Matrix analysis and applied linear algebra pdf free version free online

Matrix analysis and applied linear algebra pdf free version free online

Thank you for your participation! File loading please wait... Contents Preface . 1. Linear Equations . . . . 1 1.1 1.2 1.3 1.4 1.5 1.6 2. . . . . Row Echelon Form and Rank . Reduced Row Echelon Form . Consistency of Linear Systems Homogeneous Systems . . Nonhomogeneous Systems . . Electrical Circuits . . . . From Ancient China to Arthur Cayley Addition and Transposition . . . Linearity . . . . Why Do It This Way . . . . Matrix Multiplication . . . . Properties of Matrix Multiplication . Matrix Inversion . . . . . . Inverses of Sums and Sensitivity . . Elementary Matrices and Equivalence The LU Factorization . . . . 1 3 15 18 21 33 41 . . . . Matrix Algebra . . . . 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 3.10 4. Introduction . . . . Gaussian Elimination and Matrices . Gauss–Jordan Method . . . . Two-Point Boundary Value Problems Making Gaussian Elimination Work . Ill-Conditioned Systems . . . . Rectangular Systems and Echelon Forms . . . 2.1 2.2 2.3 2.4 2.5 2.6 3. ix 41 47 53 57 64 73 79 . . . . 79 81 89 93 95 105 115 124 131 141 Vector Spaces . . . . 159 4.1 4.2 4.3 4.4 Spaces and Subspaces . . Four Fundamental Subspaces Linear Independence . . . . Basis and Dimension . . . . 159 169 181 194 vi Contents 4.5 4.6 4.7 4.8 4.9 5. . . . . Vector Norms . . . . Matrix Norms . . . . Inner-Product Spaces . . . . Orthogonal Vectors . . . . . Gram–Schmidt Procedure . . Unitary and Orthogonal Matrices Orthogonal Reduction . . . . . Discrete Fourier Transform . . . Complementary Subspaces . . . Range-Nullspace Decomposition Orthogonal Decomposition . . . . Singular Value Decomposition . Orthogonal Projection . . . . . Why Least Squares? . . . . . . Angles between Subspaces . . . . 210 223 238 251 259 . . 269 . . . . 269 279 286 294 307 320 341 356 383 394 403 411 429 446 450 Determinants . . . . 459 6.1 6.2 7. . . . . . . Norms, Inner Products, and Orthogonality 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9 5.10 5.11 5.12 5.13 5.14 5.15 6. More about Rank . . . . . . Classical Least Squares . . . Linear Transformations . . . Change of Basis and Similarity Invariant Subspaces . . . . . Determinants . . . . Additional Properties of Determinants . . . . . . 459 475 Eigenvalues and Eigenvectors . . . . 489 7.1 7.2 7.3 7.4 7.5 7.6 7.7 7.8 7.9 Elementary Properties of Eigensystems . . . Diagonalization by Similarity Transformations Functions of Diagonalizable Matrices . . . . Systems of Differential Equations . . . . . Normal Matrices . . . . Positive Definite Matrices . . . . Nilpotent Matrices and Jordan Structure . Jordan Form . . . . Functions of Nondiagonalizable Matrices . . . . 489 505 525 541 547 558 574 587 599 Contents vii 7.10 7.11 8. Difference Equations, Limits, and Summability . . Minimum Polynomials and Krylov Methods . . . Perron–Frobenius Theory 8.1 8.2 8.3 8.4 . . . . 661 Introduction . . . . Positive Matrices . . . . Nonnegative Matrices . . . . Stochastic Matrices and Markov Chains Index 616 642 . . . . 661 663 670 687 . . . . 705 Preface Scaffolding Reacting to criticism concerning the lack of motivation in his writings, Gauss remarked that architects of great cathedrals do not obscure the beauty of their work by leaving the scaffolding in place after the construction has been completed. His philosophy epitomized the formal presentation and teaching of mathematics throughout the nineteenth and twentieth centuries, and it is still commonly found in mid-to-upper-level mathematics textbooks. The inherent efficiency and natural beauty of mathematics are compromised by straying too far from Gauss's viewpoint. But, as with most things in life, appreciation is generally preceded by some understanding seasoned with a bit of maturity, and in mathematics this comes from seeing some of the scaffolding. Purpose, Gap, and Challenge The purpose of this text is to present the contemporary theory and applications of linear algebra to university students studying mathematics, engineering, or applied science at the postcalculus level. Because linear algebra is usually encountered between basic problem solving courses such as calculus or differential equations and more advanced courses that require students to cope with mathematical rigors, the challenge in teaching applied linear algebra is to expose some of the scaffolding while conditioning students to appreciate the utility and beauty of the subject. Effectively meeting this challenge and bridging the inherent gaps between basic and more advanced mathematics are primary goals of this book. Rigor and Formalism To reveal portions of the scaffolding, narratives, examples, and summaries are used in place of the formal definition–theorem–proof development. But while well-chosen examples can be more effective in promoting understanding than rigorous proofs, and while precious classroom minutes cannot be squandered on theoretical details, I believe that all scientifically oriented students should be exposed to some degree of mathematical thought, logic, and rigor. And if logic and rigor are to reside anywhere, they have to be in the textbook. So even when logic and rigor are not the primary thrust, they are always available. Formal definition–theorem–proof designations are not used, but definitions, theorems, and proofs nevertheless exist, and they become evident as a student's maturity increases. A significant effort is made to present a linear development that avoids forward references, circular arguments, and dependence on prior knowledge of the subject. This results in some inefficiencies—e.g., the matrix 2-norm is presented x Preface before eigenvalues or singular values are thoroughly discussed. To compensate, I try to provide enough "wiggle room" so that an instructor can temper the inefficiencies by tailoring the approach to the students' prior background. Comprehensiveness and Flexibility A rather comprehensive treatment of linear algebra and its applications is presented and, consequently, the book is not meant to be devoured cover-to-cover in a typical one-semester course. However, the presentation is structured to provide flexibility in topic selection so that the text can be easily adapted to meet the demands of different course outlines without suffering breaks in continuity. Each section contains basic material paired with straightforward explanations, examples, and exercises. But every section also contains a degree of depth coupled with thought-provoking examples and exercises that can take interested students to a higher level. The exercises are formulated not only to make a student think about material from a current section, but they are designed also to pave the way for ideas in future sections in a smooth and often transparent manner. The text accommodates a variety of presentation levels by allowing instructors to select sections, discussions, examples, and exercises of appropriate sophistication. For example, traditional one-semester undergraduate courses can be taught from the basic material in Chapter 1 (Linear Equations); Chapter 2 (Rectangular Systems and Echelon Forms); Chapter 3 (Matrix Algebra); Chapter 4 (Vector Spaces); Chapter 5 (Norms, Inner Products, and Orthogonality); Chapter 6 (Determinants); and Chapter 7 (Eigenvalues and Eigenvectors). The level of the course and the degree of rigor are controlled by the selection and depth of coverage in the latter sections of Chapters 4, 5, and 7. An upper-level course might consist of a quick review of Chapters 1, 2, and 3 followed by a more in-depth treatment of Chapters 4, 5, and 7. For courses containing advanced undergraduate or graduate students, the focus can be on material in the latter sections of Chapters 4, 5, 7, and Chapter 8 (Perron–Frobenius Theory of Nonnegative Matrices). A rich two-semester course can be taught by using the text in its entirety. What Does "Applied" Mean? Most people agree that linear algebra is at the heart of applied science, but there are divergent views concerning what "applied linear algebra" really means; the academician's perspective is not always the same as that of the practitioner. In a poll conducted by SIAM in preparation for one of the triannual SIAM conferences on applied linear algebra, a diverse group of internationally recognized scientific corporations and government laboratories was asked how linear algebra finds application in their missions. The overwhelming response was that the primary use of linear algebra in applied industrial and laboratory work involves the development, analysis, and implementation of numerical algorithms along with some discrete and statistical modeling. The applications in this book tend to reflect this realization. While most of the popular "academic" applications are included, and "applications" to other areas of mathematics are honestly treated, Preface xi there is an emphasis on numerical issues designed to prepare students to use linear algebra in scientific environments outside the classroom. Computing Projects Computing projects help solidify concepts, and I include many exercises that can be incorporated into a laboratory setting. But my goal is to write a mathematics text that can last, so I don't muddy the development by marrying the material to a particular computer package or language. I am old enough to remember what happened to the FORTRAN- and APL-based calculus and linear algebra texts that came to market in the 1970s. I provide instructors with a flexible environment that allows for an ancillary computing laboratory in which any number of popular packages and lab manuals can be used in conjunction with the material in the text. History Finally, I believe that revealing only the scaffolding without teaching something about the scientific architects who erected it deprives students of an important part of their mathematical heritage. It also tends to dehumanize mathematics, which is the epitome of human endeavor. Consequently, I make an effort to say things (sometimes very human things that are not always complimentary) about the lives of the people who contributed to the development and applications of linear algebra. But, as I came to realize, this is a perilous task because writing history is frequently an interpretation of facts rather than a statement of facts. I considered documenting the sources of the historical remarks to help mitigate the inevitable challenges, but it soon became apparent that the sheer volume required to do so would skew the direction and flavor of the text. I can only assure the reader that I made an effort to be as honest as possible, and I tried to corroborate "facts." Nevertheless, there were times when interpretations had to be made, and these were no doubt influenced by my own views and experiences. Supplements Included with this text is a solutions manual and a CD-ROM. The solutions manual contains the solutions for each exercise given in the book. The solutions are constructed to be an integral part of the learning process. Rather than just providing answers, the solutions often contain details and discussions that are intended to stimulate thought and motivate material in the following sections. The CD, produced by Vickie Kearn and the people at SIAM, contains the entire book along with the solutions manual in PDF format. This electronic version of the text is completely searchable and linked. With a click of the mouse a student can jump to a referenced page, equation, theorem, definition, or proof, and then jump back to the sentence containing the reference, thereby making learning quite efficient. In addition, the CD contains material that extends historical remarks in the book and brings them to life with a large selection of xii Preface portraits, pictures, attractive graphics, and additional anecdotes. The supporting Internet site at MatrixAnalysis.com contains updates, errata, new material, and additional supplements as they become available. SIAM I thank the SIAM organization and the people who constitute it (the infrastructure as well as the general membership) for allowing me the honor of publishing my book under their name. I am dedicated to the goals, philosophy, and ideals of SIAM, and there is no other company or organization in the world that I would rather have publish this book. In particular, I am most thankful to Vickie Kearn, publisher at SIAM, for the confidence, vision, and dedication she has continually provided, and I am grateful for her patience that allowed me to write the book that I wanted to write. The talented people on the SIAM staff went far above and beyond the call of ordinary duty to make this project special. This group includes Lois Sellers (art and cover design), Michelle Montgomery and Kathleen LeBlanc (promotion and marketing), Marianne Will and Deborah Poulson (copy for CD-ROM biographies), Laura Helfrich and David Comdico (design and layout of the CD-ROM), Kelly Cuomo (linking the CDROM), and Kelly Thomas (managing editor for the book). Special thanks goes to Jean Anderson for her eagle-sharp editor's eye. Acknowledgments This book evolved over a period of several years through many different courses populated by hundreds of undergraduate and graduate students. To all my students and colleagues who have offered suggestions, corrections, criticisms, or just moral support, I offer my heartfelt thanks, and I hope to see as many of you as possible at some point in the future so that I can convey my feelings to you in person. I am particularly indebted to Michele Benzi for conversations and suggestions that led to several improvements. All writers are influenced by people who have written before them, and for this writer these include (in no particular order) Gil Strang, Jim Ortega, Charlie Van Loan, Leonid Mirsky, Ben Noble, Pete Stewart, Gene Golub, Charlie Johnson, Roger Horn, Peter Lancaster, Paul Halmos, Franz Hohn, Nick Rose, and Richard Bellman—thanks for lighting the path. I want to offer particular thanks to Richard J. Painter and Franklin A. Graybill, two exceptionally fine teachers, for giving a rough Colorado farm boy a chance to pursue his dreams. Finally, neither this book nor anything else I have done in my career would have been possible without the love, help, and unwavering support from Bethany, my friend, partner, and wife. Her multiple readings of the manuscript and suggestions were invaluable. I dedicate this book to Bethany and our children, Martin and Holly, to our granddaughter, Margaret, and to the memory of my parents, Carl and Louise Meyer. Carl D. Meyer April 19, 2000 CHAPTER 1 Linear Equations 1.1 INTRODUCTION A fundamental problem that surfaces in all mathematical sciences is that of analyzing and solving m algebraic equations in n unknowns. The study of a system of simultaneous linear equations is in a natural and indivisible alliance with the study of the rectangular array of numbers defined by the coefficients of the equations. This link seems to have been made at the outset. The earliest recorded analysis of simultaneous equations is found in the ancient Chinese book Chiu-chang Suan-shu (Nine Chapters on Arithmetic), estimated to have been written some time around 200 B.C. In the beginning of Chapter VIII, there appears a problem of the following form. Three sheafs of a good crop, two sheafs of a mediocre crop, and one sheaf of a bad crop are sold for 39 dou. Two sheafs of good, three mediocre, and one bad are sold for 34 dou; and one good, two mediocre, and three bad are sold for 26 dou. What is the price received for each sheaf of a good crop, each sheaf of a mediocre crop, and each sheaf of a bad crop? Today, this problem would be formulated as three equations in three unknowns by writing 3x + 2y + z = 39, 2x + 3y + z = 34, x + 2y + 3z = 26, where x, y, and z represent the price for one sheaf of a good, mediocre, and bad crop, respectively. The Chinese saw right to the heart of the matter. They placed the coefficients (represented by colored bamboo rods) of this system in 2 Chapter 1 Linear Equations a square array on a "counting board" and then manipulated the lines of the array according to prescribed rules of thumb. Their counting board techniques and rules of thumb found their way to Japan and eventually appeared in Europe with the colored rods having been replaced by numerals and the counting board replaced by pen and paper. In Europe, the technique became known as Gaussian elimination in honor of the German mathematician Carl Gauss, whose extensive use of it popularized the method. Because this elimination technique is fundamental, we begin the study of our subject by learning how to apply this method in order to compute solutions for linear equations. After the computational aspects have been mastered, we will turn to the more theoretical facets surrounding linear systems. 1 Carl Friedrich Gauss (1777–1855) is considered by many to have been the greatest mathematician who has ever lived, and his astounding career requires several volumes to document. He was referred to by his peers as the "prince of mathematicians." Upon Gauss's death one of them wrote that "His mind penetrated into the deepest secrets of numbers, space, and nature; He measured the course of the stars, the form and forces of the Earth; He carried within himself the evolution of mathematical sciences of a coming century." History has proven this remark to be true. 1.2 Gaussian Elimination and Matrices 3 1.2 GAUSSIAN ELIMINATION AND MATRICES The problem is to calculate, if possible, a common solution for a system of m linear algebraic equations in n unknowns a11 x1 + a12 x2 + · · · + a1n xn = b1 , a21 x1 + a22 x2 + · · · + a2n xn = b2 , . . . am1 x1 + am2 x2 + · · · + amn xn = bm , where the xi 's are the unknowns and the aij 's and the bi 's are known constants. The aij 's are called the coefficients of the system, and the set of bi 's is referred to as the right-hand side of the system. For any such system, there are exactly three possibilities for the set of solutions. Three Possibilities • UNIQUE SOLUTION: There is one and only one set of values for the xi 's that satisfies all equations simultaneously. • NO SOLUTION: There is no set of values for the xi 's that satisfies all equations simultaneously—the solution set is empty. • INFINITELY MANY SOLUTIONS: There are infinitely many different sets of values for the xi 's that satisfy all equations simultaneously. It is not difficult to prove that if a system has more than one solution, then it has infinitely many solutions. For example, it is impossible for a system to have exactly two different solutions. Part of the job in dealing with a linear system is to decide which one of these three possibilities is true. The other part of the task is to compute the solution if it is unique or to describe the set of all solutions if there are many solutions. Gaussian elimination is a tool that can be used to accomplish all of these goals. Gaussian elimination is a methodical process of systematically transforming one system into another simpler, but equivalent, system (two systems are called equivalent if they possess equal solution sets) by successively eliminating unknowns and eventually arriving at a system that is easily solvable. The elimination process relies on three simple operations by which to transform one system to another equivalent system. To describe these operations, let Ek denote the k th equation Ek : ak1 x1 + ak2 x2 + · · · + akn xn = bk 4 Chapter 1 Linear Equations and write the system as ⎧ E1 ⎫ ⎪ E2 ⎪ S= . . . ⎩ Em ⎭ For a linear system S , each of the following three elementary operations results in an equivalent system S′. (1) Interchange the ith and j th equations. That is, if ... then S′ = ... (1.2.1) (2) Replace the ith equation by a nonzero multiple of itself. That is, ... where α ≠ 0. (1.2.2) (3) Replace the j th equation by a combination of itself plus a multiple of the ith equation. That is, ... (1.2.3) 1.2 Gaussian Elimination and Matrices 5 Providing explanations for why each of these operations cannot change the solution set is left as an exercise. The most common problem encountered in practice is the one in which there are n equations in n unknowns—called a square system—for which there is a unique solution. Since Gaussian elimination is straightforward for this case, we begin here and later discuss the other possibilities. What follows is a detailed description of Gaussian elimination as applied to the following simple (but typical) square system: 2x + y + z = 1, 6x + 2y + z = −1, −2x + 2y + z = 7. (1.2.4) At each step, the strategy is to focus on one position, called the pivot position, and to eliminate all terms below this position using the three elementary operations. The coefficient in the pivot position is called a pivotal element (or simply a pivot), while the equation in which the pivot lies is referred to as the pivotal equation. Only nonzero numbers are allowed to be pivots. If a coefficient in a pivot position is ever 0, then the pivotal equation is interchanged with an equation below the pivotal equation to produce a nonzero pivot. (This is always possible for square systems possessing a unique solution.) Unless it is 0, the first coefficient of the first equation is taken as the first pivot. For example, the circled 2 in the system below is the pivot for the first step: ② x + y + z = 1, 6x + 2y + z = −1, −2x + 2y + z = 7. Step 1. Eliminate all terms below the first pivot. • Subtract three times the first equation from the second so as to produce the equivalent system: 2x + y + z = 1, − y − 2z = −4 (E2 − 3E1), −2x + 2y + z = 7. • Add the first equation to the third equation to produce the equivalent system: 2x + y + z = 1, − y − 2z = −4, 3y + 2z = 8 (E3 + E1). 6 Chapter 1 Linear Equations Step 2. Select a new pivot. • For the time being, select a new pivot by moving down and to the right. If this coefficient is not 0, then it is the next pivot. Otherwise, interchange with an equation below this position so as to bring a nonzero number into this pivotal position. In our example, −1 is the second pivot as identified below: 2x + y + z = 1, ⊖1 y − 2z = −4, 3y + 2z = 8. Step 3. Eliminate all terms below the second pivot. • Add three times the second equation to the third equation so as to produce the equivalent system: 2x + y + z = 1, − y − 2z = −4, −4z = −4 (E3 + 3E2). In general, at each step you move down and to the right to select the next pivot, then eliminate all terms below the pivot until you can no longer proceed. In this example, the third pivot is −4, but since there is nothing below the third pivot to eliminate, the process is complete. At this point, we say that the system has been triangularized. A triangular system is easily solved by a simple method known as back substitution in which the last equation is solved for the value of the last unknown and then substituted back into the penultimate equation, which is in turn solved for the penultimate unknown, etc., until each unknown has been determined. For our example, solve the last equation to obtain z = 1. Substitute z = 1 back into the second equation to obtain −y − 2(1) = −4 =⇒ y = 2. The strategy of selecting pivots in numerical computation is usually a bit more complicated than simply using the next coefficient that is down and to the right. Use the down-and-right strategy for now, and later more practical strategies will be discussed. 1.2 Gaussian Elimination and Matrices 7 Finally, substitute z = 1 and y = 2 back into the first equation 2x + y + z = 1 to get x = −1. The solution is −1, 2, 1. It should be clear that there is no reason to write down the symbols such as "x," "y," "z," and "=" at each step since we are only manipulating the coefficients. If such symbols are discarded, then a system of linear equations reduces to a rectangular array of numbers in which each horizontal line represents one equation. For example, the system in (1.2.4) reduces to the following array: ⎛ 2 1 1 1 ⎞ ⎜ 6 2 1 −1 ⎟ ⎝ −2 2 1 7 ⎠. (The line emphasizes where = appeared.) The array of coefficients—the numbers on the left-hand side of the vertical line—is called the coefficient matrix for the system. The entire array—the coefficient matrix augmented by the numbers from the right-hand side of the system—is called the augmented matrix associated with the system. If the coefficient matrix is denoted by A and the right-hand side is denoted by b , then the augmented matrix associated with the system is denoted by [A|b]. Formally, a scalar is either a real number or a complex number, and a matrix is a rectangular array of scalars. It is common practice to use uppercase boldface letters to denote matrices and to use the corresponding lowercase letters with two subscripts to denote individual entries in a matrix. For example, A = ⎛ a11 a12 · · · a1n ⎞ ⎜ a21 a22 · · · a2n ⎟ ... ⎝ am1 am2 · · · amn ⎠. The first subscript on an individual entry in a matrix designates the row (the horizontal line), and the second subscript denotes the column (the vertical line) that the entry occupies. For example, if A = ⎛ 2 1 3 4 ⎞ ⎜ 8 6 5 −9 ⎟, then a11 = 2, a12 = 1, . . . , a34 = 7. (1.2.5) ⎝ −3 8 3 7 ⎠ A submatrix of a given matrix A is an array obtained by deleting any combination of rows and columns from A. For example, B = −3 7 ... is a submatrix of the matrix A in (1.2.5) because B is the result of deleting the second row and the second and third columns of A. 8 Chapter 1 Linear Equations Matrix A is said to have shape or size m × n—pronounced "m by n"—whenever A has exactly m rows and n columns. For example, the matrix in (1.2.5) is a 3 × 4 matrix. By agreement, 1 × 1 matrices are identified with scalars and vice versa. To emphasize that matrix A has shape m × n, subscripts are sometimes placed on A as Am×n. Whenever m = n (i.e., when A has the same number of rows as columns), A is called a square matrix. Otherwise, A is said to be rectangular. Matrices consisting of a single row or a single column are often called row vectors or column vectors, respectively. The symbol Ai∗ is used to denote the ith row, while A∗j denotes the j th column of matrix A . For example, if A is the matrix in (1.2.6), then A2∗ = (8 6 5 −9) and A∗2 = (1 6 8). For a linear system of equations a11 x1 + a12 x2 + · · · + a1n xn = b1 , a21 x1 + a22 x2 + · · · + a2n xn = b2 , . . . am1 x1 + am2 x2 + · · · + amn xn = bm , Gaussian elimination can be executed on the associated augmented matrix [A|b] by performing elementary operations to the rows of [A|b]. These row operations correspond to the three elementary operations (1.2.1), (1.2.2), and (1.2.3) used to manipulate linear systems. For an m × n matrix M = ⎛ M1∗ ⎞ ... ⎝ Mm∗ ⎠, the three types of elementary row operations on M are as follows. • Type I: Interchange rows i and j to produce ... (1.2.7) • Type II: Replace row i by a nonzero multiple of itself to produce ... where α ≠ 0. (1.2.8) • Type III: Replace row j by a combination of itself plus a multiple of row i to produce ... (1.2.9) 1.2 Gaussian Elimination and Matrices 9 To solve the system (1.2.4) by using elementary row operations, start with the associated augmented matrix [A|b] and triangularize the coefficient matrix A by performing exactly the same sequence of row operations that corresponds to the elementary operations executed on the equations themselves: ⎛ ② 1 1 1 ⎞ ⎛ 2 1 1 1 ⎞ ⎜ 6 2 1 −1 ⎟ R2 − 3R1 −→ ⎜ 0 ⊖1 −2 −4 ⎟ ⎝ −2 2 1 7 ⎠ R3 + R1 ⎝ 0 3 2 8 ⎠ R3 + 3R2 ⎛ 2 1 1 1 ⎞ −→ ⎜ 0 −1 −2 −4 ⎟ ⎝ 0 0 −4 −4 ⎠. The final array represents the triangular system 2x + y + z = 1, − y − 2z = −4, − 4z = −4 that is solved by back substitution as described earlier. In general, if an n × n system has been triangularized to the form ⎛ t11 t12 · · · t1n c1 ⎞ ⎜ 0 t22 · · · t2n c2 ⎟ ... ⎝ 0 0 · · · tnn cn ⎠ (1.2.10) in which each tii ≠ 0 (i.e., there are no zero pivots), then the general algorithm for back substitution is as follows. 10 Chapter 1 Linear Equations Algorithm for Back Substitution Determine the xi 's from (1.2.10) by first setting xn = cn /tnn and then recursively computing xi = 1/tii (ci − ti,i+1 xi+1 − ti,i+2 xi+2 − · · · − tin xn) for i = n − 1, n − 2, . . . , 2, 1. One way to gauge the efficiency of an algorithm is to count the number of arithmetical operations required. For a variety of reasons, no distinction is made between additions and subtractions, and no distinction is made between multiplications and divisions. Furthermore, multiplications/divisions are usually counted separately from additions/subtractions. Even if you do not work through the details, it is important that you be aware of the operational counts for Gaussian elimination with back substitution so that you will have a basis for comparison when other algorithms are encountered. Gaussian Elimination Operation Counts Gaussian elimination with back substitution applied to an n × n system requires n3 n + n2 − multiplications/divisions 3 3 and n3 n2 5n + − additions/subtractions. 3 2 6 As n grows, the n3 /3 term dominates each of these expressions. Therefore, the important thing to remember is that Gaussian elimination with back substitution on an n × n system requires about n3 /3 multiplications/divisions and about the same number of additions/subtractions. 3 Operation counts alone may no longer be as important as they once were in gauging the efficiency of an algorithm. Older computers executed instructions sequentially, whereas some contemporary machines are capable of executing instructions in parallel so that different numerical tasks can be performed simultaneously. An algorithm that lends itself to parallelism may have a higher operational count but might nevertheless run faster on a parallel machine than an algorithm with a lesser operational count that cannot take advantage of parallelism. 1.2 Gaussian Elimination and Matrices 11 Example 1.2.1 Problem: Solve the following system using Gaussian elimination with back substitution: v − w = 3, −2u + 4v − w = 1, −2u + 5v − 4w = − 2. Solution: The associated augmented matrix is ⎛ 0 1 −1 ⎜ −2 4 −1 ⎝ −2 5 −4 ⎞3 1 ⎠. −2 Since the first pivotal position contains 0, interchange rows one and two before eliminating below the first pivot: ⎛ ⎞⎛ -2 4 3 Interchange R1 and R2 ⎜ −2 4 −1 ⎝ 0 1 1 ⎠−−−−−−− −→ −2 −2 5 −4 −2 5 ⎛ ⎛ −2 4 −1 −2 1 1 −→ ⎜ 0 −1 3 ⎠ −→ ⎜ 0 0 1 −3 0 −3 R3 − R2 0 1 −1 ⎠ 1 3 ⎠−2 R3 − R1 ⎞4 −1 1 3 ⎠. 1 −1 0 −2 −6 −1 −1 −4 Back substitution yields −6 = 3, −2 v = 3 + w = 3 + 3 = 6, 1 1 u= (1 − 4v + w) = (1 − 24 + 3) = 10. −2 −2 w= Exercises for section 1.2 1.2.1. Use Gaussian elimination with back substitution to solve the following system: x1 + x2 + x3 = 1, x1 + 2x2 + 2x3 = 1, x1 + 2x2 + 3x3 = 1. 12 Chapter 1 Linear Equations 1.2.2. Apply Gaussian elimination with back substitution to the following system: 2x1 − x2 = 0, −x1 + 2x2 − x3 = 0, −x2 + x3 = 1. 1.2.3. Use Gaussian elimination with back substitution to solve the following system: 4x2 − 3x3 = 3, −x1 + 7x2 − 5x3 = 4, −x1 + 8x2 − 6x3 = 5. 1.2.4. Solve the following system: x1 + x2 + x3 + x4 = 1, x1 + x2 + 3x3 + 3x4 = 3, x1 + x2 + 2x3 + 3x4 = 3, x1 + 3x2 + 3x3 + 3x4 = 4. 1.2.5. Consider the following three systems where the coefficients are the same for each system, but the right-hand sides are different (this situation occurs frequently): 4x − 8y + 5z = 1 0 0, 4x − 7y + 4z = 0 1 0, 3x − 4y + 2z = 0 0 1. Solve all three systems at one time by performing Gaussian elimination on an augmented matrix of the form A b1 b2 b3 . 1.2.6. Suppose that matrix B is obtained by performing a sequence of row operations on matrix A . Explain why A can be obtained by performing row operations on B . 1.2.7. Find angles α, β, and γ such that 2 sin α − cos β + 3 tan γ = 3, 4 sin α + 2 cos β − 2 tan γ = 2, 6 sin α − 3 cos β + tan γ = 9, where 0 ≤ α ≤ 2π, 0 ≤ β ≤ 2π, and 0 ≤ γ < π. 1.2 Gaussian Elimination and Matrices 13 1.2.8. The following system has no solution: −x1 + 3x2 − 2x3 = 1, −x1 + 4x2 − 3x3 = 0, −x1 + 5x2 − 4x3 = 0. Attempt to solve this system using Gaussian elimination and explain what occurs to indicate that the system is impossible to solve. 1.2.9. Attempt to solve the system −x1 + 3x2 − 2x3 = 4, −x1 + 4x2 − 3x3 = 5, −x1 + 5x2 − 4x3 = 6, using Gaussian elimination and explain why this system must have infinitely many solutions. 1.2.10. By solving a 3 × 3 system, find the coefficients in the equation of the parabola y = α+βx+γx2 that passes through the points (1, 1), (2, 2), and (3, 0). 1.2.11. Suppose that 100 insects are distributed in an enclosure consisting of four chambers with passageways between them as shown below. #3 #4 #2 #1 At the end of one minute, the insects have redistributed themselves. Assume that a minute is not enough time for an insect to visit more than one chamber and that at the end of a minute 40% of the insects in each chamber have not left the chamber they occupied at the beginning of the minute. The insects that leave a chamber disperse uniformly among the chambers that are directly accessible from the one they initially occupied—e.g., from #3, half move to #2 and half move to #4. 14 Chapter 1 Linear Equations (a) If at the end of one minute there are 12, 25, 26, and 37 insects in chambers #1, #2, #3, and #4, respectively, determine what the initial distribution had to be. (b) If the initial distribution is 20, 20, 20, 40, what is the distribution at the end of one minute? 1.2.12. Show that the three types of elementary row operations discussed on p. 8 are not independent by showing that the interchange operation (1.2.7) can be accomplished by a sequence of the other two types of row operations given in (1.2.8) and (1.2.9). 1.2.13. Suppose that [A|b] is the augmented matrix associated with a linear system. You know that performing row operations on [A|b] does not change the solution of the system. However, no mention of column operations was ever made because column operations can alter the solution. (a) Describe the effect on the solution of a linear system when columns A∗j and A∗k are interchanged. (b) Describe the effect when column A∗j is replaced by αA∗j for α = 0. (c) Describe the effect when A∗j is replaced by A∗j + αA∗k . Hint: Experiment with a 2 × 2 or 3 × 3 system. 1.2.14. Consider the n × n Hilbert ⎛ 1 ⎜ 1 2 ⎜ ⎜ H= 1 3 ⎜ . ⎜ .. ⎜ ⎝ 1 n matrix defined by 1 2 1 3 ··· 1 n 1 3 1 4 ··· 1 n+1 1 4 1 5 ··· 1 n+2 . . .. ··· . . . 1 n+1 1 n+2 ··· 1 2n−1 ⎞ ⎟ ⎟ ⎟ ⎟ ⎟ ⎟ ⎠ Express the individual entries hij in terms of i and j. 1.2.15. Verify that the operation counts given in the text for Gaussian elimination with back substitution are correct for a general 3 × 3 system. If you are up to the challenge, try to verify these counts for a general n × n system. 1.2.16. Explain why a linear system can never have exactly two different solutions. Extend your argument to explain the fact that if a system has more than one solution, then it must have infinitely many different solutions. 1.3 Gauss–Jordan Method 1.3 15 GAUSS–JORDAN METHOD The purpose of this section is to introduce a variation of Gaussian elimination 4 that is known as the Gauss–Jordan method. The two features that distinguish the Gauss–Jordan method from standard Gaussian elimination are as follows. • At each step, the pivot element is forced to be 1. • At each step, all terms above the pivot as well as all terms below the pivot are eliminated. In other words, if ⎛ a11 ⎜ a21 ⎜ . ⎝ .. a12 a22 .. . ··· ··· ... a1n a2n .. . an1 an2 · · · ann ⎞ b1 b2 ⎟ .. ⎟ . ⎠ bn is the augmented matrix associated with a linear system, then elementary row operations are used to reduce this matrix to ⎛ ⎞ 1 0 ··· 0 s1 s2 ⎟ ⎜0 1 ··· 0 ⎟ . . . . . ⎜ ⎝ .. .. ⎠ 0 0 ··· 1 sn The solution then appears in the last column (i.e., xi = si) so that this procedure circumvents the need to perform back substitution. Example 1.3.1 Problem: Apply the Gauss–Jordan method to solve the following system: 2x1 + 2x2 + 6x3 = 4, 2x1 + x2 + 7x3 = 6, −2x1 − 6x2 − 7x3 = − 1. 4 Although there has been some confusion as to which Jordan should receive credit for this algorithm, it now seems clear that the method was in fact introduced by a geodesist named Wilhelm Jordan (1842–1899) and not by the more well known mathematician Marie Ennemond Camille Jordan (1838–1922), whose name is often mistakenly associated with the technique, but who is otherwise correctly credited with other important topics in matrix analysis, the "Jordan canonical form" being the most notable. Wilhelm Jordan was born in southern Germany, educated in Stuttgart, and was a professor of geodesy at the technical college in Karlsruhe. He was a prolific writer, and he introduced his elimination scheme in the 1888 publication Handbuch der Vermessungskunde. Interestingly, a method similar to W. Jordan's variation of Gaussian elimination seems to have been discovered and described independently by an obscure Frenchman named Clasen, who appears to have published only one scientific article, which appeared in 1888—the same year as W. Jordan's Handbuch appeared. 16 Chapter 1 Linear Equations Solution: The sequence of operations is indicated in parentheses and the pivots are circled. ⎛ ⎛ ⎞ ⎞ 2 1 2 6 4 R1 /2 1 3 2 ⎜ 2 1 7 −→ ⎜ 2 1 7 6 ⎟ 6 ⎠ R2 − 2R1 −2 −6 −7 −2 −6 −7 −1 −1 R3 + 2R1 ⎞ ⎛ 2 1 1 3 1 −→ ⎜ 0 2 ⎠ (−R2) −→ ⎜ 0 −1 3 0 0 −4 −1 ⎞ ⎛ 4 1 0 4 1 0 4 1 −→ ⎜ 0 −1 −2 ⎠ −→ ⎜ 0 1 −1 1 0 0 −5 0 0 −5 −R3 /5 ⎛ ⎞ 0 1 0 0 −→ ⎜ 0 1 0 −1 ⎟. 1 0 0 1 ⎞ ⎛ ⎞ ⎛ ⎞ x1 0 Therefore, the solution is ⎜ x2 ⎟ = ⎜ −1 ⎟. 1 x3 ⎛ 1 1 −1 −4 3 1 −1 ⎞2 R1 − R2 −2 ⎠3 R3 + 4R2 ⎞4 R1 − 4R3 −2 ⎠R2 + R3 1 On the surface it may seem that there is little difference between the Gauss– Jordan method and Gaussian elimination with back substitution because eliminating terms above the pivot with Gauss–Jordan seems equivalent to performing back substitution. But this is not correct. Gauss–Jordan requires more arithmetic than Gaussian elimination with back substitution. Gauss–Jordan Operation Counts For an n × n system, the Gauss–Jordan procedure requires n3 n2 + 2 2 multiplications/divisions and n3 n − 2 2 additions/subtractions. In other words, the Gauss–Jordan method requires about n3 /2 multiplications/divisions and about the same number of additions/subtractions. Recall from the previous section that Gaussian elimination with back substitution requires only about n3 /3 multiplications/divisions and about the same 1.3 Gauss–Jordan Method 17 number of additions/subtractions. Compare this with the n3 /2 factor required by the Gauss–Jordan method, and you can see that Gauss–Jordan requires about 50% more effort than Gaussian elimination with back substitution. For small systems of the textbook variety (e.g., n = 3), these comparisons do not show a great deal of difference. However, in practical work, the systems that are encountered can be quite large, and the difference between Gauss–Jordan and Gaussian elimination with back substitution can be significant. For example, if n = 100, then n3 /3 is about 333,333, while n3 /2 is 500,000, which is a difference of 166,667 multiplications/divisions as well as that many additions/subtractions. Although the Gauss–Jordan method is not recommended for solving linear systems that arise in practical applications, it does have some theoretical advantages. Furthermore, it can be a useful technique for tasks other than computing solutions to linear systems. We will make use of the Gauss–Jordan procedure when matrix inversion is discussed—this is the primary reason for introducing Gauss–Jordan. Exercises for section 1.3 1.3.1. Use the Gauss–Jordan method to solve the following system: 4x2 − 3x3 = 3, −x1 + 7x2 − 5x3 = 4, −x1 + 8x2 − 6x3 = 5. 1.3.2. Apply the Gauss–Jordan method to the following system: x1 + x2 + x3 + x4 = 1, x1 + 2x2 + 2x3 + 2x4 = 0, x1 + 2x2 + 3x3 + 3x4 = 0, x1 + 2x2 + 3x3 + 4x4 = 0. 1.3.3. Use the Gauss–Jordan method to solve the following three systems at the same time. 2x1 − x2 = 1 0 0, −x1 + 2x2 − x3 = 0 1 0, −x2 + x3 = 0 0 1. 1.3.4. Verify that the operation counts given in the text for the Gauss–Jordan method are correct for a general 3 × 3 system. If you are up to the challenge, try to verify these counts for a general n × n system. 18 1.4 Chapter 1 Linear Equations TWO-POINT BOUNDARY VALUE PROBLEMS It was stated previously that linear systems that arise in practice can become quite large in size. The purpose of this section is to understand why this often occurs and why there is frequently a special structure to the linear systems that come from practical applications. Given an interval [a, b] and two numbers α and β, consider the general problem of trying to find a function y(t) that satisfies the differential equation u(t)y (t)+v(t)y (t)+w(t)y(t) = f (t), where y(a) = α and y(b) = β. (1.4.1) The functions u, v, w, and f are assumed to be known functions on [a, b]. Because the unknown function y(t) is specified at the boundary points a and b, problem (1.4.1) is known as a two-point boundary value problem. Such problems abound in nature and are frequently very hard to handle because it is often not possible to express y(t) in terms of elementary functions. Numerical methods are usually employed to approximate y(t) at discrete points inside [a, b]. Approximations are produced by subdividing the interval [a, b] into n + 1 equal subintervals, each of length h = (b − a)/(n + 1) as shown below. h t0 = a h t1 = a + h h t2 = a + 2h ··· ··· tn = a + nh tn+1 = b Derivative approximations at the interior nodes (grid points) ti = a + ih are ∞ made by using Taylor series expansions y(t) = k=0 y (k) (ti)(t − ti)k /k! to write y (ti)h2 y (ti)h3 + + ···, 2! 3! y (ti)h2 y (ti)h3 y(ti − h) = y(ti) − y (ti)h + − + ···, 2! 3! y(ti + h) = y(ti) + y (ti)h + (1.4.2) and then subtracting and adding these expressions to produce y (ti) = y(ti + h) − y(ti − h) + O(h3) 2h y (ti) = y(ti − h) − 2y(ti) + y(ti + h) + O(h4), h2 and where O(hp) denotes 5 5 terms containing pth and higher powers of h. The Formally, a function f (h) is O(hp) if f (h)/hp remains bounded as h → 0, but f (h)/hq becomes unbounded if q > p. This means that f goes to zero as fast as hp goes to zero. 1.4 Two-Point Boundary Value Problems 19 resulting approximations y(ti +h) − y(ti −h) y(ti −h) − 2y(ti) + y(ti +h) and y (ti) ≈ (1.4.3) 2h h2 are called centered difference approximations, and they are preferred over less accurate one-sided approximations such as y (ti) ≈ y(ti + h) − y(ti) y(t) − y(t − h) or y (ti) ≈ . h h The value h = (b − a)/(n + 1) is called the step size. Smaller step sizes produce better derivative approximations, so obtaining an accurate solution usually requires a small step size and a large number of grid points. By evaluating the centered difference approximations at each grid point and substituting the result into the original differential equation (1.4.1), a system of n linear equations in n unknowns is produced in which the unknowns are the values y(ti). A simple example can serve to illustrate this point. y (ti) ≈ Example 1.4.1 Suppose that f (t) is a known function and consider the two-point boundary value problem y (t) = f (t) on [0, 1] with y(0) = y(1) = 0. The goal is to approximate the values of y at n equally spaced grid points ti interior to [0, 1]. The step size is therefore h = 1/(n + 1). For the sake of convenience, let yi = y(ti) and fi = f (ti). Use the approximation yi−1 − 2yi + yi+1 ≈ y (ti) = fi h2 along with y0 = 0 and yn+1 = 0 to produce the system of equations −yi−1 + 2yi − yi+1 ≈ −h2 fi for i = 1, 2, . . . , n. (The signs are chosen to make the 2's positive to be consistent with later developments.) The augmented matrix associated with this system is shown below: ⎛ ⎞2 −1 0 ··· 0 0 0 −h2 f1 2 2 −1 · · · 0 0 0 −h f2 ⎟ ⎜ −1 ⎟ ⎜ 2 ··· 0 0 0 −h2 f3 ⎟ ⎜ 0 −1 ⎟ . ⎜ . .. . ⎟ . ⎜ . ⎜ . . . . ⎟ ⎜ ⎟ ⎜0 0 0 ··· 2 −1 0 −h2 fn−2 ⎟ ⎜ ⎟ ⎝0 ⎠0 0 0 · · · −1 2 −1 −h2 fn−1 0 0 0 ··· 0 −1 2 −h2 fn By solving this system, approximate values of the unknown function y at the grid points ti are obtained. Larger values of n produce smaller values of h and hence better approximations to the exact values of the yi 's. 20 Chapter 1 Linear Equations Notice the pattern of the entries in the coefficient matrix in the above example. The nonzero elements occur only on the subdiagonal, main-diagonal, and superdiagonal lines—such a system (or matrix) is said to be tridiagonal. This is characteristic in the sense that when finite difference approximations are applied to the general two-point boundary value problem, a tridiagonal system is the result. Tridiagonal systems are particularly nice in that they are inexpensive to solve. When Gaussian elimination is applied, only two multiplications/divisions are needed at each step of the triangularization process because there is at most only one nonzero entry below and to the right of each pivot. Furthermore, Gaussian elimination preserves all of the zero entries that were present in the original tridiagonal system. This makes the back substitution process cheap to execute because there are at most only two multiplications/divisions required at each substitution step. Exercise 3.10.6 contains more details. Exercises for section 1.4 1.4.1. Divide the interval [0, 1] into five equal subintervals, and apply the finite difference method in order to approximate the solution of the two-point boundary value problem y (t) = 125t, y(0) = y(1) = 0 at the four interior grid points. Compare your approximate values at the grid points with the exact solution at the grid points. Note: You should not expect very accurate approximations with only four interior grid points. 1.4.2. Divide [0, 1] into n+1 equal subintervals, and apply the finite difference approximation method to derive the linear system associated with the two-point boundary value problem y (t) − y (t) = f (t), y(0) = y(1) = 0. 1.4.3. Divide [0, 1] into five equal subintervals, and approximate the solution to y (t) − y (t) = 125t, y(0) = y(1) = 0 at the four interior grid points. Compare the approximations with the exact values at the grid points. 1.5 Making Gaussian Elimination Work 1.5 21 MAKING GAUSSIAN ELIMINATION WORK Now that you understand the basic Gaussian elimination technique, it's time to turn it into a practical algorithm that can be used for realistic applications. For pencil and paper computations where you are doing exact arithmetic, the strategy is to keep things as simple as possible (like avoiding messy fractions) in order to minimize those "stupid arithmetic errors" we are all prone to make. But very few problems in the real world are of the textbook variety, and practical applications involving linear systems usually demand the use of a computer. Computers don't care about messy fractions, and they don't introduce errors of the "stupid" variety. Computers produce a more predictable kind of error, called 6 roundoff error, and it's important to spend a little time up front to understand this kind of error and its effects on solving linear systems. Numerical computation in digital computers is performed by approximating the infinite set of real numbers with a finite set of numbers as described below. Floating-Point Numbers A t -digit, base-β floating-point number has the form f = ±.d1 d2 · · · dt × β with d1 = 0, where the base β, the exponent , and the digits 0 ≤ di ≤ β − 1 are integers. For internal machine representation, β = 2 (binary representation) is standard, but for pencil-and-paper examples it's more convenient to use β = 10. The value of t, called the precision, and the exponent can vary with the choice of hardware and software. Floating-point numbers are just adaptations of the familiar concept of scientific notation where β = 10, which will be the value used in our examples. For any fixed set of values for t, β, and , the corresponding set F of floatingpoint numbers is necessarily a finite set, so some real numbers can't be found in F. There is more than one way of approximating real numbers with floatingpoint numbers. For the remainder of this text, the following common rounding convention is adopted. Given a real number x, the floating-point approximation f l(x) is defined to be the nearest element in F to x, and in case of a tie we round away from 0. This means that for t-digit precision with β = 10, we need 6 The computer has been the single most important scientific and technological development of our century and has undoubtedly altered the course of science for all future time. The prospective young scientist or engineer who passes through a contemporary course in linear algebra and matrix theory and fails to learn at least the elementary aspects of what is involved in solving a practical linear system with a computer is missing a fundamental tool of applied mathematics. 22 Chapter 1 Linear Equations to look at digit dt+1 in x = .d1 d2 · · · dt dt+1 · · · × 10 (making sure d1 = 0) and then set .d1 d2 · · · dt × 10 if dt+1 < 5, f l(x) = ([.d1 d2 · · · dt] + 10−t) × 10 if dt+1 ≥ 5. For example, in 2 -digit, base-10 floating-point arithmetic, f l (3/80) = f l(.0375) = f l(.375 × 10−1) = .38 × 10−1 = .038. By considering η = 1/3 and ξ = 3 with t -digit base-10 arithmetic, it's easy to see that f l(η + ξ) = f l(η) + f l(ξ) and f l(ηξ) = f l(η)f l(ξ). Furthermore, several familiar rules of real arithmetic do not hold for floatingpoint arithmetic—associativity is one outstanding example. This, among other reasons, makes the analysis of floating-point computation difficult. It also means that you must be careful when working the examples and exercises in this text because although most calculators and computers can be instructed to display varying numbers of digits, most have a fixed internal precision with which all calculations are made before numbers are displayed, and this internal precision cannot be altered. Almost certainly, the internal precision of your calculator or computer is greater than the precision called for by the examples and exercises in this text. This means that each time you perform a t-digit calculation, you should manually round the result to t significant digits and reenter the rounded number before proceeding to the next calculation. In other words, don't "chain" operations in your calculator or computer. To understand how to execute Gaussian elimination using floating-point arithmetic, let's compare the use of exact arithmetic with the use of 3-digit base-10 arithmetic to solve the following system: 47x + 28y = 19, 89x + 53y = 36. Using Gaussian elimination with exact arithmetic, we multiply the first equation by the multiplier m = 89/47 and subtract the result from the second equation to produce 47 28 19 . 1/47 0 −1/47 Back substitution yields the exact solution x=1 and y = −1. Using 3-digit arithmetic, the multiplier is 89 f l(m) = f l = .189 × 101 = 1.89. 47 1.5 Making Gaussian Elimination Work Since 23 f l f l(m)f l(47) = f l(1.89 × 47) = .888 × 102 = 88.8, f l f l(m)f l(28) = f l(1.89 × 28) = .529 × 102 = 52.9, f l f l(m)f l(19) = f l(1.89 × 19) = .359 × 102 = 35.9, the first step of 3-digit Gaussian elimination is as shown below: 47 28 19 f l(89 − 88.8) f l(53 − 52.9) f l(36 − 35.9) = 47 .2 28 .1 19 .1 . The goal is to triangularize the system—to produce a zero in the circled (2,1)-position—but this cannot be accomplished with 3-digit arithmetic. Unless .2 is replaced by 0, back substitution cannot be executed. the circled value Henceforth, we will agree simply to enter 0 in the position that we are trying to annihilate, regardless of the value of the floating-point number that might actually appear. The value of the position being annihilated is generally not even computed. For example, don't even bother computing f l 89 − f l f l(m)f l(47) = f l(89 − 88.8) = .2 in the above example. Hence the result of 3-digit Gaussian elimination for this example is 47 28 19 . 0 .1 .1 Apply 3-digit back substitution to obtain the 3-digit floating-point solution .1 y = fl = 1, .1 19 − 28 −9 x = fl = fl = −.191. 47 47 The vast discrepancy between the exact solution (1, −1) and the 3-digit solution (−.191, 1) illustrates some of the problems we can expect to encounter while trying to solve linear systems with floating-point arithmetic. Sometimes using a higher precision may help, but this is not always possible because on all machines there are natural limits that make extended precision arithmetic impractical past a certain point. Even if it is possible to increase the precision, it 24 Chapter 1 Linear Equations may not buy you very much because there are many cases for which an increase in precision does not produce a comparable decrease in the accumulated roundoff error. Given any particular precision (say, t), it is not difficult to provide examples of linear systems for which the computed t-digit solution is just as bad as the one in our 3-digit example above. Although the effects of rounding can almost never be eliminated, there are some simple techniques that can help to minimize these machine induced errors. Partial Pivoting At each step, search the positions on and below the pivotal position for the coefficient of maximum magnitude. If necessary perform the appropriate row interchange to bring this maximal coefficient into the pivotal position. Illustrated below is the third step in a typical case: ⎛ ∗ ∗ ⎜0 ∗ ⎜ ⎜0 0 ⎜ 0 0 0 0 ∗ ∗ S S S ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ⎞ ∗ ∗ ⎟ ⎟ ∗ ⎟ ∗ ⎠ ∗ ∗ Search the positions in the third column marked " S " for the coefficient of maximal magnitude and, if necessary, interchange rows to bring this coefficient into the circled pivotal position. Simply stated, the strategy is to maximize the magnitude of the pivot at each step by using only row interchanges. On the surface, it is probably not apparent why partial pivoting should make a difference. The following example not only shows that partial pivoting can indeed make a great deal of difference, but it also indicates what makes this strategy effective. Example 1.5.1 It is easy to verify that the exact solution to the system −10−4 x + y = 1, x + y = 2, is given by x= 1 1.0001 and y= 1.0002 . 1.0001 If 3-digit arithmetic without partial pivoting is used, then the result is 1.5 Making Gaussian Elimination Work −10−4 1 1 1 25 1 2 R2 + 104 R1 −→ −10−4 0 1 104 1 104 because f l(1 + 104) = f l(.10001 × 105) = .100 × 105 = 104 (1.5.1) f l(2 + 104) = f l(.10002 × 105) = .100 × 105 = 104 . (1.5.2) and Back substitution now produces x=0 and y = 1. Although the computed solution for y is close to the exact solution for y, the computed solution for x is not very close to the exact solution for x —the computed solution for x is certainly not accurate to three significant figures as you might hope. If 3-digit arithmetic with partial pivoting is used, then the result is −10−4 1 1 1 1 2 −→ 2 1 R2 + 10−4 R1 1 1 −10−4 1 2 1 1 −→ 0 1 1 because and f l(1 + 10−4) = f l(.10001 × 101) = .100 × 101 = 1 f l(1 + 2 × 10−4) = f l(.10002 × 101) = .100 × 101 = 1. (1.5.3) (1.5.4) This time, back substitution produces the computed solution x=1 and y = 1, which is as close to the exact solution as one can reasonably expect—the computed solution agrees with the exact solution to three significant digits. Why did partial pivoting make a difference? The answer lies in comparing (1.5.1) and (1.5.2) with (1.5.3) and (1.5.4). Without partial pivoting the multiplier is 104 , and this is so large that it completely swamps the arithmetic involving the relatively smaller numbers 1 and 2 and prevents them from being taken into account. That is, the smaller numbers 1 and 2 are "blown away" as though they were never present so that our 3-digit computer produces the exact solution to another system, namely, −10−4 1 1 , 1 0 0 26 Chapter 1 Linear Equations which is quite different from the original system. With partial pivoting the multiplier is 10−4 , and this is small enough so that it does not swamp the numbers 1 and 2. In this case, solution to the the 3-digit computer produces the exact 7 1 system 10 11 , which is close to the original system. 2 In summary, the villain in Example 1.5.1 is the large multiplier that prevents some smaller numbers from being fully accounted for, thereby resulting in the exact solution of another system that is very different from the original system. By maximizing the magnitude of the pivot at each step, we minimize the magnitude of the associated multiplier thus helping to control the growth of numbers that emerge during the elimination process. This in turn helps circumvent some of the effects of roundoff error. The problem of growth in the elimination procedure is more deeply analyzed on p. 348. When partial pivoting is used, no multiplier ever exceeds 1 in magnitude. To see that this is the case, consider the following two typical steps in an elimination procedure: ⎛ ∗ ∗ ⎜0 ∗ ⎜ ⎜0 0 ⎜ 0 0 0 0 ∗ ∗ p q r ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ⎛ ⎞ ∗ ∗ ∗ ∗ ⎟ ⎜0 ∗ ⎟ ⎜ ∗ ⎟ −→ ⎜ 0 0 ⎜ ⎠ 0 0 ∗ R4 − (q/p)R3 ∗ R5 − (r/p)R3 0 0 ∗ ∗ p 0 0 ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ⎞ ∗ ∗ ⎟ ⎟ ∗ ⎟ ∗ ⎠ ∗ ∗ The pivot is p, while q/p and r/p are the multipliers. If partial pivoting has been employed, then |p| ≥ |q| and |p| ≥ |r| so that q ≤1 p and r ≤ 1. p By guaranteeing that no multiplier exceeds 1 in magnitude, the possibility of producing relatively large numbers that can swamp the significance of smaller numbers is much reduced, but not completely eliminated. To see that there is still more to be done, consider the following example. Example 1.5.2 The exact solution to the system −10x + 105 y = 105 , x+ 7 y = 2, Answering the question, "What system have I really solved (i.e., obtained the exact solution of), and how close is this system to the original system," is called backward error analysis, as opposed to forward analysis in which one tries to answer the question, "How close will a computed solution be to the exact solution?" Backward analysis has proven to be an effective way to analyze the numerical stability of algorithms. 1.5 Making Gaussian Elimination Work 27 is given by 1 1.0002 and y = . 1.0001 1.0001 Suppose that 3-digit arithmetic with partial pivoting is used. Since | − 10| > 1, no interchange is called for and we obtain −10 105 105 105 −10 105 −→ 1 1 2 R2 + 10−1 R1 0 104 104 x= because f l(1 + 104) = f l(.10001 × 105) = .100 × 105 = 104 and f l(2 + 104) = f l(.10002 × 105) = .100 × 105 = 104 . Back substitution yields x=0 and y = 1, which must be considered to be very bad—the computed 3-digit solution for y is not too bad, but the computed 3-digit solution for x is terrible! What is the source of difficulty in Example 1.5.2? This time, the multiplier cannot be blamed. The trouble stems from the fact that the first equation contains coefficients that are much larger than the coefficients in the second equation. That is, there is a problem of scale due to the fact that the coefficients are of different orders of magnitude. Therefore, we should somehow rescale the system before attempting to solve it. If the first equation in the above example is rescaled to insure that the coefficient of maximum magnitude is a 1, which is accomplished by multiplying the first equation by 10−5 , then the system given in Example 1.5.1 is obtained, and we know from that example that partial pivoting produces a very good approximation to the exact solution. This points to the fact that the success of partial pivoting can hinge on maintaining the proper scale among the coefficients. Therefore, the second refinement needed to make Gaussian elimination practical is a reasonable scaling strategy. Unfortunately, there is no known scaling procedure that will produce optimum results for every possible system, so we must settle for a strategy that will work most of the time. The strategy is to combine row scaling—multiplying selected rows by nonzero multipliers—with column scaling—multiplying selected columns of the coefficient matrix A by nonzero multipliers. Row scaling doesn't alter the exact solution, but column scaling does—see Exercise 1.2.13(b). Column scaling is equivalent to changing the units of the k th unknown. For example, if the units of the k th unknown xk in [A|b] are millimeters, and if the k th column of A is multiplied by . 001, then the k th ˆ | b] is x unknown in the scaled system [A ˆi = 1000xi , and thus the units of the scaled unknown x ˆk become meters. 28 Chapter 1 Linear Equations Experience has shown that the following strategy for combining row scaling with column scaling usually works reasonably well. Practical Scaling Strategy 1. 2. Choose units that are natural to the problem and do not distort the relationships between the sizes of things. These natural units are usually self-evident, and further column scaling past this point is not ordinarily attempted. Row scale the system [A|b] so that the coefficient of maximum magnitude in each row of A is equal to 1. That is, divide each equation by the coefficient of maximum magnitude. Partial pivoting together with the scaling strategy described above makes Gaussian elimination with back substitution an extremely effective tool. Over the course of time, this technique has proven to be reliable for solving a majority of linear systems encountered in practical work. Although it is not extensively used, there is an extension of partial pivoting known as complete pivoting which, in some special cases, can be more effective than partial pivoting in helping to control the effects of roundoff error. Complete Pivoting If [A|b] is the augmented matrix at the k th step of Gaussian elimination, then search the pivotal position together with every position in A that is below or to the right of the pivotal position for the coefficient of maximum magnitude. If necessary, perform the appropriate row and column interchanges to bring the coefficient of maximum magnitude into the pivotal position. Shown below is the third step in a typical situation: ⎛ ∗ 0 ⎜ ⎜ ⎜0 ⎝ 0 0 ∗ ∗ 0 0 0 ∗ ∗ S S S ∗ ∗ S S S ∗ ∗ S S S ⎞ ∗ ∗ ⎟ ⎟ ∗ ⎟ ⎠ ∗ ∗ Search the positions marked " S " for the coefficient of maximal magnitude. If necessary, interchange rows and columns to bring this maximal coefficient into the circled pivotal position. Recall from Exercise 1.2.13 that the effect of a column interchange in A is equivalent to permuting (or renaming) the associated unknowns. 1.5 Making Gaussian Elimination Work 29 You should be able to see that complete pivoting should be at least as effective as partial pivoting. Moreover, it is possible to construct specialized examples where complete pivoting is superior to partial pivoting—a famous example is presented in Exercise 1.5.7. However, one rarely encounters systems of this nature in practice. A deeper comparison between no pivoting, partial pivoting, and complete pivoting is given on p. 348. Example 1.5.3 Problem: Use 3-digit arithmetic together with complete pivoting to solve the following system: x− y = −2, −9x + 10y = 12. Solution: Since 10 is the coefficient of maximal magnitude that lies in the search pattern, interchange the first and second rows and then interchange the first and second columns: 1 −1 −2 12 −9 10 −→ −9 10 1 −1 12 −2 10 −9 12 12 10 −9 −→ −→ . −1 1 0 .1 −2 −.8 The effect of the column interchange is to rename the unknowns to x ˆ and yˆ, where x ˆ = y and yˆ = x. Back substitution yields yˆ = −8 and x ˆ = −6 so that x = yˆ = −8 and y=x ˆ = −6. In this case, the 3-digit solution and the exact solution agree. If only partial pivoting is used, the 3-digit solution will not be as accurate. However, if scaled partial pivoting is used, the result is the same as when complete pivoting is used. If the cost of using complete pivoting was nearly the same as the cost of using partial pivoting, we would always use complete pivoting. However, it is not difficult to show that complete pivoting approximately doubles the cost over straight Gaussian elimination, whereas partial pivoting adds only a negligible amount. Couple this with the fact that it is extremely rare to encounter a practical system where scaled partial pivoting is not adequate while complete pivoting is, and it is easy to understand why complete pivoting is seldom used in practice. Gaussian elimination with scaled partial pivoting is the preferred method for dense systems (i.e., not a lot of zeros) of moderate size. 30 Chapter 1 Linear Equations Exercises for section 1.5 1.5.1. Consider the following system: 10−3 x − y = 1, x + y = 0. (a) Use 3-digit arithmetic with no pivoting to solve this system. (b) Find a system that is exactly satisfied by your solution from part (a), and note how close this system is to the original system. (c) Now use partial pivoting and 3-digit arithmetic to solve the original system. (d) Find a system that is exactly satisfied by your solution from part (c), and note how close this system is to the original system. (e) Use exact arithmetic to obtain the solution to the original system, and compare the exact solution with the results of parts (a) and (c). (f) Round the exact solution to three significant digits, and compare the result with those of parts (a) and (c). 1.5.2. Consider the following system: x+ y = 3, −10x + 10 y = 105 . 5 (a) Use 4-digit arithmetic with partial pivoting and no scaling to compute a solution. (b) Use 4-digit arithmetic with complete pivoting and no scaling to compute a solution of the original system. (c) This time, row scale the original system first, and then apply partial pivoting with 4-digit arithmetic to compute a solution. (d) Now determine the exact solution, and compare it with the results of parts (a), (b), and (c). 1.5.3. With no scaling, compute the 3-digit solution of −3x + y = −2, 10x − 3y = 7, without partial pivoting and with partial pivoting. Compare your results with the exact solution. 1.5 Making Gaussian Elimination Work 31 1.5.4. Consider the following system in which the coefficient matrix is the Hilbert matrix: 1 x+ y+ 2 1 1 x+ y+ 2 3 1 1 x+ y+ 3 4 1 1 z= , 3 3 1 1 z= , 4 3 1 1 z= . 5 5 (a) First convert the coefficients to 3-digit floating-point numbers, and then use 3-digit arithmetic with partial pivoting but with no scaling to compute the solution. (b) Again use 3-digit arithmetic, but row scale the coefficients (after converting them to floating-point numbers), and then use partial pivoting to compute the solution. (c) Proceed as in part (b), but this time row scale the coefficients before each elimination step. (d) Now use exact arithmetic on the original system to determine the exact solution, and compare the result with those of parts (a), (b), and (c). 1.5.5. To see that changing units can affect a floating-point solution, consider a mining operation that extracts silica, iron, and gold from the earth. Capital (measured in dollars), operating time (in hours), and labor (in man-hours) are needed to operate the mine. To extract a pound of silica requires $.0055, .0011 hours of operating time, and .0093 man-hours of labor. For each pound of iron extracted, $.095, .01 operating hours, and .025 man-hours are required. For each pound of gold extracted, $960, 112 operating hours, and 560 man-hours are required. (a) Suppose that during 600 hours of operation, exactly $5000 and 3000 man-hours are used. Let x, y, and z denote the number of pounds of silica, iron, and gold, respectively, that are recovered during this period. Set up the linear system whose solution will yield the values for x, y, and z. (b) With no scaling, use 3-digit arithmetic and partial pivoting to compute a solution (˜ x, y˜, z˜) of the system of part (a). Then approximate the exact solution (x, y, z) by using your machine's (or calculator's) full precision with partial pivoting to solve the system in part (a), and compare this with your 3-digit solution by computing the relative error defined by er = (x − x ˜)2 + (y − y˜)2 + (z − z˜)2 x2 + y 2 + z 2 . 32 Chapter 1 Linear Equations (c) Using 3-digit arithmetic, column scale the coefficients by changing units: convert pounds of silica to tons of silica, pounds of iron to half-tons of iron, and pounds of gold to troy ounces of gold (1 lb. = 12 troy oz.). (d) Use 3-digit arithmetic with partial pivoting to solve the column scaled system of part (c). Then approximate the exact solution by using your machine's (or calculator's) full precision with partial pivoting to solve the system in part (c), and compare this with your 3-digit solution by computing the relative error er as defined in part (b). 1.5.6. Consider the system given in Example 1.5.3. (a) Use 3-digit arithmetic with partial pivoting but with no scaling to solve the system. (b) Now use partial pivoting with scaling. Does complete pivoting provide an advantage over scaled partial pivoting in this case? 1.5.7. Consider the following well-scaled matrix: ⎛ 1 0 0 ··· 0 ⎜ −1 1 0 · · · 0 ⎜ .. ⎜ . ⎜ −1 −1 1 0 ⎜ . . . . Wn = ⎜ . . . . . ⎜ . . ⎜ −1 −1 −1 . 1 ⎜ ⎝ −1 −1 −1 · · · −1 −1 −1 −1 · · · −1 −1 ⎞ 1 1 ⎟ ⎟ 0 1 ⎟ .. .. ⎟ . . ⎟ ⎟ 0 1 ⎟ ⎟ 1 1 ⎠ −1 1 0 0 (a) Reduce Wn to an upper-triangular form using Gaussian elimination with partial pivoting, and determine the element of maximal magnitude that emerges during the elimination procedure. (b) Now use complete pivoting and repeat part (a). (c) Formulate a statement comparing the results of partial pivoting with those of complete pivoting for Wn , and describe the effect this would have in determining the t -digit solution for a system whose augmented matrix is [Wn | b]. 1.5.8. Suppose that A is an n × n matrix of real numbers that has been scaled so that each entry satisfies |aij | ≤ 1, and consider reducing A to triangular form using Gaussian elimination with partial pivoting. Demonstrate that after k steps of the process, no entry can have a magnitude that exceeds 2k . Note: The previous exercise shows that there are cases where it is possible for some elements to actually attain the maximum magnitude of 2k after k steps. 1.6 Ill-Conditioned Systems 33 ILL-CONDITIONED SYSTEMS Gaussian elimination with partial pivoting on a properly scaled system is perhaps the most fundamental algorithm in the practical use of linear algebra. However, it is not a universal algorithm nor can it be used blindly. The purpose of this section is to make the point that when solving a linear system some discretion must always be exercised because there are some systems that are so inordinately sensitive to small perturbations that no numerical technique can be used with confidence. Example 1.6.1 Consider the system .835x + .667y = .168, .333x + .266y = .067, for which the exact solution is x = 1 and y = −1. If b2 = .067 is only slightly perturbed to become ˆb2 = .066, then the exact solution changes dramatically to become x̂ = −666 and ŷ = 834. This is an example of a system whose solution is extremely sensitive to a small perturbation. This sensitivity is intrinsic to the system itself and is not a result of any numerical procedure. Therefore, you cannot expect some "numerical trick" to remove the sensitivity. If the exact solution is sensitive to small perturbations, then any computed solution cannot be less so, regardless of the algorithm used. Ill-Conditioned Linear Systems A system of linear equations is said to be ill-conditioned when some small perturbation in the system can produce relatively large changes in the exact solution. Otherwise, the system is said to be wellconditioned. It is easy to visualize what causes a 2 × 2 system to be ill-conditioned. Geometrically, two equations in two unknowns represent two straight lines, and the point of intersection is the solution for the system. An ill-conditioned system represents two straight lines that are almost parallel. 34 Chapter 1 Linear Equations If two straight lines are almost parallel and if one of the lines is tilted only slightly, then the point of intersection (i.e., the solution of the associated 2 × 2 linear system) is drastically altered. L' L Perturbed Solution Original Solution Figure 1.6.1 This is illustrated in Figure 1.6.1 in which line L is slightly perturbed to become line L'. Notice how this small perturbation results in a large change in the point of intersection. This was exactly the situation for the system given in Example 1.6.1. In general, ill-conditioned systems are those that represent almost parallel lines, almost parallel planes, and generalizations of these notions. Because roundoff errors can be viewed as perturbations to the original coefficients of the system, employing even a generally good numerical technique—short of exact arithmetic—on an ill-conditioned system carries the risk of producing nonsensical results. In dealing with an ill-conditioned system, the engineer or scientist is often confronted with a much more basic (and sometimes more disturbing) problem than that of simply trying to solve the system. Even if a minor miracle could be performed so that the exact solution could be extracted, the scientist or engineer might still have a nonsensical solution that could lead to totally incorrect conclusions. The problem stems from the fact that the coefficients are often empirically obtained and are therefore known only within certain tolerances. For an ill-conditioned system, a small uncertainty in any of the coefficients can mean an extremely large uncertainty may exist in the solution. This large uncertainty can render even the exact solution totally useless. Example 1.6.2 Suppose that for the system .835x + .667y = b1 .333x + .266y = b2 the numbers b1 and b2 are the results of an experiment and must be read from the dial of a test instrument. Suppose that the dial can be read to within a tolerance of ±.001, and assume that values for b1 and b2 are read as .168 and .067, respectively. This produces the ill-conditioned system of Example 1.6.1, and it was seen in that example that the exact solution of the system is (x, y) = (1, −1). (1.6.1) However, due to the small uncertainty in reading the dial, we have that .167 ≤ b1 ≤ .169 and .066 ≤ b2 ≤ .068. (1.6.2) For example, this means that the solution associated with the reading (b1 , b2) = (.168, .067) is just as valid as the solution associated with the reading (b1 , b2) = (.167, .068), or the reading (b1 , b2) = (.169, .066), or any other reading falling in the range (1.6.2). For the reading (b1 , b2) = (.167, .068), the exact solution is (x, y) = (934, −1169), (1.6.3) while for the other reading (b1 , b2) = (.169, .066), the exact solution is (x, y) = (−932, 1167). (1.6.4) Would you be willing to be the first to fly in the plane or drive across the bridge whose design incorporated a solution to this problem? I wouldn't! There is just too much uncertainty. Since no one of the solutions (1.6.1), (1.6.3), or (1.6.4) can be preferred over any of the others, it is conceivable that totally different designs might be implemented depending on how the technician reads the last significant digit on the dial. Due to the ill-conditioned nature of an associated linear system, the successful design of the plane or bridge may depend on blind luck rather than on scientific principles. Rather than trying to extract accurate solutions from ill-conditioned systems, engineers and scientists are usually better off investing their time and resources in trying to redesign the associated experiments or their data collection methods so as to avoid producing ill-conditioned systems. There is one other discomforting aspect of ill-conditioned systems. It concerns what students refer to as "checking the answer" by substituting a computed solution back into the left-hand side of the original system of equations to see how close it comes to satisfying the system—that is, producing the right-hand side. More formally, if xc = (ξ1 ξ2 · · · ξn) is a computed solution for a system a11 x1 + a12 x2 + · · · + a1n xn = b1 , a21 x1 + a22 x2 + · · · + a2n xn = b2 , .. . an1 x1 + an2 x2 + · · · + ann xn = bn , 36 Chapter 1 Linear Equations the numbers ri = ai1 ξ1 + ai2 ξ2 + · · · + ain ξn − bi for i = 1, 2, . . . , n are called the residuals. Suppose that you compute a solution xc and substitute it back to find that all the residuals are relatively small. Does this guarantee that xc is close to the exact solution? Surprisingly, the answer is a resounding "no!" whenever the system is ill-conditioned. Example 1.6.3 For the ill-conditioned system given in Example 1.6.1, suppose that somehow you compute a solution to be ξ1 = −666 and ξ2 = 834. If you attempt to "check the error" in this computed solution by substituting it back into the original system, then you find—using exact arithmetic—that the residuals are r1 = .835ξ1 + .667ξ2 − .168 = 0, r2 = .333ξ1 + .266ξ2 − .067 = −.001. That is, the computed solution (−666, 834) exactly satisfies the first equation and comes very close to satisfying the second. On the surface, this might seem to suggest that the computed solution should be very close to the exact solution. In fact a naive person could probably be seduced into believing that the computed solution is within ±.001 of the exact solution. Obviously, this is nowhere close to being true since the exact solution is x = 1 and y = −1. This is always a shock to a student seeing this illustrated for the first time because it is counter to a novice's intuition. Unfortunately, many students leave school believing that they can always "check" the accuracy of their computations by simply substituting them back into the original equations—it is good to know that you're not among them. This raises the question, "How can I check a computed solution for accuracy?" Fortunately, if the system is well-conditioned, then the residuals do indeed provide a more effective measure of accuracy (a rigorous proof along with more insight appears in Example 5.12.2 on p. 416). But this means that you must be able to answer some additional questions. For example, how can one tell beforehand if a given system is ill-conditioned? How can one measure the extent of ill-conditioning in a linear system? One technique to determine the extent of ill-conditioning might be to experiment by slightly perturbing selected coefficients and observing how the solution changes. If a radical change in the solution is observed for a small perturbation to some set of coefficients, then you have uncovered an ill-conditioned situation. If a given perturbation does not produce a large change in the solution, then nothing can be concluded—perhaps you perturbed the wrong set of coefficients. By performing several such experiments using different sets of coefficients, a feel (but not a guarantee) for the extent of ill-conditioning can be obtained. This is expensive and not very satisfying. But before more can be said, more sophisticated tools need to be developed—the topics of sensitivity and conditioning are revisited on p. 127 and in Example 5.12.1 on p. 414. Exercises for section 1.6 1.6.1. Consider the ill-conditioned system of Example 1.6.1: .835x + .667y = .168, .333x + .266y = .067. (a) Describe the outcome when you attempt to solve the system using 5-digit arithmetic with no scaling. (b) Again using 5-digit arithmetic, first row scale the system before attempting to solve it. Describe to what extent this helps. (c) Now use 6-digit arithmetic with no scaling. Compare the results with the exact solution. (d) Using 6-digit arithmetic, compute the residuals for your solution of part (c), and interpret the results. (e) For the same solution obtained in part (c), again compute the residuals, but use 7-digit arithmetic this time, and interpret the results. (f) Formulate a concluding statement that summarizes the points made in parts (a)–(e). 1.6.2. Perturb the ill-conditioned system given in Exercise 1.6.1 above so as to form the following system: .835x + .667y = .1669995, .333x + .266y = .066601. (a) Determine the exact solution, and compare it with the exact solution of the system in Exercise 1.6.1. (b) On the basis of the results of part (a), formulate a statement concerning the necessity for the solution of an ill-conditioned system to undergo a radical change for every perturbation of the original system. 38 Chapter 1 Linear Equations 1.6.3. Consider the two straight lines determined by the graphs of the following two equations: .835x + .667y = .168, .333x + .266y = .067. (a) Use 5-digit arithmetic to compute the slopes of each of the lines, and then use 6-digit arithmetic to do the same. In each case, sketch the graphs on a common set of axes. (b) Show by diagram why a small perturbation in either of these lines can result in a large change in the solution. (c) Describe in geometrical terms the situation that must exist in order for a system to be optimally well-conditioned. 1.6.4. Using geometric considerations, rank the following three systems according to their condition. (a) 1.001x − y = .235, x + .0001y = .765. (b) 1.001x − y = .235, x + .9999y = .765. (c) 1.001x + y = .235, x + .9999y = .765. 1.6.5. Determine the exact solution of the following system: 8x + 5y + 2z = 15, 21x + 19y + 16z = 56, 39x + 48y + 53z = 140. Now change 15 to 14 in the first equation and again solve the system with exact arithmetic. Is the system ill-conditioned? 1.6.6. Show that the system v − w − x − y − z = 0, w − x − y − z = 0, x − y − z = 0, y − z = 0, z = 1, 1.6 Ill-Conditioned Systems 39 is ill-conditioned by considering the following perturbed system: v − w − x − y − z = 0, −1/15 v + w − x − y − z = 0, −1/15 v + x − y − z = 0, −1/15 v + y − z = 0, −1/15 v + z = 1. 1.6.7. Let f (x) = sin πx on [0, 1]. The object of this problem is to determine the coefficients αi of the cubic polynomial p(x) = Σ αi xi that is as close to f (x) as possible in the sense that r = ∫ [f (x) − p(x)]2 dx [illegible] is as small as possible. (a) In order to minimize r, impose the condition that ∂r/∂αi = 0 for each i = 0, 1, 2, 3, and show this results in a system of linear equations whose augmented matrix is [H4 | b], where H4 and b are given by [illegible] Any matrix Hn that has the same form as H4 is called a Hilbert matrix of order n. (b) Systems involving Hilbert matrices are badly ill-conditioned, and the ill-conditioning becomes worse as the size increases. Use exact arithmetic with Gaussian elimination to reduce H4 to triangular form. Assuming that the case in which n = 4 is typical, explain why a general system [Hn | b] will be ill-conditioned. Notice that even complete pivoting is of no help. [illegible] CHAPTER 2 Rectangular Systems and Echelon Forms 2.1 ROW ECHELON FORM AND RANK We are now ready to analyze more general linear systems consisting of m linear equations involving n unknowns a11 x1 + a12 x2 + · · · + a1n xn = b1 , a21 x1 + a22 x2 + · · · + a2n xn = b2 , .. . am1 x1 + am2 x2 + · · · + amn xn = bm , where m may be different from n. If we do not know for sure that m and n are the same, then the system is said to be rectangular. The case m = n is still allowed in the discussion—statements concerning rectangular systems also are valid for the special case of square systems. The first goal is to extend the Gaussian elimination technique from square systems to completely general rectangular systems. Recall that for a square system with a unique solution, the pivotal positions are always located along the main diagonal—the diagonal line from the upper-left-hand corner to the lower-right-hand corner—in the coefficient matrix A so that Gaussian elimination results in a reduction of A to a triangular matrix, such as the following one for the case n = 4: [illegible] 42 Chapter 2 Rectangular Systems and Echelon Forms Remember that a pivot must always be a nonzero number. For square systems possessing a unique solution, it is a fact (proven later) that one can always bring a nonzero number into each pivotal position along the main diagonal. However, in the case of a general rectangular system, it is not always possible to have the pivotal positions lying on a diagonal line in the coefficient matrix. This means that the final result of Gaussian elimination will not be triangular in form. For example, consider the following system: x1 + 2x2 + x3 + 3x4 + 3x5 = 5, 2x1 + 4x2 + 4x4 + 4x5 = 6, x1 + 2x2 + 3x3 + 5x4 + 5x5 = 9, 2x1 + 4x2 + 4x4 + 7x5 = 9. Focus your attention on the coefficient matrix [illegible] and ignore the right-hand side for a moment. Applying Gaussian elimination to A yields the following result: [illegible] In the basic elimination process, the strategy is to move down and to the right to the next pivotal position. If a zero occurs in this position, an interchange with a row below the pivotal position is executed so as to bring a nonzero number into the pivotal position. However, in this example, it is clearly impossible to bring a nonzero number into the (2, 2)-position by interchanging the second row with a lower row. In order to handle this situation, the elimination process is modified as follows. [illegible] Modified Gaussian Elimination Suppose that U is the augmented matrix associated with the system after i − 1 elimination steps have been completed. To execute the ith step, proceed as follows: • Moving from left to right in U , locate the first column that contains a nonzero entry on or below the ith position—say it is U∗j . • The pivotal position for the ith step is the (i, j)-position. • If necessary, interchange the ith row with a lower row to bring a nonzero number into the (i, j)-position, and then annihilate all entries below this pivot. • If row Ui∗ as well as all rows in U below Ui∗ consist entirely of zeros, then the elimination process is completed. Illustrated below is the result of applying this modified version of Gaussian elimination to the matrix given in (2.1.1). Example 2.1.1 Problem: Apply modified Gaussian elimination to the following matrix and circle the pivot positions: [illegible] 44 Chapter 2 Rectangular Systems and Echelon Forms Notice that the final result of applying Gaussian elimination in the above example is not a purely triangular form but rather a jagged or "stair-step" type of triangular form. Hereafter, a matrix that exhibits this stair-step structure will be said to be in row echelon form. Row Echelon Form An m × n matrix E with rows Ei∗ and columns E∗j is said to be in row echelon form provided the following two conditions hold. • If Ei∗ consists entirely of zeros, then all rows below Ei∗ are also entirely zero; i.e., all zero rows are at the bottom. • If the first nonzero entry in Ei∗ lies in the jth position, then all entries below the ith position in columns E∗1 , E∗2 , . . . , E∗j are zero. These two conditions say that the nonzero entries in an echelon form must lie on or above a stair-step line that emanates from the upper-left-hand corner and slopes down and to the right. The pivots are the first nonzero entries in each row. A typical structure for a matrix in row echelon form is illustrated below with the pivots circled. [illegible] Because of the flexibility in choosing row operations to reduce a matrix A to a row echelon form E, the entries in E are not uniquely determined by A. Nevertheless, it can be proven that the "form" of E is unique in the sense that the positions of the pivots in E (and A) are uniquely determined by the entries in A . Because the pivotal positions are unique, it follows that the number of pivots, which is the same as the number of nonzero rows in E, is also uniquely determined by the entries in A. This number is called the rank of A, and it is extremely important in the development of our subject. Rank of a Matrix Suppose Am×n is reduced by row operations to an echelon form E. The rank of A is defined to be the number rank (A) = number of pivots = number of nonzero rows in E = number of basic columns in A, where the basic columns of A are defined to be those columns in A that contain the pivotal positions. Example 2.1.2 Problem: Determine the rank, and identify the basic columns in [illegible] Solution: Reduce A to row echelon form as shown below: [illegible] Consequently, rank (A) = 2. The pivotal positions lie in the first and fourth columns so that the basic columns of A are A∗1 and A∗4 . That is, Basic Columns = [illegible] Pay particular attention to the fact that the basic columns are extracted from A and not from the row echelon form E. 2.1 Row Echelon Form and Rank 45 Exercises for section 2.1 2.1.1. Reduce each of the following matrices to row echelon form, determine the rank, and identify the basic columns. [illegible] 2.1.2. Determine which of the following matrices are in row echelon form: [illegible] 2.1.3. Suppose that A is an m × n matrix. Give a short explanation of why each of the following statements is true. (a) rank (A) ≤ min{m, n}. (b) rank (A) < m if one row in A is entirely zero. (c) rank (A) < m if one row in A is a multiple of another row. (d) rank (A) < m if one row in A is a combination of other rows. (e) rank (A) < n if one column in A is entirely zero. 2.1.4. Let A = [illegible]. (a) Use exact arithmetic to determine rank (A). (b) Now use 3-digit floating-point arithmetic (without partial pivoting or scaling) to determine rank (A). This number might be called the "3-digit numerical rank." (c) What happens if partial pivoting is incorporated? 2.1.5. How many different "forms" are possible for a 3 × 4 matrix that is in row echelon form? 2.1.6. Suppose that [A|b] is reduced to a matrix [E|c]. (a) Is [E|c] in row echelon form if E is? (b) If [E|c] is in row echelon form, must E be in row echelon form? 46 Chapter 2 Rectangular Systems and Echelon Forms 2.2 REDUCED ROW ECHELON FORM At each step of the Gauss–Jordan method, the pivot is forced to be a 1, and then all entries above and below the pivotal 1 are annihilated. If A is the coefficient matrix for a square system with a unique solution, then the end result of applying the Gauss–Jordan method to A is a matrix with 1's on the main diagonal and 0's everywhere else. That is, [illegible] But if the Gauss–Jordan technique is applied to a more general m × n matrix, then the final result is not necessarily the same as described above. The following example illustrates what typically happens in the rectangular case. Example 2.2.1 Problem: Apply Gauss–Jordan elimination to the following 4 × 5 matrix and circle the pivot positions. This is the same matrix used in Example 2.1.1: [illegible] Solution: [illegible] 2.2 Reduced Row Echelon Form 47 Compare the results of this example with the results of Example 2.1.1, and notice that the "form" of the final matrix is the same in both examples, which indeed must be the case because of the uniqueness of "form" mentioned in the previous section. The only difference is in the numerical value of some of the entries. By the nature of Gauss–Jordan elimination, each pivot is 1 and all entries above and below each pivot are 0. Consequently, the row echelon form produced by the Gauss–Jordan method contains a reduced number of nonzero entries, so it seems only natural to refer to this as a reduced row echelon form. Reduced Row Echelon Form A matrix Em×n is said to be in reduced row echelon form provided that the following three conditions hold. • E is in row echelon form. • The first nonzero entry in each row (i.e., each pivot) is 1. • All entries above each pivot are 0. A typical structure for a matrix in reduced row echelon form is illustrated below, where entries marked * can be either zero or nonzero numbers: [illegible] As previously stated, if matrix A is transformed to a row echelon form by row operations, then the "form" is uniquely determined by A, but the individual entries in the form are not unique. However, if A is transformed by row operations to a reduced row echelon form EA , then it can be shown that both the "form" as well as the individual entries in EA are uniquely determined by A. In other words, the reduced row echelon form EA produced from A is independent of whatever elimination scheme is used. Producing an unreduced form is computationally more efficient, but the uniqueness of EA makes it more useful for theoretical purposes. EA Notation For a matrix A, the symbol EA will hereafter denote the unique reduced row echelon form derived from A by means of row operations. Example 2.2.2 Problem: Determine EA , deduce rank (A), and identify the basic columns of A = [illegible] Solution: [illegible] Therefore, rank (A) = 3, and {A∗1 , A∗3 , A∗5 } are the three basic columns. The above example illustrates another important feature of EA and explains why the basic columns are indeed "basic." Each nonbasic column is expressible as a combination of basic columns. In Example 2.2.2, A∗2 = 2A∗1 and A∗4 = A∗1 + A∗3 . (2.2.1) Notice that exactly the same set of relationships hold in EA . That is, E∗2 = 2E∗1 and E∗4 = E∗1 + E∗3 . (2.2.2) This is no coincidence—it's characteristic of what happens in general. There's more to observe. The relationships between the nonbasic and basic columns in a general matrix A are usually obscure, but the relationships among the columns in EA are absolutely transparent. For example, notice that the multipliers used in the relationships (2.2.1) and (2.2.2) appear explicitly in the two nonbasic columns in EA —they are just the nonzero entries in these nonbasic columns. This is important because it means that EA can be used as a "map" or "key" to discover or unlock the hidden relationships among the columns of A . Finally, observe from Example 2.2.2 that only the basic columns to the left of a given nonbasic column are needed in order to express the nonbasic column as a combination of basic columns—e.g., representing A∗2 requires only A∗1 and not A∗3 or A∗5 , while representing A∗4 requires only A∗1 and A∗3 . This too is typical. For the time being, we accept the following statements to be true. A rigorous proof is given later on p. 136. Column Relationships in A and EA • Each nonbasic column E∗k in EA is a combination (a sum of multiples) of the basic columns in EA to the left of E∗k . That is, E∗k = µ1 E∗b1 + µ2 E∗b2 + · · · + µj E∗bj = [illegible] where the E∗bi 's are the basic columns to the left of E∗k and where the multipliers µi are the first j entries in E∗k . • The relationships that exist among the columns of A are exactly the same as the relationships that exist among the columns of EA . In particular, if A∗k is a nonbasic column in A , then A∗k = µ1 A∗b1 + µ2 A∗b2 + · · · + µj A∗bj , (2.2.3) where the A∗bi 's are the basic columns in A corresponding to the basic columns in EA , and where the multipliers µi are as described above—the first j entries in E∗k . 2.2 Reduced Row Echelon Form 51 Example 2.2.3 Problem: Write each nonbasic column as a combination of basic columns in A = [illegible] Solution: Transform A to EA as shown below. [illegible] The third and fifth columns are nonbasic. Looking at the columns in EA reveals E∗3 = 2E∗1 + E∗2 and E∗5 = −E∗1 + 2E∗2 + E∗4 . The relationships that exist among the columns of A must be exactly the same as those in EA , so A∗3 = 2A∗1 + A∗2 and A∗5 = −A∗1 + 2A∗2 + A∗4 . You can easily check the validity of these equations by direct calculation. In summary, the utility of EA lies in its ability to reveal dependencies in data stored as columns in an array A. The nonbasic columns in A represent redundant information in the sense that this information can always be expressed in terms of the data contained in the basic columns. Although data compression is not the primary reason for introducing EA , the application to these problems is clear. For a large array of data, it may be more efficient to store only "independent data" (i.e., the basic columns of A) along with the nonzero multipliers µi obtained from the nonbasic columns in EA . Then the redundant data contained in the nonbasic columns of A can always be reconstructed if and when it is called for. Exercises for section 2.2 2.2.1. Determine the reduced row echelon form for each of the following matrices and then express each nonbasic column in terms of the basic columns: [illegible] 2.2.2. Construct a matrix A whose reduced row echelon form is EA = [illegible] Is A unique? 2.2.3. Suppose that A is an m × n matrix. Give a short explanation of why rank (A) < n whenever one column in A is a combination of other columns in A . 2.2.4. Consider the following matrix: A = [illegible]. (a) Use exact arithmetic to determine EA . (b) Now use 3-digit floating-point arithmetic (without partial pivoting or scaling) to determine EA and formulate a statement concerning "near relationships" between the columns of A . 2.2.5. Consider the matrix E = [illegible]. You already know that E∗3 can be expressed in terms of E∗1 and E∗2 . However, this is not the only way to represent the column dependencies in E . Show how to write E∗1 in terms of E∗2 and E∗3 and then express E∗2 as a combination of E∗1 and E∗3 . Note: This exercise illustrates that the set of pivotal columns is not the only set that can play the role of "basic columns." Taking the basic columns to be the ones containing the pivots is a matter of convenience because everything becomes automatic that way. 2.3 Consistency of Linear Systems 53 2.3 CONSISTENCY OF LINEAR SYSTEMS A system of m linear equations in n unknowns is said to be a consistent system if it possesses at least one solution. If there are no solutions, then the system is called inconsistent. The purpose of this section is to determine conditions under which a given system will be consistent. Stating conditions for consistency of systems involving only two or three unknowns is easy. A linear equation in two unknowns represents a line in 2-space, and a linear equation in three unknowns is a plane in 3-space. Consequently, a linear system of m equations in two unknowns is consistent if and only if the m lines defined by the m equations have at least one common point of intersection. Similarly, a system of m equations in three unknowns is consistent if and only if the associated m planes have at least one common point of intersection. However, when m is large, these geometric conditions may not be easy to verify visually, and when n > 3, the generalizations of intersecting lines or planes are impossible to visualize with the eye. Rather than depending on geometry to establish consistency, we use Gaussian elimination. If the associated augmented matrix [A|b] is reduced by row operations to a matrix [E|c] that is in row echelon form, then consistency—or lack of it—becomes evident. Suppose that somewhere in the process of reducing [A|b] to [E|c] a situation arises in which the only nonzero entry in a row appears on the right-hand side, as illustrated below: [illegible] If this occurs in the ith row, then the ith equation of the associated system is 0x1 + 0x2 + · · · + 0xn = α. For α ≠ 0, this equation has no solution, and hence the original system must also be inconsistent (because row operations don't alter the solution set). The converse also holds. That is, if a system is inconsistent, then somewhere in the elimination process a row of the form (0 0 · · · 0 | α), α ≠ 0 (2.3.1) must appear. Otherwise, the back substitution process can be completed and a solution is produced. There is no inconsistency indicated when a row of the form (0 0 · · · 0 | 0) is encountered. This simply says that 0 = 0, and although 54 Chapter 2 Rectangular Systems and Echelon Forms this is no help in determining the value of any unknown, it is nevertheless a true statement, so it doesn't indicate inconsistency in the system. There are some other ways to characterize the consistency (or inconsistency) of a system. One of these is to observe that if the last column b in the augmented matrix [A|b] is a nonbasic column, then no pivot can exist in the last column, and hence the system is consistent because the situation (2.3.1) cannot occur. Conversely, if the system is consistent, then the situation (2.3.1) never occurs during Gaussian elimination and consequently the last column cannot be basic. In other words, [A|b] is consistent if and only if b is a nonbasic column. Saying that b is a nonbasic column in [A|b] is equivalent to saying that all basic columns in [A|b] lie in the coefficient matrix A . Since the number of basic columns in a matrix is the rank, consistency may also be characterized by stating that a system is consistent if and only if rank[A|b] = rank (A). Recall from the previous section the fact that each nonbasic column in [A|b] must be expressible in terms of the basic columns. Because a consistent system is characterized by the fact that the right-hand side is a nonbasic column, it follows that a system is consistent if and only if the right-hand side b is a combination of columns from the coefficient matrix A. 13 Each of the equivalent ways of saying that a system is consistent is summarized below. Consistency Each of the following is equivalent to saying that [A|b] is consistent. • In row reducing [A|b], a row of the following form never appears: (0 0 · · · 0 | α), where α ≠ 0. (2.3.2) • • b is a nonbasic column in [A|b]. rank[A|b] = rank (A). (2.3.3) (2.3.4) • b is a combination of the basic columns in A. (2.3.5) Example 2.3.1 Problem: Determine if the following system is consistent: x1 + x2 + 2x3 + 2x4 + x5 = 1, 2x1 + 2x2 + 4x3 + 4x4 + 3x5 = 1, 2x1 + 2x2 + 4x3 + 4x4 + 2x5 = 2, 3x1 + 5x2 + 8x3 + 6x4 + 5x5 = 3. 13 Statements P and Q are said to be equivalent when (P implies Q) as well as its converse (Q implies P) are true statements. This is also the meaning of the phrase "P if and only if Q." 2.3 Consistency of Linear Systems 55 Solution: Apply Gaussian elimination to the augmented matrix [A|b] as shown: ⎛ 1 1 2 2 1 │ 2 ⎜ 2 3 2 2 5 4 4 8 4 4 6 3 2 5 ⎞⎛ 1 1 1 ║ 0 ⎠−→⎝ 0 2 3 0 ⎛ 1 │ 0 −→⎝ 0 0 1 0 0 2 1 2 0 0 2 0 0 2 2 2 0 0 2 0 0 0 2 0 0 0 1 1 0 2 1 2 1 0 ⎞ 1 −1 │ 0 0 ⎠ 1 0 │. −1 0 Because a row of the form (0 0 · · · 0 | α) with α ≠ 0 never emerges, the system is consistent. We might also observe that b is a nonbasic column in [A|b] so that rank[A|b] = rank (A). Finally, by completely reducing A to EA , it is possible to verify that b is indeed a combination of the basic columns {A∗1 , A∗2 , A∗5 }. Exercises for section 2.3 2.3.1. Determine which of the following systems are consistent. x + 2y + z = 2, (a) (c) (e) 2x + 4y = 2, 3x + 6y + z = 4. x−y+z x−y−z x+y−z x+y+z = 1, = 2, = 3, = 4. 2w + x + 3y + 5z = 1, 4w + 4y + 8z = 0, w + x + 2y + 3z = 0, x + y + z = 0. (b) 2x + 2y + 4z = 0, 3x + 2y + 5z = 0, 4x + 2y + 6z = 0. (d) (f) x−y+z x−y−z x+y−z x+y+z = 1, = 2, = 3, = 2. 2w + x + 3y + 5z = 7, 4w + 4y + 8z = 8, w + x + 2y + 3z = 5, x + y + z = 3. 2.3.2. Construct a 3 × 4 matrix A and 3 × 1 columns b and c such that [A|b] is the augmented matrix for an inconsistent system, but [A|c] is the augmented matrix for a consistent system. 2.3.3. If A is an m × n matrix with rank (A) = m, explain why the system [A|b] must be consistent for every right-hand side b . 56 Chapter 2 Rectangular Systems and Echelon Forms 2.3.4. Consider two consistent systems whose augmented matrices are of the form [A|b] and [A|c]. That is, they differ only on the right-hand side. Is the system associated with [A | b + c] also consistent? Explain why. 2.3.5. Is it possible for a parabola whose equation has the form y = α+βx+γx2 to pass through the four points (0, 1), (1, 3), (2, 15), and (3, 37)? Why? 2.3.6. Consider using floating-point arithmetic (without scaling) to solve the following system: .835x + .667y = .168, .333x + .266y = .067. (a) Is the system consistent when 5-digit arithmetic is used? (b) What happens when 6-digit arithmetic is used? 2.3.7. In order to grow a certain crop, it is recommended that each square foot of ground be treated with 10 units of phosphorous, 9 units of potassium, and 19 units of nitrogen. Suppose that there are three brands of fertilizer on the market— say brand X , brand Y , and brand Z . One pound of brand X contains 2 units of phosphorous, 3 units of potassium, and 5 units of nitrogen. One pound of brand Y contains 1 unit of phosphorous, 3 units of potassium, and 4 units of nitrogen. One pound of brand Z contains only 1 unit of phosphorous and 1 unit of nitrogen. Determine whether or not it is possible to meet exactly the recommendation by applying some combination of the three brands of fertilizer. 2.3.8. Suppose that an augmented matrix [A|b] is reduced by means of Gaussian elimination to a row echelon form [E|c]. If a row of the form (0 0 · · · 0 | α), α ≠ 0 does not appear in [E|c], is it possible that rows of this form could have appeared at earlier stages in the reduction process? Why? 2.4 Homogeneous Systems 2.4 57 HOMOGENEOUS SYSTEMS A system of m linear equations in n unknowns a11 x1 + a12 x2 + · · · + a1n xn = 0, a21 x1 + a22 x2 + · · · + a2n xn = 0, .. . am1 x1 + am2 x2 + · · · + amn xn = 0, in which the right-hand side consists entirely of 0's is said to be a homogeneous system. If there is at least one nonzero number on the right-hand side, then the system is called nonhomogeneous. The purpose of this section is to examine some of the elementary aspects concerning homogeneous systems. Consistency is never an issue when dealing with homogeneous systems because the zero solution x1 = x2 = · · · = xn = 0 is always one solution regardless of the values of the coefficients. Hereafter, the solution consisting of all zeros is referred to as the trivial solution. The only question is, "Are there solutions other than the trivial solution, and if so, how can we best describe them?" As before, Gaussian elimination provides the answer. While reducing the augmented matrix [A|0] of a homogeneous system to a row echelon form using Gaussian elimination, the zero column on the righthand side can never be altered by any of the three elementary row operations. That is, any row echelon form derived from [A|0] by means of row operations must also have the form [E|0]. This means that the last column of 0's is just excess baggage that is not necessary to carry along at each step. Just reduce the coefficient matrix A to a row echelon form E, and remember that the righthand side is entirely zero when you execute back substitution. The process is best understood by considering a typical example. In order to examine the solutions of the homogeneous system x1 + 2x2 + 2x3 + 3x4 = 0, 2x1 + 4x2 + x3 + 3x4 = 0, 3x1 + 6x2 + x3 + 4x4 = 0, reduce the ⎛ 1 A = ⎝2 3 coefficient matrix to a row echelon form. ⎞⎛ ⎞ 2 2 3 1 2 2 3 1 4 1 3 ⎠−→⎝ 0 0 −3 −3 ⎠−→⎝ 0 6 1 4 0 0 −5 −5 0 (2.4.1) 2 0 0 2 −3 0 ⎞3 −3 ⎠= E. 0 Therefore, the original homogeneous system is equivalent to the following reduced homogeneous system: x1 + 2x2 + 2x3 + 3x4 = 0, − 3x3 − 3x4 = 0. (2.4.2) 58 Chapter 2 Rectangular Systems and Echelon Forms Since there are four unknowns but only two equations in this reduced system, it is impossible to extract a unique solution for each unknown. The best we can do is to pick two "basic" unknowns—which will be called the basic variables and solve for these in terms of the other two unknowns—whose values must remain arbitrary or "free," and consequently they will be referred to as the free variables. Although there are several possibilities for selecting a set of basic variables, the convention is to always solve for the unknowns corresponding to the pivotal positions—or, equivalently, the unknowns corresponding to the basic columns. In this example, the pivots (as well as the basic columns) lie in the first and third positions, so the strategy is to apply back substitution to solve the reduced system (2.4.2) for the basic variables x1 and x3 in terms of the free variables x2 and x4 . The second equation in (2.4.2) yields x3 = −x4 and substitution back into the first equation produces x1 = −2x2 − 2x3 − 3x4 , = −2x2 − 2(−x4) − 3x4 , = −2x2 − x4 . Therefore, all solutions of the original homogeneous system can be described by saying x1 = −2x2 − x4 , x2 is "free," (2.4.3) x3 = −x4 , x4 is "free." As the free variables x2 and x4 range over all possible values, the above expressions describe all possible solutions. For example, when x2 and x4 assume the values x2 = 1 and x4 = −2, then the particular solution x1 = 0, x2 = 1, x3 = 2, x4 = −2 √ is produced. When x2 = π and x4 = 2, then another particular solution √ √ √ x1 = −2π − 2, x2 = π, x3 = − 2, x4 = 2 is generated. Rather than describing the solution set as illustrated in (2.4.3), future developments will make it more convenient to express the solution set by writing ⎛ ⎞⎛ ⎞⎛ ⎞⎛ ⎞x1 −2x2 − x4 −2 −1 x2 ⎟ x2 ⎟ 1 ⎟ 0 ⎟(2.4.4) ⎝ ⎠=⎝ ⎠ = x2 ⎝ ⎠ + x4 ⎝ ⎠0 −1 x3 −x4 0 1 x4 x4 2.4 Homogeneous Systems 59 with the understanding that x2 and x4 are free variables that can range over all possible numbers. This representation will be called the general solution of the homogeneous system. This expression for the general solution emphasizes that every solution is some combination of the two particular solutions ⎛ ⎞ −2 ⎟ 1 ⎟ h1 = ⎝ ⎠ 0 0 ⎛ and ⎞ −1 ⎟ 0 ⎟ h2 = ⎝ ⎠. −1 1 The fact that h1 and h2 are each solutions is clear because h1 is produced when the free variables assume the values x2 = 1 and x4 = 0, whereas the solution h2 is generated when x2 = 0 and x4 = 1. Now consider a general homogeneous system [A|0] of m linear equations in n unknowns. If the coefficient matrix is such that rank (A) = r, then it should be apparent from the preceding discussion that there will be exactly r basic variables—corresponding to the positions of the basic columns in A —and exactly n − r free variables—corresponding to the positions of the nonbasic columns in A . Reducing A to a row echelon form using Gaussian elimination and then using back substitution to solve for the basic variables in terms of the free variables produces the general solution, which has the form x = xf1 h1 + xf2 h2 + · · · + xfn−r hn−r , (2.4.5) where xf1 , xf2 , . . . , xfn−r are the free variables and where h1 , h2 , . . . , hn−r are n × 1 columns that represent particular solutions of the system. As the free variables xfi range over all possible values, the general solution generates all possible solutions. The general solution does not depend on which row echelon form is used in the sense that using back substitution to solve for the basic variables in terms of the nonbasic variables generates a unique set of particular solutions {h1 , h2 , . . . , hn−r }, regardless of which row echelon form is used. Without going into great detail, one can argue that this is true because using back substitution in any row echelon form to solve for the basic variables must produce exactly the same result as that obtained by completely reducing A to EA and then solving the reduced homogeneous system for the basic variables. Uniqueness of EA guarantees the uniqueness of the hi 's. For example, if the coefficient matrix A associated with the system (2.4.1) is completely reduced by the Gauss–Jordan procedure to EA ⎛ 1 A = ⎝2 3 2 4 6 2 1 1 ⎞⎛ 3 1 3 ⎠−→⎝ 0 4 0 2 0 0 0 1 0 ⎞1 1 ⎠= EA , 0 60 Chapter 2 Rectangular Systems and Echelon Forms then we obtain the following reduced system: x1 + 2x2 + x4 = 0, x3 + x4 = 0. Solving for the basic variables x1 and x3 in terms of x2 and x4 produces exactly the same result as given in (2.4.3) and hence generates exactly the same general solution as shown in (2.4.4). Because it avoids the back substitution process, you may find it more convenient to use the Gauss–Jordan procedure to reduce A completely to EA and then construct the general solution directly from the entries in EA . This approach usually will be adopted in the examples and exercises. As was previously observed, all homogeneous systems are consistent because the trivial solution consisting of all zeros is always one solution. The natural question is, "When is the trivial solution the only solution?" In other words, we wish to know when a homogeneous system possesses a unique solution. The form of the general solution (2.4.5) makes the answer transparent. As long as there is at least one free variable, then it is clear from (2.4.5) that there will be an infinite number of solutions. Consequently, the trivial solution is the only solution if and only if there are no free variables. Because the number of free variables is given by n − r, where r = rank (A), the previous statement can be reformulated to say that a homogeneous system possesses a unique solution—the trivial solution—if and only if rank (A) = n. Example 2.4.1 The homogeneous system x1 + 2x2 + 2x3 = 0, 2x1 + 5x2 + 7x3 = 0, 3x1 + 6x2 + 8x3 = 0, has only the trivial solution because ⎛ ⎞⎛ 1 2 2 1 A = ⎝ 2 5 7 ⎠−→⎝ 0 3 6 8 0 2 1 0 ⎞2 3 ⎠= E 2 shows that rank (A) = n = 3. Indeed, it is also obvious from E that applying back substitution in the system [E|0] yields only the trivial solution. Example 2.4.2 Problem: Explain why the following homogeneous system has infinitely many solutions, and exhibit the general solution: x1 + 2x2 + 2x3 = 0, 2x1 + 5x2 + 7x3 = 0, 3x1 + 6x2 + 6x3 = 0. 2.4 Homogeneous Systems Solution: 61 ⎛ 1 A = ⎝2 3 2 5 6 ⎞⎛ 2 1 7 ⎠−→⎝ 0 6 0 2 1 0 ⎞2 3 ⎠= E 0 shows that rank (A) = 2 < n = 3. Since the basic columns lie in positions one and two, x1 and x2 are the basic variables while x3 is free. Using back substitution on [E|0] to solve for the basic variables in terms of the free variable produces x2 = −3x3 and x1 = −2x2 − 2x3 = 4x3 , so the general solution is ⎛ ⎞⎛ ⎞x1 4 ⎝ x2 ⎠ = x3 ⎝ −3 ⎠, where x3 is free. 1 x3 ⎛ ⎞4 That is, every solution is a multiple of the one particular solution h1 = ⎝ −3 ⎠. 1 Summary Let Am×n be the coefficient matrix for a homogeneous system of m linear equations in n unknowns, and suppose rank (A) = r. • The unknowns that correspond to the positions of the basic columns (i.e., the pivotal positions) are called the basic variables, and the unknowns corresponding to the positions of the nonbasic columns are called the free variables. • There are exactly r basic variables and n − r free variables. • To describe all solutions, reduce A to a row echelon form using Gaussian elimination, and then use back substitution to solve for the basic variables in terms of the free variables. This produces the general solution that has the form x = xf1 h1 + xf2 h2 + · · · + xfn−r hn−r , where the terms xf1 , xf2 , . . . , xfn−r are the free variables and where h1 , h2 , . . . , hn−r are n × 1 columns that represent particular solutions of the homogeneous system. The hi 's are independent of which row echelon form is used in the back substitution process. As the free variables xfi range over all possible values, the general solution generates all possible solutions. • A homogeneous system possesses a unique solution (the trivial solution) if and only if rank (A) = n —i.e., if and only if there are no free variables. 62 Chapter 2 Rectangular Systems and Echelon Forms Exercises for section 2.4 2.4.1. Determine the general solution for each of the following homogeneous systems. (a) x1 + 2x2 + x3 + 2x4 = 0, 2x1 + 4x2 + x3 + 3x4 = 0, 3x1 + 6x2 + x3 + 4x4 = 0. x1 + x2 + 2x3 (c) (b) = 0, 3x1 + 3x3 + 3x4 = 0, 2x1 + x2 + 3x3 + x4 = 0, x1 + 2x2 + 3x3 − x4 = 0. (d) 2x + y + z 4x + 2y + z 6x + 3y + z 8x + 4y + z = 0, = 0, = 0, = 0. 2x + y + z = 0, 4x + 2y + z = 0, 6x + 3y + z = 0, 8x + 5y + z = 0. 2.4.2. Among all solutions that satisfy the homogeneous system x + 2y + z = 0, 2x + 4y + z = 0, x + 2y − z = 0, determine those that also satisfy the nonlinear constraint y − xy = 2z. 2.4.3. Consider a homogeneous system whose coefficient matrix is ⎛ 1 ⎜2 ⎜ A = ⎜1 ⎝ 2 3 2 4 2 4 6 1 −1 3 2 1 3 3 5 6 7 ⎞1 8 ⎟ ⎟7 ⎟. ⎠2 −3 First transform A to an unreduced row echelon form to determine the general solution of the associated homogeneous system. Then reduce A to EA , and show that the same general solution is produced. 2.4.4. If A is the coefficient matrix for a homogeneous system consisting of four equations in eight unknowns and if there are five free variables, what is rank (A)? 2.4 Homogeneous Systems 63 2.4.5. Suppose that A is the coefficient matrix for a homogeneous system of four equations in six unknowns and suppose that A has at least one nonzero row. (a) Determine the fewest number of free variables that are possible. (b) Determine the maximum number of free variables that are possible. 2.4.6. Explain why a homogeneous system of m equations in n unknowns where m < n must always possess an infinite number of solutions. 2.4.7. Construct a homogeneous system of three equations in four unknowns that has ⎛ ⎞⎛ ⎞ −2 −3 ⎟ 1 ⎟ 0 ⎟ x2 ⎝ ⎠ + x4 ⎝ ⎠ 0 2 0 1 as its general solution. 2.4.8. If c1 and c2 are columns that represent two particular solutions of the same homogeneous system, explain why the sum c1 + c2 must also represent a solution of this system. 64 2.5 Chapter 2 Rectangular Systems and Echelon Forms NONHOMOGENEOUS SYSTEMS Recall that a system of m linear equations in n unknowns a11 x1 + a12 x2 + · · · + a1n xn = b1 , a21 x1 + a22 x2 + · · · + a2n xn = b2 , .. . am1 x1 + am2 x2 + · · · + amn xn = bm , is said to be nonhomogeneous whenever bi ≠ 0 for at least one i. Unlike homogeneous systems, a nonhomogeneous system may be inconsistent and the techniques of §2.3 must be applied in order to determine if solutions do indeed exist. Unless otherwise stated, it is assumed that all systems in this section are consistent. To describe the set of all possible solutions of a consistent nonhomogeneous system, construct a general solution by exactly the same method used for homogeneous systems as follows. • Use Gaussian elimination to reduce the associated augmented matrix [A|b] to a row echelon form [E|c]. • Identify the basic variables and the free variables in the same manner described in §2.4. • Apply back substitution to [E|c] and solve for the basic variables in terms of the free variables. • Write the result in the form x = p + xf1 h1 + xf2 h2 + · · · + xfn−r hn−r , (2.5.1) where xf1 , xf2 , . . . , xfn−r are the free variables and p, h1 , h2 , . . . , hn−r are n × 1 columns. This is the general solution of the nonhomogeneous system. As the free variables xfi range over all possible values, the general solution (2.5.1) generates all possible solutions of the system [A|b]. Just as in the homogeneous case, the columns hi and p are independent of which row echelon form [E|c] is used. Therefore, [A|b] may be completely reduced to E[A|b] by using the Gauss–Jordan method thereby avoiding the need to perform back substitution. We will use this approach whenever it is convenient. The difference between the general solution of a nonhomogeneous system and the general solution of a homogeneous system is the column p that appears 2.5 Nonhomogeneous Systems 65 in (2.5.1). To understand why p appears and where it comes from, consider the nonhomogeneous system x1 + 2x2 + 2x3 + 3x4 = 4, 2x1 + 4x2 + x3 + 3x4 = 5, 3x1 + 6x2 + x3 + 4x4 = 7, (2.5.2) in which the coefficient matrix is the same as the coefficient matrix for the homogeneous system (2.4.1) used in the previous section. If [A|b] is completely reduced by the Gauss–Jordan procedure to E[A|b] ⎛ 1 2 2 3 [A|b] = ⎝ 2 4 1 3 3 6 1 4 ⎞⎛ 4 1 2 0 1 5 ⎠−→⎝ 0 0 1 1 0 0 0 0 7 ⎞2 1 ⎠= E[A|b] , 0 then the following reduced system is obtained: x1 + 2x2 + x4 = 2, x3 + x4 = 1. Solving for the basic variables, x1 and x3 , in terms of the free variables, x2 and x4 , produces x1 = 2 − 2x2 − x4 , x2 is "free," x3 = 1 − x4 , x4 is "free." The general solution is obtained by writing these statements in the form ⎛ ⎞⎛ ⎞⎛ ⎞⎛ ⎞⎛ ⎞x1 2 − 2x2 − x4 2 −2 −1 x2 ⎟ x2 ⎟ 0 ⎟ 1 ⎟ 0 ⎟⎝ ⎠=⎝ ⎠ = ⎝ ⎠ + x2 ⎝ ⎠ + x4 ⎝ ⎠. 1 0 −1 x3 1 − x4 0 0 1 x4 x4 (2.5.3) As the free variables x2 and x4 range over all possible numbers, this generates all possible solutions of the nonhomogeneous system (2.5.2). Notice that the ⎛ ⎞2 ⎟ 0 ⎟column ⎝ ⎠in (2.5.3) is a particular solution of the nonhomogeneous system 1 0 (2.5.2)—it is the solution produced when the free variables assume the values x2 = 0 and x4 = 0. 66 Chapter 2 Rectangular Systems and Echelon Forms Furthermore, recall from (2.4.4) that the general solution of the associated homogeneous system x1 + 2x2 + 2x3 + 3x4 = 0, 2x1 + 4x2 + x3 + 3x4 = 0, 3x1 + 6x2 + x3 + 4x4 = 0, is given by (2.5.4) ⎛ ⎞⎛ ⎞⎛ ⎞−2x2 − x4 −2 −1 x2 ⎟ 1 ⎟ 0 ⎟ ⎝ ⎠ = x2 ⎝ ⎠ + x4 ⎝ ⎠. 0 −1 −x4 0 1 x4 That is, the general solution of the associated homogeneous system (2.5.4) is a part of the general solution of the original nonhomogeneous system (2.5.2). These two observations can be combined by saying that the general solution of the nonhomogeneous system is given by a particular solution plus the general 14 solution of the associated homogeneous system. To see that the previous statement is always true, suppose [A|b] represents a general m × n consistent system where rank (A) = r. Consistency guarantees that b is a nonbasic column in [A|b], and hence the basic columns in [A|b] are in the same positions as the basic columns in [A|0] so that the nonhomogeneous system and the associated homogeneous system have exactly the same set of basic variables as well as free variables. Furthermore, it is not difficult to see that E[A|0] = [EA |0] E[A|b] = [EA |c], ⎞ξ1 ⎜ .. ⎟ ⎜ . ⎟ ⎜ ⎟ ⎟ξ ⎟where c is some column of the form c = ⎜ r ⎟. This means that if you solve ⎜0⎟ ⎜ . ⎟ ⎝ .. ⎠0 the ith equation in the reduced homogeneous system for the ith basic variable xbi in terms of the free variables xfi , xfi+1 , . . . , xfn−r to produce and ⎛ xbi = αi xfi + αi+1 xfi+1 + · · · + αn−r xfn−r , (2.5.5) then the solution for the ith basic variable in the reduced nonhomogeneous system must have the form xbi = ξi + αi xfi + αi+1 xfi+1 + · · · + αn−r xfn−r . 14 (2.5.6) For those students who have studied differential equations, this statement should have a familiar ring. Exactly the same situation holds for the general solution to a linear differential equation. This is no accident—it is due to the inherent linearity in both problems. More will be said about this issue later in the text. 2.5 Nonhomogeneous Systems 67 That is, the two solutions differ only in the fact that the latter contains the constant ξi . Consider organizing the expressions (2.5.5) and (2.5.6) so as to construct the respective general solutions. If the general solution of the homogeneous system has the form x = xf1 h1 + xf2 h2 + · · · + xfn−r hn−r , then it is apparent that the general solution of the nonhomogeneous system must have a similar form x = p + xf1 h1 + xf2 h2 + · · · + xfn−r hn−r (2.5.7) in which the column p contains the constants ξi along with some 0's—the ξi 's occupy positions in p that correspond to the positions of the basic columns, and 0's occupy all other positions. The column p represents one particular solution to the nonhomogeneous system because it is the solution produced when the free variables assume the values xf1 = xf2 = · · · = xfn−r = 0. Example 2.5.1 Problem: Determine the general solution of the following nonhomogeneous system and compare it with the general solution of the associated homogeneous system: x1 + x2 + 2x3 + 2x4 + x5 = 1, 2x1 + 2x2 + 4x3 + 4x4 + 3x5 = 1, 2x1 + 2x2 + 4x3 + 4x4 + 2x5 = 2, 3x1 + 5x2 + 8x3 + 6x4 + 5x5 = 3. Solution: Reducing the augmented matrix [A|b] to E[A|b] yields ⎛ 1 ⎜2 A =⎜ 2 3 ⎛ 1 ⎜0 −→⎝ 0 ⎛ 1 ⎜0 −→⎝ 0 0 1 2 2 5 1 2 0 0 2 4 4 8 2 2 0 0 1 0 0 2 4 4 6 2 2 0 0 0 0 0 2 2 0 0 1 3 2 5 1 2 1 0 0 1 1 0 ⎛ 1 1 1 2 2 1 ⎞⎛ 1 0 0 0 0 1 ⎟ ⎜ 0 0 0 0 0 2 0 2 2 0 2 3 ⎞⎛ 1 1 1 2 2 1 0 ⎞⎛ 0 1 1 0 1 ⎟ ⎜ 0 0 0 0 1 −1 0 0 0 0 0 0 0 ⎞⎛ 1 1 0 1 2 0 0 ⎟ ⎜ 0 1 1 0 0 ⎟−→⎝ 0 0 0 0 1 −1 0 0 0 0 0 0 0 ⎞1 −1 ⎟ ⎜ 0 0 ⎞ 1 0 ⎟ ⎜ −1 0 ⎟ ⎟ 1 1 ⎟ ⎟= E[A|b] . −1 0 68 Chapter 2 Rectangular Systems and Echelon Forms Observe that the system is indeed consistent because the last column is nonbasic. Solve the reduced system for the basic variables x1 , x2 , and x5 in terms of the free variables x3 and x4 to obtain x1 = 1 − x3 − 2x4 , x2 = 1 − x3 , x3 is "free," x4 is "free," x5 = −1. The general solution to the nonhomogeneous system is ⎛ ⎞⎛ ⎞⎛ ⎞⎛ ⎞⎛ ⎞x1 1 − x3 − 2x4 −1 −2 1 1 − x3 ⎟ x2 ⎟ ⎜ −1 ⎟ ⎜ 0 ⎟ ⎜ 1 ⎟ ⎜ ⎟ ⎜ ⎟ ⎜ ⎟ ⎜ ⎟ x = ⎜ x3 ⎟ = ⎜ x3 ⎟ = ⎜ 0 ⎟ + x3 ⎜ 1 ⎟ + x4 ⎜ 0 ⎟. ⎝ ⎠⎝ ⎠⎝ ⎠⎝ ⎠⎝ ⎠0 1 x4 x4 0 0 0 x5 −1 −1 The general solution of the associated homogeneous system is ⎛ ⎞⎛ ⎞⎛ ⎞⎛ ⎞x1 −x3 − 2x4 −1 −2 x2 ⎟ ⎜ −x3 ⎟ ⎜ −1 ⎟ ⎜ 0 ⎟ ⎜ ⎟ ⎜ ⎟ ⎜ ⎟ ⎜ ⎟ x = ⎜ x3 ⎟ = ⎜ x3 ⎟ = x3 ⎜ 1 ⎟ + x4 ⎜ 0 ⎟. ⎝ ⎠⎝ ⎠⎝ ⎠⎝ ⎠0 1 x4 x4 0 0 x5 0 You should verify for yourself that ⎛ ⎞1 ⎜ 1 ⎟ ⎜ ⎟ p = ⎜ 0 ⎟ ⎝ ⎠0 −1 is indeed a particular solution to the nonhomogeneous system and that ⎛ ⎞−1 ⎜ −1 ⎟ ⎜ ⎟ h3 = ⎜ 1 ⎟ ⎝ ⎠0 0 ⎛ and ⎞−2 ⎜ 0 ⎟ ⎜ ⎟ h4 = ⎜ 0 ⎟ ⎝ ⎠1 0 are particular solutions to the associated homogeneous system. 2.5 Nonhomogeneous Systems 69 Now turn to the question, "When does a consistent system have a unique solution?" It is known from (2.5.7) that the general solution of a consistent m × n nonhomogeneous system [A|b] with rank (A) = r is given by x = p + xf1 h1 + xf2 h2 + · · · + xfn−r hn−r , where xf1 h1 + xf2 h2 + · · · + xfn−r hn−r is the general solution of the associated homogeneous system. Consequently, it is evident that the nonhomogeneous system [A|b] will have a unique solution (namely, p) if and only if there are no free variables—i.e., if and only if r = n (= number of unknowns)—this is equivalent to saying that the associated homogeneous system [A|0] has only the trivial solution. Example 2.5.2 Consider the following nonhomogeneous system: 2x1 + 4x2 + 6x3 = 2, x1 + 2x2 + 3x3 = 1, x1 + x3 = −3, 2x1 + 4x2 = 8. Reducing [A|b] to E[A|b] yields ⎛ 2 4 6 ⎜1 2 3 [A|b] = ⎝ 1 0 1 2 4 0 ⎞⎛ 2 1 0 0 1 ⎟ ⎜ 0 1 0 ⎠−→⎝ 0 0 1 −3 0 0 0 8 ⎞−2 3 ⎟ ⎠= E[A|b] . −1 0 The system is consistent because the last column is nonbasic. There are several ways to see that the system has a unique solution. Notice that rank (A) = 3 = number of unknowns, which is the same as observing that there are no free variables. Furthermore, the associated homogeneous system clearly has only the trivial solution. Finally, because we completely reduced [A|b] to E[A|b] , ⎛ it is obvious that there is only ⎞−2 one solution possible and that it is given by p = ⎝ 3 ⎠. −1 70 Chapter 2 Rectangular Systems and Echelon Forms Summary Let [A|b] be the augmented matrix for a consistent m × n nonhomogeneous system in which rank (A) = r. • Reducing [A|b] to a row echelon form using Gaussian elimination and then solving for the basic variables in terms of the free variables leads to the general solution x = p + xf1 h1 + xf2 h2 + · · · + xfn−r hn−r . As the free variables xfi range over all possible values, this general solution generates all possible solutions of the system. • Column p is a particular solution of the nonhomogeneous system. • The expression xf1 h1 + xf2 h2 + · · · + xfn−r hn−r is the general solution of the associated homogeneous system. • Column p as well as the columns hi are independent of the row echelon form to which [A|b] is reduced. • The system possesses a unique solution if and only if any of the following is true. rank (A) = n = number of unknowns. There are no free variables. The associated homogeneous system possesses only the trivial solution. Exercises for section 2.5 2.5.1. Determine the general solution for each of the following nonhomogeneous systems. 2x + y + z = 4, x1 + 2x2 + x3 + 2x4 = 3, 4x + 2y + z = 6, (a) 2x1 + 4x2 + x3 + 3x4 = 4, (b) 6x + 3y + z = 8, 3x1 + 6x2 + x3 + 4x4 = 5. 8x + 4y + z = 10. (c) x1 + x2 + 2x3 = 1, 3x1 + 3x3 + 3x4 = 6, 2x1 + x2 + 3x3 + x4 = 3, x1 + 2x2 + 3x3 − x4 = 0. 2x + y + z = 2, (d) 4x + 2y + z = 5, 6x + 3y + z = 8, 8x + 5y + z = 8. 2.5 Nonhomogeneous Systems 71 2.5.2. Among the solutions that satisfy the set of linear equations x1 + x2 + 2x3 + 2x4 + x5 = 1, 2x1 + 2x2 + 4x3 + 4x4 + 3x5 = 1, 2x1 + 2x2 + 4x3 + 4x4 + 2x5 = 2, 3x1 + 5x2 + 8x3 + 6x4 + 5x5 = 3, find all those that also satisfy the following two constraints: (x1 − x2)2 − 4x25 = 0, x23 − x25 = 0. 2.5.3. In order to grow a certain crop, it is recommended that each square foot of ground be treated with 10 units of phosphorous, 9 units of potassium, and 19 units of nitrogen. Suppose that there are three brands of fertilizer on the market—say brand X , brand Y, and brand Z. One pound of brand X contains 2 units of phosphorous, 3 units of potassium, and 5 units of nitrogen. One pound of brand Y contains 1 unit of phosphorous, 3 units of potassium, and 4 units of nitrogen. One pound of brand Z contains only 1 unit of phosphorous and 1 unit of nitrogen. (a) Take into account the obvious fact that a negative number of pounds of any brand can never be applied, and suppose that because of the way fertilizer is sold only an integral number of pounds of each brand will be applied. Under these constraints, determine all possible combinations of the three brands that can be applied to satisfy the recommendations exactly. (b) Suppose that brand X costs $1 per pound, brand Y costs $6 per pound, and brand Z costs $3 per pound. Determine the least expensive solution that will satisfy the recommendations exactly as well as the constraints of part (a). 2.5.4. Consider the following system: 2x + 2y + 3z = 0, 4x + 8y + 12z = −4, 6x + 2y + αz = 4. (a) Determine all values of α for which the system is consistent. (b) Determine all values of α for which there is a unique solution, and compute the solution for these cases. (c) Determine all values of α for which there are infinitely many different solutions, and give the general solution for these cases. 72 Chapter 2 Rectangular Systems and Echelon Forms 2.5.5. If columns s1 and s2 are particular solutions of the same nonhomogeneous system, must it be the case that the sum s1 + s2 is also a solution? 2.5.6. Suppose that [A|b] is the augmented matrix for a consistent system of m equations in n unknowns where m ≥ n. What must EA look like when the system possesses a unique solution? 2.5.7. Construct a nonhomogeneous system of three equations in four unknowns that has ⎛ ⎞⎛ ⎞⎛ ⎞1 −2 −3 ⎟ 0 ⎟ 1 ⎟ 0 ⎟ ⎝ ⎠ + x2 ⎝ ⎠ + x4 ⎝ ⎠1 0 2 0 0 1 as its general solution. 2.5.8. Consider using floating-point arithmetic (without partial pivoting or scaling) to solve the system represented by the following augmented matrix: ⎛ ⎞.835 .667 .5 .168 ⎝ .333 .266 .1994 .067 ⎠. 1.67 1.334 1.1 .436 (a) Determine the 4-digit general solution. (b) Determine the 5-digit general solution. (c) Determine the 6-digit general solution. 2.6 Electrical Circuits 2.6 73 ELECTRICAL CIRCUITS The theory of electrical circuits is an important application that naturally gives rise to rectangular systems of linear equations. Because the underlying mathematics depends on several of the concepts discussed in the preceding sections, you may find it interesting and worthwhile to make a small excursion into the elementary mathematical analysis of electrical circuits. However, the continuity of the text is not compromised by omitting this section. In a direct current circuit containing resistances and sources of electromotive force (abbreviated EMF) such as batteries, a point at which three or more conductors are joined is called a node or branch point of the circuit, and a closed conduction path is called a loop. Any part of a circuit between two adjoining nodes is called a branch of the circuit. The circuit shown in Figure 2.6.1 is a typical example that contains four nodes, seven loops, and six branches. E1 E2 R1 I1 I2 A B R5 E3 R3 2 R2 1 I5 R6 3 I3 4 I6 C I4 R4 E4 Figure 2.6.1 The problem is to relate the currents Ik in each branch to the resistances Rk 15 and the EMFs Ek . This is accomplished by using Ohm's law in conjunction with Kirchhoff 's rules to produce a system of linear equations. Ohm's Law Ohm's law states that for a current of I amps, the voltage drop (in volts) across a resistance of R ohms is given by V = IR. Kirchhoff's rules—formally stated below—are the two fundamental laws that govern the study of electrical circuits. 15 For an EMF source of magnitude E and a current I, there is always a small internal resistance in the source, and the voltage drop across it is V = E −I ×(internal resistance). But internal source resistance is usually negligible, so the voltage drop across the source can be taken as V = E. When internal resistance cannot be ignored, its effects may be incorporated into existing external resistances, or it can be treated as a separate external resistance. 74 Chapter 2 Rectangular Systems and Echelon Forms Kirchhoff's Rules NODE RULE: The algebraic sum of currents toward each node is zero. That is, the total incoming current must equal the total outgoing current. This is simply a statement of conservation of charge. LOOP RULE: The algebraic sum of the EMFs around each loop must equal the algebraic sum of the IR products in the same loop. That is, assuming internal source resistances have been accounted for, the algebraic sum of the voltage drops over the sources equals the algebraic sum of the voltage drops over the resistances in each loop. This is a statement of conservation of energy. Kirchhoff's rules may be used without knowing the directions of the currents and EMFs in advance. You may arbitrarily assign directions. If negative values emerge in the final solution, then the actual direction is opposite to that assumed. To apply the node rule, consider a current to be positive if its direction is toward the node—otherwise, consider the current to be negative. It should be clear that the node rule will always generate a homogeneous system. For example, applying the node rule to the circuit in Figure 2.6.1 yields four homogeneous equations in six unknowns—the unknowns are the Ik 's: Node 1: I1 − I2 − I5 = 0, Node 2: − I1 − I3 + I4 = 0, Node 3: I3 + I5 + I6 = 0, Node 4: I2 − I4 − I6 = 0. To apply the loop rule, some direction (clockwise or counterclockwise) must be chosen as the positive direction, and all EMFs and currents in that direction are considered positive and those in the opposite direction are negative. It is possible for a current to be considered positive for the node rule but considered negative when it is used in the loop rule. If the positive direction is considered to be clockwise in each case, then applying the loop rule to the three indicated loops A, B, and C in the circuit shown in Figure 2.6.1 produces the three nonhomogeneous equations in six unknowns—the Ik 's are treated as the unknowns, while the Rk 's and Ek 's are assumed to be known. Loop A: I1 R1 − I3 R3 + I5 R5 = E1 − E3 , Loop B: I2 R2 − I5 R5 + I6 R6 = E2 , Loop C: I3 R3 + I4 R4 − I6 R6 = E3 + E4 . 2.6 Electrical Circuits 75 There are 4 additional loops that also produce loop equations thereby making a total of 11 equations (4 nodal equations and 7 loop equations) in 6 unknowns. Although this appears to be a rather general 11 × 6 system of equations, it really is not. If the circuit is in a state of equilibrium, then the physics of the situation dictates that for each set of EMFs Ek , the corresponding currents Ik must be uniquely determined. In other words, physics guarantees that the 11 × 6 system produced by applying the two Kirchhoff rules must be consistent and possess a unique solution. Suppose that [A|b] represents the augmented matrix for the 11 × 6 system generated by Kirchhoff's rules. From the results in §2.5, we know that the system has a unique solution if and only if rank (A) = number of unknowns = 6. Furthermore, it was demonstrated in §2.3 that the system is consistent if and only if rank[A|b] = rank (A). Combining these two facts allows us to conclude that rank[A|b] = 6 so that when [A|b] is reduced to E[A|b] , there will be exactly 6 nonzero rows and 5 zero rows. Therefore, 5 of the original 11 equations are redundant in the sense that they can be "zeroed out" by forming combinations of some particular set of 6 "independent" equations. It is desirable to know beforehand which of the 11 equations will be redundant and which can act as the "independent" set. Notice that in using the node rule, the equation corresponding to node 4 is simply the negative sum of the equations for nodes 1, 2, and 3, and that the first three equations are independent in the sense that no one of the three can be written as a combination of any other two. This situation is typical. For a general circuit with n nodes, it can be demonstrated that the equations for the first n − 1 nodes are independent, and the equation for the last node is redundant. The loop rule also can generate redundant equations. Only simple loops— loops not containing smaller loops—give rise to independent equations. For example, consider the loop consisting of the three exterior branches in the circuit shown in Figure 2.6.1. Applying the loop rule to this large loop will produce no new information because the large loop can be constructed by "adding" the three simple loops A, B, and C contained within. The equation associated with the large outside loop is I1 R1 + I2 R2 + I4 R4 = E1 + E2 + E4 , which is precisely the sum of the equations that correspond to the three component loops A, B, and C. This phenomenon will hold in general so that only the simple loops need to be considered when using the loop rule. 76 Chapter 2 Rectangular Systems and Echelon Forms The point of this discussion is to conclude that the more general 11 × 6 rectangular system can be replaced by an equivalent 6 × 6 square system that has a unique solution by dropping the last nodal equation and using only the simple loop equations. This is characteristic of practical work in general. The physics of a problem together with natural constraints can usually be employed to replace a general rectangular system with one that is square and possesses a unique solution. One of the goals in our study is to understand more clearly the notion of "independence" that emerged in this application. So far, independence has been an intuitive idea, but this example helps make it clear that independence is a fundamentally important concept that deserves to be nailed down more firmly. This is done in §4.3, and the general theory for obtaining independent equations from electrical circuits is developed in Examples 4.4.6 and 4.4.7. Exercises for section 2.6 2.6.1. Suppose that Ri = i ohms and Ei = i volts in the circuit shown in Figure 2.6.1. (a) Determine the six indicated currents. (b) Select node number 1 to use as a reference point and fix its potential to be 0 volts. With respect to this reference, calculate the potentials at the other three nodes. Check your answer by verifying the loop rule for each loop in the circuit. 2.6.2. Determine the three currents indicated in the following circuit. 5Ω 8Ω I2 I1 12 volts 1Ω 1Ω 9 volts 10Ω I3 2.6.3. Determine the two unknown EMFs in the following circuit. 20 volts 6Ω 1 amp 4Ω E1 2 amps E2 2Ω 2.6 Electrical Circuits 77 2.6.4. Consider the circuit shown below and answer the following questions. R2 R3 R5 R1 R4 R6 I E (a) How many nodes does the circuit contain? (b) How many branches does the circuit contain? (c) Determine the total number of loops and then determine the number of simple loops. (d) Demonstrate that the simple loop equations form an "independent" system of equations in the sense that there are no redundant equations. (e) Verify that any three of the nodal equations constitute an "independent" system of equations. (f) Verify that the loop equation associated with the loop containing R1 , R2 , R3 , and R4 can be expressed as the sum of the two equations associated with the two simple loops contained in the larger loop. (g) Determine the indicated current I if R1 = R2 = R3 = R4 = 1 ohm, R5 = R6 = 5 ohms, and E = 5 volts. CHAPTER 3 Matrix Algebra 3.1 FROM ANCIENT CHINA TO ARTHUR CAYLEY The ancient Chinese appreciated the advantages of array manipulation in dealing with systems of linear equations, and they possessed the seed that might have germinated into a genuine theory of matrices. Unfortunately, in the year 213 B.C., emperor Shih Hoang-ti ordered that "all books be burned and all scholars be buried." It is presumed that the emperor wanted all knowledge and written records to begin with him and his regime. The edict was carried out, and it will never be known how much knowledge was lost. The book Chiu-chang Suan-shu (Nine Chapters on Arithmetic), mentioned in the introduction to Chapter 1, was compiled on the basis of remnants that survived. More than a millennium passed before further progress was documented. The Chinese counting board with its colored rods and its applications involving array manipulation to solve linear systems eventually found its way to Japan. Seki Kowa (1642–1708), whom many Japanese consider to be one of the greatest mathematicians that their country has produced, carried forward the Chinese principles involving "rule of thumb" elimination methods on arrays of numbers. His understanding of the elementary operations used in the Chinese elimination process led him to formulate the concept of what we now call the determinant. While formulating his ideas concerning the solution of linear systems, Seki Kowa anticipated the fundamental concepts of array operations that today form the basis for matrix algebra. However, there is no evidence that he developed his array operations to actually construct an algebra for matrices. From the middle 1600s to the middle 1800s, while Europe was flowering in mathematical development, the study of array manipulation was exclusively 80 Chapter 3 Matrix Algebra dedicated to the theory of determinants. Curiously, matrix algebra did not evolve along with the study of determinants. It was not until the work of the British mathematician Arthur Cayley (1821–1895) that the matrix was singled out as a separate entity, distinct from the notion of a determinant, and algebraic operations between matrices were defined. In an 1855 paper, Cayley first introduced his basic ideas that were presented mainly to simplify notation. Finally, in 1857, Cayley expanded on his original ideas and wrote A Memoir on the Theory of Matrices. This laid the foundations for the modern theory and is generally credited for being the birth of the subjects of matrix analysis and linear algebra. Arthur Cayley began his career by studying literature at Trinity College, Cambridge (1838–1842), but developed a side interest in mathematics, which he studied in his spare time. This "hobby" resulted in his first mathematical paper in 1841 when he was only 20 years old. To make a living, he entered the legal profession and practiced law for 14 years. However, his main interest was still mathematics. During the legal years alone, Cayley published almost 300 papers in mathematics. In 1850 Cayley crossed paths with James J. Sylvester, and between the two of them matrix theory was born and nurtured. The two have been referred to as the "invariant twins." Although Cayley and Sylvester shared many mathematical interests, they were quite different people, especially in their approach to mathematics. Cayley had an insatiable hunger for the subject, and he read everything that he could lay his hands on. Sylvester, on the other hand, could not stand the sight of papers written by others. Cayley never forgot anything he had read or seen—he became a living encyclopedia. Sylvester, so it is said, would frequently fail to remember even his own theorems. In 1863, Cayley was given a chair in mathematics at Cambridge University, and thereafter his mathematical output was enormous. Only Cauchy and Euler were as prolific. Cayley often said, "I really love my subject," and all indications substantiate that this was indeed the way he felt. He remained a working mathematician until his death at age 74. Because the idea of the determinant preceded concepts of matrix algebra by at least two centuries, Morris Kline says in his book Mathematical Thought from Ancient to Modern Times that "the subject of matrix theory was well developed before it was created." This must have indeed been the case because immediately after the publication of Cayley's memoir, the subjects of matrix theory and linear algebra virtually exploded and quickly evolved into a discipline that now occupies a central position in applied mathematics. 3.2 Addition and Transposition 3.2 81 ADDITION AND TRANSPOSITION In the previous chapters, matrix language and notation were used simply to formulate some of the elementary concepts surrounding linear systems. The purpose now is to turn this language into a mathematical theory. Unless otherwise stated, a scalar is a complex number. Real numbers are a subset of the complex numbers, and hence real numbers are also scalar quantities. In the early stages, there is little harm in thinking only in terms of real scalars. Later on, however, the necessity for dealing with complex numbers will be unavoidable. Throughout the text, will denote the set of real numbers, and C will denote the complex numbers. The set of all n-tuples of real numbers will be denoted by n, and the set of all complex n-tuples will be denoted by C n. For example, 2 is the set of all ordered pairs of real numbers (i.e., the standard cartesian plane), and 3 is ordinary 3-space. Analogously, m×n and C m×n denote the m × n matrices containing real numbers and complex numbers, respectively. Matrices A = [aij] and B = [bij] are defined to be equal matrices when A and B have the same shape and corresponding entries are equal. That is, aij = bij for each i = 1, 2, . . . , m and j = 1, 2, . . . , n. In particular, this definition applies to arrays such as u = (1 2 3) and v = (1 2 3). Even though u and v describe exactly the same point in 3-space, we cannot consider them to be equal matrices because they have different shapes. An array (or matrix) consisting of a single column, such as u, is called a column vector, while an array consisting of a single row, such as v, is called a row vector. Addition of Matrices If A and B are m × n matrices, the sum of A and B is defined to be the m × n matrix A + B obtained by adding corresponding entries. That is, [A + B]ij = [A]ij + [B]ij For example, −2 x 3 2 + z 3 4 −y −3 16 for each i and j. 1 − x −2 4+x 4+y = 0 z 1 18+x 4 . The great French mathematician Pierre-Simon Laplace (1749–1827) said that, "Such is the advantage of a well-constructed language that its simplified notation often becomes the source of profound theories." The theory of matrices is a testament to the validity of Laplace's statement. 82 Chapter 3 Matrix Algebra The symbol "+" is used two different ways—it denotes addition between scalars in some places and addition between matrices at other places. Although these are two distinct algebraic operations, no ambiguities will arise if the context in which "+" appears is observed. Also note that the requirement that A and B have the same shape prevents adding a row to a column, even though the two may contain the same number of entries. The matrix (−A), called the additive inverse of A, is defined to be the matrix obtained by negating each entry of A. That is, if A = [aij], then −A = [−aij]. This allows matrix subtraction to be defined in the natural way. For two matrices of the same shape, the difference A − B is defined to be the matrix A − B = A + (−B) so that [A − B]ij = [A]ij − [B]ij for each i and j. Since matrix addition is defined in terms of scalar addition, the familiar algebraic properties of scalar addition are inherited by matrix addition as detailed below. Properties of Matrix Addition For m × n matrices A, B, and C, the following properties hold. Closure property: Associative property: A + B is again an m × n matrix. (A + B) + C = A + (B + C). Commutative property: A + B = B + A. Additive identity: The m × n matrix 0 consisting of all zeros has the property that A + 0 = A. Additive inverse: The m × n matrix (−A) has the property that A + (−A) = 0. Another simple operation that is derived from scalar arithmetic is as follows. Scalar Multiplication The product of a scalar α times a matrix A, denoted by αA, is defined to be the matrix obtained by multiplying each entry of A by α. That is, [αA]ij = α[A]ij for each i and j. For example, (1 2 2 0 1 4) 2 (1 1 2 3) = (2 4 6 2 0 8 2 4 2) . The rules for combining addition and scalar multiplication are what you might suspect they should be. Some of the important ones are listed below. 3.2 Addition and Transposition 83 Properties of Scalar Multiplication For m × n matrices A and B and for scalars α and β, the following properties hold. Closure property: Associative property: αA is again an m × n matrix. (αβ)A = α(βA). Distributive property: α(A + B) = αA + αB. Scalar multiplication is distributed over matrix addition. Distributive property: (α + β)A = αA + βA. Scalar multiplication is distributed over scalar addition. Identity property: 1A = A. The number 1 is an identity element under scalar multiplication. Other properties such as αA = Aα could have been listed, but the properties singled out pave the way for the definition of a vector space on p. 160. A matrix operation that's not derived from scalar arithmetic is transposition as defined below. Transpose The transpose of Am×n is defined to be the n × m matrix AT obtained by interchanging rows and columns in A. More precisely, if A = [aij], then [AT]ij = aji. For example, (1 3 5) T = (2 4 6) (1 2 3 4 5 6) . It should be evident that for all matrices, AT T = A. Whenever a matrix contains complex entries, the operation of complex conjugation almost always accompanies the transpose operation. (Recall that the complex conjugate of z = a + ib is defined to be z̄ = a − ib.) 84 Chapter 3 Matrix Algebra Conjugate Transpose For A = [aij], the conjugate matrix is defined to be Ā = [āij], and the conjugate transpose of A is defined to be ĀT = AT . From now on, ĀT will be denoted by A∗ , so [A∗]ij = āji. For example, 1 − 4i 2 3 + 4i 1 + 3i i 2 = −1 2 − i 0 2 + i . Fact A∗ = AT whenever A contains only real entries. Sometimes the matrix A∗ is called the adjoint of A. The transpose (and conjugate transpose) operation is easily combined with matrix addition and scalar multiplication. The basic rules are given below. Properties of the Transpose If A and B are two matrices of the same shape, and if α is a scalar, then each of the following statements is true. T (A + B) = AT + BT T (αA) = αAT and and ∗ (A + B) = A∗ + B∗ , ∗ (αA) = ᾱA∗ . (3.2.1) (3.2.2) 17 Proof. We will prove that (3.2.1) and (3.2.2) hold for the transpose operation. The proofs of the statements involving conjugate transposes are similar and are left as exercises. For each i and j, it is true that [(A + B)T]ij = [A + B]ji = [A]ji + [B]ji = [AT]ij + [BT]ij = [AT + BT]ij. 17 Computers can outperform people in many respects in that they do arithmetic much faster and more accurately than we can, and they are now rather adept at symbolic computation and mechanical manipulation of formulas. But computers can't do mathematics—people still hold the monopoly. Mathematics emanates from the uniquely human capacity to reason abstractly in a creative and logical manner, and learning mathematics goes hand-in-hand with learning how to reason abstractly and create logical arguments. This is true regardless of whether your orientation is applied or theoretical. For this reason, formal proofs will appear more frequently as the text evolves, and it is expected that your level of comprehension as well as your ability to create proofs will grow as you proceed. 3.2 Addition and Transposition 85 This proves that corresponding entries in (A + B)T and AT + BT are equal, so it must be the case that (A + B)T = AT + BT . Similarly, for each i and j, [(αA)T]ij = [αA]ji = α[A]ji = α[AT]ij =⇒ (αA)T = αAT . Sometimes transposition doesn't change anything. For example, if (1 2 3 A = 2 4 5), then AT = A. 3 5 6 This is because the entries in A are symmetrically located about the main diagonal—the line from the upper-left-hand corner to the lower-right-hand corner. (λ1 0 · · · 0) Matrices of the form D = (0 . . . 0 λ2 · · · 0 . . .) . . . λn are called diagonal matrices, and they are clearly symmetric in the sense that D = DT. This is one of several kinds of symmetries described below. Symmetries Let A = [aij] be a square matrix. • A is said to be a symmetric matrix whenever A = AT , i.e., whenever aij = aji. • A is said to be a skew-symmetric matrix whenever A = −AT , i.e., whenever aij = −aji. • A is said to be a hermitian matrix whenever A = A∗ , i.e., whenever aij = āji. This is the complex analog of symmetry. • A is said to be a skew-hermitian matrix when A = −A∗ , i.e., whenever aij = −āji . This is the complex analog of skew symmetry. For example, consider (1 2 + 4i 1 − 3i A = (2 − 4i 3 8 + 6i) 1 + 3i 8 − 6i 5) and B = (1 2 + 4i 1 − 3i 2 + 4i 3 8 + 6i 1 − 3i 8 + 6i 5) Can you see that A is hermitian but not symmetric, while B is symmetric but not hermitian? Nature abounds with symmetry, and very often physical symmetry manifests itself as a symmetric matrix in a mathematical model. The following example is an illustration of this principle. 86 Chapter 3 Matrix Algebra Example 3.2.1 Consider two springs that are connected as shown in Figure 3.2.1. Node 1 k1 x1 Node 2 k2 Node 3 x2 F1 -F1 x3 -F3 F3 Figure 3.2.1 The springs at the top represent the "no tension" position in which no force is being exerted on any of the nodes. Suppose that the springs are stretched or compressed so that the nodes are displaced as indicated in the lower portion of Figure 3.2.1. Stretching or compressing the springs creates a force on each 18 node according to Hooke's law that says that the force exerted by a spring is F = kx, where x is the distance the spring is stretched or compressed and where k is a stiffness constant inherent to the spring. Suppose our springs have stiffness constants k1 and k2 , and let Fi be the force on node i when the springs are stretched or compressed. Let's agree that a displacement to the left is positive, while a displacement to the right is negative, and consider a force directed to the right to be positive while one directed to the left is negative. If node 1 is displaced x1 units, and if node 2 is displaced x2 units, then the left-hand spring is stretched (or compressed) by a total amount of x1 − x2 units, so the force on node 1 is F1 = k1 (x1 − x2). Similarly, if node 2 is displaced x2 units, and if node 3 is displaced x3 units, then the right-hand spring is stretched by a total amount of x2 − x3 units, so the force on node 3 is F3 = −k2 (x2 − x3). The minus sign indicates the force is directed to the left. The force on the lefthand side of node 2 is the opposite of the force on node 1, while the force on the right-hand side of node 2 must be the opposite of the force on node 3. That is, F2 = −F1 − F3 . 18 Hooke's law is named for Robert Hooke (1635–1703), an English physicist, but it was generally known to several people (including Newton) before Hooke's 1678 claim to it was made. Hooke was a creative person who is credited with several inventions, including the wheel barometer, but he was reputed to be a man of "terrible character." This characteristic virtually destroyed his scientific career as well as his reputation. It is said that he lacked mathematical sophistication and that he left much of his work in incomplete form, but he bitterly resented people who built on his ideas by expressing them in terms of elegant mathematical formulations. 3.2 Addition and Transposition 87 Organize the above three equations as a linear system: k1 x1 − k1 x2 = F1 , −k1 x1 + (k1 + k2)x2 − k2 x3 = F2 , −k2 x2 + k2 x3 = F3 , and observe that the coefficient matrix, called the stiffness matrix, (k1 −k1 0 K = (−k1 k1 + k2 −k2) 0 −k2 k2 is a symmetric matrix. The point of this example is that symmetry in the physical problem translates to symmetry in the mathematics by way of the symmetric matrix K. When the two springs are identical (i.e., when k1 = k2 = k), even more symmetry is present, and in this case (1 −1 0 K = k (−1 2 −1). 0 −1 1 Exercises for section 3.2 3.2.1. Determine the unknown quantities in the following expressions. T 0 3x+2 y+3 6 (a) 3X = (2 x + 3 6 9 3 0) y 3 2.2.2 . Identify each of the following as symmetric, skew-symmetric, or neither. (1 −3 3 0 −3 −3 (a) (−3 4 −3) , (b) (3 0 1) , (c) −3 3 0 1 (1 0) (d) 1 2 0 . 3.2.3. Construct an example of a 3 × 3 matrix A that satisfies the following conditions. (a) A is both symmetric and skew-symmetric. (b) A is both hermitian and symmetric. (c) A is skew-hermitian. 88 Chapter 3 Matrix Algebra 3.2.4. Explain why the set of all n × n symmetric matrices is closed under matrix addition. That is, explain why the sum of two n × n symmetric matrices is again an n × n symmetric matrix. Is the set of all n × n skew-symmetric matrices closed under matrix addition? 3.2.5. Prove that each of the following statements is true. (a) If A = [aij] is skew symmetric, then ajj = 0 for each j. (b) If A = [aij] is skew hermitian, then each ajj is a pure imaginary number—i.e., a multiple of the imaginary unit i. (c) If A is real and symmetric, then B = iA is skew hermitian. 3.2.6. Let A be any square matrix. (a) Show that A+AT is symmetric and A−AT is skew symmetric. (b) Prove that there is one and only one way to write A as the sum of a symmetric matrix and a skew-symmetric matrix. 3.2.7. If A and B are two matrices of the same shape, prove that each of the following statements is true. (a) (A + B)∗ = A∗ + B∗ . (b) (αA)∗ = ᾱA∗ . 3.2.8. Using the conventions given in Example 3.2.1, determine the stiffness matrix for a system of n identical springs, with stiffness constant k, connected in a line similar to that shown in Figure 3.2.1. 3.3 Linearity 3.3 89 LINEARITY The concept of linearity is the underlying theme of our subject. In elementary mathematics the term "linear function" refers to straight lines, but in higher mathematics linearity means something much more general. Recall that a function f is simply a rule for associating points in one set D—called the domain of f —to points in another set R—the range of f . A linear function is a particular type of function that is characterized by the following two properties. Linear Functions Suppose that D and R are sets that possess an addition operation as well as a scalar multiplication operation—i.e., a multiplication between scalars and set members. A function f that maps points in D to points in R is said to be a linear function whenever f satisfies the conditions that f (x + y) = f (x) + f (y) (3.3.1) and f (αx) = αf (x) (3.3.2) for every x and y in D and for all scalars α. These two conditions may be combined by saying that f is a linear function whenever f (αx + y) = αf (x) + f (y) for all scalars α and for all x, y ∈ D. One of the simplest linear functions is f (x) = αx, whose graph in 2 is a straight line through the origin. You should convince yourself that f is indeed a linear function according to the above definition. However, f (x) = αx + β does not qualify for the "linear function" title—it is a linear function that has been translated by a constant β. Translations of linear functions are referred to as affine functions. Virtually all information concerning affine functions can be derived from an understanding of linear functions, and consequently we will focus only on issues of linearity. In 3 , the surface described by a function of the form f (x1 , x2) = α1 x1 + α2 x2 is a plane through the origin, and it is easy to verify that f is a linear function. For β = 0, the graph of f (x1 , x2) = α1 x1 + α2 x2 + β is a plane not passing through the origin, and f is no longer a linear function—it is an affine function. 90 Chapter 3 Matrix Algebra In 2 and 3 , the graphs of linear functions are lines and planes through the origin, and there seems to be a pattern forming. Although we cannot visualize higher dimensions with our eyes, it seems reasonable to suggest that a general linear function of the form f (x1 , x2 , . . . , xn) = α1 x1 + α2 x2 + · · · + αn xn somehow represents a "linear" or "flat" surface passing through the origin 0 = (0, 0, . . . , 0) in n+1 . One of the goals of the next chapter is to learn how to better interpret and understand this statement. Linearity is encountered at every turn. For example, the familiar operations of differentiation and integration may be viewed as linear functions. Since d(f + g) df dg = + dx dx dx the differentiation operator Dx (f) = df /dx is linear. Similarly, (f + g)dx = f dx + gdx means that the integration operator I(f) = f dx is linear. There are several important matrix functions that are linear. For example, the transposition function f (Xm×n) = XT is linear because T (A + B) = AT + BT and T (αA) = αAT (recall (3.2.1) and (3.2.2)). Another matrix function that is linear is the trace function presented below. Example 3.3.1 The trace of an n × n matrix A = [aij] is defined to be the sum of the entries lying on the main diagonal of A. That is, trace (A) = a11 + a22 + · · · + ann = n aii . i=1 Problem: Show that f (Xn×n) = trace (X) is a linear function. Solution: Let's be efficient by showing that (3.3.3) holds. Let A = [aij] and B = [bij], and write f (αA + B) = trace (αA + B) = n [αA + B]ii = n (αaii + bii) = n αaii + n bii = α n aii + n bii = α trace (A) + trace (B) = αf (A) + f (B). 3.3 Linearity 91 Example 3.3.2 Consider a linear system a11 x1 + a12 x2 + · · · + a1n xn = u1 , a21 x1 + a22 x2 + · · · + a2n xn = u2 , ... am1 x1 + am2 x2 + · · · + amn xn = um , to be a function u = f (x) that maps x = (x1 x2 . . . xn) ∈ n to u = (u1 u2 . . . um) ∈ m . Problem: Show that u = f (x) is linear. Solution: Let A = [aij] be the matrix of coefficients, and write \αx1 + y1 n n n (αx2 + y2) f (αx + y) = f (...) = Σ (αxj + yj)A∗j = Σ (αxj A∗j + yj A∗j) = Σ αxj A∗j + Σ yj A∗j = α Σ xj A∗j + Σ yj A∗j = αf (x) + f (y). According to (3.3.3), the function f is linear. The following terminology will be used from now on. Linear Combinations For scalars αj and matrices Xj , the expression α1 X1 + α2 X2 + · · · + αn Xn = n αj Xj is called a linear combination of the Xj 's. 92 Chapter 3 Matrix Algebra Exercises for section 3.3 3.3.1. Each of the following is a function from 2 into 2 . Determine which are linear functions. x x y x (a) f = . (b) f = . y 1 + y y x 0 x x2 x (c) f = . (d) f = . y xy y y x x sin x (e) f = . (f) f = . y y x + y 3.3.2. Let f : 2 → 2 be a linear function. Explain why f must be linear if you know what f does to 1 and 0 . 3.3.3. Give two distinct physical examples of linear functions. 3.3.4. Determine which of the following three transformations in 2 are linear. p f(p) θ p f(p) Rotate counterclockwise through an angle θ. Reflect about the x -axis. Project onto the line y = x. 3.4 Why Do It This Way 3.4 93 WHY DO IT THIS WAY If you were given the task of formulating a definition for composing two matrices A and B in some sort of "natural" multiplicative fashion, your first attempt would probably be to compose A and B by multiplying corresponding entries—much the same way matrix addition is defined. Asked then to defend the usefulness of such a definition, you might be hard pressed to provide a truly satisfying response. Unless a person is in the right frame of mind, the issue of deciding how to best define matrix multiplication is not at all transparent, especially if it is insisted that the definition be both "natural" and "useful." The world had to wait for Arthur Cayley to come to this proper frame of mind. As mentioned in §3.1, matrix algebra appeared late in the game. Manipulation on arrays and the theory of determinants existed long before Cayley and his theory of matrices. Perhaps this can be attributed to the fact that the "correct" way to multiply two matrices eluded discovery for such a long time. 19 Around 1855, Cayley became interested in composing linear functions. In particular, he was investigating linear functions of the type discussed in Example 3.3.2. Typical examples of two such functions are x1 ax1 + bx2 x1 Ax1 + Bx2 f(x) = f and g(x) = g = = x2 cx1 + dx2 x2 Cx1 + Dx2 . Consider, as Cayley did, composing f and g to create another linear function Ax1 + Bx2 (aA + bC)x1 + (aB + bD)x2 h(x) = f g(x) = f = = Cx1 + Dx2 (cA + dC)x1 + (cB + dD)x2 . 19 It was not until 1858 that Cayley's idea to use matrices of coefficients to represent these linear functions. That is, f, g, and h are represented by a b A B aA + bC aB + bD F= , G= , and H = . c d C D cA + dC cB + dD After making this association, it was only natural for Cayley to call H the composition (or product) of F and G, and to write a b A B aA + bC aB + bD = . (3.4.1) c d C D cA + dC cB + dD In other words, the product of two matrices represents the composition of the two associated linear functions. By means of this observation, Cayley brought to life the subjects of matrix analysis and linear algebra. 19 Cayley was not the first to compose linear functions. In fact, Gauss used these compositions as early as 1801, but not in the form of an array of coefficients. Cayley was the first to make the connection between the composition of linear functions and the composition of the associated matrices. Cayley's work from 1855 to 1857 is regarded as being the birth of our subject. 94 Chapter 3 Matrix Algebra Exercises for section 3.4 Each problem in this section concerns the following three linear transformations in 2 . f(p) Rotation: Rotate points counterclockwise through an angle θ. p Reflection: Reflect points about the x -axis. p f(p) Projection: Project points onto the line y = x in a perpendicular manner. f(p) p 3.4.1. Determine the matrix associated with each of these linear functions. That is, determine the aij 's such that f (p) = f x1 x2 = a11 x1 + a12 x2 a21 x1 + a22 x2 . 3.4.2. By using matrix multiplication, determine the linear function obtained by performing a rotation followed by a reflection. 3.4.3. By using matrix multiplication, determine the linear function obtained by first performing a reflection, then a rotation, and finally a projection. 3.5 Matrix Multiplication 3.5 95 MATRIX MULTIPLICATION The purpose of this section is to further develop the concept of matrix multiplication as introduced in the previous section. In order to do this, it is helpful to begin by composing a single row with a single column. If (c1) R = (r1 r2 · · · rn) and C = (c2 ... cn), the standard inner product of R with C is defined to be the scalar RC = r1 c1 + r2 c2 + · · · + rn cn = n ri ci . i=1 For example, (2) (1 4 −2) = 2 (1)(1) + (4)(2) + (−2)(3) = 3. 3 Recall from (3.4.1) that the product of two 2 × 2 matrices a b A B F= and G = c d C D was defined naturally by writing FG = a b c d A C B D = aA + bC cA + dC aB + bD cB + dD = H. Notice that the (i, j) -entry in the product H can be described as the inner product of the ith row of F with the j th column in G. That is, h11 = F1∗ G∗1 = (a b) A C , h21 = F2∗ G∗1 = (c d) A C , h12 = F1∗ G∗2 = (a b) B D , h22 = F2∗ G∗2 = (c d) B D D . This is exactly the way that the general definition of matrix multiplication is formulated. 96 Chapter 3 Matrix Algebra Matrix Multiplication • Matrices A and B are said to be conformable for multiplication in the order AB whenever A has exactly as many columns as B has rows—i.e., A is m × p and B is p × n. • For conformable matrices Am×p = [aij] and Bp×n = [bij], the matrix product AB is the m × n matrix whose (i, j) -entry is the inner product of the ith row of A with the j th column in B. That is, [AB]ij = Ai∗ B∗j = ai1 b1j + ai2 b2j + · · · + aip bpj = aik bkj . k=1 • In case A and B fail to be conformable—i.e., A is m × p and B is q × n with p = q —then no product AB is defined. For example, if a11 a12 a13 b11 b12 b13 b14 A= and B = b21 b22 b23 b24 a21 a22 a23 2×3 b31 b32 b33 b34 3×4 ↑ inside ones match ↑ shape of the product then the product AB exists and has shape 2 × 4. Consider a typical entry of this product, say, the (2,3)-entry. The definition says [AB]23 is obtained by forming the inner product of the second row of A with the third column of B b11 b12 b13 b14 a11 a12 a13 b21 b22 b23 b24 a21 a22 a23 b31 b32 b33 b34 , so [AB]23 = A2∗ B∗3 = a21 b13 + a22 b23 + a23 b33 = 3 k=1 a2k bk3 . 3.5 Matrix Multiplication 97 For example, A = (2 −3 1 −4 0 5) , B = (1 3 −3 2 5 −1 3 0 −2 8) =⇒ AB = (8 3 −7 7 −9 4) . Notice that in spite of the fact that the product AB exists, the product BA is not defined—matrix B is 3 × 4 and A is 2 × 3, and the inside dimensions don't match in this order. Even when the products AB and BA each exist and have the same shape, they need not be equal. For example, 1 1 1 0 0 2 −2 A = , B = =⇒ AB = , BA = (3.5.1) −1 1 1 0 0 0 2 −2 This distinguishing feature is a primary difference between scalar and matrix algebra. Matrix Multiplication Is Not Commutative Matrix multiplication is a noncommutative operation—i.e., it is possible for AB = BA, even when both products exist and have the same shape. There are other major differences between multiplication of matrices and the multiplication of scalars. For scalars, αβ = 0 implies α = 0 or β = 0. (3.5.2) However, the analogous statement for matrices does not hold—the matrices given in (3.5.1) show that it is possible for AB = 0 with A = 0 and B = 0. Related to this issue is a rule sometimes known as the cancellation law. For scalars, this law says that αβ = αγ and α = 0 implies β = γ. (3.5.3) This is true because we invoke (3.5.2) to deduce that α(β − γ) = 0 implies β − γ = 0. Since (3.5.2) does not hold for matrices, we cannot expect (3.5.3) to hold for matrices. Example 3.5.1 The cancellation law (3.5.3) fails for matrix multiplication. If 1 1 2 2 3 A= , B= , and C = 1 1 2 2 1 1 then AB = 4 4 4 4 = AC but B = C in spite of the fact that A = 0. 98 Chapter 3 Matrix Algebra There are various ways to express the individual rows and columns of a matrix product. For example, the ith row of AB is [AB]i∗ = (Ai∗ B∗1 | Ai∗ B∗2 | · · · | Ai∗ B∗n) = Ai∗ B = (ai1 ai2 · · · aip) (B1∗ B2∗ ... Bp∗) = ai1 B1∗ + ai2 B2∗ + · · · + aip Bp∗ . As shown below, there are similar representations for the individual columns. Rows and Columns of a Product Suppose that A = [aij] is m × p and B = [bij] is p × n. • [AB]i∗ = Ai∗ B [(ith row of AB) = (ith row of A) ×B]. (3.5.4) • [AB]∗j = AB∗j [(j th col of AB) = A× (j th col of B)]. (3.5.5) • [AB]i∗ = ai1 B1∗ + ai2 B2∗ + · · · + aip Bp∗ = Σ aik Bk∗ . (3.5.6) • [AB]∗j = A∗1 b1j + A∗2 b2j + · · · + A∗p bpj = Σ A∗k bkj . (3.5.7) These last two equations show that rows of AB are combinations of rows of B, while columns of AB are combinations of columns of A. For example, if A = 1 3 −2 0 −4 5 and B = 3 −5 1 2 −7 2 2 −2 3 then the second row of AB is [AB]2∗ = A2∗ B = (3 0 −4) (3 −5 1 2 −7 2 2 −2 3) = (1 −7 −9), and the second column of AB is [AB]∗2 = AB∗2 = (1 −2 0 3 −4 5) (−5 −7 −2) = (9 −7). This example makes the point that it is wasteful to compute the entire product if only one row or column is called for. Although it's not necessary to compute the complete product, you may wish to verify that (1 −2 0 3 −5 1) (3 0 −4) (2 −7 2) = (9 −3 −5 5 −9 1) AB = 1 −9 2 . 3.5 Matrix Multiplication 99 Matrix multiplication provides a convenient representation for a linear system of equations. For example, the 3 × 4 system 2x1 + 3x2 + 4x3 + 8x4 = 7, 3x1 + 5x2 + 6x3 + 2x4 = 6, 4x1 + 2x2 + 4x3 + 9x4 = 4, can be written as Ax = b, where 2 A3×4 = 3 4 3 5 2 4 6 4 8 2 9 x1 x2 x4×1 = x3 x4 , and 7 b3×1 = 6 . 4 And this example generalizes to become the following statement. Linear Systems Every linear system of m equations in n unknowns a11 x1 + a12 x2 + · · · + a1n xn = b1 , a21 x1 + a22 x2 + · · · + a2n xn = b2 , ... am1 x1 + am2 x2 + · · · + amn xn = bm , can be written as a single matrix equation Ax = b in which A = (a11 a12 ... a1n), x = (x1 x2 ... xn) and b = (b1 b2 ... bm). Conversely, every matrix equation of the form Am×n xn×1 = bm×1 represents a system of m linear equations in n unknowns. The numerical solution of a linear system was presented earlier in the text without the aid of matrix multiplication because the operation of matrix multiplication is not an integral part of the arithmetical process used to extract a solution by means of Gaussian elimination. Viewing a linear system as a single matrix equation Ax = b is more of a notational convenience that can be used to uncover theoretical properties and to prove general theorems concerning linear systems. 100 Chapter 3 Matrix Algebra For example, a very concise proof of the fact (2.3.5) stating that a system of equations Am×n xn×1 = bm×1 is consistent if and only if b is a linear combination of the columns in A is obtained by noting that the system is consistent if and only if there exists a column s that satisfies (s1) b = As = (A∗1 A∗2 · · · A∗n) (s2 ... sn) = A∗1 s1 + A∗2 s2 + · · · + A∗n sn . This observation makes it possible to express the solutions of Ax = b in yet another way. [Remainder of page continues in the same extremely small print: Example 3.5.2 on flights between cities, adjacency matrices, Properties of Matrix Multiplication, Example 3.6.x, Matrix inversion, Sherman–Morrison formula, Neumann series, through "Inverses of Sums and Sensitivity" 127.]

