (the Backward splitting). Then

$$\begin{aligned} M_F u^{(k+1/2)} &= N_F u^{(k)} + b \\ M_B u^{(k+1)} &= N_B u^{(k+1/2)} + b, \end{aligned}$$

or,

$$\begin{aligned} u^{(k+1)} &= M_B^{-1} [N_B u^{(k+1/2)} + b] \\ &= M_B^{-1} [N_B M_F^{-1} [N_F u^{(k)} + b] + b] \\ &= \underbrace{\left(U + \frac{1}{\omega}D\right)^{-1}}_{M_B^{-1}} \left[ \underbrace{\left(\frac{1-\omega}{\omega}D - L\right)}_{N_B} \underbrace{\left(L + \frac{1}{\omega}D\right)^{-1}}_{M_F^{-1}} \left[ \underbrace{\left(\frac{1-\omega}{\omega}D - U\right)}_{N_F} u^{(k)} + b \right] + b \right] \\ &= \left(U + \frac{1}{\omega}D\right)^{-1} \left[ \left(\frac{2-\omega}{\omega}D - \left(L + \frac{1}{\omega}D\right)\right) \left(L + \frac{1}{\omega}D\right)^{-1} \left[ \left(\frac{1-\omega}{\omega}D - U\right) u^{(k)} + b \right] + b \right] \\ &= \left(U + \frac{1}{\omega}D\right)^{-1} \left[ \left(\frac{2-\omega}{\omega}D \left(L + \frac{1}{\omega}D\right)^{-1} - I\right) \left[ \left(\frac{1-\omega}{\omega}D - U\right) u^{(k)} + b \right] + b \right] \\ &= (\omega U + D)^{-1} \left[ ((1-\omega)D - \omega L) (\omega L + D)^{-1} \left[ ((1-\omega)D - \omega U) u^{(k)} + b \right] + b \right] \\ &= (\omega U + D)^{-1} \left[ ((1-\omega)D - \omega L) (\omega L + D)^{-1} \left[ ((1-\omega)D - \omega U) u^{(k)} + \omega b \right] + \omega b \right] \end{aligned}$$

(One of these days I'll work out all the details!)

## 18.6 Theory

### 18.6.1 Some useful facts

Recall that for  $A \in \mathbb{C}^{n \times n}$  its spectral radius is denoted by  $\rho(A)$ . It equals to the magnitude of the eigenvalue of  $A$  that is largest in magnitude.

**Theorem 18.9** Let  $\|\cdot\|$  be matrix norm induced by a vector norm  $\|\cdot\|$ . Then for any  $A \in \mathbb{C}^{m \times n}$   $\rho(A) \leq \|A\|$ .

**Homework 18.10** Prove the above theorem.

 [SEE ANSWER](#)

**Theorem 18.11** Given a matrix  $A \in \mathbb{C}^{n \times n}$  and  $\varepsilon > 0$ , there exists a consistent matrix norm  $\|\cdot\|$  such that  $\|A\| \leq \rho(A) + \varepsilon$ .

**Proof:** Let

$$A = UTU^H = \left( u_0 \mid U_1 \right) \left( \begin{array}{c|c} \rho(A) & t_{01}^T \\ \hline 0 & T_{11} \end{array} \right) \left( u_0 \mid U_1 \right)^H$$

be a Schur decomposition of  $A$ . Define  $\|\cdot\|$  by

$$\|X\| = \left\| U^T X U \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \varepsilon/\|A\|_2 I \end{array} \right) \right\|_2.$$

(You can show that this is a matrix norm.) Then

$$\begin{aligned} \|A\| &= \left\| U^T A U \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \varepsilon/\|A\|_2 I \end{array} \right) \right\|_2 = \left\| \left( \begin{array}{c|c} \rho(A) & t_{01}^T \\ \hline 0 & T_{11} \end{array} \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \varepsilon/\|A\|_2 I \end{array} \right) \right\|_2 \\ &= \left\| \left[ \left( \begin{array}{c|c} \rho(A) & 0 \\ \hline 0 & 0 \end{array} \right) + \left( \begin{array}{c|c} 0 & t_{01}^T \\ \hline 0 & T_{11} \end{array} \right) \right] \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \varepsilon/\|A\|_2 I \end{array} \right) \right\|_2 \\ &= \left\| \left( \begin{array}{c|c} \rho(A) & 0 \\ \hline 0 & 0 \end{array} \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \varepsilon/\|A\|_2 I \end{array} \right) + \left( \begin{array}{c|c} 0 & t_{01}^T \\ \hline 0 & T_{11} \end{array} \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & \varepsilon/\|A\|_2 I \end{array} \right) \right\|_2 \\ &\leq \left\| \left( \begin{array}{c|c} \rho(A) & 0 \\ \hline 0 & 0 \end{array} \right) \right\|_2 + \left\| \left( \begin{array}{c|c} 0 & t_{01}^T \\ \hline 0 & T_{11} \end{array} \right) \left( \begin{array}{c|c} 0 & 0 \\ \hline 0 & \varepsilon/\|A\|_2 I \end{array} \right) \right\|_2 \\ &\leq \left\| \left( \begin{array}{c|c} \rho(A) & 0 \\ \hline 0 & 0 \end{array} \right) \right\|_2 + \varepsilon \left\| \left( \begin{array}{c|c} 0 & t_{01}^T \\ \hline 0 & T_{11} \end{array} \right) \right\|_2 \left\| \left( \begin{array}{c|c} 0 & 0 \\ \hline 0 & 1/\|A\|_2 I \end{array} \right) \right\|_2 \\ &\leq \left\| \left( \begin{array}{c|c} \rho(A) & 0 \\ \hline 0 & 0 \end{array} \right) \right\|_2 + \varepsilon \left\| \left( \begin{array}{c|c} \rho(A) & t_{01}^T \\ \hline 0 & T_{11} \end{array} \right) \right\|_2 \left\| \left( \begin{array}{c|c} 0 & 0 \\ \hline 0 & 1/\|A\|_2 I \end{array} \right) \right\|_2 \\ &\leq \rho(A) + \varepsilon \|A\|_2 / \|A\|_2 = \rho(A) + \varepsilon. \end{aligned}$$

**Theorem 18.12** The iteration

$$x^{(k+1)} = M^{-1}(Nk^{(k)} + b)$$

converges for any  $x^{(0)}$  if and only if

$$\rho(M^{-1}N) < 1.$$

**Proof:** Recall that for any consistent matrix norm  $\|\cdot\|$

$$\|x^{(k)} - x\| \leq \|M^{-1}N\|^k \|x^{(0)} - x\|$$

.

( $\Leftarrow$ ) Assume  $\rho(M^{-1}N) < 1$ . Pick  $\varepsilon$  such that  $\rho(M^{-1}N) + \varepsilon < 1$ . Finally, pick  $\|\cdot\|$  such that  $\|M^{-1}N\| \leq \rho(M^{-1}N) + \varepsilon$ . Then  $\|M^{-1}N\| < 1$  and hence the sequence converges.

( $\Rightarrow$ ) We will show that if  $\rho(M^{-1}N) \geq 1$  then there exists a  $x^{(0)}$  such that the iteration does not converge.

Let  $y \neq 0$  have the property that  $M^{-1}Ny = \rho(M^{-1}N)y$ . Choose  $x^{(0)} = y + x$  where  $Ax = b$ . Then

$$\begin{aligned}\|x^{(1)} - x\| &= \|M^{-1}N(x^{(0)} - x)\| = \|M^{-1}Ny\| = \|\rho(M^{-1}N)y\| \\ &= \rho(M^{-1}N)\|y\| = \rho(M^{-1}N)\|x^{(0)} - x\| \geq \|x^{(0)} - x\|.\end{aligned}$$

Extending this, one finds that for all  $k$ ,  $\|x^{(k)} - x\| = \|M^{-1}N(x^{(0)} - x)\|$ . Thus, the sequence does not converge.

**Definition 18.13** A matrix  $A \in \mathbb{C}^{n \times n}$  is said to be diagonally dominant if for all  $i$ ,  $0 \leq i < n$

$$|\alpha_{i,i}| \geq \sum_{j \neq i} |\alpha_{i,j}|.$$

It is said to be strictly diagonally dominant if for all  $i$ ,  $0 \leq i < n$

$$|\alpha_{i,i}| > \sum_{j \neq i} |\alpha_{i,j}|.$$

**Theorem 18.14 Gershgorin Disc Theorem** Let  $A \in \mathbb{C}^{n \times n}$ . Define

$$\mathcal{D}_i = \left\{ \beta : |\beta - \alpha_{ii}| \leq \sum_{j \neq i} |\alpha_{ij}| \right\}$$

Then  $\lambda \in \Lambda(A)$  implies there exists a  $k$  such that  $\lambda \in \mathcal{D}_k$ .

**Proof:** Let  $\lambda \in \Lambda(A)$  and  $u \neq 0$  be such that  $Au = \lambda u$ . Then  $(\lambda I - A)u = 0$ . Let  $k$  be such that  $|v_k| = \max_{j=0}^{n-1} |v_j|$ . In other words,  $v_k$  is the element in  $u$  of largest magnitude. Then, looking at the  $k$ th row of  $(\lambda I - A)u$  we find that

$$(\lambda - \alpha_{kk})v_k - \sum_{j \neq k} \alpha_{k,j}v_j = 0.$$

Hence

$$|\lambda - \alpha_{kk}| = \left| \sum_{j \neq k} \alpha_{k,j} \frac{v_j}{v_k} \right| \leq \sum_{j \neq k} |\alpha_{k,j}| \left| \frac{v_j}{v_k} \right| \leq \sum_{j \neq k} |\alpha_{k,j}|$$

which implies that  $\lambda \in \mathcal{D}_k$ .

**Example 18.15** In the case of the example in the previous section, where  $A$  is given by (18.2) and the splitting yields the Jacobi iteration, the Gershgorin Disc Theorem implies that  $\|M^{-1}N\|_2 \leq 1$ . We would like to prove that  $\|M^{-1}N\|_2 < 1$ , which is true, but a little trickier to prove...

**Definition 18.16** A matrix  $A \in \mathbb{C}^{n \times n}$  is said to be reducible if there exists a permutation matrix  $P$  such that

$$PAP^T = \left( \begin{array}{c|c} T_{TL} & T_{TR} \\ \hline 0 & T_{BR} \end{array} \right),$$

where  $T_{TL}$  (and hence  $T_{BR}$ ) is square of size  $b \times b$  where  $0 < b < n$ . A matrix that is not reducible is irreducible.

In other words, a matrix is reducible if and only if there exists a symmetric permutation of the rows and columns that leaves the matrix block upper triangular.

**Homework 18.17** A symmetric matrix  $A \in \mathbb{C}^{n \times n}$  is reducible if and only if there exists a permutation matrix  $P$  such that

$$PAP^T = \left( \begin{array}{c|c} T_{TL} & 0 \\ \hline 0 & T_{BR} \end{array} \right),$$

where  $T_{TL}$  (and hence  $T_{BR}$ ) is square of size  $b \times b$  where  $0 < b < n$ .

In other words, a symmetric matrix is reducible if and only if there exists a symmetric permutation of the rows and columns that leaves the matrix block diagonal.

**Theorem 18.18** If  $A$  is strictly diagonally dominant, then the Jacobi iteration converges.

**Proof:** Let

$$A = \underbrace{\begin{pmatrix} \alpha_{0,0} & 0 & \cdots & 0 \\ 0 & \alpha_{1,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{n-1,n-1} \end{pmatrix}}_M - \underbrace{\begin{pmatrix} 0 & -\alpha_{0,0} & \cdots & -\alpha_{0,n-1} \\ -\alpha_{1,0} & 0 & \cdots & -\alpha_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_{n-1,0} & -\alpha_{n-1,1} & \cdots & 0 \end{pmatrix}}_N.$$

Then

$$\begin{aligned} M^{-1}N &= \begin{pmatrix} \alpha_{0,0} & 0 & \cdots & 0 \\ 0 & \alpha_{1,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{n-1,n-1} \end{pmatrix}^{-1} \begin{pmatrix} 0 & -\alpha_{0,1} & \cdots & -\alpha_{0,n-1} \\ -\alpha_{1,0} & 0 & \cdots & -\alpha_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_{n-1,0} & -\alpha_{n-1,1} & \cdots & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\frac{\alpha_{0,1}}{\alpha_{0,0}} & \cdots & -\frac{\alpha_{0,n-1}}{\alpha_{0,0}} \\ -\frac{\alpha_{1,0}}{\alpha_{1,1}} & 0 & \cdots & -\frac{\alpha_{1,n-1}}{\alpha_{1,1}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\alpha_{n-1,0}}{\alpha_{n-1,n-1}} & -\frac{\alpha_{n-1,1}}{\alpha_{n-1,n-1}} & \cdots & 0 \end{pmatrix} \end{aligned}$$

Since  $A$  is strictly diagonally dominant, we know that

$$\sum_{j \neq i} \frac{|\alpha_{i,j}|}{|\alpha_{i,i}|} < 1.$$

By the Gershgorin Disk Theorem we therefore know that  $\rho(M^{-1}N) < 1$ . Hence we know there exists a norm  $\|\cdot\|$  such that  $\|M^{-1}N\| < 1$ , and hence the Jacobi iteration converges.

**Theorem 18.19** *If  $A$  is strictly diagonally dominant, then the Gram-Schmidt iteration converges.*

**Proof:**

**Theorem 18.20** *If  $A$  is weakly diagonally dominant, irreducible, and has at least one row for which*

$$|\alpha_{k,k}| > \sum_{j \neq k} |\alpha_{k,j}|,$$

*then the Jacobi and Gram-Schmidt iterations converge.*

**Theorem 18.21** *If  $A$  is SPD then SOR converges iff  $0 < \omega < 2$ .*

Proofs of these last few observations will be added to this notes in the future.

---





## Notes on Descent Methods and the Conjugate Gradient Method

**These notes are quite incomplete. More to come in the future!**

In this note, we assume that the  $n \times n$  matrix  $A$  is Symmetric Positive Definite. We are interested in iterative solving the linear system  $Ax = b$ . We will do so by creating a sequence of approximate solutions,  $\{x^{(0)}, x^{(1)}, \dots\}$  that, hopefully, converge to the true solution  $x$ .

## Outline

<b>Outline</b> . . . . .	<b>314</b>
<b>19.1. Basics</b> . . . . .	<b>315</b>
<b>19.2. Descent Methods</b> . . . . .	<b>315</b>
<b>19.3. Relation to Splitting Methods</b> . . . . .	<b>316</b>
<b>19.4. Method of Steepest Descent</b> . . . . .	<b>317</b>
<b>19.5. Preconditioning</b> . . . . .	<b>318</b>
<b>19.6. Methods of A-conjugate Directions</b> . . . . .	<b>318</b>
<b>19.7. Conjugate Gradient Method</b> . . . . .	<b>320</b>

## 19.1 Basics

The basic idea is to look at the problem of solving  $Ax = b$  instead as the minimization problem that finds the vector  $x$  that minimizes the function  $f(x) = \frac{1}{2}x^T Ax - x^T b$ .

**Theorem 19.1** *Let  $A$  be SPD and assume that  $Ax = b$ . Then the vector  $x$  minimizes the function  $f(y) = \frac{1}{2}y^T Ay - y^T b$ .*

**Proof:** Let  $Ax = b$ . Then

$$\begin{aligned} f(y) = \frac{1}{2}y^T Ay - y^T b &= \frac{1}{2}y^T Ay - y^T Ax \\ &= \frac{1}{2}y^T Ay - y^T Ax + \frac{1}{2}x^T Ax - \frac{1}{2}x^T Ax \\ &= \frac{1}{2}(y-x)^T A(y-x) - \frac{1}{2}x^T Ax. \end{aligned}$$

Since  $x^T Ax$  is independent of  $y$ , this is clearly minimized when  $y = x$ .

An alternative way to look at this is to note that the function is minimized when the gradient is zero and that  $\nabla f(x) = Ax - b$ .

## 19.2 Descent Methods

The basic idea behind a descent method is that at the  $k$ th iteration one has an approximation to  $x$ ,  $x^{(k)}$ . One would like to create a better approximation,  $x^{(k+1)}$ . To do so, one picks a *search direction*,  $p^{(k)}$ , and lets  $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ . In other words, one searches for a minimum along a line defined by the current iterate,  $x^{(k)}$ , and the search direction,  $p^{(k)}$ . One then picks  $\alpha_k$  so that, preferably,  $f(x^{(k+1)}) \leq f(x^{(k)})$ . Typically, one picks  $\alpha_k$  to exactly minimize the function along the line (leading to *exact* descent methods). Now

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha_k p^{(k)}) = \frac{1}{2}(x^{(k)} + \alpha_k p^{(k)})^T A(x^{(k)} + \alpha_k p^{(k)}) - (x^{(k)} + \alpha_k p^{(k)})^T b \\ &= \frac{1}{2}x^{(k)T} Ax^{(k)} + \alpha_k p^{(k)T} Ax^{(k)} + \frac{1}{2}\alpha_k^2 p^{(k)T} A p^{(k)} - x^{(k)T} b - \alpha_k p^{(k)T} b \\ &= \frac{1}{2}x^{(k)T} Ax^{(k)} - x^{(k)T} b + \alpha_k p^{(k)T} Ax^{(k)} + \frac{1}{2}\alpha_k^2 p^{(k)T} A p^{(k)} - \alpha_k p^{(k)T} b \\ &= f(x^{(k)}) + \frac{1}{2}\alpha_k^2 p^{(k)T} A p^{(k)} + \alpha_k p^{(k)T} (Ax^{(k)} - b) \\ &= f(x^{(k)}) + \frac{1}{2}p^{(k)T} A p^{(k)} \alpha_k^2 - p^{(k)T} r^{(k)} \alpha_k, \end{aligned}$$

where  $r^{(k)} = b - Ax^{(k)}$ , the residual. This is a quadratic equation in  $\alpha_k$  (since this is the only free variable). Thus, minimizing this expression exactly requires the derivative with respect to  $\alpha_k$  to be zero:

$$0 = \frac{df(x^{(k)} + \alpha_k p^{(k)})}{d\alpha_k} = p^{(k)T} A p^{(k)} \alpha_k - p^{(k)T} r^{(k)}.$$

Given: $A, b, x^{(0)}, p^{(0)}$	Given: $A, b, x^{(0)}, p^{(0)}$	Given: $A, b, x, p$
$r^{(0)} = b - Ax^{(0)}$	$r^{(0)} = b - Ax^{(0)}$	$r = b - Ax$
for $k = 0, 1, \dots$	for $k = 0, 1, \dots$	for $k = 0, 1, \dots$
$\alpha_k = \frac{p^{(k)T} r^{(k)}}{p^{(k)T} A p^{(k)}}$	$\alpha_k = \frac{p^{(k)T} r^{(k)}}{p^{(k)T} A p^{(k)}}$	$q = Ap$
$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$	$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$	$\alpha = \frac{p^T r}{p^T q}$
$r^{(k+1)} = b - Ax^{(k+1)}$	$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)}$	$x = x + \alpha p$
$p^{(k+1)} = \text{next direction}$	$p^{(k+1)} = \text{next direction}$	$r = r - \alpha q$
endfor	endfor	$p = \text{next direction}$
		endfor

Figure 19.1: Three basic descent method. The formulation on the middle follows from the one on the left by virtue of  $b - Ax^{(k+1)} = b - A(x^{(k)} + \alpha_k p^{(k)}) = r^{(k)} - \alpha_k A p^{(k)}$ . It is preferred because it requires only one matrix-vector multiplication per iteration (to form  $A p^{(k)}$ ). The one on the right overwrites the various vectors to reduce storage.

and hence, for exact descent methods,

$$\alpha_k = \frac{p^{(k)T} r^{(k)}}{p^{(k)T} A p^{(k)}} \quad \text{and} \quad x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}.$$

A question now becomes how to pick the search directions  $\{p^{(0)}, p^{(1)}, \dots\}$ .

Basic decent methods based on these ideas are given in Figure 19.1. What is missing in those algorithms is a stopping criteria. Notice that  $\alpha_k = 0$  if  $p^{(k)T} r^{(k)} = 0$ . But this does not necessarily mean that we have converged, since it merely means that  $p^{(k)}$  is orthogonal to  $r^{(k)}$ . What we will see is that commonly used methods pick  $p^{(k)}$  not to be orthogonal to  $r^{(k)}$ . Then  $p^{(k)T} r^{(k)} = 0$  implies  $r^{(k)} = 0$  and this condition, cheap to compute, can be used as a stopping criteria.

## 19.3 Relation to Splitting Methods

Let us do something really simple:

- Pick  $p^{(k)} = e_{k \bmod n}$ .
- $p^{(0)} = e_0$ .
- $p^{(0)T} A p^{(0)} = e_0^T A e_0 = \alpha_{0,0}$  (the  $(0,0)$  element in  $A$ ).
- $r^{(0)} = Ax^{(0)} - b$ .
- $p^{(0)T} r^{(0)} = e_0^T (Ax^{(0)} - b) = e_0^T A x^{(0)} - e_0^T b = \hat{a}_0^T x^{(0)} - \beta_0$ , where  $\hat{a}_i$  denotes the  $i$ th row of  $A$ .

- $x^{(1)} = x^{(0)} + \alpha_0 p^{(0)} = x^{(0)} + \frac{p^{(0)T} r^{(0)}}{p^{(0)T} A p^{(0)}} e_0 = x^{(0)} + \frac{\beta_0 - \tilde{a}_0^T x^{(0)}}{\alpha_{0,0}} e_0$ . This means that only the first element of  $x^{(0)}$  changes, and it changes to

$$x_0^{(1)} = x_0^{(0)} + \frac{1}{\alpha_{0,0}} \left( \beta_0 - \sum_{j=0}^{n-1} \alpha_{0,j} x_j^{(0)} \right) = \frac{1}{\alpha_{0,0}} \left( \beta_0 - \sum_{j=1}^{n-1} \alpha_{0,j} x_j^{(0)} \right).$$

Careful contemplation then reveals that this is exactly how the first element in vector  $x$  is changed in the Gauss-Seidel method!

We leave it to the reader to similarly deduce that  $x^{(n)}$  is exactly the vector one gets from applying one step of Gauss-Seidel (updating all elements of  $x$  once):

$$x^{(n)} = D^{-1}((L + L^T)x^{(0)} + b)$$

where  $D$  and  $-(L + L^T)$  equal the diagonal and off-diagonal parts of  $A$  so that  $A = M - N = D - (L + L^T)$ .

Given that choosing the search directions cyclically as  $\{e_0, e_1, \dots\}$  yields the Gauss-Seidel iteration, one can imagine picking  $x^{(k+1)} = x^{(k)} + \omega \alpha_k p^{(k)}$ , with some relaxation factor  $\omega$ . This would yield SOR if we continue to use, cyclically, the search directions  $\{e_0, e_1, \dots\}$ .

When the next iterate is chosen to equal the exact minimum along the line defined by the search direction, the method is called an *exact* descent direction and it is, clearly, guaranteed that  $f(x^{(k+1)}) \leq f(x^{(k)})$ . When a relaxation parameter  $\omega \neq 1$  is used, this is no longer the case, and the method is called *inexact*.

## 19.4 Method of Steepest Descent

For a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  that we are trying to minimize, for a given  $x$ , the direction in which the function most rapidly increases in value at  $x$  is given by the gradient,

$$\nabla f(x).$$

Thus, the direction in which it decreases most rapidly is

$$-\nabla f(x).$$

For our function

$$f(x) = \frac{1}{2} x^T A x - x^T b$$

this direction of steepest descent is given by

$$-\nabla f(x) = -(Ax - b) = b - Ax.$$

Thus, recalling that  $r^{(k)} = b - Ax^{(k)}$ , the direction of steepest descent at  $x^{(k)}$  is given by  $p^{(k)} = r^{(k)} = b - Ax^{(k)}$ , the residual, so that the iteration becomes

$$x^{(k+1)} = x^{(k)} + \underbrace{\frac{p^{(k)T} r^{(k)}}{p^{(k)T} A p^{(k)}}}_{\alpha_k} r^{(k)} = x^{(k)} + \underbrace{\frac{p^{(k)T} r^{(k)}}{p^{(k)T} q^{(k)}}}_{\alpha_k} r^{(k)},$$

where  $q^{(k)} = A p^{(k)}$ . We again notice that

$$r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + \alpha_k p^{(k)}) = b - Ax^{(k)} - \alpha_k A p^{(k)} = r^{(k)} - \alpha_k q^{(k)}.$$

These insights explain the algorithm in Figure 19.2 (left).

## 19.5 Preconditioning

For general nonlinear  $f(x)$ , using the direction of steepest descent as the search direction is often effective. Not necessarily so for our problem, if  $A$  is relatively ill-conditioned.

A picture clarifies this (on the blackboard). The key insight is that if  $\kappa(A) = \lambda_0/\lambda_{n-1}$  (the ratio between the largest and smallest eigenvalues or, equivalently, the ratio between the largest and smallest singular value), then convergence can take many iterations.

What would happen if  $\lambda_0 = \dots = \lambda_{n-1}$ ? Then  $A = Q\Lambda Q^T$  is the spectral decomposition of  $A$  and  $A = Q(\lambda_0 I)Q^T = \lambda_0 I$  and

$$x^{(1)} = x^{(0)} - \frac{r^{(0)T} r^{(0)}}{r^{(0)T} \lambda_0 I r^{(0)}} r^{(0)} = x^{(0)} - \frac{1}{\lambda_0} r^{(0)} = x^{(0)} - \frac{1}{\lambda_0} (\lambda_0 x^{(0)} - b) = x^{(0)} - x^{(0)} + \frac{1}{\lambda_0} b = \frac{1}{\lambda_0} b,$$

which is the solution to  $\lambda_0 Ix = b$ . Thus, the iteration converges in one step.

The point we are trying to make is that if  $A$  is well-conditioned, then the method of steepest descent converges faster. Now,  $Ax = b$  is equivalent to  $M^{-1}Ax = M^{-1}b$ . Hence, one can define a new problem with the same solution and, hopefully, a better condition number by letting  $\tilde{A} = M^{-1}A$  and  $\tilde{b} = M^{-1}b$ . The problem is that our methods are defined for SPD matrices. Generally speaking,  $M^{-1}A$  will not be SPD.

If one chooses  $M$  to be SPD with Cholesky factorization  $M = LL^T$ , then one can also transform the problem into  $L^{-1}AL^{-T}L^T x = L^{-1}b$ . One then solves  $\tilde{A}\tilde{x} = \tilde{b}$  where  $\tilde{A} = L^{-1}AL^{-T}$ ,  $\tilde{x} = L^T x$ , and  $\tilde{b} = L^{-1}b$  and eventually transforms the solution of this back to solution of the original problem by solving  $L^T x = \tilde{x}$ . If  $M$  is chosen carefully,  $\kappa(L^{-1}AL^{-T})$  can be greatly improved. The best choice would be  $M = A$ , of course, but that is not realistic. The matrix  $M$  is called the preconditioner. Some careful rearrangement takes the method of steepest descent on the transformed problem to the much simpler preconditioned algorithm in Figure 19.2.

## 19.6 Methods of A-conjugate Directions

We now borrow heavily from the explanation in Golub and Van Loan [21].

If we start with  $x^{(0)} = 0$ , then

$$x^{(k+1)} = \alpha_0 p^{(0)} + \dots + \alpha_k p^{(k)}.$$

Thus,  $x^{(k+1)} \in \mathcal{S}(p^{(0)}, \dots, p^{(k)})$ . It would be nice if each  $x^{(k+1)}$  satisfied

$$f(x^{(k+1)}) = \min_{x \in \mathcal{S}(p^{(0)}, \dots, p^{(k)})} f(x) \quad (19.1)$$

and the search directions were linearly independent. The primary reason is that then the descent direction method is guaranteed to complete in at most  $n$  iterations, since then

$$\mathcal{S}(p^{(0)}, \dots, p^{(n-1)}) = \mathbb{R}^n$$

so that

$$f(x^{(n)}) = \min_{x \in \mathcal{S}(p^{(0)}, \dots, p^{(n-1)})} f(x) = \min_{x \in \mathbb{R}^n} f(x)$$

so that  $Ax^{(n)} = b$ .