and from Exercise 3.8.2 when a single column is perturbed. This discussion resolves, at least in part, an issue raised in §1.6—namely, "What mechanism determines the extent to which a nonsingular system Ax = b is ill-conditioned?" To see how, an aggregate measure of the magnitude of the entries in A is needed, and one common measure is A = max i |aij | = the maximum absolute row sum. (3.8.7) This is one example of a matrix norm, a detailed discussion of which is given in §5.1. Theoretical properties specific to (3.8.7) are developed on pp. 280 and 283, and one property established there is the fact that XY ≤ X Y for all conformable matrices X and Y. But let's keep things on an intuitive level for the time being and defer the details. Using the norm (3.8.7), the approximation (3.8.6) insures that if B is sufficiently small, then $ -1 $ ≈ $A−1 BA−1 ≤ $A−1 B $A−1 $, so we interpret x − far from y, we can write $ ≈ $ since x − $A−1 B A $A−1 $ B $A−1 B A−1 A The term on the left is the relative change in B is small, then the ratio A / B is the relative change in A. The number κ = $A−1 $ A is called a condition number for A. In other words, if κ is small relative to 1 (i.e., if A is well conditioned), then a small relative change (or error) in A cannot produce a large relative change (or error) in the inverse, but if κ is large (i.e., if A is ill conditioned), then a small relative change (or error) in A can possibly (but not necessarily) result in a large relative change (or error) in the inverse. The situation for linear systems is similar. If the coefficients in a nonsingular system Ax = b are slightly perturbed to produce the system (A + B)x̃ = b, then x = A−1 b and x̃ = (A + B)−1 b so that (3.8.6) implies that x − x̃ = A−1 b − (A + B)−1 b ≈ A−1 BA−1 b = A−1 Bx. For column vectors, (3.8.7) reduces to x = max |xi |, and we have x − x̃ ≲ A−1 B x, so x̃ x 3.8 Chapter 3 Matrix Algebra so the relative change in the solution is κ relative to the relative change in A. Hence, if κ is small, a small relative change in A cannot produce a large relative change in x, but for larger values of κ, a small relative change in A can possibly result in a large relative change in x. Below is a summary of these observations. Sensitivity and Conditioning • A nonsingular matrix A is said to be ill conditioned if a small relative change in A can cause a large relative change in A−1. The degree of ill-conditioning is gauged by a condition number κ = A A−1, where ‖ is a matrix norm. • The sensitivity of the solution of Ax = b to perturbations (or errors) in A is measured by the extent to which A is an ill-conditioned matrix. More is said in Example 5.12.1 on p. 414. Example 3.8.2 It was demonstrated in Example 1.6.1 that the system .835x + .667y = .168, .333x + .266y = .067, is sensitive to small perturbations. We can understand this in the current context by examining the condition number of the coefficient matrix. If the matrix norm (3.8.7) is employed with A = .835 .333 .667 .266 and A−1 = −266000 333000 667000 −835000 , then the condition number for A is κ = A A−1 = 1.502(1668000) = 2,505,336 ≈ 2.5 × 10^6. Since the right-hand side of (3.8.8) is only an estimate of the relative error in the solution, the exact value of κ is not as important as its order of magnitude. Because κ is of order 10^6, (3.8.8) holds the possibility that the relative change (or error) in the solution can be about a million times larger than the relative change (or error) in A. Therefore, we must consider A and the associated linear system to be ill conditioned. A Rule of Thumb. If Gaussian elimination with partial pivoting is used to solve a well-scaled nonsingular system Ax = b using t-digit floating-point arithmetic, then, assuming no other source of error exists, it can be argued that when κ is of order 10^p, the computed solution is expected to be accurate to at least t − p significant digits, more or less. In other words, one expects to lose roughly p significant figures. For example, if Gaussian elimination with 8-digit arithmetic is used to solve the 2 × 2 system given above, then only about t − p = 8 − 6 = 2 significant figures of accuracy should be expected. This doesn't preclude the possibility of getting lucky and attaining a higher degree of accuracy—it just says that you shouldn't bet the farm on it. The complete story of conditioning has not yet been told. As pointed out earlier, it's about three times more costly to compute A−1 than to solve Ax = b, so it doesn't make sense to compute A−1 just to estimate the condition of A. Questions concerning condition estimation without explicitly computing an inverse still need to be addressed. Furthermore, liberties allowed by using the ≈ and ≲ symbols produce results that are intuitively correct but not rigorous. Rigor will eventually be attained—see Example 5.12.1 on p. 414.

Exercises for section 3.8 3.8.1. Suppose you are given that ... 2 A = ... and A−1 = ... 2. (a) Use the Sherman–Morrison formula to determine the inverse of the matrix B that is obtained by changing the (3, 2)-entry in A from 0 to 2. (b) Let C be the matrix that agrees with A except that c32 = 2 and c33 = 2. Use the Sherman–Morrison formula to find C−1. 3.8.2. Suppose A and B are nonsingular matrices in which B is obtained from A by replacing A∗j with another column b. Use the Sherman–Morrison formula to show that B−1 = A−1 − (A−1 b − ej)[A−1]j∗ / [A−1]j∗ b. 3.8.3. Suppose the coefficient matrix of a nonsingular system Ax = b is updated to produce another nonsingular system (A + cdT)z = b, where b, c, d ∈ ℜn×1, and let y be the solution of Ay = c. Show that z = x − y dT x / (1 + dT y). 3.8.4. (a) Use the Sherman–Morrison formula to prove that if A is nonsingular, then A + αei ejT is nonsingular for a sufficiently small α. (b) Use part (a) to prove that I + E is nonsingular when all εij's are sufficiently small in magnitude. This is an alternative to using the Neumann series argument. 3.8.5. For given matrices A and B, where A is nonsingular, explain why A + εB is also nonsingular when the real number ε is constrained to a sufficiently small interval about the origin. In other words, prove that small perturbations of nonsingular matrices are also nonsingular. 3.8.6. Derive the Sherman–Morrison–Woodbury formula. 3.8.7. Using the norm (3.8.7), rank the following matrices according to their degree of ill-conditioning. 3.8.8. Suppose that the entries in A(t), x(t), and b(t) are differentiable functions of a real variable t such that A(t)x(t) = b(t). (a) Assuming that A(t)−1 exists, explain why dA(t)−1/dt = −A(t)−1 A'(t)A(t)−1. (b) Derive the equation x'(t) = A(t)−1 b'(t) − A(t)−1 A'(t)x(t). This shows that A−1 magnifies both the change in A and the change in b, and thus it confirms the observation derived from (3.8.8) saying that the sensitivity of a nonsingular system to small perturbations is directly related to the magnitude of the entries in A−1.

3.9 ELEMENTARY MATRICES AND EQUIVALENCE

A common theme in mathematics is to break complicated objects into more elementary components, such as factoring large polynomials into products of smaller polynomials. The purpose of this section is to lay the groundwork for similar ideas in matrix algebra by considering how a general matrix might be factored into a product of more "elementary" matrices. Elementary Matrices Matrices of the form I − uvT, where u and v are n × 1 columns such that vT u ≠ 1 are called elementary matrices, and we know from (3.8.1) that all such matrices are nonsingular and (I − uvT)−1 = I − uvT / (vT u − 1). (3.9.1) Notice that inverses of elementary matrices are elementary matrices. We are primarily interested in the elementary matrices associated with the three elementary row (or column) operations hereafter referred to as follows. • Type I is interchanging rows (columns) i and j. • Type II is multiplying row (column) i by α ≠ 0. • Type III is adding a multiple of row i (column i) to row j (column j). An elementary matrix of Type I, II, or III is created by performing an elementary operation of Type I, II, or III to an identity matrix. For example, the matrices E1, E2, and E3 are elementary matrices of Types I, II, and III, respectively, because E1 arises by interchanging rows 1 and 2 in I3, whereas E2 is generated by multiplying row 2 in I3 by α, and E3 is constructed by multiplying row 1 in I3 by α and adding the result to row 3. The matrices in (3.9.2) also can be generated by column operations. For example, E3 can be obtained by adding α times the third column of I3 to the first column. The fact that E1, E2, and E3 are of the form (3.9.1) follows by using the unit columns ei to write E1 = I − uuT, where u = e1 − e2, E2 = I − (1−α)e2 e2T, and E3 = I+αe3 e1T. 132 Chapter 3 Matrix Algebra These observations generalize to matrices of arbitrary size. One of our objectives is to remove the arrows from Gaussian elimination because the inability to do "arrow algebra" limits the theoretical analysis. For example, while it makes sense to add two equations together, there is no meaningful analog for arrows—reducing A → B and C → D by row operations does not guarantee that A + C → B + D is possible. The following properties are the mechanisms needed to remove the arrows from elimination processes. Properties of Elementary Matrices • When used as a left-hand multiplier, an elementary matrix of Type I, II, or III executes the corresponding row operation. • When used as a right-hand multiplier, an elementary matrix of Type I, II, or III executes the corresponding column operation. Proof. A proof for Type III operations is given—the other two cases are left to the reader. Using I + αej eiT as a left-hand multiplier on an arbitrary matrix A produces (I + αej eiT)A = A + αej Ai∗. This is exactly the matrix produced by a Type III row operation in which the ith row of A is multiplied by α and added to the jth row. When I + αej eiT is used as a right-hand multiplier on A, the result is A(I + αej eiT) = A + αA∗j eiT. This is the result of a Type III column operation in which the jth column of A is multiplied by α and then added to the ith column. Example 3.9.1 The sequence of row operations used to reduce A to EA can be described as a sequence of left-hand multiplications by elementary matrices as indicated below. The reduction can be accomplished by a sequence of left-hand multiplications with the corresponding elementary matrices as shown below. The product of these elementary matrices is P = E4E3E2E1, and you can verify that PA = EA. This is the answer to the question of why nonsingular matrices are precisely those matrices that can be factored as a product of elementary matrices.

Products of Elementary Matrices • A is a nonsingular matrix if and only if A is the product of elementary matrices of Type I, II, or III. (3.9.3) Proof. If A is nonsingular, then the Gauss–Jordan technique reduces A to I by row operations. If G1, G2, ..., Gk is the sequence of elementary matrices that corresponds to the elementary row operations employed, then Gk ··· G2 G1 A = I or, equivalently, A = G1−1 G2−1 ··· Gk−1. Since the inverse of an elementary matrix is again an elementary matrix of the same type, this proves that A is the product of elementary matrices of Type I, II, or III. Conversely, if A = E1 E2 ··· Ek is a product of elementary matrices, then A must be nonsingular because the Ei's are nonsingular, and a product of nonsingular matrices is also nonsingular. 134 Chapter 3 Matrix Algebra Equivalence • Whenever B can be derived from A by a combination of elementary row and column operations, we write A ∼ B, and we say that A and B are equivalent matrices. Since elementary row and column operations are left-hand and right-hand multiplication by elementary matrices, respectively, and in view of (3.9.3), we can say that A ∼ B ⇐⇒ PAQ = B for nonsingular P and Q. Whenever B can be obtained from A by performing a sequence of elementary row operations only, we write A ∼row B, and we say that A and B are row equivalent. In other words, A ∼row B ⇐⇒ PA = B for a nonsingular P. Whenever B can be obtained from A by performing a sequence of column operations only, we write A ∼col B, and we say that A and B are column equivalent. In other words, A ∼col B ⇐⇒ AQ = B for a nonsingular Q. If it's possible to go from A to B by elementary row and column operations, then clearly it's possible to start with B and get back to A because elementary operations are reversible—i.e., PAQ = B ⟹ P−1 BQ−1 = A. It therefore makes sense to talk about the equivalence of a pair of matrices without regard to order. In other words, A ∼ B ⟺ B ∼ A. Furthermore, it's not difficult to see that each type of equivalence is transitive in the sense that A∼B and B ∼ C ⟹ A ∼ C. In §2.2 it was stated that each matrix A possesses a unique reduced row echelon form EA, and we accepted this fact because it is intuitively evident. However, we are now in a position to understand a rigorous proof. Example 3.9.2 Problem: Prove that EA is uniquely determined by A. Solution: Without loss of generality, we may assume that A is square—otherwise the appropriate number of zero rows or columns can be adjoined to A so as not to alter the nature of the argument. Suppose that A ∼row E1 and A ∼row E2, where E1 and E2 are both in reduced row echelon form. Consequently, E1 ∼row E2, and hence there is a nonsingular matrix P such that PE1 = E2. (3.9.4) Furthermore, by permuting the rows of E1 and E2 to force the pivotal 1's to occupy the diagonal positions, we see that E1 ∼row T1 and E2 ∼row T2, where T1 and T2 are upper-triangular matrices in which the basic columns in each Ti occupy the same positions as the basic columns in Ei. For example, if E = ..., then T = ... Each Ti has the property that Ti2 = Ti because there is a permutation matrix Qi (a product of elementary interchange matrices of Type I) such that QiTiQiT = (Iri Ji; 0 0) or, equivalently, Ti = QiT (Iri Ji; 0 0) Qi (see Exercise 3.9.4) implies Ti2 = Ti. It follows from (3.9.5) that T1 ∼row T2, so there is a nonsingular matrix R such that RT1 = T2. Thus T2 = RT1 = RT1T1 = T2T1 and T1 = R−1T2 = R−1T2T2 = T1T2. Because T1 and T2 are both upper triangular, T2T1 and T1T2 have the same diagonal entries, and hence T1 and T2 have the same diagonal. Therefore, the positions of the basic columns (i.e., the pivotal positions) in T1 agree with those in T2, and hence E1 and E2 have basic columns in exactly the same positions. This means there is a permutation matrix Q such that E1Q = (Ir J1; 0 0) and E2Q = (Ir J2; 0 0). Using (3.9.4) yields PE1Q = E2Q, or (P11 P12; P21 P22)(Ir J1; 0 0) = (Ir J2; 0 0), which in turn implies that P11 = Ir and P11J1 = J2. Consequently, J1 = J2, and it follows that E1 = E2. In passing, notice that the uniqueness of EA implies the uniqueness of the row pivot positions in any other row echelon form derived from A. If A ∼row U1 and A ∼row U2, where U1 and U2 are row echelon forms with different pivot positions, then Gauss–Jordan reduction applied to U1 and U2 would lead to two different reduced echelon forms, which is impossible. In §2.2 we observed the fact that the column relationships in a matrix A are exactly the same as the column relationships in EA. This observation is a special case of the more general result presented below. 136 Chapter 3 Matrix Algebra Column and Row Relationships • If A ∼row B, then linear relationships existing among columns of A also hold among corresponding columns of B. That is, B∗k = Σ αj B∗j if and only if A∗k = Σ αj A∗j. (3.9.6) • In particular, the column relationships in A and EA must be identical, so the nonbasic columns in A must be linear combinations of the basic columns in A as described in (2.2.3). • If A ∼col B, then linear relationships existing among rows of A must also hold among corresponding rows of B. Summary. Row equivalence preserves column relationships, and column equivalence preserves row relationships. Proof. If A ∼row B, then PA = B for some nonsingular P. Recall from (3.5.5) that the jth column in B is given by B∗j = (PA)∗j = PA∗j. Therefore, if A∗k = Σ αj A∗j, then multiplication by P on the left produces B∗k = Σ αj B∗j. Conversely, if B∗k = Σ αj B∗j, then multiplication on the left by P−1 produces A∗k = Σ αj A∗j. The statement concerning column equivalence follows by considering transposes. The reduced row echelon form EA is as far as we can go in reducing A by using only row operations. However, if we are allowed to use row operations in conjunction with column operations, then, as described below, the end result of a complete reduction is much simpler. Rank Normal Form If A is an m × n matrix such that rank(A) = r, then A ∼ Nr = (Ir 0; 0 0). (3.9.7) Nr is called the rank normal form for A, and it is the end product of a complete reduction of A by using both row and column operations. 3.9 Elementary Matrices and Equivalence 137 Proof. It is always true that A ∼row EA, so there is a nonsingular matrix P such that PA = EA. If rank (A) = r, then the basic columns in EA are the r unit columns. Apply column interchanges to EA so as to move these r unit columns to the far left-hand side. If Q1 is the product of the elementary matrices corresponding to these column interchanges, then PAQ1 has the form PAQ1 = EAQ1 = (Ir J; 0 0). Multiplying both sides of this equation on the right by the nonsingular matrix Q2 = (Ir −J; 0 I) produces PAQ1Q2 = (Ir 0; 0 0). Thus A ∼ Nr, because P and Q = Q1Q2 are nonsingular. Example 3.9.3 Problem: Explain why rank (A; 0 B) = rank(A) + rank(B). Solution: If rank(A) = r and rank (B) = s, then A ∼ Nr and B ∼ Ns. Consequently, (A 0; 0 B) ∼ (Nr 0; 0 Ns) ⟹ rank (A 0; 0 B) = rank (Nr 0; 0 Ns) = r + s. Given matrices A and B, how do we decide whether or not A ∼ B, A ∼row B, or A ∼col B? We could use a trial-and-error approach by attempting to reduce A to B by elementary operations, but this would be silly because there are easy tests, as described below. Testing for Equivalence For m × n matrices A and B the following statements are true. • A ∼ B if and only if rank (A) = rank (B). (3.9.8) • A ∼row B if and only if EA = EB. (3.9.9) • A ∼col B if and only if EAT = EBT. (3.9.10) Corollary. Multiplication by nonsingular matrices cannot change rank. 138 Chapter 3 Matrix Algebra Proof. To establish the validity of (3.9.8), observe that rank (A) = rank (B) implies A ∼ Nr and B ∼ Nr. Therefore, A ∼ Nr ∼ B. Conversely, if A ∼ B, where rank(A) = r and rank(B) = s, then A ∼ Nr and B ∼ Ns, and hence Nr ∼ A ∼ B ∼ Ns. Clearly, Nr ∼ Ns implies r = s. To prove (3.9.9), suppose first that A ∼row B. Because B ∼row EB, it follows that A ∼row EB. But a matrix has a uniquely determined reduced row echelon form by Example 3.9.2, so it must be the case that EB = EA. Conversely, if EA = EB, then row row A ∼ EA = EB ∼ B ⟹ A ∼row B. The proof of (3.9.10) follows from (3.9.9) by considering transposes because A ∼col B ⟺ AQ = B ⟺ (AQ)T = BT ⟺ QT AT = BT ⟺ AT ∼row BT. Example 3.9.4 Problem: Are the relationships that exist among the columns in A the same as the column relationships in B, and are the row relationships in A the same as the row relationships in B, where A = ..., B = ...? Solution: Straightforward computation reveals that EA = EB = ..., and hence A ∼row B. Therefore, the column relationships in A and B must be identical, and they must be the same as those in EA. Examining EA reveals that E∗3 = 2E∗1 + E∗2, so it must be the case that A∗3 = 2A∗1 + A∗2 and B∗3 = 2B∗1 + B∗2. The row relationships in A and B are different because EAT ≠ EBT. On the surface, it may not seem plausible that a matrix and its transpose should have the same rank. After all, if A is 3 × 100, then A can have as many as 100 basic columns, but AT can have at most three. Nevertheless, we can now demonstrate that rank (A) = rank (AT). Transposition and Rank • Transposition does not change the rank—i.e., for all m × n matrices, rank (A) = rank (AT) and rank (A) = rank (A∗). (3.9.11) Proof. Let rank(A) = r, and let P and Q be nonsingular matrices such that PAQ = Nr. Applying the reverse order law for transposition produces QT AT PT = NrT. Since QT and PT are nonsingular, it follows that AT ∼ NrT, and therefore rank (AT) = rank (NrT) = r = rank (A). Similarly, for the conjugate transpose, rank (A∗) = rank (A).

Exercises for section 3.9 3.9.1. Suppose that A is an m × n matrix. (a) If [A|Im] is row reduced to a matrix [B|P], explain why P is a nonsingular matrix such that PA = B. (b) If [A; In] is column reduced to [C; Q], explain why Q is a nonsingular matrix such that AQ = C. (c) Find a nonsingular matrix P such that PA = EA, where A = ... (d) Find nonsingular matrices P and Q such that PAQ is in rank normal form. 3.9.2. Consider the two matrices A = ... and B = ... (a) Are A and B equivalent? (b) Are A and B row equivalent? (c) Are A and B column equivalent? 3.9.3. If A ∼row B, explain why the basic columns in A occupy exactly the same positions as the basic columns in B. 3.9.4. A product of elementary interchange matrices—i.e., elementary matrices of Type I—is called a permutation matrix. If P is a permutation matrix, explain why PT = P−1. 3.9.5. If An×n is a nonsingular matrix, which (if any) of the following statements are true? (a) A ∼ A−1. (b) A ∼row A−1. (c) A ∼col A−1. (d) A ∼ I. (e) A ∼row I. (f) A ∼col I. 3.9.6. Which (if any) of the following statements are true? (a) A ∼ B ⟹ AT ∼ BT. (b) A ∼row B ⟹ AT ∼row BT. (c) A ∼row B ⟹ AT ∼col BT. (d) A ∼row B ⟹ A ∼ B. (e) A ∼col B ⟹ A ∼ B. (f) A ∼ B ⟹ A ∼row B. 3.9.7. Show that every elementary matrix of Type I can be written as a product of elementary matrices of Types II and III. Hint: Recall Exercise 2.1.3 on p. 14. 3.9.8. If rank (Am×n) = r, show that there exist matrices Bm×r and Cr×n such that A = BC, where rank (B) = rank (C) = r. Such a factorization is called a full-rank factorization. Hint: Consider the basic columns of A and the nonzero rows of EA. 3.9.9. Prove that rank(Am×n) = 1 if and only if there are nonzero columns um×1 and vn×1 such that A = uvT. 3.9.10. Prove that if rank (An×n) = 1, then A2 = τA, where τ = trace (A).

3.10 THE LU FACTORIZATION

We have now come full circle, and we are back to where the text began—solving a nonsingular system of linear equations using Gaussian elimination with back substitution. This time, however, the goal is to describe and understand the process in the context of matrices. If Ax = b is a nonsingular system, then the object of Gaussian elimination is to reduce A to an upper-triangular matrix using elementary row operations. If no zero pivots are encountered, then row interchanges are not necessary, and the reduction can be accomplished by using only elementary row operations of Type III. For example, consider reducing the matrix A = (2 2 2; 4 7 7; 6 18 22) to upper-triangular form as shown below. These are also the three elementary row operations that could be represented by left-hand multiplication with elementary matrices Gi, and the product of all of these Gi's is G3G2G1 = ... We learned in the previous section that each of these Type III operations can be executed by means of a left-hand multiplication with the corresponding elementary matrix Gi, and the product of all of these Gi's is ... In other words, G3 G2 G1 A = U, so that A = G1−1 G2−1 G3−1 U = LU, where L is the lower-triangular matrix L = G1−1 G2−1 G3−1 = (1 0 0; 2 1 0; 3 4 1). Thus A = LU is a product of a lower-triangular matrix L and an upper-triangular matrix U. Naturally, this is called an LU factorization of A. 142 Chapter 3 Matrix Algebra Observe that U is the end product of Gaussian elimination and has the pivots on its diagonal, while L has 1's on its diagonal. Moreover, L has the remarkable property that below its diagonal, each entry ℓij is precisely the multiplier used in the elimination (3.10.1) to annihilate the (i, j)-position. This is characteristic of what happens in general. To develop the general theory, it's convenient to introduce the concept of an elementary lower-triangular matrix, which is defined to be an n × n triangular matrix of the form Tk = I − ck ekT, where ck is a column with zeros in the first k positions. In particular, if ck = (0, ..., 0, μk+1, ..., μn)T, then Tk = ... By observing that ekT ck = 0, the formula (3.8.1) for the inverse of an elementary matrix produces Tk−1 = I + ck ekT, which is also an elementary lower-triangular matrix. The utility of elementary lower-triangular matrices lies in the fact that all of the Type III row operations needed to annihilate the entries below the k th pivot can be accomplished with one multiplication by Tk. If A(k−1) = ... is the partially triangularized result after k − 1 elimination steps, then Tk A(k−1) = (I − ck ekT)A(k−1) = A(k−1) − ck ekT A(k−1) = ..., where ck contains the multipliers used to annihilate those entries below akk. Notice that Tk does not alter the first k − 1 columns of A(k−1) because ekT [A(k−1)]∗j = 0 whenever j ≤ k − 1. Therefore, if no row interchanges are required, then reducing A to an upper-triangular matrix U by Gaussian elimination is equivalent to executing a sequence of n − 1 left-hand multiplications with elementary lower-triangular matrices. That is, Tn−1 ··· T2 T1 A = U, and hence A = T1−1 T2−1 ··· Tn−1−1 U. (3.10.4) Making use of the fact that ejT ck = 0 whenever j ≤ k and applying (3.10.3) reveals that T1−1 T2−1 ··· Tn−1−1 = (I + c1 e1T)(I + c2 e2T) ··· (I + cn−1 en−1T) = I + c1 e1T + c2 e2T + ··· + cn−1 en−1T. By observing that ... where the ℓij's are the multipliers used at the kth stage to annihilate the entries below the kth pivot, it now follows from (3.10.4) and (3.10.5) that A = LU, 144 Chapter 3 Matrix Algebra where L = I + c1e1T + ··· + cn−1en−1T = (1 0 ··· 0; ℓ21 1 ··· 0; ... ; ℓn1 ℓn2 ··· 1) is the lower-triangular matrix with 1's on the diagonal, and where ℓij is precisely the multiplier used to annihilate the (i, j)-position during Gaussian elimination. Thus the factorization A = LU can be viewed as the matrix formulation of Gaussian elimination, with the understanding that no row interchanges are used. LU Factorization If A is an n × n matrix such that a zero pivot is never encountered when applying Gaussian elimination with Type III operations, then A can be factored as the product A = LU, where the following hold. • L is lower triangular and U is upper triangular. (3.10.6) • ℓii = 1 and uii ≠ 0 for each i = 1, 2, ..., n. (3.10.7) • Below the diagonal of L, the entry ℓij is the multiple of row j that is subtracted from row i in order to annihilate the (i, j)-position during Gaussian elimination. • U is the final result of Gaussian elimination applied to A. • The matrices L and U are uniquely determined by properties (3.10.7) and (3.10.8). The decomposition of A into A = LU is called the LU factorization of A, and the matrices L and U are called the LU factors of A. Proof. Except for the statement concerning the uniqueness of the LU factors, each point has already been established. To prove uniqueness, observe that LU factors must be nonsingular because they have nonzero diagonals. If L1U1 = A = L2U2 are two LU factorizations for A, then L2−1L1 = U2U1−1. Notice that L2−1L1 is lower triangular, while U2U1−1 is upper triangular because the inverse of a matrix of the form (3.10.6) is again of the same form, and the product of two lower (upper) triangular matrices is also lower (upper) triangular. Consequently, (3.10.9) implies L2−1L1 = D = U2U1−1 must be a diagonal matrix. However, [L2]ii = 1 = [L2−1]ii, so it must be the case that L2−1L1 = I, and thus L1 = L2. Similarly, U1 = U2. Example 3.10.1 Once L and U are known, there is usually no need to manipulate with A. This together with the fact that the multipliers used in Gaussian elimination occur in just the right places in L means that A can be successively overwritten with the information in L and U as Gaussian elimination evolves. The rule is to store the multiplier ℓij in the position it annihilates—namely, the (i, j)-position of the array. For a 3 × 3 matrix, the result looks like this: ... For example, generating the LU factorization of A = (2 2 2; 4 7 7; 6 18 22) by successively overwriting a single 3 × 3 array would evolve as shown below. This is an important feature in practical computation because it guarantees that an LU factorization requires no more computer memory than that required to store the original matrix A. Once the LU factors for a nonsingular matrix An×n have been obtained, it's relatively easy to solve a linear system Ax = b. By rewriting Ax = b as L(Ux) = b and setting y = Ux, we see that Ax = b is equivalent to the two triangular systems Ly = b and Ux = y. First, the lower-triangular system Ly = b is solved for y by forward substitution. That is, if ... then y1 = b1, y2 = b2 − ℓ21y1, y3 = b3 − ℓ31y1 − ℓ32y2, etc. The forward substitution algorithm can be written more concisely as y1 = b1 and yi = bi − Σ ℓik yk for i = 2, 3, ..., n. After y is known, the upper-triangular system Ux = y is solved using the standard back substitution procedure by starting with xn = yn/unn, and setting xi = (yi − Σ uik xk)/uii for i = n − 1, n − 2, ..., 1. It can be verified that only n² multiplications/divisions and n² − n additions/subtractions are required when (3.10.10) and (3.10.11) are used to solve the two triangular systems Ly = b and Ux = y, so it's relatively cheap to solve Ax = b once L and U are known—recall from §1.2 that these operation counts are about n³/3 when we start from scratch. If only one system Ax = b is to be solved, then there is no significant difference between the technique of reducing the augmented matrix [A|b] to a row echelon form and the LU factorization method presented here. However, suppose it later becomes necessary to solve other systems Ax = b̃ with the same coefficient matrix but with different right-hand sides, which is frequently the case in applied work. If the LU factors of A were computed and saved when the original system was solved, then they need not be recomputed, and the solutions to all subsequent systems Ax = b̃ are therefore relatively inexpensive to obtain. That is, the operation counts for each subsequent system are on the order of n², whereas these counts would be on the order of n³/3 if we would start from scratch each time. Summary • To solve a nonsingular system Ax = b using the LU factorization A = LU, first solve Ly = b for y with the forward substitution algorithm (3.10.10), and then solve Ux = y for x with the back substitution procedure (3.10.11). • The advantage of this approach is that once the LU factors for A have been computed, any other linear system Ax = b̃ can be solved with only n² multiplications/divisions and n² − n additions/subtractions. 3.10 The LU Factorization 147 Example 3.10.2 Problem 1: Use the LU factorization of A to solve Ax = b, where A = (2 2 2; 4 7 7; 6 18 22) and b = (12; 24; 12). Problem 2: Suppose that after solving the original system new information is received that changes b to b̃ = (6; 24; 70). Use the LU factors of A to solve the updated system Ax = b̃. Solution 1: The LU factors of the coefficient matrix were determined in Example 3.10.1 to be L = (1 0 0; 2 1 0; 3 4 1) and U = (2 2 2; 0 3 3; 0 0 4). The strategy is to set Ux = y and solve Ax = L(Ux) = b by solving the two triangular systems Ly = b and Ux = y. First solve the lower-triangular system Ly = b by using forward substitution: y1 = 12, y2 = 24 − 2y1 = 0, y3 = 12 − 3y1 − 4y2 = −24. Now use back substitution to solve the upper-triangular system Ux = y: x3 = −24/4 = −6, x2 = (0 − 3x3)/3 = 6, x1 = (12 − 2x2 − 2x3)/2 = 6. Solution 2: To solve the updated system Ax = b̃, simply repeat the forward and backward substitution steps with b replaced by b̃. Solving Ly = b̃ with forward substitution gives ... and solving Ux = y with back substitution yields the updated solution ... Example 3.10.3 Computing A−1. Although matrix inversion is not used for solving Ax = b, there are a few applications where explicit knowledge of A−1 is desirable. Problem: Explain how to use the LU factors for a nonsingular matrix An×n to compute A−1 efficiently. Solution: The strategy is to solve the matrix equation AX = I. Recall from (3.5.5) that AA−1 = I ⟺ A[A−1]∗j = ej, so the jth column of A−1 is the solution of a system Axj = ej. Each of these n systems has the same coefficient matrix, so, once the LU factors for A are known, each system Axj = LUxj = ej can be solved by the standard two-step process. (1) Set yj = Uxj, and solve Lyj = ej for yj by forward substitution. (2) Solve Uxj = yj for xj = [A−1]∗j by back substitution. This method has at least two advantages: it's efficient, and any code written to solve Ax = b can also be used to compute A−1. Note: A tempting alternate solution might be to use the fact A−1 = (LU)−1 = U−1L−1. But computing U−1 and L−1 explicitly and then multiplying the results is not as computationally efficient as the method just described. Not all nonsingular matrices possess an LU factorization. For example, there is clearly no nonzero value of u11 that will satisfy (0 1; 1 1) = (1 0; ℓ21 1)(u11 u12; 0 u22). The problem here is the zero pivot in the (1,1)-position. Our development of the LU factorization using elementary lower-triangular matrices shows that if no zero pivots emerge, then no row interchanges are necessary, and the LU factorization can indeed be carried to completion. The converse is also true (its proof is left as an exercise), so we can say that a nonsingular matrix A has an LU factorization if and only if a zero pivot does not emerge during row reduction to upper-triangular form with Type III operations. Although it is a bit more theoretical, there is another interesting way to characterize the existence of LU factors. This characterization is given in terms of the leading principal submatrices of A that are defined to be those submatrices taken from the upper-left-hand corner of A. That is, A1 = a11, A2 = (a11 a12; a21 a22), ..., Ak = ..., ..., An = A. 3.10 The LU Factorization 149 Existence of LU Factors Each of the following statements is equivalent to saying that a nonsingular matrix An×n possesses an LU factorization. • A zero pivot does not emerge during row reduction to upper-triangular form with Type III operations. • Each leading principal submatrix Ak is nonsingular. (3.10.12) Proof. We will prove the statement concerning the leading principal submatrices and leave the proof concerning the nonzero pivots as an exercise. Assume first that A possesses an LU factorization and partition A as A = LU = (L11 0; L21 L22)(U11 U12; 0 U22) = (L11U11 ∗; ∗ ∗), where L11 and U11 are each k × k. Thus Ak = L11U11 must be nonsingular because L11 and U11 are each nonsingular—they are triangular with nonzero diagonal entries. Conversely, suppose that each leading principal submatrix in A is nonsingular. Use induction to prove that each Ak possesses an LU factorization. For k = 1, this statement is clearly true because if A1 = (a11) is nonsingular, then A1 = (1)(a11) is its LU factorization. Now assume that Ak has an LU factorization and show that this together with the nonsingularity condition implies Ak+1 must also possess an LU factorization. If Ak = LkUk is the LU factorization for Ak, then Ak−1 = Uk−1Lk−1 so that Ak+1 = (Ak b; cT αk+1) = (Lk 0; cT Uk−1 1)(Uk Lk−1b; 0 αk+1 − cT Ak−1 b), where cT and b contain the first k components of Ak+1∗ and A∗k+1, respectively. Observe that this is the LU factorization for Ak+1 because Lk+1 = (Lk 0; cT Uk−1 1) and Uk+1 = (Uk Lk−1b; 0 αk+1 − cT Ak−1 b) are lower- and upper-triangular matrices, respectively, and L has 1's on its diagonal while the diagonal entries of U are nonzero. The fact that αk+1 − cT Ak−1 b ≠ 0 follows because Ak+1 and Lk+1 are each nonsingular, so Uk+1 = Lk+1−1 Ak+1 must also be nonsingular. Therefore, the nonsingularity of the leading principal submatrices implies that each Ak possesses an LU factorization, and hence An = A must have an LU factorization. Up to this point we have avoided dealing with row interchanges because if a row interchange is needed to remove a zero pivot, then no LU factorization is possible. However, we know from the discussion in §1.5 that practical computation necessitates row interchanges in the form of partial pivoting. So even if no zero pivots emerge, it is usually the case that we must still somehow account for row interchanges. To understand the effects of row interchanges in the framework of an LU decomposition, let Tk = I − ck ekT be an elementary lower-triangular matrix as described in (3.10.3), and let E = I − uuT with u = ek+i − ek+j be the Type I elementary interchange matrix associated with an interchange of rows k + i and k + j. Notice that ekT E = ekT because ekT has 0's in positions k + i and k + j. This together with the fact that E2 = I guarantees ETkE = E2 − Eck ekT E = I − c̃k ekT, where c̃k = Eck. In other words, the matrix T̃k = ETkE = I − c̃k ekT (3.10.14) agrees with Tk in all entries except that the multipliers μk+i and μk+j have traded places. As before, assume we are row reducing an n × n nonsingular matrix A, but suppose that an interchange of rows k + i and k + j is necessary immediately after the k th stage so that the sequence of left-hand multiplications ETkTk−1 ··· T1 is applied to A. Since E2 = I, we may insert E2 to the right of each T to obtain ETkTk−1 ··· T1 = ETkE2Tk−1E2 ··· E2T1E2 = (ETkE)(ETk−1E) ··· (ET1E)E = T̃kT̃k−1 ··· T̃1E. In such a manner, the necessary interchange matrices E can be "factored" to the far-right-hand side, and the matrices T̃j retain the desirable feature of being elementary lower-triangular matrices. Furthermore, (3.10.14) implies that T̃kT̃k−1 ··· T̃1 differs from TkTk−1 ··· T1 only in the sense that the multipliers in rows k + i and k + j have traded places. Therefore, row interchanges in Gaussian elimination can be accounted for by writing T̃n−1 ··· T̃2T̃1PA = U, where P is the product of all elementary interchange matrices used during the reduction and where the T̃k's are elementary lower-triangular matrices in which the multipliers have been permuted according to the row interchanges that were implemented. Since all of the T̃k's are lower triangular, (3.10.4) through (3.10.6) imply that PA = LU, where L = T̃1−1T̃2−1 ··· T̃n−1−1 = I + c̃1e1T + ··· + c̃n−1en−1T. 3.10 The LU Factorization 151 When row interchanges are allowed, zero pivots can always be avoided when the original matrix A is nonsingular. Consequently, we may conclude that for every nonsingular matrix A, there exists a permutation matrix P (a product of elementary interchange matrices) such that PA has an LU factorization. Furthermore, because of the observation in (3.10.14) concerning how the multipliers in T̃k trade places when a row interchange occurs, and because T̃k−1 = I + c̃k ekT, it is not difficult to see that the same line of reasoning used to arrive at (3.10.6) can be applied to conclude that the multipliers in the matrix L in (3.10.15) are permuted according to the row interchanges that are executed. More specifically, if rows k and k+i are interchanged to create the kth pivot, then the multipliers (ℓk1, ℓk2, ..., ℓk,k−1) and (ℓk+i,1, ℓk+i,2, ..., ℓk+i,k−1) trade places in the formation of L. This means that we can proceed just as in the case when no interchanges are used and successively overwrite the array originally containing A with each multiplier replacing the position it annihilates. Whenever a row interchange occurs, the corresponding multipliers will be correctly interchanged as well. The permutation matrix P is simply the cumulative record of the various row interchanges used, and the information in P is easily accounted for by a simple technique that is illustrated in the following example. Example 3.10.4 Problem: Use partial pivoting on the matrix A = (1 2 −3 4; 4 8 12 −8; 2 3 2 1; −3 −1 1 −4) and determine the LU decomposition PA = LU, where P is the associated permutation matrix. Solution: As explained earlier, the strategy is to successively overwrite the array A with components from L and U. For the sake of clarity, the multipliers ℓij are shown in boldface type. Adjoin a "permutation counter column" p that is initially set to the natural order 1,2,3,4. Permuting components of p as the various row interchanges are executed will accumulate the desired permutation. The matrix P is obtained by executing the final permutation residing in p to the rows of an appropriate size identity matrix. ... Therefore, L = (1 0 0 0; −3/4 1 0 0; 1/4 0 1 0; 1/2 −1/5 1/3 1), U = (4 8 12 −8; 0 5 10 −10; 0 0 −6 6; 0 0 0 1), P = (0 1 0 0; 0 0 0 1; 1 0 0 0; 0 0 1 0). It is easy to combine the advantages of partial pivoting with the LU decomposition in order to solve a nonsingular system Ax = b. Because permutation matrices are nonsingular, the system Ax = b is equivalent to PAx = Pb, and hence we can employ the LU solution techniques discussed earlier to solve this permuted system. That is, if we have already performed the factorization PA = LU—as illustrated in Example 3.10.4—then we can solve Ax = b with the following steps. 1. Set Ly = Pb and solve for y by forward substitution. 2. Set Ux = y and solve for x by back substitution. It should be evident that the permutation matrix P is not really needed. All that is necessary is knowledge of the LU factors along with the final permutation contained in the permutation counter column p illustrated in Example 3.10.4. The column b̃ = Pb is simply a rearrangement of the components of b according to the final permutation shown in p. In other words, the strategy is to first permute b into b̃ according to the permutation p, and then solve Ly = b̃ followed by Ux = y. Example 3.10.5 Problem: Use the LU decomposition obtained with partial pivoting to solve the system Ax = b, where A = (1 2 −3 4; 4 8 12 −8; 2 3 2 1; −3 −1 1 −4) and b = (3; 60; 1; 5). Solution: The LU decomposition with partial pivoting was computed in Example 3.10.4. Permute the components in b according to the permutation p = (2 4 1 3), and call the result b̃. Now solve Ly = b̃ by applying forward substitution ... Then solve Ux = y by applying back substitution ... 3.10 The LU Factorization 153 LU Factorization with Row Interchanges • For each nonsingular matrix A, there exists a permutation matrix P such that PA possesses an LU factorization PA = LU. • To compute L, U, and P, successively overwrite the array originally containing A. Replace each entry being annihilated with the multiplier used to annihilate it. Each time a row interchange is required, interchange the entire rows in the array so that the multipliers and the U-entries are automatically interchanged in the correct manner. • Although the entire permutation matrix P is rarely called for, it can be constructed by permuting the rows of the identity matrix I according to the various interchanges used. These interchanges can be accumulated in a "permutation counter column" p that is initially in natural order (1, 2, ..., n)—see Example 3.10.4. • To solve a nonsingular linear system Ax = b using the LU decomposition with partial pivoting, permute the components in b to construct b̃ according to the sequence of interchanges used—i.e., according to p—and then solve Ly = b̃ by forward substitution followed by the solution of Ux = y using back substitution. Example 3.10.6 The LDU factorization. There's some asymmetry in an LU factorization because the lower factor has 1's on its diagonal while the upper factor has a nonunit diagonal. This is easily remedied by factoring the diagonal entries out of the upper factor as shown below. Setting D = diag (u11, u22, ..., unn) (the diagonal matrix of pivots) and redefining U to be the rightmost upper-triangular matrix shown above allows any LU factorization to be written as A = LDU, where L and U are lower- and upper-triangular matrices with 1's on both of their diagonals. This is called the LDU factorization of A. It is uniquely determined, and when A is symmetric, the LDU factorization is A = LDLT (Exercise 3.10.9). Example 3.10.7 The Cholesky Factorization. A symmetric matrix A possessing an LU factorization in which each pivot is positive is said to be positive definite. Problem: Prove that A is positive definite if and only if A can be uniquely factored as A = RTR, where R is an upper-triangular matrix with positive diagonal entries. This is known as the Cholesky factorization of A, and R is called the Cholesky factor of A. Solution: If A is positive definite, then, as pointed out in Example 3.10.6, it has an LDU factorization A = LDLT in which D = diag (p1, p2, ..., pn) is the diagonal matrix containing the pivots pi > 0. Setting R = D1/2 LT where D1/2 = diag(√p1, √p2, ..., √pn) yields the desired factorization because A = LD1/2D1/2LT = RTR, and R is upper triangular with positive diagonal entries. Conversely, if A = RRT, where R is lower triangular with a positive diagonal, then factoring the diagonal entries out of R as illustrated in Example 3.10.6 produces R = LD, where L is lower triangular with a unit diagonal and D is the diagonal matrix whose diagonal entries are the rii's. Consequently, A = LD2LT is the LDU factorization for A, and thus the pivots must be positive because they are the diagonal entries in D2. We have now proven that A is positive definite if and only if it has a Cholesky factorization. To see why such a factorization is unique, suppose A = R1TR1 = R2TR2, and factor out the diagonal entries as illustrated in Example 3.10.6 to write R1 = D1L1T and R2 = D2L2T, where each Ri is lower triangular with a unit diagonal and Di contains the diagonal of Ri so that A = L1D12L1T = L2D22L2T. The uniqueness of the LDU factors insures that L1 = L2 and D1 = D2, so R1 = R2. Note: More is said about the Cholesky factorization and positive definite matrices on pp. 313, 345, and 559. André-Louis Cholesky (1875–1918) was a French military officer. Although originally assigned to an artillery branch, Cholesky later became attached to the Geodesic Section of the Geographic Service in France where he became noticed for his extraordinary intelligence and his facility for mathematics. From 1905 to 1909 Cholesky was involved with the problem of adjusting the triangularization grid for France. This was a huge computational task, and there were arguments as to what computational techniques should be employed. It was during this period that Cholesky invented the ingenious procedure for solving a positive definite system of equations that is the basis for the matrix factorization that now bears his name. Unfortunately, Cholesky's mathematical talents were never allowed to flower. In 1914 war broke out, and Cholesky was again placed in the artillery group—but this time as the commander. On August 31, 1918, Major Cholesky was killed in battle. Cholesky never had time to publish his clever computational methods—they were carried forward by word-of-mouth. Issues surrounding the Cholesky factorization have been independently rediscovered several times by people who were unaware of Cholesky, and, in some circles, the Cholesky factorization is known as the square root method. 3.10 The LU Factorization 155