Given: $A, b, x^{(0)}$	Given: $A, b, x^{(0)}$	Given: $A, b, x^{(0)}$
$r^{(0)} = b - Ax^{(0)}, p^{(0)} = r^{(0)}$	$\tilde{A} = L^{-1}AL^{-T}, \tilde{b} = L^{-1}b, \tilde{x}^{(0)} = L^T x^{(0)}$	$r^{(0)} = b - Ax^{(0)}, p^{(0)} = M^{-1}r^{(0)}$
for $k = 0, 1, \dots$	for $k = 0, 1, \dots$	for $k = 0, 1, \dots$
$q^{(k)} = Ap^{(k)}$	$\tilde{q}^{(k)} = \tilde{A}\tilde{p}^{(k)}$	$q^{(k)} = Ap^{(k)}$
$\alpha_k = \frac{p^{(k)T}r^{(k)}}{p^{(k)T}q^{(k)}}$	$\tilde{\alpha}_k = \frac{\tilde{p}^{(k)T}\tilde{r}^{(k)}}{\tilde{p}^{(k)T}\tilde{q}^{(k)}}$	$\alpha_k = \frac{p^{(0)T}r^{(0)}}{p^{(0)T}q^{(0)}}$
$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$	$\tilde{x}^{(k+1)} = \tilde{x}^{(k)} + \tilde{\alpha}_k \tilde{p}^{(k)}$	$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$
$r^{(k+1)} = r^{(k)} - \alpha_k q^{(k)}$	$\tilde{r}^{(k+1)} = \tilde{r}^{(k)} - \tilde{\alpha}_k \tilde{q}^{(k)}$	$r^{(k+1)} = r^{(k)} - \alpha_k q^{(k)}$
$p^{(k+1)} = r^{(k+1)}$	$\tilde{p}^{(k+1)} = \tilde{r}^{(k+1)}$	$p^{(k+1)} = M^{-1}r^{(k+1)}$
endfor	endfor	endfor
	$x^{(k+1)} = L^{-T}\tilde{x}^{(k+1)}$	

Figure 19.2: Left: method of steepest decent. Middle: method of steepest decent with transformed problem. Right: preconditioned method of steepest decent. It can be checked that the  $x^{(k)}$  computed by the middle algorithm is exactly the  $x^{(k)}$  computed by the one on the right. Of course, the computation  $x^{(k+1)} = L^{-T}\tilde{x}^{(k+1)}$  needs only be done once, after convergence, in the algorithm in the middle.

Unfortunately, the method of steepest descent does not have this property. The iterate  $x^{(k+1)}$  minimizes  $f(x)$  where  $x$  is constraint to be on the line  $x^{(k)} + \alpha p^{(k)}$ . Because in each step  $f(x^{(k+1)}) \leq f(x^{(k)})$ , a slightly stronger result holds: It also minimizes  $f(x)$  where  $x$  is constraint to be on the union of lines  $x^{(j)} + \alpha p^{(j)}$ ,  $j = 0, \dots, k$ . However, that is not the same as it minimizing over all vectors in  $\mathcal{S}(p^{(0)}, \dots, p^{(k)})$ .

We can write (19.1) more concisely: Let  $P^{(k)} = \begin{pmatrix} p^{(0)} & p^{(1)} & \dots & p^{(k)} \end{pmatrix}$ . Then

$$\begin{aligned}
\min_{x \in \mathcal{S}(p^{(0)}, \dots, p^{(k)})} f(x) &= \min_y f(P^{(k)}y) = \min_y f\left(\begin{pmatrix} P^{(k-1)} & p^{(k)} \end{pmatrix} \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix}\right) \\
&= \min_y \left[ \frac{1}{2} \left[ \begin{pmatrix} P^{(k-1)} & p^{(k)} \end{pmatrix} \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix} \right]^T A \begin{pmatrix} P^{(k-1)} & p^{(k)} \end{pmatrix} \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix} \right. \\
&\quad \left. - \left[ \begin{pmatrix} P^{(k-1)} & p^{(k)} \end{pmatrix} \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix} \right]^T b \right] \\
&= \min_y \left[ \frac{1}{2} [y_0^T P^{(k-1)T} + \psi_1 p^{(k)T}] A [P^{(k-1)} y_0 + \psi_1 p^{(k)}] \right. \\
&\quad \left. - [y_0^T P^{(k-1)T} + \psi_1 p^{(k)T}] b \right] \\
&= \min_y \left[ \frac{1}{2} y_0^T P^{(k-1)T} A P^{(k-1)} y_0 + \psi_1 y_0^T P^{(k-1)T} A p^{(k)} + \frac{1}{2} \psi_1^2 p^{(k)T} A p^{(k)} - y_0^T P^{(k-1)T} b - \psi_1 p^{(k)T} b \right]
\end{aligned}$$

$$= \min_y \left[ \frac{1}{2} y_0^T P^{(k-1)T} A P^{(k-1)} y_0 - y_0^T P^{(k-1)T} b + \psi_1 y_0^T P^{(k-1)T} A p^{(k)} + \frac{1}{2} \psi_1^2 p^{(k)T} A p^{(k)} - \psi_1 p^{(k)T} b \right].$$

where  $y = \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix}$ . Now, if

$$P^{(k-1)T} A p^{(k)} = 0$$

then

$$\begin{aligned} \min_{x \in S(p^{(0)}, \dots, p^{(k)})} f(x) &= \min_y f(P^{(k)} y) \\ &= \min_y \left[ \frac{1}{2} y_0^T P^{(k-1)T} A P^{(k-1)} y_0 - y_0^T P^{(k-1)T} b + \frac{1}{2} \psi_1^2 p^{(k)T} A p^{(k)} - \psi_1 p^{(k)T} b \right] \\ &= \min_{y_0} \left[ \frac{1}{2} y_0^T P^{(k-1)T} A P^{(k-1)} y_0 - y_0^T P^{(k-1)T} b \right] + \min_{\psi_1} \left[ \frac{1}{2} \psi_1^2 p^{(k)T} A p^{(k)} - \psi_1 p^{(k)T} b \right] \\ &= \min_{y_0} f(P^{(k-1)} y_0) + 0. \end{aligned}$$

where the minimizing  $\psi_1$  is given by

$$\psi_1 = \frac{p^{(k)T} b}{p^{(k)T} A p^{(k)}}$$

so that

$$x^{(k+1)} = P^{(k-1)} y_0 + \psi_1 p^{(k)} = x^{(k)} + \psi_1 p^{(k)}.$$

Since

$$P^{(k-1)T} A p^{(k)}$$

is equivalent to  $p^{(j)T} A p^{(k)} = 0$ ,  $j = 0, \dots, k-1$ , we are looking for a sequence of directions that are A-conjugate:

**Definition 19.2** Let  $p^{(0)}, \dots, p^{(k-1)} \in \mathbb{R}^n$  and  $A$  be SPD. Then this sequence of vectors is said to be A-conjugate if  $p^{(j)T} A p^{(k)} = 0$  if and only if  $j \neq k$ .

## 19.7 Conjugate Gradient Method

The Conjugate Gradient method (CG) is a descent method that picks the search directions very carefully. Specifically, it picks them to be "A-conjugate", meaning that  $p_i^T A p_j = 0$  if  $i \neq j$ . It also starts with  $x^{(0)} = 0$  so that  $p^{(0)} = r^{(0)} = b$ .

Notice that, if  $x^{(0)} = 0$  then, by nature, descent methods yield

$$x^{(k+1)} = \alpha_0 p^{(0)} + \dots + \alpha_k p^{(k)}.$$

Also, the residual

$$r^{(k+1)} = b - A x^{(k+1)} = b - A(\alpha_0 p^{(0)} + \dots + \alpha_k p^{(k)}) = b - \alpha_0 A p^{(0)} - \dots - \alpha_k A p^{(k)}.$$



<p>Given: <math>A, b, x^{(0)} = 0</math>  <math>r^{(0)} = b, p^{(0)} = r^{(0)}</math>          for <math>k = 0, 1, \dots</math>  <math>q^{(k)} = Ap^{(k)}</math>  <math>\alpha_k = \frac{p^{(k)T} r^{(k)}}{p^{(k)T} q^{(k)}}</math>  <math>x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}</math>  <math>r^{(k+1)} = r^{(k)} - \alpha_k q^{(k)}</math>  <math>\gamma_k = -\frac{r^{(k+1)T} Ap^{(k)}}{p^{(k)T} Ap^{(k)}}</math>  <math>p^{(k+1)} = r^{(k+1)} + \gamma_k p^{(k)}</math>          endfor</p>	<p>Given: <math>A, b, x^{(0)} = 0</math></p>	<p>Given: <math>A, b, x^{(0)} = 0</math></p>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------	----------------------------------------------

Figure 19.3: Left: Basic Conjugate Gradient Method.

Now, if  $r^{(k+1)} = 0$ , then we know that  $x^{(k+1)}$  solves  $Ax = b$ , and we are done. So, we can assume that  $r^{(k+1)} \neq 0$ . How can we construct a new  $p^{(k+1)}$  such that  $p^{(k+1)T} Ap^{(k)} = 0$ . For the Conjugate Gradient method we pick

$$p^{(k+1)} = r^{(k+1)} + \gamma_k p^{(k)}$$

such that

$$p^{(k+1)T} Ap^{(k)} = 0.$$

Now,

$$0 = p^{(k+1)T} Ap^{(k)} = (r^{(k+1)} + \gamma_k p^{(k)})^T Ap^{(k)} = r^{(k+1)T} Ap^{(k)} + \gamma_k p^{(k)T} Ap^{(k)}$$

so that

$$\gamma_k = -\frac{r^{(k+1)T} Ap^{(k)}}{p^{(k)T} Ap^{(k)}}.$$

Now, it can be shown that not only  $p^{(k+1)T} Ap^{(k)}$ , but also  $p^{(k+1)T} Ap^{(j)}$ , for  $j = 0, 1, \dots, k$ . Thus, the Conjugate Gradient Method is an A-conjugate method.

Working these insights into the basic descent method algorithm yields the most basic form of the Conjugate Gradient method, summarized in Figure 19.3 (left).



# Chapter 20

## Notes on Lanczos Methods

**These notes are quite incomplete. More to come in the future!**

In this note, we give a brief overview of the ideas behind Lanczos Methods. These methods are useful when computing (some) eigenvalues and eigenvectors of a sparse matrix.

## Outline

<b>Outline</b> . . . . .	<b>324</b>
<b>20.1. Krylov Subspaces</b> . . . . .	<b>325</b>
<b>20.2. The Lanczos Method</b> . . . . .	<b>325</b>

## 20.1 Krylov Subspaces

Consider matrix  $A$ , some initial unit vector  $q_0$ , and the sequence  $x_0, x_1, \dots$  defined by

$$x_{k+1} := Ax_k, \quad k = 0, 1, \dots,$$

where  $x_0 = q_0$ . We already saw this sequence when we discussed the power method.

**Definition 20.1** Given square matrix  $A$  and initial vector  $x_0$ , the Krylov subspace  $\mathcal{K}(A, x_0, k)$  is defined by

$$\mathcal{K}(A, x_0, k) = \mathcal{S}(x_0, Ax_0, \dots, A^{k-1}x_0) = \mathcal{S}(x_0, x_1, \dots, x_{k-1}),$$

where  $x_{j+1} = Ax_j$ , for  $j = 0, 1, \dots$ . The Krylov matrix is defined by

$$K(A, x_0, k) = \left( x_0 \mid Ax_0 \mid \dots \mid A^{k-1}x_0 \right) = \left( x_0 \mid x_1 \mid \dots \mid x_{k-1} \right)$$

## 20.2 The Lanczos Method

From our discussion of the Power Method, we notice that it is convenient to instead work with mutually orthonormal vectors. Consider the QR factorization of  $K(A, q_0, n)$ :

$$K(A, q_0, n) = \left( q_0 \mid x_1 \mid \dots \mid x_{n-1} \right) = \left( q_0 \mid q_1 \mid \dots \mid q_{n-1} \right) \begin{pmatrix} 1 & \rho_{0,1} & \dots & \rho_{0,n-1} \\ 0 & \rho_{1,1} & \dots & \rho_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \rho_{n-1,n-1} \end{pmatrix} = QR$$

Notice that

$$AK(A, q_0, n) = A \left( q_0 \mid x_1 \mid \dots \mid x_{n-1} \right) = \left( x_1 \mid \dots \mid x_{n-1} \mid Ax_{n-1} \right).$$

Partition

$$K(A, q_0, n) = \left( q_0 \mid X_1 \right), Q = \left( q_0 \mid Q_1 \right), \text{ and } R = \begin{pmatrix} 1 & r_{01}^T \\ 0 & R_{11} \end{pmatrix}.$$

Then

$$K(A, q_0, n) = A \left( q_0 \mid X_1 \right) = \left( X_1 \mid Ax_{n-1} \right)$$

and hence

$$K(A, q_0, n) = A \left( q_0 \mid Q_1 \right) \begin{pmatrix} 1 & r_{01}^T \\ 0 & R_{11} \end{pmatrix} = \left( Q_1 R_{11} - q_0 r_{01}^T \mid Ax_{n-1} \right)$$

Now, notice that

$$\begin{aligned} Q^T A Q R &= \left( q_0 \mid Q_1 \right)^T A \left( q_0 \mid Q_1 \right) R \\ &= \left( q_0 \mid Q_1 \right)^T \left( Q_1 R_{11} - q_0 r_{01}^T \mid Ax_{n-1} \right) \end{aligned}$$

$$\begin{aligned}
&= \left( \left( q_0 \mid Q_1 \right)^T (Q_1 R_{11} - q_0 r_{01}^T) \mid \left( q_0 \mid Q_1 \right)^T A x_{n-1} \right) \\
&= \left( \frac{r_{01}^T}{R_{11}} \mid \frac{q_0^T A x_{n-1}}{Q_1^T A x_{n-1}} \right).
\end{aligned}$$

Finally, we see that (if  $R$  is invertible)

$$Q^T A Q = \underbrace{\left( \frac{r_{01}^T}{R_{11}} \mid \frac{q_0^T A x_{n-1}}{Q_1^T A x_{n-1}} \right)}_{\text{upperHessenberg!}} R^{-1} = T.$$

We emphasize that

- $T = Q^T A Q$  is upperHessenberg (and if  $A$  is symmetric, it is tridiagonal).
- $Q$  is generated by the initial vector  $x_0 = q_0$ .

From the Implicit Q Theorem, we thus know that  $T$  and  $Q$  are uniquely determined! Finally,  $Q^T A Q$  is a unitary similarity transformations so that the eigenvalues can be computed from  $T$  and then the eigenvectors of  $T$  can be transformed back to eigenvectors of  $Q$ .

One way for computing the tridiagonal matrix is, of course, to use Householder transformations. However, that would quickly fill zeroes that exist in matrix  $A$ . Instead, we compute it more direction. We will do so under the assumption that  $A$  is symmetric and that therefore  $T$  is tridiagonal.

Consider  $AQ = QT$  and partition

$$Q = \left( Q_L \mid q_M \mid Q_R \right) \quad \text{and} \quad T = \left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML} e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM} e_0 & T_{BR} \end{array} \right)$$

The idea is that  $Q_L$ ,  $q_M$  have already been computed, as have  $T_{TL}$  and  $\tau_{ML}$ . Also, we assume that a temporary vector  $u$  holds  $Aq_M - \tau_{ML} Q_L e_L$ . Initially,  $Q_L$  has zero rows and  $q_M$  equals the  $q_1$  with which we started our discussion. Also, initially,  $T_{TL}$  is  $0 \times 0$  so that  $\tau_{ML}$  doesn't really exist.

Now, we repartition

$$\left( Q_L \mid q_M \mid Q_R \right) \rightarrow \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right)$$

and

$$\left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML} e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM} e_0 & T_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c|c} T_{00} & \star & 0 & 0 \\ \hline \tau_{10} e_L^T & \tau_{11} & \star & 0 \\ \hline 0 & \tau_{21} & \tau_{22} & \star \\ \hline 0 & 0 & \tau_{32} e_0 & T_{33} \end{array} \right),$$

which one could visualize as

$$\left( \begin{array}{cc|c|ccc} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \tau_{ML} & 0 & 0 & 0 \\ \hline 0 & \tau_{ML} & \tau_{MM} & \tau_{BM} & 0 & 0 \\ \hline 0 & 0 & \tau_{BM} & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{array} \right) \rightarrow \left( \begin{array}{cc|c|ccc} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \tau_{10} & 0 & 0 & 0 \\ \hline 0 & \tau_{10} & \tau_{11} & \tau_{21} & 0 & 0 \\ \hline 0 & 0 & \tau_{21} & \tau_{22} & \tau_{32} & 0 \\ \hline 0 & 0 & 0 & \tau_{32} & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{array} \right),$$

so that

$$A \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right) = \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right) \begin{pmatrix} T_{00} & \star & 0 & 0 \\ \tau_{10}e_L^T & \tau_{11} & \star & 0 \\ \hline 0 & \tau_{21} & \tau_{22} & \star \\ \hline 0 & 0 & \tau_{32}e_0 & T_{33} \end{pmatrix}$$

Notice that then

$$Aq_1 = \tau_{10}Q_0e_L + \tau_{11}q_1 + \tau_{21}q_2.$$

We assumed that  $u$  already contained  $u := Aq_M - \tau_{ML}Q_Le_L$  which means that  $u = Aq_1 - \tau_{10}Q_0e_L$ . (Notice that  $Q_0e_L$  is merely the last column of  $Q_0$ .) Next,

$$q_1^T u = q_1^T (Aq_1 - \tau_{10}Q_0e_L) = \tau_{11}q_1^T q_1 + \tau_{21}q_1^T q_2 = \tau_{11},$$

which means that  $\tau_{11} := q_1^T u$ . This allows us to update  $u := u - \tau_{11}q_1 = Aq_1 - \tau_{10}Q_0e_L - \tau_{11}q_1$ . We notice that now  $\tau_{21}q_2 = u$  and hence we can choose  $\tau_{21} := \|q_2\|_2$ , since  $q_2$  must have length one, and then  $q_2 := u/\tau_{21}$ . We then create a new vector  $u := Aq_2 - \tau_{21}q_1$ . This allows us to move where we are in the matrices forward:

$$\left( Q_L \mid q_M \mid Q_R \right) \leftarrow \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right)$$

and

$$\left( \begin{array}{cc|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML}e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM}e_0 & T_{BR} \end{array} \right) \leftarrow \left( \begin{array}{cc|c|ccc} T_{00} & \star & 0 & 0 \\ \hline \tau_{10}e_L^T & \tau_{11} & \star & 0 \\ \hline 0 & \tau_{21} & \tau_{22} & \star \\ \hline 0 & 0 & \tau_{32}e_0 & T_{33} \end{array} \right),$$

which one could visualize as

$$\left( \begin{array}{ccc|cc} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \tau_{ML} & 0 & 0 & 0 \\ \hline 0 & \tau_{ML} & \tau_{MM} & \tau_{BM} & 0 & 0 \\ \hline 0 & 0 & \tau_{BM} & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{array} \right) \leftarrow \left( \begin{array}{ccc|cc} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \tau_{10} & 0 & 0 & 0 \\ \hline 0 & \tau_{10} & \tau_{11} & \tau_{21} & 0 & 0 \\ \hline 0 & 0 & \tau_{21} & \tau_{22} & \tau_{32} & 0 \\ \hline 0 & 0 & 0 & \tau_{32} & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{array} \right).$$

These observations are captured in the algorithm in Figure 20.1.

**Algorithm:**  $[Q, T] := \text{LANCZOS\_UNB\_VAR1}(A, q_0, Q, T)$

**Partition**  $Q \rightarrow \left( Q_L \mid q_M \mid Q_R \right), T \rightarrow \left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML}e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM}e_0 & T_{BR} \end{array} \right)$

**where**  $Q_L$  has 0 columns,  $T_{TL}$  is  $0 \times 0$

$q_M := q_0; \quad u = Aq_0$

**while**  $n(Q_L) < n(Q)$  **do**

**Repartition**

$\left( Q_L \mid q_M \mid Q_R \right) \rightarrow \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right),$   
 $\left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML}e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM}e_0 & T_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c|c} T_{00} & \star & 0 & 0 \\ \hline \tau_{10}e_L^T & \tau_{11} & \star & \star \\ \hline 0 & \tau_{21} & \tau_{22} & \star \\ \hline 0 & 0 & \tau_{32}e_0 & T_{33} \end{array} \right)$

**where**  $q_2$  has 1 column,  $\tau_{22}$  is  $1 \times 1$

$\tau_{11} := q_1^T u$

$u := u - \tau_{11}q_1$

$\tau_{21} := \|q_2\|_2$

$q_2 := u/\tau_{21}$

$u := Aq_2 - \tau_{21}q_1$

**Continue with**

$\left( Q_L \mid q_M \mid Q_R \right) \rightarrow \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right),$   
 $\left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML}e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM}e_0 & T_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c|c} T_{00} & \star & 0 & 0 \\ \hline \tau_{10}e_L^T & \tau_{11} & \star & \star \\ \hline 0 & \tau_{21} & \tau_{22} & \star \\ \hline 0 & 0 & \tau_{32}e_0 & T_{33} \end{array} \right)$

**endwhile**

Figure 20.1: Lanczos algorithm for symmetric  $Q$ . It computes  $Q$  and  $T$  such that  $AQ = QT$ . It lacks a stopping criteria.



Consider

$$A \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right) = \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right) \begin{pmatrix} T_{00} & \star & 0 & 0 \\ \tau_{10}e_L^T & \tau_{11} & \star & 0 \\ 0 & \tau_{21} & \tau_{22} & \star \\ 0 & 0 & \tau_{32}e_0 & T_{33} \end{pmatrix}$$

- If  $q_2$  represents the  $k$ th column of  $Q$ , then it is in the direction of  $(A^{k-1}q_0)^\perp$ , the component of  $A^{k-1}q_0$  that is orthogonal to the columns of  $\left( Q_0 \mid q_1 \right)$ .
- As  $k$  gets large,  $A^{k-1}q_0$  is essentially in the space spanned by  $\left( Q_0 \mid q_1 \right)$ .
- $q_2$  is computed from

$$\tau_{21}q_2 = Aq_1 - \tau_{10}Q_0e_L - \tau_{11}q_1.$$

- The length of  $Aq_1 - \tau_{10}Q_0e_L - \tau_{11}q_1$  must be very small as  $k$  gets large since  $A^{k-1}q_0$  eventually lies approximately in the direction of  $A^{k-2}q_0$ .
- This means that  $\tau_{21} = \|Aq_1 - \tau_{10}Q_0e_L - \tau_{11}q_1\|_2$  will be small.
- Hence

$$A \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right) \approx \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right) \begin{pmatrix} T_{00} & \star & 0 & 0 \\ \tau_{10}e_L^T & \tau_{11} & 0 & 0 \\ 0 & 0 & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix}$$

or, equivalently,

$$A \left( Q_0 \mid q_1 \right) \approx \left( Q_0 \mid q_1 \right) \begin{pmatrix} T_{00} & \star \\ \tau_{10}e_L^T & \tau_{11} \end{pmatrix} = S,$$

where  $S$  is a  $(k-1) \times (k-1)$  tridiagonal matrix.

Here the space spanned by  $\left( Q_0 \mid q_1 \right)$ ,  $\mathcal{C}(\left( Q_0 \mid q_1 \right))$ , is said to be an *invariant subspace* of matrix  $A$ .

- If we compute the Spectral Decomposition of  $S$ :

$$S = Q_S \Lambda_S Q_S^T$$

then

$$A \left[ \left( Q_0 \mid q_1 \right) Q_S \right] \approx \left[ \left( Q_0 \mid q_1 \right) Q_S \right] \Lambda_S.$$

- We conclude that, approximately,
  - the diagonal elements of  $\Lambda_S$  equal eigenvalues of  $A$  and
  - the columns of  $\left[ \left( Q_0 \mid q_1 \right) Q_S \right]$  equal the corresponding eigenvectors.

The algorithm now, with stopping criteria, is given in Figure 20.2.

**Algorithm:**  $[Q_L, T_{TL}] := \text{LANCZOS\_UNB\_VAR1}(A, q_0, Q, T)$

**Partition**  $Q \rightarrow \left( Q_L \mid q_M \mid Q_R \right), T \rightarrow \left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML}e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM}e_0 & T_{BR} \end{array} \right)$

**where**  $Q_L$  has 0 columns,  $T_{TL}$  is  $0 \times 0$

$q_M := q_0; \quad u = Aq_0$

**while**  $|\tau_{ML}| > \text{tolerance}$  **do**

**Repartition**

$\left( Q_L \mid q_M \mid Q_R \right) \rightarrow \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right),$   
 $\left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML}e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM}e_0 & T_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c|c} T_{00} & \star & 0 & 0 \\ \hline \tau_{10}e_L^T & \tau_{11} & \star & \star \\ \hline 0 & \tau_{21} & \tau_{22} & \star \\ \hline 0 & 0 & \tau_{32}e_0 & T_{33} \end{array} \right)$

**where**  $q_2$  has 1 column,  $\tau_{22}$  is  $1 \times 1$

$\tau_{11} := q_1^T u$

$u := u - \tau_{11}q_1$

$\tau_{21} := \|q_2\|_2$

**if**  $|\tau_{21}| > \text{tolerance}$

$\left| \begin{array}{l} q_2 := u/\tau_{21} \\ u := Aq_2 - \tau_{21}q_1 \end{array} \right.$

**Continue with**

$\left( Q_L \mid q_M \mid Q_R \right) \rightarrow \left( Q_0 \mid q_1 \mid q_2 \mid Q_3 \right),$   
 $\left( \begin{array}{c|c|c} T_{TL} & \star & 0 \\ \hline \tau_{ML}e_L^T & \tau_{MM} & \star \\ \hline 0 & \tau_{BM}e_0 & T_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c|c} T_{00} & \star & 0 & 0 \\ \hline \tau_{10}e_L^T & \tau_{11} & \star & \star \\ \hline 0 & \tau_{21} & \tau_{22} & \star \\ \hline 0 & 0 & \tau_{32}e_0 & T_{33} \end{array} \right)$

**endwhile**

Figure 20.2: Lanczos algorithm for symmetric  $Q$ . It computes  $Q_L$  and  $T_{TL}$  such that  $AQ_L \approx Q_L T_{TL}$ .

**More Chapters to be Added in the Future!**



# Answers

## Chapter 1. Notes on Simple Vector and Matrix Operations (Answers)

### Homework 1.1 Partition A

$$A = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right) = \begin{pmatrix} \frac{\hat{a}_0^T}{\hat{a}_1^T} \\ \vdots \\ \frac{\hat{a}_{m-1}^T}{\hat{a}_{m-1}^T} \end{pmatrix}.$$

Convince yourself that the following hold:

$$\bullet \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right)^T = \begin{pmatrix} \frac{a_0^T}{a_1^T} \\ \vdots \\ \frac{a_{m-1}^T}{a_{m-1}^T} \end{pmatrix}.$$

$$\bullet \begin{pmatrix} \frac{\hat{a}_0^T}{\hat{a}_1^T} \\ \vdots \\ \frac{\hat{a}_{m-1}^T}{\hat{a}_{m-1}^T} \end{pmatrix}^T = \left( \hat{a}_0 \mid \hat{a}_1 \mid \cdots \mid \hat{a}_{n-1} \right).$$

$$\bullet \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right)^H = \begin{pmatrix} \frac{a_0^H}{a_1^H} \\ \vdots \\ \frac{a_{m-1}^H}{a_{m-1}^H} \end{pmatrix}.$$

$$\bullet \begin{pmatrix} \overline{\hat{a}_0^T} \\ \overline{\hat{a}_1^T} \\ \vdots \\ \overline{\hat{a}_{m-1}^T} \end{pmatrix}^H = \left( \overline{\hat{a}_0} \mid \overline{\hat{a}_1} \mid \cdots \mid \overline{\hat{a}_{n-1}} \right).$$

[🔙 BACK TO TEXT](#)

**Homework 1.2** Partition  $x$  into subvectors:

$$x = \begin{pmatrix} \overline{x_0} \\ \overline{x_1} \\ \vdots \\ \overline{x_{N-1}} \end{pmatrix}.$$

Convince yourself that the following hold:

$$\bullet \bar{x} = \begin{pmatrix} \overline{\overline{x_0}} \\ \overline{\overline{x_1}} \\ \vdots \\ \overline{\overline{x_{N-1}}} \end{pmatrix}.$$

$$\bullet x^T = \left( x_0^T \mid x_1^T \mid \cdots \mid x_{N-1}^T \right).$$

$$\bullet x^H = \left( x_0^H \mid x_1^H \mid \cdots \mid x_{N-1}^H \right).$$

[🔙 BACK TO TEXT](#)

**Homework 1.3** Partition  $A$

$$A = \begin{pmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,N-1} \end{pmatrix},$$

where  $A_{i,j} \in \mathbb{C}^{m_i \times n_j}$ . Here  $\sum_{i=0}^{M-1} m_i = m$  and  $\sum_{j=0}^{N-1} n_j = n$ .

Convince yourself that the following hold:

$$\begin{aligned}
& \bullet \begin{pmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,N-1} \end{pmatrix}^T = \begin{pmatrix} A_{0,0}^T & A_{1,0}^T & \cdots & A_{M-1,0}^T \\ A_{0,1}^T & A_{1,1}^T & \cdots & A_{M-1,1}^T \\ \vdots & \vdots & \cdots & \vdots \\ A_{0,N-1}^T & A_{1,N-1}^T & \cdots & A_{M-1,N-1}^T \end{pmatrix}. \\
& \bullet \begin{pmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,N-1} \\ \vdots & \vdots & \cdots & \vdots \\ A_{M-1,0} & A_{M-1,1} & \cdots & A_{M-1,N-1} \end{pmatrix}^H = \begin{pmatrix} A_{0,0}^H & A_{1,0}^H & \cdots & A_{M-1,0}^H \\ A_{0,1}^H & A_{1,1}^H & \cdots & A_{M-1,1}^H \\ \vdots & \vdots & \cdots & \vdots \\ A_{0,N-1}^H & A_{1,N-1}^H & \cdots & A_{M-1,N-1}^H \end{pmatrix}.
\end{aligned}$$

[👉 BACK TO TEXT](#)

**Homework 1.4** Convince yourself of the following:

- $\alpha x^T = \left( \alpha x_0 \mid \alpha x_1 \mid \cdots \mid \alpha x_{n-1} \right).$
- $(\alpha x)^T = \alpha x^T.$
- $(\alpha x)^H = \bar{\alpha} x^H.$
- $\alpha \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} = \begin{pmatrix} \alpha x_0 \\ \alpha x_1 \\ \vdots \\ \alpha x_{N-1} \end{pmatrix}$

[👉 BACK TO TEXT](#)

**Homework 1.5** Convince yourself of the following:

$$\bullet \alpha \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} + \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{pmatrix} = \begin{pmatrix} \alpha x_0 + y_0 \\ \alpha x_1 + y_1 \\ \vdots \\ \alpha x_{N-1} + y_{N-1} \end{pmatrix}. \text{ (Provided } x_i, y_i \in \mathbb{C}^{n_i} \text{ and } \sum_{i=0}^{N-1} n_i = n.)$$

[👉 BACK TO TEXT](#)

**Homework 1.6** Convince yourself of the following:

$$\bullet \quad \left( \begin{array}{c} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{array} \right)^H \left( \begin{array}{c} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{array} \right) = \sum_{i=0}^{N-1} x_i^H y_i. \text{ (Provided } x_i, y_i \in \mathbb{C}^{n_i} \text{ and } \sum_{i=0}^{N-1} n_i = n.)$$

[👉 BACK TO TEXT](#)

**Homework 1.7** Prove that  $x^H y = \overline{y^H x}$ .

[👉 BACK TO TEXT](#)



## Chapter 2. Notes on Vector and Matrix Norms (Answers)

**Homework 2.2** Prove that if  $v : \mathbb{C}^n \rightarrow \mathbb{R}$  is a norm, then  $v(0) = 0$  (where the first 0 denotes the zero vector in  $\mathbb{C}^n$ ).

**Answer:** Let  $x \in \mathbb{C}^n$  and  $\vec{0}$  the zero vector of size  $n$  and 0 the scalar zero. Then

$$\begin{aligned} v(\vec{0}) &= v(0 \cdot x) & 0 \cdot x &= \vec{0} \\ &= |0|v(x) & v(\cdot) \text{ is homogeneous} \\ &= 0 & \text{algebra} \end{aligned}$$

[👉 BACK TO TEXT](#)

**Homework 2.7** The vector 1-norm is a norm.

**Answer:** We show that the three conditions are met:

Let  $x, y \in \mathbb{C}^n$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $x \neq 0 \Rightarrow \|x\|_1 > 0$  ( $\|\cdot\|_1$  is positive definite):

Notice that  $x \neq 0$  means that at least one of its components is nonzero. Let's assume that  $\chi_j \neq 0$ . Then

$$\|x\|_1 = |\chi_0| + \cdots + |\chi_{n-1}| \geq |\chi_j| > 0.$$

- $\|\alpha x\|_1 = |\alpha| \|x\|_1$  ( $\|\cdot\|_1$  is homogeneous):

$$\begin{aligned} \|\alpha x\|_1 &= |\alpha \chi_0| + \cdots + |\alpha \chi_{n-1}| = |\alpha| |\chi_0| + \cdots + |\alpha| |\chi_{n-1}| = |\alpha| (|\chi_0| + \cdots + |\chi_{n-1}|) = \\ &= |\alpha| (|\chi_0| + \cdots + |\chi_{n-1}|) = |\alpha| \|x\|_1. \end{aligned}$$

- $\|x + y\|_1 \leq \|x\|_1 + \|y\|_1$  ( $\|\cdot\|_1$  obeys the triangle inequality).

$$\begin{aligned} \|x + y\|_1 &= |\chi_0 + \psi_0| + |\chi_1 + \psi_1| + \cdots + |\chi_{n-1} + \psi_{n-1}| \\ &\leq |\chi_0| + |\psi_0| + |\chi_1| + |\psi_1| + \cdots + |\chi_{n-1}| + |\psi_{n-1}| \\ &= |\chi_0| + |\chi_1| + \cdots + |\chi_{n-1}| + |\psi_0| + |\psi_1| + \cdots + |\psi_{n-1}| \\ &= \|x\|_1 + \|y\|_1. \end{aligned}$$

[👉 BACK TO TEXT](#)

**Homework 2.9** The vector  $\infty$ -norm is a norm.

**Answer:** We show that the three conditions are met:

Let  $x, y \in \mathbb{C}^n$  and  $\alpha \in \mathbb{C}$  be arbitrarily chosen. Then

- $x \neq 0 \Rightarrow \|x\|_\infty > 0$  ( $\|\cdot\|_\infty$  is positive definite):

Notice that  $x \neq 0$  means that at least one of its components is nonzero. Let's assume that  $\chi_j \neq 0$ . Then

$$\|x\|_\infty = \max_i |\chi_i| \geq |\chi_j| > 0.$$

- $\|\alpha x\|_\infty = |\alpha| \|x\|_\infty$  ( $\|\cdot\|_\infty$  is homogeneous):

$$\|\alpha x\|_\infty = \max_i |\alpha \chi_i| = \max_i |\alpha| |\chi_i| = |\alpha| \max_i |\chi_i| = |\alpha| \|x\|_\infty.$$

- $\|x+y\|_\infty \leq \|x\|_\infty + \|y\|_\infty$  ( $\|\cdot\|_\infty$  obeys the triangle inequality).

$$\begin{aligned} \|x+y\|_\infty &= \max_i |\chi_i + \psi_i| \\ &\leq \max_i (|\chi_i| + |\psi_i|) \\ &\leq \max_i (|\chi_i| + \max_j |\psi_j|) \\ &= \max_i |\chi_i| + \max_j |\psi_j| = \|x\|_\infty + \|y\|_\infty. \end{aligned}$$

➡ BACK TO TEXT

**Homework 2.13** Show that the Frobenius norm is a norm.

**Answer:** The answer is to realize that if  $A = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1} \end{pmatrix}$  then

$$\|A\|_F = \sqrt{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |\alpha_{i,j}|^2} = \sqrt{\sum_{j=0}^{n-1} \sum_{i=0}^{m-1} |\alpha_{i,j}|^2} = \sqrt{\sum_{j=0}^{n-1} \|a_j\|_2^2} = \sqrt{\left\| \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} \right\|_2^2}.$$

In other words, it equals the vector 2-norm of the vector that is created by stacking the columns of  $A$  on top of each other. The fact that the Frobenius norm is a norm then comes from realizing this connection and exploiting it.

Alternatively, just grind through the three conditions!

➡ BACK TO TEXT

**Homework 2.20** Let  $A \in \mathbb{C}^{m \times n}$  and partition  $A = \begin{pmatrix} \widehat{a}_0^T \\ \widehat{a}_1^T \\ \vdots \\ \widehat{a}_{m-1}^T \end{pmatrix}$ . Show that

$$\|A\|_\infty = \max_{0 \leq i < m} \|\widehat{a}_i\|_1 = \max_{0 \leq i < m} (|\alpha_{i,0}| + |\alpha_{i,1}| + \cdots + |\alpha_{i,n-1}|)$$

**Answer:** Partition  $A = \begin{pmatrix} \frac{\hat{a}_0^T}{\hat{a}_1^T} \\ \vdots \\ \frac{\hat{a}_{m-1}^T}{\hat{a}_{m-1}^T} \end{pmatrix}$ . Then

$$\begin{aligned}
\|A\|_\infty &= \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{\|x\|_\infty=1} \left\| \begin{pmatrix} \frac{\hat{a}_0^T}{\hat{a}_1^T} \\ \vdots \\ \frac{\hat{a}_{m-1}^T}{\hat{a}_{m-1}^T} \end{pmatrix} x \right\|_\infty = \max_{\|x\|_\infty=1} \left\| \begin{pmatrix} \frac{\hat{a}_0^T x}{\hat{a}_1^T x} \\ \vdots \\ \frac{\hat{a}_{m-1}^T x}{\hat{a}_{m-1}^T x} \end{pmatrix} \right\|_\infty \\
&= \max_{\|x\|_\infty=1} \left( \max_i |a_i^T x| \right) = \max_{\|x\|_\infty=1} \max_i \left| \sum_{p=0}^{n-1} \alpha_{i,p} \chi_p \right| \leq \max_{\|x\|_\infty=1} \max_i \sum_{p=0}^{n-1} |\alpha_{i,p} \chi_p| \\
&= \max_{\|x\|_\infty=1} \max_i \sum_{p=0}^{n-1} (|\alpha_{i,p}| |\chi_p|) \leq \max_{\|x\|_\infty=1} \max_i \sum_{p=0}^{n-1} (|\alpha_{i,p}| (\max_k |\chi_k|)) \leq \max_{\|x\|_\infty=1} \max_i \sum_{p=0}^{n-1} (|\alpha_{i,p}| \|x\|_\infty) \\
&= \max_i \sum_{p=0}^{n-1} (|\alpha_{i,p}|) = \|\hat{a}_i\|_1
\end{aligned}$$

so that  $\|A\|_\infty \leq \max_i \|\hat{a}_i\|_1$ .

We also want to show that  $\|A\|_\infty \geq \max_i \|\hat{a}_i\|_1$ . Let  $k$  be such that  $\max_i \|\hat{a}_i\|_1 = \|\hat{a}_k\|_1$  and pick  $y = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{n-1} \end{pmatrix}$  so that  $\hat{a}_k^T y = |\alpha_{k,0}| + |\alpha_{k,1}| + \cdots + |\alpha_{k,n-1}| = \|\hat{a}_k\|_1$ . (This is a matter of picking  $\psi_i$  so that  $|\psi_i| = 1$  and  $\psi_i \alpha_{k,i} = |\alpha_{k,i}|$ .) Then

$$\begin{aligned}
\|A\|_\infty &= \max_{\|x\|_1=1} \|Ax\|_\infty = \max_{\|x\|_1=1} \left\| \begin{pmatrix} \frac{\hat{a}_0^T}{\hat{a}_1^T} \\ \vdots \\ \frac{\hat{a}_{m-1}^T}{\hat{a}_{m-1}^T} \end{pmatrix} x \right\|_\infty \geq \left\| \begin{pmatrix} \frac{\hat{a}_0^T}{\hat{a}_1^T} \\ \vdots \\ \frac{\hat{a}_{m-1}^T}{\hat{a}_{m-1}^T} \end{pmatrix} y \right\|_\infty \\
&= \left\| \begin{pmatrix} \frac{\hat{a}_0^T y}{\hat{a}_1^T y} \\ \vdots \\ \frac{\hat{a}_{m-1}^T y}{\hat{a}_{m-1}^T y} \end{pmatrix} \right\|_\infty \geq |\hat{a}_k^T y| = \hat{a}_k^T y = \|\hat{a}_k\|_1 = \max_i \|\hat{a}_i\|_1
\end{aligned}$$

 BACK TO TEXT

**Homework 2.21** Let  $y \in \mathbb{C}^m$  and  $x \in \mathbb{C}^n$ . Show that  $\|yx^H\|_2 = \|y\|_2\|x\|_2$ .

**Answer:** Answer needs to be filled in.

🔍 BACK TO TEXT

**Homework 2.25** Show that  $\|Ax\|_\mu \leq \|A\|_{\mu,\nu}\|x\|_\nu$ .

**Answer:** W.l.o.g. let  $x \neq 0$ .

$$\|A\|_{\mu,\nu} = \max_{y \neq 0} \frac{\|Ay\|_\mu}{\|y\|_\nu} \geq \frac{\|Ax\|_\mu}{\|x\|_\nu}.$$

Rearranging this establishes the result.

🔍 BACK TO TEXT

**Homework 2.26** Show that  $\|AB\|_\mu \leq \|A\|_{\mu,\nu}\|B\|_\nu$ .

**Answer:**

$$\|AB\|_{\mu,\nu} = \max_{\|x\|_\nu=1} \|ABx\|_\mu \leq \max_{\|x\|_\nu=1} \|A\|_{\mu,\nu}\|Bx\|_\nu = \|A\|_{\mu,\nu} \max_{\|x\|_\nu=1} \|Bx\|_\nu = \|A\|_{\mu,\nu}\|B\|_\nu$$

🔍 BACK TO TEXT

**Homework 2.27** Show that the Frobenius norm,  $\|\cdot\|_F$ , is submultiplicative.

**Answer:**

$$\begin{aligned} \|AB\|_F^2 &= \left\| \begin{pmatrix} \hat{a}_0^H \\ \hat{a}_1^H \\ \vdots \\ \hat{a}_{m-1}^H \end{pmatrix} \begin{pmatrix} b_0 & b_1 & \cdots & b_{n-1} \end{pmatrix} \right\|_F^2 = \left\| \begin{pmatrix} \hat{a}_0^H b_0 & \hat{a}_0^H b_1 & \cdots & \hat{a}_0^H b_{n-1} \\ \hat{a}_1^H b_0 & \hat{a}_1^H b_1 & \cdots & \hat{a}_1^H b_{n-1} \\ \vdots & \vdots & & \vdots \\ \hat{a}_{m-1}^H b_0 & \hat{a}_{m-1}^H b_1 & \cdots & \hat{a}_{m-1}^H b_{n-1} \end{pmatrix} \right\|_F^2 = \sum_i \sum_j |\hat{a}_i^H b_j|^2 \\ &\leq \sum_i \sum_j \|\hat{a}_i^H\|_2^2 \|b_j\|^2 \quad (\text{Cauchy-Schwartz}) \\ &= \left( \sum_i \|\hat{a}_i\|_2^2 \right) \left( \sum_j \|b_j\|^2 \right) = \left( \sum_i \hat{a}_i^H \hat{a}_i \right) \left( \sum_j b_j^H b_j \right) \\ &\leq \left( \sum_i \sum_j |\hat{a}_i^H \hat{a}_j| \right) \left( \sum_i \sum_j |b_i^H b_j| \right) = \|A\|_F^2 \|B\|_F^2. \end{aligned}$$

so that  $\|AB\|_F^2 \leq \|A\|_F^2 \|B\|_F^2$ . Taking the square-root of both sides established the desired result.

🔍 BACK TO TEXT

**Homework 2.28** Let  $\|\cdot\|$  be a matrix norm induced by the  $\|\cdots\|$  vector norm. Show that  $\kappa(A) = \|A\|\|A^{-1}\| \geq 1$ .

**Answer:**

$$\|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|.$$

But

$$\|I\| = \max_{\|x\|=1} \|Ix\| = \max_{\|x\|=1} \|x\| = 1.$$

Hence  $1 \leq \|A\|\|A^{-1}\|$ .

[👉 BACK TO TEXT](#)

## Chapter 3. Notes on Orthogonality and the SVD (Answers)

**Homework 3.4** Let  $Q \in \mathbb{C}^{m \times n}$  (with  $n \leq m$ ). Partition  $Q = \left( q_0 \mid q_1 \mid \cdots \mid q_{n-1} \right)$ . Show that  $Q$  is an orthonormal matrix if and only if  $q_0, q_1, \dots, q_{n-1}$  are mutually orthonormal.

**Answer:** Answer needs to be filled in.

[👉 BACK TO TEXT](#)

**Homework 3.6** Let  $Q \in \mathbb{C}^{m \times m}$ . Show that if  $Q$  is unitary then  $Q^{-1} = Q^H$  and  $QQ^H = I$ .

**Answer:** If  $Q$  is unitary, then  $Q^H Q = I$ . If  $A, B \in \mathbb{C}^{m \times m}$ , the matrix  $B$  such that  $BA = I$  is the inverse of  $A$ . Hence  $Q^{-1} = Q^H$ . Also, if  $BA = I$  then  $AB = I$  and hence  $QQ^H = I$ .

[👉 BACK TO TEXT](#)

**Homework 3.7** Let  $Q_0, Q_1 \in \mathbb{C}^{m \times m}$  both be unitary. Show that their product,  $Q_0 Q_1$ , is unitary.

**Answer:** Obviously,  $Q_0 Q_1$  is a square matrix.

$$(Q_0 Q_1)^H (Q_0 Q_1) = Q_1^H \underbrace{Q_0^H Q_0}_I Q_1 = \underbrace{Q_1^H Q_1}_I = I.$$

Hence  $Q_0 Q_1$  is unitary.

[👉 BACK TO TEXT](#)

**Homework 3.8** Let  $Q_0, Q_1, \dots, Q_{k-1} \in \mathbb{C}^{m \times m}$  all be unitary. Show that their product,  $Q_0 Q_1 \cdots Q_{k-1}$ , is unitary.

**Answer:** Strictly speaking, we should do a proof by induction. But instead we will make the more informal argument that

$$\begin{aligned} (Q_0 Q_1 \cdots Q_{k-1})^H Q_0 Q_1 \cdots Q_{k-1} &= Q_{k-1}^H \cdots Q_1^H Q_0^H Q_0 Q_1 \cdots Q_{k-1} \\ &= Q_{k-1}^H \cdots Q_1^H \underbrace{Q_0^H Q_0}_I Q_1 \cdots Q_{k-1} = I. \end{aligned}$$

$\underbrace{\underbrace{\underbrace{I}_I}_I}_I$

[👉 BACK TO TEXT](#)

**Homework 3.9** Let  $U \in \mathbb{C}^{m \times m}$  be unitary and  $x \in \mathbb{C}^m$ , then  $\|Ux\|_2 = \|x\|_2$ .

**Answer:**  $\|Ux\|_2^2 = (Ux)^H(Ux) = x^H \underbrace{U^H U}_I x = x^H x = \|x\|_2^2$ . Hence  $\|Ux\|_2 = \|x\|_2$ .

🔗 BACK TO TEXT

**Homework 3.10** Let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary matrices and  $A \in \mathbb{C}^{m \times n}$ . Then

$$\|UA\|_2 = \|AV\|_2 = \|A\|_2.$$

**Answer:**

- $\|UA\|_2 = \max_{\|x\|_2=1} \|UAx\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \|A\|_2$ .
- $\|AV\|_2 = \max_{\|x\|_2=1} \|AVx\|_2 = \max_{\|Vx\|_2=1} \|A(Vx)\|_2 = \max_{\|y\|_2=1} \|Ay\|_2 = \|A\|_2$ .

🔗 BACK TO TEXT

**Homework 3.11** Let  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary matrices and  $A \in \mathbb{C}^{m \times n}$ . Then  $\|UA\|_F = \|AV\|_F = \|A\|_F$ .

**Answer:**

- Partition  $A = \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right)$ . Then it is easy to show that  $\|A\|_F^2 = \sum_{j=0}^{n-1} \|a_j\|_2^2$ . Thus

$$\begin{aligned} \|UA\|_F^2 &= \|U \left( a_0 \mid a_1 \mid \cdots \mid a_{n-1} \right)\|_F^2 = \left\| \left( Ua_0 \mid Ua_1 \mid \cdots \mid Ua_{n-1} \right) \right\|_F^2 \\ &= \sum_{j=0}^{n-1} \|Ua_j\|_2^2 = \sum_{j=0}^{n-1} \|a_j\|_2^2 = \|A\|_F^2. \end{aligned}$$

Hence  $\|UA\|_F = \|A\|_F$ .

- Partition  $A = \begin{pmatrix} \hat{a}_0^T \\ \hat{a}_1^T \\ \vdots \\ \hat{a}_{m-1}^T \end{pmatrix}$ . Then it is easy to show that  $\|A\|_F^2 = \sum_{i=0}^{m-1} \|(\hat{a}_i^T)^H\|_2^2$ . Thus

$$\begin{aligned} \|AV\|_F^2 &= \left\| \begin{pmatrix} \hat{a}_0^T \\ \hat{a}_1^T \\ \vdots \\ \hat{a}_{m-1}^T \end{pmatrix} V \right\|_F^2 = \left\| \begin{pmatrix} \hat{a}_0^T V \\ \hat{a}_1^T V \\ \vdots \\ \hat{a}_{m-1}^T V \end{pmatrix} \right\|_F^2 = \sum_{i=0}^{m-1} \|(\hat{a}_i^T V)^H\|_2^2 \\ &= \sum_{i=0}^{m-1} \|V^H (\hat{a}_i^T)^H\|_2^2 = \sum_{i=0}^{m-1} \|(\hat{a}_i^T)^H\|_2^2 = \|A\|_F^2. \end{aligned}$$

Hence  $\|AV\|_F = \|A\|_F$ .

**Homework 3.13** Let  $D = \text{diag}(\delta_0, \dots, \delta_{n-1})$ . Show that  $\|D\|_2 = \max_{i=0}^{n-1} |\delta_i|$ .

**Answer:**

$$\begin{aligned}
 \|D\|_2^2 &= \max_{\|x\|_2=1} \|Dx\|_2^2 = \max_{\|x\|_2=1} \left\| \begin{pmatrix} \delta_0 & 0 & \cdots & 0 \\ 0 & \delta_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_{n-1} \end{pmatrix} \begin{pmatrix} \chi_0 \\ \chi_1 \\ \vdots \\ \chi_{n-1} \end{pmatrix} \right\|_2^2 \\
 &= \max_{\|x\|_2=1} \left\| \begin{pmatrix} \delta_0 \chi_0 \\ \delta_1 \chi_1 \\ \vdots \\ \delta_{n-1} \chi_{n-1} \end{pmatrix} \right\|_2^2 = \max_{\|x\|_2=1} \left[ \sum_{j=0}^{n-1} |\delta_j \chi_j|^2 \right] = \max_{\|x\|_2=1} \left[ \sum_{j=0}^{n-1} [|\delta_j|^2 |\chi_j|^2] \right] \\
 &\leq \max_{\|x\|_2=1} \left[ \sum_{j=0}^{n-1} \left[ \max_{i=0}^{n-1} |\delta_i|^2 |\chi_j|^2 \right] \right] = \max_{\|x\|_2=1} \left[ \max_{i=0}^{n-1} |\delta_i|^2 \sum_{j=0}^{n-1} |\chi_j|^2 \right] = \left( \max_{i=0}^{n-1} |\delta_i| \right)^2 \max_{\|x\|_2=1} \|x\|_2^2 \\
 &= \left( \max_{i=0}^{n-1} |\delta_i| \right)^2.
 \end{aligned}$$

so that  $\|D\|_2 \leq \max_{i=0}^{n-1} |\delta_i|$ .

Also, choose  $j$  so that  $|\delta_j| = \max_{i=0}^{n-1} |\delta_i|$ . Then

$$\|D\|_2 = \max_{\|x\|_2=1} \|Dx\|_2 \geq \|De_j\|_2 = \|\delta_j e_j\|_2 = |\delta_j| \|e_j\|_2 = |\delta_j| = \max_{i=0}^{n-1} |\delta_i|.$$

so that  $\max_{i=0}^{n-1} |\delta_i| \leq \|D\|_2 \leq \max_{i=0}^{n-1} |\delta_i|$ , which implies that  $\|D\|_2 = \max_{i=0}^{n-1} |\delta_i|$ .

**Homework 3.14** Let  $A = \begin{pmatrix} A_T \\ 0 \end{pmatrix}$ . Use the SVD of  $A$  to show that  $\|A\|_2 = \|A_T\|_2$ .

**Answer:** Let  $A_T = U_T \Sigma_T V_T^H$  be the SVD of  $A$ . Then

$$A = \begin{pmatrix} A_T \\ 0 \end{pmatrix} = \begin{pmatrix} U_T \Sigma_T V_T^H \\ 0 \end{pmatrix} = \begin{pmatrix} U_T \\ 0 \end{pmatrix} \Sigma_T V_T^H,$$

which is the SVD of  $A$ . As a result, clearly the largest singular value of  $A_T$  equals the largest singular value of  $A$  and hence  $\|A\|_2 = \|A_T\|_2$ .



**Homework 3.15** Assume that  $U \in \mathbb{C}^{m \times m}$  and  $V \in \mathbb{C}^{n \times n}$  be unitary matrices. Let  $A, B \in \mathbb{C}^{m \times n}$  with  $B = UAV^H$ . Show that the singular values of  $A$  equal the singular values of  $B$ .

**Answer:** Let  $A = U_A \Sigma_A V_A^H$  be the SVD of  $A$ . Then  $B = U U_A \Sigma_A V_A^H V^H = (U U_A) \Sigma_A (V V_A)^H$  where both  $U U_A$  and  $V V_A$  are unitary. This gives us the SVD for  $B$  and it shows that the singular values of  $B$  equal the singular values of  $A$ .

🔗 BACK TO TEXT

**Homework 3.16** Let  $A \in \mathbb{C}^{m \times n}$  with  $A = \left( \begin{array}{c|c} \sigma_0 & 0 \\ \hline 0 & B \end{array} \right)$  and assume that  $\|A\|_2 = \sigma_0$ . Show that  $\|B\|_2 \leq \|A\|_2$ . (Hint: Use the SVD of  $B$ .)

**Answer:** Let  $B = U_B \Sigma_B V_B^H$  be the SVD of  $B$ . Then

$$A = \left( \begin{array}{c|c} \sigma_0 & 0 \\ \hline 0 & B \end{array} \right) = \left( \begin{array}{c|c} \sigma_0 & 0 \\ \hline 0 & U_B \Sigma_B V_B^H \end{array} \right) = \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & U_B \end{array} \right) \left( \begin{array}{c|c} \sigma_0 & 0 \\ \hline 0 & \Sigma_B \end{array} \right) \left( \begin{array}{c|c} 1 & 0 \\ \hline 0 & V_B \end{array} \right)^H$$

which shows the relationship between the SVD of  $A$  and  $B$ . Since  $\|A\|_2 = \sigma_0$ , it must be the case that the diagonal entries of  $\Sigma_B$  are less than or equal to  $\sigma_0$  in magnitude, which means that  $\|B\|_2 \leq \|A\|_2$ .