Exercises for section 3.10 3.10.1. Let A = (1 4 5; 4 18 26; 3 16 30). (a) Determine the LU factors of A. (b) Use the LU factors to solve Ax1 = b1 as well as Ax2 = b2, where b1 = (6; 0; −6) and b2 = (6; 6; 12). (c) Use the LU factors to determine A−1. 3.10.2. Let A and b be the matrices A = (1 2 4 17; 3 6 −12 3; 2 3 −3 2; 0 2 −2 6) and b = (17; 3; 3; 4). (a) Explain why A does not have an LU factorization. (b) Use partial pivoting and find the permutation matrix P as well as the LU factors such that PA = LU. (c) Use the information in P, L, and U to solve Ax = b. 3.10.3. Determine all values of ξ for which A = (ξ 2 0; 1 ξ 1; 0 1 ξ) fails to have an LU factorization. 3.10.4. If A is a nonsingular matrix that possesses an LU factorization, prove that the pivot that emerges after (k + 1) stages of standard Gaussian elimination using only Type III operations is given by pk+1 = ak+1,k+1 − cT Ak−1 b, where Ak and Ak+1 = (Ak b; cT ak+1,k+1) are the leading principal submatrices of orders k and k + 1, respectively. Use this to deduce that all pivots must be nonzero when an LU factorization for A exists. 3.10.5. If A is a matrix that contains only integer entries and all of its pivots are 1, explain why A−1 must also be an integer matrix. Note: This fact can be used to construct random integer matrices that possess integer inverses by randomly generating integer matrices L and U with unit diagonals and then constructing the product A = LU. 3.10.6. Consider the tridiagonal matrix T = (β1 γ1 0 0; α1 β2 γ2 0; 0 α2 β3 γ3; 0 0 α3 β4). (a) Assuming that T possesses an LU factorization, verify that it is given by L = ..., U = ..., where the πi's are generated by the recursion formula π1 = β1 and πi+1 = βi+1 − αiγi/πi. Note: This holds for tridiagonal matrices of arbitrary size thereby making this particular LU factorization particularly easy to compute. (b) Apply the recursion formula given above to obtain the LU factorization of T = ... 3.10.7. An n × n matrix A such that aij = 0 whenever |i − j| > w is called a band matrix with bandwidth w. (a) Prove that if A is a band matrix with bandwidth w and if A has an LU factorization A = LU, then L inherits the lower band structure of A, and U inherits the upper band structure of A. (b) Illustrate this by using a generic 5 × 5 matrix with bandwidth w = 2. 3.10.8. (a) Construct an example of a nonsingular symmetric matrix that fails to possess an LU (or LDU) factorization. (b) Construct an example of a nonsingular symmetric matrix that has an LU factorization but is not positive definite. 3.10.9. (a) Determine the LDU factors for A = (1 4 5; 4 18 26; 3 16 30) (this is the same matrix used in Exercise 3.10.1). (b) Prove that if a matrix has an LDU factorization, then the LDU factors are uniquely determined. (c) If A is symmetric and possesses an LDU factorization, explain why it must be given by A = LDLT. 3.10.10. Explain why A = (1 2 3; 2 8 12; 3 12 27) is positive definite, and then find the Cholesky factor R.

CHAPTER 4 VECTOR SPACES

4.1 SPACES AND SUBSPACES

After matrix theory became established toward the end of the nineteenth century, it was realized that many mathematical entities that were considered to be quite different from matrices were in fact quite similar. For example, objects such as points in the plane ℜ2, points in 3-space ℜ3, polynomials, continuous functions, and differentiable functions (to name only a few) were recognized to satisfy the same additive properties and scalar multiplication properties given in §3.2 for matrices. Rather than studying each topic separately, it was reasoned that it is more efficient and productive to study many topics at one time by studying the common properties that they satisfy. This eventually led to the axiomatic definition of a vector space. A vector space involves four things—two sets V and F, and two algebraic operations called vector addition and scalar multiplication. • V is a nonempty set of objects called vectors. Although V can be quite general, we will usually consider V to be a set of n-tuples or a set of matrices. • F is a scalar field—for us F is either the field ℜ of real numbers or the field C of complex numbers. • Vector addition (denoted by x + y) is an operation between elements of V. • Scalar multiplication (denoted by αx) is an operation between elements of F and V. The formal definition of a vector space stipulates how these four things relate to each other. In essence, the requirements are that vector addition and scalar multiplication must obey exactly the same properties given in §3.2 for matrices. 160 Chapter 4 Vector Spaces Vector Space Definition The set V is called a vector space over F when the vector addition and scalar multiplication operations satisfy the following properties. (A1) x+y ∈ V for all x, y ∈ V. This is called the closure property for vector addition. (A2) (x + y) + z = x + (y + z) for every x, y, z ∈ V. (A3) x + y = y + x for every x, y ∈ V. (A4) There is an element 0 ∈ V such that x + 0 = x for every x ∈ V. (A5) For each x ∈ V, there is an element (−x) ∈ V such that x + (−x) = 0. (M1) αx ∈ V for all α ∈ F and x ∈ V. This is the closure property for scalar multiplication. (M2) (αβ)x = α(βx) for all α, β ∈ F and every x ∈ V. (M3) α(x + y) = αx + αy for every α ∈ F and all x, y ∈ V. (M4) (α + β)x = αx + βx for all α, β ∈ F and every x ∈ V. (M5) 1x = x for every x ∈ V. A theoretical algebraic treatment of the subject would concentrate on the logical consequences of these defining properties, but the objectives in this text are different, so we will not dwell on the axiomatic development. Nevertheless, it is important to recognize some of the more significant examples and to understand why they are indeed vector spaces. The idea of defining a vector space by using a set of abstract axioms was contained in a general theory published in 1844 by Hermann Grassmann (1808–1887), a theologian and philosopher from Stettin, Poland, who was a self-taught mathematician. But Grassmann's work was originally ignored because he tried to construct a highly abstract self-contained theory, independent of the rest of mathematics, containing nonstandard terminology and notation, and he had a tendency to mix mathematics with obscure philosophy. Grassmann published a complete revision of his work in 1862 but with no more success. Only later was it realized that he had formulated the concepts we now refer to as linear dependence, bases, and dimension. The Italian mathematician Giuseppe Peano (1858–1932) was one of the few people who noticed Grassmann's work, and in 1888 Peano published a condensed interpretation of it. In a small chapter at the end, Peano gave an axiomatic definition of a vector space similar to the one above, but this drew little attention outside of a small group in Italy. The current definition is derived from the 1918 work of the German mathematician Hermann Weyl (1885–1955). Even though Weyl's definition is closer to Peano's than to Grassmann's, Weyl did not mention his Italian predecessor, but he did acknowledge Grassmann's "epoch making work." Weyl's success with the idea was due in part to the fact that he thought of vector spaces in terms of geometry, whereas Grassmann and Peano treated them as abstract algebraic structures. As we will see, it's the geometry that's important. 4.1 Spaces and Subspaces 161 Example 4.1.1 Because (A1)–(A5) are generalized versions of the five additive properties of matrix addition, and (M1)–(M5) are generalizations of the five scalar multiplication properties given in §3.2, we can say that the following hold. • The set ℜm×n of m × n real matrices is a vector space over ℜ. • The set Cm×n of m × n complex matrices is a vector space over C. Example 4.1.2 The real coordinate spaces ℜ1×n and ℜn×1 are special cases of the preceding example, and these will be the primary objects of our attention. In the context of vector spaces, it usually makes no difference whether a coordinate vector is depicted as a row or as a column. When the row or column distinction is irrelevant, or when it is clear from the context, we will use the common symbol ℜn to designate a coordinate space. In those cases where it is important to distinguish between rows and columns, we will explicitly write ℜ1×n or ℜn×1. Similar remarks hold for complex coordinate spaces. Although the coordinate spaces will be our primary concern, be aware that there are many other types of mathematical structures that are vector spaces—this was the underlying motivation for the abstraction at the outset. Listed below are a few examples. Example 4.1.3 With function addition and scalar multiplication defined by (f + g)(x) = f(x) + g(x) and (αf)(x) = αf(x), the following sets are vector spaces over ℜ: • The set of functions mapping the interval [0, 1] into ℜ. • The set of all real-valued continuous functions defined on [0, 1]. • The set of real-valued functions that are differentiable on [0, 1]. • The set of all polynomials with real coefficients. 162 Chapter 4 Vector Spaces Example 4.1.4 Consider the vector space ℜ2, and let L = {(x, y) | y = αx} be a line through the origin. L is a subset of ℜ2, but L is a special kind of subset because L also satisfies the properties (A1)–(A5) and (M1)–(M5) that define a vector space. This shows that it is possible for one vector space to properly contain other vector spaces. Subspaces Let S be a nonempty subset of a vector space V over F (symbolically, S ⊆ V). If S is also a vector space over F using the same addition and scalar multiplication operations, then S is said to be a subspace of V. It's not necessary to check all 10 of the defining conditions in order to determine if a subset is also a subspace—only the closure conditions (A1) and (M1) need to be considered. That is, a nonempty subset S of a vector space V is a subspace of V if and only if (A1) x, y ∈ S ⟹ x + y ∈ S and (M1) x ∈ S ⟹ αx ∈ S for all α ∈ F. Proof. If S is a subset of V, then S automatically inherits all of the vector space properties of V except (A1), (A4), (A5), and (M1). However, (A1) together with (M1) implies (A4) and (A5). To prove this, observe that (M1) implies (−x) = (−1)x ∈ S for all x ∈ S so that (A5) holds. Since x and (−x) are now both in S, (A1) insures that x + (−x) ∈ S, and thus 0 ∈ S. Example 4.1.5 Given a vector space V, the set Z = {0} containing only the zero vector is a subspace of V because (A1) and (M1) are trivially satisfied. Naturally, this subspace is called the trivial subspace. Vector addition in ℜ2 and ℜ3 is easily visualized by using the parallelogram law, which states that for two vectors u and v, the sum u + v is the vector defined by the diagonal of the parallelogram as shown in Figure 4.1.1. 4.1 Spaces and Subspaces 163 We have already observed that straight lines through the origin in ℜ2 are subspaces, but what about straight lines not through the origin? No—they cannot be subspaces because subspaces must contain the zero vector (i.e., they must pass through the origin). What about curved lines through the origin—can some of them be subspaces of ℜ2? Again the answer is "No!" As depicted in Figure 4.1.2, the parallelogram law indicates why the closure property (A1) cannot be satisfied for lines with a curvature because there are points u and v on the curve for which u + v (the diagonal of the corresponding parallelogram) is not on the curve. Consequently, the only proper subspaces of ℜ2 are the trivial subspace and lines through the origin. Similarly, the only proper subspaces of ℜ3 are the trivial subspace, lines through the origin, and planes through the origin. In ℜ3, the trivial subspace and lines through the origin are again the only possibilities, but there is also another one—planes through the origin. If P is a plane through the origin in ℜ3, then, as shown in Figure 4.1.3, the parallelogram law guarantees that the closure property for addition (A1) holds—the parallelogram defined by any two vectors in P is also in P so that if u, v ∈ P, then u + v ∈ P. The closure property for scalar multiplication (M1) holds because multiplying any vector by a scalar merely stretches it, but its angular orientation does not change so that if u ∈ P, then αu ∈ P for all scalars α. Lines and surfaces in ℜ3 that have curvature cannot be subspaces for essentially the same reason depicted in Figure 4.1.2. So the only proper subspaces of ℜ3 are the trivial subspace, lines through the origin, and planes through the origin. The concept of a subspace now has an obvious interpretation in the visual spaces ℜ2 and ℜ3—subspaces are the flat surfaces passing through the origin. Flatness Although we can't use our eyes to see "flatness" in higher dimensions, our minds can conceive it through the notion of a subspace. From now on, think of flat surfaces passing through the origin whenever you encounter the term "subspace." For a set of vectors S = {v1, v2, ..., vr} from a vector space V, the set of all possible linear combinations of the vi's is denoted by span(S) = {α1v1 + α2v2 + ··· + αrvr | αi ∈ F}. Notice that span (S) is a subspace of V because the two closure properties (A1) and (M1) are satisfied. That is, if x = Σ ξi vi and y = Σ ηi vi are two linear combinations from span (S), then the sum x + y = Σ(ξi + ηi)vi is also a linear combination in span (S), and for any scalar β, βx = Σ(βξi)vi is also a linear combination in span (S). 164 Chapter 4 Vector Spaces For example, if u ≠ 0 is a vector in ℜ3, then span {u} is the straight line passing through the origin and u. If S = {u, v}, where u and v are two nonzero vectors in ℜ3 not lying on the same line, then, as shown in Figure 4.1.4, span (S) is the plane passing through the origin and the points u and v. As we will soon see, all subspaces of ℜn are of the type span (S), so it is worthwhile to introduce the following terminology. Spanning Sets • For a set of vectors S = {v1, v2, ..., vr}, the subspace span (S) = {α1v1 + α2v2 + ··· + αrvr} generated by forming all linear combinations of vectors from S is called the space spanned by S. • If V is a vector space such that V = span (S), we say S is a spanning set for V. In other words, S spans V whenever each v ∈ V is a linear combination of vectors from S. Example 4.1.6 (i) In Figure 4.1.4, S = {u, v} is a spanning set for the indicated plane. (ii) S = {(1; 1), (2; 2)} spans the line y = x in ℜ2. (iii) The unit vectors {e1, e2, ..., en} in ℜn are a spanning set for ℜn. (iv) The finite set S = {1, x, x2, ..., xn} spans the space of all polynomials such that deg p(x) ≤ n, and the infinite set S = {1, x, x2, ...} spans the space of all polynomials. Example 4.1.7 Problem: For a set of vectors S = {a1, a2, ..., an} from a subspace V ⊆ ℜm×1, let A be the matrix containing the ai's as its columns. Explain why S spans V if and only if for each b ∈ V there corresponds a column x such that Ax = b (i.e., if and only if Ax = b is a consistent system for every b ∈ V). Solution: By definition, S spans V if and only if for each b ∈ V there exist scalars αi such that b = α1a1 + α2a2 + ··· + αnan = A(α1; α2; ...; αn). Note: This simple observation often is quite helpful. For example, to test whether or not S = {(1 1 1), (1 −1 −1), (3 1 1)} spans ℜ3, place these rows as columns in a matrix A, and ask, "Is the system Ax = b consistent for all b ∈ ℜ3?" Recall from (2.3.4) that Ax = b is consistent if and only if rank[A|b] = rank (A). In this case, rank (A) = 2, but rank[A|b] = 3 for some b's (e.g., b1 = 0, b2 = 1, b3 = 0), so S doesn't span ℜ3. On the other hand, S′ = {(1 1 1), (1 −1 −1), (3 1 2)} is a spanning set for ℜ3 because A = ... is nonsingular, so Ax = b is consistent for all b (the solution is x = A−1b). As shown below, it's possible to "add" two subspaces to generate another. Sum of Subspaces If X and Y are subspaces of a vector space V, then the sum of X and Y is defined to be the set of all possible sums of vectors from X with vectors from Y. That is, X + Y = {x + y | x ∈ X and y ∈ Y}. • The sum X + Y is again a subspace of V. (4.1.1) • If SX, SY span X, Y, then SX ∪ SY spans X + Y. (4.1.2) 4.1 Spaces and Subspaces 167 Proof. To prove (4.1.1), demonstrate that the two closure properties (A1) and (M1) hold for S = X + Y. To show (A1) is valid, observe that if u, v ∈ S, then u = x1 + y1 and v = x2 + y2, where x1, x2 ∈ X and y1, y2 ∈ Y. Because X and Y are closed with respect to addition, it follows that x1 + x2 ∈ X and y1 + y2 ∈ Y, and therefore u + v = (x1 + x2) + (y1 + y2) ∈ S. To verify (M1), observe that X and Y are both closed with respect to scalar multiplication so that αx1 ∈ X and αy1 ∈ Y for all α, and consequently αu = αx1 + αy1 ∈ S for all α. To prove (4.1.2), suppose SX = {x1, x2, ..., xr} and SY = {y1, y2, ..., yt}, and write z ∈ X + Y ⟺ z = x + y with x ∈ X, y ∈ Y ⟺ z = Σ αi xi + Σ βi yi ⟺ z ∈ span (SX ∪ SY). Example 4.1.8 If X ⊆ ℜ2 and Y ⊆ ℜ2 are subspaces defined by two different lines through the origin, then X + Y = ℜ2. This follows from the parallelogram law—sketch a picture for yourself.

Exercises for section 4.1 4.1.1. Determine which of the following subsets of ℜn are in fact subspaces of ℜn (n > 2). (a) {x | xi ≥ 0}, (b) {x | x1 = 0}, (c) {x | x1x2 = 0}, (d) {x | Σ xj = 0}, (e) {x | Σ xj = 1}, (f) {x | Ax = b, where Am×n ≠ 0 and bm×1 ≠ 0}. 4.1.2. Determine which of the following subsets of ℜn×n are in fact subspaces of ℜn×n. (a) The symmetric matrices. (b) The diagonal matrices. (c) The nonsingular matrices. (d) The singular matrices. (e) The triangular matrices. (f) The upper-triangular matrices. (g) All matrices that commute with a given matrix A. (h) All matrices such that A2 = A. (i) All matrices such that trace (A) = 0. 4.1.3. If X is a plane passing through the origin in ℜ3 and Y is the line through the origin that is perpendicular to X, what is X + Y? 168 Chapter 4 Vector Spaces 4.1.4. Why must a real or complex nonzero vector space contain an infinite number of vectors? 4.1.5. Sketch a picture in ℜ3 of the subspace spanned by each of the following. (a) {(1; 3; 2), (2; 6; 4), (−3; −9; −6)}, (b) {(−4; 0; 0), (0; 5; 0), (1; 1; 0)}, (c) {(1; 0; 0), (1; 1; 0), (1; 1; 1)}. 4.1.6. Which of the following are spanning sets for ℜ3? (a) {(1 1 1)} (b) {(1 0 0), (0 0 1)} (c) {(1 0 0), (0 1 0), (0 0 1), (1 1 1)} (d) {(1 2 1), (2 0 −1), (4 4 1)} (e) {(1 2 1), (2 0 −1), (4 4 0)}. 4.1.7. For a vector space V, and for M, N ⊆ V, explain why span (M ∪ N) = span (M) + span (N). 4.1.8. Let X and Y be two subspaces of a vector space V. (a) Prove that the intersection X ∩ Y is also a subspace of V. (b) Show that the union X ∪ Y need not be a subspace of V. 4.1.9. For A ∈ ℜm×n and S ⊆ ℜn×1, the set A(S) = {Ax | x ∈ S} contains all possible products of A with vectors from S. We refer to A(S) as the set of images of S under A. (a) If S is a subspace of ℜn, prove A(S) is a subspace of ℜm. (b) If s1, s2, ..., sk spans S, show As1, As2, ..., Ask spans A(S). 4.1.10. With the usual addition and multiplication, determine whether or not the following sets are vector spaces over the real numbers. (a) ℜ, (b) C, (c) The rational numbers. 4.1.11. Let M = {m1, m2, ..., mr} and N = {m1, m2, ..., mr, v} be two sets of vectors from the same vector space. Prove that span (M) = span (N) if and only if v ∈ span (M). 4.1.12. For a set of vectors S = {v1, v2, ..., vn}, prove that span (S) is the intersection of all subspaces that contain S. Hint: For M = ∩ V, prove that span (S) ⊆ M and M ⊆ span (S).

4.2 FOUR FUNDAMENTAL SUBSPACES

The closure properties (A1) and (M1) on p. 162 that characterize the notion of a subspace have much the same "feel" as the definition of a linear function as stated on p. 89, but there's more to it than just a "similar feel." Subspaces are intimately related to linear functions as explained below. Subspaces and Linear Functions For a linear function f mapping ℜn into ℜm, let R(f) denote the range of f. That is, R(f) = {f(x) | x ∈ ℜn} ⊆ ℜm is the set of all "images" as x varies freely over ℜn. • The range of every linear function f : ℜn → ℜm is a subspace of ℜm, and every subspace of ℜm is the range of some linear function. For this reason, subspaces of ℜm are sometimes called linear spaces. Proof. If f : ℜn → ℜm is a linear function, then the range of f is a subspace of ℜm because the closure properties (A1) and (M1) are satisfied. Establish (A1) by showing that y1, y2 ∈ R(f) ⟹ y1 + y2 ∈ R(f). If y1, y2 ∈ R(f), then there must be vectors x1, x2 ∈ ℜn such that y1 = f(x1) and y2 = f(x2), so it follows from the linearity of f that y1 + y2 = f(x1) + f(x2) = f(x1 + x2) ∈ R(f). Similarly, establish (M1) by showing that if y ∈ R(f), then αy ∈ R(f) for all scalars α by using the definition of range along with the linearity of f to write y ∈ R(f) ⟹ y = f(x) for some x ∈ ℜn ⟹ αy = αf(x) = f(αx) ∈ R(f). Now prove that every subspace V of ℜm is the range of some linear function f : ℜn → ℜm. Suppose that {v1, v2, ..., vn} is a spanning set for V so that V = {α1v1 + ··· + αnvn | αi ∈ ℜ}. (4.2.1) Stack the vi's as columns in a matrix Am×n = [v1 | v2 | ··· | vn], and put the αi's in an n × 1 column x = (α1, α2, ..., αn)T to write V = {Ax | x ∈ ℜn×1}. (4.2.2) The function f(x) = Ax is linear (recall Example 3.6.1, p. 106), and we have that R(f) = {Ax | x ∈ ℜn×1} = V. 170 Chapter 4 Vector Spaces In particular, this result means that every matrix A ∈ ℜm×n generates a subspace of ℜm by means of the range of the linear function f(x) = Ax. Likewise, the transpose of A ∈ ℜm×n defines a subspace of ℜn by means of the range of f(y) = ATy. These two "range spaces" are two of the four fundamental subspaces defined by a matrix. Range Spaces The range of a matrix A ∈ m×n is defined to be the subspace R (A) of m that is generated by the range of f (x) = Ax. That is, R (A) = {Ax | x ∈ n } ⊆ m . Similarly, the range of AT is the subspace of n defined by R AT = {AT y | y ∈ m } ⊆ n . Because R (A) is the set of all "images" of vectors x ∈ m under transformation by A, some people call R (A) the image space of A. The observation (4.2.2) that every matrix–vector product Ax (i.e., every image) is a linear combination of the columns of A provides a useful characterization of the range spaces. Allowing the components of x = (ξ1 , ξ2 , . . . , ξn)T to vary freely and writing ⎛ ⎞ξ1 n ⎜ ξ2 ⎟ = Ax = A∗1 | A∗2 | · · · | A∗n ⎜ ξj A∗j . ⎝ . ⎠. j=1 ξn shows that the set of all images Ax is the same as the set of all linear combinations of the columns of A. Therefore, R (A) is nothing more than the space spanned by the columns of A. That's why R (A) is often called the column space of A. Likewise, R AT is the space spanned by the columns of AT. But the columns of AT are just the rows of A (stacked upright), so R AT is simply 25 the space spanned by the rows of A. Consequently, R AT is also known as the row space of A. Below is a summary. 24 25 For ease of exposition, the discussion in this section is in terms of real matrices and real spaces, but all results have complex analogs obtained by replacing AT by A∗ . Strictly speaking, the range of AT is a set of columns, while the row space of A is a set of rows. However, no logical difficulties are encountered by considering them to be the same. 4.2 Four Fundamental Subspaces 171 Column and Row Spaces For A ∈ m×n , the following statements are true. • • • • Example 4.2.1 R (A) = the space spanned by the columns of A (column space). R AT = the space spanned by the rows of A (row space). b ∈ R (A) ⟺ b = Ax for some x. a ∈ R AT ⟺ aT = yT A for some yT . (4.2.3) (4.2.4) 1 2 3 Problem: Describe R (A) and R AT for A = 2 4 6 . Solution: R (A) = span {A∗1 , A∗2 , A∗3 } = {α1 A∗1 +α2 A∗2 +α3 A∗3 | αi ∈ }, but since A∗2 = 2A∗1 and A∗3 = 3A∗1 , it's clear that every linear combination of A∗1 , A∗2 , and A∗3 reduces to a multiple of A∗1 , so R (A) = span {A∗1 } . 2 Geometrically, (A) is the line in through the origin and the point (1, 2). R T = span {A1∗ , A2∗ } = {α1 A1∗ + α2 A2∗ | α1 , α2 ∈ } . But Similarly, R A A2∗ = 2A1∗ implies that combination of A1∗ and A2∗ reduces to a every multiple of A1∗ , so R AT = span {A1∗ } , and this is a line in 3 through the origin and the point (1, 2, 3). There are times when it is desirable to know whether or not two matrices have the same row space or the same range. The following theorem provides the solution to this problem. Equal Ranges For two matrices A and B of the same shape: row • R AT = R BT if and only if A ∼ B. • col R (A) = R (B) if and only if A ∼ B. row (4.2.5) (4.2.6) Proof. To prove (4.2.5), first assume A ∼ B so there exists a nonsingular that matrix P such that PA = B. To see that R AT = R BT , use (4.2.4) to write a ∈ R AT ⟺ aT = yT A = yT P−1 PA for some yT ⟺ aT = zT B for zT = yT P−1 ⟺ a ∈ R BT . 172 Chapter 4 Vector Spaces Conversely, if R AT = R BT , then span {A1∗ , A2∗ , . . . , Am∗ } = span {B1∗ , B2∗ , . . . , Bm∗ } , so each row of B is a combination of the rows of A, and vice versa. On the basis of this fact, it can be argued that it is possible to reduce A to B by using row only row operations (the tedious details are omitted), and thus A ∼ B. The T T proof of (4.2.6) follows by replacing A and B with A and B . Example 4.2.2 Testing Spanning Sets. Two sets {a1 , a2 , . . . , ar } and {b1 , b2 , . . . , bs } in n span the same subspace if and only if the nonzero rows of EA agree with the nonzero rows of EB , where A and B are the matrices containing the ai 's and bi 's as rows. This is a corollary of (4.2.5) because zero rows are irrelevant in considering the row space of a matrix, and we already know from (3.9.9) that row A ∼ B if and only if EA = EB . Problem: Determine whether or not the following sets span the same subspace: ⎧⎛ ⎞⎛ ⎞⎛ ⎞⎫ ⎧⎛ ⎞⎛ ⎞⎫ 1 2 3 ⎪ 0 1 ⎪ ⎪ ⎪ ⎨ ⎬⎨ ⎬ ⎪2 ⎟⎜4 ⎟⎜6 ⎟⎜0 ⎟⎜2 ⎟A = ⎝ ⎠, ⎝ ⎠, ⎝ ⎠, B = ⎝ ⎠, ⎝ ⎠. 1 1 ⎪ 3 ⎪ ⎪ ⎪ ⎩ 2 ⎭ ⎩ 1 ⎭ 3 3 4 1 4 Solution: Place the vectors ⎛ 1 2 A = ⎝2 4 3 6 and 0 0 B= 1 2 as rows in matrices ⎞⎛ 2 3 1 2 1 3⎠→ ⎝0 0 1 4 0 0 1 3 1 4 → 1 0 2 0 A and B, and compute ⎞0 1 1 1 ⎠= EA 0 0 0 1 1 1 = EB . Hence span {A} = span {B} because the nonzero rows in EA and EB agree. We already know that the rows of A span R AT , and the columns of A span R (A), but it's often possible to span these spaces with fewer vectors than the full set of rows and columns. Spanning the Row Space and Range Let A be an m × n matrix, and let U be any row echelon form derived from A. Spanning sets for the row and column spaces are as follows: • The nonzero rows of U span R AT . (4.2.7) • The basic columns in A span R (A). (4.2.8) 4.2 Four Fundamental Subspaces 173 Proof. Statement (4.2.7) is an immediate consequence of (4.2.5). To prove (4.2.8), suppose that the basic columns in A are in positions b1 , b2 , . . . , br , and the nonbasic columns occupy positions n1 , n2 , . . . , nt , and let Q1 be the permutation matrix that permutes all of the basic columns in A to the left-hand side so that AQ1 = (Bm×r Nm×t) , where B contains the basic columns and N contains the nonbasic columns. Since the nonbasic columns are linear combinations of the basic columns—recall (2.2.3)—we can annihilate the nonbasic columns in N using elementary column operations. In other words, there is a nonsingular matrix Q2 such that (B N) Q2 = (B 0) . Thus Q = Q1 Q2 is a nonsingular matrix such that AQ = AQ1 Q2 = (B N) Q2 = (B 0) , and col hence A ∼ (B Example 4.2.3 0). The conclusion (4.2.8) now follows from (4.2.6). Problem: Determine spanning sets for R (A) and R AT , where ⎛ ⎞1 2 2 3 A = ⎝2 4 1 3⎠. 3 6 1 4 Solution: Reducing A to any row echelon form U provides the solution—the basic columns in A correspond to the pivotal positions the nonzero in U, and rows of U span the row space of A. Using EA = ⎧⎛ ⎞⎛ ⎞⎫ 2 ⎬⎨ 1 R (A) = span ⎝ 2 ⎠, ⎝ 1 ⎠⎝ ⎠ 3 1 and R A T 1 0 0 2 0 0 0 1 0 1 1 0 produces ⎧⎛ ⎞⎛ ⎞⎫ 1 0 ⎪ ⎨ ⎬ ⎪2 ⎟⎜0 ⎟= span ⎝ ⎠, ⎝ ⎠. 1 ⎪ ⎪ ⎩ 0 ⎭ 1 1 So far, only two of the four fundamental subspaces associated each with matrix A ∈ m×n have been discussed, namely, R (A) and R AT . To see where the other two fundamental subspaces come from, consider again a general linear function f mapping n into m , and focus on N (f) = {x | f (x) = 0} (the set of vectors that are mapped to 0). N (f) is called the nullspace of f (some texts call it the kernel of f), and it's easy to see that N (f) is a subspace of n because the closure properties (A1) and (M1) are satisfied. Indeed, if x1 , x2 ∈ N (f), then f (x1) = 0 and f (x2) = 0, so the linearity of f produces f (x1 + x2) = f (x1) + f (x2) = 0 + 0 = 0 ⟹ x1 + x2 ∈ N (f). (A1) Similarly, if α ∈ , and if x ∈ N (f), then f (x) = 0 and linearity implies f (αx) = αf (x) = α0 = 0 ⟹ αx ∈ N (f). (M1) T By considering the linear functions f (x) = Ax and g(y) = A y, the m×n other two fundamental subspaces defined by A are obtained. They are ∈ n N (f) = {xn×1 | Ax = 0} ⊆ and N (g) = ym×1 | AT y = 0 ⊆ m . 174 Chapter 4 Vector Spaces Nullspace • • For an m × n matrix A, the set N (A) = {xn×1 | Ax = 0} ⊆ n is called the nullspace of A. In other words, N (A) is simply the set of all solutions to the homogeneous system Ax = 0. m The set N AT = ym×1 | AT y = 0T ⊆ is called the lefthand nullspace of A because N A is the set of all solutions to the left-hand homogeneous system yT A = 0T . Example 4.2.4 Problem: Determine a spanning set for N (A), where A = 1 2 2 4 3 6 . Solution: N (A) is merely the general solution of Ax = 0, and this is determined by reducing A to a row echelon formU. As discussed in §2.4, any such U will suffice, so we will use EA = 10 02 03 . Consequently, x1 = −2x2 − 3x3 , where x2 and x3 are free, so the general solution of Ax = 0 is ⎛ ⎞⎛ ⎞⎛ ⎞⎛ ⎞x1 −2x2 − 3x3 −2 −3 ⎝ x2 ⎠= ⎝ ⎠= x2 ⎝ 1 ⎠+ x3 ⎝ 0 ⎠. x2 0 1 x3 x3 In other words, N (A) is the set of all possible linear combinations of the vectors ⎛ ⎞−2 h1 = ⎝ 1 ⎠0 ⎛ and ⎞−3 h2 = ⎝ 0 ⎠, 1 and therefore span {h1 , h2 } = N (A). For this example, N (A) is the plane in 3 that passes through the origin and the two points h1 and h2 . Example 4.2.4 indicates the general technique for determining a spanning set for N (A). Below is a formal statement of this procedure. 4.2 Four Fundamental Subspaces 175 Spanning the Nullspace To determine a spanning set for N (A), where rank (Am×n) = r, row reduce A to a row echelon form U, and solve Ux = 0 for the basic variables in terms of the free variables to produce the general

[illegible]