🔗 BACK TO TEXT

**Homework 3.17** Prove Lemma 3.12 for  $m \leq n$ .

**Answer:** You can use the following as an outline for your proof:

**Proof:** First, let us observe that if  $A = 0$  (the zero matrix) then the theorem trivially holds:  $A = UDV^H$  where  $U = I_{m \times m}$ ,  $V = I_{n \times n}$ , and  $D = \left( \begin{array}{c|c} & \\ \hline & 0 \end{array} \right)$ , so that  $D_{TL}$  is  $0 \times 0$ . Thus, w.l.o.g. assume that  $A \neq 0$ .

We will employ a proof by induction on  $m$ .

- **Base case:**  $m = 1$ . In this case  $A = \left( \begin{array}{c} \hat{a}_0^T \end{array} \right)$  where  $\hat{a}_0^T \in \mathbb{R}^{1 \times n}$  is its only row. By assumption,  $\hat{a}_0^T \neq 0$ . Then

$$A = \left( \begin{array}{c} \hat{a}_0^T \end{array} \right) = \left( \begin{array}{c} 1 \end{array} \right) (\|\hat{a}_0^T\|_2) \left( \begin{array}{c} v_0 \end{array} \right)^H$$

where  $v_0 = (\hat{a}_0^T)^H / \|\hat{a}_0^T\|_2$ . Choose  $V_1 \in \mathbb{C}^{n \times (n-1)}$  so that  $V = \left( \begin{array}{c|c} v_0 & V_1 \end{array} \right)$  is unitary. Then

$$A = \left( \begin{array}{c} \hat{a}_0^T \end{array} \right) = \left( \begin{array}{c} 1 \end{array} \right) \left( \begin{array}{c|c} \|\hat{a}_0^T\|_2 & 0 \end{array} \right) \left( \begin{array}{c|c} v_0 & V_1 \end{array} \right)^H = UDV^H$$

where  $D_{TL} = \left( \begin{array}{c} \delta_0 \end{array} \right) = \left( \begin{array}{c} \|\hat{a}_0^T\|_2 \end{array} \right)$  and  $U = \left( \begin{array}{c} 1 \end{array} \right)$ .

- **Inductive step:** Similarly modify the inductive step of the proof of the theorem.

- **By the Principle of Mathematical Induction** the result holds for all matrices  $A \in \mathbb{C}^{m \times n}$  with  $m \geq n$ .

[👉 BACK TO TEXT](#)

**Homework 3.35** Show that if  $A \in \mathbb{C}^{m \times m}$  is nonsingular, then

- $\|A\|_2 = \sigma_0$ , the largest singular value;
- $\|A^{-1}\|_2 = 1/\sigma_{m-1}$ , the inverse of the smallest singular value; and
- $\kappa_2(A) = \sigma_0/\sigma_{m-1}$ .

**Answer:** Answer needs to be filled in

[👉 BACK TO TEXT](#)

## Chapter 4. Notes on Gram-Schmidt QR Factorization (Answers)

**Homework 4.1** • What happens in the Gram-Schmidt algorithm if the columns of  $A$  are NOT linearly independent?

**Answer:** If  $a_j$  is the first column such that  $\{a_0, \dots, a_j\}$  are linearly dependent, then  $a_j^\perp$  will equal the zero vector and the process breaks down.

• How might one fix this?

**Answer:** When a vector with  $a_j^\perp$  is encountered, the columns can be rearranged so that that column (or those columns) come last.

• How can the Gram-Schmidt algorithm be used to identify which columns of  $A$  are linearly independent?

**Answer:** Again, if  $a_j^\perp = 0$  for some  $j$ , then the columns are linearly dependent.

[🔗 BACK TO TEXT](#)

**Homework 4.2 Homework 4.3** Convince yourself that the relation between the vectors  $\{a_j\}$  and  $\{q_j\}$  in the algorithms in Figure 4.2 is given by

$$\left( \begin{array}{c|c|c|c} a_0 & a_1 & \cdots & a_{n-1} \end{array} \right) = \left( \begin{array}{c|c|c|c} q_0 & q_1 & \cdots & q_{n-1} \end{array} \right) \left( \begin{array}{c|c|c|c} \rho_{0,0} & \rho_{0,1} & \cdots & \rho_{0,n-1} \\ \hline 0 & \rho_{1,1} & \cdots & \rho_{1,n-1} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & \rho_{n-1,n-1} \end{array} \right),$$

where

$$q_i^H q_j = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \rho_{i,j} = \begin{cases} q_i^H a_j & \text{for } i < j \\ \|a_j - \sum_{i=0}^{j-1} \rho_{i,j} q_i\|_2 & \text{for } i = j \\ 0 & \text{otherwise.} \end{cases}$$

**Answer:** Just watch the video for this lecture!

[🔗 BACK TO TEXT](#)

**Homework 4.5 Homework 4.6** Let  $A$  have linearly independent columns and let  $A = QR$  be a QR factorization of  $A$ . Partition

$$A \rightarrow \left( \begin{array}{c|c} A_L & A_R \end{array} \right), \quad Q \rightarrow \left( \begin{array}{c|c} Q_L & Q_R \end{array} \right), \quad \text{and} \quad R \rightarrow \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right),$$

where  $A_L$  and  $Q_L$  have  $k$  columns and  $R_{TL}$  is  $k \times k$ . Show that

1.  $A_L = Q_L R_{TL}$ :  $Q_L R_{TL}$  equals the QR factorization of  $A_L$ ,
2.  $C(A_L) = C(Q_L)$ : the first  $k$  columns of  $Q$  form an orthonormal basis for the space spanned by the first  $k$  columns of  $A$ .
3.  $R_{TR} = Q_L^H A_R$ ,
4.  $(A_R - Q_L R_{TR})^H Q_L = 0$ ,
5.  $A_R - Q_L R_{TR} = Q_R R_{BR}$ , and
6.  $C(A_R - Q_L R_{TR}) = C(Q_R)$ .

**Answer:** Consider the fact that  $A = QR$ . Then

$$\left( \begin{array}{c|c} A_L & A_R \end{array} \right) = \left( \begin{array}{c|c} Q_L & Q_R \end{array} \right) \left( \begin{array}{c|c} R_{TL} & R_{TR} \\ \hline 0 & R_{BR} \end{array} \right) = \left( \begin{array}{c|c} Q_L R_{TL} & Q_L R_{TR} + Q_R R_{BR} \end{array} \right).$$

Hence

$$A_L = Q_L R_{TL} \quad \text{and} \quad A_R = Q_L R_{TR} + Q_R R_{BR}.$$

- The left equation answers 1.
- Rearranging the right equation yields  $A_R - Q_L R_{TR} = Q_R R_{BR}$ , which answers 5.
- $C(A_L) = C(Q_L)$  can be shown by showing that  $C(A_L) \subset C(Q_L)$  and  $C(Q_L) \subset C(A_L)$ :

$C(A_L) \subset C(Q_L)$ : Let  $y \in C(A_L)$ . Then there exists  $x$  such that  $A_L x = y$ . But then  $Q_L R_{TL} x = y$  and hence  $Q_L (R_{TL} x) = y$  which means that  $y \in C(Q_L)$ .

$C(Q_L) \subset C(A_L)$ : Let  $y \in C(Q_L)$ . Then there exists  $x$  such that  $Q_L x = y$ . But then  $A_L R_{TL}^{-1} x = y$  and hence  $A_L (R_{TL}^{-1} x) = y$  which means that  $y \in C(A_L)$ . (Notice that  $R_{TL}$  is nonsingular because it is a triangular matrix that has only nonzeros on its diagonal.)

This answer 2.

- Take  $A_R - Q_L R_{TR} = Q_R R_{BR}$ . and multiply both side by  $Q_L^H$ :

$$Q_L^H (A_R - Q_L R_{TR}) = Q_L^H Q_R R_{BR}$$

is equivalent to

$$Q_L^H A_R - \underbrace{Q_L^H Q_L}_I R_{TR} = \underbrace{Q_L^H Q_R}_0 R_{BR} = 0.$$

Rearranging yields 3.

- Since  $A_R - Q_L R_{TR} = Q_R R_{BR}$  we find that  $(A_R - Q_L R_{TR})^H Q_L = (Q_R R_{BR})^H Q_L$  and

$$(A_R - Q_L R_{TR})^H Q_L = R_{BR}^H Q_R^H Q_L = 0.$$

- The proof of 6. follows similar to the proof of 2.

➡ BACK TO TEXT

## Chapter 6. Notes on Householder QR Factorization (Answers)

**Homework 6.2** Show that if  $H$  is a reflector, then

- $HH = I$  (reflecting the reflection of a vector results in the original vector),
- $H = H^H$ , and
- $H^H H = I$  (a reflection is a unitary matrix and thus preserves the norm).

**Answer:**

[👉 BACK TO TEXT](#)

**Homework 6.4** Show that if  $x \in \mathbb{R}^n$ ,  $v = x \mp \|x\|_2 e_0$ , and  $\tau = v^T v / 2$  then  $(I - \frac{1}{\tau} v v^T)x = \pm \|x\|_2 e_0$ .

[👉 BACK TO TEXT](#)

**Homework 6.5** Verify that

$$\left( I - \frac{1}{\tau} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \right) \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \rho \\ 0 \end{pmatrix}$$

where  $\tau = u^H u / 2 = (1 + u_2^H u_2) / 2$  and  $\rho = \pm \|x\|_2$ .

Hint:  $\rho \bar{\rho} = |\rho|^2 = \|x\|_2^2$  since  $H$  preserves the norm. Also,  $\|x\|_2^2 = |\chi_1|^2 + \|x_2\|_2^2$  and  $\sqrt{\frac{z}{\bar{z}}} = \frac{z}{|z|}$ .

**Answer:**

[👉 BACK TO TEXT](#)

**Homework 6.6** Show that

$$\left( \begin{array}{c|c} I & 0 \\ \hline 0 & \left( I - \frac{1}{\tau_1} \begin{pmatrix} 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 1 \\ u_2 \end{pmatrix}^H \right) \end{array} \right) = \left( I - \frac{1}{\tau_1} \begin{pmatrix} 0 \\ 1 \\ u_2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ u_2 \end{pmatrix}^H \right).$$

**Answer:**

[👉 BACK TO TEXT](#)

**Homework 6.11** If  $m = n$  then  $Q$  could be accumulated by the sequence

$$Q = (\cdots((IH_0)H_1)\cdots H_{n-1}).$$

Give a high-level reason why this would be (much) more expensive than the algorithm in Figure 6.6.

**Answer:**

[👉 BACK TO TEXT](#)

**Homework 6.12** Assuming all inverses exist, show that

$$\left( \begin{array}{c|c} T_{00} & t_{01} \\ \hline 0 & \tau_1 \end{array} \right)^{-1} = \left( \begin{array}{c|c} T_{00}^{-1} & -T_{00}^{-1}t_{01}/\tau_1 \\ \hline 0 & 1/\tau_1 \end{array} \right).$$

**Answer:**

$$\left( \begin{array}{c|c} T_{00} & t_{01} \\ \hline 0 & \tau_1 \end{array} \right) \left( \begin{array}{c|c} T_{00}^{-1} & -T_{00}^{-1}t_{01}/\tau_1 \\ \hline 0 & 1/\tau_1 \end{array} \right) = \left( \begin{array}{c|c} T_{00}T_{00}^{-1} & -T_{00}T_{00}^{-1}t_{01}/\tau_1 + \tau_1/\tau_1 \\ \hline 0 & \tau_1/\tau_1 \end{array} \right) = \left( \begin{array}{c|c} I & 0 \\ \hline 0 & 1 \end{array} \right).$$

[👉 BACK TO TEXT](#)

**Homework 6.13 Homework 6.14** Consider  $u_1 \in \mathbb{C}^m$  with  $u_1 \neq 0$  (the zero vector),  $U_0 \in \mathbb{C}^{m \times k}$ , and non-singular  $T_{00} \in \mathbb{C}^{k \times k}$ . Define  $\tau_1 = (u_1^H u_1)/2$ , so that

$$H_1 = I - \frac{1}{\tau_1} u_1 u_1^H$$

equals a Householder transformation, and let

$$Q_0 = I - U_0 T_{00}^{-1} U_0^H.$$

Show that

$$Q_0 H_1 = (I - U_0 T_{00}^{-1} U_0^H) \left( I - \frac{1}{\tau_1} u_1 u_1^H \right) = I - \left( \begin{array}{c|c} U_0 & u_1 \end{array} \right) \left( \begin{array}{c|c} T_{00} & t_{01} \\ \hline 0 & \tau_1 \end{array} \right)^{-1} \left( \begin{array}{c|c} U_0 & u_1 \end{array} \right)^H,$$

where  $t_{01} = Q_0^H u_1$ .

**Answer:**

$$\begin{aligned} Q_0 H_1 &= (I - U_0 T_{00}^{-1} U_0^H) \left( I - \frac{1}{\tau_1} u_1 u_1^H \right) \\ &= I - U_0 T_{00}^{-1} U_0^H - \frac{1}{\tau_1} u_1 u_1^H + U_0 T_{00}^{-1} U_0^H \frac{1}{\tau_1} u_1 u_1^H \\ &= I - U_0 T_{00}^{-1} U_0^H - \frac{1}{\tau_1} u_1 u_1^H + \frac{1}{\tau_1} U_0 T_{00}^{-1} t_{01} u_1^H \end{aligned}$$

Also

$$\begin{aligned}
& I - \left( U_0 \mid u_1 \right) \left( \begin{array}{c|c} T_{00} & t_{01} \\ \hline 0 & \tau_1 \end{array} \right)^{-1} \left( U_0 \mid u_1 \right)^H \\
&= I - \left( U_0 \mid u_1 \right) \left( \begin{array}{c|c} T_{00}^{-1} & -T_{00}^{-1}t_{01}/\tau_1 \\ \hline 0 & 1/\tau_1 \end{array} \right) \left( U_0 \mid u_1 \right)^H \\
&= I - \left( U_0 \mid u_1 \right) \left( \begin{array}{c} T_{00}^{-1}U_0^H - T_{00}^{-1}t_{01}u_1^H/\tau_1 \\ \hline 1/\tau_1 u_1^H \end{array} \right) \\
&= I - U_0(T_{00}^{-1}U_0^H - T_{00}^{-1}t_{01}u_1^H/\tau_1) - 1/\tau_1 u_1 u_1^H \\
&= I - U_0 T_{00}^{-1} U_0^H - \frac{1}{\tau_1} u_1 u_1^H + \frac{1}{\tau_1} U_0 T_{00}^{-1} t_{01} u_1^H
\end{aligned}$$

[👉 BACK TO TEXT](#)

**Homework 6.14 Homework 6.15** Consider  $u_i \in \mathbb{C}^m$  with  $u_i \neq 0$  (the zero vector). Define  $\tau_i = (u_i^H u_i)/2$ , so that

$$H_i = I - \frac{1}{\tau_i} u_i u_i^H$$

equals a Householder transformation, and let

$$U = \left( u_0 \mid u_1 \mid \cdots \mid u_{k-1} \right).$$

Show that

$$H_0 H_1 \cdots H_{k-1} = I - U T^{-1} U^H,$$

where  $T$  is an upper triangular matrix.

**Answer:** The results follows from Homework 6.14 via a proof by induction.

[👉 BACK TO TEXT](#)

## Chapter 7. Notes on Solving Linear Least-squares Problems (Answers)

**Homework 7.2** Let  $A \in \mathbb{C}^{m \times n}$  with  $m < n$  have linearly independent rows. Show that there exist a lower triangular matrix  $L_L \in \mathbb{C}^{m \times m}$  and a matrix  $Q_T \in \mathbb{C}^{m \times n}$  with orthonormal rows such that  $A = L_L Q_T$ , noting that  $L_L$  does not have any zeroes on the diagonal. Letting  $L = \left( L_L \mid 0 \right)$  be  $\mathbb{C}^{m \times n}$  and unitary  $Q = \begin{pmatrix} Q_T \\ Q_B \end{pmatrix}$ , reason that  $A = LQ$ .

Don't overthink the problem: use results you have seen before.

**Answer:** We know that  $A^H \in \mathbb{C}^{n \times m}$  with linearly independent columns (and  $n \leq m$ ). Hence there exists a unitary matrix  $\check{Q} = \text{FlaOneByTwo} Q_L Q_R$  and  $R = \begin{pmatrix} R_T \\ 0 \end{pmatrix}$  with upper triangular matrix  $R_T$  that has no zeroes on its diagonal such that  $A^H = \check{Q} R_T$ . It is easy to see that the desired  $Q_T$  equals  $Q_T = \check{Q}_L^H$  and the desired  $L_L$  equals  $R_T^H$ .

[BACK TO TEXT](#)

**Homework 7.3** Let  $A \in \mathbb{C}^{m \times n}$  with  $m < n$  have linearly independent rows. Consider

$$\|Ax - y\|_2 = \min_z \|Az - y\|_2.$$

Use the fact that  $A = L_L Q_T$ , where  $L_L \in \mathbb{C}^{m \times m}$  is lower triangular and  $Q_T$  has orthonormal rows, to argue that any vector of the form  $Q_T^H L_L^{-1} y + Q_B^H w_B$  (where  $w_B$  is any vector in  $\mathbb{C}^{n-m}$ ) is a solution to the LLS problem. Here  $Q = \begin{pmatrix} Q_T \\ Q_B \end{pmatrix}$ .

**Answer:**

$$\begin{aligned} \min_z \|Az - y\|_2 &= \min_z \|L_Q z - y\|_2 \\ &= \min_{\substack{w \\ z = Q^H w}} \|Lw - y\|_2 \\ &= \min_{\substack{w \\ z = Q^H w}} \left\| \begin{pmatrix} L_L & 0 \end{pmatrix} \begin{pmatrix} w_T \\ w_B \end{pmatrix} - y \right\|_2 \\ &= \min_{\substack{w \\ z = Q^H w}} \|L_L w_T - y\|_2. \end{aligned}$$

Hence  $w_T = L_L^{-1} y$  minimizes. But then

$$z = Q^H w = \begin{pmatrix} Q_T^H & Q_B^H \end{pmatrix} \begin{pmatrix} w_T \\ w_B \end{pmatrix} = Q_T^H w_T + Q_B^H w_B.$$



describes all solutions to the LLS problem.

[👉 BACK TO TEXT](#)

**Homework 7.4** Continuing Exercise 7.2, use Figure 7.1 to give a Classical Gram-Schmidt inspired algorithm for computing  $L_L$  and  $Q_T$ . (The best way to check you got the algorithm right is to implement it!)

**Answer:**

[👉 BACK TO TEXT](#)

**Homework 7.5** Continuing Exercise 7.2, use Figure 7.2 to give a Householder QR factorization inspired algorithm for computing  $L$  and  $Q$ , leaving  $L$  in the lower triangular part of  $A$  and  $Q$  stored as Householder vectors above the diagonal of  $A$ . (The best way to check you got the algorithm right is to implement it!)

**Answer:**

[👉 BACK TO TEXT](#)

## Chapter 8. Notes on the Condition of a Problem (Answers)

**Homework 8.1** Show that, if  $A$  is a nonsingular matrix, for a consistent matrix norm,  $\kappa(A) \geq 1$ .

**Answer:** Hmmm, I need to go and check what the exact definition of a consistent matrix norm is here... What I mean is a matrix norm induced by a vector norm. The reason is that then  $\|I\| = 1$ .

Let  $\|\cdot\|$  be the norm that is used to define  $\kappa(A) = \|AA^{-1}\|$ . Then

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \kappa(A).$$

👉 BACK TO TEXT

**Homework 8.2** If  $A$  has linearly independent columns, show that  $\|(A^H A)^{-1} A^H\|_2 = 1/\sigma_{n-1}$ , where  $\sigma_{n-1}$  equals the smallest singular value of  $A$ . Hint: Use the SVD of  $A$ .

**Answer:** Let  $A = U\Sigma V^H$  be the reduced SVD of  $A$ . Then

$$\begin{aligned} \|(A^H A)^{-1} A^H\|_2 &= \|((U\Sigma V^H)^H U\Sigma V^H)^{-1} (U\Sigma V^H)^H\|_2 \\ &= \|(V\Sigma U^H U\Sigma V^H)^{-1} V\Sigma U^H\|_2 \\ &= \|(V\Sigma^{-1}\Sigma^{-1}V^H)V\Sigma U^H\|_2 \\ &= \|V\Sigma^{-1}U^H\|_2 \\ &= \|\Sigma^{-1}\|_2 \\ &= 1/\sigma_{n-1} \end{aligned}$$

(since the two norm of a diagonal matrix equals its largest diagonal element in absolute value).

👉 BACK TO TEXT

**Homework 8.3** Let  $A$  have linearly independent columns. Show that  $\kappa_2(A^H A) = \kappa_2(A)^2$ .

**Answer:** Let  $A = U\Sigma V^H$  be the reduced SVD of  $A$ . Then

$$\begin{aligned} \kappa_2(A^H A) &= \|A^H A\|_2 \|(A^H A)^{-1}\|_2 \\ &= \|(U\Sigma V^H)^H U\Sigma V^H\|_2 \|((U\Sigma V^H)^H U\Sigma V^H)^{-1}\|_2 \\ &= \|V\Sigma^2 V^H\|_2 \|V(\Sigma^{-1})^2 V^H\|_2 \\ &= \|\Sigma^2\|_2 \|(\Sigma^{-1})^2\|_2 \\ &= \frac{\sigma_0^2}{\sigma_{n-1}^2} = \left(\frac{\sigma_0}{\sigma_{n-1}}\right)^2 = \kappa_2(A)^2. \end{aligned}$$

👉 BACK TO TEXT

**Homework 8.4** Let  $A \in \mathbb{C}^{n \times n}$  have linearly independent columns.

- Show that  $Ax = y$  if and only if  $A^H Ax = A^H y$ .

**Answer:** Since  $A$  has linearly independent columns and is square, we know that  $A^{-1}$  and  $A^{-H}$  exist. If  $Ax = y$ , then multiplying both sides by  $A^H$  yields  $A^H Ax = A^H y$ . If  $A^H Ax = A^H y$  then multiplying both sides by  $A^{-H}$  yields  $Ax = y$ .

- Reason that using the method of normal equations to solve  $Ax = y$  has a condition number of  $\kappa_2(A)^2$ .

[👉 BACK TO TEXT](#)

**Homework 8.5** Let  $U \in \mathbb{C}^{n \times n}$  be unitary. Show that  $\kappa_2(U) = 1$ .

**Answer:** The SVD of  $U$  is given by  $U = U\Sigma V^H$  where  $\Sigma = I$  and  $V = I$ . Hence  $\sigma_0 = \sigma_{n-1} = 1$  and  $\kappa_2(U) = \sigma_0/\sigma_{n-1} = 1$ .

[👉 BACK TO TEXT](#)

**Homework 8.6** Characterize the set of all square matrices  $A$  with  $\kappa_2(A) = 1$ .

**Answer:** If  $\kappa_2(A) = 1$  then  $\sigma_0/\sigma_{n-1} = 1$  which means that  $\sigma_0 = \sigma_{n-1}$  and hence  $\sigma_0 = \sigma_1 = \dots = \sigma_{n-1}$ . Hence the singular value decomposition of  $A$  must equal  $A = U\Sigma V^H = \sigma_0 UV^H$ . But  $UV^H$  is unitary since  $U$  and  $V^H$  are, and hence we conclude that  $A$  must be a nonzero multiple of a unitary matrix.

[👉 BACK TO TEXT](#)

## Chapter 9. Notes on the Stability of an Algorithm (Answers)

**Homework 9.3** Assume a floating point number system with  $\beta = 2$  and a mantissa with  $t$  digits so that a typical positive number is written as  $.d_0d_1 \dots d_{t-1} \times 2^e$ , with  $d_i \in \{0, 1\}$ .

- Write the number 1 as a floating point number.

**Answer:**  $. \underbrace{10 \dots 0}_t \times 2^1$ .  
digits

- What is the largest positive real number  $\mathbf{u}$  (represented as a binary fraction) such that the floating point representation of  $1 + \mathbf{u}$  equals the floating point representation of 1? (Assume rounded arithmetic.)

**Answer:**  $. \underbrace{10 \dots 0}_t \times 2^1 + . \underbrace{00 \dots 0}_t 1 \times 2^1 = . \underbrace{10 \dots 0}_t 1 \times 2^1$  which rounds to  $. \underbrace{10 \dots 0}_t \times 2^1 = 1$ .  
digits digits digits digits

It is not hard to see that any larger number rounds to  $. \underbrace{10 \dots 01}_t \times 2^1 > 1$ .  
digits

- Show that  $\mathbf{u} = \frac{1}{2}2^{1-t}$ .

**Answer:**  $. \underbrace{0 \dots 00}_t 1 \times 2^1 = .1 \times 2^{1-t} = \frac{1}{2} \times 2^{1-t}$ .  
digits

[👉 BACK TO TEXT](#)

**Homework 9.10** Prove Lemma 9.9.

**Answer:** Let  $C = AB$ . Then the  $(i, j)$  entry in  $|C|$  is given by

$$|\gamma_{i,j}| = \left| \sum_{p=0}^k \alpha_{i,p} \beta_{p,j} \right| \leq \sum_{p=0}^k |\alpha_{i,p} \beta_{p,j}| = \sum_{p=0}^k |\alpha_{i,p}| |\beta_{p,j}|$$

which equals the  $(i, j)$  entry of  $|A||B|$ . Thus  $|AB| \leq |A||B|$ .

[👉 BACK TO TEXT](#)

**Homework 9.12** Prove Theorem 9.11.

**Answer:**

Show that if  $|A| \leq |B|$  then  $\|A\|_1 \leq \|B\|_1$ :

Let

$$A = \left( a_0 \mid \cdots \mid a_{n-1} \right) \quad \text{and} \quad B = \left( b_0 \mid \cdots \mid b_{n-1} \right).$$

Then

$$\begin{aligned} \|A\|_1 &= \max_{0 \leq j < n} \|a_j\|_1 = \max_{0 \leq j < n} \left( \sum_{i=0}^{m-1} |\alpha_{i,j}| \right) \\ &= \left( \sum_{i=0}^{m-1} |\alpha_{i,k}| \right) \text{ where } k \text{ is the index that maximizes} \\ &\leq \left( \sum_{i=0}^{m-1} |\beta_{i,k}| \right) \text{ since } |A| \leq |B| \\ &\leq \max_{0 \leq j < n} \left( \sum_{i=0}^{m-1} |\beta_{i,j}| \right) = \max_{0 \leq j < n} \|b_j\|_1 = \|B\|_1. \end{aligned}$$

Show that if  $|A| \leq |B|$  then  $\|A\|_\infty \leq \|B\|_\infty$ :

Note:  $\|A\|_\infty = \|A^T\|_1$  and  $\|B\|_\infty = \|B^T\|_1$ . Also, if  $|A| \leq |B|$  then, clearly,  $|A^T| \leq |B^T|$ . Hence  $\|A\|_\infty = \|A^T\|_1 \leq \|B^T\|_1 = \|B\|_\infty$ .

Show that if  $|A| \leq |B|$  then  $\|A\|_F \leq \|B\|_F$ :

$$\|A\|_F^2 = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \alpha_{i,j}^2 \leq \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \beta_{i,j}^2 = \|B\|_F^2.$$

Hence  $\|A\|_F \leq \|B\|_F$ .

[👉 BACK TO TEXT](#)

**Homework 9.13** Repeat the above steps for the computation

$$\kappa := ((\chi_0 \Psi_0 + \chi_1 \Psi_1) + \chi_2 \Psi_2),$$

computing in the indicated order.

**Answer:**

[👉 BACK TO TEXT](#)

**Homework 9.15** Complete the proof of Lemma 9.14.

**Answer:** We merely need to fill in the details for Case 1 in the proof:

**Case 1:**  $\prod_{i=0}^n (1 + \varepsilon_i)^{\pm 1} = (\prod_{i=0}^{n-1} (1 + \varepsilon_i)^{\pm 1}) (1 + \varepsilon_n)$ . By the I.H. there exists a  $\theta_n$  such that  $(1 + \theta_n) = \prod_{i=0}^{n-1} (1 + \varepsilon_i)^{\pm 1}$  and  $|\theta_n| \leq n\mathbf{u}/(1 - n\mathbf{u})$ . Then

$$\left( \prod_{i=0}^{n-1} (1 + \varepsilon_i)^{\pm 1} \right) (1 + \varepsilon_n) = (1 + \theta_n)(1 + \varepsilon_n) = 1 + \underbrace{\theta_n + \varepsilon_n + \theta_n \varepsilon_n}_{\theta_{n+1}},$$

which tells us how to pick  $\theta_{n+1}$ . Now

$$\begin{aligned} |\theta_{n+1}| &= |\theta_n + \varepsilon_n + \theta_n \varepsilon_n| \leq |\theta_n| + |\varepsilon_n| + |\theta_n| |\varepsilon_n| \leq \frac{n\mathbf{u}}{1 - n\mathbf{u}} + \mathbf{u} + \frac{n\mathbf{u}}{1 - n\mathbf{u}} \mathbf{u} \\ &= \frac{n\mathbf{u} + \mathbf{u}(1 - n\mathbf{u}) + n\mathbf{u}^2}{1 - n\mathbf{u}} = \frac{(n+1) + \mathbf{u}}{1 - n\mathbf{u}} \leq \frac{(n+1) + \mathbf{u}}{1 - (n+1)\mathbf{u}}. \end{aligned}$$

👉 BACK TO TEXT

**Homework 9.18** Prove Lemma 9.17.

**Answer:**

$$\gamma_n = \frac{n\mathbf{u}}{1 - n\mathbf{u}} \leq \frac{(n+b)\mathbf{u}}{1 - n\mathbf{u}} \leq \frac{(n+b)\mathbf{u}}{1 - (n+b)\mathbf{u}} = \gamma_{n+b}.$$

$$\begin{aligned} \gamma_n + \gamma_b + \gamma_n \gamma_b &= \frac{n\mathbf{u}}{1 - n\mathbf{u}} + \frac{b\mathbf{u}}{1 - b\mathbf{u}} + \frac{n\mathbf{u}}{(1 - n\mathbf{u})(1 - b\mathbf{u})} \\ &= \frac{n\mathbf{u}(1 - b\mathbf{u}) + (1 - n\mathbf{u})b\mathbf{u} + bn\mathbf{u}^2}{(1 - n\mathbf{u})(1 - b\mathbf{u})} \\ &= \frac{n\mathbf{u} - bn\mathbf{u}^2 + b\mathbf{u} - bn\mathbf{u}^2 + bn\mathbf{u}^2}{1 - (n+b)\mathbf{u} + bn\mathbf{u}^2} = \frac{(n+b)\mathbf{u} - bn\mathbf{u}^2}{1 - (n+b)\mathbf{u} + bn\mathbf{u}^2} \\ &\leq \frac{(n+b)\mathbf{u}}{1 - (n+b)\mathbf{u} + bn\mathbf{u}^2} \leq \frac{(n+b)\mathbf{u}}{1 - (n+b)\mathbf{u}} = \gamma_{n+b}. \end{aligned}$$

👉 BACK TO TEXT

**Homework 9.19** Let  $k \geq 0$  and assume that  $|\varepsilon_1|, |\varepsilon_2| \leq \mathbf{u}$ , with  $\varepsilon_1 = 0$  if  $k = 0$ . Show that

$$\left( \frac{I + \Sigma^{(k)} \mid 0}{0 \mid (1 + \varepsilon_1)} \right) (1 + \varepsilon_2) = (I + \Sigma^{(k+1)}).$$

Hint: reason the case where  $k = 0$  separately from the case where  $k > 0$ .

**Answer:**

Case:  $k = 0$ . Then

$$\left( \frac{I + \Sigma^{(k)} \mid 0}{0 \mid (1 + \varepsilon_1)} \right) (1 + \varepsilon_2) = (1 + 0)(1 + \varepsilon_2) = (1 + \varepsilon_2) = (1 + \theta_1) = (I + \Sigma^{(1)}).$$

Case:  $k = 1$ . Then

$$\begin{aligned}
 \left( \begin{array}{c|c} I + \Sigma^{(k)} & 0 \\ \hline 0 & (1 + \varepsilon_1) \end{array} \right) (1 + \varepsilon_2) &= \left( \begin{array}{c|c} 1 + \theta_1 & 0 \\ \hline 0 & (1 + \varepsilon_1) \end{array} \right) (1 + \varepsilon_2) \\
 &= \left( \begin{array}{c|c} (1 + \theta_1)(1 + \varepsilon_2) & 0 \\ \hline 0 & (1 + \varepsilon_1)(1 + \varepsilon_2) \end{array} \right) \\
 &= \left( \begin{array}{c|c} (1 + \theta_2) & 0 \\ \hline 0 & (1 + \theta_2) \end{array} \right) = (I + \Sigma^{(2)}).
 \end{aligned}$$

Case:  $k > 1$ . Notice that

$$\begin{aligned}
 (I + \Sigma^{(k)})(1 + \varepsilon_2) &= \text{diag}((1 + \theta_k), (1 + \theta_k), (1 + \theta_{k-1}), \dots, (1 + \theta_2))(1 + \varepsilon_2) \\
 &= \text{diag}((1 + \theta_{k+1}), (1 + \theta_{k+1}), (1 + \theta_k), \dots, (1 + \theta_3))
 \end{aligned}$$

Then

$$\begin{aligned}
 \left( \begin{array}{c|c} I + \Sigma^{(k)} & 0 \\ \hline 0 & (1 + \varepsilon_1) \end{array} \right) (1 + \varepsilon_2) &= \left( \begin{array}{c|c} (I + \Sigma^{(k)})(1 + \varepsilon_2) & 0 \\ \hline 0 & (1 + \varepsilon_1)(1 + \varepsilon_2) \end{array} \right) \\
 &= \left( \begin{array}{c|c} (I + \Sigma^{(k)})(1 + \varepsilon_2) & 0 \\ \hline 0 & (1 + \theta_2) \end{array} \right) = (I + \Sigma^{(k+1)}).
 \end{aligned}$$

 BACK TO TEXT

**Homework 9.23** Prove R1-B.

**Answer:** From Theorem 9.20 we know that

$$\check{\mathbf{x}} = \mathbf{x}^T (I + \Sigma^{(n)}) \mathbf{y} = (\mathbf{x} + \frac{\Sigma^{(n)} \mathbf{x}}{\delta \mathbf{x}})^T \mathbf{y}.$$

Then

$$\begin{aligned}
 |\delta \mathbf{x}| &= |\Sigma^{(n)} \mathbf{x}| = \left| \begin{pmatrix} \theta_n \chi_0 \\ \theta_n \chi_1 \\ \theta_{n-1} \chi_2 \\ \vdots \\ \theta_2 \chi_{n-1} \end{pmatrix} \right| = \begin{pmatrix} |\theta_n \chi_0| \\ |\theta_n \chi_1| \\ |\theta_{n-1} \chi_2| \\ \vdots \\ |\theta_2 \chi_{n-1}| \end{pmatrix} = \begin{pmatrix} |\theta_n| |\chi_0| \\ |\theta_n| |\chi_1| \\ |\theta_{n-1}| |\chi_2| \\ \vdots \\ |\theta_2| |\chi_{n-1}| \end{pmatrix} \\
 &\leq \begin{pmatrix} |\theta_n| |\chi_0| \\ |\theta_n| |\chi_1| \\ |\theta_n| |\chi_2| \\ \vdots \\ |\theta_n| |\chi_{n-1}| \end{pmatrix} = |\theta_n| \begin{pmatrix} |\chi_0| \\ |\chi_1| \\ |\chi_2| \\ \vdots \\ |\chi_{n-1}| \end{pmatrix} \leq \gamma_n |\mathbf{x}|.
 \end{aligned}$$

(Note: strictly speaking, one should probably treat the case  $n = 1$  separately.)!.

 [BACK TO TEXT](#)

**Homework 9.25** In the above theorem, could one instead prove the result

$$\check{y} = A(x + \delta x),$$

where  $\delta x$  is “small”?

**Answer:** The answer is “no”. The reason is that for each individual element of  $y$

$$\check{\psi}_i = a_i^T(x + \delta x)$$

which would appear to support that

$$\begin{pmatrix} \check{\psi}_0 \\ \check{\psi}_1 \\ \vdots \\ \check{\psi}_{m-1} \end{pmatrix} = \begin{pmatrix} a_0^T(x + \delta x) \\ a_1^T(x + \delta x) \\ \vdots \\ a_{m-1}^T(x + \delta x) \end{pmatrix}.$$

However, the  $\delta x$  for each entry  $\check{\psi}_i$  is different, meaning that we cannot factor out  $x + \delta x$  to find that  $\check{y} = A(x + \delta x)$ .

 [BACK TO TEXT](#)

**Homework 9.27** In the above theorem, could one instead prove the result

$$\check{C} = (A + \Delta A)(B + \Delta B),$$

where  $\Delta A$  and  $\Delta B$  are “small”?

**Answer:** The answer is “no” for reasons similar to why the answer is “no” for Exercise 9.25.

 [BACK TO TEXT](#)



## **Chapter 10. Notes on Performance (Answers)**

No exercises to answer yet.

## Chapter 11. Notes on Gaussian Elimination and LU Factorization (Answers)

**Homework 11.6** Show that

$$\left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right)^{-1} = \left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right)$$

**Answer:**

$$\left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right) \left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right) = \left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} + l_{21} & I \end{array} \right) = \left( \begin{array}{c|c|c} I_k & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & 0 & I \end{array} \right)$$

[BACK TO TEXT](#)

**Homework 11.7** Let  $\tilde{L}_k = L_0 L_1 \dots L_k$ . Assume that  $\tilde{L}_k$  has the form  $\tilde{L}_{k-1} = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 1 & 0 \\ \hline L_{20} & 0 & I \end{array} \right)$ , where  $\tilde{L}_{00}$  is  $k \times k$ . Show that  $\tilde{L}_k$  is given by  $\tilde{L}_k = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 1 & 0 \\ \hline L_{20} & l_{21} & I \end{array} \right)$  .. (Recall:  $\hat{L}_k = \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right)$ .)

**Answer:**

$$\begin{aligned} \tilde{L}_k &= \tilde{L}_{k-1} L_k = \tilde{L}_{k-1} \hat{L}_k^{-1} \\ &= \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 1 & 0 \\ \hline L_{20} & 0 & I \end{array} \right) \cdot \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -l_{21} & I \end{array} \right)^{-1} = \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 1 & 0 \\ \hline L_{20} & 0 & I \end{array} \right) \cdot \left( \begin{array}{c|c|c} I & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & l_{21} & I \end{array} \right) \\ &= \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline l_{10}^T & 1 & 0 \\ \hline L_{20} & l_{21} & I \end{array} \right). \end{aligned}$$

[BACK TO TEXT](#)

```

function [ A_out ] = LU_unb_var5( A )

[ ATL, ATR, ...
  ABL, ABR ] = FLA_Part_2x2( A, ...
                             0, 0, 'FLA_TL' );

while ( size( ATL, 1 ) < size( A, 1 ) )

    [ A00,  a01,      A02,  ...
      a10t, alpha11, a12t, ...
      A20,  a21,      A22 ] = FLA_Repart_2x2_to_3x3( ATL, ATR, ...
                                                       ABL, ABR, ...
                                                       1, 1, 'FLA_BR' );

    %-----%

    a21 = a21/ alpha11;
    A22 = A22 - a21 * a12t;

    %-----%

    [ ATL, ATR, ...
      ABL, ABR ] = FLA_Cont_with_3x3_to_2x2( A00,  a01,      A02,  ...
                                              a10t, alpha11, a12t, ...
                                              A20,  a21,      A22, ...
                                              'FLA_TL' );

end

A_out = [ ATL, ATR
          ABL, ABR ];

return

```

Figure 11.3: Answer to Exercise 11.12.

**Homework 11.12** Implement LU factorization with partial pivoting with the FLAME@lab API, in M-script.

**Answer:** See Figure 11.3.

[🔗 BACK TO TEXT](#)

**Homework 11.13** Derive an algorithm for solving  $Ux = y$ , overwriting  $y$  with the solution, that casts most computation in terms of DOT products. Hint: Partition

$$U \rightarrow \left( \begin{array}{c|c} v_{11} & u_{12}^T \\ \hline 0 & U_{22} \end{array} \right).$$

Call this Variant 1 and use Figure 11.6 to state the algorithm.

**Algorithm:** Solve  $Uz = y$ , overwriting  $y$  (Variant 1)

**Partition**  $U \rightarrow \left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right), y \rightarrow \left( \begin{array}{c} y_T \\ y_B \end{array} \right)$

**where**  $U_{BR}$  is  $0 \times 0$ ,  $y_B$  has 0 rows

**while**  $m(U_{BR}) < m(U)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ y_B \end{array} \right) \rightarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ y_2 \end{array} \right)$$

---


$$\psi_1 := \psi_1 - u_{12}^T x_2$$

$$\psi_1 := \psi_1 / v_{11}$$


---

**Continue with**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ y_B \end{array} \right) \leftarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ y_2 \end{array} \right)$$

**endwhile**

**Algorithm:** Solve  $Uz = y$ , overwriting  $y$  (Variant 2)

**Partition**  $U \rightarrow \left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right), y \rightarrow \left( \begin{array}{c} y_T \\ y_B \end{array} \right)$

**where**  $U_{BR}$  is  $0 \times 0$ ,  $y_B$  has 0 rows

**while**  $m(U_{BR}) < m(U)$  **do**

**Repartition**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ y_B \end{array} \right) \rightarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ y_2 \end{array} \right)$$

---


$$\psi_1 := \chi_1 / v_{11}$$

$$y_0 := y_0 - \chi_1 u_{01}$$


---

**Continue with**

$$\left( \begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c|c|c} U_{00} & u_{01} & U_{02} \\ \hline u_{10}^T & v_{11} & u_{12}^T \\ \hline U_{20} & u_{21} & U_{22} \end{array} \right),$$

$$\left( \begin{array}{c} y_T \\ y_B \end{array} \right) \leftarrow \left( \begin{array}{c} y_0 \\ \hline \psi_1 \\ y_2 \end{array} \right)$$

**endwhile**

Figure 6 (Answer): Algorithms for the solution of upper triangular system  $Ux = y$  that overwrite  $y$  with  $x$ .

**Answer:**

Partition

$$\left( \begin{array}{c|c} v_{11} & u_{12}^T \\ \hline 0 & U_{22} \end{array} \right) \begin{pmatrix} \chi_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix}.$$

Multiplying this out yields

$$\begin{pmatrix} v_{11}\chi_1 + u_{12}^T x_2 \\ U_{22}x_2 \end{pmatrix} = \begin{pmatrix} \psi_1 \\ y_2 \end{pmatrix}.$$

So, if we assume that  $x_2$  has already been computed and has overwritten  $y_2$ , then  $\chi_1$  can be computed as

$$\chi_1 = (\psi_1 - u_{12}^T x_2) / v_{11}$$

which can then overwrite  $\psi_1$ . The resulting algorithm is given in Figure 11.6 (Answer) (left).

➡ BACK TO TEXT

**Homework 11.14** Derive an algorithm for solving  $Ux = y$ , overwriting  $y$  with the solution, that casts most computation in terms of AXPY operations. Call this Variant 2 and use Figure 11.6 to state the algorithm.

**Answer:** Partition

$$\left( \begin{array}{c|c} U_{00} & u_{01} \\ \hline 0 & v_{11} \end{array} \right) \begin{pmatrix} x_0 \\ \chi_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix}.$$

Multiplying this out yields

$$\begin{pmatrix} U_{00}x_0 + u_{01}\chi_1 \\ v_{11}\chi_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ \psi_1 \end{pmatrix}.$$

So,  $\chi_1 = \psi_1 / v_{11}$  after which  $x_0$  can be computed by solving  $U_{00}x_0 = y_0 - \chi_1 u_{01}$ . The resulting algorithm is given in Figure 11.6 (Answer) (right).

➡ BACK TO TEXT

**Homework 11.15** If  $A$  is an  $n \times n$  matrix, show that the cost of Variant 1 is approximately  $\frac{2}{3}n^3$  flops.

**Answer:** During the  $k$ th iteration,  $L_{00}$  is  $k \times k$ , for  $k = 0, \dots, n-1$ . Then the (approximate) cost of the steps are given by

- Solve  $L_{00}u_{01} = a_{01}$  for  $u_{01}$ , overwriting  $a_{01}$  with the result. Cost:  $k^2$  flops.
- Solve  $l_{10}^T U_{00} = a_{10}^T$  (or, equivalently,  $U_{00}^T (l_{10}^T)^T = (a_{10}^T)^T$  for  $l_{10}^T$ ), overwriting  $a_{10}^T$  with the result. Cost:  $k^2$  flops.
- Compute  $v_{11} = \alpha_{11} - l_{10}^T u_{01}$ , overwriting  $\alpha_{11}$  with the result. Cost:  $2k$  flops.

Thus, the total cost is given by

$$\sum_{k=0}^{n-1} (k^2 + k^2 + 2k) \approx 2 \sum_{k=0}^{n-1} k^2 \approx 2 \frac{1}{3} n^3 = \frac{2}{3} n^3.$$

➡ BACK TO TEXT

**Homework 11.16** Derive the up-looking variant for computing the LU factorization.

**Answer:** Consider the loop invariant:

$$\left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array} \right) = \left( \begin{array}{c|c} L \setminus U_{TL} & U_{TR} \\ \hline \hat{A}_{BL} & \hat{A}_{BR} \end{array} \right)$$

$$\wedge \begin{array}{c|c} L_{TL}U_{TL} = \hat{A}_{TL} & L_{TL}U_{TR} = \hat{A}_{TR} \\ \hline \cancel{L_{BL}U_{TL} = \hat{A}_{BL}} & \cancel{L_{BL}U_{TR} + L_{BR}U_{BR} = \hat{A}_{BR}} \end{array}$$

At the top of the loop, after repartitioning,  $A$  contains

$$\begin{array}{c|c|c} L \setminus U_{00} & u_{01} & U_{02} \\ \hline \hat{a}_{10}^T & \hat{\alpha}_{11} & \hat{a}_{12}^T \\ \hline \hat{A}_{20} & \hat{a}_{21} & \hat{A}_{22} \end{array}$$

while at the bottom it must contain

$$\begin{array}{c|c|c} L \setminus U_{00} & u_{01} & \hat{A}_{02} \\ \hline l_{10}^T & v_{11} & u_{12}^T \\ \hline \hat{A}_{20} & \hat{a}_{21} & \hat{A}_{22} \end{array}$$

where the entries in blue are to be computed. Now, considering  $LU = \hat{A}$  we notice that

$$\begin{array}{c|c|c} L_{00}U_{00} = \hat{A}_{00} & L_{00}u_{01} = \hat{a}_{01} & L_{00}U_{02} = \hat{A}_{02} \\ \hline l_{10}^T U_{00} = \hat{a}_{10}^T & l_{10}^T u_{01} + v_{11} = \hat{\alpha}_{11} & l_{10}^T U_{02} + u_{12}^T = \hat{a}_{12}^T \\ \hline L_{20}U_{00} = \hat{A}_{20} & L_{20}u_{01} + v_{11}l_{21} = \hat{a}_{21} & L_{20}U_{02} + l_{21}u_{12}^T + L_{22}U_{22} = \hat{A}_{22} \end{array}$$

The equalities in yellow can be used to compute the desired parts of  $L$  and  $U$ :

- Solve  $l_{10}^T U_{00} = \hat{a}_{10}^T$  for  $l_{10}^T$ , overwriting  $a_{10}^T$  with the result.
- Compute  $v_{11} = \hat{\alpha}_{11} - l_{10}^T u_{01}$ , overwriting  $\alpha_{11}$  with the result.
- Compute  $u_{12}^T := \hat{a}_{12}^T - l_{10}^T U_{02}$ , overwriting  $a_{12}^T$  with the result.

➡ BACK TO TEXT

**Homework 11.17** Implement all five LU factorization algorithms with the FLAME@lab API, in M-script.

➡ BACK TO TEXT

**Homework 11.18** Which of the five variants can be modified to incorporate partial pivoting?

**Answer:** Variants 2, 4, and 5.

[👉 BACK TO TEXT](#)

**Homework 11.23** Apply LU with partial pivoting to

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ -1 & -1 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & & \cdots & 1 & 1 \\ -1 & -1 & & \cdots & -1 & 1 \end{pmatrix}.$$

Pivot only when necessary.

**Answer:** Notice that no pivoting is necessary: Eliminating the entries below the diagonal in the first column yields:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 2 \\ 0 & -1 & 1 & \cdots & 0 & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & -1 & & \cdots & 1 & 2 \\ 0 & -1 & & \cdots & -1 & 2 \end{pmatrix}.$$

Eliminating the entries below the diagonal in the second column yields:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 2 \\ 0 & 0 & 1 & \cdots & 0 & 4 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & & \cdots & 1 & 4 \\ 0 & 0 & & \cdots & -1 & 4 \end{pmatrix}.$$

Eliminating the entries below the diagonal in the  $(n - 1)$ st column yields:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 2 \\ 0 & 0 & 1 & \cdots & 0 & 4 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & & \cdots & 1 & 2^{n-2} \\ 0 & 0 & & \cdots & 0 & 2^{n-1} \end{pmatrix}.$$

 [BACK TO TEXT](#)



## Chapter 12. Notes on Cholesky Factorization (Answers)

**Homework 12.3** Let  $B \in \mathbb{C}^{m \times n}$  have linearly independent columns. Prove that  $A = B^H B$  is HPD.

**Answer:** Let  $x \in \mathbb{C}^m$  be a nonzero vector. Then  $x^H B^H B x = (Bx)^H (Bx)$ . Since  $B$  has linearly independent columns we know that  $Bx \neq 0$ . Hence  $(Bx)^H Bx > 0$ .

[BACK TO TEXT](#)

**Homework 12.4** Let  $A \in \mathbb{C}^{m \times m}$  be HPD. Show that its diagonal elements are real and positive.

**Answer:** Let  $e_j$  be the  $j$ th unit basis vectors. Then  $0 < e_j^H A e_j = \alpha_{j,j}$ .

[BACK TO TEXT](#)

**Homework 12.14** Implement the Cholesky factorization with M-script.

[BACK TO TEXT](#)

**Homework 12.15** Consider  $B \in \mathbb{C}^{m \times n}$  with linearly independent columns. Recall that  $B$  has a QR factorization,  $B = QR$  where  $Q$  has orthonormal columns and  $R$  is an upper triangular matrix with positive diagonal elements. How are the Cholesky factorization of  $B^H B$  and the QR factorization of  $B$  related?

**Answer:**

$$B^H B = (QR)^H QR = R^H \underbrace{Q^H Q}_I R = \underbrace{R^H}_L \underbrace{R}_{L^H}.$$

[BACK TO TEXT](#)

**Homework 12.16** Let  $A$  be SPD and partition

$$A \rightarrow \left( \begin{array}{c|c} A_{00} & a_{10} \\ \hline a_{10}^T & \alpha_{11} \end{array} \right)$$

(Hint: For this exercise, use techniques similar to those in Section 12.4.)

1. Show that  $A_{00}$  is SPD.

**Answer:** Assume that  $A$  is  $n \times n$  so that  $A_{00}$  is  $(n-1) \times (n-1)$ . Let  $x_0 \in \mathbb{R}^{n-1}$  be a nonzero vector. Then

$$x_0^T A_{00} x_0 = \left( \begin{array}{c} x_0 \\ 0 \end{array} \right)^T \left( \begin{array}{c|c} A_{00} & a_{10} \\ \hline a_{10}^T & \alpha_{11} \end{array} \right) \left( \begin{array}{c} x_0 \\ 0 \end{array} \right) > 0$$

since  $\left( \begin{array}{c} x_0 \\ 0 \end{array} \right)^T$  is a nonzero vector and  $A$  is SPD.

2. Assuming that  $A_{00} = L_{00}L_{00}^T$ , where  $L_{00}$  is lower triangular and nonsingular, argue that the assignment  $l_{10}^T := a_{10}^T L_{00}^{-T}$  is well-defined.

**Answer:** The computation is well-defined because  $L_{00}^{-1}$  exists and hence  $l_{10}^T := a_{10}^T L_{00}^{-T}$  uniquely computes  $l_{10}^T$ .

3. Assuming that  $A_{00}$  is SPD,  $A_{00} = L_{00}L_{00}^T$  where  $L_{00}$  is lower triangular and nonsingular, and  $l_{10}^T = a_{10}^T L_{00}^{-T}$ , show that  $\alpha_{11} - l_{10}^T l_{10} > 0$  so that  $\lambda_{11} := \sqrt{\alpha_{11} - l_{10}^T l_{10}}$  is well-defined.

**Answer:** We want to show that  $\alpha_{11} - l_{10}^T l_{10} > 0$ . To do so, we are going to construct a nonzero vector  $x$  so that  $x^T A x = \alpha_{11} - l_{10}^T l_{10}$ , at which point we can invoke the fact that  $A$  is SPD.