solution if and only if rank (A) = n, −1 T in which case the unique solution is x = AT A A b. • When Ax = b is consistent and has a unique solution, then the same is true for AT Ax = AT b, and the unique solution to both −1 T systems is given by x = AT A A b. Example 4.5.1 Caution! Use of the product AT A or the normal equations is not recommended for numerical computation. Any sensitivity to small perturbations that is present in the underlying matrix A is magnified by forming the product AT A. In other words, if Ax = b is somewhat ill-conditioned, then the associated system of normal equations AT Ax = AT b will be ill-conditioned to an even greater extent, and the theoretical properties surrounding AT A and the normal equations may be lost in practical applications. For example, consider the nonsingular system Ax = b, where 3 6 9 A= and b = 1 2.01 3.01 If Gaussian elimination with 3-digit floating-point arithmetic is used to solve Ax = b, then the 3-digit solution is (1, 1), and this agrees with the exact 4.5 More about Rank 217 solution. However if 3-digit arithmetic is used to form the associated system of normal equations, the result is 10 20 20 40 x1 x2 = 30 60.1 . The 3-digit representation of AT A is singular, and consequently the associated system of normal equations is inconsistent. For these reasons, the normal equations are often avoided in numerical computations. Nevertheless, the normal equations are an important theoretical idea that leads to practical tools of fundamental importance such as the method of least squares developed in §4.6 and §5.13. Because the concept of rank is at the heart of our subject, it's important to understand rank from a variety of different viewpoints. The statement below is 29 one more way to think about rank. Rank and the Largest Nonsingular Submatrix The rank of a matrix Am×n is precisely the order of a maximal square nonsingular submatrix of A. In other words, to say rank (A) = r means that there is at least one r × r nonsingular submatrix in A, and there are no nonsingular submatrices of larger order. Proof. First demonstrate that there exists an r × r nonsingular submatrix in A, and then show there can be no nonsingular submatrix of larger order. Begin with the fact that there must be a maximal linearly independent set of r rows in A as well as a maximal independent set of r columns, and prove that the submatrix Mr×r lying on the intersection of these r rows and r columns is nonsingular. The r independent rows can be permuted to the top, and the remaining rows can be annihilated using row operations. This is the last characterization of rank presented in this text, but historically this was the essence of the first definition (p. 44) of rank given by Georg Frobenius (p. 662) in 1879. 216 Chapter 4 Vector Spaces Rank isn't changed by row or column operations, so r = rank (A) = rank (M), and thus M is nonsingular. Now suppose that W is any other nonsingular submatrix of A, and let P and Q be permutation matrices such that PAQ = W ... Therefore, rank (A) ≥ rank (W), and consequently W can be no larger than r × r, or else rank (A) would be larger than r. ... Example 4.5.2 Problem: Determine the rank of A = 2 3 4 6 11 ... Solution: rank (A) = 2 because there is at least one 2 × 2 nonsingular submatrix (e.g., there is the one lying on the intersection of rows 1 and 2 with columns 1 and 2), and there is no larger nonsingular submatrix (the entire matrix is singular). Notice that not all 2 × 2 matrices are nonsingular (e.g., consider the one lying on the intersection of rows 1 and 2 with columns 1 and 2). Earlier in this section we saw that it is impossible to increase the rank by means of matrix multiplication—i.e., (4.5.2) says rank (AE) ≤ rank (A). In a certain sense there is a dual statement for matrix addition that says that it is impossible to decrease the rank by means of a "small" matrix addition—i.e., rank (A + E) ≥ rank (A) whenever E has entries of sufficiently small magnitude. Small Perturbations Can't Reduce Rank If A and E are m × n matrices such that E has entries of sufficiently small magnitude, then rank (A + E) ≥ rank (A). (4.5.7) The term "sufficiently small" is further clarified in Exercise 5.12.4. 4.5 More about Rank 217 Proof. Suppose rank (A) = r, and let P and Q be nonsingular matrices that reduce A to rank normal form—i.e., PAQ = Ir 0 0 0 . If P and Q are applied to E to form PEQ = E11 E12 E21 E22 , where E11 is r × r, then applying ... PAQ + PEQ = Ir + E11 E12 E21 E22 . (4.5.8) If the magnitude of the entries in E are small enough to insure that Ek11 → 0 as k → ∞, then the discussion of the Neumann series on p. 126 insures that I + E11 is nonsingular. (Exercise 5.12.4 gives a condition on the size of E11 to insure this.) This allows the right-hand side of (4.5.8) to be further reduced by writing ... In other words, A + E ∼ ... and therefore rank (A + E) = rank (Ir + E11) + rank (S) (recall Example 3.9.3) = rank (A) + rank (S) (4.5.11) ≥ rank (A). Example 4.5.3 A Pitfall in Solving Singular Systems. Solving Ax = b with floating-point arithmetic produces the exact solution of a perturbed system whose coefficient matrix is A + E. If A is nonsingular, and if we are using a stable algorithm (an algorithm that insures that the entries in E have small magnitudes), then (4.5.9) guarantees that we are finding the exact solution to a nearby system that is also nonsingular. On the other hand, if A is singular, then perturbations of even the slightest magnitude can increase the rank, thereby producing a system with fewer free variables than the original system theoretically demands, so even a stable algorithm can result in a significant loss of information. But what are the chances that this will actually occur in practice? To answer this, recall from (4.5.11) that rank (A + E) = rank (A) + rank (S), where S = E22 − E21 (I + E11)−1 E12 . 218 Chapter 4 Vector Spaces If the rank is not to jump, then the perturbation E must be such that S = 0, which is equivalent to saying E22 = E21 (I + E11)−1 E12 . Clearly, this requires the existence of a very specific (and quite special) relationship among the entries of E, and a random perturbation will almost never produce such a relationship. Although rounding errors cannot be considered to be truly random, they are random enough so as to make the possibility that S = 0 very unlikely. Consequently, when A is singular, the small perturbation E due to roundoff makes the possibility that rank (A + E) > rank (A) very likely. The moral is to avoid floating-point solutions of singular systems. Singular problems can often be distilled down to a nonsingular core or to nonsingular pieces, and these are the components you should be dealing with. Since no more significant characterizations of rank will be given, it is appropriate to conclude this section with a summary of all of the different ways we have developed to say "rank." Summary of Rank For A ∈ m×n , each of the following statements is true. • rank (A) = The number of nonzero rows in any row echelon form that is row equivalent to A. • rank (A) = The number of pivots in the process of reducing A to a row echelon form with row operations. • rank (A) = The number of basic columns in A (or basic rows in AT). • rank (A) = The number of independent columns in A—i.e., the size of a maximal independent set of columns from A. • rank (A) = The number of independent rows in A—i.e., the size of a maximal independent set of rows from A. • rank (A) = dim R (A). • rank (A) = dim R AT . • rank (A) = n − dim N (A). • rank (A) = m − dim N AT . • rank (A) = The size of the largest nonsingular submatrix in A. For A ∈ C m×n , replace ∗T with ∗. 4.5 More about Rank 219 Exercises for section 4.5 4.5.1. Verify that rank AT A = rank (A) = rank AAT for ... (b) A = ... 4.5.2. Determine dim N (A) ∩ R (B) for ... 4.5.3. For the matrices given in Exercise 4.5.2, use the procedure described on p. 211 to determine a basis for N (A) ∩ R (B). 4.5.4. If A1 A2 ··· Ak is a product of square matrices such that some Ai is singular, explain why the entire product must be singular. 4.5.5. For A ∈ m×n , explain why AT A = 0 implies A = 0. 4.5.6. Find rank (A) and all nonsingular submatrices of maximal order in ... 4.5.7. Is it possible that rank (AB) < rank (A) and rank (AB) < rank (B) for the same pair of matrices? 4.5.8. Is rank (AB) = rank (BA) when both products are defined? Why? 4.5.9. Explain why rank (AB) = rank (A) − dim N (A) ∩ R (B). 4.5.10. Explain why dim N (Am×n Bn×p) = dim N (B) + dim R (B) ∩ N (A). 220 Chapter 4 Vector Spaces 4.5.11. Sylvester's law of nullity given by James J. Sylvester in 1884, states that for square matrices A and B, max {ν(A), ν(B)} ≤ ν(AB) ≤ ν(A) + ν(B), where ν(∗) = dim N (∗) denotes the nullity. (a) Establish the validity of Sylvester's law. (b) Show Sylvester's law is not valid for rectangular matrices because ν(A) > ν(AB) is possible. Is ν(B) > ν(AB) possible? 4.5.12. For matrices Am×n and Bn×p , prove each of the following statements: (a) rank (AB) = rank (A) and R (AB) = R (A) if rank (B) = n. (b) rank (AB) = rank (B) and N (AB) = N (B) if rank (A) = n. 4.5.13. Perform the following calculations using the matrices: ... (a) Find rank (A), and solve Ax = b using exact arithmetic. (b) Find rank AT A , and solve AT Ax = AT b exactly. (c) Find rank (A), and solve Ax = b with 3-digit arithmetic. (d) Find AT A, AT b, and the solution of AT Ax = AT b with 3-digit arithmetic. 4.5.14. Prove that if the entries of Fr×r satisfy ... for each i (i.e., F is diagonally dominant), then I + F is nonsingular. Hint: Use the triangle inequality for scalars |α + β| ≤ |α| + |β| to show N (I + F) = 0. 4.5.15. Construct an example to show that rank X 0 0 Y ... 4.5.16. Prove that if A is nonsingular, then ... 4.5.17. Prove that each of the following statements is true ... 4.5.18. If X satisfies ... 4.5.19. Let A and B be n × n matrices ... 4.5.20. Moore–Penrose Inverse. For A ∈ m×n such that rank (A) = r, let A = BC be the full rank factorization of A in which Bm×r and Cr×n ... The matrix ... is called the Moore–Penrose inverse of A. Some authors refer to A† as the pseudoinverse or the generalized inverse of A. A more elegant treatment is given on p. 423, but it's worthwhile to introduce the idea here so that it can be used and viewed from different perspectives. (a) Explain why the matrix BT ACT is nonsingular. (b) Verify that x = A† b solves the normal equations AT Ax = AT b (as well as Ax = b when it is consistent). (c) Show that the general solution for AT Ax = AT b (as well as Ax = b when it is consistent) can be described as x = A† b + I − A† A h, 30 This is in honor of Eliakim H. Moore (1862–1932) and Roger Penrose (a famous contemporary English mathematical physicist). Each formulated a concept of generalized matrix inversion—Moore's work was published in 1922, and Penrose's work appeared in 1955. E. H. Moore is considered by many to be America's first great mathematician. 222 Chapter 4 Vector Spaces where h is a "free variable" vector in n . Hint: Verify AA† A = A. (d) Verify that R I − A† A = N (A). (e) If A is square and nonsingular, explain why A† = A−1 . (f) Verify that A† satisfies the Penrose equations: AA† A = A, A† AA† = A† , (AA†)T = AA† , (A† A)T = A† A. Penrose originally defined A† to be the unique solution to these four equations. 4.6 Classical Least Squares 4.6 223 CLASSICAL LEAST SQUARES The following problem arises in almost all areas where mathematics is applied. At discrete points ti (often points in time), observations bi of some phenomenon are made, and the results are recorded as a set of ordered pairs D = {(t1 , b1), (t2 , b2), ... , (tm , bm)} . On the basis of these observations, the problem is to make estimations or predictions at points (times) t̂ that are between or beyond the observation points ti . A standard approach is to find the equation of a curve y = f (t) that closely fits the points in D so that the phenomenon can be estimated at any nonobservation point t̂ with the value ŷ = f (t̂). Let's begin by fitting a straight line to the points in D. Once this is understood, it will be relatively easy to see how to fit the data with curved lines. ... Figure 4.6.1 The strategy is to determine the coefficients α and β in the equation of the line f (t) = α + βt that best fits the points (ti , bi) in the sense that the sum of the squares of the vertical errors ε1 , ε2 , ... , εm indicated in Figure 4.6.1 is minimal. The distance from the point (ti , bi) to the line is εi = |f (ti) − bi | = |α + βti − bi |. If we consider only vertical errors because there is a tacit assumption that only the observations bi are subject to error or variation. The ti 's are assumed to be errorless constants—think of them as being exact points in time (as they often are). If the ti 's are also subject to variation, then horizontal as well as vertical errors have to be considered in Figure 4.6.1, and a more complicated theory known as total least squares (p. 223) emerges. The least squares line L obtained by minimizing only vertical deviations will not be the closest line to points in D in terms of perpendicular distance, but L is the best line for the purpose of linear estimation—see §5.14 (p. 446). 224 Chapter 4 Vector Spaces minimal. The distance from (ti , bi) to a line f (t) = α + βt is εi = |f (ti) − bi | = |α + βti − bi |, so that the objective is to find values for α and β such that ... is minimal. Minimization techniques from calculus tell us that the minimum value must occur at a solution to the two equations ... Rearranging terms produces two equations in the two unknowns α and β ... (4.6.1) By setting A = ..., we see that the two equations (4.6.1) have the matrix form AT Ax = AT b. In other words, (4.6.1) is the system of normal equations associated with the system Ax = b (see p. 213). The (4.6.1) is consistent, and the normal equations always possess a solution. ... Finally, notice that the total sum of squares of the errors is given by ... 4.6 Classical Least Squares 225 Example 4.6.1 Problem: A small company has been in business for four years and has recorded annual sales (in tens of thousands of dollars) as follows. Year 1 2 3 4 Sales 23 27 30 34 When this data is plotted as shown in Figure 4.6.2, we see that although the points do not exactly lie on a straight line, there nevertheless appears to be a linear trend. Predict the sales for any future year if this trend continues. 34 33 32 Sales 31 30 29 28 27 26 25 24 23 22 0 1 2 3 4 Year Figure 4.6.2 Solution: Determine the line f (t) = α + βt that best fits the data in the sense of least squares. ... so the normal equations AT Ax = AT b are ... α = 19.5 and β = 3.6. Therefore, the estimated sales for year five is $375,000. To get a feel for how close the least squares line comes to 226 Chapter 4 Vector Spaces passing through the data points, let ε = Ax − b, and compute the sum of the squares of the errors to be ... = .2. General Least Squares Problem For A ∈ m×n and b ∈ m , let ε = ε(x) = Ax − b. The general least squares problem is to find a vector x that minimizes the quantity ... = (Ax − b)T (Ax − b). Any vector that provides a minimum value for this expression is called a least squares solution. • The set of all least squares solutions is precisely the set of solutions to the system of normal equations AT Ax = AT b. • There is a unique least squares solution if and only if rank (A) = n, in which case it is x = (AT A)−1 AT b. • If Ax = b is consistent, then the solution set for Ax = b is the same as the set of least squares solutions. 32 Proof. First prove that if x minimizes ε(x), then x must satisfy the normal equations. Begin by using xT AT b = bT Ax (scalars are symmetric) to write ... (4.6.2) To determine vectors x that minimize the expression (4.6.2), we will again use minimization techniques from calculus and differentiate the function ... (4.6.3) with respect to each xi . Differentiating matrix functions is similar to differentiating scalar functions (see Exercise 3.5.9) in the sense that ... 32 A more modern development not relying on calculus is given in §5.13 on p. 437, but the more traditional approach is given here because it's worthwhile to view least squares from both perspectives. 4.6 Classical Least Squares 227 Applying these rules to the function in (4.6.3) produces ... Since ... we see that ... and hence ... for i = 1, 2, ..., n. Since the minimum value of f (x1 , x2 , ..., xn) must occur at a point where all derivatives vanish, it follows that AT Ax = AT b. ... Using (4.6.2) produces ... So we must show that the function f in (4.6.3) attains its minimum value at each solution to AT Ax = AT b. Observe that if z is a solution to the normal equations, then f (z) = bT b − zT AT b. For any other y ∈ n , let u = y − z, so y = z + u, and observe that f (y) = f (z) + vT v, where v = Au. Since vT v ≥ 0, it follows that f (z) ≤ f (y) for all y ∈ n , and thus f attains its minimum value at each solution of the normal equations. The remaining statements in the theorem follow from the properties established on p. 213. The classical least squares problem discussed at the beginning of this section and illustrated in Example 4.6.1 is part of a broader topic known as linear regression, which is the study of situations where attempts are made to express one variable y as a linear combination of other variables t1 , t2 , ..., tn . In practice, hypothesizing that y is linearly related to t1 , t2 , ..., tn means that one assumes the existence of a set of constants {α0 , α1 , ..., αn } (called parameters) such that y = α0 + α1 t1 + α2 t2 + ··· + αn tn + ε, where ε is a "random function" whose values "average out" to zero in some sense. Practical problems almost always involve more variables than we wish to consider, but it is frequently fair to assume that the effect of variables of lesser significance will indeed "average out" to zero. The random function ε accounts for this assumption. In other words, a linear hypothesis is the supposition that the expected (or mean) value of y at each point where the phenomenon can be observed is given by a linear equation E(y) = α0 + α1 t1 + α2 t2 + ··· + αn tn . 228 Chapter 4 Vector Spaces To help seat these ideas, consider the problem of predicting the amount of weight that a pint of ice cream loses when it is stored at very low temperatures. There are many factors that may contribute to weight loss—e.g., storage temperature, storage time, humidity, atmospheric pressure, butterfat content, the amount of corn syrup, the amounts of various gums (guar gum, carob bean gum, locust bean gum, cellulose gum), and the never-ending list of other additives and preservatives. It is reasonable to believe that storage time and temperature are the primary factors, so to predict weight loss we will make a linear hypothesis of the form y = α0 + α1 t1 + α2 t2 + ε, where y = weight loss (grams), t1 = storage time (weeks), t2 = storage temperature (o F), and ε is a random function to account for all other factors. The assumption is that all other factors "average out" to zero, so the expected (or mean) weight loss at each point (t1 , t2) is E(y) = α0 + α1 t1 + α2 t2 . (4.6.4) Suppose that we conduct an experiment in which values for weight loss are measured for various values of storage time and temperature as shown below. Time (weeks) If 1 1 1 2 2 2 3 3 Temp (o F) −10 −5 0 −10 −5 0 −10 −5 0 Loss (grams) .15 .18 .20 .17 .19 .22 .20 .23 .25 3 If ... and if we were lucky enough to exactly observe the mean weight loss each time (i.e., if bi = E(yi)), then equation (4.6.4) would insure that Ax = b is a consistent system, so we could solve for the unknown parameters α0 , α1 , and α2 . However, it is virtually impossible to observe the exact value of the mean weight loss for a given storage time and temperature, and almost certainly the system defined by Ax = b will be inconsistent—especially when the number of observations greatly exceeds the number of parameters. Since we can't solve Ax = b to find exact values for the αi 's, the best we can hope for is a set of "good estimates" for these parameters. 4.6 Classical Least Squares 229 The famous Gauss–Markov theorem (developed on p. 448) states that under certain reasonable assumptions concerning the random error function ε, the "best" estimates for the αi 's are obtained by minimizing the sum of squares (Ax − b)T (Ax − b). In other words, the least squares estimates are the "best" way to estimate the αi 's. Returning to our ice cream example, it can be verified that b ∉ R (A), so, as expected, the system Ax = b is not consistent, and we cannot determine exact values for α0 , α1 , and α2 . The best we can do is to determine least squares estimates for the αi 's by solving the associated normal equations AT Ax = AT b, which in this example are ... The solution is ... and the estimating equation for mean weight loss becomes ŷ = .174 + .025t1 + .005t2 . For example, the mean weight loss of a pint of ice cream that is stored for nine weeks at a temperature of −35o F is estimated to be ŷ = .174 + .025(9) + .005(−35) = .224 grams. Example 4.6.2 Least Squares Curve Fitting Problem: Find a polynomial p(t) = α0 + α1 t + α2 t2 + ··· + αn−1 tn−1 with a specified degree that comes as close as possible in the sense of least squares to passing through a set of data points D = {(t1 , b1), (t2 , b2), ..., (tm , bm)} , where the ti 's are distinct numbers, and n ≤ m. 230 Chapter 4 Vector Spaces ... Solution: For the εi 's indicated in Figure 4.6.3, the objective is to minimize the sum of squares ... In other words, the least squares polynomial is unique because Am×n is the Vandermonde matrix of Example 4.3.4 with n ≤ m, so rank (A) = n, and the least squares polynomial is given by x = AT A A b. Note: We know from Example 4.3.5 on p. 186 that the Lagrange interpolation polynomial ℓ(t) of degree m − 1 will exactly fit the data—i.e., it passes through each point in D. So why would one want to settle for a least squares fit when an exact fit is possible? One answer stems from the fact that in practical work the observations bi are rarely exact due to small errors arising from imprecise measurements or from simplifying assumptions. For this reason, it is the trend of the observations that needs to be fitted and not the observations themselves. To hit the data points, the interpolation polynomial ℓ(t) is usually forced to oscillate between or beyond the data points, and as m becomes larger the oscillations can become more pronounced. Consequently, ℓ(t) is generally not useful in making estimations concerning the trend of the observations—Example 4.6.3 drives this point home. In addition to exactly hitting a prescribed set of data points, an interpolation polynomial called the Hermite polynomial (p. 607) can be constructed to have specified derivatives at each data point. While this helps, it still is not as good as least squares for making estimations on the basis of observations. Example 4.6.3 A missile is fired from enemy territory, and its position in flight is observed by radar tracking devices at the following positions. Position down range (miles) 0 250 500 750 1000 Height (miles) 0 8 15 19 20 Suppose that our intelligence sources indicate that enemy missiles are programmed to follow a parabolic flight path—a fact that seems to be consistent with the diagram obtained by plotting the observations on the coordinate system shown in Figure 4.6.4. 20 15 b = Height 10 5 0 0 250 500 750 1000 t = Range Figure 4.6.4 Problem: Predict how far down range the missile will land. 232 Chapter 4 Vector Spaces Solution: Determine the parabola f (t) = α0 + α1 t + α2 t2 that best fits the observed data in the least squares sense. Then estimate where the missile will land by determining the roots of f (i.e., determine where the parabola crosses the horizontal axis). As it stands, the problem will involve numbers having relatively large magnitudes in conjunction with relatively small ones. Consequently, it is better to first scale the data by considering one unit to be 1000 miles. ... Solving the system AT Ax = AT b ... so the least squares parabola is ... and the roots of f(t) ... It follows that the missile will land 2044 miles down range. The sum of the squares of the errors associated with the least squares solution is ... Using the Lagrange interpolation polynomial ... If we use ℓ(t) to estimate the coordinate in figure ... 4.6 Classical Least Squares 233 Least Squares vs. Lagrange Interpolation. Instead of using least squares, fit the observations exactly by using the fourth-degree Lagrange interpolation polynomial described in Example 4.3.5 on p. 186 (you can verify that ... is the fourth-degree polynomial that passes through all points). ... As the graph in Figure 4.6.5 indicates, ℓ(t) has only one real nonnegative root, so it is worthless for predicting where the missile will land. This is characteristic of Lagrange interpolation. y = ℓ(t) Figure 4.6.5 Computational Note: Theoretically, the least squares solutions of Ax = b are exactly the solutions of the normal equations AT Ax = AT b, but forming and solving the normal equations to compute least squares solutions with floating-point arithmetic is not recommended. As pointed out in Example 4.5.1 on p. 214, any sensitivities to small perturbations that are present in the underlying problem are magnified by forming the normal equations. In other words, if the underlying problem is somewhat ill-conditioned, then the system of normal equations will be ill-conditioned to an even greater extent. Numerically stable techniques that avoid the normal equations are presented in Example 5.5.3 on p. 313 and Example 5.7.3 on p. 346. Epilogue While viewing a region in the Taurus constellation on January 1, 1801, Giuseppe Piazzi, an astronomer and director of the Palermo observatory, observed a small "star" that he had never seen before. As Piazzi and others continued to watch this new "star"—which was really an asteroid—they noticed that it was in fact moving, and they concluded that a new "planet" had been discovered. However, their new "planet" completely disappeared in the autumn of 1801. Well-known astronomers of the time joined the search to relocate the lost "planet," but all efforts were in vain. 234 Chapter 4 Vector Spaces In September of 1801 Carl F. Gauss decided to take up the challenge of finding this lost "planet." Gauss allowed for the possibility of an elliptical orbit rather than constraining it to be circular—which was an assumption of the others—and he proceeded to develop the method of least squares. By December the task was completed, and Gauss informed the scientific community not only where the lost "planet" was located, but he also predicted its position at future times. They looked, and it was exactly where Gauss had predicted it would be! The asteroid was named Ceres, and Gauss's contribution was recognized by naming another minor asteroid Gaussia. This extraordinary feat of locating a tiny and distant heavenly body from apparently insufficient data astounded the scientific community. Furthermore, Gauss refused to reveal his methods, and there were those who even accused him of sorcery. These events led directly to Gauss's fame throughout the entire European community, and they helped to establish his reputation as a mathematical and scientific genius of the highest order. Gauss waited until 1809, when he published his Theoria Motus Corporum Coelestium In Sectionibus Conicis Solem Ambientium, to systematically develop the theory of least squares and his methods of orbit calculation. This was in keeping with Gauss's philosophy to publish nothing but well-polished work of lasting significance. When criticized for not revealing more motivational aspects in his writings, Gauss remarked that architects of great cathedrals do not obscure the beauty of their work by leaving the scaffolds in place after the construction has been completed. Gauss's theory of least squares approximation has indeed proven to be a great mathematical cathedral of lasting beauty and significance. Exercises for section 4.6 4.6.1. Hooke's law says that the displacement y of an ideal spring is proportional to the force x that is applied—i.e., y = kx for some constant k. Consider a spring in which k is unknown. Various masses are attached, and the resulting displacements shown in Figure 4.6.6 are observed. Using these observations, determine the least squares estimate for k. ... Figure 4.6.6 4.6.2. Show that the slope of the line that passes through the origin in 2 and comes closest in the least squares sense to passing through the points ... 4.6.3. A small company has been in business for three years and has recorded annual profits (in thousands of dollars) as follows. Year 1 2 3 Sales 7 4 3 Assuming that there is a linear trend in the declining profits, predict the year and the month in which the company begins to lose money. 4.6.4. An economist hypothesizes that the change (in dollars) in the price of a loaf of bread is primarily a linear combination of the change in the price of a bushel of wheat and the change in the minimum wage. That is, if B is the change in bread prices, W is the change in wheat prices, and M is the change in the minimum wage, then B = αW + βM. Suppose that for three consecutive years the change in bread prices, wheat prices, and the minimum wage are as shown below. Year 1 Year 2 Year 3 B +$1 +$1 +$1 W +$1 +$2 0$ M +$1 0$ −$1 Use the theory of least squares to estimate the change in the price of bread in Year 4 if wheat prices and the minimum wage each fall by $1. 4.6.5. Suppose that a researcher hypothesizes that the weight loss of a pint of ice cream during storage is primarily a linear function of time. That is, y = α0 + α1 t + ε, where y = the weight loss in grams, t = the storage time in weeks, and ε is a random error function whose mean value is 0. Suppose that an experiment is conducted, and the following data is obtained. Time (t) 1 2 3 4 5 6 7 8 Loss (y) .15 .21 .30 .41 .49 .59 .72 .83 (a) Determine the least squares estimates for the parameters α0 and α1 . (b) Predict the mean weight loss for a pint of ice cream that is stored for 20 weeks. 236 Chapter 4 Vector Spaces 4.6.6. After studying a certain type of cancer, a researcher hypothesizes that in the short run the number (y) of malignant cells in a particular tissue grows exponentially with time (t). That is, y = α0 eα1 t . Determine least squares estimates for the parameters α0 and α1 from the researcher's observed data given below. t (days) 1 2 3 4 5 y (cells) 16 27 45 74 122 Hint: What common transformation converts an exponential function into a linear function? 4.6.7. Using least squares techniques, fit the following data x −5 −4 −3 −2 −1 0 1 2 3 4 5 y 2 7 9 12 13 14 14 13 10 8 4 with a line y = α0 + α1 x and then fit the data with a quadratic y = α0 + α1 x + α2 x2 . Determine which of these two curves best fits the data by computing the sum of the squares of the errors in each case. 4.6.8. Consider the time (T) it takes for a runner to complete a marathon (26 miles and 385 yards). Many factors such as height, weight, age, previous training, etc. can influence an athlete's performance, but experience has shown that the following three factors are particularly important: x1 = Ponderal index = height (in.) / [weight (lbs.)]1/3 , x2 = Miles run the previous 8 weeks, x3 = Age (years). A linear model hypothesizes that the time T (in minutes) is given by T = α0 + α1 x1 + α2 x2 + α3 x3 + ε, where ε is a random function accounting for all other factors and whose mean value is assumed to be zero. On the basis of the five observations given below, estimate the expected marathon time for a 43-year-old runner of height 74 in., weight 180 lbs., who has run 450 miles during the previous eight weeks. T x1 x2 x3 181 193 212 221 248 13.1 12.9 13.5 13.8 13.1 619 803 207 409 482 23 42 31 38 45 What is your personal predicted mean marathon time? 4.6 Classical Least Squares 237 4.6.9. For A ∈ m×n and b ∈ m , prove that x2 is a least squares solution for Ax = b if and only if x2 is part of a solution to the larger system ... In some applications it is easier to work with this larger system. 4.6.10. Note: It is not uncommon to encounter least squares problems in which A is extremely large but very sparse (mostly zero entries). For these situations, the system (4.6.5) will usually contain significantly fewer nonzero entries than the system of normal equations, thereby helping to overcome the memory requirements that plague these problems. Using (4.6.5) also eliminates the undesirable need to explicitly form the product AT A—recall from Example 4.5.1 that forming AT A can cause loss of significant information. 4.6.10. In many least squares applications, the underlying data matrix Am×n does not have independent columns—i.e., rank (A) < n—so the corresponding system of normal equations AT Ax = AT b will fail to have a unique solution. This means that in an associated linear estimation problem of the form y = α1 t1 + α2 t2 + ··· + αn tn + ε there will be infinitely many least squares estimates for the parameters αi , and hence there will be infinitely many estimates for the mean value of y at any given point (t1 , t2 , ..., tn)—which is clearly an undesirable situation. In order to remedy this problem, we restrict ourselves to making estimates only at those points (t1 , t2 , ..., tn) that are in the row space of A. If ... then we say that ŷ = ... is an estimate of y. Explain why the estimate defined by ... is unique regardless of which least squares solution x is used. 238 4.7 Chapter 4 Vector Spaces LINEAR TRANSFORMATIONS The connection between linear functions and matrices is at the heart of our subject. As explained on p. 93, matrix algebra grew out of Cayley's observation that the composition of two linear functions can be represented by the multiplication of two matrices. It's now time to look deeper into such matters and to formalize the connections between vector spaces, and linear functions defined on vector spaces. This is the point at which linear algebra, as the study of linear functions on vector spaces, begins in earnest. Linear Transformations Let U and V be vector spaces over a field F (or C for us). • A linear transformation from U into V is defined to be a linear function T mapping U into V. That is, T(x + y) = T(x) + T(y) and T(αx) = αT(x) (4.7.1) or, equivalently, T(αx + y) = αT(x) + T(y) for all x, y ∈ U, α ∈ F. • A linear operator on U is defined to be a linear transformation from U into itself—i.e., a linear function mapping U back into U. Example 4.7.1 • The function 0(x) = 0 that maps all vectors in a space U to the zero vector in another space V is a linear transformation from U into V, and, not surprisingly, it is called the zero transformation. • The function I(x) = x that maps every vector from a space U back to itself is a linear operator on U. I is called the identity operator on U. • For A ∈ m×n and x ∈ n×1 , the function T(x) = Ax is a linear transformation from n into m because matrix multiplication satisfies A(αx + y) = αAx + Ay. T is a linear operator on n if A is n × n. • If W is the vector space of all functions from to , and if V is the space of all differentiable functions from to , then the mapping D(f) = df /dx is a linear transformation from V into W because d(αf + g)/dx = α df /dx + dg/dx. • If V is the space of all continuous functions from into , then the mapping defined by T(f) = ∫ f (t)dt is a linear operator on V because ∫ [αf (t) + g(t)]dt = α ∫ f (t)dt + ∫ g(t)dt. • 0 0 0 4.7 Linear Transformations 239 • The rotator Q that rotates vectors u in 2 counterclockwise through an angle θ, as shown in Figure 4.7.1, is a linear operator on 2 because the "action" of Q on u can be described by matrix multiplication in the sense that the coordinates of the rotated vector Q(u) are given by ... • The projector P that maps each point v = (x, y, z) ∈ 3 to its orthogonal projection (x, y, 0) in the xy-plane, as depicted in Figure 4.7.2, is a linear operator on 3 because if u = (u1 , u2 , u3) and v = (v1 , v2 , v3), then P(αu + v) = ... = αP(u) + P(v). • The reflector R that maps each vector v = (x, y, z) ∈ 3 to its reflection R(v) = (x, y, −z) about the xy-plane, as shown in Figure 4.7.3, is a linear operator on 3 . Figure 4.7.1 Figure 4.7.2 Figure 4.7.3 • Just as the rotator Q is represented by a matrix [Q] = ..., the projector P and the reflector R can be represented by matrices ... and the respective actions of P and R are given by matrix multiplication [P][v] and [R][v]. 240 Chapter 4 Vector Spaces It would be wrong to infer from Example 4.7.1 that all linear transformations can be represented by matrices (of finite size). For example, the differential and integral operators do not have matrix representations because they are defined on infinite-dimensional spaces. But linear transformations on finite-dimensional spaces will always have matrix representations. To see why, the concept of "coordinates" in higher dimensions must first be understood. Recall that if B = {u1 , u2 , ..., un } is a basis for a vector space U, then each v ∈ U can be written as v = α1 u1 + α2 u2 + ··· + αn un . The αi 's in this expansion are uniquely determined by v because if v = Σ αi ui = Σ βi ui , then 0 = Σ (αi − βi) ui , and this implies αi − βi = 0 (i.e., αi = βi) for each i because B is an independent set. Coordinates of a Vector Let B = {u1 , u2 , ..., un } be a basis for a vector space U, and let v ∈ U. The coefficients αi in the expansion v = α1 u1 + α2 u2 + ··· + αn un are called the coordinates of v with respect to B, and, from now on, [v]B will denote the column vector ... Caution! Order is important. If B' is a permutation of B, then [v]B' is the corresponding permutation of [v]B . From now on, S = {e1 , e2 , ..., en } will denote the standard basis of unit vectors (in natural order) for n (or C n). If no other basis is explicitly mentioned, then the standard basis is assumed. For example, if no basis is mentioned, and if we write v = ..., then it is understood that this is the representation with respect to S in the sense that v = [v]S = 8e1 + 4e2 − 6e3 . The standard coordinates of a vector are its coordinates with respect to S. So, 8, 4, and −6 are the standard coordinates of v in the above example. Example 4.7.2 Problem: If v is a vector in 3 whose standard coordinates are ..., determine the coordinates of v with respect to the basis ... Solution: The object is to find the three unknowns α1 , α2 , and α3 such that α1 u1 + α2 u2 + α3 u3 = v. This is simply a 3 × 3 system of linear equations ... 4.7 Linear Transformations 241 The general rule for making a change of coordinates is given on p. 252. Linear transformations possess coordinates in the same way vectors do because linear transformations from U to V also form a vector space. Space of Linear Transformations • For each pair of vector spaces U and V over F, the set L(U, V) of all linear transformations from U to V is a vector space over F. • Let B = {u1 , u2 , ..., un } and B' = {v1 , v2 , ..., vm } be bases for U and V, respectively, and let Bji be the linear transformation from U into V defined by Bji (u) = ξj vi , where (ξ1 , ξ2 , ..., ξn)T = [u]B . That is, pick off the jth coordinate of u, and attach it to vi . = BL = {Bji }i=1...m j=1...n is a basis for L(U, V). = dim L(U, V) = (dim U) (dim V). Proof. L(U, V) is a vector space because the defining properties on p. 160 are satisfied—details are omitted. Prove BL is a basis by demonstrating that it is a linearly independent spanning set for L(U, V). To establish linear independence, suppose Σj,i ηji Bji = 0 for scalars ηji , and observe that for each uk ∈ B, Bji (uk) = vi if j = k, 0 if j ≠ k. ... and hence ηki = 0 for each i and k because B' is an independent set. To show BL spans L(U, V), suppose T ∈ L(U, V), and determine the action of T on any u ∈ U by using u = Σ ξj uj and T(uj) = Σ αij vi to write ... This holds for all u ∈ U, so T = Σ αij Bji , and thus BL spans L(U, V). It now makes sense to talk about the coordinates of T ∈ L(U, V) with respect to the basis BL . In fact, the rule for determining these coordinates is contained in the proof above, where it was demonstrated that T = Σj,i αij Bji in which the coordinates αij are precisely the scalars in T(uj) = Σi αij vi = α1j v1 + α2j v2 + ··· + αmj vm . This suggests that rather than listing all coordinates αij in a single column containing all the coordinates, it's more natural to arrange the αij 's as an m × n matrix in which the jth column contains the coordinates of T(uj) with respect to B'. These ideas are summarized below. Coordinate Matrix Representations Let B = {u1 , u2 , ..., un } and B' = {v1 , v2 , ..., vm } be bases for U and V, respectively. The coordinate matrix of T ∈ L(U, V) with respect to the pair (B, B') is defined to be the m × n matrix [T]BB' = ([T(u1)]B' | [T(u2)]B' | ··· | [T(un)]B'). (4.7.3) In other words, if T(uj) = α1j v1 + α2j v2 + ··· + αmj vm , then ... When T is a linear operator on U, and when there is only one basis involved, [T]B is used in place of [T]BB to denote the (necessarily square) coordinate matrix of T with respect to B. 4.7 Linear Transformations 243 Example 4.7.3 Problem: If P is the projector defined in Example 4.7.1 that maps each point v = (x, y, z) ∈ 3 to its orthogonal projection P(v) = (x, y, 0) in the xy-plane, determine the coordinate matrix [P]B with respect to the basis ... Solution: According to (4.7.3), the jth column in [P]B is [P(uj)]B . Therefore, ... so that ... Example 4.7.4 Problem: Consider the same problem given in Example 4.7.3, but use different bases—say, ... For the projector defined by P(x, y, z) = (x, y, 0), determine [P]BB' . Solution: Determine the coordinates of each P(uj) with respect to B', as shown below: ... Therefore, according to (4.7.3), ... At the heart of linear algebra is the realization that the theory of finite-dimensional linear transformations is essentially the same as the theory of matrices. This is due primarily to the fundamental fact that the action of a linear transformation T on a vector u is precisely matrix multiplication between the coordinates of T and the coordinates of u. 244 Chapter 4 Vector Spaces Action as Matrix Multiplication Let T ∈ L(U, V), and let B and B' be bases for U and V, respectively. For each u ∈ U, the action of T on u is given by matrix multiplication between their coordinates in the sense that [T(u)]B' = [T]BB' [u]B . (4.7.4) Proof. Let B = {u1 , u2 , ..., un } and B' = {v1 , v2 , ..., vm }. If u = Σ ξj uj and T(uj) = Σ αij vi , then ... so ... In other words, ... Example 4.7.5 Problem: Show how the action of the operator D(p(t)) = dp/dt on the space P3 of polynomials of degree three or less is given by matrix multiplication. Solution: The coordinate matrix of D with respect to the basis B = {1, t, t2 , t3 } is ... If p = p(t) = α0 + α1 t + α2 t2 + α3 t3 , then D(p) = α1 + 2α2 t + 3α3 t2 so that ... The action of D is accomplished by means of matrix multiplication because ... For T ∈ L(U, V) and L ∈ L(V, W), the composition of L with T is defined to be the function C : U → W such that C(x) = L(T(x)), and this composition, denoted by C = LT, is also a linear transformation because C(αx + y) = L(T(αx + y)) = L(αT(x) + T(y)) = αL(T(x)) + L(T(y)) = αC(x) + C(y). Consequently, if B, B', and B'' are bases for U, V, and W, respectively, then C must have a coordinate matrix representation with respect to (B, B''), so it's only natural to ask how [C]BB'' is related to [L]B'B'' and [T]BB' . Recall that the motivation behind the definition of matrix multiplication given on p. 93 was based on the need to represent the composition of two linear transformations, so it should be no surprise to discover that [C]BB'' = [L]B'B'' [T]BB' . This, along with the other properties given below, makes it clear that studying linear transformations on finite-dimensional spaces amounts to studying matrix algebra. 246 Chapter 4 Vector Spaces Connections with Matrix Algebra • If T, L ∈ L(U, V), and if B and B' are bases for U and V, then • [αT]BB' = α[T]BB' for scalars α, (4.7.5) • [T + L]BB' = [T]BB' + [L]BB' . (4.7.6) • If T ∈ L(U, V) and L ∈ L(V, W), and if B, B', and B'' are bases for U, V, and W, respectively, then LT ∈ L(U, W), and • [LT]BB'' = [L]B'B'' [T]BB' . (4.7.7) • If T ∈ L(U, U) is invertible in the sense that TT−1 = T−1 T = I for some T−1 ∈ L(U, U), then for every basis B of U, • [T−1]B = [T]B −1 . (4.7.8) Proof. The first three properties (4.7.5)–(4.7.7) follow directly from (4.7.4). For example, to prove (4.7.7), let u be any vector in U, and write ... This is true for all u ∈ U, so [LT]BB'' = [L]B'B'' [T]BB' (see Exercise 3.5.5). Proving (4.7.5) and (4.7.6) is similar—details are omitted. To prove (4.7.8), note that if dim U = n, then [I]B = In for all bases B, so property (4.7.7) implies In = [I]B = [TT−1]B = [T]B [T−1]B , and thus [T−1]B = [T]B −1 . Example 4.7.6 Problem: Form the composition C = LT of the two linear transformations T : 3 → 2 and L : 2 → 2 defined by T(x, y, z) = (x + y, y − z) and L(u, v) = (2u − v, u), and then verify (4.7.7) and (4.7.8) using the standard bases S2 and S3 for 2 and 3 , respectively. Solution: The composition C : 3 → 2 is the linear transformation C(x, y, z) = L(T(x, y, z)) = L(x + y, y − z) = (2x + y + z, x + y). The coordinate matrix representations of C, L, and T are ... Property (4.7.7) is verified because ... Find [L−1]S2 by looking for scalars βij in L−1 (1, 0) = β11 (1, 0) + β21 (0, 1) ... so that L−1 = ... 4.7 Linear Transformations 247 so that ... Since ... L−1 (u, v) = (v, 2v − u) ... Exercises for section 4.7 4.7.1. Determine which of the following functions are linear operators on 2 . (a) T(x, y) = (x, 1 + y), (b) T(x, y) = (y, x), (c) T(x, y) = (0, xy), (d) T(x, y) = (x2 , y2), (e) T(x, y) = (x, sin y), (f) T(x, y) = (x + y, x − y). 4.7.2. For A ∈ n×n , determine which of the following functions are linear transformations. (a) T(Xn×n) = AX − XA, (b) T(xn×1) = Ax + b for b ≠ 0, (c) T(A) = AT , (d) T(Xn×n) = (X + XT)/2. 4.7.3. Explain why T(0) = 0 for every linear transformation T. 4.7.4. Determine which of the following mappings are linear operators on Pn , the vector space of polynomials of degree n or less. (a) T = ξ Dk , the kth-order differentiation operator (i.e., Dk p(t) = dk p/dtk). (b) T = an−1 Dn−1 + ··· + a1 D + a0 I, where the ak 's are constants. (c) T(p(t)) = tn p'(0) + t. 4.7.5. Let v be a fixed vector in n×1 and let T : n×1 → n be the mapping defined by T(x) = vT x (i.e., the standard inner product). (a) Is T a linear operator? (b) Is T a linear transformation? 4.7.6. For the operator T : 2 → 2 defined by T(x, y) = (x + y, −2x + 4y), determine [T]B , where B is the basis B = {(1, 1), (1, 2)}. 4.7.7. Let T : 2 → 3 be the linear transformation defined by T(x, y) = (x + 3y, 0, 2x − 4y). (a) Determine [T]SS' , where S and S' are the standard bases for 2 and 3 , respectively. (b) Determine [T]SS'' , where S'' is the basis for 3 obtained by permuting the standard basis according to S'' = {e3 , e2 , e1 }. 4.7.8. Let T be the operator on 3 defined by T(x, y, z) = (x − y, y − x, x − z) and consider the vector v = ... and the basis ... (a) Determine [T]B and [v]B . (b) Compute [T(v)]B , and then verify that [T]B [v]B = [T(v)]B . 248 Chapter 4 Vector Spaces 4.7.9. For A ∈ n×n , let T be the linear operator on n×1 defined by T(x) = Ax. That is, T is the operator defined by matrix multiplication. With respect to the standard basis S, show that [T]S = A. 4.7.10. If T is a linear operator on a space V with basis B, explain why [Tk]B = [T]Bk for all nonnegative integers k. 4.7.11. Let P be the projector that maps each point v ∈ 2 to its orthogonal projection on the line y = x as depicted in Figure 4.7.4. Figure 4.7.4 (a) Determine the coordinate matrix of P with respect to the standard basis. (b) Determine the orthogonal projection of v = (α, β) onto the line y = x. 4.7 Linear Transformations 249 4.7.12. For the standard basis S = {...} of 2×2 , determine the matrix representation [T]S for each of the following linear operators on 2×2 , and then verify [T(U)]S = [T]S [U]S for U = ... (a) T(X2×2) = (X + XT)/2. (b) T(X2×2) = AX − XA, where A = ... 4.7.13. For P2 and P3 (the spaces of polynomials of degrees less than or equal to two and three, respectively), let S : P2 → P3 be the linear transformation defined by S(p) = ∫0t p(x)dx. Determine [S]BB' , where B = {1, t, t2 } and B' = {1, t, t2 , t3 }. 4.7.14. Let Q be the linear operator on 2 that rotates each point counterclockwise through an angle θ, and let R be the linear operator on 2 that reflects each point about the x-axis. (a) Determine the matrix of the composition [RQ]S relative to the standard basis S. (b) Relative to the standard basis, determine the matrix of the linear operator that rotates each point in 2 counterclockwise through an angle 2θ. 4.7.15. Let P : U → V and Q : U → V be two linear transformations, and let B and B' be arbitrary bases for U and V, respectively. (a) Provide the details to explain why [P+Q]BB' = [P]BB' + [Q]BB' . (b) Provide the details to explain why [αP]BB' = α[P]BB' , where α is an arbitrary scalar. 4.7.16. Let I be the identity operator on an n-dimensional space V. (a) Explain why ... regardless of the choice of basis B. (b) Let B = {xi }ni=1 and B' = {yi }ni=1 be two different bases for V, and let T be the linear operator on V that maps vectors from B' to vectors in B according to the rule T(yi) = xi for i = 1, 2, ..., n. Explain why [I]BB' = [T]B = [T]B' . (c) When V = 3 , determine [I]BB' for ... 250 Chapter 4 Vector Spaces 4.7.17. Let T be the linear operator on 2 defined by T(x, y) = (2x − y, −x + 2y) ... 4.7.18. Let T be the linear operator on 3 defined by T(x, y, z) = (2x − y, −x + 2y − z, z − y). (a) Determine T−1 (x, y, z). (b) Determine [T−1]S , where S is the standard basis for 3 . 4.7.19. Let T be a linear operator on an n-dimensional space V. Show that the following statements are equivalent. (1) T−1 exists. (2) T is a one-to-one mapping (i.e., T(x) = T(y) =⇒ x = y). (3) N (T) = {0}. (4) T is an onto mapping (i.e., for each v ∈ V, there is an x ∈ V such that T(x) = v). Hint: Show that (1) =⇒ (2) =⇒ (3) =⇒ (4) =⇒ (2), and then show (2) and (4) =⇒ (1). 4.7.20. Let V be an n-dimensional space with a basis B = {ui }ni=1 . (a) Prove that a set of vectors {x1 , x2 , ..., xr } ⊆ V is linearly independent if and only if the set of coordinate vectors ... (b) If T is a linear operator on V, then the range of T is the set R (T) = {T(x) | x ∈ V}. Suppose that the basic columns of [T]B occur in positions b1 , b2 , ..., br . Explain why {T(ub1), T(ub2), ..., T(ubr)} is a basis for R (T). 4.8 Change of Basis and Similarity 4.8 251 CHANGE OF BASIS and SIMILARITY By their nature, coordinate matrix representations are basis dependent. However, it's desirable to study linear transformations without reference to particular bases because some bases may force a coordinate matrix representation to exhibit special properties that are not present in the coordinate matrix relative to other bases. To divorce the study from the choice of bases it's necessary to somehow identify properties of coordinate matrices that are invariant among all bases—these are properties intrinsic to the transformation itself, and they are the ones on which to focus. The purpose of this section is to learn how to sort out these basis-independent properties. The discussion is limited to a single finite-dimensional space V and to linear operators on V. Begin by examining how the coordinates of v ∈ V change as the basis for V changes. Consider two different bases B = {x1 , x2 , ..., xn } and B' = {y1 , y2 , ..., yn } for V. It's an elementary fact that the coordinate matrix representation of the identity operator I on V with respect to B is the identity matrix, but if I is the identity operator and B, B' are different bases, then [I]BB' is not the identity matrix. The matrix [I]BB' is called the change of basis matrix ... For example, if [v]B = ..., then [v]B' = [I]BB' [v]B . ... Change of Basis Let B = {x1 , x2 , ..., xn } and B' = {y1 , y2 , ..., yn } be bases for V, and let I be the identity operator on V. • If v ∈ V, then [v]B = P[v]B' , where P = [I]B'B ... is nonsingular. (4.8.1) • If T is an operator on V, then [T]B' = P−1 [T]B P. (4.8.2) P is called the change of basis matrix from B' to B. ... Solution: ... Example 4.8.1 Problem: For the space P2 of polynomials of degree 2 or less, determine the change of basis matrix P from B to B', where B = {1, t, t2 } and B' = {1, 1 + t, 1 + t + t2 }, and then find the coordinates of q(t) = 3 + 2t + 4t2 with respect to B'. Solution: ... To independently check that these coordinates are correct, simply verify that q(t) = 1(1) + (−2)(1 + t) + 4(1 + t + t2). It's now rather easy to describe how the coordinate matrix of a linear operator changes as the underlying basis changes. Changing Matrix Coordinates Let A be a linear operator on V, and let B and B' be two bases for V. The coordinate matrices [A]B and [A]B' are related as follows. [A]B = P−1 [A]B' P, where P = [I]BB' . (4.8.3) That is, the different coordinate matrix representations of A are similar matrices. Proof. Let B = {x1 , x2 , ..., xn } and B' = {y1 , y2 , ..., yn }, and observe that for each j, ... 254 Chapter 4 Vector Spaces In other words, [A]B = P−1 [A]B' P ... Since ... Example 4.8.2 Problem: Consider the linear operator A(x, y) = (y, −2x + 3y) on 2 along with the two bases S = {(1, 0), (0, 1)} and S' = {(1, 1), (1, 2)}. First compute the coordinate matrix [A]S as well as the change of basis matrix Q from S' to S, and then use these two matrices to determine [A]S' . Solution: The matrix of A relative to S is obtained by computing A(e1) = A(1, 0) = (0, −2) = (0)e1 + (−2)e2 , A(e2) = A(0, 1) = (1, 3) = (1)e1 + (3)e2 , so that [A]S = ... According to (4.8.3), [A]S' = Q−1 [A]S Q ... Notice that [A]S' is a diagonal matrix, whereas [A]S is not. This shows that the representation of A relative to S' is more revealing. 4.8 Change of Basis and Similarity 255 the one whose coordinate matrix is as simple as possible is the most revealing one. ... Similarity • Matrices Bn×n and Cn×n are said to be similar matrices whenever there exists a nonsingular matrix Q such that B = Q−1 CQ. We write B ≃ C to denote that B and C are similar. • The linear operator f : n×n → n×n defined by f (C) = Q−1 CQ is called a similarity transformation. 256 Chapter 4 Vector Spaces ... Invariant Properties ... Example 4.8.3 Problem: Show that the trace of a square matrix is a similarity invariant ... Solution: ... trace (AB) = trace (BA) ... 4.8.1. Explain why rank is a similarity invariant. 4.8.2. Explain why similarity is transitive ... 4.8.3. ... Exercises for section 4.8 ... 4.8.10. Suppose that B = {x1 , x2 , ..., xn } and B' = {y1 , y2 , ..., yn } are bases for an n-dimensional subspace V ⊆ m×1 , and let Xm×n and Ym×n be the matrices whose columns are the vectors from B and B', respectively. ... 4.8.11. ... nilpotent ... 4.8.12. ... 4.9 INVARIANT SUBSPACES For a linear operator T on a vector space V, and for X ⊆ V, T(X) = {T(x) | x ∈ X} is the set of all possible images of vectors from X under the transformation T. Notice that T(V) = R (T). When X is a subspace of V, it follows that T(X) is also a subspace of V, but T(X) is usually not related to X. However, in some special cases it can happen that T(X) ⊆ X, and such subspaces are the focus of this section. Invariant Subspaces • For a linear operator T on V, a subspace X ⊆ V is said to be an invariant subspace under T whenever T(X) ⊆ X. • In such a situation, T can be considered as a linear operator on X by forgetting about everything else in V and restricting T to act only on vectors from X. Hereafter, this restricted operator will be denoted by T/X . Example 4.9.1 Problem: For ... show that the subspace X spanned by B = {x1 , x2 } is an invariant subspace under A. Then describe the restriction A/X and determine the coordinate matrix of A/X relative to B. Solution: Observe that Ax1 = 2x1 ∈ X and Ax2 = x1 + 2x2 ∈ X, so the image of any x = αx1 + βx2 ∈ X is back in X because Ax = A(αx1 + βx2) = αAx1 + βAx2 = 2αx1 + β(x1 + 2x2) = (2α + β)x1 + 2βx2 . This equation completely describes the action of A restricted to X, so A/X (x) = (2α + β)x1 + 2βx2 for each x = αx1 + βx2 ∈ X. Since A/X (x1) = 2x1 and A/X (x2) = x1 + 2x2 , we have ... 260 Chapter 4 Vector Spaces The invariant subspaces for a linear operator T are important because they produce simplified coordinate matrix representations of T. To understand how this occurs, suppose X is an invariant subspace under T, and let BX = {x1 , x2 , ..., xr } be a basis for X that is part of a basis B = {x1 , ..., xr , y1 , ..., yq } for the entire space V. To compute [T]B , recall from the definition of coordinate matrices that [T]B = ([T(x1)]B | ··· | [T(xr)]B | [T(y1)]B | ··· | [T(yq)]B). (4.9.1) Because each T(xj) is contained in X, only the first r vectors from B are needed to represent each T(xj), so, for j = 1, 2, ..., r, T(xj) = Σ αij xi and [T(xj)]B = ... (4.9.2) The space Y = span {y1 , ..., yq } may not be an invariant subspace for T, so all the basis vectors in B may be needed to represent the T(yj)'s. ... Using (4.9.2) and (4.9.3) in (4.9.1) produces the block-triangular matrix ... The equations T(xj) = Σ αij xi for j = 1, 2, ..., r also mean that [T/X]BX = ..., so the matrix in (4.9.5) can be written as ... In other words, (4.9.6) says that the matrix representation for T can be made to be block triangular whenever a basis for an invariant subspace is available. The more invariant subspaces we can find, the more tools we have to construct simplified matrix representations. For example, if the space Y in (4.9.3) is also an invariant subspace for T, then T(yj) ∈ Y for each j = 1, 2, . . . , q, and only the yi 's are needed to represent T(yj) in (4.9.3). Consequently, the βij 's are all zero, and [T]B has the block-diagonal form ⎛ 0 T/ Y Bx 0 Ar×r * ⎞ [T]B = ⎝ 0 Cq×q 0 T/ Y By This notion easily generalizes in the sense that if B = BX ∪BY ∪···∪BZ is a basis for V, where BX , BY , . . . , BZ are bases for invariant subspaces under T that have dimensions r1 , r2 , . . . , rk , respectively, then [T]B has the block-diagonal form (A 0 ··· 0 \r1 ×r1 Bz2 ×r2 ··· 0 [T]B = ⎜ . . . . ⎟ , ⎝ . ⎠ . . . 0 0 ··· Crk ×rk , * ° C= T/ Z Bz The situations discussed above are also reversible in the sense that if the matrix representation of T has a block-triangular form Ar×r Br×q [T]B = 0 Cq×q relative to some basis B = {u1 , u2 , . . . , ur , w1 , w2 , . . . , wq }, 262 Chapter 4 Vector Spaces then the r-dimensional subspace U = span {u1 , u2 , . . . , ur } spanned by the first r vectors in B must be an invariant subspace under T. Furthermore, if the matrix representation of T has a block-diagonal form [T]B = Ar×r 0 0 Cq×q relative to B, then both U = span {u1 , u2 , . . . , ur } and W = span {w1 , w2 , . . . , wq } must be invariant subspaces for T. The details are left as exercises. The general statement concerning invariant subspaces and coordinate matrix representations is given below. Invariant Subspaces and Matrix Representations Let T be a linear operator on an n-dimensional space V, and let X , Y, . . . , Z be subspaces of V with respective dimensions r1 , r2 , . . . , rk and bases BX , BY , . . . , BZ . Furthermore, suppose that i ri = n and B = BX ∪ BY ∪···∪ BZ is a basis for V. • The subspace X is an invariant subspace under T if and only if [T]B has the block-triangular form [T]B = Ar×r B 0 C, in which case A = T/ X BX (4.9.7) The hypotheses X , Y, . . . , Z are invariant subspaces under T if and only if [T]B has the block-diagonal form .... in which case ... . (4.9.8) ··· Ck ×rk BZ An important corollary concerns the special case in which the linear operator T is for ect an n × n matrix and T(v) = Av is a matrix–vector multiplication. 4.9 Invariant Subspaces 263 Triangular and Block Diagonal Forms When T is an n × n matrix, the following two statements are true. • Q is a nonsingular matrix such that Q−1 TQ = Br×q C (4.9.9) if and only if the first r columns in Q span an invariant subspace under T. • Q is a nonsingular matrix such that Q−1 TQ = Br×r ... (4.9.10) if and only if Q = Q1 | Q2 | · · · | Qk in which Qi is n × ri , and the columns of each Qi span an invariant subspace under T. Proof. We know from that if B = {q1 , q2 , . . . , qn } is a basis for Example 4.8.3 n , and if Q = q1 q2 ··· qn is the matrix containing the vectors from B as its columns, then [T]B = Q−1 TQ. Statements (4.9.9) and (4.9.10) are now direct consequences of statements (4.9.7) and (4.9.8), respectively. Example 4.9.2 Problem: For ... 3 Solution: The first two columns of Q span a space that is invariant under T, and this allows us to conclude that T = Q−1 TQ must be in block-triangular form. This is easy to verify by computing ... Notice that the diagonal blocks must be the matrices of the restrictions in the sense that ... Example 4.9.4 Problem: Find all subspaces of 2 that are invariant under 0 1 A= −2 3 Solution: The trivial subspace {0} is the only zero-dimensional invariant subspace, and the entire space 2 is the only two-dimensional invariant subspace. The real problem is to find all one-dimensional invariant subspaces. If M is a one-dimensional subspace spanned by x = 0 such that A(M) ⊆ M, then Ax ∈ M =⇒ there is a scalar λ such that Ax = λx =⇒ (A − λI) x = 0. In other words, M ⊆ N (A − λI) . Since x = 0, it follows that N (A − λI) = 0 and ... Consequently, λ must be a scalar such that (A − λI) is a singular matrix. Row operations produce ... and it is clear that (A − λI) is singular if and only if ... The two real-valued scalars λ = 1 and λ = 2, and straightforward computation of the associated one-dimensional invariant subspaces M1 = N (A − I) = span ... and M2 = N (A − 2I) = span ... In general, scalars λ for which (A − λI) is singular are called the eigenvalues of A, and the nonzero vectors in N (A − λI) are known as the associated eigenvectors for A. As this example indicates, eigenvalues and eigenvectors are of fundamental importance in identifying invariant subspaces and reducing matrices by means of similarity transformations. Eigenvalues and eigenvectors are discussed at length in Chapter 7. Exercises for section 4.9 4.9.1. Let T be an arbitrary linear operator on a vector space V. (a) Is the trivial subspace {0} invariant under T? (b) Is the entire space V invariant under T? 4.9.2. Describe all of the invariant subspaces for the identity operator I on a space V. 4.9.3. Let T be the linear operator on 4 defined by T(x1, x2, x3, x4) = (x1 + x2 + 2x3 − x4 , x2 + x4 , 2x3 − x4 , x3 + x4), and let X = span {e1 , e2 } be the subspace that is spanned by the first two unit vectors in 4 . (a) Explain why X is invariant under T. (b) Determine T/ X {e , e } . (c) Describe the structure of [T]B , where B is any basis obtained from an extension of {e1 , e2 } . 4.9.4. Let T and Q be the matrices ... (a) Explain why the columns of Q are a basis for 4 . (b) Verify that X = span {Q∗1 , Q∗2 } and Y = span {Q∗3 , Q∗4 } are each invariant subspaces under T. (c) Describe the structure of Q−1 TQ without doing any computation. (d) Now compute the product Q−1 TQ to determine ... 4.9.5. Let T be a linear operator on a space V, and suppose that B = {u1 , . . . , ur , w1 , . . . , wq } is a basis for V such that [T]B has the block-diagonal form ... Explain why U = span {u1 , . . . , ur } and W = span {w1 , . . . , wq } must each be invariant subspaces under T. 4.9.6. If Tn×n and Pn×n are matrices such that ... 4.9.7. If A is an n × n matrix and λ is a scalar such that (A − λI) is singular (i.e., λ is an eigenvalue), explain why the associated space of eigenvectors N (A − λI) is an invariant subspace under A. 4.9.8. Consider the matrix A = ... (a) Determine the eigenvalues of A. (b) Identify all of the invariant subspaces under A by finding all eigenvectors. CHAPTER 5 Norms, Inner Products, and Orthogonality VECTOR NORMS 5.1 A significant portion of linear algebra is in fact geometric in nature because much of the subject grew out of the need to generalize the basic geometry of 2 and 3 to nonvisual higher-dimensional spaces. The usual approach is to coordinatize geometrical concepts in 2 and 3 , and then extend statements concerning ordered pairs and triples to ordered n-tuples in n and C n . For example, the length of a vector u ∈ 2 or v ∈ 3 is obtained from the Pythagorean theorem by computing the length of the hypotenuse of a right triangle as shown in Figure 5.1.1. v = (x,y,z) ||v|| = x2 + y2 + z2 ... This measure of length, u = x2 + y2 and v = x2 + y2 + z2 , is called the euclidean norm in 2 and 3 , and there is an obvious extension to higher dimensions. Euclidean Vector Norm For a vector x n×1 , the euclidean norm of x is defined to be ... whenever x ∈ C n×1 . For example, ... The definition of euclidean norm guarantees that for all scalars α, αx = |α| x . (5.1.1) Given a vector x = 0, it's frequently convenient to have another vector that points in the same direction as x (i.e., is a positive multiple of x) but has unit length. To construct such a vector, we normalize x by setting u = x/ x . From (5.1.1), it's easy to see that u = x / x = 1. (5.1.2) By convention, column vectors are used throughout this chapter. But there is nothing special about columns because, with the appropriate interpretation, all statements concerning columns will also hold for rows. 5.1 Vector Norms 271 • The distance between vectors in 3 can be visualized with the aid of the parallelogram law as shown in Figure 5.1.2, so for vectors in n and C n , the distance between u and v is naturally defined to be u − v . u ||u - v|| v Figure 5.1.2 Standard Inner Product The scalar terms defined by ... are called the standard inner products for n and C n , respectively. 34 The Cauchy–Bunyakovskii–Schwarz (CBS) inequality is one of the most important inequalities in mathematics. It relates inner product to norm. 34 The Cauchy–Bunyakovskii–Schwarz inequality is named in honor of the three men who played a role in its development. The basic inequality for real numbers is attributed to Cauchy in 1821, whereas Schwarz and Bunyakovskii contributed by later formulating useful generalizations of the inequality involving integrals of functions. Augustin-Louis Cauchy (1789–1857) was a French mathematician who is generally regarded as being the founder of mathematical analysis—including the theory of complex functions. Although deeply embroiled in political turmoil for much of his life (he was a partisan of the Bourbons), Cauchy emerged as one of the most prolific mathematicians of all time. He authored at least 789 mathematical papers, and his collected works fill 27 volumes—this is on a par with Cayley and second only to Euler. It is said that more theorems, concepts, and methods bear Cauchy's name than any other mathematician. Victor Bunyakovskii (1804–1889) was a Russian professor of mathematics at St. Petersburg, and in 1859 he extended Cauchy's inequality for discrete sums to integrals of continuous functions. His contribution was overlooked by western mathematicians for many years, and his name is often omitted in classical texts that simply refer to the Cauchy–Schwarz inequality. Hermann Amandus Schwarz (1843–1921) was a student and successor of the famous German mathematician Karl Weierstrass at the University of Berlin. Schwarz independently generalized Cauchy's inequality just as Bunyakovskii had done earlier. 272 Chapter 5 Norms, Inner Products, and Orthogonality Cauchy–Bunyakovskii–Schwarz (CBS) Inequality |x∗ y| ≤ x y for all x, y ∈ C n×1 . (5.1.3) Equality holds if and only if y = αx for α = x∗ y/x∗ x. Proof. Set α = x∗ y/x∗ x = x∗ y/ x 2 (assume x = 0 because there is nothing to prove if x = 0) and observe that x∗ (αx − y) = 0, so 0 ≤ αx − y 2 = (αx − y)∗ (αx − y) = ... (5.1.4) Since y∗ x = x∗ y, it follows that ... and (5.1.3) follows. Equality holds ... Now, 0 < x implies that 0 ≤ αx − y 2 = ... The CBS inequality is one of the most important inequalities in mathematics ... Triangle Inequality x + y ≤ x + y for every x, y ∈ C n . (5.1.5) ... Backward Triangle Inequality ... (5.1.6) ... 274 Chapter 5 Norms, Inner Products, and Orthogonality There are notions of length other than the euclidean measure. For example, urban dwellers navigate on a grid of city blocks with one-way streets, so they are prone to measure distances in the city not as the crow flies but rather in terms of lengths on a directed grid. ... p-Norms For p ≥ 1, the p-norm of x ∈ C n is defined as xp = (i=1 |xi |p)1/p . It can be proven that the following properties are valid for all p-norms ... Generalized Triangle Inequality ... (5.1.8) ... 5.1 Vector Norms 275 To see that limp→∞ xp = maxi |xi | , proceed as follows. ... 1 Example 5.1.2 To get a feel for the 1-, 2-, and ∞-norms, it helps to know the shapes and relative sizes of the unit p-spheres ... Figure 5.1.4 ... Vector Norms A norm for a real or complex vector space V is a function mapping V into that satisfies the following conditions. x ≥ 0 and x = 0 ⇐⇒ x = 0, αx = |α| x for all scalars α, x + y ≤ x + y . (5.1.9) ... Equivalent Norms ... (5.1.10) ... Exercises for section 5.1 5.1.1. Find the 1-, 2-, and ∞-norms of x = ... 5.1.2. Consider the euclidean norm with u = ... 5.1.3. ... 5.1.4. ... 5.1.5. ... 5.1.6. ... 5.1.7. ... 5.1.8. ... 5.1.9. ... 5.1.10. ... 5.1.11. ... 5.1.12. ... 5.1.13. ... 5.2 Matrix Norms Because C m×n is a vector space of dimension mn, magnitudes of matrices A ∈ C m×n can be "measured" by employing any vector norm on C mn . ... Frobenius Matrix Norm The Frobenius norm of A ∈ C m×n is defined by the equations ... (5.2.1) ... Induced Matrix Norms A vector norm that is defined on C p for p = m, n induces a matrix norm on C m×n by setting A = max Ax ... (5.2.4) ... Matrix 2-Norm The matrix norm induced by the euclidean vector norm is A 2 = max Ax 2 = λmax , (5.2.7) ... Properties of the 2-Norm ... (5.2.8)–(5.2.12) ... Matrix 1-Norm and Matrix ∞-Norm The matrix norms induced by the vector 1-norm and ∞-norm are as follows. A 1 = max x 1 =1 Ax 1 = max j i |aij | = the largest absolute column sum. A ∞ = max x ∞ =1 Ax ∞ = max i j |aij | = the largest absolute row sum. (5.2.15) (5.2.16) ... Exercises for section 5.2 5.2.1. ... 5.2.10. ... 5.3 Inner-Product Spaces The euclidean norm, which naturally came first, is a coordinate-dependent function. But by isolating its important properties we quickly moved to the more general coordinate-free definition of a vector norm given in (5.1.9) on p. 275. ... Inner Product An inner product on a real (or complex) vector space V is a function that maps each ordered pair of vectors x, y to a real (or complex) scalar x y such that the following four properties hold. x x is real with x x ≥ 0, and x x = 0 if and only if x = 0, x αy = α x y for all scalars α, (5.3.1) x y + z = x y + x z , x y = y x (for real spaces, this becomes x y = y x). ... Example 5.3.1 ... (5.3.2) ... (5.3.3) ... Parallelogram Identity For a given norm on a vector space V, there exists an inner product on V such that = if and only if the parallelogram identity x + y 2 + x − y 2 = 2 x 2 + y 2 (5.3.7) holds for all x, y ∈ V. ... Example 5.3.3 ... (5.3.9) ... 5.3 Inner-Product Spaces 291 ... Example 5.3.4 ... (5.3.10) ... norm in (5.3.5). Virtually all important statements concerning n or C n with the standard inner product remain valid for general inner-product spaces—e.g., consider the statement and proof of the general CBS inequality. Advanced or more theoretical texts prefer a development in terms of general inner-product spaces. However, the focus of this text is matrices and the coordinate spaces n and C n , so subsequent discussions will usually be phrased in terms of n or C n and their standard inner products. But remember that extensions to more general inner-product spaces are always lurking in the background, and we will not hesitate to use these generalities or general inner-product notation when they serve our purpose. Exercises for section 5.3 5.3.1. For x = x1 x2 x3 , y = y1 y2 y3 , determine which of the following are inner products for 3×1 . (a) x y = x1 y1 + x3 y3 , (b) x y = x1 y1 − x2 y2 + x3 y3 , (c) x y = 2x1 y1 + x2 y2 + 4x3 y3 , (d) x y = x21 y12 + x22 y22 + x23 y32 . 5.3.2. For a general inner-product space V, explain why each of the following statements must be true. (a) If x y = 0 for all x ∈ V, then y = 0. (b) αx y = α x y for all x, y ∈ V and for all scalars α. (c) x + y z = x z + y z for all x, y, z ∈ V. 5.3.3. Let V be an inner-product space with an inner product x y . Explain why the function defined by = satisfies the first two norm properties in (5.2.3) on p. 280. 2 5.3.4. For a real inner-product space with = , derive the inequality 2 x y ≤ 2 x + y . 2 Hint: Consider x − y. 5.3 Inner-Product Spaces 293 5.3.5. For n × n matrices A and B, explain why each of the following inequalities is valid. (a) |trace (B)| ≤ n [trace (B∗ B)] . (b) trace B2 ≤ trace BT B for real matrices. T trace AT A + trace BT B (c) trace A B ≤ for real matrices. 2 2 5.3.6. Extend the proof given on p. 290 concerning the parallelogram identity (5.3.7) to include complex spaces. Hint: If V is a complex space with a norm that satisfies the parallelogram identity, let 2 x yr = 2 x + y − x − y , 4 and prove that x y = x yr + i x iy r (the polarization identity) (5.3.10) is an inner product on V. n 5.3.7. Explain why there does not exist an inner product on C (n ≥ 2) such that ∞ = . 5.3.8. Explain why the Frobenius matrix norm on C n×n must satisfy the parallelogram identity. 5.3.9. For n ≥ 2, is either the matrix 1-, 2-, or ∞-norm generated by an inner product on C n×n ? 294 5.4 Chapter 5 Norms, Inner Products, and Orthogonality ORTHOGONAL VECTORS Two vectors in 3 are orthogonal (perpendicular) if the angle between them is a right angle (90◦). But the visual concept of a right angle is not at our disposal in higher dimensions, so we must dig a little deeper. The essence of perpendicularity in 2 and 3 is embodied in the classical Pythagorean theorem, v || u - v || u || v || || u || 2 2 2 which says that u and v are orthogonal if and only if u + v = u − v . 39 2 But u = uT u for all u ∈ 3 , and uT v = vT u, so we can rewrite the Pythagorean statement as 2 2 2 T 0 = u + v − u − v = uT u + vT v − (u − v) (u − v) = uT u + vT v − uT u − uT v − vT u + vT v = 2uT v. Therefore, u and v are orthogonal vectors in 3 if and only if uT v = 0. The natural extension of this provides us with a definition in more general spaces. Orthogonality In an inner-product space V, two vectors x, y ∈ V are said to be orthogonal (to each other) whenever x y = 0, and this is denoted by writing x ⊥ y. Example 5.4.1 For n with the standard inner product, x ⊥ y ⇐⇒ xT y = 0. • For C n with the standard inner product, x ⊥ y ⇐⇒ x∗ y = 0. ⎛ x= 39 • ⎞1 ⎝ −2 ⎠3 −1 ⎛ is orthogonal to y = ⎞4 ⎝ 1⎠−2 −4 because xT y = 0. Throughout this section, only norms generated by an underlying inner product 2 = are used, so distinguishing subscripts on the norm notation can be omitted. 5.4 Orthogonal Vectors 295 In spite of the fact that uT v = 0, the vectors u = i 3 1 not orthogonal because u∗ v = 0. i and v = 0 1 are Now that "right angles" in higher dimensions make sense, how can more general angles be defined? Proceed just as before, but use the law of cosines rather than the Pythagorean theorem. Recall that v -v || || v || || u u θ || u|| 2 2 2 the law of cosines in 2 or 3 says u − v = u +v −2 u v cos θ. If u and v are orthogonal, then this reduces to the Pythagorean theorem. But, in general, 2 cos θ = 2 2 T u + v − u − v uT u + vT v − (u − v) (u − v) = 2 u v 2 u v 2uT v uT v = . 2 u v u v This easily extends to higher dimensions because if x, y are vectors from any real inner-product space, then the general CBS inequality (5.3.4) on p. 287 guarantees that x y / x y is a number in the interval [−1, 1], and hence there is a unique value θ in [0, π] such that cos θ = x y / x y. Angles In a real inner-product space V, the radian measure of the angle between nonzero vectors x, y ∈ V is defined to be the number θ ∈ [0, π] such that x y cos θ = . (5.4.1) x y 296 Chapter 5 Norms, Inner Products, and Orthogonality Example 5.4.2 n T In ⎛ , cos y. For example, to determine the angle between ⎞θ = x y/ x ⎛ −4 1 2 2 x = ⎝ 21 ⎠and y = ⎝ 02 ⎠, compute cos θ = 2/(5)(3) = 2/15, and use the inverse cosine function to conclude that θ = 1.437 radians (rounded). Example 5.4.3 Linear Correlation. Suppose that an experiment is conducted, and the resulting observations are recorded in two data vectors ⎛ ⎞⎛ ⎞y1 x1 ⎜ x2 ⎟ ⎜ y2 ⎟ ⎜ x= ⎝ ... ⎠, y = ⎝ .. ⎠, . xn yn ⎛ ⎞1 ⎜ 1 ⎟ and let e = ⎝ ... ⎠. 1 Problem: Determine to what extent the yi 's are linearly related to the xi 's. That is, measure how close y is to being a linear combination β0 e + β1 x. Solution: The cosine as defined in (5.4.1) does the job. To understand how, let μx and σx be the mean and standard deviation of the data in x. That is, 2 x − μx e2 eT x i xi i (xi − μx) √ μx = = and σx = = . n n n n The mean is a measure of central tendency, and the standard deviation measures the extent to which the data is spread. Frequently, raw data from different sources is difficult to compare because the units of measure are different—e.g., one researcher may use the metric system while another uses American units. To compensate, data is almost always first "standardized" into unitless quantities. The standardization of a vector x for which σx = 0 is defined to be zx = x − μx e . σx Entries in zx are often referred to as standard scores or z-scores. All stan√ dardized vectors have the properties that z = n, μz = 0, and σz = 1. Furthermore, it's not difficult to verify that for vectors x and y such that σx = 0 and σy = 0, it's the case that zx = zy ⇐⇒ ∃ constants β0 , β1 such that y = β0 e + β1 x, where β1 > 0, zx = −zy ⇐⇒ ∃ constants β0 , β1 such that y = β0 e + β1 x, where β1 < 0. • In other words, y = β0 e + β1 x for some β0 and β1 if and only if zx = ±zy , in which case we say y is perfectly linearly correlated with x. 5.4 Orthogonal Vectors 297 Since zx varies continuously with x, the existence of a "near" linear relationship between x and y is equivalent √ to zx being "close" to ±zy in some sense. The fact that zx = ±zy = n means zx and ±zy differ only in orientation, so a natural measure of how close zx is to ±zy is cos θ, where θ is the angle between zx and zy . The number T ρxy = cos θ = zx T zy (x − μx e) (y − μy e) zx T zy = = zx zy n x − μx e y − μy e is called the coefficient of linear correlation, and the following facts are now immediate. • ρxy = 0 if and only if x and y are orthogonal, in which case we say that x and y are completely uncorrelated. • |ρxy | = 1 if and only if y is perfectly correlated with x. That is, |ρxy | = 1 if and only if there exists a linear relationship y = β0 e + β1 x. • When β1 > 0, we say that y is positively correlated with x. When β1 < 0, we say that y is negatively correlated with x. |ρxy | measures the degree to which y is linearly related to x. In other words, |ρxy | ≈ 1 if and only if y ≈ β0 e + β1 x for some β0 and β1 . Positive correlation is measured by the degree to which ρxy ≈ 1. Negative correlation is measured by the degree to which ρxy ≈ −1. If the data in x and y are plotted in 2 as points (xi , yi), then, as depicted in Figure 5.4.1, ρxy ≈ 1 means that the points lie near a straight line with positive slope, while ρxy ≈ −1 means that the points lie near a line with negative slope, and ρxy ≈ 0 means that the points do not lie near a straight line. . . .. ρxy ≈ 1 Positive Correlation . . . . ρxy ≈ −1 Negative Correlation . . . . ρxy ≈ 0 No Correlation Figure 5.4.1 If |ρxy | ≈ 1, then the theory of least squares as presented in §4.6 can be used to determine a "best-fitting" straight line. 298 Chapter 5 Norms, Inner Products, and Orthogonality Orthonormal Sets B = {u1 , u2 , . . . , un } is called an orthonormal set whenever ui = 1 for each i, and ui ⊥ uj for all i = j. In other words, ui uj = 1 0 when i = j, when i = j. • Every orthonormal set is linearly independent. • Every orthonormal set of n vectors from an n-dimensional space V is an orthonormal basis for V. (5.4.2) Proof. The second point follows from the first. To prove the first statement, suppose B = {u1 , u2 , . . . , un } is orthonormal. If 0 = α1 u1 + α2 u2 + · · · + αn un , use the properties of an inner product to write 0 = ui 0 = ui α1 u1 + α2 u2 + · · · + αn un = α1 ui u1 + · · · + αi ui ui + · · · + αn ui un = αi ui = αi Example 5.4.4 2 for each i. 1 The set B = −1 0 u1 = 1 , u2 = 1 1 , u3 = −1 ! −1 2 is a set of mutually orthogonal vectors because uTi uj = 0 for i = j, but B is not an orthonormal set—each vector does not have unit length. However, it's easy to convert an orthogonal set (not containing a zero vector) set by √ simply √ into an orthonormal √ normalizing each vector. Since u = 2, u = 3, and u = 6, it 2 3 √ √ 1 √ follows that B = u1 / 2, u2 / 3, u3 / 6 is orthonormal. The most common orthonormal basis is S = {e1 , e2 , . . . , en } , the standard basis for n and C n , and, as illustrated below for 2 and 3 , these orthonormal vectors are directed along the standard coordinate axes. y z e2 e3 e1 e2 y x e1 x 5.4 Orthogonal Vectors 299 Another orthonormal basis B need not be directed in the same way as S, but that's the only significant difference because it's geometrically evident that B must amount to some rotation of S. Consequently, we should expect general orthonormal bases to provide essentially the same advantages as the standard basis. For example, an important function of the standard basis S for n is to provide coordinate representations by writing ⎛ ⎞x1 ⎜ x2 ⎟ ⎜ ⎟x = [x]S = ⎜ ⎟ ⎝ ... ⎠to mean x = x1 e1 + x2 e2 + · · · + xn en . xn With respect to a general basis B = {u1 , u2 , . . . , un } , the coordinates of x are the scalars ξi in the representation x = ξ1 u1 + ξ2 u2 + · · · + ξn un , and, as illustrated in Example 4.7.2, finding the ξi 's requires solving an n × n system, a nuisance we would like to avoid. But if B is an orthonormal basis, then the ξi 's are readily available because ui x = ui ξ1 u1 + ξ2 u2 + · · · + ξn un = n 40 2 expansion of x. j=1 ξj ui uj = ξi ui = ξi . This yields the Fourier Fourier Expansions If B = {u1 , u2 , . . . , un } is an orthonormal basis for an inner-product space V, then each x ∈ V can be expressed as x = u1 x u1 + u2 x u2 + · · · + un x un . (5.4.3) This is called the Fourier expansion of x. The scalars ξi = ui x are the coordinates of x with respect to B, and they are called the Fourier coefficients. Geometrically, the Fourier expansion resolves x into n mutually orthogonal vectors ui x ui , each of which represents the orthogonal projection of x onto the space (line) spanned by ui . (More is said in Example 5.13.1 on p. 431 and Exercise 5.13.11.) 