Rather than just giving the answer, we go through the steps that will give insight into the thought process leading up to the answer as well. Consider

$$\begin{aligned} \begin{pmatrix} x_0 \\ \chi_1 \end{pmatrix}^T \begin{pmatrix} A_{00} & a_{10} \\ a_{10}^T & \alpha_{11} \end{pmatrix} \begin{pmatrix} x_0 \\ \chi_1 \end{pmatrix} &= \begin{pmatrix} x_0 \\ \chi_1 \end{pmatrix}^T \begin{pmatrix} A_{00}x_0 + \chi_1 a_{10} \\ a_{10}^T x_0 + \alpha_{11} \chi_1 \end{pmatrix} \\ &= x_0^T A_{00} x_0 + \chi_1 a_{10}^T x_0 + \chi_1 a_{10}^T x_0 + \alpha_{11} \chi_1^2 \\ &= x_0^T A_{00} x_0 + \chi_1 x_0^T a_{10} + \chi_1 a_{10}^T x_0 + \alpha_{11} \chi_1^2. \end{aligned}$$

Since we are trying to get to  $\alpha_{11} - l_{10}^T l_{10}$ , perhaps we should pick  $\chi_1 = 1$ . Then

$$\begin{pmatrix} x_0 \\ 1 \end{pmatrix}^T \begin{pmatrix} A_{00} & a_{10} \\ a_{10}^T & \alpha_{11} \end{pmatrix} \begin{pmatrix} x_0 \\ 1 \end{pmatrix} = x_0^T A_{00} x_0 + x_0^T a_{10} + a_{10}^T x_0 + \alpha_{11}.$$

The question now becomes how to pick  $x_0$  so that

$$x_0^T A_{00} x_0 + x_0^T a_{10} + a_{10}^T x_0 = -l_{10}^T l_{10}.$$

Let's try  $x_0 = -L_{00}^{-1} l_{10}$  and recall that  $l_{10}^T = a_{10}^T L_{00}^{-T}$  so that  $L_{00} l_{10} = a_{10}$ . Then

$$\begin{aligned} x_0^T A_{00} x_0 + x_0^T a_{10} + a_{10}^T x_0 &= x_0^T L_{00} L_{00}^T x_0 + x_0^T a_{10} + a_{10}^T x_0 \\ &= (-L_{00}^{-1} l_{10})^T L_{00} L_{00}^T (-L_{00}^{-1} l_{10}) + (-L_{00}^{-1} l_{10})^T (L_{00} l_{10}) + (L_{00} l_{10})^T (-L_{00}^{-1} l_{10}) \\ &= l_{10}^T L_{00}^{-T} L_{00} L_{00}^T L_{00}^{-1} l_{10} - l_{10}^T L_{00}^{-T} L_{00} l_{10} - l_{10}^T L_{00}^T L_{00}^{-1} l_{10} \\ &= -l_{10}^T l_{10}. \end{aligned}$$

We now put all these insights together:

$$\begin{aligned} 0 &< \begin{pmatrix} -L_{00}^{-1} l_{10} \\ 1 \end{pmatrix}^T \begin{pmatrix} A_{00} & a_{10} \\ a_{10}^T & \alpha_{11} \end{pmatrix} \begin{pmatrix} -L_{00}^{-1} l_{10} \\ 1 \end{pmatrix} \\ &= (-L_{00}^{-1} l_{10})^T L_{00} L_{00}^T (-L_{00}^{-1} l_{10}) + (-L_{00}^{-1} l_{10})^T (L_{00} l_{10}) + (L_{00} l_{10})^T (-L_{00}^{-1} l_{10}) + \alpha_{11} \\ &= \alpha_{11} - l_{10}^T l_{10}. \end{aligned}$$

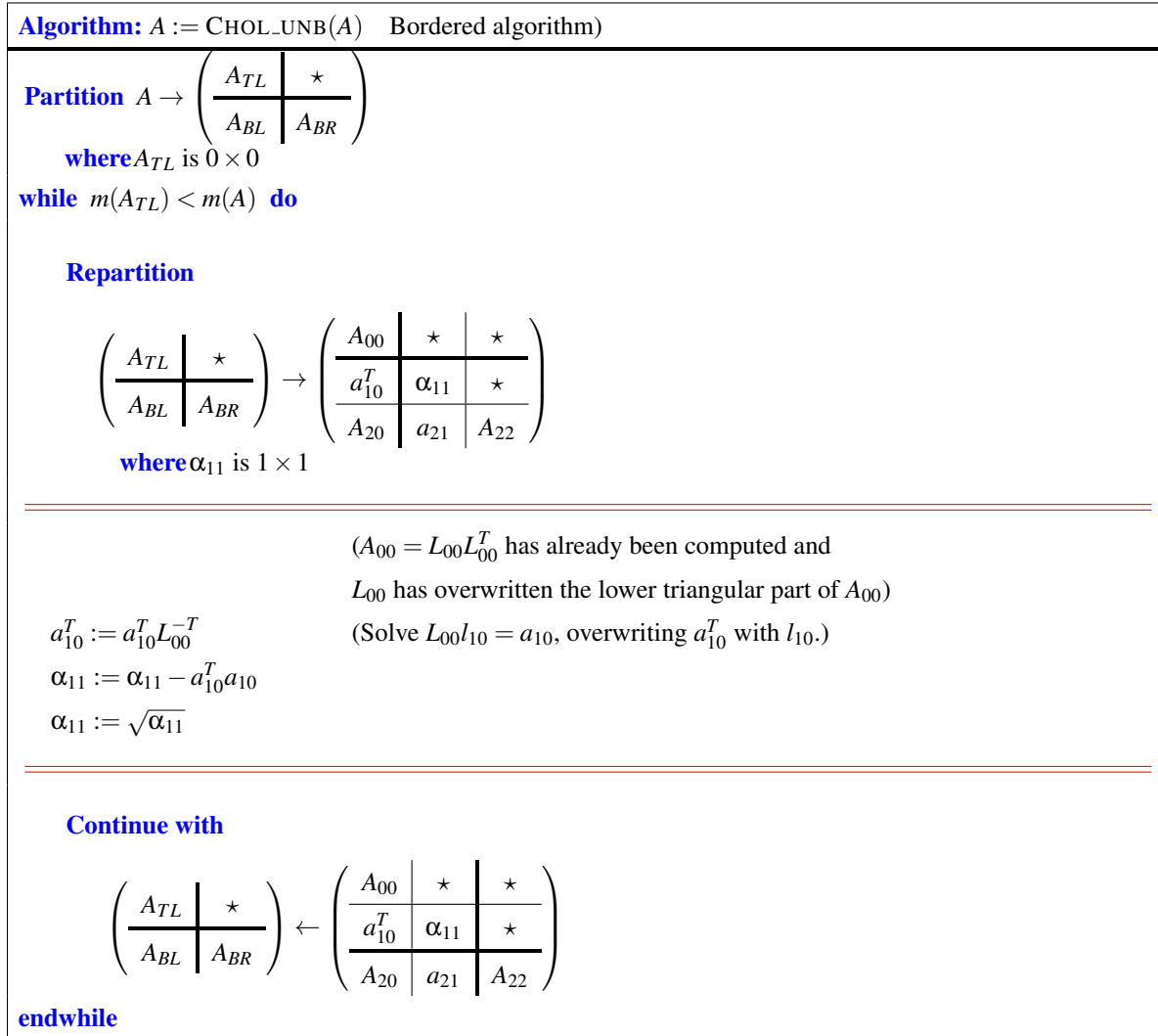


Figure 12.4: Answer for Homework 12.17.

4. Show that

$$\left( \begin{array}{c|c} A_{00} & a_{10} \\ \hline a_{10}^T & \alpha_{11} \end{array} \right) = \left( \begin{array}{c|c} L_{00} & 0 \\ \hline l_{10}^T & \lambda_{11} \end{array} \right) \left( \begin{array}{c|c} L_{00} & 0 \\ \hline l_{10}^T & \lambda_{11} \end{array} \right)^T$$

**Answer:** This is just a matter of multiplying it all out.

[🔙 BACK TO TEXT](#)

**Homework 12.17** Use the results in the last exercise to give an alternative proof by induction of the Cholesky Factorization Theorem and to give an algorithm for computing it by filling in Figure 12.3. This algorithm is often referred to as the *bordered* Cholesky factorization algorithm.

**Answer:**

Proof by induction.

**Base case:**  $n = 1$ . Clearly the result is true for a  $1 \times 1$  matrix  $A = \alpha_{11}$ : In this case, the fact that  $A$  is SPD means that  $\alpha_{11}$  is real and positive and a Cholesky factor is then given by  $\lambda_{11} = \sqrt{\alpha_{11}}$ , with uniqueness if we insist that  $\lambda_{11}$  is positive.

**Inductive step:** Inductive Hypothesis (I.H.): Assume the result is true for SPD matrix  $A \in \mathbb{C}^{(n-1) \times (n-1)}$ .

We will show that it holds for  $A \in \mathbb{C}^{n \times n}$ . Let  $A \in \mathbb{C}^{n \times n}$  be SPD. Partition  $A$  and  $L$  like

$$A = \left( \begin{array}{c|c} A_{00} & \star \\ \hline a_{10}^T & \alpha_{11} \end{array} \right) \quad \text{and} \quad L = \left( \begin{array}{c|c} L_{00} & 0 \\ \hline l_{10}^T & \lambda_{11} \end{array} \right).$$

Assume that  $A_{00} = L_{00}L_{00}^T$  is the Cholesky factorization of  $A_{00}$ . By the I.H., we know this exists since  $A_{00}$  is  $(n-1) \times (n-1)$  and that  $L_{00}$  is nonsingular. By the previous homework, we then know that

- $l_{10}^T = a_{10}^T L_{00}^{-T}$  is well-defined,
- $\lambda_{11} = \sqrt{\alpha_{11} - l_{10}^T l_{10}}$  is well-defined, and
- $A = LL^T$ .

Hence  $L$  is the desired Cholesky factor of  $A$ .

**By the principle of mathematical induction**, the theorem holds.

The algorithm is given in 12.4.

 [BACK TO TEXT](#)

**Homework 12.18** Show that the cost of the bordered algorithm is, approximately,  $\frac{1}{3}n^3$  flops.

**Answer:** During the  $k$ th iteration,  $k = 0, \dots, n-1$  assume that  $A_{00}$  is  $k \times k$ . Then

- $a_{10}^T = a_{10}^T L_{00}^{-T}$  (implemented as the lower triangular solve  $L_{00}l_{10} = a_{10}$ ) requires, approximately,  $k^2$  flops.
- The cost of update  $\alpha_{11}$  can be ignored.

Thus, the total cost equals (approximately)

$$\sum_{k=0}^{n-1} k^2 \approx \int_0^n x^2 dx = \frac{1}{3}n^3.$$

 [BACK TO TEXT](#)

## Chapter 13. Notes on Eigenvalues and Eigenvectors (Answers)

**Homework 13.3** Eigenvectors are not unique.

**Answer:** If  $Ax = \lambda x$  for  $x \neq 0$ , then  $A(\alpha x) = \alpha Ax = \alpha \lambda x = \lambda(\alpha x)$ . Hence any (nonzero) scalar multiple of  $x$  is also an eigenvector. This demonstrates we care about the *direction* of an eigenvector rather than its length.

[👉 BACK TO TEXT](#)

**Homework 13.4** Let  $\lambda$  be an eigenvalue of  $A$  and let  $\mathcal{E}_\lambda(A) = \{x \in \mathbb{C}^m | Ax = \lambda x\}$  denote the set of all eigenvectors of  $A$  associated with  $\lambda$  (including the zero vector, which is not really considered an eigenvector). Show that this set is a (nontrivial) subspace of  $\mathbb{C}^m$ .

**Answer:**

- $0 \in \mathcal{E}_\lambda(A)$ : (since we explicitly include it).
- $x \in \mathcal{E}_\lambda(A)$  implies  $\alpha x \in \mathcal{E}_\lambda(A)$ : (by the last exercise).
- $x, y \in \mathcal{E}_\lambda(A)$  implies  $x + y \in \mathcal{E}_\lambda(A)$ :

$$A(x + y) = Ax + Ay = \lambda x + \lambda y = \lambda(x + y).$$

[👉 BACK TO TEXT](#)

**Homework 13.8** The eigenvalues of a diagonal matrix equal the values on its diagonal. The eigenvalues of a triangular matrix equal the values on its diagonal.

**Answer:** Since a diagonal matrix is a special case of a triangular matrix, it suffices to prove that the eigenvalues of a triangular matrix are the values on its diagonal.

If  $A$  is triangular, so is  $A - \lambda I$ . By definition,  $\lambda$  is an eigenvalue of  $A$  if and only if  $A - \lambda I$  is singular. But a triangular matrix is singular if and only if it has a zero on its diagonal. The triangular matrix  $A - \lambda I$  has a zero on its diagonal if and only if  $\lambda$  equals one of the diagonal elements of  $A$ .

[👉 BACK TO TEXT](#)

**Homework 13.19** Prove Lemma 13.18. Then generalize it to a result for block upper triangular matrices:

$$A = \left( \begin{array}{c|c|c|c} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ \hline 0 & A_{1,1} & \cdots & A_{1,N-1} \\ \hline 0 & 0 & \ddots & \vdots \\ \hline 0 & 0 & \cdots & A_{N-1,N-1} \end{array} \right).$$

**Answer:**

**Lemma 13.18** Let  $A \in \mathbb{C}^{m \times m}$  be of form  $A = \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline 0 & A_{BR} \end{array} \right)$ . Assume that  $Q_{TL}$  and  $Q_{BR}$  are unitary “of appropriate size”. Then

$$A = \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right)^H \left( \begin{array}{c|c} Q_{TL}A_{TL}Q_{TL}^H & Q_{TL}A_{TR}Q_{BR}^H \\ \hline 0 & Q_{BR}A_{BR}Q_{BR}^H \end{array} \right) \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right).$$

**Proof:**

$$\begin{aligned} & \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right)^H \left( \begin{array}{c|c} Q_{TL}A_{TL}Q_{TL}^H & Q_{TL}A_{TR}Q_{BR}^H \\ \hline 0 & Q_{BR}A_{BR}Q_{BR}^H \end{array} \right) \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right) \\ &= \left( \begin{array}{c|c} Q_{TL}^H Q_{TL} A_{TL} Q_{TL}^H Q_{TL} & Q_{TL}^H Q_{TL} A_{TR} Q_{BR}^H Q_{BR} \\ \hline 0 & Q_{BR}^H Q_{BR} A_{BR} Q_{BR}^H Q_{BR} \end{array} \right) \\ &= \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline 0 & A_{BR} \end{array} \right) = A. \end{aligned}$$

By simple extension  $A = Q^H B Q$  where

$$Q = \left( \begin{array}{c|c|c|c} Q_{0,0} & 0 & \cdots & 0 \\ \hline 0 & Q_{1,1} & \cdots & 0 \\ \hline 0 & 0 & \ddots & \vdots \\ \hline 0 & 0 & \cdots & Q_{N-1,N-1} \end{array} \right)$$

and

$$B = \left( \begin{array}{c|c|c|c} Q_{0,0}A_{0,0}Q_{0,0}^H & Q_{0,0}A_{0,1}Q_{1,1}^H & \cdots & Q_{0,0}A_{0,N-1}Q_{N-1,N-1}^H \\ \hline 0 & Q_{1,1}A_{1,1}Q_{1,1}^H & \cdots & Q_{1,1}A_{1,N-1}Q_{N-1,N-1}^H \\ \hline 0 & 0 & \ddots & \vdots \\ \hline 0 & 0 & \cdots & Q_{N-1,N-1}A_{N-1,N-1}Q_{N-1,N-1}^H \end{array} \right).$$

➡ BACK TO TEXT

**Homework 13.21** Prove Corollary 13.20. Then generalize it to a result for block upper triangular matrices.

**Answer:**

**Colollary 13.20** Let  $A \in \mathbb{C}^{m \times m}$  be of for  $A = \left( \begin{array}{c|c} A_{TL} & A_{TR} \\ \hline 0 & A_{BR} \end{array} \right)$ . Then  $\Lambda(A) = \Lambda(A_{TL}) \cup \Lambda(A_{BR})$ .

**Proof:** This follows immediately from Lemma 13.18. Let  $A_{TL} = Q_{TL}T_{TL}Q_{TL}^H$  and  $A_{BR} = Q_{BR}T_{BR}Q_{BR}^H$  be the Shur decompositions of the diagonal blocks. Then  $\Lambda(A_{TL})$  equals the diagonal elements of  $T_{TL}$  and  $\Lambda(A_{BR})$  equals the diagonal elements of  $T_{BR}$ . Now, by Lemma 13.18 the Schur decomposition of  $A$  is given by

$$\begin{aligned} A &= \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right)^H \left( \begin{array}{c|c} Q_{TL}A_{TL}Q_{TL}^H & Q_{TL}A_{TR}Q_{BR}^H \\ \hline 0 & Q_{BR}A_{BR}Q_{BR}^H \end{array} \right) \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right) \\ &= \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right)^H \underbrace{\left( \begin{array}{c|c} T_{TL} & Q_{TL}A_{TR}Q_{BR}^H \\ \hline 0 & T_{BR} \end{array} \right)}_{T_A} \left( \begin{array}{c|c} Q_{TL} & 0 \\ \hline 0 & Q_{BR} \end{array} \right). \end{aligned}$$

Hence the diagonal elements of  $T_A$  (which is upper triangular) equal the elements of  $\Lambda(A)$ . The set of those elements is clearly  $\Lambda(A_{TL}) \cup \Lambda(A_{BR})$ .

The generalization is that if

$$A = \left( \begin{array}{c|c|c|c} A_{0,0} & A_{0,1} & \cdots & A_{0,N-1} \\ \hline 0 & A_{1,1} & \cdots & A_{1,N-1} \\ \hline 0 & 0 & \ddots & \vdots \\ \hline 0 & 0 & \cdots & A_{N-1,N-1} \end{array} \right)$$

where the blocks on the diagonal are square, then

$$\Lambda(A) = \Lambda(A_{0,0}) \cup \Lambda(A_{1,1}) \cup \cdots \cup \Lambda(A_{N-1,N-1}).$$

[➡ BACK TO TEXT](#)

**Homework 13.24** Let  $A$  be Hermitian and  $\lambda$  and  $\mu$  be distinct eigenvalues with eigenvectors  $x_\lambda$  and  $x_\mu$ , respectively. Then  $x_\lambda^H x_\mu = 0$ . (In other words, the eigenvectors of a Hermitian matrix corresponding to distinct eigenvalues are orthogonal.)

**Answer:** Assume that  $Ax_\lambda = \lambda x_\lambda$  and  $Ax_\mu = \mu x_\mu$ , for nonzero vectors  $x_\lambda$  and  $x_\mu$  and  $\lambda \neq \mu$ . Then

$$x_\mu^H Ax_\lambda = \lambda x_\mu^H x_\lambda$$

and

$$x_\lambda^H Ax_\mu = \mu x_\lambda^H x_\mu.$$

Because  $A$  is Hermitian and  $\lambda$  is real

$$\mu x_\lambda^H x_\mu = x_\lambda^H Ax_\mu = (x_\mu^H Ax_\lambda)^H = (\lambda x_\mu^H x_\lambda)^H = \lambda x_\lambda^H x_\mu.$$

Hence

$$\mu x_\lambda^H x_\mu = \lambda x_\lambda^H x_\mu.$$

If  $x_\lambda^H x_\mu \neq 0$  then  $\mu = \lambda$ , which is a contradiction.

[➡ BACK TO TEXT](#)

**Homework 13.25** Let  $A \in \mathbb{C}^{m \times m}$  be a Hermitian matrix,  $A = Q\Lambda Q^H$  its Spectral Decomposition, and  $A = U\Sigma V^H$  its SVD. Relate  $Q$ ,  $U$ ,  $V$ ,  $\Lambda$ , and  $\Sigma$ .

**Answer:** I am going to answer this by showing how to take the Spectral decomposition of  $A$ , and turn this into the SVD of  $A$ . Observations:

- We will assume all eigenvalues are nonzero. It should be pretty obvious how to fix the below if some of them are zero.
- $Q$  is unitary and  $\Lambda$  is diagonal. Thus,  $A = Q\Lambda Q^H$  is the SVD *except* that the diagonal elements of  $\Lambda$  are not necessarily nonnegative *and* they are not ordered from largest to smallest in magnitude.
- We can fix the fact that they are not nonnegative with the following observation:

$$\begin{aligned}
 A &= Q\Lambda Q^H = Q \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} \end{pmatrix} Q^H \\
 &= Q \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} \end{pmatrix} \underbrace{\begin{pmatrix} \text{sign}(\lambda_0) & 0 & \cdots & 0 \\ 0 & \text{sign}(\lambda_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{sign}(\lambda_{n-1}) \end{pmatrix}}_I \begin{pmatrix} \text{sign}(\lambda_0) & 0 & \cdots & 0 \\ 0 & \text{sign}(\lambda_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{sign}(\lambda_{n-1}) \end{pmatrix} Q^H \\
 &= Q \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} \end{pmatrix} \underbrace{\begin{pmatrix} \text{sign}(\lambda_0) & 0 & \cdots & 0 \\ 0 & \text{sign}(\lambda_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{sign}(\lambda_{n-1}) \end{pmatrix}}_{\tilde{\Lambda}} \underbrace{\begin{pmatrix} \text{sign}(\lambda_0) & 0 & \cdots & 0 \\ 0 & \text{sign}(\lambda_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{sign}(\lambda_{n-1}) \end{pmatrix}}_{\tilde{Q}^H} Q^H \\
 &= Q\tilde{\Lambda}\tilde{Q}^H.
 \end{aligned}$$

Now  $\tilde{\Lambda}$  has nonnegative diagonal entries and  $\tilde{Q}$  is unitary since it is the product of two unitary matrices.

- Next, we fix the fact that the entries of  $\tilde{\Lambda}$  are not ordered from largest to smallest. We do so by noting that there is a permutation matrix  $P$  such that  $\Sigma = P\tilde{\Lambda}P^H$  equals the diagonal matrix with the values ordered from the largest to smallest. Then

$$A = Q\tilde{\Lambda}\tilde{Q}^H$$



$$\begin{aligned}
&= Q \underbrace{P^H P}_I \tilde{\Lambda} \underbrace{P^H P}_I \tilde{Q}^H \\
&= \underbrace{QP^H}_U \underbrace{P\tilde{\Lambda}P^H}_\Sigma \underbrace{P\tilde{Q}^H}_{V^H} \\
&= U\Sigma V^H.
\end{aligned}$$

[👉 BACK TO TEXT](#)

**Homework 13.26** Let  $A \in \mathbb{C}^{m \times m}$  and  $A = U\Sigma V^H$  its SVD. Relate the Spectral decompositions of  $A^H A$  and  $AA^H$  to  $U$ ,  $V$ , and  $\Sigma$ .

**Answer:**

$$A^H A = (U\Sigma V^H)^H U\Sigma V^H = V\Sigma U^H U\Sigma V^H = V\Sigma^2 V^H.$$

Thus, the eigenvalues of  $A^H A$  equal the square of the singular values of  $A$ .

$$AA^H = U\Sigma V^H (U\Sigma V^H)^H = U\Sigma V^H V\Sigma U^H = U\Sigma^2 U^H.$$

Thus, the eigenvalues of  $AA^H$  equal the square of the singular values of  $A$ .

[👉 BACK TO TEXT](#)

## Chapter 14. Notes on the Power Method and Related Methods (Answers)

**Homework 14.4** Prove Lemma 14.3.

**Answer:** We need to show that

- Let  $y \neq 0$ . Show that  $\|y\|_X > 0$ : Let  $z = Xy$ . Then  $z \neq 0$  since  $X$  is nonsingular. Hence

$$\|y\|_X = \|Xy\| = \|z\| > 0.$$

- Show that if  $\alpha \in \mathbb{C}$  and  $y \in \mathbb{C}^m$  then  $\|\alpha y\|_X = |\alpha| \|y\|_X$ :

$$\|\alpha y\|_X = \|X(\alpha y)\| = \|\alpha Xy\| = |\alpha| \|Xy\| = |\alpha| \|y\|_X.$$

- Show that if  $x, y \in \mathbb{C}^m$  then  $\|x + y\|_X \leq \|x\|_X + \|y\|_X$ :

$$\|x + y\|_X = \|X(x + y)\| = \|Xx + Xy\| \leq \|Xx\| + \|Xy\| = \|x\|_X + \|y\|_X.$$

[👉 BACK TO TEXT](#)

**Homework 14.7** Assume that

$$|\lambda_0| \geq |\lambda_1| \geq \cdots \geq |\lambda_{m-2}| > |\lambda_{m-1}| > 0.$$

Show that

$$\left| \frac{1}{\lambda_{m-1}} \right| > \left| \frac{1}{\lambda_{m-2}} \right| \geq \left| \frac{1}{\lambda_{m-3}} \right| \geq \cdots \geq \left| \frac{1}{\lambda_0} \right|.$$

**Answer:** This follows immediately from the fact that if  $\alpha > 0$  and  $\beta > 0$  then

- $\alpha > \beta$  implies that  $1/\beta > 1/\alpha$  and
- $\alpha \geq \beta$  implies that  $1/\beta \geq 1/\alpha$ .

[👉 BACK TO TEXT](#)

**Homework 14.9** Prove Lemma 14.8.

**Answer:** If  $Ax = \lambda x$  then  $(A - \mu I)x = Ax - \mu x = \lambda x - \mu x = (\lambda - \mu)x$ .

[👉 BACK TO TEXT](#)

**Homework 14.11** Prove Lemma 14.10.

**Answer:**

$$A - \mu I = X\Lambda X^{-1} - \mu XX^{-1} = X(\Lambda - \mu I)X^{-1}.$$

[👉 BACK TO TEXT](#)

## Chapter 15. Notes on the QR Algorithm and other Dense Eigen-solvers(Answers)

**Homework 15.2** Prove that in Figure 15.3,  $\widehat{V}^{(k)} = V^{(k)}$ , and  $\widehat{A}^{(k)} = A^{(k)}$ ,  $k = 0, \dots$

**Answer:** This requires a proof by induction.

- Base case:  $k = 0$ . We need to show that  $\widehat{V}^{(0)} = V^{(0)}$  and  $\widehat{A}^{(0)} = A^{(0)}$ . This is trivially true.
- Inductive step: Inductive Hypothesis (IH):  $\widehat{V}^{(k)} = V^{(k)}$  and  $\widehat{A}^{(k)} = A^{(k)}$ .

We need to show that  $\widehat{V}^{(k+1)} = V^{(k+1)}$  and  $\widehat{A}^{(k+1)} = A^{(k+1)}$ . Notice that

$$A\widehat{V}^{(k)} = \widehat{V}^{(k+1)}\widehat{R}^{(k+1)} \text{ (QR factorization)}$$

so that

$$\widehat{A}^{(k)} = \widehat{V}^{(k)T}A\widehat{V}^{(k)} = \widehat{V}^{(k)T}\widehat{V}^{(k+1)}\widehat{R}^{(k+1)}$$

Since  $\widehat{V}^{(k)T}\widehat{V}^{(k+1)}$  this means that  $\widehat{A}^{(k)} = \widehat{V}^{(k)T}\widehat{V}^{(k+1)}\widehat{R}^{(k+1)}$  is a QR factorization of  $\widehat{A}^{(k)}$ .

Now, by the I.H.,

$$A^{(k)} = \widehat{A}^{(k)}\widehat{V}^{(k)T}\widehat{V}^{(k+1)}\widehat{R}^{(k+1)}$$

and in the algorithm on the right

$$A^{(k)} = Q^{(k+1)}R^{(k+1)}.$$

By the uniqueness of the QR factorization, this means that

$$Q^{(k+1)} = \widehat{V}^{(k)T}\widehat{V}^{(k+1)}.$$

But then

$$V^{(k+1)} = V^{(k)}Q^{(k+1)} = \widehat{V}^{(k)}Q^{(k+1)} = \underbrace{\widehat{V}^{(k)}\widehat{V}^{(k)T}}_I \widehat{V}^{(k+1)} = \widehat{V}^{(k+1)}.$$

Finally,

$$\begin{aligned} \widehat{A}^{(k+1)} &= \widehat{V}^{(k+1)T}A\widehat{V}^{(k+1)} = V^{(k+1)T}AV^{(k+1)} = (V^{(k)}Q^{(k+1)})^T A(V^{(k)}Q^{(k+1)}) \\ &= Q^{(k+1)T}V^{(k)T}AV^{(k)}Q^{(k+1)} = \underbrace{Q^{(k+1)T}A^{(k)}}_{R^{(k+1)}} Q^{(k+1)} = R^{(k+1)}Q^{(k+1)} = A^{(k+1)} \end{aligned}$$

- By the Principle of Mathematical Induction the result holds for all  $k$ .

➡ BACK TO TEXT

**Homework 15.3** Prove that in Figure 15.4,  $\widehat{V}^{(k)} = V^{(k)}$ , and  $\widehat{A}^{(k)} = A^{(k)}$ ,  $k = 1, \dots$

**Answer:** This requires a proof by induction.

- Base case:  $k = 0$ . We need to show that  $\widehat{V}^{(0)} = V^{(0)}$  and  $\widehat{A}^{(0)} = A^{(0)}$ . This is trivially true.
- Inductive step: Inductive Hypothesis (IH):  $\widehat{V}^{(k)} = V^{(k)}$  and  $\widehat{A}^{(k)} = A^{(k)}$ .

We need to show that  $\widehat{V}^{(k+1)} = V^{(k+1)}$  and  $\widehat{A}^{(k+1)} = A^{(k+1)}$ .

The proof of the inductive step is a minor modification of the last proof *except* that we need to show that  $\widehat{\mu}_k = \mu_k$ :

$$\begin{aligned}\widehat{\mu}_k &= \widehat{v}_{n-1}^{(k)T} A \widehat{v}_{n-1}^{(k)} (V^{(k)} e_{n-1})^T A (V^{(k)} e_{n-1}) = e_{n-1} V^{(k)T} A V^{(k)} e_{n-1} \\ &= \underbrace{e_{n-1} \widehat{A}^{(k)} e_{n-1} = e_{n-1} A^{(k)} e_{n-1}}_{\text{by I.H.}} = \alpha_{n-1, n-1}^{(k)} = \mu_k\end{aligned}$$

- By the Principle of Mathematical Induction the result holds for all  $k$ .

➡ BACK TO TEXT

**Homework 15.5** Prove the above theorem.

**Answer:** I believe we already proved this in “Notes on Eigenvalues and Eigenvectors”.

➡ BACK TO TEXT

**Homework 15.9** Give all the details of the above proof.

**Answer:** Assume that  $q_1, \dots, q_k$  and the column indexed with  $k-1$  of  $B$  have been shown to be uniquely determined under the stated assumptions. We now show that then  $q_{k+1}$  and the column indexed by  $k$  of  $B$  are uniquely determined. (This is the inductive step in the proof.) Then

$$Aq_k = \beta_{0,k}q_0 + \beta_{1,k}q_1 + \dots + \beta_{k,k}q_k + \beta_{k+1,k}q_{k+1}.$$

We can determine  $\beta_{0,k}$  through  $\beta_{k,k}$  by observing that

$$q_j^T Aq_k = \beta_{j,k}$$

for  $j = 0, \dots, k$ . Then

$$\beta_{k+1,k}q_{k+1} = Aq_k - (\beta_{0,k}q_0 + \beta_{1,k}q_1 + \dots + \beta_{k,k}q_k) = \tilde{q}_{k+1}.$$

Since it is assumed that  $\beta_{k+1,k} > 0$ , it can be determined as

$$\beta_{k+1,k} = \|\tilde{q}_{k+1}\|_2$$

and then

$$q_{k+1} = \tilde{q}_{k+1} / \beta_{k+1,k}.$$

This way, the columns of  $Q$  and  $B$  can be determined, one-by-one.

➡ BACK TO TEXT

**Homework 15.11** In the above discussion, show that  $\alpha_{11}^2 + 2\alpha_{31}^2 + \alpha_{33}^2 = \hat{\alpha}_{11}^2 + \hat{\alpha}_{33}^2$ .

**Answer:** If  $\hat{A}_{31} = J_{31}A_{31}J_{31}^T$  then  $\|\hat{A}_{31}\|_F^2 = \|A_{31}\|_F^2$  because multiplying on the left and/or the right by a unitary matrix preserves the Frobenius norm of a matrix. Hence

$$\left\| \begin{pmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{31} & \alpha_{33} \end{pmatrix} \right\|_F^2 = \left\| \begin{pmatrix} \hat{\alpha}_{11} & 0 \\ 0 & \hat{\alpha}_{33} \end{pmatrix} \right\|_F^2.$$

But

$$\left\| \begin{pmatrix} \alpha_{11} & \alpha_{31} \\ \alpha_{31} & \alpha_{33} \end{pmatrix} \right\|_F^2 = \alpha_{11}^2 + 2\alpha_{31}^2 + \alpha_{33}^2$$

and

$$\left\| \begin{pmatrix} \hat{\alpha}_{11} & 0 \\ 0 & \hat{\alpha}_{33} \end{pmatrix} \right\|_F^2 = \hat{\alpha}_{11}^2 + \hat{\alpha}_{33}^2,$$

which proves the result.

[👉 BACK TO TEXT](#)

**Homework 15.14** Give the approximate total cost for reducing a nonsymmetric  $A \in \mathbb{R}^{n \times n}$  to upper-Hessenberg form.

**Answer:** To prove this, assume that in the current iteration of the algorithm  $A_{00}$  is  $k \times k$ . The update in the current iteration is then

$[u_{21}, \tau_1, a_{21}] := \text{HouseV}(a_{21})$	lower order cost, ignore
$y_{01} := A_{02}u_{21}$	$2k(n-k-1)$ flops since $A_{02}$ is $k \times (n-k-1)$ .
$A_{02} := A_{02} - \frac{1}{\tau}y_{01}u_{21}^T$	$2k(n-k-1)$ flops since $A_{02}$ is $k \times (n-k-1)$ .
$\psi_{11} := a_{12}^T u_{21}$	lower order cost, ignore
$a_{12}^T := a_{12}^T + \frac{\psi_{11}}{\tau}u_{21}^T$	lower order cost, ignore
$y_{21} := A_{22}u_{21}$	$2(n-k-1)^2$ flops since $A_{22}$ is $(n-k-1) \times (n-k-1)$ .
$\beta := u_{21}^T y_{21} / 2$	lower order cost, ignore
$z_{21} := (y_{21} - \beta u_{21} / \tau) / \tau_1$	lower order cost, ignore
$w_{21} := (A_{22}^T u_{21} - \beta u_{21} / \tau) / \tau$	$2(n-k-1)^2$ flops since $A_{22}$ is $(n-k-1) \times (n-k-1)$ and the significant cost is in the matrix-vector multiply.
$A_{22} := A_{22} - (u_{21}w_{21}^T + z_{21}u_{21}^T)$	$2 \times 2k(n-k-1)$ flops since $A_{22}$ is $(n-k-1) \times (n-k-1)$ , which is updated by two rank-1 updates.

Thus, the total cost in flops is, approximately,

$$\begin{aligned}
& \sum_{k=0}^{n-1} [2k(n-k-1) + 2k(n-k-1) + 2(n-k-1)^2 + 2(n-k-1)^2 + 2 \times 2(n-k-1)^2] \\
&= \sum_{k=0}^{n-1} [4k(n-k-1) + 4(n-k-1)^2] + \sum_{k=0}^{n-1} 4(n-k-1)^2 \\
&= \sum_{k=0}^{n-1} \underbrace{[4k(n-k-1) + 4(n-k-1)^2]}_{= 4(k+n-k-1)(n-k-1)} + 4(n-k-1)^2 \\
&\quad = 4(n-1)(n-k-1) \\
&\quad \approx 4n(n-k-1) \\
&\approx 4n \sum_{k=0}^{n-1} (n-k-1) + 4 \sum_{k=0}^{n-1} (n-k-1)^2 = 4n \sum_{j=0}^{n-1} j + 4 \sum_{j=0}^{n-1} j^2 \\
&\approx 4n \frac{n(n-1)}{2} + 4 \int_{j=0}^n j^2 dj \approx 2n^3 + \frac{4}{3}n^3 = \frac{10}{3}n^3.
\end{aligned}$$

Thus, the cost is (approximately)

$$\frac{10}{3}n^3 \text{ flops.}$$

 BACK TO TEXT

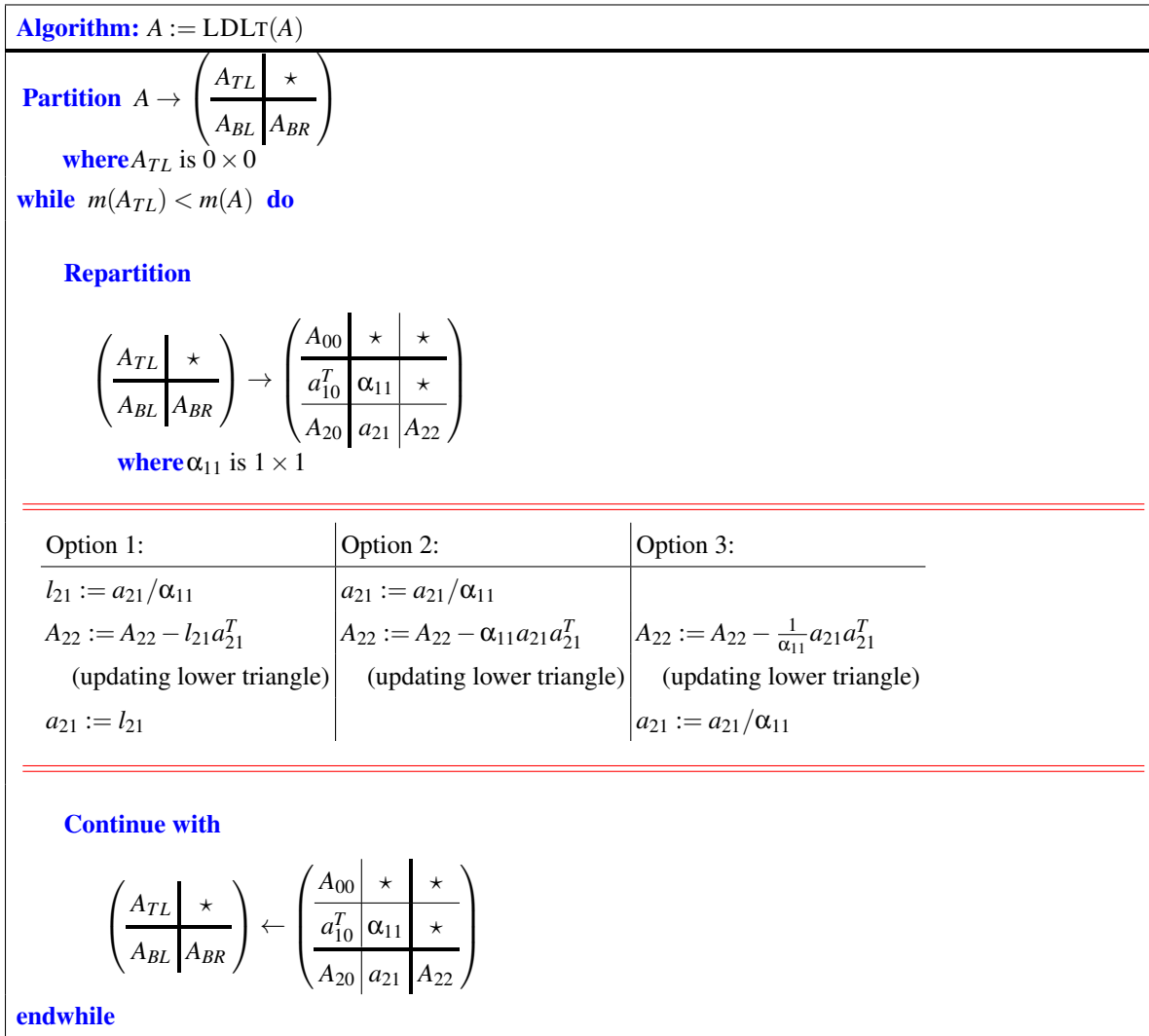


Figure 16.5: Unblocked algorithm for computing  $A \rightarrow LDL^T$ , overwriting  $A$ .

## Chapter 16. Notes on the Method of Relatively Robust Representations (Answers)

**Homework 16.1** Modify the algorithm in Figure 16.1 so that it computes the  $LDL^T$  factorization. (Fill in Figure 16.3.)

**Answer:**

Three possible answers are given in Figure 16.5.

[BACK TO TEXT](#)

**Homework 16.2** Modify the algorithm in Figure 16.2 so that it computes the  $LDL^T$  factorization of a tridiagonal matrix. (Fill in Figure 16.4.) What is the approximate cost, in floating point operations, of

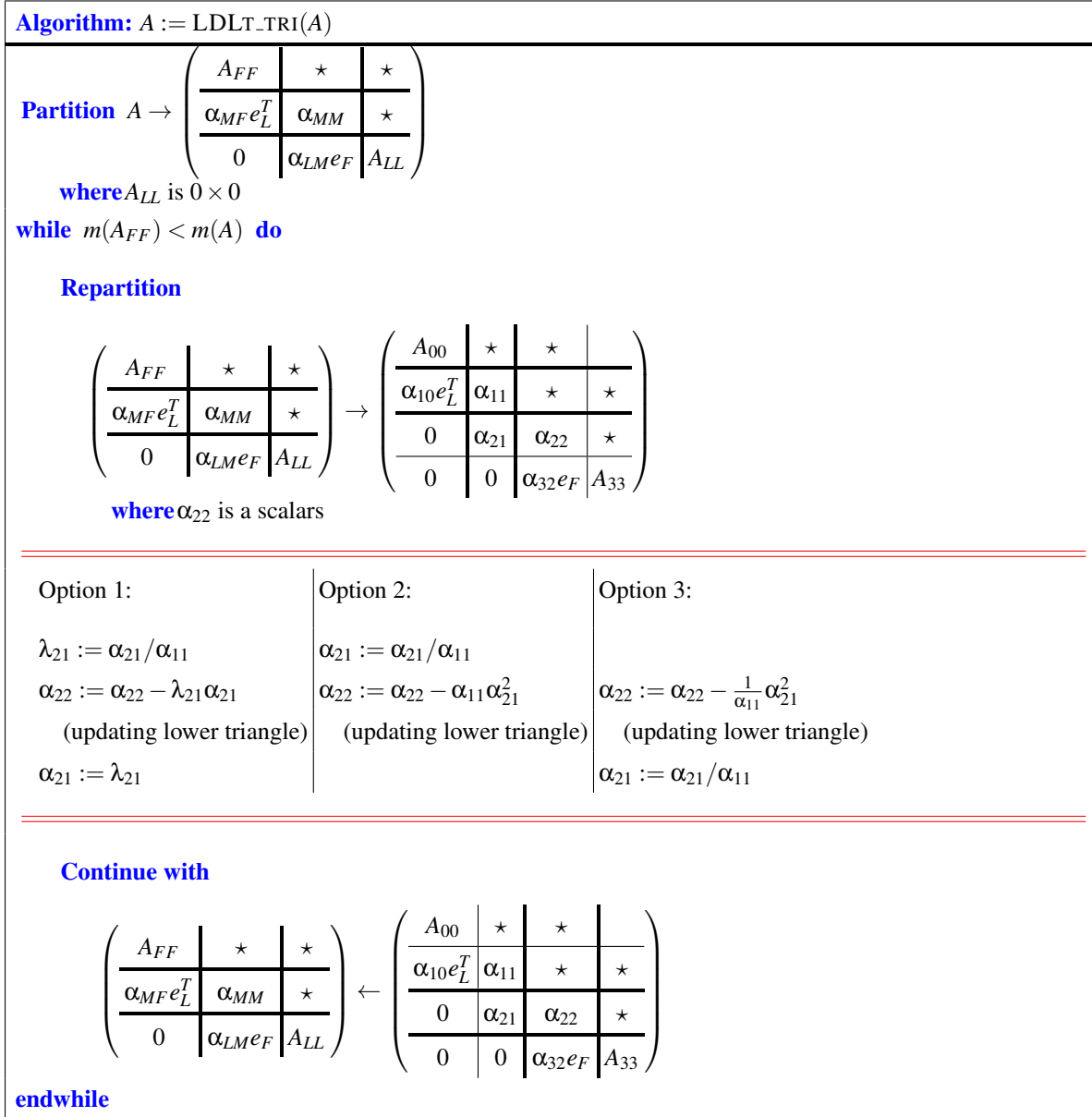


Figure 16.6: Algorithm for computing the the  $LDL^T$  factorization of a tridiagonal matrix.

computing the  $LDL^T$  factorization of a tridiagonal matrix? Count a divide, multiply, and add/subtract as a floating point operation each. Show how you came up with the algorithm, similar to how we derived the algorithm for the tridiagonal Cholesky factorization.

**Answer:**

Three possible algorithms are given in Figure 16.6.

The key insight is to recognize that, relative to the algorithm in Figure 16.5,  $a_{21} = \begin{pmatrix} \alpha_{21} \\ 0 \end{pmatrix}$  so that,



for example,  $a_{21} := a_{21}/\alpha_{11}$  becomes

$$\left( \begin{array}{c|c} \alpha_{21} & \\ \hline 0 & \end{array} \right) := \left( \begin{array}{c|c} \alpha_{21} & \\ \hline 0 & \end{array} \right) / \alpha_{11} = \left( \begin{array}{c|c} \alpha_{21}/\alpha_{11} & \\ \hline 0 & \end{array} \right).$$

Then, an update like  $A_{22} := A_{22} - \alpha_{11}a_{21}a_{21}^T$  becomes

$$\begin{aligned} \left( \begin{array}{c|c} \alpha_{22} & \star \\ \hline \alpha_{32}e_F & A_{33} \end{array} \right) &:= \left( \begin{array}{c|c} \alpha_{22} & \star \\ \hline \alpha_{32}e_F & A_{33} \end{array} \right) - \alpha_{11} \left( \begin{array}{c|c} \alpha_{21} & \\ \hline 0 & \end{array} \right) \left( \begin{array}{c|c} \alpha_{21} & \\ \hline 0 & \end{array} \right)^T \\ &= \left( \begin{array}{c|c} \alpha_{22} - \alpha_{11}\alpha_{21}^2 & \star \\ \hline \alpha_{32}e_F & A_{33} \end{array} \right). \end{aligned}$$

 [BACK TO TEXT](#)

**Homework 16.3** Derive an algorithm that, given an indefinite matrix  $A$ , computes  $A = UDU^T$ . Overwrite only the upper triangular part of  $A$ . (Fill in Figure 16.5.) Show how you came up with the algorithm, similar to how we derived the algorithm for  $LDL^T$ .

**Answer:** Three possible algorithms are given in Figure 16.7.

Partition

$$A \rightarrow \left( \begin{array}{c|c} A_{00} & a_{01} \\ \hline a_{01}^T & \alpha_{11} \end{array} \right) U \rightarrow \left( \begin{array}{c|c} U_{00} & u_{01} \\ \hline 0 & 1 \end{array} \right), \text{ and } D \rightarrow \left( \begin{array}{c|c} D_{00} & 0 \\ \hline 0 & \delta_1 \end{array} \right)$$

Then

$$\begin{aligned} \left( \begin{array}{c|c} A_{00} & a_{01} \\ \hline a_{01}^T & \alpha_{11} \end{array} \right) &= \left( \begin{array}{c|c} U_{00} & u_{01} \\ \hline 0 & 1 \end{array} \right) \left( \begin{array}{c|c} D_{00} & 0 \\ \hline 0 & \delta_1 \end{array} \right) \left( \begin{array}{c|c} U_{00} & u_{01} \\ \hline 0 & 1 \end{array} \right)^T \\ &= \left( \begin{array}{c|c} U_{00}D_{00} & u_{01}\delta_1 \\ \hline 0 & \delta_1 \end{array} \right) \left( \begin{array}{c|c} U_{00}^T & 0 \\ \hline u_{01}^T & 1 \end{array} \right) \\ &= \left( \begin{array}{c|c} U_{00}D_{00}U_{00}^T + \delta_1 u_{01}u_{01}^T & u_{01}\delta_1 \\ \hline \star & \delta_1 \end{array} \right) \end{aligned}$$

This suggests the following algorithm for overwriting the strictly upper triangular part of  $A$  with the strictly upper triangular part of  $U$  and the diagonal of  $A$  with  $D$ :

- Partition  $A \rightarrow \left( \begin{array}{c|c} A_{00} & a_{01} \\ \hline a_{01}^T & \alpha_{11} \end{array} \right)$ .
- $\alpha_{11} := \delta_{11} = \alpha_{11}$  (no-op).

<b>Algorithm:</b> $A := \text{UDUT}(A)$		
<b>Partition</b> $A \rightarrow \left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline \star & A_{BR} \end{array} \right)$ <b>where</b> $A_{BR}$ is $0 \times 0$ <b>while</b> $m(A_{BR}) < m(A)$ <b>do</b>		
<b>Repartition</b> $\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline \star & A_{BR} \end{array} \right) \rightarrow \left( \begin{array}{c c c} A_{00} & a_{01} & A_{02} \\ \hline \star & \alpha_{11} & a_{12}^T \\ \hline \star & \star & A_{22} \end{array} \right)$ <b>where</b> $\alpha_{11}$ is $1 \times 1$		
Option 1:	Option 2:	Option 3:
$u_{01} := a_{01}/\alpha_{11}$	$u_{01} := a_{01}/\alpha_{11}$	
$A_{00} := A_{00} - u_{01}a_{01}^T$ (updating upper triangle)	$A_{00} := A_{00} - \alpha_{11}a_{01}a_{01}^T$ (updating upper triangle)	$A_{00} := A_{00} - \frac{1}{\alpha_{11}}a_{01}a_{01}^T$ (updating upper triangle)
$a_{01} := u_{01}$		$a_{01} := a_{01}/\alpha_{11}$
<b>Continue with</b> $\left( \begin{array}{c c} A_{TL} & A_{TR} \\ \hline \star & A_{BR} \end{array} \right) \leftarrow \left( \begin{array}{c c c} A_{00} & a_{01} & A_{02} \\ \hline \star & \alpha_{11} & a_{12}^T \\ \hline \star & \star & A_{22} \end{array} \right)$		
<b>endwhile</b>		

Figure 16.7: Algorithm for computing the the  $UDU^T$  factorization of a tridiagonal matrix.

- Compute  $u_{01} := u_{01}/\alpha_{11}$ .
- Update  $A_{22} := A_{22} - u_{01}a_{01}^T$  (updating only the upper triangular part).
- $a_{01} := u_{01}$ .
- Continue with computing  $A_{00} \rightarrow U_{00}D_{00}L_{00}^T$ .

This algorithm will complete as long as  $\delta_{11} \neq 0$ ,

[👉 BACK TO TEXT](#)

**Homework 16.4** Derive an algorithm that, given an indefinite tridiagonal matrix  $A$ , computes  $A = UDU^T$ . Overwrite only the upper triangular part of  $A$ . (Fill in Figure 16.6.) Show how you came up with the algorithm, similar to how we derived the algorithm for  $LDL^T$ .

**Answer:** Three possible algorithms are given in Figure 16.8.

The key insight is to recognize that, relative to the algorithm in Figure 16.7,  $\alpha_{11} = \alpha_{22}$  and  $a_{10} = \begin{pmatrix} 0 \\ \alpha_{12} \end{pmatrix}$  so that, for example,  $a_{10} := a_{10}/\alpha_{11}$  becomes

$$\begin{pmatrix} 0 \\ \alpha_{12} \end{pmatrix} := \begin{pmatrix} 0 \\ \alpha_{12} \end{pmatrix} / \alpha_{22} = \begin{pmatrix} 0 \\ \alpha_{12}/\alpha_{22} \end{pmatrix}.$$

Then, an update like  $A_{00} := A_{00} - \alpha_{11}a_{01}a_{01}^T$  becomes

$$\begin{aligned} \left( \begin{array}{c|c} A_{00} & \alpha_{01}e_L \\ \hline \star & \alpha_{11} \end{array} \right) &:= \left( \begin{array}{c|c} A_{00} & \alpha_{01}e_L \\ \hline \star & \alpha_{11} \end{array} \right) - \alpha_{22} \begin{pmatrix} 0 \\ \alpha_{12} \end{pmatrix} \begin{pmatrix} 0 \\ \alpha_{12} \end{pmatrix}^T \\ &= \left( \begin{array}{c|c} A_{00} & \alpha_{01}e_L \\ \hline \star & \alpha_{11} - \alpha_{12}^2 \end{array} \right). \end{aligned}$$

[👉 BACK TO TEXT](#)

**Homework 16.5** Show that, provided  $\phi_1$  is chosen appropriately,

$$\begin{aligned} \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right) \left( \begin{array}{c|c|c} D_{00} & 0 & 0 \\ \hline 0 & \phi_1 & 0 \\ \hline 0 & 0 & E_{22} \end{array} \right) \left( \begin{array}{c|c|c} L_{00} & 0 & 0 \\ \hline \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ \hline 0 & 0 & U_{22} \end{array} \right)^T \\ = \left( \begin{array}{c|c|c} A_{00} & \alpha_{01}e_L & 0 \\ \hline \alpha_{01}e_L^T & \alpha_{11} & \alpha_{21}e_F^T \\ \hline 0 & \alpha_{21}e_F & A_{22} \end{array} \right). \end{aligned}$$

<b>Algorithm:</b> $A := \text{UDUT\_TRI}(A)$		
<b>Partition</b> $A \rightarrow \left( \begin{array}{c c c} A_{FF} & \alpha_{FM}e_L^T & 0 \\ \hline \star & \alpha_{MM} & \alpha_{ML}e_F^T \\ \hline \star & \star & A_{LL} \end{array} \right)$		
<b>where</b> $A_{FF}$ is $0 \times 0$		
<b>while</b> $m(A_{LL}) < m(A)$ <b>do</b>		
<b>Repartition</b>		
$\left( \begin{array}{c c c} A_{FF} & \alpha_{FM}e_L & 0 \\ \hline \star & \alpha_{MM} & \alpha_{ML}e_F^T \\ \hline \star & \star & A_{LL} \end{array} \right) \rightarrow \left( \begin{array}{c c c c} A_{00} & \alpha_{01}e_L & 0 & 0 \\ \hline \star & \alpha_{11} & \alpha_{12} & 0 \\ \hline \star & \star & \alpha_{22} & \alpha_{23}e_F^T \\ \hline \star & \star & \star & A_{33} \end{array} \right)$		
<b>where</b>		
<hr/>		
Option 1:	Option 2:	Option 3:
$v_{12} := \alpha_{12}/\alpha_{22}$	$\alpha_{12} := \alpha_{12}/\alpha_{22}$	
$\alpha_{11} := \alpha_{11} - v_{12}\alpha_{12}^T$	$\alpha_{11} := \alpha_{11} - \alpha_{22}\alpha_{12}^2$	$\alpha_{11} := \alpha_{11} - \frac{1}{\alpha_{22}}\alpha_{12}^2$
(updating upper triangle)	(updating upper triangle)	(updating upper triangle)
$\alpha_{12} := v_{12}$		$\alpha_{12} := \alpha_{12}/\alpha_{22}$
<hr/>		
<b>Continue with</b>		
$\left( \begin{array}{c c c} A_{FF} & \alpha_{FM}e_L & 0 \\ \hline \star & \alpha_{MM} & \alpha_{ML}e_F^T \\ \hline \star & \star & A_{LL} \end{array} \right) \leftarrow \left( \begin{array}{c c c c} A_{00} & \alpha_{01}e_L & 0 & 0 \\ \hline \star & \alpha_{11} & \alpha_{12} & 0 \\ \hline \star & \star & \alpha_{22} & \alpha_{23}e_F^T \\ \hline \star & \star & \star & A_{33} \end{array} \right)$		
<b>endwhile</b>		

Figure 16.8: Algorithm for computing the the  $UDU^T$  factorization of a tridiagonal matrix.

(Hint: multiply out  $A = LDL^T$  and  $A = UEU^T$  with the partitioned matrices first. Then multiply out the above. Compare and match...) How should  $\phi_1$  be chosen? What is the cost of computing the twisted factorization given that you have already computed the  $LDL^T$  and  $UDU^T$  factorizations? A "Big O" estimate is sufficient. Be sure to take into account what  $e_L^T D_{00} e_L$  and  $e_F^T E_{22} e_F$  equal in your cost estimate.