40 Jean Baptiste Joseph Fourier (1768–1830) was a French mathematician and physicist who, while studying heat flow, developed expansions similar to (5.4.3). Fourier's work dealt with special infinite-dimensional inner-product spaces involving trigonometric functions as discussed in Example 5.4.6. Although they were apparently used earlier by Daniel Bernoulli (1700–1782) to solve problems concerned with vibrating strings, these orthogonal expansions became known as Fourier series, and they are now a fundamental tool in applied mathematics. Born the son of a tailor, Fourier was orphaned at the age of eight. Although he showed a great aptitude for mathematics at an early age, he was denied his dream of entering the French artillery because of his "low birth." Instead, he trained for the priesthood, but he never took his vows. However, his talents did not go unrecognized, and he later became a favorite of Napoleon. Fourier's work is now considered as marking an epoch in the history of both pure and applied mathematics. The next time you are in Paris, check out Fourier's plaque on the first level of the Eiffel Tower. 300 Chapter 5 Norms, Inner Products, and Orthogonality Example 5.4.5 Problem: Determine the Fourier expansion of x = −1 2 1 with respect to the standard inner product and the orthonormal basis given in Example 5.4.4 ⎛ ⎛ ⎞ ⎛ ⎞ ⎛ ⎞⎞ 1 1 −1 ⎬ ⎨ 1 1 1 B = u1 = √ ⎝ −1 ⎠, u2 = √ ⎝ 1 ⎠, u3 = √ ⎝ −1 ⎠. ⎭ ⎩ 2 3 6 0 1 2 Solution: The Fourier coefficients are −3 ξ1 = u1 x = √ , 2 so 2 ξ2 = u2 x = √ , 3 1 ξ3 = u3 x = √ , 6 ⎛ ⎞⎛ ⎞⎛ ⎞−3 2 −1 1 1 1 x = ξ1 u1 + ξ2 u2 + ξ3 u3 = ⎝ 3 ⎠+ ⎝ 2 ⎠+ ⎝ −1 ⎠. 2 3 6 0 2 2 You may find it instructive to sketch a picture of these vectors in 3 . Example 5.4.6 Fourier Series. Let V be the inner-product space of real-valued functions that are integrable on the interval (−π, π) and where the inner product and norm are given by f |g = π f (t)g(t)dt and −π f = 1/2 π f 2 (t)dt . −π It's straightforward to verify that the set of trigonometric functions B = {1, cos t, cos 2t, . . . , sin t, sin 2t, sin 3t, . . .} is a set of mutually orthogonal vectors, so normalizing each vector produces the orthonormal set ! 1 sin t sin 2t sin 3t cos t cos 2t B = √ , √ , √ ,..., √ , √ , √ ,... . π π π π π 2π Given an arbitrary f ∈ V, we construct its Fourier expansion ∞ ∞ cos kt sin kt 1 F (t) = α0 √ + αk √ + βk √ , π π 2π k=1 k=1 (5.4.4) 5.4 Orthogonal Vectors 301 where the Fourier coefficients are given by π 1 1 √ f =√ f (t)dt , 2π 2π −π π cos kt 1 √ αk = f =√ f (t) cos kt dt for k = 1, 2, 3, . . . , π π −π π sin kt 1 √ βk = f =√ f (t) sin kt dt for k = 1, 2, 3, ... π π −π α0 = Substituting these coefficients in (5.4.4) produces the infinite series ∞ F (t) = a0 (an cos nt + bn sin nt) , + 2 n=1 (5.4.5) where an = 1 π π f (t) cos nt dt and −π bn = 1 π π f (t) sin nt dt. (5.4.6) −π The series F (t) in (5.4.5) is called the Fourier series expansion for f (t), but, unlike the situation in finite-dimensional spaces, F (t) need not agree with the original function f (t). After all, F is periodic, so there is no hope of agreement when f is not periodic. However, the following statement is true. • If f (t) is a periodic function with period 2π that is sectionally continu41 ous on the interval (−π, π), then the Fourier series F (t) converges to f (t) at each t ∈ (−π, π), where f is continuous. If f is discontinuous at t0 but possesses left-hand and right-hand derivatives at t0 , then F (t0) converges to the average value F (t0) = + f (t− 0) + f (t0) , 2 + − where f (t− 0) and f (t0) denote the one-sided limits f (t0) = limt→t− f (t) f (t). and f (t+ 0) = limt→t+ 0 For example, the square wave function defined by f (t) = 41 −1 1 0 when −π < t < 0, when 0 < t < π, A function f is sectionally continuous on (a, b) when f has only a finite number of discontinuities in (a, b) and the one-sided limits exist at each point of discontinuity as well as at the end points a and b. 302 Chapter 5 Norms, Inner Products, and Orthogonality and illustrated in Figure 5.4.2, satisfies these conditions. The value of f at t = 0 is irrelevant—it's not even necessary that f (0) be defined. 1 −π π −1 Figure 5.4.2 To find the Fourier series expansion for f, compute the coefficients in (5.4.6) as 1 an = π π 1 f (t) cos nt dt = π −π 0 1 − cos nt dt + π −π π cos nt dt 0 = 0, 1 π 1 0 1 π f (t) sin nt dt = − sin nt dt + sin nt dt π −π π −π π 0 2 0 when n is even, = (1 − cos nπ) = 4/nπ when n is odd, nπ bn = so that F (t) = ∞ 4 4 4 4 sin t + sin 3t + sin 5t + · · · = sin(2n − 1)t. π 3π 5π (2n − 1)π n=1 For each t ∈ (−π, π), except t = 0, it must be the case that F (t) = f (t), and F (0) = f (0−) + f (0+) = 0. 2 Not only does F (t) agree with f (t) everywhere f is defined, but F also provides a periodic extension of f in the sense that the graph of F (t) is the entire square wave depicted in Figure 5.4.2—the values at the points of discontinuity (the jumps) are F (±nπ) = 0. 5.4 Orthogonal Vectors 303 Exercises for section 5.4 5.4.1. Using the standard inner product, determine which of the following pairs are orthogonal ⎛ vectors ⎞in the indicated ⎛ space. ⎞1 −2 (a) x = ⎝ −3 ⎠and y = ⎝ 2 ⎠in 3 , 4 2 ⎛ ⎞ ⎛ ⎞ i 0 ⎜1 + i⎟ ⎜1 + i⎟ (b) x = ⎝ ⎠and y = ⎝ ⎠in C 4 , 2 −2 1−i 1−i ⎛ ⎞ ⎛ ⎞ 1 4 ⎜ −2 ⎟ ⎜ 2 ⎟ (c) x = ⎝ ⎠and y = ⎝ ⎠in 4 , 3 −1 ⎛ 4 ⎞ ⎛ 1 ⎞ 1+i 1−i (d) x = ⎝ 1 ⎠and y = ⎝ −3 ⎠in C 3 , i −i ⎛ ⎞ ⎛ ⎞y1 0 ⎜ 0 ⎟ ⎜ y2 ⎟ in n . ⎟ (e) x = ⎜ ⎝ ... ⎠and y = ⎝ .. ⎠in . . . 0 yn 5.4.2. Find two vectors of unit norm that are orthogonal to u = 3 −2 . 5.4.3. Consider the following set of three vectors. ⎧ ⎛ ⎞ ⎛ ⎞ ⎛ ⎞⎫ 1 1 −1 ⎪ ⎪ 1 ⎬ ⎨ ⎜ −1 ⎟ ⎜ 1 ⎟ ⎜ −1 ⎟ x1 = ⎝ , x = , x = ⎠ ⎝ ⎠ ⎝ ⎠ . 2 3 0 1 2 ⎪ ⎪ ⎭ ⎩ 2 0 0 (a) Using the standard inner product in 4 , verify that these vectors are mutually orthogonal. (b) Find a nonzero vector x4 such that {x1 , x2 , x3 , x4 } is a set of mutually orthogonal vectors. (c) Convert the resulting set into an orthonormal basis for 4 . 5.4.4. Using the standard inner product, determine the Fourier expansion of x with respect to B, where ⎧ ⎛ ⎞⎛ ⎞⎛ ⎞⎫ 1 1 1 ⎬ ⎨ 1 1 1 x = ⎝ 0 ⎠and B = √ ⎝ −1 ⎠, √ ⎝ 1 ⎠, √ ⎝ −1 ⎠. ⎭ ⎩ 2 ⎭ 3 6 −2 0 1 2 304 Chapter 5 Norms, Inner Products, and Orthogonality 5.4.5. With respect to the inner product for matrices given by (5.3.2), verify that the set ! 1 1 1 1 1 −1 1 1 0 1 1 0 √ √ B= , , , 1 2 1 2 −1 1 2 1 0 2 0 −1 is an orthonormal basis for 2×2 , and then compute the Fourier expan sion of A = 11 11 with respect to B. 2 5.4.6. Determine the angle between x = −1 1 1 and y = 1 2 . 5.4.7. Given an orthonormal basis B for a space V, explain why the Fourier expansion for x ∈ V is uniquely determined by B. 5.4.8. Explain why the columns of Un×n are an orthonormal basis for C n if and only if U∗ = U−1 . Such matrices are said to be unitary—their properties are studied in a later section. 5.4.9. Matrices with the property A∗ A = AA∗ are said to be normal. Notice that hermitian matrices as well as real symmetric matrices are included in the class of normal matrices. Prove that if A is normal, then R (A) ⊥ N (A)—i.e., every vector in R (A) is orthogonal to every vector in N (A). Hint: Recall equations (4.5.5) and (4.5.6). 5.4.10. Using the trace inner product described in Example 5.3.1, determine the angle between the following pairs of matrices. 1 0 1 1 (a) I = and B = . 0 1 1 1 1 3 2 −2 (b) A = and B = . 2 4 2 0 5.4.11. Why is the definition for cos θ given in (5.4.1) not good for C n ? Explain how to define cos θ so that it makes sense in C n . 5.4.12. If {u1 , u2 , . . . , un } is an orthonormal basis for an inner-product space V, explain why x y = x ui ui y i holds for every x, y ∈ V. 5.4 Orthogonal Vectors 305 2 5.4.13. Consider a real inner-product space, where = . (a) Prove that if x = y , then (x + y) ⊥ (x − y). (b) For the standard inner product in 2 , draw a picture of this. That is, sketch the location of x + y and x − y for two vectors with equal norms. 5.4.14. Pythagorean Theorem. Let V be a general inner-product space in 2 which = . (a) When V is a real space, prove that x ⊥ y if and only if 2 2 2 x + y = x + y . (Something would be wrong if this were not true because this is where the definition of orthogonality originated.) (b) Construct an example to show that one of the implications in part (a) does not hold when V is a complex space. (c) When V is a complex space, prove that x ⊥ y if and only if 2 2 2 αx + βy = αx + βy for all scalars α and β. 5.4.15. Let B = {u1 , u2 , . . . , u n } be an orthonormal basis for an inner-product space V, and let x = i ξi ui be the Fourier expansion of x ∈ V. (a) If V is a real space, and if θi is the angle between ui and x, explain why ξi = x cos θi . Sketch a picture of this in 2 or 3 to show why the component ξi ui represents the orthogonal projection of x onto the line determined by ui , and thus illustrate the fact that a Fourier expansion is nothing more than simply resolving x into mutually orthogonal components. n 42 2 2 (b) Derive Parseval's identity, which says i=1 |ξi | = x . 5.4.16. Let B = {u1 , u2 , . . . , uk } be an orthonormal set in an n-dimensional 43 inner-product space V. Derive Bessel's inequality, which says that if x ∈ V and ξi = ui x , then k 2 |ξi |2 ≤ x . i=1 Explain why equality holds if and only if x ∈ span {u1 , u2 , . . . , uk } . k Hint: Consider x − i=1 ξi ui 2 . 42 43 This result appeared in the second of the five mathematical publications by Marc-Antoine Parseval des Chˆenes (1755–1836). Parseval was a royalist who had to flee from France when Napoleon ordered his arrest for publishing poetry against the regime. This inequality is named in honor of the German astronomer and mathematician Friedrich Wilhelm Bessel (1784–1846), who devoted his life to understanding the motions of the stars. In the process he introduced several useful mathematical ideas. 306 Chapter 5 Norms, Inner Products, and Orthogonality 5.4.17. Construct an example using the standard inner product in n to show that two vectors x and y can have an angle between them that is close to π/2 without xT y being close to 0. Hint: Consider n to be large, and use the vector e of all 1's for one of the vectors. 5.4.18. It was demonstrated in Example 5.4.3 that y is linearly correlated with x in the sense that y ≈ β0 e + β1 x if and only if the standardization vectors zx and zy are "close" in the sense that they are almost on the same line in n . Explain why simply measuring zx − zy 2 does not always gauge the degree of linear correlation. 5.4.19. Let θ be the angle between two vectors x and y from a real innerproduct space. (a) Prove that cos θ = 1 if and only if y = αx for α > 0. (b) Prove that cos θ = −1 if and only if y = αx for α < 0. Hint: Use the generalization of Exercise 5.1.9. 5.4.20. With respect to the orthonormal set B= ! 1 cos t cos 2t sin t sin 2t sin 3t √ , √ , √ ,..., √ , √ , √ ,... , π π π π π 2π determine the Fourier series expansion of the saw-toothed function defined by f (t) = t for −π < t < π. The periodic extension of this function is depicted in Figure 5.4.3. π −π π −π Figure 5.4.3 5.5 Gram–Schmidt Procedure 5.5 307 GRAM–SCHMIDT PROCEDURE As discussed in §5.4, orthonormal bases possess significant advantages over bases that are not orthonormal. The spaces n and C n clearly possess orthonormal bases (e.g., the standard basis), but what about other spaces? Does every finitedimensional space possess an orthonormal basis, and, if so, how can one be 44 produced? The Gram–Schmidt orthogonalization procedure developed below answers these questions. Let B = {x1 , x2 , . . . , xn } be an arbitrary basis (not necessarily orthonormal) 1/2 for an n-dimensional inner-product space S, and remember that = . Objective: Use B to construct an orthonormal basis O = {u1 , u2 , . . . , un } for S. Strategy: Construct O sequentially so that Ok = {u1 , u2 , . . . , uk } is an orthonormal basis for Sk = span {x1 , x2 , . . . , xk } for k = 1, . . . , n. For k = 1, simply take u1 = x1 / x1 . It's clear that O1 = {u1 } is an orthonormal set whose span agrees with that of S1 = {x1 } . Now reason inductively. Suppose that Ok = {u1 , u2 , . . . , uk } is an orthonormal basis for Sk = span {x1 , x2 , . . . , xk } , and consider the problem of finding one additional vector uk+1 such that Ok+1 = {u1 , u2 , . . . , uk , uk+1 } is an orthonormal basis for Sk+1 = span {x1 , x2 , . . . , xk , xk+1 } . For this to hold, the Fourier expansion (p. 299) of xk+1 with respect to Ok+1 must be xk+1 = k+1 ui xk+1 ui , i=1 which in turn implies that uk+1 k xk+1 − i=1 ui xk+1 ui . = uk+1 xk+1 (5.5.1) Since uk+1 = 1, it follows from (5.5.1) that k | uk+1 xk+1 | = xk+1 − ui xk+1 ui , i=1 44 Jorgen P. Gram (1850–1916) was a Danish actuary who implicitly presented the essence of orthogonalization procedure in 1883. Gram was apparently unaware that Pierre-Simon Laplace (1749–1827) had earlier used the method. Today, Gram is remembered primarily for his development of this process, but in earlier times his name was also associated with the matrix product A∗ A that historically was referred to as the Gram matrix of A. Erhard Schmidt (1876–1959) was a student of Hermann Schwarz (of CBS inequality fame) and the great German mathematician David Hilbert. Schmidt explicitly employed the orthogonalization process in 1907 in his study of integral equations, which in turn led to the development of what are now called Hilbert spaces. Schmidt made significant use of the orthogonalization process to develop the geometry of Hilbert Spaces, and thus it came to bear Schmidt's name. 308 Chapter 5 Norms, Inner Products, and Orthogonality k so uk+1 xk+1 = eiθ xk+1 − i=1 ui xk+1 ui for some 0 ≤ θ < 2π, and k xk+1 − i=1 ui xk+1 ui . uk+1 = k eiθ xk+1 − i=1 ui xk+1 ui Since the value of θ in the scalar eiθ neither affects span {u1 , u2 , . . . , uk+1 } nor the facts that uk+1 = 1 and uk+1 ui = 0 for all i ≤ k, we can arbitrarily define uk+1 to be the vector corresponding to the θ = 0 or, equivalently, eiθ = 1. For the sake of convenience, let k νk+1 = xk+1 − i=1 ui xk+1 ui so that we can write k xk+1 − i=1 ui xk+1 ui x1 u1 = x1 and uk+1 = for k > 0. (5.5.2) x1 νk+1 This sequence of vectors is called the Gram–Schmidt sequence. A straightforward induction argument proves that Ok = {u1 , u2 , . . . , uk } is indeed an orthonormal basis for span {x1 , x2 , . . . , xk } for each k = 1, 2, ... Details are called for in Exercise 5.5.7. The orthogonalization procedure defined by (5.5.2) is valid for any innerproduct space, but if we concentrate on subspaces of m or C m with the standard inner product and euclidean norm, then we can formulate (5.5.2) in terms of matrices. Suppose that B = {x1 , x2 , . . . , xn } is a basis for an n-dimensional subspace S of C m×1 so that the Gram–Schmidt sequence (5.5.2) becomes k−1 xk − i=1 (u∗i xk) ui x1 for k = 2, 3, . . . , n. (5.5.3) u1 = x1 and uk = k−1 x1 xk − i=1 (u∗i xk) ui To express this in matrix notation, set U1 = 0m×1 and Uk = u1 | u2 | · · · | uk−1 m×k−1 and notice that ⎛ u∗ x ⎞ U∗k xk = ⎜ ⎝ 1 k u∗2 xk .. . u∗k−1 xk Since xk − k−1 for k > 1, ⎞ ⎟ ⎠and Uk U∗k xk = k−1 ui (u∗i xk) ui . i=1 k−1 (u∗i xk) ui = xk − Uk U∗k xk = (I − Uk U∗k) xk , i=1 the vectors in (5.5.3) can be concisely written as (I − Uk U∗k) xk uk = for k = 1, 2, . . . , n. (I − Uk U∗k) xk Below is a summary. 5.5 Gram–Schmidt Procedure 309 Gram–Schmidt Orthogonalization Procedure If B = {x1 , x2 , . . . , xn } is a basis for a general inner-product space S, then the Gram–Schmidt sequence defined by x1 u1 = x1 and uk = k−1 xk − i=1 ui xk ui for k = 2, . . . , n uk = k−1 xk − i=1 ui xk ui is an orthonormal basis for S. When S is an n-dimensional subspace of C m×1 , the Gram–Schmidt sequence can be expressed as uk = (I − Uk U∗k) xk (I − Uk U∗k) xk for k = 1, 2, . . . , n (5.5.4) in which U1 = 0m×1 and Uk = u1 | u2 | · · · | uk−1 m×k−1 for k > 1. Example 5.5.1 Classical Gram–Schmidt Algorithm. The following formal algorithm is the straightforward or "classical" implementation of the Gram–Schmidt procedure. Interpret a ← b to mean that "a is defined to be (or overwritten by) b." For k = 1: x1 u1 ← x1 For k > 1: uk ← xk − uk ← uk uk k−1 i=1 (u∗i xk)ui (See Exercise 5.5.10 for other formulations of the Gram–Schmidt algorithm.) Problem: Use the classical formulation of the Gram–Schmidt procedure given above to find an orthonormal basis for the space spanned by the following three linearly independent vectors. ⎛ ⎞1 ⎜ 0 ⎟x1 = ⎝ ⎠, 0 −1 ⎛ ⎞1 ⎜ 2 ⎟x2 = ⎝ ⎠, 0 −1 ⎛ ⎞3 ⎜ 1 ⎟x3 = ⎝ ⎠. 1 −1 310 Chapter 5 Norms, Inner Products, and Orthogonality Solution: k = 1: k = 2: k = 3: Thus ⎛ ⎞1 1 ⎜ 0 ⎟x1 u1 ← =√ ⎝ ⎠ 0 x1 2 −1 ⎛ ⎞0 2 ⎜ ⎟ u2 ← x2 − (uT1 x2)u1 = ⎝ ⎠, 0 0 ⎛ 0 u2 ⎜ 1 ⎟ =⎝ ⎠u2 ← 0 u2 0 ⎛ ⎞ 1 1 u 1 0 0 ⎜ ⎟ 3 u3 ← x3 − (uT1 x3)u1 − (uT2 x3)u2 = ⎝ ⎠, u3 ← =√ ⎝ ⎠1 u3 3 1 1 1 ⎛ ⎞⎛ ⎞⎛ ⎞0 1 1 1 ⎜ 0 ⎟ 1 ⎜1⎟ 1 ⎜0 ⎟ u1 = √ ⎝ ⎠, u2 = ⎝ ⎠, u3 = √ ⎝ ⎠0 0 2 3 1 0 −1 1 is the desired orthonormal basis. The Gram–Schmidt process frequently appears in the disguised form of a matrix factorization. To see this, let Am×n = a1 | a2 | · · · | an be a matrix with linearly independent columns. When Gram–Schmidt is applied to the columns of A, the result is an orthonormal basis {q1 , q2 , . . . , qn } for R (A), where a1 q1 = ν1 and qk = ak − k−1 i=1 qi ak qi νk for k > 1, where ν1 = a1 and νk = ak − k−1 i=1 qi ak qi for k > 1. The above relationships can be rewritten as a1 = ν1 q1 and ak = q1 ak q1 + · · · + qk−1 ak qk−1 + νk qk for k > 1, which in turn can be expressed in matrix form by writing ⎛ν a1 | a2 | · · · | an ⎜ = q1 | q2 | · · · | qn ⎜ ⎜ ⎝ 1 0 0 .. . 0 q1 a2 q1 a3 · · · q1 an ⎞ν2 q2 a3 · · · q2 an ⎟ ⎟ 0 ν3 · · · q3 an ⎟ . ⎟ . . .. ⎠ .. . . 0 0 ··· νn This says that it's possible to factor a matrix with independent columns as Am×n = Qm×n Rn×n , where the columns of Q are an orthonormal basis for R (A) and R is an upper-triangular matrix with positive diagonal elements. 5.5 Gram–Schmidt Procedure 311 The factorization A = QR is called the QR factorization for A, and it is uniquely determined by A (Exercise 5.5.8). When A and Q are not square, some authors emphasize the point by calling A = QR the rectangular QR factorization—the case when A and Q are square is further discussed on p. 345. Below is a summary of the above observations. QR Factorization Every matrix Am×n with linearly independent columns can be uniquely factored as A = QR in which the columns of Qm×n are an orthonormal basis for R (A) and Rn×n is an upper-triangular matrix with positive diagonal entries. • The QR factorization is the complete "road map" of the Gram– Schmidt process because the columns of Q = q1 | q2 | · · · | qn are the result of applying the Gram–Schmidt procedure to the columns of A = a1 | a2 | · · · | an and R is given by ⎛ ⎞ ν1 q∗1 a2 q∗1 a3 · · · q∗1 an ⎜ 0 ν2 q∗2 a3 · · · q∗2 an ⎟ ⎟ R=⎜ ⎜ 0 0 ν3 · · · q∗3 an ⎟ , ⎝ . . . . . ⎠ 0 0 0 · · · νn where ν1 = a1 and νk = ak − k−1 i=1 qi ak qi for k > 1. Example 5.5.2 Problem: Determine the QR factors of A = ⎛ 0 −20 −14 ⎞ ⎝ 3 27 −4 ⎠ 4 11 −2 . Solution: Using the standard inner product for n , apply the Gram–Schmidt procedure to the columns of A by setting k−1 ak −i=1 qi ak qi q1 = and qk = for k = 2, 3, νk where ν1 = a1 and νk = ak − k−1 i=1 qi ak qi . The computation of these quantities can be organized as follows. 312 Chapter 5 Norms, Inner Products, and Orthogonality k = 1: r11 ← a1 = 5 and q1 ← a1/r11 = (0 3/5 4/5) k = 2: r12 ← qT1 a2 = 25 q2 ← a2 − r12 q1 = (−20 12 −9) r22 ← q2 = 25 and q2 ← q2/r22 = (−20/25 12/25 −9/25) k = 3: r13 ← qT1 a3 = −4 and r23 ← qT2 a3 = 10 q3 ← a3 − r13 q1 − r23 q2 = −2/5 (15 −16 12) r33 ← q3 = 10 and q3 ← q3/r33 = (−15 16 −12) /25. Therefore, Q = (0 −20 −15 ; 15 12 16 ; 20 −9 −12) /25 and R = (5 25 −4 ; 0 25 10 ; 0 0 10). We now have two important matrix factorizations, namely, the LU factorization, discussed in §3.10 on p. 141 and the QR factorization. They are not the same, but some striking analogies exist. • Each factorization represents a reduction to upper-triangular form—LU by Gaussian elimination, and QR by Gram–Schmidt. In particular, the LU factorization is the complete "road map" of Gaussian elimination applied to a square nonsingular matrix, whereas QR is the complete road map of Gram– Schmidt applied to a matrix with linearly independent columns. • When they exist, both factorizations A = LU and A = QR are uniquely determined by A. • Once the LU factors (assuming they exist) of a nonsingular matrix A are known, the solution of Ax = b is easily computed—solve Ly = b by forward substitution, and then solve Ux = y by back substitution (see p. 146). The QR factors can be used in a similar manner. If A ∈ n×n is nonsingular, then QT = Q−1 (because Q has orthonormal columns), so Ax = b ⇐⇒ QRx = b ⇐⇒ Rx = QT b, which is also a triangular system that is solved by back substitution. 5.5 Gram–Schmidt Procedure 313 While the LU and QR factors can be used in more or less the same way to solve nonsingular systems, things are different for singular and rectangular cases because Ax = b might be inconsistent, in which case a least squares solution as described in §4.6, (p. 223) may be desired. Unfortunately, the LU factors of A don't exist when A is rectangular. And even if A is square and has an LU factorization, the LU factors of A are not much help in solving the system of normal equations AT Ax = AT b that produces least squares solutions. But the QR factors of Am×n always exist as long as A has linearly independent columns, and as demonstrated in the following example, the QR factors provide the least squares solution of an inconsistent system in exactly the same way as they provide the solution of a consistent system. Example 5.5.3 Application to the Least Squares Problem. If Ax = b is a possibly inconsistent (real) system, then, as discussed on p. 226, the set of all least squares solutions is the set of solutions to the system of normal equations AT Ax = AT b. (5.5.5) But computing AT A and then performing an LU factorization of AT A to solve (5.5.5) is generally not advisable. First, it's inefficient and, second, as pointed out in Example 4.5.1, computing AT A with floating-point arithmetic can result in a loss of significant information. The QR approach doesn't suffer from either of these objections. Suppose that rank (Am×n) = n (so that there is a unique least squares solution), and let A = QR be the QR factorization. Because the columns of Q are an orthonormal set, it follows that QT Q = In , so AT A = (QR)T (QR) = RT QT QR = RT R. (5.5.6) Consequently, the normal equations (5.5.5) can be written as RT Rx = RT QT b. (5.5.7) But RT is nonsingular (it is triangular with positive diagonal entries), so (5.5.7) simplifies to become Rx = QT b. (5.5.8) This is just an upper-triangular system that is efficiently solved by back substitution. In other words, most of the work involved in solving the least squares problem is in computing the QR factorization of A. Finally, notice that −1 T x = R−1 QT b = AT A A b is the solution of Ax = b when the system is consistent as well as the least squares solution when the system is inconsistent (see p. 214). That is, with the approach described here, there is no need to distinguish between consistent and inconsistent systems—in both cases the solution is obtained from the same formula x = R−1 QT b. Below is a formal summary. 314 Chapter 5 Norms, Inner Products, and Orthogonality Linear Systems and the QR Factorization If rank (Am×n) = n, and if A = QR is the QR factorization, then the solution of the nonsingular triangular system Rx = QT b (5.5.9) is either the solution or the least squares solution of Ax = b depending on whether or not Ax = b is consistent. It's worthwhile to reemphasize that the QR approach to the least squares problem obviates the need to explicitly compute the product AT A. But if AT A is ever needed, it is retrievable from the factorization AT A = RT R. In fact, this is the Cholesky factorization of AT A as discussed in Example 3.10.7, p. 154. The Gram–Schmidt procedure is a powerful theoretical tool, but it's not a good numerical algorithm when implemented in the straightforward or "classical" sense. When floating-point arithmetic is used, the classical Gram–Schmidt algorithm applied to a set of vectors that is not already close to being an orthogonal set can produce a set of vectors that is far from being an orthogonal set. To see this, consider the following example. Example 5.5.4 Problem: Using 3-digit floating-point arithmetic, apply the classical Gram– Schmidt algorithm to the set ⎧⎛ 1 ⎞ ⎛ 1 ⎞ ⎛ 1 ⎞⎫ x1 = ⎝ 10−3 ⎠ , x2 = ⎝ 10−3 ⎠ , x3 = ⎝ 0 ⎠ , 10−3 0 10−3 Solution: k = 1: fl x1 = 1, so u1 ← x1 . k = 2: fl uT1 x2 = 1, so ⎛ 0 ⎞ u1 ⎛ 0 ⎞ u2 ← x2 − uT1 x2 u1 = ⎝ 0 ⎠ and u2 ← fl = ⎝ 0 ⎠ . −10−3 u2 −1 k = 3: fl uT1 x3 = 1 and fl uT2 x3 = −10−3 , so ⎛ 0 ⎞ u3 ⎛ 0 ⎞ u3 ← x3 − uT1 x3 u1 − uT2 x3 u2 = ⎝ −10−3 ⎠ and u3 ← fl = ⎝ −.709 ⎠ . −10−3 u3 −.709 5.5 Gram–Schmidt Procedure 315 Therefore, classical Gram–Schmidt with 3-digit arithmetic returns ⎛ 1 ⎞ ⎛ 0 ⎞ ⎛ 0 ⎞ u1 = ⎝ 10−3 ⎠ , u2 = ⎝ 0 ⎠ , u3 = ⎝ −.709 ⎠ , (5.5.10) 10−3 −1 −.709 which is unsatisfactory because u2 and u3 are far from being orthogonal. It's possible to improve the numerical stability of the orthogonalization process by rearranging the order of the calculations. Recall from (5.5.4) that uk = (I − Uk UkT) xk , where U1 = 0 and Uk = u1 | u2 | · · · | uk−1 for k > 1. If E1 = I and Ei = I − ui−1 ui−1T for i > 1, then the orthogonality of the ui 's insures that Ek · · · E2 E1 = I − u1 uT1 − u2 uT2 − · · · − uk−1 uTk−1 = I − Uk UkT , so the Gram–Schmidt sequence can also be expressed as uk = Ek · · · E2 E1 xk for k = 1, 2, . . . , n. This means that the Gram–Schmidt sequence can be generated as follows: Normalize 1-st (x1 , x2 , . . . , xn) Apply E2 = (u1 , x2 , . . . , xn) Normalize 2-nd = (u1 , E2 x2 , E2 x3 , . . . , E2 xn) Apply E3 (u1 , u2 , E2 x3 , . . . , E2 xn) Normalize 3rd = (u1 , u2 , E3 E2 x3 , . . . , E3 E2 xn) (u1 , u2 , u3 , E3 E2 x4 , . . . , E3 E2 xn) , etc. While there is no theoretical difference, this "modified" algorithm is numerically more stable than the classical algorithm when floating-point arithmetic is used. The k th step of the classical algorithm alters only the k th vector, but the k th step of the modified algorithm "updates" all vectors from the k th through the last, and conditioning the unorthogonalized tail in this way makes a difference. 316 Chapter 5 Norms, Inner Products, and Orthogonality Modified Gram–Schmidt Algorithm For a linearly independent set {x1 , x2 , . . . , xn } ⊂ C m×1 , the Gram–Schmidt sequence given on p. 309 can be alternately described as uk = Ek · · · E2 E1 xk with E1 = I, Ei = I − ui−1 u∗i−1 for i > 1, and this sequence is generated by the following algorithm. For k = 1: u1 ← x1 / x1 and uj ← xj for j = 2, 3, . . . , n For k > 1: uj ← Ek uj = uj − u∗k−1 uj uk−1 for j = k, k + 1, . . . , n uk ← uk / uk (An alternate implementation is given in Exercise 5.10.10.) To see that the modified version of Gram–Schmidt can indeed make a difference when floating-point arithmetic is used, consider the following example. Example 5.5.5 Problem: Use 3-digit floating-point arithmetic, and apply the modified Gram–Schmidt algorithm to the set given in Example 5.5.4 (p. 314), and then compare the results of the modified algorithm with those of the classical algorithm. Solution: ⎧⎛ 1 ⎞ ⎛ 1 ⎞ ⎛ 1 ⎞⎫ x1 = ⎝ 10−3 ⎠ , x2 = ⎝ 10−3 ⎠ , x3 = ⎝ 0 ⎠ ⎩ ⎭ 10−3 0 10−3 k = 1: fl x1 = 1, so {u1 , u2 , u3 } ← {x1 , x2 , x3 } . k = 2: fl uT1 u2 = 1 and fl uT1 u3 = 1, so ⎛ 0 ⎞ ⎛ 0 ⎞ u2 ← u2 − uT1 u2 u1 = ⎝ 0 ⎠ , u3 ← u3 − uT1 u3 u1 = ⎝ −10−3 ⎠ , −10−3 0 and ⎛ 0 ⎞ u2 u2 ← = ⎝ 0 ⎠ . u2 −1 k = 3: uT2 u3 = 0, so ⎛ 0 ⎞ ⎛ 0 ⎞ u3 u3 ← u3 − uT2 u3 u2 = ⎝ −10−3 ⎠ and u3 ← = ⎝ −1 ⎠ . 0 u3 0 5.5 Gram–Schmidt Procedure 317 Thus the modified Gram–Schmidt algorithm produces ⎛ 1 ⎞ ⎛ 0 ⎞ ⎛ 0 ⎞ u1 = ⎝ 10−3 ⎠ , u2 = ⎝ 0 ⎠ , u3 = ⎝ −1 ⎠ , (5.5.11) 10−3 −1 0 which is as good as one can expect using 3-digit arithmetic. Comparing (5.5.11) with the result (5.5.10) obtained in Example 5.5.4 illustrates the advantage possessed by modified Gram–Schmidt algorithm over the classical algorithm. Below is a summary of some facts concerning the modified Gram–Schmidt algorithm compared with the classical implementation. Summary • When the Gram–Schmidt procedures (classical or modified) are applied to the columns of A using exact arithmetic, each produces an orthonormal basis for R (A). • For computing a QR factorization in floating-point arithmetic, the modified algorithm produces results that are at least as good as and often better than the classical algorithm, but the modified algorithm is not unconditionally stable—there are situations in which it fails to produce a set of columns that are nearly orthogonal. • For solving the least square problem with floating-point arithmetic, the modified procedure is a numerically stable algorithm in the sense that the method described in Example 5.5.3 returns a result that is the exact solution of a nearby least squares problem. However, the Householder method described on p. 346 is just as stable and needs slightly fewer arithmetic operations. Exercises for section 5.5 ⎧⎛ 1 ⎞ ⎛ 1 ⎞ ⎛ −1 ⎞⎫ ⎨ 0 2 1 ⎬ 5.5.1. Let S = ⎝ ⎠ , ⎝ ⎠ , ⎝ ⎠ . ⎩ 0 1 1 ⎭ 0 1 1 (a) Use the classical Gram–Schmidt algorithm (with exact arithmetic) to determine an orthonormal basis for span (S) . (b) Verify directly that the Gram–Schmidt sequence produced in part (a) is indeed an orthonormal basis for span (S) . (c) Repeat part (a) using the modified Gram–Schmidt algorithm, and compare the results. 5.5.2. Use the Gram–Schmidt procedure to find an orthonormal basis for the four fundamental subspaces of A = ⎛ 1 −2 3 −1 ⎞ ⎝ 2 −4 6 −2 ⎠ . 3 −6 9 −3 318 Chapter 5 Norms, Inner Products, and Orthogonality 5.5.3. Apply the Gram–Schmidt procedure with the standard inner product for C 3 to ⎧⎛ i ⎞ ⎛ 0 ⎞ ⎛ 0 ⎞⎫ ⎝ i ⎠ , ⎝ i ⎠ , ⎝ 0 ⎠ . ⎩ i i i ⎭ 5.5.4. Explain what happens when the Gram–Schmidt process is applied to an orthonormal set of vectors. 5.5.5. Explain what happens when the Gram–Schmidt process is applied to a linearly dependent set of vectors. 5.5.6. Let A = ⎛ 1 0 −1 ⎞ ⎜ 1 2 1 ⎟ ⎝ 1 1 −3 ⎠ 0 1 1 and b = ⎛ 1 ⎞ ⎜ 1 ⎟ ⎝ 1 ⎠ 1 . (a) Determine the rectangular QR factorization of A. (b) Use the QR factors from part (a) to determine the least squares solution to Ax = b. 5.5.7. Given a linearly independent set of vectors S = {x1 , x2 , . . . , xn } in an inner-product space, let S = {u1 , u2 , . . . , un } be the Gram–Schmidt sequence defined in (5.5.2). Prove that if x = Σ ... then Ok is related to the set {x1 , x2 , . . . , xk } for each k, and (a) ... 5.5.8. Explain why if A = QR is the rectangular QR factorization of A, then R is unique. Hint: Recall Example 5.5.8 ... 5.5.9. Use 3-digit floating-point arithmetic, apply the modified Gram–Schmidt algorithm to the set {x1 , x2 , x3 } ... (a) Again use 3-digit floating-point arithmetic, but ... apply the modified Gram–Schmidt algorithm ... 5.5.10. Use Legendre's polynomials ... 5.5.11. Let V be the inner-product space of real-valued continuous functions defined on the interval [−1, 1], where the inner product is defined by f g = ∫ f(x)g(x)dx, and let S be the subspace of V that is spanned by the three linearly independent polynomials q0 = 1, q1 = x, q2 = x2 . (a) Use the Gram–Schmidt process to determine an orthonormal set of polynomials {p0 , p1 , p2 } that spans S. These polynomials are the first three normalized Legendre polynomials. (b) Verify that pn satisfies Legendre's differential equation (1 − x2)y − 2xy + n(n + 1)y = 0 for n = 0, 1, 2. This equation and its solutions are of considerable importance in applied mathematics. 45 Adrien–Marie Legendre (1752–1833) was one of the most eminent French mathematicians of the eighteenth century. His primary work in higher mathematics concerned number theory and the study of elliptic functions. But he was also instrumental in the development of the theory of least squares, and some people believe that Legendre should receive the credit that is often afforded to Gauss for the introduction of the method of least squares. Like Gauss and many other successful mathematicians, Legendre spent substantial time engaged in diligent and painstaking computation. It is reported that in 1824 Legendre refused to vote for the government's candidate for Institut National, so his pension was stopped, and he died in poverty. 320 5.6 Chapter 5 Norms, Inner Products, and Orthogonality UNITARY AND ORTHOGONAL MATRICES The purpose of this section is to examine square matrices whose columns (or rows) are orthonormal. The standard inner product and the euclidean 2-norm are the only ones used in this section, so distinguishing subscripts are omitted. Unitary and Orthogonal Matrices • A unitary matrix is defined to be a complex matrix Un×n whose columns (or rows) constitute an orthonormal basis for C n . • An orthogonal matrix is defined to be a real matrix Pn×n whose columns (or rows) constitute an orthonormal basis for n . Unitary and orthogonal matrices have some nice features, one of which is the fact that they are easy to invert. To see why, notice that the columns of U = u1 | u2 | · · · | un are an orthonormal set if and only if [U∗ U]ij = u∗i uj = 1 when i = j, 0 when i = j ⇐⇒ U∗ U = I ⇐⇒ U−1 = U∗ . Notice that because U∗ U = I ⇐⇒ UU∗ = I, the columns of U are orthonormal if and only if the rows of U are orthonormal, and this is why the definitions of unitary and orthogonal matrices can be stated either in terms of orthonormal columns or orthonormal rows. Another nice feature is that multiplication by a unitary matrix doesn't change the length of a vector. Only the direction can be altered because Ux 2 = x∗ U∗ Ux = x∗ x = x 2 ∀ x ∈ C n . (5.6.1) Conversely, if (5.6.1) holds, then U must be unitary. To see this, set x = ei in (5.6.1) to observe u∗i ui = 1 for each i, and then set x = ej + ek for j = k to obtain 0 = u∗j uk + u∗k uj = 2 Re (u∗j uk). By setting x = ej + iek in (5.6.1), it also follows that 0 = 2 Im (u∗j uk), so u∗j uk = 0 for each j = k, and thus (5.6.1) guarantees that U is unitary. 5.6 Unitary and Orthogonal Matrices 321 In the case of orthogonal matrices, everything is real so that (·)∗ can be replaced by (·)T . Below is a summary of these observations. Characterizations The following statements are equivalent to saying that a complex matrix Un×n is unitary. • U has orthonormal columns. • U has orthonormal rows. • U−1 = U∗ . • Ux 2 = x 2 for every x ∈ C n×1 . The following statements are equivalent to saying that a real matrix Pn×n is orthogonal. • P has orthonormal columns. • P has orthonormal rows. • P−1 = PT . • Px 2 = x 2 for every x ∈ n×1 . Example 5.6.1 • The identity matrix I is an orthogonal matrix. • All permutation matrices (products of elementary interchange matrices) are orthogonal—recall Exercise 3.9.4. • The matrix P = (1/√2 1/√3 −1/√6 ; −1/√2 1/√3 −1/√6 ; 0 1/√3 2/√6) is an orthogonal matrix because PT P = PPT = I or, equivalently, because the columns (and rows) constitute an orthonormal set. • The matrix U = (1/2) (1 + i −1 + i ; 1 + i 1 − i) is unitary because U∗ U = UU∗ = I or, equivalently, because the columns (and rows) are an orthonormal set. • An orthogonal matrix can be considered to be unitary, but a unitary matrix is generally not orthogonal. In general, a linear operator T on a vector space V with the property that Tx = x for all x ∈ V is called an isometry on V. The isometries on n are precisely the orthogonal matrices, and the isometries on C n are the unitary matrices. The term "isometry" has an advantage in that it can be used to treat the real and complex cases simultaneously, but for clarity we will often revert back to the more cumbersome "orthogonal" and "unitary" terminology. 322 Chapter 5 Norms, Inner Products, and Orthogonality The geometrical concepts of projection, reflection, and rotation are among the most fundamental of all linear transformations in 2 and 3 (see Example 4.7.1 for three simple examples), so pursuing these ideas in higher dimensions is only natural. The reflector and rotator given in Example 4.7.1 are isometries (because they preserve length), but the projector is not. We are about to see that the same is true in more general settings. Elementary Orthogonal Projectors For a vector u ∈ C n×1 such that u = 1, a matrix of the form Q = I − uu∗ (5.6.2) is called an elementary orthogonal projector. More general projectors are discussed on pp. 386 and 429. To understand the nature of elementary projectors consider the situation in 3 . Suppose that u3×1 = 1, and let u⊥ denote the space (the plane through the origin) consisting of all vectors that are perpendicular to u—we call u⊥ the orthogonal complement of u (a more general definition appears on p. 403). The matrix P = uuT is the orthogonal projector onto u in the sense that P maps each x ∈ 3×1 to its orthogonal projection in u as shown in Figure 5.6.1. u⊥ (I - uuT)x u x 0 uuTx Figure 5.6.1 To see this, observe that each x can be resolved into two components x = (I − Qx + Qx) where Q = I − uuT. ... Notice that (I − Q)x = u(uT x) is on the line determined by u, and Qx is in the plane u⊥ because uT Qx = 0. The "elementary" in (5.6.2) stems from the fact that the projection is onto the one-dimensional space span {u}. ... Therefore, the orthogonal projection of x onto the subspace spanned by u is (5.6.4) ... and the orthogonal projection of x onto the orthogonal complement u⊥ is ... (5.6.5) ... Geometry of Elementary Projectors ... Example 5.6.2 Problem: Determine the orthogonal projection of a vector v onto the space span {u} ... Elementary Reflectors For un×1 = 0, the elementary reflector about u⊥ is defined to be R = I − 2 uu∗/u∗u (5.6.6) or, equivalently, R = I − 2uu∗ when u = 1. (5.6.7) Elementary reflectors are also called Householder transformations. ... Alston Scott Householder (1904–1993) was one of the first people to appreciate and promote the use of elementary reflectors for numerical applications. ... Householder joined the Mathematics Division of Oak Ridge National Laboratory in 1946 and became its director in 1948. ... Properties of Elementary Reflectors ... R = R∗ = R−1 , i.e., R is unitary, hermitian, and involutory ... Reflector Theorem If vn×1 and un×1 are vectors of the same length with v = u ... Example 5.6.3 ... Example 5.6.4 Rotations in 2 and 3 ... Problem: ... yz-plane ... Solution: ... (5.6.13) ... Figure 5.6.5 ... Figure 5.6.6 ... Figure 5.6.7 Just set the x-coordinate of each vertex to 0 (i.e., ignore the x-coordinates), plot the (y, z)-coordinates on the yz-plane (the screen), and 5.6 Unitary and Orthogonal Matrices 331 draw edges between appropriate vertices. For example, suppose that the vertices of the polytope in Figure 5.6.5 are numbered as indicated below in Figure 5.6.7, z 5 6 10 9 7 8 1 4 y 2 x 3 Figure 5.6.7 and suppose that the associated vertex matrix is v1 x 0 V= y⎛ 0 z 0 ⎛ v2 1 0 0 v3 1 1 0 v4 0 1 0 v5 0 0 1 v6 1 0 1 v7 1 .8 1 v8 1 1 .8 v9 .8 1 1 v10 ⎞ 0 1 . 1 There are 15 edges, and they can be recorded in an edge matrix E= e1 1 2 e2 2 3 e3 3 4 e4 4 1 e5 1 5 e6 2 6 e7 3 8 e8 4 10 e9 5 6 e10 6 7 e11 7 8 e12 7 9 e13 8 9 e14 9 10 e15 10 5 in which the k th column represents an edge between the indicated pair of vertices. To display the image of the polytope in Figure 5.6.7 on a monitor, (i) drop the first row from V, (ii) plot the remaining yz-coordinates on the screen, (iii) draw edges between appropriate vertices as dictated by the information in the edge matrix E. To display the image of the polytope after it has been rotated counterclockwise around the x-, y-, and z-axes by 90◦ , 45◦ , and 60◦ , respectively, use the orthogonal matrix P = Pz Py Px determined in (5.6.15) and compute the product ⎛ 0 .354 .707 PV = ⎝ 0 .612 1.22 0 −.707 0 .354 .612 .707 .866 −.5 0 1.22 1.5 .112 .602 −.707 −.141 1.4 1.5 .825 .602 0 .141 ⎞ 1.22 .112 ⎠ . .707 Now proceed as before—(i) ignore the first row of PV, (ii) plot the points in the second and third row of PV as yz-coordinates on the monitor, (iii) draw edges between appropriate vertices as indicated by the edge matrix E. 332 Chapter 5 Norms, Inner Products, and Orthogonality Problem: In addition to rotation, how can a polytope (or its image on a computer monitor) be translated? Solution: Translation of a polytope to a different point in space is accomplished by adding a constant to each of its coordinates. For example, to translate the polytope shown in Figure 5.6.7 to the location where vertex 1 is at pT = (x0 , y0 , z0) instead of at the origin, just add p to every point. In particular, if e is the column of 1's, the translated vertex matrix is ⎞ x0 x0 · · · x0 Vtrans = Vorig + ⎝ y0 y0 · · · y0 ⎠ = Vorig + peT (a rank-1 update). z 0 z 0 · · · z0 Of course, the edge matrix is not affected by translation. Problem: How can a polytope (or its image on a computer monitor) be scaled? Solution: Simply multiply every coordinate by the desired scaling factor. For example, to scale an image by a factor α, form the scaled vertex matrix Vscaled = αVorig , and then connect the scaled vertices with appropriate edges as dictated by the edge matrix E. Problem: How can the faces of a polytope that are hidden from the viewer's perspective be detected so that they can be omitted from the drawing on the screen? Solution: A complete discussion of this tricky problem would carry us too far astray, but one clever solution relying on the cross product of vectors in 3 is presented in Exercise 5.6.21 for the case of convex polytopes. Rotations in higher dimensions are straightforward generalizations of rotations in 3 . Recall from p. 328 that rotation around any particular axis in 3 amounts to rotation in the complementary plane, and the associated 3 × 3 rotator is constructed by embedding a 2 × 2 rotator in the appropriate position in a 3 × 3 identity matrix. For example, rotation around the y-axis is rotation in the xz-plane, and the corresponding rotator is produced by embedding cos θ sin θ − sin θ cos θ in the " xz-position" of I3×3 to form ⎛ cos θ Py = ⎝ 0 − sin θ 0 1 0 ⎞sin θ 0 ⎠ . cos θ These observations directly extend to higher dimensions. 5.6 Unitary and Orthogonal Matrices 333 Plane Rotations Orthogonal matrices of the form ⎛ ⎞ Pij = ⎝ col j ↓ col i ↓ 1 .. . c s 1 −s ... c 1 ... ⎞ ←− row i ⎞ ←− row j ⎠ 1 in which c2 + s2 = 1 are called plane rotation matrices because they perform a rotation in the (i, j)-plane of n . The entries c and s are meant to suggest cosine and sine, respectively, but designating a rotation angle θ as is done in 2 and 3 is not useful in higher dimensions. 48 Pij Plane rotations matrices Pij are also called Givens rotations. Applying to 0 = x ∈ n rotates the (i, j)-coordinates of x in the sense that ⎞ x1 . .. ⎞ cxi + sxj ⎞←− i ⎞ .. ⎞Pij x = ⎞ . ⎞. ⎞ −sxi + cxj ⎞←− j ⎞ ⎠ .. . xn If xi and xj are not both zero, and if we set xi xj c= and s = , 2 2 2 xi + xj xi + x2j 48 (5.6.16) J. Wallace Givens, Jr. (1910–1993) pioneered the use of plane rotations in the early days of automatic matrix computations. Givens graduated from Lynchburg College in 1928, and he completed his Ph.D. at Princeton University in 1936. After spending three years at the Institute for Advanced Study in Princeton as an assistant of O. Veblen, Givens accepted an appointment at Cornell University but later moved to Northwestern University. In addition to his academic career, Givens was the Director of the Applied Mathematics Division at Argonne National Laboratory and, like his counterpart A. S. Householder (p. 324) at Oak Ridge National Laboratory, Givens served as an early president of SIAM. 334 Chapter 5 Norms, Inner Products, and Orthogonality then ⎛ x1 .. . ⎞ x2 + x2 ⎞←− i ⎞ i j ⎞ .. ⎞ Pij x = ⎞ . ⎞. ⎞←− j ⎞ 0 ⎞ .. ⎠ . xn This means that we can selectively annihilate any component—the j th in this case—by a rotation in the (i, j)-plane without affecting any entry except xi and xj . Consequently, plane rotations can be applied to annihilate all components below any particular "pivot." For example, to annihilate all entries below the first position in x, apply a sequence of plane rotations as follows: ⎛√ ⎞ P12 x = ⎞ ⎛ 2 x2 1 +x2 0 x3 x4 .. . xn ⎞ ⎛√ ⎞ ⎞ ⎞ P13 P12 x = ⎞ ⎞ ⎛ 2 2 x2 1 +x2 +x3 0 0 x4 ... xn ⎞ ⎞. . . , ⎞ ⎞ x ⎞ 0 ⎞ ⎞P1n · · ·P13 P12 x = ⎞ 0. ⎞ ⎝ . ⎠ . 0 The product of plane rotations is generally not another plane rotation, but such a product is always an orthogonal matrix, and hence it is an isometry. If we are willing to interpret "rotation in n " as a sequence of plane rotations, then we can say that it is always possible to "rotate" each nonzero vector onto the first coordinate axis. Recall from (5.6.11) that we can also do this with a reflection. More generally, the following statement is true. Rotations in n Every nonzero vector x ∈ n can be rotated to the ith coordinate axis by a sequence of n − 1 plane rotations. In other words, there is an orthogonal matrix P such that Px = x ei , where P has the form P = Pin · · · Pi,i+1 Pi,i−1 · · · Pi1 . (5.6.17) 5.6 Unitary and Orthogonal Matrices 335 Example 5.6.6 Problem: If x ∈ n is a vector such that x = 1, explain how to use plane rotations to construct an orthonormal basis for n that contains x. Solution: This is almost the same problem as that posed in Example 5.6.3, and as explained there, the goal is to construct an orthogonal matrix Q such that Q∗1 = x. But this time we need to use plane rotations rather than an elementary reflector. Equation (5.6.17) asserts that we can build an orthogonal matrix from a sequence of plane rotations P = P1n · · · P13 P12 such that Px = e1 . Thus x = PT e1 = PT∗1 , and hence the columns of Q = PT serve the purpose. For example, to extend ⎛ −1 1 ⎞ 2 ⎞x= ⎝ 0 3 −2 4 to an orthonormal basis for 4 , sequentially annihilate the second and fourth components of x by using (5.6.16) to construct the following plane rotations: √ √ ⎛ ⎛ ⎛√ 2/√5 0 0 5 −1/√5 −1 ⎞ −2/5 −1/5 0 0 ⎞1 ⎞ 2 ⎞1 ⎞ 0 ⎞P12 x = ⎝ ⎠, ⎞ ⎞= ⎝ 0 0 0 1 0 3 0 3 −2 0 0 0 1 −2 ⎛√ ⎞⎛√ ⎞ ⎛ 5/3 0 0 −2/3 5 1 1 0 0 ⎞1 ⎞ 0 ⎞0 ⎞ 0 P14 P12 x = ⎝ ⎠ = ⎝ ⎠ 0 0 1 √0 0 0 3 2/3 0 0 −2 0 5/3 Therefore, the columns of √ ⎛ ⎛ −1/3 −2/√5 0 −2/3√5 4/3 5 ⎞ 2/3 −1/5 0 T Q = (P14 P12) = PT12 PT14 = ⎝ 0 0 1 √0 −2/3 0 0 5/3 are an orthonormal set containing the specified vector x. Exercises for section 5.6 5.6.1. Determine which of the following matrices are isometries. √ ⎛ √ ⎞ 1/√2 −1/√2 0√ 1 0 1 (a) ⎝ 1/√6 1/√6 −2/√6 ⎠. (b) ⎝ 1 0 −1 ⎠. 1/3 1/3 1/3 3 0 1 0 ⎛ ⎞ eiθ1 0 ··· 0 0 0 1 0 ⎞ 0 eiθ2 · · · 0 ⎞1 0 0 0 ⎞(c) ⎞ (d) ⎞ .. . . ⎞. ⎝ .. . ⎠ 0 0 0 1 . . . ⎠0 1 0 0 0 0 · · · eiθn 336 Chapter 5 Norms, Inner Products, and Orthogonality ⎛ 1+i √ ⎞ 3 5.6.2. Is ⎞ ⎝ i √ 3 ⎠1+i √ 6 ⎞a unitary matrix? −2 i ⎠√ 6 5.6.3. (a) How many 3 × 3 matrices are both diagonal and orthogonal? (b) How many n × n matrices are both diagonal and orthogonal? (c) How many n × n matrices are both diagonal and unitary? 5.6.4. (a) Under what conditions on the real numbers α and β will α+β β−α P= α−β β+α be an orthogonal matrix? (b) Under what conditions on the real numbers α and β will ⎛ 0 α 0 iβ ⎞ α 0 iβ 0 ⎞U=⎝ 0 iβ 0 α iβ 0 α 0 be a unitary matrix? 5.6.5. Let U (a) (b) (c) and V be two n × n unitary (orthogonal) matrices. Explain why the product UV must be unitary (orthogonal). Explain why the sum U + V need not be unitary (orthogonal). 0 Explain why Un×n must be unitary (orthogonal). 0 V m×m 5.6.6. Cayley Transformation. Prove, as Cayley did in 1846, that if A is skew hermitian (or real skew symmetric), then U = (I − A)(I + A)−1 = (I + A)−1 (I − A) is unitary (orthogonal) by first showing that (I + A)−1 exists for skewhermitian matrices, and (I − A)(I + A)−1 = (I + A)−1 (I − A) (recall Exercise 3.7.6). Note: There is a more direct approach, but it requires the diagonalization theorem for normal matrices—see Exercise 7.5.5. 5.6.7. Suppose that R and S are elementary reflectors. 0 (a) Is 0I R an elementary reflector? 0 (b) Is R an elementary reflector? 0 S 5.6 Unitary and Orthogonal Matrices 5.6.8. 337 (a) Explain why the standard inner product is invariant under a unitary transformation. That is, if U is any unitary matrix, and if u = Ux and v = Uy, then u∗ v = x∗ y. (b) Given any two vectors x, y ∈ n , explain why the angle between them is invariant under an orthogonal transformation. That is, if u = Px and v = Py, where P is an orthogonal matrix, then cos θu,v = cos θx,y . 5.6.9. Let Um×r be a matrix with orthonormal columns, and let Vk×n be a matrix with orthonormal rows. For an arbitrary A ∈ C r×k , solve the following problems using the matrix 2-norm (p. 281) and the Frobenius matrix norm (p. 279). (a) Determine the values of U2 , V2 , UF , and VF . (b) Show that UAV2 = A2 . (Hint: Start with UA2 .) (c) Show that UAVF = AF . Note: In particular, these properties are valid when U and V are unitary matrices. Because of parts (b) and (c), the 2-norm and the F norm are said to be unitarily invariant norms. ⎛ 5.6.10. Let u = (a) (b) (c) (d) ⎞−2 ⎞ 1 ⎞3 −1 ⎛ and v = Determine Determine Determine Determine the the the the ⎞ 1 ⎛ 4 ⎠. 0 −1 orthogonal orthogonal orthogonal orthogonal projection projection projection projection of of of of u v u v onto onto onto onto span {v} . span {u} . v⊥ . u⊥ . 5.6.11. Consider elementary orthogonal projectors Q = I − uu∗ . (a) Prove that Q is singular. (b) Now prove that if Q is n × n, then rank (Q) = n − 1. Hint: Recall Exercise 4.4.10. 5.6.12. For vectors u, x ∈ C n such that u = 1, let p be the orthogonal projection of x onto span {u} . Explain why p ≤ x with equality holding if and only if x is a scalar multiple of u. 338 Chapter 5 Norms, Inner Products, and Orthogonality 1 5.6.13. Let x = (1/3) −2 . −2 (a) Determine an elementary reflector R such that Rx lies on the x-axis. (b) Verify by direct computation that your reflector R is symmetric, orthogonal, and involutory. (c) Extend x to an orthonormal basis for 3 by using an elementary reflector. 5.6.14. Let R = I − 2uu∗ , where un×1 = 1. If x is a fixed point for R in the sense that Rx = x, and if n > 1, prove that x must be orthogonal to u, and then sketch a picture of this situation in 3 . 5.6.15. Let x, y ∈ n×1 be vectors such that x = y but x = y. Explain how to construct an elementary reflector R such that Rx = y. Hint: The vector u that defines R can be determined visually in 3 by considering Figure 5.6.2. 5.6.16. Let xn×1 be a vector such that x = 1, and partition x as x= x1 ˜ x , ˜ is n − 1 × 1. where x (a) If the entries of x are real, and if x1 = 1, show that P= x1 ˜ x x ˜T I − α˜ xx ˜T , where α= 1 1 − x1 is an orthogonal matrix. (b) Suppose that the entries of x are complex. If |x1 | = 1, and if μ is the number defined in (5.6.10), show that the matrix U= x1 ˜ x μ2 x ˜∗ μ(I − α˜ xx ˜∗) , where α= 1 1 − |x1 | is unitary. Note: These results provide an easy way to extend a given vector to an orthonormal basis for the entire space n or C n . 5.6 Unitary and Orthogonal Matrices 339 5.6.17. Perform the following sequence of rotations in 3 beginning with ⎛ ⎞ 1 v0 = ⎝ 1 ⎠. −1 1. Rotate v0 counterclockwise 45◦ around the x-axis to produce v1 . 2. Rotate v1 clockwise 90◦ around the y-axis to produce v2 . 3. Rotate v2 counterclockwise 30◦ around the z-axis to produce v3 . Determine the coordinates of v3 as well as an orthogonal matrix Q such that Qv0 = v3 . 5.6.18. Does it matter in what order rotations in 3 are performed? For example, suppose that a vector v ∈ 3 is first rotated counterclockwise around the x-axis through an angle θ, and then that vector is rotated counterclockwise around the y-axis through an angle φ. Is the result the same as first rotating v counterclockwise around the y-axis through an angle φ followed by a rotation counterclockwise around the x-axis through an angle θ? 5.6.19. For each nonzero vector u ∈ C n , prove that dim u⊥ = n − 1. 5.6.20. A matrix satisfying A2 = I is said to be an involution or an involutory matrix , and a matrix P satisfying P2 = P is called a projector or is said to be an idempotent matrix —properties of such matrices are developed on p. 386. Show that there is a one-to-one correspondence between the set of involutions and the set of projectors in C n×n . Hint: Consider the relationship between the projectors in (5.6.6) and the reflectors (which are involutions) in (5.6.7) on p. 324. 5.6.21. When using a computer to generate and display a three-dimensional convex polytope such as the one in Example 5.6.4, it is desirable to not draw those faces that should be hidden from the perspective of a viewer positioned as shown in Figure 5.6.6. The operation of cross product in 3 (usually introduced in elementary calculus courses) can be used to decide which faces are visible and which are not. Recall that if ⎛ ⎞⎛ ⎞ u1 v1 u2 v3 − u3 v2 u = ⎝ u2 ⎠and v = ⎝ v2 ⎠, then u × v = ⎝ u3 v1 − u1 v3 ⎠, u3 v3 u1 v2 − u2 v1 340 Chapter 5 Norms, Inner Products, and Orthogonality and u × v is a vector orthogonal to both u and v. The direction of u × v is determined from the so-called right-hand rule as illustrated in Figure 5.6.8. Figure 5.6.8 Assume the origin is interior to the polytope, and consider a particular face and three vertices p0 , p1 , and p2 on the face that are positioned as shown in Figure 5.6.9. The vector n = (p1 − p0) × (p2 − p1) is orthogonal to the face, and it points in the outward direction. Figure 5.6.9 Explain why the outside of the face is visible from the perspective indicated in Figure 5.6.6 if and only if the first component of the outward normal vector n is positive. In other words, the face is drawn if and only if n1 > 0. 5.7 Orthogonal Reduction 5.7 341 ORTHOGONAL REDUCTION We know that a matrix A can be reduced to row echelon form by elementary row operations. This is Gaussian elimination, and, as explained on p. 143, the basic "Gaussian transformation" is an elementary lower triangular matrix Tk whose action annihilates all entries below the k th pivot at the k th elimination step. But Gaussian elimination is not the only way to reduce a matrix. Elementary reflectors Rk can be used in place of elementary lower triangular matrices Tk to annihilate all entries below the k th pivot at the k th elimination step, or a sequence of plane rotation matrices can accomplish the same purpose. When reflectors are used, the process is usually called Householder reduction, and it proceeds as follows. For Am×n = [A∗1 | A∗2 | · · · | A∗n] , use x = A∗1 in (5.6.10) to construct the elementary reflector R1 = I − 2 uu∗ , u∗ u so that where u = A∗1 ± μ A∗1 e1 , ⎞t11 ⎞ 0 ⎞ ⎞= ∓μ A∗1 e1 = ⎞ ⎝ ... ⎠. (5.7.1) ⎛ R1 A∗1 (5.7.2) 0 Applying R1 to A yields ⎛ ⎞ R1 A=[R1 A∗1 | R1 A∗2 | · · · | R1 A∗n]= ⎞ ⎝ · · · t1n ⎞t11 t12 0 .. ∗ .. . ··· ⎞ ∗ ⎞= t11 .. ⎞0 . ⎠0 ∗ ··· ∗ tT1 , A2 where A2 is m − 1 × n − 1. Thus all entries below the (1, 1)-position are annihilated. Now apply the same procedure to A2 to construct an elementary ˆ 2 that annihilates all entries below the (1, 1)-position in A2 . If we reflector R 1 0 set R2 = 0 R ˆ 2 , then R2 R1 is an orthogonal matrix (Exercise 5.6.5) such that ⎛ ⎞t12 t13 · · · t1n t11 ⎞ ⎞ 0 t22 t23 · · · t2n ⎞T ⎞ ⎞t11 t1 ⎞ R2 R1 A = =⎞ ˆ ⎞ ⎞0 0 ∗ · · · ∗ 0 R2 A2 ⎞ . ⎞. . . ⎞. .. .. ⎠0 0 ∗ ··· ∗ ˜ k−1 T The result after k − 1 steps is Rk−1 · · · R2 R1 A = Tk−1 . At step 0 Ak ˆ k is constructed in a manner similar to (5.7.1) k an elementary reflector R 342 Chapter 5 Norms, Inner Products, and Orthogonality to annihilate below the (1, 1)-position in Ak , and Rk is defined all entries Ik−1 0 as Rk = ˆ k , which is another elementary reflector (Exercise 5.6.7). 0 R Eventually, all of the rows or all of the columns will be exhausted, so the final result is one of the two following upper-trapezoidal forms: ⎞ ··· ∗⎞ ⎞ ⎞ ··· ∗ ⎞ ⎞0 ⎞. .. ⎞.. n×n⎞ .. ⎞. . ⎞ ⎞ ⎞ ⎞ = ⎞0 0 ··· ∗ ⎞ ⎞ ⎞0 0 ··· 0 ⎞⎞. . ⎞. ⎞ . . 0 0 ··· 0 ⎛∗ Rn · · · R2 R1 Am×n ∗ ∗ ⎞Rm−1 · · · R2 R1 Am×n ∗ ∗ ··· ∗ ⎞0 ∗ ··· ∗ =⎞ . . ⎞. ... . . 0 0 ··· ∗) ∗+ , when ∗ ∗ .. . ··· .. ⎞ ∗ ∗ ⎞ .. ⎞. ⎠ ∗ ··· ∗ m > n, when m < n. m×m If m = n, then the final form is an upper-triangular matrix. A product of elementary reflectors is not necessarily another elementary reflector, but a product of unitary (orthogonal) matrices is again unitary (orthogonal) (Exercise 5.6.5). The elementary reflectors Ri described above are unitary (orthogonal in the real case) matrices, so every product Rk Rk−1 · · · R2 R1 is a unitary matrix, and thus we arrive at the following important conclusion. Orthogonal Reduction • For every A ∈ C m×n , there exists a unitary matrix P such that PA = T (5.7.3) has an upper-trapezoidal form. When P is constructed as a product of elementary reflectors as described above, the process is called Householder reduction. • If A is square, then T is upper triangular, and if A is real, then the P can be taken to be an orthogonal matrix. 5.7 Orthogonal Reduction 343 Example 5.7.1 Problem: Use Householder reduction to find an orthogonal matrix P such that PA = T is upper triangular with positive diagonal entries, where ⎛ ⎞0 −20 −14 A = ⎝3 27 −4 ⎠. 4 11 −2 Solution: To annihilate the entries below the (1, 1)-position and to guarantee that t11 is positive, equations (5.7.1) and (5.7.2) dictate that we set ⎛ ⎞−5 u1 uT u1 = A∗1 − A∗1 e1 = A∗1 − 5e1 = ⎝ 3 ⎠and R1 = I − 2 T 1 . u1 u1 4 To compute a reflector-by-matrix product RA = [RA∗1 | RA∗2 | · · · | RA∗n] , it's wasted effort to actually determine the entries in R = I−2uuT /uT u. Simply compute uT A∗j and then T u A∗j RA∗j = A∗j − 2 u for each j = 1, 2, . . . , n. (5.7.4) uT u By using this observation we obtain ⎞ 25 −4 5 R1 A = [R1 A∗1 | R1 A∗2 | R1 A∗3] = ⎝ 0 0 −10 ⎠. 0 −25 −10 To annihilate the entry below the (2, 2)-position, set 0 −10 −1 A2 = and u2 = [A2]∗1 − [A2]∗1 e1 = 25 . −25 −10 −1 ˆ 2 = I − 2u2 uT /uT u2 and R2 = 1 ˆ0 (neither is explicitly computed), If R 2 2 0 R2 then ⎛ ⎞ 5 25 −4 25 10 ˆ 2 A2 = R and R2 R1 A = T = ⎝ 0 25 10 ⎠. 0 10 0 0 10 ˆ k = I − 2ˆ ˆ T /ˆ ˆ is an elementary reflector, then so is If R uu uT u uuT I 0 Rk = , ˆ k = I − 2 uT u with u = u ˆ 0 R and consequently the product of any sequence of these Rk 's can be formed by using the observation (5.7.4). In this example, ⎛ ⎞0 15 20 1 ⎞ P = R2 R1 = −20 12 −9 ⎠. 25 −15 −16 12 You may wish to check that P really is an orthogonal matrix and PA = T. 344 Chapter 5 Norms, Inner Products, and Orthogonality Elementary reflectors are not the only type of orthogonal matrices that can be used to reduce a matrix to an upper-trapezoidal form. Plane rotation matrices are also orthogonal, and, as explained on p. 334, plane rotation matrices can be used to selectively annihilate any component in a given column, so a sequence of plane rotations can be used to annihilate all elements below a particular pivot. This means that a matrix A ∈ m×n can be reduced to an upper-trapezoidal form strictly by using plane rotations—such a process is usually called a Givens reduction. Example 5.7.2 Problem: Use Givens reduction (i.e., use plane rotations) to reduce the matrix ⎛ ⎞0 −20 −14 A = ⎝3 27 −4 ⎠4 11 −2 to upper-triangular form. Also compute an orthogonal matrix P such that PA = T is upper triangular. Solution: The plane rotation that uses the (1,1)-entry to annihilate the (2,1)entry is determined from (5.6.16) to be ⎛ ⎞⎛ ⎞0 1 0 3 27 −4 P12 = ⎝ −1 0 0 ⎠so that P12 A = ⎝ 0 20 14 ⎠. 0 0 1 4 11 −2 Now use the (1,1)-entry in P12 A to annihilate the (3,1)-entry in P12 A. The job is again obtained from (5.6.16) to be ⎛ ⎞ 3 0 1⎛ P13 = 0 5 5 −4 0 in P12 A to annihilate the (3,1)-entry in P12 A. The the job is again obtained from (5.6.16) to be ⎛ ⎞ 5 25 −4 4 20 14 ⎠. 0 ⎠so that P13 P12 A = ⎝ 0 0 −15 2 3 Finally, using the (2,2)-entry in P13 P12 A to annihilate the (3,2)-entry produces ⎛ ⎞⎛ ⎞ 5 25 −4 5 0 0 1⎛ P23 = 0 4 −3 ⎠so that P23 P13 P12 A = T = ⎝ 0 25 10 ⎠. 5 0 0 10 0 3 4 Since plane rotation matrices are orthogonal, and since the product of orthogonal matrices is again orthogonal, it must be the case that ⎛ ⎞0 15 20 1 ⎞ P = P23 P13 P12 = −20 12 −9 ⎠25 −15 −16 12 is an orthogonal matrix such that PA = T. 5.7 Orthogonal Reduction 345 Householder and Givens reductions are closely related to the results produced by applying the Gram–Schmidt process (p. 307) to the columns of A. When A is nonsingular, Householder, Givens, and Gram–Schmidt each produce an orthogonal matrix Q and an upper-triangular matrix R such that A = QR (Q = PT in the case of orthogonal reduction). The upper-triangular matrix R produced by the Gram–Schmidt algorithm has positive diagonal entries, and, as illustrated in Examples 5.7.1 and 5.7.2, we can also force this to be true using the Householder or Givens reduction. This feature makes Q and R unique. QR Factorization For each nonsingular A ∈ n×n , there is a unique orthogonal matrix Q and a unique upper-triangular matrix R with positive diagonal entries such that A = QR. This "square" QR factorization is a special case of the more general "rectangular" QR factorization discussed on p. 311. Proof. Only uniqueness needs to be proven. If there are two QR factorizations A = Q1 R1 = Q2 R2 , −1 let U = QT2 Q1 = R2 R−1 is upper triangular with positive 1 . The matrix R2 R1 diagonal entries (Exercises 3.5.8 and 3.7.4) while QT2 Q1 is an orthogonal matrix (Exercise 5.6.5), and therefore U is an upper-triangular matrix whose columns are an orthonormal set and whose diagonal entries are positive. Considering the first column of U we see that ⎛ ⎞u11 ⎞ 0 ⎞⎞ . ⎞= 1 =⇒ u11 = ±1 and u11 > 0 =⇒ u11 = 1, ⎝ .. ⎠0 so that U∗1 = e1 . A similar argument together with the fact that the columns of U are mutually orthogonal produces UT∗1 U∗2 = 0 =⇒ u12 = 0 =⇒ u22 = 1 =⇒ U∗2 = e2 . Proceeding inductively establishes that U∗k = ek for each k (i.e., U = I), and therefore Q1 = Q2 and R1 = R2 . 346 Chapter 5 Norms, Inner Products, and Orthogonality Example 5.7.3 Orthogonal Reduction and Least Squares. Orthogonal reduction can be used to solve the least squares problem associated with an inconsistent system Ax = b in which A ∈ m×n and m ≥ n (the most common case). If ε denotes the difference ε = Ax − b, then, as described on p. 226, the general least squares problem is to find a vector x that minimizes the quantity m 2 ε2i = εT ε = ε , i=1 where is the standard euclidean vector norm. Suppose that A is reduced to an upper-trapezoidal matrix T by an orthogonal matrix P, and write Rn×n cn×1 PA = T = and Pb = 0 d in which R is an upper-triangular matrix. An orthogonal matrix is an isometry— recall (5.6.1)—so that 2 R Rx − c 2 c 2 2 2 ε = Pε = P(Ax − b) = x− = 0 d d 2 2 = Rx − c + d . 2 2 Consequently, ε is minimized when x is a vector such that Rx − c is minimal or, in other words, x is a least squares solution for Ax = b if and only if x is a least squares solution for Rx = c. Full-Rank Case. In a majority of applications the coefficient matrix A has linearly independent columns so rank (Am×n) = n. Because multiplication by a nonsingular matrix P does not change the rank, n = rank (A) = rank (PA) = rank (T) = rank (Rn×n). Thus R is nonsingular, and we have established the following fact. • If A has linearly independent columns, then the (unique) least squares solution for Ax = b is obtained by solving the nonsingular triangular system Rx = c for x. As pointed out in Example 4.5.1, computing the matrix product AT A is to be avoided when floating-point computation is used because of the possible loss of significant information. Notice that the method based on orthogonal reduction sidesteps this potential problem because the normal equations AT Ax = AT b are avoided and the product AT A is never explicitly computed. Householder reduction (or Givens reduction for sparse problems) is a numerically stable algorithm (see the discussion following this example) for solving the full-rank least squares problem, and, if the computations are properly ordered, it is an attractive alternative to the method of Example 5.5.3 that is based on the modified Gram–Schmidt procedure. 5.7 Orthogonal Reduction 347 We now have four different ways to reduce a matrix to an upper-triangular (or trapezoidal) form. (1) Gaussian elimination; (2) Gram–Schmidt procedure; (3) Householder reduction; and (4) Givens reduction. It's natural to try to compare them and to sort out the advantages and disadvantages of each. First consider numerical stability. This is a complicated issue, but you can nevertheless gain an intuitive feel for the situation by considering the effect of applying a sequence of "elementary reduction" matrices to a small perturbation of A. Let E be a matrix such that EF is small relative to AF (the Frobenius norm was introduced on p. 279), and consider Pk · · · P2 P1 (A + E) = (Pk · · · P2 P1 A) + (Pk · · · P2 P1 E) = PA + PE. If each Pi is an orthogonal matrix, then the product P = Pk · · · P2 P1 is also an orthogonal matrix (Exercise 5.6.5), and consequently PEF = EF (Exercise 5.6.9). In other words, a sequence of orthogonal transformations cannot magnify the magnitude of E, and you might think of E as representing the effects of roundoff error. This suggests that Householder and Givens reductions should be numerically stable algorithms. On the other hand, if the Pi 's are elementary matrices of Type I, II, or III, then the product P = Pk · · · P2 P1 can be any nonsingular matrix—recall (3.9.3). Nonsingular matrices are not generally norm preserving (i.e., it is possible that PEF > EF), so the possibility of E being magnified is generally present in elimination methods, and this suggests the possibility of numerical instability. Strictly speaking, an algorithm is considered to be numerically stable if, under floating-point arithmetic, it always returns an answer that is the exact solution of a nearby problem. To give an intuitive argument that the Householder or Givens reduction is a stable algorithm for producing the QR factorization of An×n , suppose that Q and R are the exact QR factors, and suppose that floating-point arithmetic produces an orthogonal matrix Q + E and an uppertriangular matrix R + F that are the exact QR factors of a different matrix ˜ = (Q + E)(R + F) = QR + QF + ER + EF = A + QF + ER + EF. A If E and F account for the roundoff errors, and if their entries are small relative to those in A, then the entries in EF are negligible, and ˜ ≈ A + QF + ER. A But since Q is orthogonal, QFF = FF and AF = QRF = RF , and this means that neither QF nor ER can contain entries that are large ˜ ≈ A, and this is what is required to conclude relative to those in A. Hence A that the algorithm is stable. Gaussian elimination is not a stable algorithm because, as alluded to in §1.5, problems arise due to the growth of the magnitude of the numbers that can occur 348 Chapter 5 Norms, Inner Products, and Orthogonality during the process. To see this from a heuristic point of view, consider the LU factorization of A = LU, and suppose that floating-point Gaussian elimination with no pivoting returns matrices L + E and U + F that are the exact LU factors of a somewhat different matrix ˜ = (L + E)(U + F) = LU + LF + EU + EF = A + LF + EU + EF. A If E and F account for the roundoff errors, and if their entries are small relative to those in A, then the entries in EF are negligible, and ˜ ≈ A + LF + EU A (using no pivoting). However, if L or U contains entries that are large relative to those in A (and this is certainly possible), then LF or EU can contain entries that are significant. In other words, Gaussian elimination with no pivoting can return the ˜ that is not very close to the original matrix LU factorization of a matrix A A, and this is what it means to say that an algorithm is unstable. We saw on p. 26 that if partial pivoting is employed, then no multiplier can exceed 1 in magnitude, and hence no entry of L can be greater than 1 in magnitude (recall that the subdiagonal entries of L are in fact the multipliers). Consequently, L cannot greatly magnify the entries of F, so, if the rows of A have been reordered according to the partial pivoting strategy, then ˜ ≈ A + EU A (using partial pivoting). ˜ ≈ A, so the issue boils down to the degree Numerical stability requires that A to which U magnifies the entries in E —i.e., the issue rests on the magnitude of the entries in U. Unfortunately, partial pivoting may not be enough to control the growth of all entries in U. For example, when Gaussian elimination with partial pivoting is applied to ⎛ ⎞1 0 0 ··· 0 0 1 ⎞ −1 1 0 ··· 0 0 1 ⎞ ⎞.. ⎞ ⎞. ⎞ −1 −1 1 0 0 1 ⎞ .. .. ⎞.. ⎞Wn = ⎞ . . . . . .⎞ .. ⎞ .. ⎞ −1 −1 −1 . . . 1 0 1 ⎞⎞ ⎞ −1 −1 −1 · · · −1 1 1 ⎠−1 −1 −1 ··· −1 −1 1 the largest entry in U is unn = 2n−1 . However, if complete pivoting is used on Wn , then no entry in the process exceeds 2 in magnitude (Exercises 1.5.7 and 1.5.8). In general, it has been proven that if complete pivoting is used on a wellscaled matrix An×n for which max |aij | = 1, then no entry of U can exceed 5.7 Orthogonal Reduction 349 1/2 γ = n1/2 21 31/2 41/3 · · · n1/n−1 in magnitude. Since γ is a slow growing function of n, the entries in U won't greatly magnify the entries of E, so ˜ ≈A A (using complete pivoting). In other words, Gaussian elimination with complete pivoting is stable, but Gaussian elimination with partial pivoting is not. Fortunately, in practical work it is rare to encounter problems such as the matrix Wn in which partial pivoting fails to control the growth in the U factor, so scaled partial pivoting is generally considered to be a "practically stable" algorithm. Algorithms based on the Gram–Schmidt procedure are more complicated. First, the Gram–Schmidt algorithms differ from Householder and Givens reductions in that the Gram–Schmidt procedures are not a sequential application of elementary orthogonal transformations. Second, as an algorithm to produce the QR factorization even the modified Gram–Schmidt technique can return a Q factor that is far from being orthogonal, and the intuitive stability argument used earlier is not valid. As an algorithm to return the QR factorization of A, the modified Gram–Schmidt procedure has been proven to be unstable, but as an algorithm used to solve the least squares problem (see Example 5.5.3), it is stable—i.e., stability of modified Gram–Schmidt is problem dependent. Summary of Numerical Stability • Gaussian elimination with scaled partial pivoting is theoretically unstable, but it is "practically stable"—i.e., stable for most practical problems. • Complete pivoting makes Gaussian elimination unconditionally stable. For the QR factorization, the Gram–Schmidt procedure (classical or modified) is not stable. However, the modified Gram–Schmidt procedure is a stable algorithm for solving the least squares problem. Householder and Givens reductions are unconditionally stable algorithms for computing the QR factorization. • • For the algorithms under consideration, the number of multiplicative operations is about the same as the number of additive operations, so computational effort is gauged by counting only multiplicative operations. For the sake of comparison, lower-order terms are not significant, and when they are neglected the following approximations are obtained. 350 Chapter 5 Norms, Inner Products, and Orthogonality Summary of Computational Effort The approximate number of multiplications/divisions required to reduce an n × n matrix to an upper-triangular form is as follows. • Gaussian elimination (scaled partial pivoting) ≈ n3 /3. • Gram–Schmidt procedure (classical and modified) ≈ n3 . • Householder reduction ≈ 2n3 /3. • Givens reduction ≈ 4n3 /3. It's not surprising that the unconditionally stable methods tend to be more costly—there is no free lunch. No one triangularization technique can be considered optimal, and each has found a place in practical work. For example, in solving unstructured linear systems, the probability of Gaussian elimination with scaled partial pivoting failing is not high enough to justify the higher cost of using the safer Householder or Givens reduction, or even complete pivoting. Although much the same is true for the full-rank least squares problem, Householder reduction or modified Gram–Schmidt is frequently used as a safeguard against sensitivities that often accompany least squares problems. For the purpose of computing an orthonormal basis for R (A) in which A is unstructured and dense (not many zeros), Householder reduction is preferred—the Gram–Schmidt procedures are unstable for this purpose and Givens reduction is too costly. Givens reduction is useful when the matrix being reduced is highly structured or sparse (many zeros). Example 5.7.4 Reduction to Hessenberg Form. For reasons alluded to in §4.8 and §4.9, it is often desirable to triangularize a square matrix A by means of a similarity transformation—i.e., find a nonsingular matrix P such that P−1 AP = T is upper triangular. But this is a computationally difficult task, so we will try to do the next best thing, which is to find a similarity transformation that will reduce A to a matrix in which all entries below the first subdiagonal are zero. Such a matrix is said to be in upper-Hessenberg form—illustrated below is a 5 × 5 Hessenberg form. ⎞⎛ ∗ ∗ ∗ ∗ ∗⎞ ∗ ∗ ∗ ∗ ∗ ⎞ ⎞⎞ H = ⎞0 ∗ ∗ ∗ ∗⎞ ⎝ ⎠⎝ 0 0 ∗ ∗ ∗ 0 0 0 ∗ ∗ 5.7 Orthogonal Reduction 351 Problem: Reduce A ∈ n×n to upper-Hessenberg form by means of an orthogonal similarity transformation—i.e., construct an orthogonal matrix P such that PT AP = H is upper Hessenberg. Solution: At each step, use Householder reduction on entries below the main ˆ ∗1 denote the entries of the first column that are diagonal. Begin by letting A below the (1,1)-position—this is illustrated below for n = 5 : ⎛ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ⎞∗ ⎞ A=⎞ ⎞∗ ⎞⎝∗ ⎠∗ ⎞ ∗ ⎞ ⎞∗ ⎞ ⎞= ∗⎠∗ a11 ˆ 1∗ A ˆ ∗1 A A1 . ˆ 1 is an elementary reflector determined according to (5.7.1) for which If R ⎛ ⎞∗ 0 ˆ ˆ ⎝ R1 A∗1 = 00 ⎠, then R1 = 01 R1 is an orthogonal matrix such that ˆ 1 0 R1 AR1 = 0 = 1 0 ˆ1 R a11 ˆ ˆ ∗1 R1 A a11 ˆ ∗1 A ˆ 1∗ A A1 ⎛ ˆ 1∗ R ˆ1 A ˆ 1 A1 R ˆ1 R ˆ1 R 1 0 ˆ1 0 R ∗ ∗ ∗ ∗ 0 ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ∗ ⎞∗ ⎞ =⎞ ⎞0 ⎝0 ⎠∗ ⎞ ⎞∗ ⎞ ⎞∗ ⎞ ⎠∗ ˆ ˆ 1 A1 R At the second step, repeat the process on A2 = R ⎞1 to obtain an orthogo∗ ∗ ∗ I 0 ˆ 2 such that R ˆ2 = ⎞ ˆ 2 A2 R ∗ ∗ ∗ ⎞nal matrix R ⎠. Matrix R2 = 02 R ⎞∗ ˆ 0 0 ∗ ∗ ∗ ∗ ∗ ∗ 2 is an orthogonal matrix such that ⎛ ∗ ⎞∗ ⎞ ⎞ R2 R1 AR1 R2 = ⎞ 0 ⎞ ⎝0 0 ∗ ∗ ∗ ∗ ∗ ∗ ∗ 0 0 ∗ ∗ ∗ ∗ ∗ ∗ ⎞ ∗ ⎞ ⎞ ⎞. ∗ ⎞ ⎞∗ ⎠∗ After n − 2 of these steps, the product P = R1 R2 · · · Rn−2 is an orthogonal matrix such that PT AP = H is in upper-Hessenberg form. 352 Chapter 5 Norms, Inner Products, and Orthogonality Note: If A is a symmetric matrix, then HT = (PT AP)T = PT AT P = H, so H is symmetric. But as illustrated below for n = 5, a symmetric Hessenberg form is a tridiagonal matrix, ⎛ ∗ ⎞∗ ⎞ H = PT AP = ⎞ 0 ⎞ ⎝ 0 0 ∗ 0 ∗ ∗ ∗ 0 ∗ 0 0 ⎞ 0 0 0 0 ⎞ ⎞ ∗ 0 ⎞, ⎠ ∗ ∗ ∗ ∗ so the following useful corollary is obtained. • Every real-symmetric matrix is orthogonally similar to a tridiagonal matrix, and Householder reduction can be used to compute this tridiagonal matrix. However, the Lanczos technique discussed on p. 651 can be much more efficient. Example 5.7.5 Problem: Compute the QR factors of a nonsingular upper-Hessenberg matrix H ∈ n×n . Solution: Due to its smaller multiplication count, Householder reduction is generally preferred over Givens reduction. The exception is for matrices that have a zero pattern that can be exploited by the Givens method but not by the Householder method. A Hessenberg matrix H is such an example. The first step of Householder reduction completely destroys most of the zeros in H, but applying plane rotations does not. This is illustrated below for a 5 × 5 Hessenberg form—remember that the action of Pk,k+1 affects only the k th and (k + 1)st rows. ⎛ ∗ ⎞∗ ⎞ ⎞0 ⎝ 0 0 ∗ ∗ ∗ 0 0 ⎞ ⎛ ∗ ∗ ∗ ∗ ∗ ∗ ⎞ ⎞0 ⎞P12 ⎞ ∗ ∗ ∗ ⎞ −−−→ ⎞0 ⎠ ⎝ ∗ ∗ ∗ 0 0 ∗ ∗ ∗ 0 ⎛ ∗ ⎞0 P34 ⎞ −−−→ ⎞0 ⎝ 0 0 ∗ ∗ ∗ 0 0 ∗ ∗ ∗ ∗ ∗ 0 ∗ ∗ ∗ ∗ ∗ ∗ 0 0 0 ∗ ∗ 0 0 ∗ ∗ ∗ ∗ ∗ ⎞ ∗ ∗ ∗ ⎞ ⎞0 ⎞P23 ⎞ ∗ ⎞ −−−→ ⎞0 ⎝ ⎠ ∗ 0 ∗ 0 ⎞ ∗ ∗ ∗ ⎞ ⎞0 ⎞P45 ⎞ ∗ ⎞ −−−→ ⎞0 ⎝ ⎠ ∗ 0 ∗ 0 ∗ 0 ∗ 0 0 0 ∗ ∗ ∗ 0 0 ∗ ∗ 0 0 0 ∗ ∗ ∗ 0 0 ⎞ ∗ ∗ ⎞ ∗ ⎞ ⎞ ∗ ∗ ⎞ ⎠ ∗ ∗ ∗ ∗ ⎞ ∗ ∗ ⎞ ∗ ⎞ ⎠ ∗ ∗ 0 ∗ ∗ ∗ ∗ ∗ ∗ ∗ In general, Pn−1,n · · · P23 P12 H = R is upper triangular in which all diagonal entries, except possibly the last, are positive—the last diagonal can be made positive by the technique illustrated in Example 5.7.2. Thus we obtain an orthogonal matrix P such that PH = R, or H = QR in which Q = PT . 5.7 Orthogonal Reduction 353 Example 5.7.6 49 Jacobi Reduction. Given a real-symmetric matrix A, the result of Example 5.7.4 shows that Householder reduction can be used to construct an orthogonal matrix P such that PT AP = T is tridiagonal. Can we do better?—i.e., can we construct an orthogonal matrix P such that PT AP = D is a diagonal matrix? Indeed we can, and much of the material in Chapter 7 concerning eigenvalues and eigenvectors is devoted to this problem. But in the present context, this fact can be constructively established by means of Jacobi's diagonalization algorithm. Jacobi's Idea. If A ∈ n×n is symmetric, then a plane rotation matrix can be applied to reduce the magnitude of the off-diagonal entries. In particular, suppose that aij = 0 is the off-diagonal entry of maximal magnitude, and let A denote the matrix obtained by setting each akk = 0. If Pij is the plane rotation matrix described on p. 333 in which c = cos θ and s = sin θ, where cot 2θ = (aii − ajj)/2aij , and if B = PTij APij , then (1) bij = bji = 0 (2) 2 B F (3) B F ≤ 2 = (i.e., aij is annihilated), 2 A F − 2a2ij , 2 2 1− 2 A F . n −n Proof. The entries of B = PTij APij that lay on the intersection of the ith and j th rows with the ith and j th columns can be described by bii bij cos θ sin θ aii aij cos θ − sin θ ˆ ˆ B= = PT AP. bji bjj aji ajj − sin θ cos θ sin θ cos θ Use the identities cos 2θ = cos2 θ − sin2 θ and sin 2θ = 2 cos θ sin θ to verify ˆ F = A ˆ F = PT AP ˆ F (recall Exercise bij = bji = 0, and recall that B 49 Karl Gustav Jacob Jacobi (1804–1851) first presented this method in 1846, and it was popular for a time. But the twentieth-century development of electronic computers sparked tremendous interest in numerical algorithms for diagonalizing symmetric matrices, and Jacobi's method quickly fell out of favor because it could not compete with newer procedures—at least on the traditional sequential machines. However, the emergence of multiprocessor parallel computers has resurrected interest in Jacobi's method because of the inherent parallelism in the algorithm. Jacobi was born in Potsdam, Germany, educated at the University of Berlin, and employed as a professor at the University of K¨ onigsberg. During his prolific career he made contributions that are still important facets of contemporary mathematics. His accomplishments include the development of elliptic functions; a systematic development and presentation of the theory of determinants; contributions to the theory of rotating liquids; and theorems in the areas of differential equations, calculus of variations, and number theory. In contrast to his great contemporary Gauss, who disliked teaching and was anything but inspiring, Jacobi was regarded as a great teacher (the introduction of the student seminar method is credited to him), and he advocated the view that "the sole end of science is the honor of the human mind, and that under this title a question about numbers is worth as much as a question about the system of the world." Jacobi once defended his excessive devotion to work by saying that "Only cabbages have no nerves, no worries. And what do they get out of their perfect wellbeing?" Jacobi suffered a breakdown from overwork in 1843, and he died at the relatively young age of 46. 354 Chapter 5 Norms, Inner Products, and Orthogonality 5.6.9) to produce the conclusion b2ii + b2jj = a2ii + 2a2ij + a2jj . Now use the fact that bkk = akk for all k = i, j together with BF = AF to write 2 2 2 B F = BF − b2kk = BF − b2kk − b2ii + b2jj k = 2 AF = 2 A F − a2kk − k =i,j a2ii 2 + 2a2ij + a2jj = AF − a2kk − 2a2ij k =i,j − k 2a2ij . Furthermore, since a2pq ≤ a2ij for all p = q, A F = 2 a2pq ≤ p =q = 2 A F =⇒ −a2ij ≤ − p =q so 2 B F − 2a2ij ≤ A F , n2 − n 2 a2ij = (n2 − n)a2ij 2 A F A −2 2 F = n −n 2 2 1− 2 n −n A F . 2 Jacobi's Diagonalization Algorithm. Start with A0 = A, and produce a sequence of matrices Ak = PTk Ak−1 Pk , where at the k th step Pk is a plane rotation constructed to annihilate the maximal off-diagonal entry in Ak−1 . In particular, if aij is the entry of maximal magnitude in Ak−1 , then Pk is the rotator in the (i, j)-plane defined by setting 1 σ (aii − ajj) s= √ and c = √ = 1 − s2 , where σ = . 2aij 1 + σ2 1 + σ2 For n > 2 we have Ak F ≤ 2 1− 2 n2 − n k A F → 0 as k → ∞. Therefore, if P(k) is the orthogonal matrix defined by P(k) = P1 P2 · · · Pk , then T lim P(k) AP(k) = lim Ak = D k→∞ k→∞ is a diagonal matrix. Exercises for section 5.7 5.7.1. (a) Using Householder reduction, compute the QR factors of ⎛ ⎞1 19 −34 A = ⎝ −2 −5 20 ⎠. 2 8 37 (b) Repeat part (a) using Givens reduction. 5.7 Orthogonal Reduction 355 5.7.2. For A ∈ m×n , suppose that rank (A) = n, and let P be an orthogonal matrix such that Rn×n PA = T = , 0 where R is an upper-triangular matrix. If PT is partitioned as PT = [Xm×n | Y] , explain why the columns of X constitute an orthonormal basis for R (A). 5.7.3. By using Householder reduction, find an orthonormal basis for R (A), where ⎛ ⎞4 −3 4 ⎞ 2 −14 −3 ⎞A=⎝ ⎠. −2 14 0 1 −7 15 5.7.4. Use Householder reduction to compute the least squares solution for Ax = b, where ⎞ ⎛ ⎞ ⎛ 4 −3 4 5 ⎞ 2 −14 −3 ⎞ −15 ⎞A=⎝ and b = ⎝ ⎠ ⎠. −2 14 0 0 1 −7 15 30 Hint: Make use of the factors you computed in Exercise 5.7.3. 5.7.5. If A = QR is the QR factorization for A, explain why AF = RF , where F is the Frobenius matrix norm introduced on p. 279. 5.7.6. Find an orthogonal matrix P such that PT AP = H is in upperHessenberg form, where ⎛ ⎞−2 3 −4 A = ⎝ 3 −25 50 ⎠. −4 50 25 5.7.7. Let H be an upper-Hessenberg matrix, and suppose that H = QR, where R is a nonsingular upper-triangular matrix. Prove that

Q as well as the product RQ must also be in upper-Hessenberg form. 5.7.8. Approximately how many multiplications are needed to reduce an n × n nonsingular upper-Hessenberg matrix to upper-triangular form by using plane rotations? 356 5.8 Chapter 5 Norms, Inner Products, and Orthogonality DISCRETE FOURIER TRANSFORM For a positive integer n, the complex numbers ω = e2πi/n = cos 2π/n + i sin 2π/n are called the nth roots of unity because they represent all solutions to z n = 1. Geometrically, they are the vertices of a regular polygon of n sides as depicted in Figure 5.8.1 for n = 3 and n = 6. ω2 ω ω3 1 ω2 1 ω ω3 ω5 n=6 Figure 5.8.1 The roots of unity are cyclic in the sense that if k ≥ n, then ωk = ωk (mod n), where k (mod n) denotes the remainder when k is divided by n—for example, when n = 6, ω6 = 1, ω7 = ω, ω8 = ω2, ω9 = ω3, ... The numbers 1, ξ, ξ2, . . . , ξn−1, where ξ = e−2πi/n = cos 2π/n − i sin 2π/n = ω̄, are also the nth roots of unity, but, as depicted in Figure 5.8.2 for n = 3 and n = 6, they are listed in clockwise order around the unit circle rather than counterclockwise. ξ2 ξ ξ2 ξ5 ξ3 1 1 ξ2 ξ3 ξ n=6 Figure 5.8.2 The following identities will be useful in our development. If k is an integer, then 1 = |ξk |2 = ξ̄k ξk implies that ξ−k = ξ̄k = ωk. (5.8.1) 5.8 Discrete Fourier Transform 357 Furthermore, the fact that 1 + ξk + ξ2k + · · · + ξ(n−2)k + ξ(n−1)k = 0 implies 1 + ξk + ξ2k + · · · + ξ(n−1)k = 0 whenever ξk ≠ 1. (5.8.2) Fourier Matrix The n × n matrix whose (j, k)-entry is ξjk = ω−jk for 0 ≤ j, k ≤ n−1 is called the Fourier matrix of order n, and it has the form Fn = (1 1 1 ··· 1 ; 1 ξ ξ2 ··· ξn−1 ; 1 ξ2 ξ4 ··· ξn−2 ; ... ; 1 ξn−1 ξn−2 ··· ξ) n×n. Note. Throughout this section entries are indexed from 0 to n − 1. For example, the upper left-hand entry of Fn is considered to be in the (0, 0) position (rather than the (1, 1) position), and the lower right-hand entry is in the (n − 1, n − 1) position. When the context makes it clear, the subscript n on Fn is omitted. 50 The Fourier matrix is a special case of the Vandermonde matrix introduced in Example 4.3.4. Using (5.8.1) and (5.8.2), we see that the inner product of any two columns in Fn, say, the rth and sth, is F∗∗r F∗s = Σ ξ̄jr ξjs = Σ ξ−jr ξjs = Σ ξj(s−r) = 0. In other words, the columns of Fn are mutually orthogonal. Furthermore, each column of Fn has norm √n because ‖F∗k‖22 = Σ |ξjk|2 = Σ 1 = n, and consequently every column of Fn can be normalized by multiplying by the same scalar—namely, 1/√n. This means that (1/√n)Fn is a unitary matrix. Since it is also true that FTn = Fn, we have (1/√n Fn)−1 = (1/√n Fn)∗ = 1/√n F̄n, and therefore F−1n = F̄n/n. But (5.8.1) says that ξk = ω−k, so it must be the case that F̄n = (ω jk) = (1 1 ··· 1 ; 1 ω ω2 ··· ωn−1 ; ...) n×n. 358 Chapter 5 Norms, Inner Products, and Orthogonality Example 5.8.1 The Fourier matrices of orders 2 and 4 are given by F2 = (1 1 ; 1 −1) and F4 = (1 1 1 1 ; 1 −i −1 i ; 1 −1 1 −1 ; 1 i −1 −i), and their inverses are F2−1 = F̄2/2 = 1/2 (1 1 ; 1 −1) and F4−1 = F̄4/4 = 1/4 (1 1 1 1 ; 1 i −1 −i ; 1 −1 1 −1 ; 1 −i −1 i). Discrete Fourier Transform Given a vector xn×1, the product Fn x is called the discrete Fourier transform of x, and Fn−1 x is called the inverse transform of x. The k th entries in Fn x and Fn−1 x are given by [Fn x]k = Σ xj ξjk and [Fn−1 x]k = 1/n Σ xj ωjk. (5.8.3) 5.8 Discrete Fourier Transform 359 Example 5.8.2 Problem: Computing the Inverse Transform. Explain why any algorithm or program designed to compute the discrete Fourier transform of a vector x can also be used to compute the inverse transform of x. Solution: Call such an algorithm FFT (see p. 373 for a specific example). The fact that Fn−1 x = F̄n x/n means that FFT will return the inverse transform of x by executing the following three steps: (1) x ← x̄ (compute x̄). (2) x ← FFT(x) (compute Fn x̄). (3) x ← (1/n)x̄ (compute n−1 F̄n x = Fn−1 x). For example, computing the inverse transform of x = (i 0 −i 0)T is accomplished as follows—recall that F4 was given in Example 5.8.1. x̄ = (−i 0 i 0), F4 x̄ = (0 −2i 0 −2i), 1/4 F4x̄ = (0 i/2 0 i/2) = F4−1 x. You may wish to check that this answer agrees with the result obtained by directly multiplying F4−1 times x, where F4−1 is given in Example 5.8.1. Example 5.8.3 Signal Processing. Suppose that a microphone is placed under a hovering helicopter, and suppose that Figure 5.8.3 represents the sound signal that is recorded during 1 second of time. 6 4 2 0 -2 -4 -6 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 Chapter 5 Norms, Inner Products, and Orthogonality It seems reasonable to expect that the signal should have oscillatory components together with some random noise contamination. That is, we expect the signal to have the form y(τ) = Σ αk cos 2πfkτ + βk sin 2πfkτ + Noise. k But due to the noise contamination, the oscillatory nature of the signal is only barely apparent—the characteristic "chop-a chop-a chop-a" is not completely clear. To reveal the oscillatory components, the magic of the Fourier transformation is employed. Let x be the vector obtained by sampling the signal at n equally spaced points between time τ = 0 and τ = 1 (n = 512 in our case), and let y = (2/n)x = a + ib, where a = (2/n)Re (Fn x) and b = (2/n)Im (Fn x). Using only the first n/2 = 256 entries in a and ib, we plot the points in {(0, a0), (1, a1), . . . , (255, a255)} and {(0, ib0), (1, ib1), . . . , (255, ib255)} to produce the two graphs shown in Figure 5.8.4. 5.8 Discrete Fourier Transform 361 Real Axis 1.5 0.5 -0.5 0 50 100 150 200 250 300 Imaginary Axis 0.5 0 -0.5 -1 -1.5 -2 0 50 100 150 200 250 300 Frequency 0 Figure 5.8.4 Now there are some obvious characteristics—the plot of a in the top graph of Figure 5.8.4 has a spike of height approximately 1 at entry 80, and the plot of ib in the bottom graph has a spike of height approximately −2 at entry 50. These two spikes indicate that the signal is made up primarily of two oscillatory components. The spike in the real vector a indicates that one of the oscillatory components is a cosine of frequency 80 Hz (or period = 1/80) whose amplitude is approximately 1, and the spike in the imaginary vector ib indicates there is a sine component with frequency 50 Hz and amplitude of about 2. In other words, the Fourier transform indicates that the signal is y(τ) = cos 2π(80τ) + 2 sin 2π(50τ) + Noise. In truth, the data shown in Figure 5.8.3 was artificially generated by contaminating the function y(τ) = cos 2π(80τ) + 2 sin 2π(50τ) with some normally distributed zero-mean noise, and therefore the plot of (2/n)Fn x shown in Figure 5.8.4 does indeed accurately reflect the true nature of the signal. To understand why Fn reveals the hidden frequencies, let cos 2πft and sin 2πft denote the discrete cosine and discrete sine vectors cos 2πf·0/n ... sin 2πf·(n−1)/n, where f is an integer. Note that ... cos 2πft = (e2πift + e−2πift)/2 and sin 2πft = (e2πift − e−2πift)/2i ... Since the discrete exponential vectors e2πift and e−2πift are defined in the natural way as e2πift = cos 2πft + i sin 2πft ... which, in terms of the columns of the Fourier matrix, are F∗n−f and F∗f ... The Fourier transform of a discrete cosine or sine vector is therefore essentially a vector of zeros except for entries in positions f and n − f ... Consequently, if f1 and f2 are integers between 0 and n/2, the Fourier transform of αcos 2πf1t + βsin 2πf2t is ... Notice that this result is exactly what is depicted in Figure 5.8.4 ... 5.8 Discrete Fourier Transform 363 ... The result is that Fn acts as a filter—picking out oscillatory components of a signal. ... (the discrete sine and cosine vectors) ... 5.8 Discrete Fourier Transform ... Example 5.8.4 ... Convolution ... For p = (p0, p1, ..., pn−1)T and q = (q0, ..., qn−1)T, the convolution p ⊛ q is defined ... a(x)b(x) ... (5.8.8) ... The Convolution Theorem ... Fast Fourier Transform ... 5.8 Discrete Fourier Transform ... F2r = ... (Fr DrFr ; Fr −DrFr) ... (5.8.13) ... Example 5.8.5 ... Exercises for section 5.8 ... 5.9 Chapter 5 Norms, Inner Products, and Orthogonality COMPLEMENTARY SUBSPACES The sum of two subspaces X and Y of a vector space V was defined on p. 166 to be the set X + Y = {x + y | x ∈ X and y ∈ Y}, and it was established that X + Y is another subspace of V. For example, consider the two subspaces of ℜ3 shown in Figure 5.9.1 in which X is a plane through the origin and Y is a line through the origin. Figure 5.9.1 Notice that X and Y are disjoint in the sense that X ∩ Y = 0. The parallelogram law for vector addition makes it clear that X + Y = ℜ3 because each vector in ℜ3 can be written as "something from X plus something from Y." Thus ℜ3 is resolved into a pair of disjoint components X and Y. These ideas generalize as described below. Complementary Subspaces Subspaces X, Y of a space V are said to be complementary whenever V = X + Y and X ∩ Y = 0, (5.9.1) in which case V is said to be the direct sum of X and Y, and this is denoted by writing V = X ⊕ Y. • For a vector space V with subspaces X, Y having respective bases BX and BY, the following statements are equivalent. ◦ V = X ⊕ Y. (5.9.2) ◦ For each v ∈ V there are unique vectors x ∈ X and y ∈ Y such that v = x + y. (5.9.3) ◦ BX ∩ BY = φ and BX ∪ BY is a basis for V. (5.9.4) 384 Chapter 5 Norms, Inner Products, and Orthogonality Prove these by arguing (5.9.2) =⇒ (5.9.3) =⇒ (5.9.4) =⇒ (5.9.2). Proof of (5.9.2) =⇒ (5.9.3). First recall from (4.4.19) that dim V = dim (X + Y) = dim X + dim Y − dim (X ∩ Y). If V = X ⊕ Y, then X ∩ Y = 0, and thus dim V = dim X + dim Y. To prove (5.9.3), suppose there are two ways to represent v ∈ V as "something from X plus something from Y." If v = x1 + y1 = x2 + y2, where x1, x2 ∈ X and y1, y2 ∈ Y, then x1 − x2 = y2 − y1 ∈ X ∩ Y = 0 ... so x1 = x2 and y1 = y2. Proof of (5.9.3) =⇒ (5.9.4). The hypothesis insures that V = X + Y, and we know from (4.1.2) that BX ∪ BY spans X + Y, so BX ∪ BY must be a spanning set for V. To prove BX ∪ BY is linearly independent, let BX = {x1, x2, . . . , xr} and BY = {y1, y2, . . . , ys}, and suppose that 0 = Σ αi xi + Σ βj yj. This is one way to express 0 as "something from X plus something from Y," while 0 = 0 + 0 is another way. Consequently, (5.9.3) guarantees that Σ αi xi = 0 and Σ βj yj = 0, and hence α1 = α2 = · · · = αr = 0 and β1 = β2 = · · · = βs = 0 because BX and BY are both independent. Therefore, BX ∪ BY is linearly independent, and hence it is a basis for V. Proof of (5.9.4) =⇒ (5.9.2). If BX ∪ BY is a basis for V, then BX ∪ BY is a linearly independent set. This together with the fact that BX and BY always spans X + Y means BX ∪ BY is a basis for X + Y as well as for V. Consequently, V = X + Y, and hence dim V = dim X + dim Y = dim (X + Y) = dim X + dim Y − dim (X ∩ Y), so dim (X ∩ Y) = 0 or, equivalently, X ∩ Y = 0. If V = X ⊕ Y, then (5.9.3) says there is one and only one way to write v ∈ V as v = x + y with x ∈ X and y ∈ Y. The vector x is called the "X-component" of v, and y is the "Y-component" of v. These two components of v have a definite geometrical interpretation. Look back at Figure 5.9.1 in which ℜ3 = X ⊕ Y, where X is a plane and Y is a line outside of X, and notice that x (the X-component of v) is the result of projecting v onto X along a line parallel to Y, and y (the Y-component of v) is obtained by projecting v onto Y along a line parallel to X. This leads to the following formal definition of a projection. 5.9 Complementary Subspaces 385 Projection Suppose that V = X ⊕ Y so that for each v ∈ V there are unique vectors x ∈ X and y ∈ Y such that v = x + y. • The vector x is called the projection of v onto X along Y. • The vector y is called the projection of v onto Y along X. It's clear that if X ⊥ Y in Figure 5.9.1, then this notion of projection agrees with the concept of orthogonal projection that was discussed on p. 322. The phrase "oblique projection" is sometimes used to emphasize the fact that X and Y are not orthogonal subspaces. In this text the word "projection" is synonymous with the term "oblique projection." If it is known that X ⊥ Y, then we explicitly say "orthogonal projection." Orthogonal projections are discussed in detail on p. 429. Given a pair of complementary subspaces X and Y of ℜn and an arbitrary vector v ∈ ℜn = X ⊕ Y, how can the projection of v onto X be computed? One way is to build a projector (a projection operator) that is a matrix Pn×n with the property that for each v ∈ ℜn, the product Pv is the projection of v onto X along Y. Let BX = {x1, x2, . . . , xr} and BY = {y1, y2, . . . , yn−r} be respective bases for X and Y so that BX ∪ BY is a basis for ℜn—recall (5.9.4). This guarantees that if the xi's and yi's are placed as columns in B = (x1 x2 · · · xr | y1 y2 · · · yn−r) = (X | Y), then B is nonsingular. If Pn×n is to have the property that Pv is the projection of v onto X along Y for every v ∈ ℜn, then (5.9.3) implies that Pxi = xi, i = 1, 2, . . . , r and Pyj = 0, j = 1, 2, . . . , n − r, so PB = P (X | Y) = (PX | PY) = (X | 0) and, consequently, P = (X | 0) B−1 = B (I 0 ; 0 0) B−1. (5.9.5) 386 Chapter 5 Norms, Inner Products, and Orthogonality It is possible that there can be more than one matrix P with the property that Pv is the projection of v onto X along Y for each v ∈ ℜn ... To argue that P is independent of which pair of bases is used, suppose P1 and P2 are two such matrices ... The projection of v onto X along Y is therefore P1v = P2v for all v, and hence P1 = P2. ... Since the complementary subspace ... Q = I − P ... Projectors ... Let X and Y be complementary subspaces of a vector space V so that each v ∈ V can be uniquely resolved as v = x + y, where x ∈ X and y ∈ Y. The unique linear operator P defined by Pv = x is called the projector onto X along Y, and P has the following properties. • P2 = P (P is idempotent). (5.9.6) • I − P is the complementary projector onto Y along X. (5.9.7) • R (P) = {x | Px = x} (R (P) is the set of "fixed points" for P). (5.9.8) • R (P) = N (I − P) = X and R (I − P) = N (P) = Y. (5.9.9) • If V = ℜn or Cn, then P is given by P = (X | 0) (X | Y)−1 = (X | Y) (I 0 ; 0 0) (X | Y)−1, (5.9.10) where the columns of X and Y are respective bases for X and Y. Other formulas for P are given on p. 634. Proof. Some of these properties have already been derived in the context of ℜn. But since the concepts of projections and projectors are valid for all vector spaces, more general arguments that do not rely on properties of ℜn will be provided. Uniqueness is evident because if P1 and P2 both satisfy the defining condition, then P1v = P2v for every v ∈ V, and thus P1 = P2. The linearity of P can be established by using the uniqueness ... (5.9.6) ... Consequently P2 = P ... 5.9 Complementary Subspaces 387 ... (5.9.7) ... (5.9.8) ... To prove (5.9.9), ... It's easy to see that ... conversely ... The implications are left as exercises ... Notice that ... P2 = P ... Example 5.9.1 Problem: Let X and Y be the subspaces of ℜ3 that are spanned by BX = ... and BY = ..., respectively. Explain why X and Y are complementary, and then determine the projector onto X along Y. What is the projection of v onto X along Y? Solution: BX ∪ BY is linearly independent, so they are complementary ... P = B (I 0 ; 0 0) B−1 ... 388 Chapter 5 Norms, Inner Products, and Orthogonality ... Angle between Complementary Subspaces. The angle between nonzero vectors u and v in ℜn was defined on p. 295 to be the number 0 ≤ θ ≤ π/2 such that cos θ = vTu/‖v‖2‖u‖2. It's natural to try to extend this idea to somehow make sense of angles between complementary subspaces of ℜn ... Example 5.9.2 ... 5.9 Complementary Subspaces 389 ... ‖P‖2 = 1/sin θ ... Exercises for section 5.9 ... 5.9.1. ... 5.9.2. ... 5.9.3. ... 5.9.4. ... 5.9.5. ... 5.9.6. ... 5.9.7. ... 5.9.8. ... 5.9.9. ... 5.9.10. ... 5.9.11. ... 5.9.12. ... 5.9.13. ... 5.9.14. ... 5.9.15. ... 5.9.16. ... 5.9.17. ... 5.10 RANGE-NULLSPACE DECOMPOSITION Since there are infinitely many different pairs of complementary subspaces in ℜn (or Cn), is some pair more "natural" than the rest? Without reference to anything else the question is hard to answer. But if we start with a given matrix An×n, then there is a very natural direct sum decomposition of ℜn defined by fundamental subspaces associated with powers of A. The rank plus nullity theorem on p. 199 says that dim R (A) + dim N (A) = n, so it's reasonable to ask about the possibility of R (A) and N (A) being complementary subspaces. If A is nonsingular, then it's trivially true that R (A) and N (A) are complementary, but when A is singular, this need not be the case because R (A) and N (A) need not be disjoint. For example, 0 1 ; 0 0 ... Range-Nullspace Decomposition For every singular matrix An×n, there exists a positive integer k such that R (Ak) and N (Ak) are complementary subspaces. That is, ℜn = R (Ak) ⊕ N (Ak). (5.10.1) The smallest positive integer k for which (5.10.1) holds is called the index of A. For nonsingular matrices we define index(A) = 0. Proof. First observe that as A is powered the nullspaces grow and the ranges shrink—recall Exercise 4.2.12. ... 5.10 Range-Nullspace Decomposition 395 ... (5.10.2) ... (5.10.3) ... (5.10.4) ... Rank Stabilization ... Nilpotent Matrices ... index(N) ... Core-Nilpotent Decomposition If A is an n × n singular matrix of index k such that rank (Ak) = r, then there exists a nonsingular matrix Q such that Q−1AQ = (Cr×r 0 ; 0 N) (5.10.5) in which C is nonsingular, and N is nilpotent of index k. In other words, A is similar to a 2 × 2 block-diagonal matrix containing a nonsingular "core" and a nilpotent component. The block-diagonal matrix in (5.10.5) is called a core-nilpotent decomposition of A. ... Example 5.10.2 ... Example 5.10.3 ... 398 Chapter 5 Norms, Inner Products, and Orthogonality ... Example 5.10.4 ... Let T be a linear operator ... Since V = R ⊕ N, ... T = E ⊕ F in which E = T/R