**Answer:**

$$\begin{aligned}
& \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10} e_L^T & 1 & 0 \\ 0 & \lambda_{21} e_F & L_{22} \end{pmatrix} \begin{pmatrix} D_{00} & 0 & 0 \\ 0 & \delta_1 & 0 \\ 0 & 0 & D_{22} \end{pmatrix} \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10} e_L^T & 1 & 0 \\ 0 & \lambda_{21} e_F & L_{22} \end{pmatrix}^T \\
&= \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10} e_L^T & 1 & 0 \\ 0 & \lambda_{21} e_F & L_{22} \end{pmatrix} \begin{pmatrix} D_{00} & 0 & 0 \\ 0 & \delta_1 & 0 \\ 0 & 0 & D_{22} \end{pmatrix} \begin{pmatrix} L_{00}^T & 0 & \lambda_{10} e_L \\ 0 & 1 & \lambda_{21} e_F^T \\ 0 & & L_{22}^T \end{pmatrix} \\
&= \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10} e_L^T & 1 & 0 \\ 0 & \lambda_{21} e_F & L_{22} \end{pmatrix} \begin{pmatrix} D_{00} L_{00}^T & \lambda_{10} D_{00} e_L & 0 \\ 0 & \delta_1 & \lambda_{21} \delta_1 e_F^T \\ 0 & 0 & D_{22} L_{22}^T \end{pmatrix} \\
&= \begin{pmatrix} L_{00} D_{00} L_{00}^T & \lambda_{10} L_{00} D_{00} e_L & 0 \\ \lambda_{10} e_L^T D_{00} L_{00}^T & \lambda_{10}^2 e_L^T D_{00} e_L + \delta_1 & \lambda_{21} \delta_1 e_F^T \\ 0 & \lambda_{21} \delta_1 e_F & \lambda_{21}^2 \delta_1 e_F e_F^T + L_{22} D_{22} L_{22}^T \end{pmatrix} = \begin{pmatrix} A_{00} & \alpha_{01} e_L & 0 \\ \alpha_{01} e_L^T & \alpha_{11} & \alpha_{21} e_F^T \\ 0 & \alpha_{21} e_F & A_{22} \end{pmatrix}. \quad (16.1)
\end{aligned}$$

and

$$\begin{aligned}
& \begin{pmatrix} U_{00} & v_{01} e_L & 0 \\ 0 & 1 & v_{21} e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} E_{00} & 0 & 0 \\ 0 & \varepsilon_1 & 0 \\ 0 & 0 & E_{22} \end{pmatrix} \begin{pmatrix} U_{00} & v_{01} e_L & 0 \\ 0 & 1 & v_{21} e_F^T \\ 0 & 0 & U_{22} \end{pmatrix}^T \\
&= \begin{pmatrix} U_{00} & v_{01} e_L & 0 \\ 0 & 1 & v_{21} e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} E_{00} & 0 & 0 \\ 0 & \varepsilon_1 & 0 \\ 0 & 0 & E_{22} \end{pmatrix} \begin{pmatrix} U_{00}^T & 0 & 0 \\ v_{01} e_L^T & 1 & 0 \\ 0 & v_{21} e_F & U_{22}^T \end{pmatrix} \\
&= \begin{pmatrix} U_{00} & v_{01} e_L & 0 \\ 0 & 1 & v_{21} e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} E_{00} U_{00}^T & 0 & 0 \\ \varepsilon_1 v_{01} e_L^T & \varepsilon_1 & 0 \\ 0 & v_{21} E_{22} e_F & E_{22} U_{22}^T \end{pmatrix} \\
&= \begin{pmatrix} U_{00} E_{00} U_{00}^T + v_{01}^2 \varepsilon_1 e_L e_L^T & v_{01} \varepsilon_1 e_L & 0 \\ v_{01} \varepsilon_1 e_L^T & \varepsilon_1 + v_{01}^2 e_F^T E_{22} e_F & \varepsilon_1 v_{21} e_F^T E_{22} U_{22}^T \\ 0 & v_{21} U_{22}^T E_{22} e_F & U_{22} E_{22} U_{22}^T \end{pmatrix} = \begin{pmatrix} A_{00} & \alpha_{01} e_L & 0 \\ \alpha_{01} e_L^T & \alpha_{11} & \alpha_{21} e_F^T \\ 0 & \alpha_{21} e_F & A_{22} \end{pmatrix}. \quad (16.2)
\end{aligned}$$

Finally,

$$\begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10} e_L^T & 1 & v_{21} e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} D_{00} & 0 & 0 \\ 0 & \phi_1 & 0 \\ 0 & 0 & E_{22} \end{pmatrix} \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10} e_L^T & 1 & v_{21} e_F^T \\ 0 & 0 & U_{22} \end{pmatrix}^T$$

$$\begin{aligned}
&= \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} D_{00} & 0 & 0 \\ 0 & \phi_1 & 0 \\ 0 & 0 & E_{22} \end{pmatrix} \begin{pmatrix} L_{00}^T & \lambda_{10}e_L & 0 \\ 0 & 1 & 0 \\ 0 & v_{21}e_F & U_{22}^T \end{pmatrix} \\
&= \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} D_{00}L_{00}^T & \lambda_{10}D_{00}e_L & 0 \\ 0 & \phi_1 & 0 \\ 0 & v_{21}E_{22}e_F & E_{22}U_{22}^T \end{pmatrix} \\
&= \begin{pmatrix} L_{00}D_{00}L_{00}^T & \lambda_{10}L_{00}D_{00}e_L & 0 \\ \lambda_{10}e_L^TD_{00}L_{00}^T & \lambda_{10}^2e_L^TD_{00}e_L + \phi_1 + v_{21}^2e_F^TE_{22}e_F & v_{21}e_F^TE_{22}U_{22}^T \\ 0 & v_{21}U_{22}E_{22}e_F & U_{22}E_{22}U_{22}^T \end{pmatrix} = \begin{pmatrix} A_{00} & \alpha_{01}e_L & 0 \\ \alpha_{01}e_L^T & \alpha_{11} & \alpha_{21}e_F^T \\ 0 & \alpha_{21}e_F & A_{22} \end{pmatrix}. \quad (16.3)
\end{aligned}$$

The equivalence of the submatrices highlighted in yellow in (16.1) and (??) justify the submatrices highlighted in yellow in (16.3).

The equivalence of the submatrices highlighted in grey in (16.1), (??, and (16.3) tell us that

$$\begin{aligned}
\alpha_{11} &= \lambda_{10}^2 e_L^T D_{00} e_L + \delta_1 \\
\alpha_{11} &= v_{01}^2 e_F^T E_{22} e_F + \varepsilon_1 \\
\alpha_{11} &= \lambda_{10}^2 e_L^T D_{00} e_L + \phi_1 + v_{21}^2 e_F^T E_{22} e_F.
\end{aligned}$$

or

$$\begin{aligned}
\alpha_{11} - \delta_1 &= \lambda_{10}^2 e_L^T D_{00} e_L \\
\alpha_{11} - \varepsilon_1 &= v_{01}^2 e_F^T E_{22} e_F
\end{aligned}$$

so that

$$\alpha_{11} = (\alpha_{11} - \delta_1) + \phi_1 + (\alpha_{11} - \varepsilon_1).$$

Solving this for  $\phi_1$  yields

$$\phi_1 = \delta_1 + \varepsilon_1 - \alpha_{11}.$$

Notice that, given the factorizations  $A = LDL^T$  and  $A = UEU^T$  the cost of computing the twisted factorization is  $O(1)$ .

➡ BACK TO TEXT

**Homework 16.6** Compute  $x_0$ ,  $x_1$ , and  $x_2$  so that

$$\begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} D_{00} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & E_{22} \end{pmatrix} \underbrace{\begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix}^T \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}}_{\text{Hint: } \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

where  $x = \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix}$  is not a zero vector. What is the cost of this computation, given that  $L_{00}$  and  $U_{22}$  have special structure?

**Answer:** Choose  $x_0$ ,  $\chi_1$ , and  $x_2$  so that

$$\begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix}^T \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

or, equivalently,

$$\begin{pmatrix} L_{00}^T & \lambda_{10}e_L & 0 \\ 0 & 1 & 0 \\ 0 & v_{21}e_F & U_{22}^T \end{pmatrix} \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

We conclude that  $\chi_1 = 1$  and

$$\begin{aligned} L_{00}^T x_0 &= -\lambda_{10}e_L \\ U_{22}^T x_2 &= -v_{21}e_F \end{aligned}$$

so that  $x_0$  and  $x_2$  can be computed via solves with a bidiagonal upper triangular matrix ( $L_{00}^T$ ) and bidiagonal lower triangular matrix ( $U_{00}^T$ ), respectively.

Then

$$\begin{aligned} & \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} D_{00} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & E_{22} \end{pmatrix} \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix}^T \begin{pmatrix} x_0 \\ \chi_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} D_{00} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & E_{22} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} L_{00} & 0 & 0 \\ \lambda_{10}e_L^T & 1 & v_{21}e_F^T \\ 0 & 0 & U_{22} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \end{aligned}$$

Now, consider solving  $L_{00}^T x_0 = -\lambda_{10} e_L$ :

$$\begin{pmatrix} \times & \times & 0 & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & \times & \times & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -\lambda_{10} \end{pmatrix}$$

A moment of reflection shows that if  $L_{00}$  is  $k \times k$ , then solving  $L_{00}^T x_0 = -\lambda_{10} e_L$  requires  $O(k)$  flops. Similarly, since then  $U_{22}$  is then  $(n - k - 1) \times (n - k - 1)$ , then solving  $U_{22}^T x_2 = -v_{21} e_F$  requires  $O(n - k - 1)$  flops. Thus, computing  $x$  requires  $O(n)$  flops.

Thus,

- Given tridiagonal  $\hat{A}$  computing all eigenvalues requires  $O(n^2)$  computation. (We have not discussed this in detail...)
- Then, for each eigenvalue
  - Computing  $A := \hat{A} - \lambda I$  requires  $O(n)$  flops.
  - Factoring  $A = LDL^T$  requires  $O(n)$  flops.
  - Factoring  $A = UEU^T$  requires  $O(n)$  flops.
  - Computing *all*  $\phi_1$  so that the smallest can be chose requires  $O(n)$  flops.
  - Computing the eigenvector of  $\hat{A}$  associated with  $\lambda$  from the twisted factorization requires  $O(n)$  computation.

Thus, computing all eigenvalues and eigenvectors of a tridiagonal matrix via this method requires  $O(n^2)$  computation. This is *much* better than computing these via the tridiagonal QR algorithm.

➡ BACK TO TEXT



## Chapter 17 Notes on Computing the SVD (Answers)

**Homework 17.2** If  $A = U\Sigma V^T$  is the SVD of  $A$  then  $A^T A = V\Sigma^2 V^T$  is the Spectral Decomposition of  $A^T A$ .

[👉 BACK TO TEXT](#)

**Homework 17.3 Homework 17.4** Give the approximate total cost for reducing  $A \in \mathbb{R}^{n \times n}$  to bidiagonal form.

[👉 BACK TO TEXT](#)

## Chapter 18. Notes on Splitting Methods (Answers)

**Homework 18.10** Prove the above theorem.





**Answer:** Let  $\lambda$  equal the eigenvector such that  $|\lambda| = \rho(A)$ . Then

$$\|A\| = \max_{\|y\|=1} \|Ay\| \geq \max_{\substack{\|x\|=1 \\ Ax=\lambda x}} \|Ax\| = \max_{\substack{\|x\|=1 \\ Ax=\lambda x}} \|\lambda x\| = \max_{\substack{\|x\|=1 \\ Ax=\lambda x}} |\lambda| \|x\| = |\lambda| = \rho(A).$$

[👉 BACK TO TEXT](#)

## How to Download

Videos associated with these notes can be viewed in one of three ways:

-  **YouTube** links to the video uploaded to YouTube.
-  **Download from UT Box** links to the video uploaded to UT-Austin's "UT Box" File Sharing Service".
-  **View After Local Download** links to the video downloaded to your own computer in directory Video within the same directory in which you stored this document. You can download videos by first linking on  **Download from UT Box**. Alternatively, visit the **UT Box directory** in which the videos are stored and download some or all.



# Bibliography

- [1] E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, Jack J. Dongarra, J. Du Croz, S. Hammarling, A. Greenbaum, A. McKenney, and D. Sorensen. *LAPACK Users' guide (third ed.)*. SIAM, Philadelphia, PA, USA, 1999.
- [2] E. Anderson, Z. Bai, J. Demmel, J. E. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. E. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, 1992.
- [3] Paolo Bientinesi. *Mechanical Derivation and Systematic Analysis of Correct Linear Algebra Algorithms*. PhD thesis, Department of Computer Sciences, The University of Texas, 2006. Technical Report TR-06-46. September 2006.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [4] Paolo Bientinesi, John A. Gunnels, Margaret E. Myers, Enrique S. Quintana-Ortí, and Robert A. van de Geijn. The science of deriving dense linear algebra algorithms. *ACM Trans. Math. Soft.*, 31(1):1–26, March 2005.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [5] Paolo Bientinesi, Enrique S. Quintana-Ortí, and Robert A. van de Geijn. Representing linear algebra algorithms in code: The FLAME APIs. *ACM Trans. Math. Soft.*, 31(1):27–59, March 2005.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [6] Paolo Bientinesi and Robert A. van de Geijn. The science of deriving stability analyses. FLAME Working Note #33. Technical Report AICES-2008-2, Aachen Institute for Computational Engineering Sciences, RWTH Aachen, November 2008.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [7] Paolo Bientinesi and Robert A. van de Geijn. Goal-oriented and modular stability analysis. *SIAM J. Matrix Anal. & Appl.*, 32(1):286–308, 2011.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [8] Paolo Bientinesi and Robert A. van de Geijn. Goal-oriented and modular stability analysis. *SIAM J. Matrix Anal. Appl.*, 32(1):286–308, March 2011.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.  
We suggest you read FLAME Working Note #33 for more details.
- [9] Christian Bischof and Charles Van Loan. The WY representation for products of Householder matrices. *SIAM J. Sci. Stat. Comput.*, 8(1):s2–s13, Jan. 1987.
- [10] Basic linear algebra subprograms technical forum standard. *International Journal of High Performance Applications and Supercomputing*, 16(1), 2002.

- [11] James W. Demmel. *Applied Numerical Linear Algebra*. SIAM, 1997.
- [12] I. S. Dhillon. *A New  $O(n^2)$  Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*. PhD thesis, Computer Science Division, University of California, Berkeley, California, May 1997. Available as UC Berkeley Technical Report No. UCB//CSD-97-971.
- [13] I. S. Dhillon. Reliable computation of the condition number of a tridiagonal matrix in  $O(n)$  time. *SIAM J. Matrix Anal. Appl.*, 19(3):776–796, July 1998.
- [14] I. S. Dhillon and B. N. Parlett. Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices. *Lin. Alg. Appl.*, 387:1–28, August 2004.
- [15] Inderjit S. Dhillon, Beresford N. Parlett, and Christof Vömel. The design and implementation of the MRRR algorithm. *ACM Trans. Math. Soft.*, 32(4):533–560, December 2006.
- [16] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart. *LINPACK Users' Guide*. SIAM, Philadelphia, 1979.
- [17] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Iain Duff. A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Soft.*, 16(1):1–17, March 1990.
- [18] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Richard J. Hanson. An extended set of FORTRAN basic linear algebra subprograms. *ACM Trans. Math. Soft.*, 14(1):1–17, March 1988.
- [19] Jack J. Dongarra, Iain S. Duff, Danny C. Sorensen, and Henk A. van der Vorst. *Solving Linear Systems on Vector and Shared Memory Computers*. SIAM, Philadelphia, PA, 1991.
- [20] Jack J. Dongarra, Sven J. Hammarling, and Danny C. Sorensen. Block reduction of matrices to condensed forms for eigenvalue computations. *Journal of Computational and Applied Mathematics*, 27, 1989.
- [21] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 3rd edition, 1996.
- [22] Kazushige Goto and Robert van de Geijn. Anatomy of high-performance matrix multiplication. *ACM Trans. Math. Soft.*, 34(3):12:1–12:25, May 2008.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [23] John A. Gunnels. *A Systematic Approach to the Design and Analysis of Parallel Dense Linear Algebra Algorithms*. PhD thesis, Department of Computer Sciences, The University of Texas, December 2001.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [24] John A. Gunnels, Fred G. Gustavson, Greg M. Henry, and Robert A. van de Geijn. FLAME: Formal Linear Algebra Methods Environment. *ACM Trans. Math. Soft.*, 27(4):422–455, December 2001.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [25] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

- [26] C. G. J. Jacobi. Über ein leichtes Verfahren, die in der Theorie der Säkular-störungen vorkommenden Gleichungen numerisch aufzulösen. *Crelle's Journal*, 30:51–94, 1846.
- [27] Thierry Joffrain, Tze Meng Low, Enrique S. Quintana-Ortí, Robert van de Geijn, and Field G. Van Zee. Accumulating Householder transformations, revisited. *ACM Trans. Math. Softw.*, 32(2):169–179, June 2006.
- [28] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh. Basic linear algebra subprograms for Fortran usage. *ACM Trans. Math. Soft.*, 5(3):308–323, Sept. 1979.
- [29] Margaret E. Myers, Pierce M. van de Geijn, and Robert A. van de Geijn. *Linear Algebra: Foundations to Frontiers - Notes to LAFF With*. Self published, 2014.  
Download from <http://www.ulaff.net>.
- [30] Jack Poulson, Bryan Marker, Robert A. van de Geijn, Jeff R. Hammond, and Nichols A. Romero. Elemental: A new framework for distributed memory dense matrix computations. *ACM Trans. Math. Softw.*, 39(2):13:1–13:24, February 2013.
- [31] C. Puglisi. Modification of the Householder method based on the compact WY representation. *SIAM J. Sci. Stat. Comput.*, 13:723–726, 1992.
- [32] Gregorio Quintana-Ortí and Robert van de Geijn. Improving the performance of reduction to Hessenberg form. *ACM Trans. Math. Softw.*, 32(2):180–194, June 2006.
- [33] Robert Schreiber and Charles Van Loan. A storage-efficient WY representation for products of Householder transformations. *SIAM J. Sci. Stat. Comput.*, 10(1):53–57, Jan. 1989.
- [34] G. W. Stewart. *Matrix Algorithms Volume I: Basic Decompositions*. SIAM, Philadelphia, PA, USA, 1998.
- [35] G. W. Stewart. *Matrix Algorithms Volume II: Eigensystems*. SIAM, Philadelphia, PA, USA, 2001.
- [36] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, 1997.
- [37] Robert van de Geijn and Kazushige Goto. *Encyclopedia of Parallel Computing*, chapter BLAS (Basic Linear Algebra Subprograms), pages Part 2, 157–164. Springer, 2011.
- [38] Robert A. van de Geijn and Enrique S. Quintana-Ortí. *The Science of Programming Matrix Computations*. [www.lulu.com/contents/contents/1911788/](http://www.lulu.com/contents/contents/1911788/), 2008.
- [39] Field G. Van Zee. libflame: *The Complete Reference*. [www.lulu.com](http://www.lulu.com), 2012.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [40] Field G. Van Zee, Ernie Chan, Robert van de Geijn, Enrique S. Quintana-Ortí, and Gregorio Quintana-Ortí. The libflame library for dense matrix computations. *IEEE Computation in Science & Engineering*, 11(6):56–62, 2009.
- [41] Field G. Van Zee and Robert A. van de Geijn. BLIS: A framework for rapid instantiation of BLAS functionality. *ACM Trans. Math. Soft.*, 2015. To appear.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.

- [42] Field G. Van Zee, Robert A. van de Geijn, and Gregorio Quintana-Ortí. Restructuring the tridiagonal and bidiagonal QR algorithms for performance. *ACM Trans. Math. Soft.*, 40(3):18:1–18:34, April 2014.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [43] Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí, and G. Joseph Elizondo. Families of algorithms for reducing a matrix to condensed form. *ACM Trans. Math. Soft.*, 39(1), 2012.  
Download from <http://www.cs.utexas.edu/users/flame/web/FLAMEPublications.html>.
- [44] H. F. Walker. Implementation of the GMRES method using Householder transformations. *SIAM J. Sci. Stat. Comput.*, 9(1):152–163, 1988.
- [45] David S. Watkins. *Fundamentals of Matrix Computations, 3rd Edition*. Wiley, third edition, 2010.
- [46] Stephen J. Wright. A collection of problems for which Gaussian elimination with partial pivoting is unstable. *SIAM J. Sci. Comput.*, 14(1):231–238, 1993.



# Index

## Greek letters

$\alpha$ , 4  
 $\beta$ , 4  
 $\gamma$ , 4  
 $\delta$ , 4  
 $\varepsilon$ , 4  
 $\eta$ , 4  
 $\kappa$ , 4  
 $\lambda$ , 4  
 $\mu$ , 4  
 $\nu$ , 4  
 $\omega$ , 4  
 $\pi$ , 4  
 $\psi$ , 4  
 $\rho$ , 4  
 $\sigma$ , 4  
 $\tau$ , 4  
 $\theta$ , 4  
 $\upsilon$ , 4  
 $\chi$ , 4  
 $\xi$ , 4  
 $\zeta$ , 4

$|\cdot|$  (e.g.,  $|\alpha|$ ), 5

$\mathbb{C}$ , 3

$\mathbb{C}^{m \times n}$ , 3

$\mathbb{C}^n$ , 3

$\kappa(\cdot)$  (e.g.,  $\kappa(A)$ ), 33

$H$  (e.g.,  $A^H$ ), 6

$c$  (e.g.,  $A^c$ ), 6

$|\cdot|$  (e.g.,  $|\alpha|$ ), 25

$\|\cdot\|_F$  (e.g.,  $\|A\|_F$ ), 27

$\|\cdot\|_\infty$  (e.g.,  $\|A\|_\infty$ ), 30

$\|\cdot\|_\infty$  (e.g.,  $\|x\|_\infty$ ), 27

$\|\cdot\|_\mu$  (e.g.,  $\|x\|_\mu$ ), 28

$\|\cdot\|_{\mu,\nu}$  (e.g.,  $\|A\|_{\mu,\nu}$ ), 28

$\|\cdot\|_\nu$  (e.g.,  $\|x\|_\mu$ ), 28

$\|\cdot\|_1$  (e.g.,  $\|A\|_1$ ), 30

$\|\cdot\|_1$  (e.g.,  $\|x\|_1$ ), 27

$\|\cdot\|_p$  (e.g.,  $\|A\|_p$ ), 29

$\|\cdot\|_2$  (e.g.,  $\|A\|_2$ ), 30

$\|\cdot\|_2$  (e.g.,  $\|x\|_2$ ), 25

$\overline{\phantom{x}}$  (e.g.,  $\overline{\alpha}$ ), 5

$\mathbb{R}$ , 3

$\mathbb{R}^{m \times n}$ , 3

$\mathbb{R}^n$ , 3

$T$  (e.g.,  $A^T$ ), 5

$\hat{\phantom{x}}$  (e.g.,  $\hat{a}_i^T$ ), 3

absolute value, 5, 25

axpy, 10

cost, 11, 12

blocked matrix-matrix multiplication, 22

bordered Cholesky factorization algorithm, 204

Cauchy-Schwartz inequality, 25

CGS, 59

Cholesky factorization

bordered algorithm, 204

other algorithm, 204

Classical Gram-Schmidt, 59

code skeleton, 77

complete pivoting

LU factorization, 174

complex conjugate, 5

complex scalar

absolute value, 5

condition number, 33, 54

conformal partitioning, 22

- conjugate, 5
  - complex, 5
  - matrix, 5
  - scalar, 5
  - vector, 5
- dot, 11
- dot product, 11
- ”dot” product, 12
- eigenpair
  - definition, 217
- eigenvalue, 215–222
  - definition, 217
- eigenvector, 215–222
  - definition, 217
- equivalence of norms, 33
- FLAME
  - API, 73–81
  - notation, 75
- FLAME API, 73–81
- FLAME notation, 13
- floating point operations, 10
- flop, 10
- Frobenius norm, 27
  - definition, 27
- Gauss transform, 162–164
- Gaussian elimination, 159–193
- gemm
  - cost, 20
- gemm, 18
- gemv
  - cost, 14
- gemv, 13
- ger
  - cost, 17
- ger, 15
- Gershgorin Disc Theorem, 309
- Gram-Schmidt
  - Classical, 59
  - cost, 70
  - Madified, 64
- Gram-Schmidt orthogonalization
  - implementation, 75
- Gram-Schmidt QR factorization, 57–72
- Greek letters
  - use of, 4
- Hermitian transpose
  - vector, 6
- homogeneous, 25
- Householder notation, 3
- Householder transformation, 85
- $i$ , 5
- Greek letters
  - entity matrix, 4
- induced matrix norm, 28
- infinity norm
  - vector, 27
- $\infty$ -norm
  - definition, 27
  - matrix, 30
  - vector, 27
- inner product, 11
- invariant subspace, 329
- inverse
  - Moore-Penrose generalized, 50
  - pseudo, 50
- LAFF Notes, x
- linear system
  - conditioning, 32
- linear system solve, 175
  - triangular, 175–178
- low-rank approximation, 50
- lower triangular solve, 175–178
- LU decomposition, 161
- LU factorization, 159–193
  - complete pivoting, 174
  - cost, 164–165
  - existence, 161
  - existence proof, 173–174
  - partial pivoting, 165–172
    - algorithm, 167
- LU factorization:derivation, 161–162
- MAC, 21
- Matlab, 75
- matrix
  - condition number, 33, 54
  - conjugate, 5

- $\infty$ -norm, 30
- low-rank approximation, 50
- 1-norm, 30
- orthogonal, 35–55
- orthogormal, 37
- permutation, 165
- spectrum, 217
- transpose, 5–8
- 2-norm, 30
- unitary, 37
- matrix norm, 23–33
  - Frobenius, 27
  - definition, 27
  - induced, 28
  - submultiplicative, 31
  - definition, 31
- matrix p-norm
  - definition, 29
- matrix-matrix multiplication, 18
  - blocked, 22
  - cost, 20
  - element-by-element, 18
  - via matrix-vector multiplications, 19
  - via rank-1 updates, 20
  - via row-vector times matrix multiplications, 19
- matrix-matrix operation
  - gemm, 18
  - matrix-matrix multiplication, 18
- matrix-matrix product, 18
- matrix-vector multiplication, 12
  - algorithm, 14
  - by columns, 14
  - by rows, 14
  - cost, 14
  - via axpy, 14
  - via dot, 14
- matrix-vector operation, 12
  - gemv, 13
  - ger, 15
  - matrix-vector multiplication, 12
  - rank-1 update, 15
- matrix-vector product, 12
- max, 27, 28
- MGS, 64
- Modified Gram-Schmidt, 64
- Moore-Penrose generalized inverse, 50
- multiplication
  - matrix-matrix, 18
    - blocked, 22
    - cost, 20
    - element-by-element, 18
    - via matrix-vector multiplications, 19
    - via rank-1 updates, 20
    - via row-vector times matrix multiplications, 19
  - matrix-vector, 12
    - cost, 14
- multiply-accumulate, 21
- norm, 23
  - matrix, 23
  - submultiplicative, 31
  - vector, 23–33
  - definition, 25
- norms
  - equivalence, 33
- notation
  - FLAME, 13
  - Householder, 3
- Octave, 75
- 1-norm
  - definition, 27
  - matrix, 30
  - vector, 27
- orthonormal basis, 57
- outer product, 15
- p-norm
  - matrix
    - definition, 29
  - vector, 27
- partial pivoting
  - LU factorization, 165–172
- partitioning
  - conformal, 22
- permutation
  - matrix, 165
- positive definite, 25
- preface, ix
- product
  - dot, 11
  - inner, 11

- matrix-matrix, 18
  - matrix-vector, 12
  - outer, 15
- projection
  - onto column space, 49
- pseudo inverse, 50
- QR factorization, 61
  - Gram-Schmidt, 57–72
- rank-1 update, 15
  - algorithm, 17
  - by columns, 17
  - by rows, 17
  - cost, 17
  - via axpy, 17
  - via dot, 17
- Reflector, 85
- reflector, 85
- residual, 303
- Roman letters
  - use of, 4
- scal, 9
  - cost, 10
- scaled vector addition, 10
  - cost, 11, 12
- scaling
  - vector
    - cost, 10
- singular value, 41
- Singular Value Decomposition, 35, 39, 41–55
  - geometric interpretation, 41
  - reduced, 45
  - theorem, 41
- solve
  - triangular, 175–178
- Spark webpage, 75
- spectral radius
  - definition, 217
- spectrum
  - definition, 217
- submultiplicative matrix norm, 31
  - definition, 31
- sup, 28
- supremum, 28
- transpose, 5–8
- matrix, 6
- vector, 6
- triangle inequality, 25
- triangular solve, 175
  - lower, 175–178
  - upper, 178
- triangular system solve, 178
- 2-norm
  - definition, 25
  - matrix, 30
  - vector, 25–26
- Greek letters
  - it basis vector, 4
- upper triangular solve, 178
- vector
  - 2-norm (length)
    - definition, 25
  - complex conjugate, 5
  - conjugate, 5
  - Hermitian transpose, 6
  - infinity norm, 27
  - $\infty$ -norm, 27
  - $\infty$ -norm, 27
    - definition, 27
  - 1-norm, 27
    - definition, 27
  - orthogonal, 37
  - orthonormal, 37
  - p-norm, 27
  - perpendicular, 37
  - scaling
    - cost, 10
  - transpose, 6
  - 2-norm (length), 25–26
- vector addition
  - scaled, 10
- vector norm, 23–33
  - definition, 25
- vector-vector operation, 9
  - axpy, 10
  - dot, 11
  - scal, 9