is an invertible operator on R, and F is a nilpotent operator on N . Since V = R ⊕ N , each x ∈ V can be expressed as x = r + n with r ∈ R and n ∈ N . This allows us to formulate the concept of the direct sum of E and F by defining E ⊕ F to be the linear operator on V such that (E ⊕ F)(x) = E(r) + F(n) for each x ∈ V. Therefore, T(x) = T(r + n) = T(r) + T(n) = (T/)(r) + (T/)(n) R N = E(r) + F(n) = (E ⊕ F)(x) for each x ∈ V. In other words, T = E ⊕ F in which E = T/ is invertible and F = T/ is N nilpotent. Example 5.10.5 Drazin Inverse. Inverting the nonsingular core C and neglecting the nilpotent part N in the core-nilpotent decomposition (5.10.5) produces a natural generalization of matrix inversion. More precisely, if A=Q C 0 0 N Q−1 , then AD = Q C−1 0 0 0 Q−1 (5.10.6) defines the Drazin inverse of A. Even though the components in a corenilpotent decomposition are not uniquely defined by A, it can be proven that AD is unique and has the following properties. 55 • AD = A−1 when A is nonsingular (the nilpotent part is not present). • AD AAD = AD , AAD = AD A, Ak+1 AD = Ak , where k = index(A). 55 These three properties served as Michael P. Drazin's original definition in 1968. Initially, 400 Chapter 5 • • • Norms, Inner Products, and Orthogonality If Ax = b is a consistent system of linear equations in which b ∈ R Ak , then x = AD b is the unique solution that belongs to R Ak (Exercise 5.10.9). k AAD is the projector onto R Ak along N A , and I − AAD is the complementary projector onto N Ak along R Ak (Exercise 5.10.10). If A is considered as a linear operator on n , then, with respect to a basis k BR for R A , C is the matrix representation for the restricted operator A/ (see p. 263). Thus A/ is invertible. Moreover, k k R(A) 1 D A / R(Ak) R(A) 2 −1 BR =C 3 = A/ R(Ak) −1 4 , BR so AD/ −1 = A /R(Ak) . R(Ak) In other words, AD is the inverse of A on R Ak , and AD is the zero operator on N Ak , so, in the context of Example 5.10.4, ⊕ A/ R(Ak) N (Ak) A = A/ and AD = A/ ⊕ 0/ . R(Ak) N (Ak) Exercises for section 5.10 5.10.1. If A is a square matrix of index k > 0, prove that index(Ak) = 1. 5.10.2. If A is a nilpotent matrix of index k, describe the components in a core-nilpotent decomposition of A. 5.10.3. Prove that if A is a symmetric matrix, then index(A) ≤ 1. 5.10.4. A ∈ C n×n is said to be a normal matrix whenever AA∗ = A∗ A. Prove that if A is normal, then index(A) ≤ 1. Note: All symmetric matrices are normal, so the result of this exercise includes the result of Exercise 5.10.3 as a special case. Drazin's concept attracted little interest—perhaps due to Drazin's abstract algebraic presentation. But eventually Drazin's generalized inverse was recognized to be a useful tool for analyzing nonorthogonal types of problems involving singular matrices. In this respect, the Drazin inverse is complementary to the Moore–Penrose pseudoinverse discussed in Exercise 4.5.20 and on p. 423 because the Moore–Penrose pseudoinverse is more useful in applications where orthogonality is somehow wired in (e.g., least squares). 5.10 Range-Nullspace Decomposition 401 5.10.5. Find a core-nilpotent decomposition and the Drazin inverse of ⎛ −2 A=⎝ 4 3 0 2 2 ⎞−4 4⎠. 2 5.10.6. For a square matrix A, any scalar λ that makes A − λI singular is called an eigenvalue for A. The index of an eigenvalue λ is defined to be the index of the associated matrix A − λI. In other words, index(λ) = index(A − λI). Determine the eigenvalues and the index of each eigenvalue for the following matrices: ⎛ 1 ⎜0 (a) J = ⎜ 0 ⎝ 0 0 0 1 0 0 0 0 1 0 0 0 0 0 2 0 ⎞0 0 ⎟0 ⎟. 0 ⎠ 0 2 ⎛ 1 ⎜0 (b) J = ⎜ 0 ⎝ 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 2 0 ⎞0 0 ⎟0 ⎟. 0 ⎠ 1 2 5.10.7. Let P be a projector different from the identity. (a) Explain why index(P) = 1. What is the index of I? (b) Determine the core-nilpotent decomposition for P. 5.10.8. Let N be a nilpotent matrix of index k, and suppose that x is a vector such that Nk−1 x = 0. Prove that the set C = {x, Nx, N2 x, . . . , Nk−1 x} is a linearly independent set. C is sometimes called a Jordan chain or a Krylov sequence. 5.10.9. Let A (a) (b) (c) be a square matrix of index k, and let b ∈ R Ak . Explain why the linear system Ax = b must be consistent. Explain why x = AD b is the unique solution in R Ak . Explain why the general solution is given by AD b + N (A). 5.10.10. Suppose that A is a square matrix of index k, and let AD be the Drazin inverse of A as defined Explain why AAD k in Example k5.10.5. is the projector onto R A along N A . What does I − AAD project onto and along? 402 Chapter 5 Norms, Inner Products, and Orthogonality 5.10.11. An algebraic group is a set G together with an associative operation between its elements such that G is closed with respect to this operation; G possesses an identity element E (which can be proven to be unique); and every member A ∈ G has an inverse A# (which can be proven to be unique). These are essentially the axioms (A1), (A2), (A4), and (A5) in the definition of a vector space given on p. 160. A matrix group is a set of square matrices that forms an algebraic group under ordinary matrix multiplication. (a) Show that the set of n × n nonsingular matrices is a matrix group. (b) Show that the set of n × n unitary matrices is a subgroup of the n × n nonsingular matrices. 5 6 α α (c) Show that the set G = α = 0 is a matrix group. α α In particular, what does the identity element E ∈ G look like, and what does the inverse A# of A ∈ G look like? 5.10.12. For singular matrices, prove that the following statements are equivalent. (a) A is a group matrix (i.e., A belongs to a matrix group). (b) R (A) ∩ N (A) = 0. (c) R (A) and N (A) are complementary subspaces. (d) index(A) = 1. (e) There are nonsingular matrices Qn×n and Cr×r such that −1 Q AQ = Cr×r 0 0 0 , where r = rank (A). 5.10.13. Let A ∈ G for some matrix group G. (a) Show that the identity element E ∈ G is the projector onto R (A) along N (A) by arguing that E must be of the form E=Q Ir×r 0 0 0 Q−1 . (b) Show that the group inverse of A (the inverse of A in G) must be of the form −1 0 C # A =Q Q−1 . 0 0 5.11 Orthogonal Decomposition 5.11 403 ORTHOGONAL DECOMPOSITION The orthogonal complement of a single vector x was defined on p. 322 to be the set of all vectors orthogonal to x. Below is the natural extension of this idea. Orthogonal Complement For a subset M of an inner-product space V, the orthogonal complement M⊥ (pronounced "M perp") of M is defined to be the set of all vectors in V that are orthogonal to every vector in M . That is, M⊥ = x ∈ V m x = 0 for all m ∈ M . For example, if M = {x} is a single vector in 2 , then, as illustrated in Figure 5.11.1, M⊥ is the line through the origin that is perpendicular to x. If M is a plane through the origin in 3 , then M⊥ is the line through the origin that is perpendicular to the plane. Figure 5.11.1 ⊥ Notice that M is a subspace of V even if M is not a subspace because M⊥ is closed with respect to vector addition and scalar multiplication (Exercise 5.11.4). But if M is a subspace, then M and M⊥ decompose V as described below. Orthogonal Complementary Subspaces If M is a subspace of a finite-dimensional inner-product space V, then V = M ⊕ M⊥ . (5.11.1) Furthermore, if N is a subspace such that V = M ⊕ N and N ⊥ M (every vector in N is orthogonal to every vector in M), then N = M⊥ . (5.11.2) 404 Chapter 5 Norms, Inner Products, and Orthogonality Proof. Observe that M ∩ M⊥ = 0 because if x ∈ M and x ∈ M⊥ , then x must be orthogonal to itself, and x x = 0 implies x = 0. To prove that M ⊕ M⊥ = V, suppose that BM and BM⊥ are orthonormal bases for M and M⊥ , respectively. Since M and M⊥ are disjoint, BM ∪ BM⊥ is an orthonormal basis for some subspace S = M ⊕ M⊥ ⊆ V. If S = V, then the basis extension technique of Example 4.4.5 followed by the Gram–Schmidt orthogonalization procedure of §5.5 yields a nonempty set of vectors E such that BM ∪ BM⊥ ∪ E is an orthonormal basis for V. Consequently, E ⊥ BM == E ⊥ M =⇒ E ⊆ M⊥ == E ⊆ span (BM⊥) . But this is impossible because BM ∪BM⊥ ∪E is linearly independent. Therefore, E is the empty set, and thus V = M ⊕ M⊥ . To prove statement (5.11.2), note that N ⊥ M implies N ⊆ M⊥ , and coupling this with the fact that M ⊕ M⊥ = V = M ⊕ N together with (4.4.19) insures dim N = dim V − dim M = dim M⊥ . Example 5.11.1 Problem: Let Um×m = U1 | U2 be a partitioned orthogonal matrix. Explain why R (U1) and R (U2) must be orthogonal complements of each other. Solution: Statement (5.9.4) insures that m = R (U1) ⊕ R (U2), and we know that R (U1) ⊥ R (U2) because the columns of U are an orthonormal set. ⊥ Therefore, (5.11.2) guarantees that R (U2) = R (U1) . Perp Operation If M is a subspace of an n-dimensional inner-product space, then the following statements are true. • dim M⊥ = n − dim M. (5.11.3) • ⊥ M⊥ = M. (5.11.4) Proof. Property (5.11.3) follows from the fact that M and M⊥ are comple⊥ mentary subspaces—recall (4.4.19). To prove (5.11.4), first show that M⊥ ⊆ M. If x ∈ M⊥ , then (5.11.1) implies x = m + n, where m ∈ M and ⊥ n ∈ M , so 0 = n x = n m + n = n m + n n = n n == n = 0 == x ∈ M, ⊥ and thus M⊥ ⊆ M. We know from (5.11.3) that dim M⊥ = n − dim M and ⊥ ⊥ dim M⊥ = n−dim M⊥ , so dim M⊥ = dim M. Therefore, (4.4.6) guarantees ⊥ that M⊥ = M. 5.11 Orthogonal Decomposition 405 We are now in a position to understand why the four fundamental subspaces associated with a matrix A ∈ m×n are indeed "fundamental." First consider ⊥ R (A) , and observe that for all y ∈ n , ⊥ x ∈ R (A) == Ay x = 0 == yT AT x = 0 7 8 == y AT x = 0 == AT x = 0 (Exercise 5.3.2) == x ∈ N AT . ⊥ Therefore, R (A) = N AT . Perping both sides of this equation and replac 56 ing A by AT produces R AT = N (A) . Combining these observations produces one of the fundamental theorems of linear algebra. Orthogonal Decomposition Theorem For every A ∈ m×n , ⊥ R (A) = N AT ⊥ N (A) = R AT . and (5.11.5) In light of (5.11.1), this means that every matrix A ∈ m×n produces an orthogonal decomposition of m and n in the sense that and ⊥ m = R (A) ⊕ R (A) = R (A) ⊕ N AT , (5.11.6) ⊥ n = N (A) ⊕ N (A) = N (A) ⊕ R AT . (5.11.7) Theorems without hypotheses tend to be extreme in the sense that they either say very little or they reveal a lot. The orthogonal decomposition theorem has no hypothesis—it holds for all matrices—so, does it really say something significant? Yes, it does, and here's part of the reason why. In addition to telling us how to decompose m and n in terms of the four fundamental subspaces of A, the orthogonal decomposition theorem also tells us how to decompose A itself into more basic components. Suppose that rank (A) = r, and let BR(A) = {u1 , u2 , . . . , ur } and BN (AT) = {ur+1 , ur+2 , . . . , um } be orthonormal bases for R (A) and N AT , respectively, and let BR(AT) = {v1 , v2 , . . . , vr } 56 and BN (A) = {vr+1 , vr+2 , . . . , vn } Here, as well as throughout the rest of this section, ()T can be replaced by ()∗ whenever m×n is replaced by C m×n . 406 Chapter 5 Norms, Inner Products, and Orthogonality be orthonormal bases for R AT and N (A), respectively. It follows that BR(A) ∪ BN (AT) and BR(AT) ∪ BN (A) are orthonormal bases for m and n , respectively, and hence Um×m = u1 | u2 | · · · | um and Vn×n = v1 | v2 | · · · | vn (5.11.8) are orthogonal matrices. Now consider the product R = UT AV, and notice that rij = uTi Avj . However, uTi A = 0 for i = r + 1, . . . , m and Avj = 0 for j = r + 1, . . . , n, so ⎛ uT Av · · · uT Av 0 · · · 0 ⎞1 1 ..⎟ . | | | T ur Av1 R = UT AV = | | | 0 | .. ⎝ . 0 1 .. . r .. . .. . · · · uTr Avr 0 · · · 0 .. 0 .. . ⎠· · · 0 0 In other words, A can be factored as A = URVT = U Cr×r 0 0 0 .. ⎟.. ⎟· · · 0 ⎟ ⎠ · · · 0 ⎟.. ⎟. .. . . ⎠· · · 0 (5.11.9) VT . (5.11.10) Moreover, C is nonsingular because it is r × r and C 0 rank (C) = rank = rank UT AV = rank (A) = r. 0 0 For lack of a better name, we will refer to (5.11.10) as a URV factorization. We have just observed that every set of orthonormal bases for the four fundamental subspaces defines a URV factorization. The situation is also reversible in the sense that every URV factorization of A defines an orthonormal basis for each fundamental subspace. Starting with orthogonal matrices U = U1 | U2 and V = V1 | V2 together with a nonsingular matrix Cr×r such that (5.11.10) holds, use the fact that right-hand multiplication by a nonsingular matrix does not alter the range (Exercise 4.5.12) to observe R (A) = R (UR) = R (U1 C | 0) = R (U1 C) = R (U1). ⊥ ⊥ By (5.11.5) and Example 5.11.1, N AT = R (A) = R (U1) = R (U2). Similarly, left-hand multiplication by a nonsingular matrix does not change the nullspace, so the second equation in (5.11.5) along with Example 5.11.1 yields CV1T ⊥ N (A) = N RVT = N = N CV1T = N V1T = R (V1) = R (V2), 0 ⊥ ⊥ and R AT = N (A) = R (V2) = R (V1). A summary is given below. 5.11 Orthogonal Decomposition 407 URV Factorization For each A ∈ m×n of rank r, there are orthogonal matrices Um×m and Vn×n and a nonsingular matrix Cr×r such that A = URVT = U • • • • Cr×r 0 0 0 VT . (5.11.11) m×n The first r columns in U are an orthonormal basis for R (A). The last m−r columns of U are an orthonormal basis for N AT . The first r columns in V are an orthonormal basis for R AT . The last n − r columns of V are an orthonormal basis for N (A). Each different collection of orthonormal bases for the four fundamental subspaces of A produces a different URV factorization of A. In the complex case, replace ()T by ()∗ and "orthogonal" by "unitary." Example 5.11.2 Problem: Explain how to make C lower triangular in (5.11.11). Solution: Apply Householder (orGivens) reduction to produce an orthogonal B matrix Pm×m such that PA = 0 , where B is r × n of rank r. Householder (or Givens) reduction applied to BT results in an orthogonal matrix Qn×n and a nonsingular upper-triangular matrix T such that QBT = Tr×r 0 == B = TT | 0 Q == B 0 = TT 0 0 0 Q, T 0 T T so A = PT B = P Q is a URV factorization. 0 0 0 Note: C can in fact be made diagonal—see (p. 412). Have you noticed the duality that has emerged concerning the use of fundamental subspaces of A to decompose n (or C n)? On one hand there is the range-nullspace decomposition (p. 394), and on the other is the orthogonal decomposition theorem (p. 405). Each produces a decomposition of A. The range-nullspace decomposition of n produces the core-nilpotent decomposition of A (p. 397), and the orthogonal decomposition theorem produces the URV factorization. In the next section, the URV factorization specializes to become 408 Chapter 5 Norms, Inner Products, and Orthogonality the singular value decomposition (p. 412), and in a somewhat parallel manner, the core-nilpotent decomposition paves the way to the Jordan form (p. 590). These two parallel tracks constitute the backbone for the theory of modern linear algebra, so it's worthwhile to take a moment and reflect on them. The range-nullspace decomposition decomposes n with square matrices while the orthogonal decomposition theorem does it with rectangular matrices. So does this mean that the range-nullspace decomposition is a special case of, or somehow weaker than, the orthogonal decomposition theorem? No! Even for square matrices they are not very comparable because each says something that the other doesn't. The core-nilpotent decomposition (and eventually the Jordan form) is obtained by a similarity transformation, and, as discussed in §§4.8–4.9, similarity is the primary mechanism for revealing characteristics of A that are independent of bases or coordinate systems. The URV factorization has little to say about such things because it is generally not a similarity transformation. Orthogonal decomposition has the advantage whenever orthogonality is naturally built into a problem—such as least squares applications. And, as discussed in §5.7, orthogonal methods often produce numerically stable algorithms for floating-point computation, whereas similarity transformations are generally not well suited for numerical computations. The value of similarity is mainly on the theoretical side of the coin. So when do we get the best of both worlds—i.e., when is a URV factorization also a core-nilpotent decomposition? First, A must be square and, second, (5.11.11) must be a similarity transformation, so U = V. Surprisingly, this happens for a rather large class of matrices described below. Range Perpendicular to Nullspace For rank (An×n) = r, the following statements are equivalent: • R (A) ⊥ N (A), (5.11.12) • R (A) = R AT , (5.11.13) T • N (A) = N A , (5.11.14) Cr×r 0 • A=U (5.11.15) UT 0 0 in which U is orthogonal and C is nonsingular. Such matrices will be called RPN matrices, short for "range perpendicular to nullspace." Some authors call them range-symmetric or EP matrices. Nonsingular matrices are trivially RPN because they have a zero nullspace. For complex matrices, replace ()T by ()∗ and "orthogonal" by "unitary." Proof. The fact that (5.11.12) == (5.11.13) == (5.11.14) is a direct consequence of (5.11.5). It suffices to prove (5.11.15) == (5.11.13). If (5.11.15) is a 5.11 Orthogonal Decomposition 409 URV with V = U = U1 |U2), then R (A) = R (U1) = R (V1) = Tfactorization R A . Conversely, if R (A) = R AT , perping both sides and using equation (5.11.5) produces N (A) = N AT , so (5.11.8) yields a URV factorization with U = V. Example 5.11.3 A ∈ C n×n is called a normal matrix whenever AA∗ = A∗ A. As illustrated in Figure 5.11.2, normal matrices fill the niche between hermitian and (complex) RPN matrices in the sense that real-symmetric ⇒ hermitian ⇒ normal ⇒ RPN, with no implication being reversible—details are called for in Exercise 5.11.13. RPN Normal Hermitian Real-Symmetric Nonsingular Figure 5.11.2 Exercises for section 5.11 2 5.11.1. Verify the orthogonal decomposition theorem for A = 1 −1 −1 −2 1 0 −1 . 5.11.2. For an inner-product space V, what is V ⊥ ? What is 0⊥ ? ⎧⎛ ⎞⎛ ⎞⎫ 2 ⎬ ⎨ 1 2 ⎠⎝4 ⎠⎭ 5.11.3. Find a basis for the orthogonal complement of M = span , 1 . ⎩⎝ 0 ⎭ 3 6 5.11.4. For every inner-product space V, prove that if M ⊆ V, then M⊥ is a subspace of V. 5.11.5. If M and N are subspaces of an n-dimensional inner-product space, prove that the following statements are true. (a) M ⊆ N =⇒ N ⊥ ⊆ M⊥ . (b) (M + N)⊥ = M⊥ ∩ N ⊥ . (c) (M ∩ N)⊥ = M⊥ + N ⊥ . 410 Chapter 5 Norms, Inner Products, and Orthogonality 5.11.6. Explain why the rank plus nullity theorem on p. 199 is a corollary of the orthogonal decomposition theorem. 5.11.7. Suppose A = URVT is a URV factorization ofan m ×n matrix of rank r, and suppose U is partitioned as U = U1 | U2 , where U1 is m ×r. Prove that P = U1 UT1 is the projector onto R (A) along N AT . In this case, P is said to be an orthogonal projector because its range is orthogonal to its nullspace. What is the orthogonal projector onto N AT along R (A)? (Orthogonal projectors are discussed in more detail on p. 429.) 5.11.8. Use the Householder reduction method as described in Example 5.11.2 to compute a URV factorization as well as orthonormal bases for the four fundamental subspaces of A = −4 2 −4 −2 −2 1 −4 2 −4 −2 1 −2 . 5.11.9. Compute a URV factorization for the matrix given in Exercise 5.11.8 by using elementary row operations together with Gram–Schmidt orthogonalization. Are the results the same as those of Exercise 5.11.8? 5.11.10. For the matrix A of Exercise 5.11.8, find vectors x ∈ R (A) and T y ∈ N AT such that v = x + y, where v = (3 3 3) . Is there more than one choice for x and y? 5.11.11. Construct a square matrix such that R (A) ∩ N (A) = 0, but R (A) is not orthogonal to N (A). 5.11.12. For An×n singular, explain why R (A) ⊥ N (A) implies index(A) = 1, but not conversely. 5.11.13. Prove that real-symmetric matrix ⇒ hermitian ⇒ normal ⇒ (complex) RPN. Construct examples to show that none of the implications is reversible. 5.11.14. Let A be a normal matrix. (a) Prove that R (A − λI) ⊥ N (A − λI) for every scalar λ. (b) Let λ and μ be scalars such that A − λI and A − μI are singular matrices—such scalars are called eigenvalues of A. Prove that if λ = μ, then N (A − λI) ⊥ N (A − μI). 5.12 Singular Value Decomposition 5.12 411 SINGULAR VALUE DECOMPOSITION For an m × n matrix A of rank r, Example 5.11.2 shows how to build a URV factorization Cr×r 0 T A = URV = U VT 0 0 m×n in which C is triangular. The purpose of this section is to prove that it's possible to do even better by showing that C can be made to be diagonal . To see how, let σ1 = A2 = C2 (Exercise 5.6.9), and recall from the proof of (5.2.7) on p. 281 that C2 = Cx2 for some vector x such that (CT C − λI)x = 0, x2 = 1 and λ = xT CT Cx = σ12 . (5.12.1) Set y = Cx/Cx2 = Cx/σ1 , and let Ry = y | Y and Rx = x | X be elementary reflectors having y and x as their first columns, respectively—recall Example 5.6.3. Reflectors are orthogonal matrices, so xT X = 0 and YT y = 0, and these together with (5.12.1) yield CX = where xT CT CX λxT X = =0 σ1 σ1 and YT Cx = σ1 YT y = 0. Coupling these facts with yT Cx = yT (σ1 y) = σ1 and Ry = RTy produces T T y σ1 0 y Cx yT CX Ry CRx = = C x|X = YT Cx YT CX 0 C2 YT with σ1 ≥ C2 2 (because σ1 = C2 = max{σ1 , C2 } by (5.2.12)). Repeating the process on C2 yields reflectors Sy , Sx such that σ2 0 Sy C2 Sx = , where σ2 ≥ C3 2 . 0 C3 If P2 and Q2 are the orthogonal matrices ⎛ σ1 1 0 1 0 P2 = Ry , Q2 = Rx , then P2 CQ2 = ⎝ 0 0 Sx 0 Sy 0 0 σ2 0 ⎞0 0 ⎠C3 in which σ1 ≥ σ2 ≥ C3 2 . Continuing for r − 1 times produces orthogonal matrices Pr−1 and Qr−1 such that Pr−1 CQr−1 = diag (σ1 , σ2 , . . . , σr) = D, ˜ T and V ˜ are the orthogonal matrices where σ1 ≥ σ2 ≥ · · · ≥ σr . If U ˜ T = Pr−1 0 UT and V ˜ T AV ˜ = V Qr−1 0 , then U ˜ = D 0 , U 0 I 0 I 0 0 and thus the singular value decomposition (SVD) is derived. 57 57 The SVD has been independently discovered and rediscovered several times. Those credited with the early developments include Eugenio Beltrami (1835–1899) in 1873, M. E. Camille Jordan (1838–1922) in 1875, James J. Sylvester (1814–1897) in 1889, L. Autonne in 1913, and C. Eckart and G. Young in 1936. 412 Chapter 5 Norms, Inner Products, and Orthogonality Singular Value Decomposition For each A ∈ m×n of rank r, there are orthogonal matrices Um×m , Vn×n and a diagonal matrix Dr×r = diag (σ1 , σ2 , . . . , σr) such that A=U D 0 0 0 VT with σ1 ≥ σ2 ≥ · · · ≥ σr > 0. (5.12.2) m×n The σi 's are called the nonzero singular values of A. When r < p = min{m, n}, A is said to have p − r additional zero singular values. The factorization in (5.12.2) is called a singular value decomposition of A, and the columns in U and V are called left-hand and right-hand singular vectors for A, respectively. While the constructive method used to derive the SVD can be used as an algorithm, more sophisticated techniques exist, and all good matrix computation packages contain numerically stable SVD implementations. However, the details of a practical SVD algorithm are too complicated to be discussed at this point. The SVD is valid for complex matrices when ()T is replaced by ()∗ , and it can be shown that the singular values are unique, but the singular vectors are not. In the language of Chapter 7, the σi2 's are the eigenvalues of AT A, and the singular vectors are specialized sets of eigenvectors for AT A—see the summary on p. 555. In fact, the practical algorithm for computing the SVD is an implementation of the QR iteration (p. 535) that is cleverly applied to AT A without ever explicitly computing AT A. Singular values reveal something about the geometry of linear transformations because the singular values σ1 ≥ σ2 ≥ · · · ≥ σn of a matrix A tell us how much distortion can occur under transformation by A. They do so by giving us an explicit picture of how A distorts the unit sphere. To develop this, suppose that A ∈ n×n is nonsingular (Exercise 5.12.5 treats the singular and rectangular case), and let S2 = {x | x2 = 1} be the unit 2-sphere in n . The nature of the image A(S2) is revealed by considering the singular value decompositions A = UDVT and A−1 = VD−1 UT with D = diag (σ1 , σ2 , . . . , σn) , where U and V are orthogonal matrices. For each y ∈ A(S2) there is an x ∈ S2 such that y = Ax, so, with w = UT y, 2 2 2 2 2 1 = x = A−1 Ax = A−1 y = VD−1 UT y = D−1 UT y 2 2 = D−1 w 2 = w12 σ12 + 2 w22 σ22 + · · · + wr2 . σr2 2 2 (5.12.3) 5.12 Singular Value Decomposition 413 This means that UT A(S2) is an ellipsoid whose k th semiaxis has length σk . Because orthogonal transformations are isometries (length preserving transformations), UT can only affect the orientation of A(S2) , so A(S2) is also an ellipsoid whose k th semiaxis has length σk . Furthermore, (5.12.3) implies that the ellipsoid UT A(S2) is in standard position—i.e., its axes are directed along the standard basis vectors ek . Since U maps UT A(S2) to A(S2), and since Uek = U∗k , it follows that the axes of A(S2) are directed along the left-hand singular vectors defined by the columns of U. Therefore, the k th semiaxis of A(S2) is σk U∗k . Finally, since AV = UD implies AV∗k = σk U∗k , the righthand singular vector V∗k is a point on S2 that is mapped to the k th semiaxis vector on the ellipsoid A(S2). The picture in 3 looks like Figure 5.12.1. σ2 U∗2 V∗2 1 V∗1 σ1 U∗1 V∗3 A σ3 U∗3 Figure 5.12.1 The degree of distortion of the unit sphere under transformation by A is therefore measured by κ2 = σ1 /σn , the ratio of the largest singular value to the smallest singular value. Moreover, from the discussion of induced matrix norms (p. 280) and the unitary invariance of the 2-norm (Exercise 5.6.9), max Ax2 = A2 = UDVT 2 = D2 = σ1 x2 =1 and min Ax2 = x2 =1 1 1 1 = = = σn . A−1 2 VD−1 UT 2 D−1 2 In other words, longest and shortest vectors on A(S2) have respective lengths σ1 = A2 and σn = 1/ A−1 2 (this justifies Figure 5.2.1 on p. 281), so κ2 = A2 A−1 2 . This is called the 2-norm condition number of A. Different norms result in condition numbers with different values but with more or less the same order of magnitude as κ2 (see Exercise 5.12.3), so the qualitative information about distortion is the same. Below is a summary. 414 Chapter 5 Norms, Inner Products, and Orthogonality Image of the Unit Sphere For a nonsingular An×n having singular values σ1 ≥ σ2 ≥ · · · ≥ σn and an SVD A = UDVT with D = diag (σ1 , σ2 , . . . , σn) , the image of the unit 2-sphere is an ellipsoid whose k th semiaxis is given by σk U∗k (see Figure 5.12.1). Furthermore, V∗k is a point on the unit sphere such that AV∗k = σk U∗k . In particular, • σ1 = AV∗1 2 = max Ax2 = A2 , (5.12.4) x2 =1 • σn = AV∗n 2 = min Ax2 = 1/A−1 2 . x2 =1 (5.12.5) The degree of distortion of the unit sphere under transformation by A is measured by the 2-norm condition number σ1 • κ2 = = A2 A−1 2 ≥ 1. (5.12.6) σn Notice that κ2 = 1 if and only if A is an orthogonal matrix. The amount of distortion of the unit sphere under transformation by A determines the degree to which uncertainties in a linear system Ax = b can be magnified. This is explained in the following example. Example 5.12.1 Uncertainties in Linear Systems. Systems of linear equations Ax = b arising in practical work almost always come with built-in uncertainties due to modeling errors (because assumptions are almost always necessary), data collection errors (because infinitely √ precise gauges don't exist), and data entry errors (because numbers like 2, π, and 2/3 can't be entered exactly). In addition, roundoff error in floating-point computation is a prevalent source of uncertainty. In all cases it's important to estimate the degree of uncertainty in the solution of Ax = b. This is not difficult when A is known exactly and all uncertainty resides in the right-hand side. Even if this is not the case, it's sometimes possible to aggregate uncertainties and shift all of them to the right-hand side. Problem: Let Ax = b be a nonsingular system in which A is known exactly ˜ Estimate but b is subject to an uncertainty e, and consider A˜ x = b − e = b. 58 ˜ / x in x in terms of the relative uncertainty the relative uncertainty x − x ˜ b = e / b in b. Use any vector norm and its induced matrix b − b/ norm (p. 280). 58 ˜ by itself may not be meaningful. For example, an Knowing the absolute uncertainty x − x absolute uncertainty of a half of an inch might be fine when measuring the distance between the earth and the moon, but it's not good in the practice of eye surgery. 5.12 Singular Value Decomposition 415 ˜ = A−1 e to write Solution: Use b = Ax ≤ A x with x − x −1 A e A A−1 e ˜ x − x e = ≤ =κ , (5.12.7) x x b b where κ = A A−1 is a condition number as discussed earlier (κ = σ1 /σn ˜) ≤ A (x − x ˜) and if the 2-norm is used). Furthermore, e = A(x − x x x ≤ A−1 b imply ˜ x − x e e 1 e ≥ ≥ = . x A x A A−1 b κ b This with (5.12.7) yields the following bounds on the relative uncertainty: κ−1 ˜ e x − x e ≤ ≤κ , b x b where κ = A A−1 . (5.12.8) In other words, when A is well conditioned (i.e., when κ is small—see the rule of thumb in Example 3.8.2 to get a feeling of what "small" and "large" might mean), (5.12.8) insures that small relative uncertainties in b cannot greatly affect the solution, but when A is ill conditioned (i.e., when κ is large), a relatively small uncertainty in b might result in a relatively large uncertainty in x. To be more sure, the following problem needs to be addressed. Problem: Can equality be realized in each bound in (5.12.8) for every nonsingular A, and if so, how? Solution: Use the 2-norm, and let A = UDVT be an SVD so AV∗k = σk U∗k for each k. If b and e are directed along left-hand singular vectors associated with σ1 and σn , respectively—say, b = βU∗1 and e = U∗n , then x = A−1 b = A−1 (βU∗1) = βV∗1 σ1 and ˜ = A−1 e = A−1 (U∗n) = x− x V∗n , σn so ˜ 2 x − x = x2 σ1 σn e2 || = κ2 |β| b2 when b = βU∗1 and e = U∗n . Thus the upper bound (the worst case) in (5.12.8) is attainable for all A. The lower bound (the best case) is realized in the opposite situation when b and e are directed along U∗n and U∗1 , respectively. If b = βU∗n and e = U∗1 , ˜ = σ1−1 V∗1 , so then the same argument yields x = σn−1 βV∗n and x − x ˜ 2 x − x = x2 σn σ1 e2 || = κ−1 2 |β| b2 when b = βU∗n and e = U∗1 . 416 Chapter 5 Norms, Inner Products, and Orthogonality Therefore, if A is well conditioned, then relatively small uncertainties in b can't produce relatively large uncertainties in x. But when A is ill conditioned, it's possible for relatively small uncertainties in b to have relatively large effects on x, and it's also possible for large uncertainties in b to have almost no effect on x. Since the direction of e is almost always unknown, we must guard against the worst case and proceed with caution when dealing with ill-conditioned matrices. Problem: What if there are uncertainties in both sides of Ax = b? Solution: Use calculus to analyze the situation by considering the entries of A = A(t) and b = b(t) to be differentiable functions of a variable t, and compute the relative size of the derivative of x = x(t) by differentiating b = Ax to obtain b = (Ax) = A x + Ax (with denoting d /dt), so x = A−1 b − A−1 A x ≤ A−1 b + A−1 A x ≤ A−1 b + A−1 A x . Consequently, −1 A b x ≤ + A−1 A x x b A ≤ A A−1 + A A−1 A x A b A b A ≤κ +κ =κ + . b A b A In other words, the relative sensitivity of the solution is the sum of the relative sensitivities of A and b magnified by κ = A A−1 . A discrete analog of the above inequality is developed in Exercise 5.12.12. Conclusion: In all cases, the credibility of the solution to Ax = b in the face of uncertainties must be gauged in relation to the condition of A. As the next example shows, the condition number is pivotal also in determining whether or not the residual r = b − A˜ x is a reliable indicator of the ˜. accuracy of an approximate solution x Example 5.12.2 ˜ is a computed (or otherwise approxiChecking an Answer. Suppose that x mate) solution for a nonsingular system Ax = b, and suppose the accuracy of ˜ is "checked" by computing the residual r = b − A˜ x x. If r = 0, exactly, ˜ must be the exact solution. But if r is not exactly zero—say, r2 is then x ˜ is accurate to roughly t zero to t significant digits—are we guaranteed that x significant figures? This question was briefly examined in Example 1.6.3, but it's worth another look. Problem: To what extent does the size of the residual reflect the accuracy of an approximate solution? 5.12 Singular Value Decomposition 417 Solution: Without realizing it, we answered this question in Example 5.12.1. ˜ relative to the exact solution x, write r = b − A˜ To bound the accuracy of x x as A˜ x = b − r, and apply (5.12.8) with e = r to obtain ˜ r2 r2 x − x κ−1 ≤ , where κ = A2 A−1 2 . (5.12.9) ≤κ b2 x b2 Therefore, for a well-conditioned A, the residual r is relatively small if and ˜ is relatively accurate. However, as demonstrated in Example 5.12.1, only if x equality on either side of (5.12.9) is possible, so, when A is ill conditioned, a ˜ can produce a small residual r, and a very very inaccurate approximation x

AA† = A† AAT for all A ∈ m×n . A† = AT (AAT)† = (AT A)† AT for all A ∈ m×n. R A† = R AT = R A† A , and N A† = N AT = N AA† . (PAQ)† = QT A† PT when P and Q are orthogonal matrices, † but in general (AB) = B† A† (the reverse-order law fails). (AT A)† = A† (AT)† and (AAT)† = (AT)† A† . (f) (g) (h) (i) (AT A)−1 AT AT (AAT)−1 when rank (Am×n) = n, when rank (Am×n) = m. 5.12.17. Explain why A† = AD if and only if A is an RPN matrix. 5.12.18. Let X, Y ∈ m×n be such that R (X) ⊥ R (Y). (a) Establish the Pythagorean theorem for matrices by proving 2 2 2 X + YF = XF + YF . (b) Give an example to show that the result of part (a) does not hold for the matrix 2-norm. (c) Demonstrate that A† is the best approximate inverse for A in the sense that A† is the matrix of smallest Frobenius norm that minimizes I − AXF . 5.13 Orthogonal Projection 5.13 429 ORTHOGONAL PROJECTION As discussed in §5.9, every pair of complementary subspaces defines a projector. But when the complementary subspaces happen to be orthogonal complements, the resulting projector has some particularly nice properties, and the purpose of this section is to develop this special case in more detail. Discussions are in the context of real spaces, but generalizations to complex spaces are straightforward by replacing ()T by ()∗ and "orthogonal matrix" by "unitary matrix." If M is a subspace of an inner-product space V, then V = M ⊕ M⊥ by (5.11.1), and each v ∈ V can be written uniquely as v = m + n, where m ∈ M and n ∈ M⊥ by (5.9.3). The vector m was defined on p. 385 to be the projection of v onto M along M⊥ , so the following definitions are natural. Orthogonal Projection For v ∈ V, let v = m + n, where m ∈ M and n ∈ M⊥ . • m is called the orthogonal projection of v onto M. • The projector PM onto M along M⊥ is called the orthogonal projector onto M. • PM is the unique linear operator such that PM v = m (see p. 386). These ideas are illustrated illustrated in Figure 5.13.1 for V = 3 . Figure 5.13.1 Given an arbitrary pair of complementary subspaces M, N of n , formula (5.9.12) on p. 386 says that the projector P onto M along N is given by I 0 −1 −1 P = M|N M|N = M|0 M|N , (5.13.1) 0 0 where the columns of M and N constitute bases for M and N , respectively. So, how does this expression simplify when N = M⊥ ? To answer the question, 430 Chapter 5 Norms, Inner Products, and Orthogonality T observe that if N = M⊥ , then NT M = 0 and N M = 0. Furthermore, if T T dim M = r, then M M is r × r, and rank M M = rank (M) = r by (4.5.4), so MT M is nonsingular. Therefore, if the columns of N are chosen to be an orthonormal basis for M⊥ , then ⎛ ⎛ −1 T ⎞−1 T ⎞ M MT M M M −1 ⎠ ⎠M|N = 0 ⎠. ⇐= M | N =⎝ 0 I NT NT This together with (5.13.1) says the orthogonal projector onto M is given by ⎛ −1 T ⎞ M MT M ⎠ = M MT M −1 MT . PM = M | 0 ⎝ (5.13.2) NT As discussed in §5.9, the projector associated with any given pair of complementary subspaces is unique, and it doesn't matter which bases are used to −1 T form M and N in (5.13.1). Consequently, formula PM = M MT M M is independent of the choice of M —just as long as its columns constitute some basis for M. In particular, the columns of M need not be an orthonormal basis for M. But if they are, then MT M = I, and (5.13.2) becomes PM = MMT . Moreover, if the columns of M and N constitute orthonormal bases for M and M⊥ , respectively, then U = M | N is an orthogonal matrix, and (5.13.1) becomes Ir 0 PM = U UT . 0 0 In other words, every orthogonal projector is orthogonally similar to a diagonal matrix in which the diagonal entries are 1's and 0's. Below is a summary of the formulas used to build orthogonal projectors. Constructing Orthogonal Projectors Let M be an r-dimensional subspace of n , and let the columns of Mn×r and Nn×n−r be bases for M and M⊥ , respectively. The orthogonal projectors onto M and M⊥ are −1 T −1 T • PM = M MT M M and PM⊥ = N NT N N . (5.13.3) If M and N contain orthonormal bases for M and M⊥ , then • • PM = MMT and PM⊥ = NNT . Ir 0 PM = U UT , where U = M | N . 0 0 • PM⊥ = I − PM in all cases. Note: Extensions of (5.13.3) appear on p. 634. (5.13.4) (5.13.5) (5.13.6) 5.13 Orthogonal Projection 431 Example 5.13.1 Problem: Let un×1 = 0, and consider the line L = span {u} . Construct the orthogonal projector onto L, and then determine the orthogonal projection of a vector xn×1 onto L. Solution: The vector u by itself is a basis for L, so, according to (5.13.3), −1 T uuT PL = u uT u u = T u u is the orthogonal projector onto L. The orthogonal projection of a vector x onto L is therefore given by T uuT u x PL x = T x = u. u. u uT u Note: If u2 = 1, then PL = uuT , so PL x = uuT x = (uT x)u, and PL x2 = |uT x| u2 = |uT x|. This yields a geometrical interpretation for the magnitude of the standard inner product. It says that if u is a vector of unit length in L, then, as illustrated in Figure 5.13.2, |uT x| is the length of the orthogonal projection of x onto the line spanned by u. x L PL x 0 T |u x| Figure 5.13.2 Finally, notice that since PL = uuT is the orthogonal projector onto L, it must be the case that PL⊥ = I − PL = I − uuT is the orthogonal projector onto L⊥ . This was called an elementary orthogonal projector on p. 322—go back and reexamine Figure 5.6.1. Example 5.13.2 Volume, Gram–Schmidt, and QR. A solid in m with parallel opposing faces whose adjacent sides are defined by vectors from a linearly independent set {x1 , x2 , . . . , xn } is called an n-dimensional parallelepiped. As shown in the shaded portions of Figure 5.13.3, a two-dimensional parallelepiped is a parallelogram, and a three-dimensional parallelepiped is a skewed rectangular box. 432 Chapter 5 Norms, Inner Products, and Orthogonality x3 x1 x2 x1 x2 (I − P2)x2 (I − P3)x3 x1 Figure 5.13.3 Problem: Determine the volumes of a two-dimensional and a three-dimensional parallelepiped, and then make the natural extension to define the volume of an n-dimensional parallelepiped. Solution: In the two-dimensional case, volume is area, and it's evident from Figure 5.13.3 that the area of the shaded parallelogram is the same as the area of the dotted rectangle. The width of the dotted rectangle is ν1 = x1 2 , and the height is ν2 = (I − P2)x2 2 , where P2 is the orthogonal projector onto the space (line) spanned by x1 , and I − P2 is the orthogonal projector onto ⊥ span {x1 } . In other words, the area, V2 , of the parallelogram is the length of its base times its projected height, ν2 , so V2 = x1 2 (I − P2)x2 2 = ν1 ν2 . Similarly, the volume of a three-dimensional parallelepiped is the area of its base times its projected height. The area of the base was just determined to be V2 = x1 2 (I − P2)x2 2 = ν1 ν2 , and it's evident from Figure 5.13.3 that the projected height is ν3 = (I − P3)x3 2 , where P3 is the orthogonal projector onto span {x1 , x2 } . Therefore, the volume of the parallelepiped generated by {x1 , x2 , x3 } is V3 = x1 2 (I − P2)x2 2 (I − P3)x3 2 = ν1 ν2 ν3 . It's now clear how to inductively define V4 , V5 , etc. In general, the volume of the parallelepiped generated by a linearly independent set {x1 , x2 , . . . , xn } is Vn = x1 2 (I − P2)x2 2 (I − P3)x3 2 · · · (I − Pn)xn 2 = ν1 ν2 · · · νn , where Pk is the orthogonal projector onto span {x1 , x2 , . . . , xk−1 } , and where ν1 = x1 2 and νk = (I − Pk)xk 2 for k > 1. (5.13.7) Note that if {x1 , x2 , . . . , xn } is an orthogonal set, Vn = x1 2 x2 2 · · · xn 2 , which is what we would expect. 5.13 Orthogonal Projection 433 Connections with Gram–Schmidt and QR. Recall from (5.5.4) on p. 309 that the vectors in the Gram–Schmidt sequence generated from a linearly independent set {x1 , x2 , . . . , xn } ⊂ m are u1 = x1 / x1 2 and I − Uk UTk xk uk = I − Uk UT xk , k where Uk = u1 | u2 | · · · | uk−1 for k > 1. 2 Since {u1 , u2 , . . . , uk−1 } is an orthonormal basis for span {x1 , x2 , . . . , xk−1 } , it follows from (5.13.4) that Uk UTk must be the orthogonal projector onto T T span k = Pk and (I − Pk)xk = (I − Uk Uk)xk , } . Hence Uk U {x1 , x2 , . . .T. , xk−1 th so I − Uk Uk xk 2 = νk is the k projected height in (5.13.7). This means that when the Gram–Schmidt equations are written in the form of a QR factorization as explained on p. 311, the diagonal elements of the upper-triangular matrix R are the νk 's. Consequently, the product of the diagonal entries in R is the volume of the parallelepiped generated by the xk 's. But the QR factor ization of A = x1 | x2 | · · · | xn is unique (Exercise 5.5.8), so it doesn't matter whether Gram–Schmidt or another method is used to determine the QR factors. Therefore, we arrive at the following conclusion. • If Am×n = Qm×n Rn×n is the (rectangular) QR factorization of a matrix with linearly independent columns, then the volume of the n-dimensional parallelepiped generated by the columns of A is Vn = ν1 ν2 · · · νn , where the νk 's are the diagonal elements of R. We will see on p. 468 what this means in terms of determinants. Of course, not all projectors are orthogonal projectors, so a natural question to ask is, "What characteristic features distinguish orthogonal projectors from more general oblique projectors?" Some answers are given below. Orthogonal Projectors Suppose that P ∈ n×n is a projector—i.e., P2 = P. The following statements are equivalent to saying that P is an orthogonal projector. • R (P) ⊥ N (P). • P =P • P2 = 1 for the matrix 2-norm (p. 281). T (5.13.8) (i.e., orthogonal projector ⇐⇒ P = P = P). (5.13.9) 2 T (5.13.10) Proof. Every projector projects vectors onto its range along (parallel to) its nullspace, so statement (5.13.8) is essentially a restatement of the definition of an orthogonal projector. To prove (5.13.9), note that if P is an orthogonal projector, then (5.13.3) insures that P is symmetric. Conversely, if a projector 434 Chapter 5 Norms, Inner Products, and Orthogonality P is symmetric, then it must be an orthogonal projector because (5.11.5) on p. 405 allows us to write P = PT =⇒ R (P) = R PT =⇒ R (P) ⊥ N (P). To see why (5.13.10) characterizes projectors that are orthogonal, refer back to Example 5.9.2 on p. 389 (or look ahead to (5.15.3)) and note that P2 = 1/ sin θ, where θ is the angle between R (P) and N (P). This makes it clear that P2 ≥ 1 for all projectors, and P2 = 1 if and only if θ = π/2, (i.e., if and only if R (P) ⊥ N (P)). Example 5.13.3 Problem: For A ∈ m×n such that rank (A) = r, describe the orthogonal projectors onto each of the four fundamental subspaces of A. Solution 1: Let Bm×r and Nn×n−r be matrices whose columns are bases for R (A) and N (A), respectively—e.g., B might contain the basic columns of ⊥ A. The orthogonal decomposition theorem on p. 405 says R (A) = N AT ⊥ and N (A) = R AT , so, by making use of (5.13.3) and (5.13.6), we can write −1 T PR(A) = B BT B B , −1 T PN (AT) = PR(A)⊥ = I − PR(A) = I − B BT B B , −1 T PN (A) = N NT N N , −1 T PR(AT) = PN (A)⊥ = I − PN (A) = I − N NT N N . Note: If rank (A) = n, then all columns of A are basic and −1 T PR(A) = A AT A A . (5.13.11) Solution 2: Another way to describe these projectors is to make use of the Moore–Penrose pseudoinverse A† (p. 423). Recall that if A has a URV factorization −1 C 0 0 C T † A=U V , then A = V UT , 0 0 0 0 where U = U1 | U2 and V = V1 | V2 are orthogonal matrices in which the columns of U1 and V1 constitute orthonormal bases for R (A) and R AT , respectively, and the columns of U2 and V2 are orthonormal bases for N AT and N (A), respectively. Computing the products AA† and A† A reveals I 0 I 0 † T T † AA = U U = U1 U1 and A A = V VT = V1 V1T , 0 0 0 0 5.13 Orthogonal Projection 435 so, according to (5.13.4), PR(A) = U1 UT1 = AA† , PN (AT) = I − PR(A) = I − AA† , PR(AT) = V1 V1T = A† A, PN (A) = I − PR(AT) = I − A† A. (5.13.12) The notion of orthogonal projection in higher-dimensional spaces is consistent with the visual geometry in 2 and 3 . In particular, it is visually evident from Figure 5.13.4 that if M is a subspace of 3 , and if b is a vector outside of M, then the point in M that is closest to b is p = PM b, the orthogonal projection of b onto M. b min b − m2 m∈M 0 p = PM b Figure 5.13.4 The situation is exactly the same in higher dimensions. But rather than using our eyes to understand why, we use mathematics—it's surprising just how easy it is to "see" such things in abstract spaces. Closest Point Theorem Let M be a subspace of an inner-product space V, and let b be a vector in V. The unique vector in M that is closest to b is p = PM b, the orthogonal projection of b onto M. In other words, min b − m2 = b − PM b2 = dist (b, M). (5.13.13) m∈M This is called the orthogonal distance between b and M. Proof. If p = PM b, then p − m ∈ M for all m ∈ M, and b − p = (I − PM)b ∈ M⊥ , 2 2 so (p − m) ⊥ (b − p). The Pythagorean theorem says x + y = x + y whenever x ⊥ y (recall Exercise 5.4.14), and hence 2 2 2 2 2 b − m2 = b − p + p − m2 = b − p2 + p − m2 ≥ p − m2 . 2 436 Chapter 5 Norms, Inner Products, and Orthogonality In other words, minm∈M b − m2 = b − p2 . Now argue that there is not : ∈ M such that another point in M that is as close to b as p is. If m : 2 = b − p2 , then by using the Pythagorean theorem again we see b − m 2 2 2 2 : 2 = b − p + p − m : 2 = b − p2 + p − m : 2 =⇒ p − m : 2 = 0, b − m : = p. and thus m Example 5.13.4 n×n To illustrate some with the inner product Tof the previous ideas, consider A B = trace A B . If Sn is the subspace of n × n real-symmetric matrices, then each of the following statements is true. • Sn⊥ = the subspace Kn of n × n skew-symmetric matrices. Sn ⊥ Kn because for all S ∈ Sn and K ∈ Kn , T S K = trace ST K = −trace SKT = −trace SKT = −trace KST = −trace ST K = − S K =⇒ S K = 0. n×n = Sn ⊕ Kn because every A ∈ n×n can be uniquely expressed as the sum of a symmetric and a skew-symmetric matrix by writing A= A + AT A − AT + 2 2 (recall (5.9.3) and Exercise 3.2.6). • The orthogonal projection of A ∈ n×n onto Sn is P(A) = (A + AT)/2. • The closest symmetric matrix to A ∈ n×n is P(A) = (A + AT)/2. • The distance from A ∈ n×n to Sn (the deviation from symmetry) is dist(A, Sn) = A−P(A)F = (A−AT)/2 F = trace (AT A)−trace (A2) . 2 Example 5.13.5 Affine Projections. If v = 0 is a vector in a space V, and if M is a subspace of V, then the set of points A = v + M is called an affine space in V. Strictly speaking, A is not a subspace (e.g., it doesn't contain 0), but, as depicted in Figure 5.13.5, A is the translate of a subspace—i.e., A is just a copy of M that has been translated away from the origin through v. Consequently, notions such as projection onto A and points closest to A are analogous to the corresponding concepts for subspaces. 5.13 Orthogonal Projection 437 Problem: For b ∈ V, determine the point p in A = v + M that is closest to b. In other words, explain how to project b orthogonally onto A. Solution: The trick is to subtract v from b as well as from everything in A to put things back into the context of subspaces where we already know the answers. As illustrated in Figure 5.13.5, this moves A back down to M, and it translates v → 0, b → (b − v), and p → (p − v). b A=v+M p b−v v M p-v p−v 0 0 Figure 5.13.5 If p is to be the orthogonal projection of b onto A, then p − v must be the orthogonal projection of b − v onto M, so p − v = PM (b − v) =⇒ p = v + PM (b − v), (5.13.14) and thus p is the point in A that is closest to b. Applications to the solution of linear systems are developed in Exercises 5.13.17–5.13.22. We are now in a position to replace the classical calculus-based theory of least squares presented in §4.6 with a more modern vector space development. In addition to being straightforward, the modern geometrical approach puts the entire least squares picture in much sharper focus. Viewing concepts from more than one perspective generally produces deeper understanding, and this is particularly true for the theory of least squares. Recall from p. 226 that for an inconsistent system Am×n x = b, the object of the least squares problem is to find vectors x that minimize the quantity 2 (Ax − b)T (Ax − b) = Ax − b2 . (5.13.15) The classical development in §4.6 relies on calculus to argue that the set of vectors x that minimize (5.13.15) is exactly the set that solves the (always consistent) system of normal equations AT Ax = AT b. In the context of the closest point theorem the least squares problem asks for vectors x such that Ax is as close 438 Chapter 5 Norms, Inner Products, and Orthogonality to b as possible. But Ax is always a vector in R (A), and the closest point theorem says that the vector in R (A) that is closest to b is PR(A) b, the orthogonal projection of b onto R (A). Figure 5.13.6 illustrates the situation in 3 . b min Ax − b2 = PR(A) b − b2 x∈n R (A) 0 PR(A) b Figure 5.13.6 So the least squares problem boils down to finding vectors x such that Ax = PR(A) b. But this system is equivalent to the system of normal equations because Ax = PR(A) b ⇐⇒ PR(A) Ax = PR(A) b ⇐⇒ PR(A) (Ax − b) = 0 ⊥ ⇐⇒ (Ax − b) ∈ N PR(A) = R (A) = N AT ⇐⇒ AT (Ax − b) = 0 ⇐⇒ AT Ax = AT b. Characterizing the set of least squares solutions as the solutions to Ax = PR(A) b makes it obvious that x = A† b is a particular least squares solution because (5.13.12) insures AA† = PR(A) , and thus A(A† b) = PR(A) b. Furthermore, since A† b is a particular solution of Ax = PR(A) b, the general solution—i.e., the set of all least squares solutions—must be the affine space S = A† b + N (A). Finally, the fact that A† b is the least squares solution of minimal norm follows from Example 5.13.5 together with ⊥ R A† = R AT = N (A) (see part (g) of Exercise 5.12.16) because (5.13.14) insures that the point in S that is closest to the origin is p = A† b + PN (A) (0 − A† b) = A† b. The classical development in §4.6 based on partial differentiation is not easily generalized to cover the case of complex matrices, but the vector space approach given in this example trivially extends to complex matrices by simply replacing ()T by ()∗ . Below is a summary of some of the major points concerning the theory of least squares. 5.13 Orthogonal Projection 439 Least Squares Solutions : is a Each of the following four statements is equivalent to saying that x least squares solution for a possibly inconsistent linear system Ax = b. • A: x − b2 = minn Ax − b2 . (5.13.16) x∈ • • • A: x = PR(A) b. T T A A: x=A b † (5.13.17) ∗ ∗ (A A: x = A b when A ∈ C m×n †). : ∈ A b + N (A) (A b is the minimal 2-norm LSS). x (5.13.18) (5.13.19) Caution! These are valuable theoretical characterizations, but none is recommended for floating-point computation. Directly solving (5.13.17) or (5.13.18) or explicitly computing A† can be inefficient and numerically unstable. Computational issues are discussed in Example 4.5.1 on p. 214; Example 5.5.3 on p. 313; and Example 5.7.3 on p. 346. The least squares story will not be complete until the following fundamental question is answered: "Why is the method of least squares the best way to make estimates of physical phenomena in the face of uncertainty?" This is the focal point of the next section. Exercises for section 5.13 5.13.1. Find the orthogonal projection of b onto M = span {u} , and then deT termine the orthogonal projection of b onto M⊥ , where b = (4 8) T and u = (3 1) . ⎛ ⎞⎛ ⎞1 2 0 1 5.13.2. Let A = ⎝ 2 4 1 ⎠and b = ⎝ 1 ⎠. 1 2 0 1 (a) Compute the orthogonal projectors onto each of the four fundamental subspaces associated with A. ⊥ (b) Find the point in N (A) that is closest to b. 5.13.3. For an orthogonal projector P, prove that Px2 = x2 if and only if x ∈ R (P). 5.13.4. Explain why AT PR(A) = AT for all A ∈ m×n. 440 Chapter 5 Norms, Inner Products, and Orthogonality r 5.13.5. Explain why PM = i=1 ui ui T whenever B = {u1 , u2 , . . . , ur } is an orthonormal basis for M ⊆ n×1 . 5.13.6. Explain how to use orthogonal reduction techniques to compute the orthogonal projectors onto each of the four fundamental subspaces of a matrix A ∈ m×n . 5.13.7. (a) Describe all 2 × 2 orthogonal projectors in 2×2 . (b) Describe all 2 × 2 projectors in 2×2 . 5.13.8. The line L in n passing through two distinct points u and v is L = u + span {u − v} . If u = 0 and v = αu, then L is a line not passing through the origin—i.e., L is not a subspace. Sketch a picture in 2 or 3 to visualize this, and then explain how to project a vector b orthogonally onto L. : is a least 5.13.9. Explain why x squares solution for Ax = b if and only if A: x − b2 = PN (AT) b 2 . : is a least squares solution for 5.13.10. Prove that if ε = A: x − b, where x 2 2 2 Ax = b, then ε2 = b2 − PR(A) b 2 . 5.13.11. Let M be an r -dimensional subspace of n . We know from (5.4.3) that if B = {u1 , u2 , . . . , ur } is an orthonormal basis for M, and if x ∈ M, thenx is equal to its Fourier expansion with respect to B. r That is, x = i=1 (ui T x)ui . However, if x ∈ / M, then equality is not possible (why?), so the question that arises is, "What does the Fourier expansion on the right-hand side of this expression represent?" Answer r T this question by showing that the Fourier expansion (u x)ui is i i=1 the point in M that is closest to x in the euclidean norm. In other r T words, show that (u x)u = P x. i i M i=1 5.13.12. Determine the orthogonal projection of b onto M, where ⎛ ⎞5 ⎜2 ⎟b=⎜ ⎟5 3 and ⎡⎛ ⎞−3/5 ⎞ ⎤ ⎛ 0 ⎞M = span ⎝ ⎠, ⎣ 4/5 0 ⎝ ⎞ 0 ⎝0 ⎜⎝ ⎠, 0 1 Hint: Is this spanning set in fact an orthonormal basis? ⎟⎞ 4/5 ⎟ ⎜⎟ 0 ⎜⎠. 3/5 ⎦ ⎠ 0 ⎝ 5.13 Orthogonal Projection 441 5.13.13. Let M and N be subspaces of a vector space V, and consider the associated orthogonal projectors PM and PN . (a) Prove that PM PN = 0 if and only if M ⊥ N . (b) Is it true that PM PN = 0 if and only if PN PM = 0? Why? 5.13.14. Let M and N be subspaces of the same vector space, and let PM and PN be orthogonal projectors onto M and N , respectively. (a) Prove that R (PM + PN) = R (PM) + R (PN) = M + N . Hint: Use Exercise 4.2.9 along with (4.5.5). (b) Explain why M ⊥ N if and only if PM PN = 0. (c) Explain why PM + PN is an orthogonal projector if and only if PM PN = 0, in which case R (PM + PN) = M ⊕ N and M ⊥ N . Hint: Recall Exercise 5.9.17. 59 5.13.15. Anderson–Duffin Formula. Prove that if M and N are subspaces of the same vector space, then the orthogonal projector onto M ∩ N is given by PM∩N = 2PM (PM + PN)† PN . Hint: Use (5.13.12) and Exercise 5.13.14 to show PM (PM + PN)† PN = PN (PM + PN)† PM . Argue that if Z = 2PM (PM +PN)† PM , then Z = PM∩N Z = PM∩N . 5.13.16. Given a square matrix X, the matrix exponential eX is defined as eX = I + X + ∞ Xn X3 X2 + + ··· = . 2! 3! n! n=0 It can be shown that this series converges for all X, and it is legitimate to differentiate and integrate itterm by term to produce the statements deAt /dt = AeAt = eAt A and eAt A dt = eAt . T (a) Use the fact that limt→∞ e−A At = 0 for all A ∈ m×n to T ∞ show A† = 0 e−A At AT dt. ∞ k+1 k+1 (b) If limt→∞ e−A t = 0, show AD = 0 e−A t Ak dt, where 60 k = index(A). −At (c) For nonsingular = 0, then ∞ −At matrices, show that if limt→∞ e −1 dt. A = 0 e 59 60 W. N. Anderson, Jr., and R. J. Duffin discovered this formula for the orthogonal projector onto an intersection in 1969. They called PM (PM + PN)† PN the parallel sum of PM and PN because it is the matrix generalization of the scalar function r1 r2 /(r1 + r2) = r1 (r1 + r2)−1 r2 that is the resistance of a circuit composed of two resistors r1 and r2 connected in parallel. The simple elegance of the Anderson–Duffin formula makes it one of the innumerable little sparkling facets in the jewel that is linear algebra. A more useful integral representation for AD is given in Exercise 7.9.22 (p. 615). 442 Chapter 5 Norms, Inner Products, and Orthogonality 5.13.17. An affine space v + M ⊆ n for which dim M = n − 1 is called a hyperplane. For example, a hyperplane in 2 is a line (not necessarily through the origin), and a hyperplane in 3 is a plane (not necessarily through the origin). The ith equation Ai∗ x = bi in a linear system Am×n x = b is a hyperplane in n , so the solutions of Ax = b occur at the intersection of the m hyperplanes defined by the rows of A. (a) Prove that for a given scalar β and a nonzero vector u ∈ n , the set H = {x | uT x = β} is a hyperplane in n . (b) Explain why projection of b ∈ n onto H is T the orthogonal T p = b − u b − β/u u u. 5.13.18. For u, w ∈ n such that uT w = 0, let M = u⊥ and W = span {w} . (a) Explain why n = M ⊕ W. (b) For b ∈ n×1 , explain why the oblique projection of b onto M along W is given by p = b − uT b/uT ww. (c) For a given scalar β, let H be the hyperplane in n defined by H = {x | uT x = β}—see Exercise 5.13.17. Explain why the oblique projection of b onto H along W should be given by p = b − uT b − β/uT w w. 5.13.19. Kaczmarz's system 61 Projection Method. The solution of a nonsingular a11 a21 a12 a22 x1 x2 = b1 b2 is the intersection of the two hyperplanes (lines in this case) defined by H1 ={(x1 , x2) | a11 x1 + a12 x2 = b1 } , H2 = {(x1 , x2) | a21 x1 + a22 x2 = b2 } . It's visually evident that by starting with an arbitrary point p0 and alternately projecting orthogonally onto H1 and H2 as depicted in Figure 5.13.7, the resulting sequence of projections {p1 , p2 , p3 , p4 , . . . } converges to H1 ∩ H2 , the solution of Ax = b. 61 Although this idea has probably occurred to many people down through the ages, credit is usually given to Stefan Kaczmarz, who published his results in 1937. Kaczmarz was among a school of bright young Polish mathematicians who were beginning to flower in the first part of the twentieth century. Tragically, this group was decimated by Hitler's invasion of Poland, and Kaczmarz himself was killed in military action while trying to defend his country. 5.13 Orthogonal Projection 443 Figure 5.13.7 This idea can be generalized by using Exercise 5.13.17. For a consistent system An×r x = b with rank (A) = r, scale the rows so that Ai∗ 2 = 1 for each i, and let Hi = {x | Ai∗ x = bi } be the hyperplane defined by the ith equation. Begin with an arbitrary vector p0 ∈ r×1 , and successively perform orthogonal projections onto each hyperplane to generate the following sequence: T p1 = p0 − (A1∗ p0 − b1) (A1∗) (project p0 onto H1), T (project p1 onto H2), p2 = p1 − (A2∗ p1 − b2) (A2∗) .. . .. . T pn = pn−1 − (An∗ pn−1 − bn) (An∗) (project pn−1 onto Hn). When all n hyperplanes have been used, continue by repeating the process. For example, on the second pass project pn onto H1 ; then project pn+1 onto H2 , etc. For an arbitrary p0 , the entire Kaczmarz sequence is generated by executing the following double loop: For k = 0, 1, 2, 3, . . . For i = 1, 2, . . . , n T pkn+i = pkn+i−1 − (Ai∗ pkn+i−1 − bi) (Ai∗) Prove that the Kaczmarz sequence converges to the solution of Ax = b 2 2 2 by showing pkn+i − x2 = pkn+i−1 − x2 − (Ai∗ pkn+i−1 − bi) . 5.13.20. Oblique Projection Method. Assume that a nonsingular system An×n x = b has been row scaled so that Ai∗ 2 = 1 for each i, and let Hi = {x | Ai∗ x = bi } be the hyperplane defined by the ith equation—see Exercise 5.13.17. In theory, the system can be solved by making n−1 oblique projections of the type described in Exercise 5.13.18 because if an arbitrary point p1 in H1 is projected obliquely onto H2 along H1 to produce p2 , then p2 is in H1 ∩H2 . If p2 is projected onto H3 along H1 ∩ H2 to produce p3 , then p3 ∈ H1 ∩ H2 ∩ H3 , and so forth until pn ∈ ∩ni=1 Hi . This is similar to Kaczmarz's method given in Exercise 5.13.19, but here we are projecting obliquely instead of orthogonally. However, projecting pk onto Hk+1 along ∩ki=1 Hi is difficult because 444 Chapter 5 Norms, Inner Products, and Orthogonality ∩ki=1 Hi is generally unknown. This problem is overcome by modifying the procedure as follows—use Figure 5.13.8 with n = 3 as a guide. Figure 5.13.8 (1) (1) (1) ⊂ H1 such that Step 0. Begin with any set p1 , p2 , . . . , pn (1) (1) (1) (1) (1) (1) is linearly independent p1 − p2 , p1 − p3 , . . . , p1 − pn (1) (1) and A2∗ p1 − pk = 0 for k = 2, 3, . . . , n. (1) (1) (1) (1) Step 1. In turn, project p1 onto H2 through p2 , p3 , . . . , pn (2) (2) (2) produce p2 , p3 , . . . , pn ⊂ H1 ∩ H2 (see Figure 5.13.8). (2) (2) (2) to (2) Step 2. Project p2 onto H3 through p3 , p4 , . . . , pn to produce (3) (3) (3) p3 , p4 , . . . , pn ⊂ H1 ∩ H2 ∩ H3 . And so the process continues. (n−1) (n−1) (n) Step n−1. Project pn−1 through pn to produce pn ∈ ∩ni=1 Hi . (n) Of course, x = pn is the solution of the system. For any initial set {x1 , x2 , . . . , xn } ⊂ H1 satisfying the properties described in Step 0, explain why the following algorithm performs the computations described in Steps 1, 2, . . . , n − 1. For i = 2 to n For j = i to n xj ← xj − x ← xn (Ai∗ xi−1 − bi) (xi−1 − xj) Ai∗ (xi−1 − xj) (the solution of the system) 5.13.21. Let M be a subspace of n , and let R = I − 2PM . Prove that the orthogonal distance between any point x ∈ n and M⊥ is the same as the orthogonal distance between Rx and M⊥ . In other words, prove that R reflects everything in n about M⊥ . Naturally, R is called the reflector about M⊥ . The elementary reflectors I − 2uuT /uT u discussed on p. 324 are special cases—go back and look at Figure 5.6.2. 5.13 Orthogonal Projection 445 5.13.22. Cimmino's Reflection Method. In 1938 the Italian mathematician Gianfranco Cimmino used the following elementary observation to construct an iterative algorithm for solving linear systems. For a 2 × 2 system Ax = b, let H1 and H2 be the two lines (hyperplanes) defined by the two equations. For an arbitrary guess r0 , let r1 be the reflection of r0 about the line H1 , and let r2 be the reflection of r0 about the line H2 . As illustrated in Figure 5.13.9, the three points r0 , r1 , and r2 lie on a circle whose center is H1 ∩ H2 (the solution of the system). Figure 5.13.9 The mean value m = (r1 + r2)/2 is strictly inside the circle, so m is a better approximation to the solution than r0 . It's visually evident that iteration produces a sequence that converges to the solution of Ax = b. Prove this in general by using the following blueprint. (a) For a scalar β and a vector u ∈ n such that u2 = 1, consider the hyperplane H = {x | uT x = β} (Exercise 5.13.17). Use (5.6.8) to show that the reflection of a vector b about H is r = b − 2(uT b − β)u. (b) For a system Ax = b in which the rows of A ∈ n×r have been scaled so that Ai∗ 2 = 1 for each i, let Hi = {x | Ai∗ x = bi } be the hyperplane defined by the ith equation. If r0 ∈ r×1 is an arbitrary vector, and if ri is the reflection of r0 about Hi , explain why the mean value of the reflections {r1 , r2 , . . . , rn } is m = r0 − (2/n)AT ε, where ε = Ar0 − b. (c) Iterating part (b) produces mk = mk−1 − (2/n)AT εk−1 , where εk−1 = Amk−1 − b. Show that if A is nonsingular, and if k x = A−1 b, then x − mk = I − (2/n)AT A (x − m0). Note: k It can be proven that I − (2/n)AT A → 0 as k → ∞, so mk → x for all m0 . In fact, mk converges even if A is rank deficient—if consistent, it converges to a solution, and, if inconsistent, the limit is a least squares solution. Cimmino's method also works with weighted means. If W = diag (w1 , w2 , . . . , wn), where wi > 0 and wi = 1, then mk = mk−1 − ωAT Wεk−1 is a convergent sequence in which 0 < ω < 2 is a "relaxation parameter" that can be adjusted to alter the rate of convergence. 446 5.14 Chapter 5 Norms, Inner Products, and Orthogonality WHY LEAST SQUARES? Drawing inferences about natural phenomena based upon physical observations and estimating characteristics of large populations by examining small samples are fundamental concerns of applied science. Numerical characteristics of a phenomenon or population are often called parameters, and the goal is to design functions or rules called estimators that use observations or samples to estimate parameters of interest. For example, the mean height h of all people is a parameter of the world's population, and one way of estimating h is to observe the mean height of a sample of k people. In other words, if hi is the height of ˆ defined by the ith person in a sample, the function h k 1 ˆ 1 , h2 , . . . , hk) = h(h hi k i=1 ˆ is a linear estimator because h ˆ is a linear is an estimator for h. Moreover, h function of the observations. Good estimators should possess at least two properties—they should be unbiased and they should have minimal variance. For example, consider estimating the center of a circle drawn on a wall by asking Larry, Moe, and Curly to each throw one dart at the circle. To decide which estimator is best, we need to know more about each thrower's style. While being able to throw a tight pattern, it is known that Larry tends to have a left-hand bias in his style. Moe doesn't suffer from a bias, but he tends to throw a rather large pattern. However, Curly can throw a tight pattern without a bias. Typical patterns are shown below. Larry Moe Curly Although Larry has a small variance, he is an unacceptable estimator because he is biased in the sense that his average is significantly different than the center. Moe and Curly are each unbiased estimators because they have an average that is the center, but Curly is clearly

(A) j th ↓ a1k . . kth ↓ · · · a1k · · · a1n a11 · · · . . . n . . ˜ = ai1 · · · aik · · · aik · · · ain = 0 = det (A) aik ˚ Aij . . . . . i=1 . . . an1 · · · ank · · · ank · · · ann Thus the inner product of the k th column of An×n with the cofactors of the j th column of A is zero. A similar result holds for rows. Combining these observations with (6.2.5) and (6.2.6) produces n n det (A) if k = i, det (A) if k = j, ˚ ˚ akj Aij = and aik Aij = 0 if k = i, 0 if k = j, j=1 i=1 which is equivalent to saying that A[adj (A)] = [adj (A)]A = det (A) I. Example 6.2.9 Differential Equations and Determinants. A system of n homogeneous first-order linear differential equations d xi (t) = ai1 (t)x1 (t) + ai2 (t)x2 (t) + · · · + ain (t)xn (t), dt i = 1, 2, . . . , n can be expressed in matrix notation by writing x (t) a (t) a (t) · · · a (t) x (t) 11 12 1n 1 1 x2 (t) a21 (t) a22 (t) · · · a2n (t) x2 (t) . . . . xn (t) an1 (t) an2 (t) · · · ann (t) xn (t) or, equivalently, x = Ax. Let S = {w1 (t), w2 (t), . . . , wn (t)} be a set of n × 1 vectors that are solutions to x = Ax, and place these solutions as columns in a matrix W(t)n×n = [w1 (t) | w2 (t) | · · · | wn (t)] so that W = AW. Problem: Prove that if w(t) = det (W), (called the Wronskian (p. 474)), then t trace A(ξ) dξ w(t) = w(ξ0) e ξ0 , where ξ0 is an arbitrary constant. (6.2.8) 482 Chapter 6 Determinants n Solution: By (6.1.19), d w(t)/dt = i=1 det (Di), where w w12 · · · w1n 11 . . . . · · · . w · · · win Di = i1 wi2 = W + ei eTi W − ei eTi W. . . . . . . wn1 wn2 · · · wnn Notice that −ei eTi W subtracts Wi∗ from the ith row while +ei eTi W adds Wi∗ to the ith row. Use the fact that W = AW to write Di = W+ei eTi W −ei eTi W = W+ei eTi AW−ei eTi W = I+ei eTi A − eTi W, and apply formula (6.2.2) for the determinant of a rank-one updated matrix together with the product rule (6.1.15) to produce det (Di) = 1 + eTi Aei − eTi ei det (W) = aii (t) w(t), so d w(t) det (Di) = = dt i=1 n n aii (t) w(t) = trace A(t) w(t). i=1 In other words, w(t) satisfies the first-order differential equation w= τ w, where t τ (ξ) dξ τ = trace A(t), and the solution of this equation is w(t) = w(ξ0) e ξ0 . Consequences: In addition to its aesthetic elegance, (6.2.8) is a useful result because it is the basis for the following theorems. • If x = Ax has a set of solutions S = {w1 (t), w2 (t), . . . , wn (t)} that is t linearly independent at some point ξ0 ∈ (a, b), and if ξ0 τ (ξ) dξ is finite for t ∈ (a, b), then S must be linearly independent at every point t ∈ (a, b). • If A is a constant matrix, and if S is a set of n solutions that is linearly independent at some value t = ξ0 , then S must be linearly independent for all values of t. Proof. If S is linearly independent at ξ0 , then W(ξ0) is t nonsingular, so t τ (ξ) dξ w(ξ0) = 0. If ξ0 τ (ξ) dξ is finite when t ∈ (a, b), then e ξ0 is finite and nonzero on (a, b), so by (6.2.8), w(t) = 0 on (a, b). Therefore, W(t) is nonsingular for t ∈ (a, b), and thus S is linearly independent at each t ∈ (a, b). Exercises for section 6.2 6.2.1. Use a cofactor expansion to evaluate each of the following determinants. 0 1 1 1 0 0 −2 3 2 1 1 1 2 1 0 1 1 1 0 (c) (b) (a) 6 2 1 , . , 2 1 1 0 1 −1 1 −2 2 1 1 1 0 0 2 −3 0 6.2 Additional Properties of Determinants 483 6.2.2. Use determinants to compute the following inverses. −1 −1 0 0 −2 3 2 1 1 1 2 1 0 (a) 6 2 1 . (b) . −1 1 2 1 −2 2 1 0 2 −3 0 6.2.3. (a) Use Cramer's rule to solve x1 + x2 + x3 = 1, x1 + x2 = α, x2 + x3 = β. (b) Evaluate limt→∞ x2 (t), where x2 (t) is defined by the system x1 + tx2 + t2 x3 = t4 , t2 x1 + x2 + tx3 = t3 , tx1 + t2 x2 + x3 = 0. 6.2.4. Is the following equation a valid derivation of Cramer's rule for solving a nonsingular system Ax = b, where Ai is as described on p. 476? det (Ai) = det A−1 Ai = det e1 · · · ei−1 x ei+1 · · · en = xi . det (A) 6.2.5. (a) By example, show that det (A + B) = det (A) + det (B). (b) Using square matrices, construct an example that shows that A B det = det (A)det (D) − det (B)det (C). C D 6.2.6. Suppose onto rank (Bm×n) = n, and let Q be the orthogonal projector N BT . For A = [B | cn×1] , prove cT Qc = det AT A /det BT B . 6.2.7. If An×n is a nonsingular matrix, and if D and C are n × k matrices, explain how to use (6.2.1) to derive the formula det A + CDT = det (A)det Ik + DT A−1 C . Note: This is a generalization of (6.2.3) because if ci and di are the ith columns of C and D, respectively, then A + CDT = A + c1 dT1 + c2 dT2 + · · · + ck dTk . 484 Chapter 6 Determinants 6.2.8. Explain why A is singular if and only if A[adj (A)] = 0. 6.2.9. For a nonsingular linear system Ax = b, explain why each component of the solution must vary continuously with the entries of A. 6.2.10. For scalars α, explain why adj (αA) = αn−1 adj (A) . Hint: Recall Exercise 6.1.11. 6.2.11. For an (a) (b) (c) n × n matrix A, prove that the following statements are true. If rank (A) < n − 1, then adj (A) = 0. If rank (A) = n − 1, then rank (adj (A)) = 1. If rank (A) = n, then rank (adj (A)) = n. 6.2.12. In 1812, Cauchy discovered the formula that says that if A is n × n, n−1 then det (adj (A)) = [det (A)] . Establish Cauchy's formula. 6.2.13. For the following tridiagonal matrix, An , let Dn = det (An), and derive the formula Dn = 2Dn−1 − Dn−2 to deduce that Dn = n + 1. 2 −1 0 · · · 0 −1 2 −1 · · · 0 . . . . An = 0 · · · −1 2 −1 0 · · · 0 −1 2 n×n 6.2.14. By considering rank-one updated matrices, derive the following formulas. 1+α1 α1 1 · · · 1 1+α2 1 · · · 1 1 + αi α2 (a) . . . . = α . . . . i . . . 1 1+αn 1 · · · (b) (c) α β β . . . β αn β α β . . β β α . . · · · · · · . . . β β · · · 1 + α1 α1 . . . α1 α2 1 + α2 . . . α2 β β n (α − β) 1+ β = . . 0 . α n×n · · · . . . αn αn . . · · · 1 + αn if α = β, if α = β. = 1 + α1 + α2 + · · · + αn . 6.2 Additional Properties of Determinants 485 6.2.15. A bordered matrix has the form B = A yT x α in which An×n is T nonsingular, x is a column, y is a row, and α is a scalar. Explain why the following statements must be true. (a) A T y x = −det A + xyT . (b) −1 A T y x = −yT adj (A) x. 0 6.2.16. If B is m × n and C is n × m, explain why (6.2.1) guarantees that λm det (λIn − CB) = λn det (λIm − BC) is true for all scalars λ. 6.2.17. For a square matrix A and column vectors c and d, derive the following two extensions of formula (6.2.3). (a) If Ax = c, then det A + cdT =det (A) 1 + dT x . (b) If yT A = dT , then det A + cdT = det (A) 1 + yT c . 6.2.18. Describe the determinant of an elementary reflector (p. 324) and a plane rotation (p. 333), and then explain how to find det (A) using Householder reduction (p. 341) and Givens reduction (Example 5.7.2). 6.2.19. Suppose that A is a nonsingular matrix whose entries are integers. Prove that the entries in A−1 are integers if and only if det (A) = ±1. 6.2.20. Let A = I − 2uvT be a matrix in which u and v are column vectors with integer entries. (a) Prove that A−1 has integer entries if and only if vT u = 0 or 1. (b) A matrix is said to be involutory whenever A−1 = A. Explain why A = I − 2uvT is involutory when vT u = 1. 6.2.21. Use induction to argue that a cofactor expansion of det (An×n) requires 1 1 1 c(n) = n! 1 + + + · · · + 2! 3! (n − 1)! multiplications for n ≥ 2. Assume a computer will do 1,000,000 multiplications per second, and neglect all other operations to estimate how long it will take to evaluate the determinant of a 100 × 100 matrix using cofactor expansions. Hint: Recall the series expansion for ex , and use 100! ≈ 9.33 × 10157 . 486 Chapter 6 Determinants 6.2.22. Determine all values of λ for which the matrix A−λI is singular, where 0 −3 −2 A= 2 5 2 . −2 −3 0 Hint: If p(λ) = λn + αn−1 λn−1 + · · · + α1 λ + α0 is a monic polynomial with integer coefficients, then the integer roots of p(λ) are a subset of the factors of α0 . 6.2.23. Suppose that f1 (t), f2 (t), . . . , fn (t) are solutions of nth-order linear differential equation y (n) + p1 (t)y (n−1) + · · · + pn−1 (t)y + pn (t)y = 0, and let w(t) be the Wronskian f2 (t) · · · fn (t) f1 (t) f2 (t) · · · fn (t) f1 (t) . w(t) = ... (n−1) (n−1) (n−1) f1 (t) f2 (t) · · · fn (t) By converting the nth-order equation into a system of n first-order equations with the substitutions x1 = y, x2 = y , . . . , xn = y (n−1) , show that w(t) = w(ξ0) e − t ξ0 p1 (ξ) dξ for an arbitrary constant ξ0 . 6.2.24. Evaluate the Vandermonde determinant by showing n−1 1 x1 x21 · · · x1 n−1 2 1 x2 x2 · · · x2 . (xj − xi). . . . . = . . . · · · . j>i . 1 xn x2n · · · xn−1 n When is this nonzero (compare 1 λ λ2 1 x2 x22 polynomial p(λ) = ... 1 x x.2 k k with Example 4.3.4)? Hint: For the λk−1 xk−1 2 . . , use induction to find the · · · . · · · xk−1 · · · · · · k k×k degree of p(λ), the roots of p(λ), and the coefficient of λk−1 in p(λ). 6.2.25. Suppose that each entry in An×n = [aij (x)] is a differentiable function of a real variable x. Use formula (6.1.19) to derive the formula n n d det (A) d aij ˚ = Aij . dx dx j=1 i=1 6.2 Additional Properties of Determinants 487 6.2.26. Consider the entries of A to be independent variables, and use formula (6.1.19) to derive the formula ∂ det (A) ˚ = Aij . ∂aij 6.2.27. Laplace's Expansion. In 1772, the French mathematician Pierre-Simon Laplace (1749–1827) presented the following generalized version of the cofactor expansion. For an n × n matrix A, let A(i1 i2 · · · ik | j1 j2 · · · jk) = the k × k submatrix of A that lies on the intersection of rows i1 , i2 , . . . , ik with columns j1 , j2 , . . . , jk , and let M (i1 i2 · · · ik | j1 j2 · · · jk) = the n − k × n − k minor determinant obtained by deleting rows i1 , i2 , . . . , ik and columns j1 , j2 , . . . , jk from A. The cofactor of A(i1 · · · ik | j1 · · · jk) is defined to be the signed minor ˚ A(i1 · · · ik | j1 · · · jk) = (−1)i1 +···+ik +j1 +···+jk M (i1 · · · ik | j1 · · · jk) . This is consistent with the definition of cofactor given earlier because if A(i | j) = aij , then ˚ A(i | j) = (−1)i+j M (i | j) = (−1)i+j Mij = ˚ Aij . For each fixed set of row indices 1 ≤ i1 < · · · < ik ≤ n, det (A) = det A(i1 · · · ik | j1 · · · jk) ˚ A(i1 · · · ik | j1 · · · jk) . 1≤j1 0. In other words, the mathematical model would be grossly inconsistent with reality if the symmetric matrix A in (7.6.2) were not positive definite. It turns out that A is positive definite because there is a Cholesky factorization A = RT R with r −1/r * R= 1 T mL 1 r2 −1/r2 .. rn−1 .. −1/rn−1 rn * with rk = 2− k−1 , k and thus we are insured that each λk > 0. In fact, since A is a tridiagonal Toeplitz matrix, the results of Example 7.2.5 (p. 514) can be used to show that kπ 2T kπ 4T λk = 1 − cos = sin2 (see Exercise 7.2.18). mL n+1 mL 2(n + 1) 562 Chapter 7 Eigenvalues and Eigenvectors Therefore, √ √ = αk cos t λk + βk sin t λk zk √ zk (0) = c˜k =⇒ zk = c˜k cos t λk , (7.6.3) zk (0) = 0 and for P = x1 | x2 | · · · | xn , y = Pz = z1 x1 + z2 x2 + · · · + zn xn = n √ c˜j cos t λk xj . (7.6.4) j=1 This means that every possible mode of vibration is a combination of modes determined by the eigenvectors xj . To understand this more clearly, suppose that the beads are initially positioned according to the components of xj —i.e., ˜ = PT c = PT xj = ej , so (7.6.3) and (7.6.4) reduce to c = y(0) = xj . Then c √ if k = j =⇒ y = cos t√λ x . zk = cos t λk (7.6.5) k j 0 if k = j In other words, when y(0) = xj , the j th eigenpair (λj , xj) completely determines the mode of vibration because the amplitudes are determined by xj , and each bead vibrates with a common frequency f = λj /2π. This type of motion (7.6.5) is called a normal mode of vibration. In these terms, equation (7.6.4) translates to say that every possible mode of vibration is a combination of the normal modes. For example, when n = 3, the matrix in (7.6.2) is 2 −1 0 λ1 = (T /mL)(2) √ T A= λ2 = (T /mL)(2 − √2) , −1 2 −1 with mL 0 −1 2 λ3 = (T /mL)(2 + 2) and a complete orthonormal set of eigenvectors is 1 √1 √1 1 1 x1 = √ 0 , x2 = 2 , x3 = − 2 . 2 2 2 −1 1 1 The three corresponding normal modes are shown in Figure 7.6.3. Mode for (λ1 , x1) Mode for (λ2 , x2) Figure 7.6.3 Mode for (λ3 , x3) 7.6 Positive Definite Matrices 563 Example 7.6.2 Discrete Laplacian. According to the laws of physics, the temperature at time t at a point (x, y, z) in a solid body is a function u(x, y, z, t) satisfying the diffusion equation ∂u = K∇2 u, ∂t where ∇2 u = ∂2u ∂2u ∂2u + 2 + 2 ∂x2 ∂y ∂z is the Laplacian of u and K is a constant of thermal diffusivity. At steady state the temperature at each point does not vary with time, so ∂u/∂t = 0 and u = u(x, y, z) satisfy Laplace's equation ∇2 u = 0. Solutions of this equation are often called harmonic functions. The nonhomogeneous equation ∇2 u = f (Poisson's equation) is addressed in Exercise 7.6.9. To keep things simple, let's confine our attention to the following two-dimensional problem. Problem: For a square plate as shown in Figure 7.6.4(a), explain how to numerically determine the steady-state temperature at interior grid points when the temperature around the boundary is prescribed to be u(x, y) = g(x, y) for a given function g. In other words, explain how to extract a numerical solution to ∇2 u = 0 in the interior of the square when u(x, y) = g(x, y) on the square's 76 boundary. This is called a Dirichlet problem. Solution: Discretize the problem by overlaying the plate with a square mesh containing n2 interior points at equally spaced intervals of length h. As illustrated in Figure 7.6.4(b) for n = 4, label the grid points using a rowwise ordering scheme—i.e., label them as you would label matrix entries. ∇ u = 0 in the interior 2 + h 00 01 02 03 04 05 10 11 12 13 14 15 20 21 22 23 24 25 30 31 32 33 34 35 40 41 42 43 44 45 51 52 53 54 55 ,- u(x, y) = g(x, y) on the boundary u(x, y) = g(x, y) on the boundary u(x, y) = g(x, y) on the boundary . u(x, y) = g(x, y) on the boundary 50 + ,- . h (a) (b) Figure 7.6.4 76 Johann Peter Gustav Lejeune Dirichlet (1805–1859) held the chair at G¨ ottingen previously occupied by Gauss. Because of his work on the convergence of trigonometric series, Dirichlet is generally considered to be the founder of the theory of Fourier series, but much of the groundwork was laid by S. D. Poisson (p. 572) who was Dirichlet's Ph.D. advisor. 564 Chapter 7 Eigenvalues and Eigenvectors Approximate ∂ 2 u/∂x2 and ∂ 2 u/∂y 2 at the interior grid points (xi , yj) by using the second-order centered difference formula (1.4.3) developed on p. 19 to write ∂ 2 u u(xi − h, yj) − 2u(xi , yj) + u(xi + h, yj) = + O(h2) , 2 ∂x (xi ,yj) h2 ∂ 2 u u(xi , yj − h) − 2u(xi , yj) + u(xi , yj + h) = + O(h2) . ∂y 2 (xi ,yj) h2 (7.6.6) Adopt the notation uij = u(xi , yj), and add the expressions in (7.6.6) using ∇2 u|(xi ,yj) = 0 for interior points (xi , yj) to produce 4uij = (ui−1,j + ui+1,j + ui,j−1 + ui,j+1) + O(h4) for i, j = 1, 2, . . . , n. In other words, the steady-state temperature at an interior grid point is approximately the average of the steady-state temperatures at the four neighboring grid points as illustrated in Figure 7.6.5. i − 1, j i, j − 1 ij uij = i, j + 1 ui−1,j + ui+1,j + ui,j−1 + ui,j+1 + O(h4) 4 i + 1, j Figure 7.6.5 4 If the O(h) terms are neglected, the resulting five-point difference equations, 4uij − (ui−1,j + ui+1,j + ui,j−1 + ui,j+1) = 0 for i, j = 1, 2, . . . , n, constitute an n2 × n2 linear system Lu = g in which the unknowns are the uij 's, and the right-hand side contains boundary values. For example, a mesh with nine interior points produces the 9 × 9 system in Figure 7.6.6. 00 01 02 03 04 10 11 12 13 14 20 21 22 23 24 30 31 32 33 34 40 41 42 43 44 4 −1 0 4 −1 −1 4 0 −1 −1 0 0 0 −1 0 −1 0 0 0 0 0 0 0 0 −1 0 0 0 −1 0 0 0 −1 0 0 0 0 0 0 4 −1 0 −1 4 −1 0 −1 4 −1 0 −1 4 −1 0 0 −1 0 0 −1 0 0 0 −1 0 0 0 −1 4 −1 −1 4 0 −1 Figure 7.6.6 0 g01 + g10 u11 0 u12 g02 0 u13 g03 + g14 0 u21 g20 0 0 u22 = 0 −1 u23 g24 u g + g 0 31 30 41 u32 g42 −1 u33 g43 + g34 4 7.6 Positive Definite Matrices 565 The coefficient matrix of this system is the discrete Laplacian, and in general it has the symmetric block-tridiagonal form T −I 4 −1 −I T −I −1 4 −1 . . . . . . L= . . . . . . with T = −I T −I −1 4 −1 −I T n2 ×n2 −1 4 n×n In addition, L is positive definite. In fact, the discrete Laplacian is a primary example of how positive definite matrices arise in practice. Note that L is the two-dimensional version of the one-dimensional finite-difference matrix in Example 1.4.1 (p. 19). Problem: Show L is positive definite by explicitly exhibiting its eigenvalues. Solution: Example 7.2.5 (p. 514) insures that the n eigenvalues of T are iπ λi = 4 − 2 cos , i = 1, 2, . . . , n. (7.6.7) n+1 If U is an orthogonal matrix such that UT TU = D = diag and if B is the n2 × n2 block-diagonal orthogonal matrix D −I U 0 · · · 0 −I D 0 U · · · 0 . . ˜ = B= , then BT LB = L . . . . 0 0 · · · U (λ1 , λ2 , . . . , λn) , −I . . −I . . D −I −I D Consider the permutation obtained by placing the numbers 1, 2, . . . , n2 rowwise in a square matrix, and then reordering them by listing the entries columnwise. For example, when n = 3 this permutation is generated as follows: 1 2 3 ˜. v = (1, 2, 3, 4, 5, 6, 7, 8, 9) → A = 4 5 6 → (1, 4, 7, 2, 5, 8, 3, 6, 9) = v 7 8 9 Equivalently, this can be described in terms of wrapping and unwrapping rows by wrap unwrap writing v− −−→A −→ AT −−−−→˜ v. If P is the associated n2 × n2 permutation matrix, then λi −1 T1 0 · · · 0 −1 λi −1 0 T2 · · · 0 . . . T˜ . . . P LP = ... with Ti = ... −1 λi −1 0 0 · · · Tn −1 λi n×n 566 Chapter 7 Eigenvalues and Eigenvectors If you try it on the 9 × 9 case, you will see why it works. Now, Ti is another tridiagonal Toeplitz matrix, so Example 7.2.5 (p. 514) again applies to yield σ (Ti) = {λi − 2 cos (jπ/n + 1) , j = 1, 2, . . . , n} . This together with (7.6.7) produces the n2 eigenvalues of L as % & iπ jπ λij = 4 − 2 cos + cos , n+1 n+1 i, j = 1, 2, . . . , n, or, by using the identity 1 − cos θ = 2 sin2 (θ/2), % λij = 4 sin2 iπ 2(n + 1) + sin2 jπ 2(n + 1) & , i, j = 1, 2, . . . , n. (7.6.8) Since each λij is positive, L must be positive definite. As a corollary, L is nonsingular, and hence Lu = g yields a unique solution for the steady-state temperatures on the square plate (otherwise something would be amiss). At first glance it's tempting to think that statements about positive definite matrices translate to positive semidefinite matrices simply by replacing the word "positive" by "nonnegative," but this is not always true. When A has zero eigenvalues (i.e., when A is singular) there is no LU factorization, and, unlike the positive definite case, having nonnegative leading principal minors doesn't insure 0 0 that A is positive semidefinite—e.g., consider A = 0 −1 . The positive definite properties that have semidefinite analogues are listed below. Positive Semidefinite Matrices For real-symmetric matrices such that rank (An×n) = r, the following statements are equivalent, so any one of them can serve as the definition of a positive semidefinite matrix. • xT Ax ≥ 0 for all x ∈ n×1 (the most common definition). (7.6.9) • • All eigenvalues of A are nonnegative. A = BT B for some B with rank (B) = r. • All principal minors of A are nonnegative. (7.6.10) (7.6.11) (7.6.12) For hermitian matrices, replace (%)T by (%)∗ and by C. Proof of (7.6.9) =⇒ (7.6.10). The hypothesis insures xT Ax ≥ 0 for eigenvectors of A. If (λ, x) is an eigenpair, then λ = xT Ax/xT x = Bx2 / x2 ≥ 0. Proof of (7.6.10) =⇒ (7.6.11). Similar to the positive definite case, if each λi ≥ 0, write A = PD1/2 D1/2 PT = BT B, where B = D1/2 PT has rank r. 7.6 Positive Definite Matrices 567 Proof of (7.6.11) =⇒ (7.6.12). If Pk is a principal submatrix of A, then Pk % % % = QT AQ = QT BT BQ = FT % F % == Pk = FT F for a permutation matrix Q. Thus det (Pk) = det FT F ≥ 0 (Exercise 6.1.10). Proof of (7.6.12) =⇒ (7.6.9). If Ak is the leading k × k principal submatrix of A, and if {µ1 , µ2 , . . . , µk } are the eigenvalues (including repetitions) of Ak , then I + Ak has eigenvalues { + µ1 , + µ2 , . . . , + µk }, so, for every > 0, det (I + Ak) = (+ µ1)(+ µ2) · · · (+ µk) = k + s1 k−1 + · · · + sk−1 + sk > 0 because sj is the j th symmetric function of the µi 's (p. 494), and, by (7.1.6), sj is the sum of the j × j principal minors of Ak , which are principal minors of A. In other words, each leading principal minor of I + A is positive, so I+A is positive definite by the results on p. 559. Consequently, for each nonzero x ∈ n×1 , we must have xT (I + A)x > 0 for every > 0. Let → 0+ (i.e., through positive values) to conclude that xT Ax ≥ 0 for each x ∈ n×1 . Quadratic Forms For a vector x ∈ n×1 and a matrix A ∈ n×n , the scalar function defined by n n T f (x) = x Ax = aij xi xj (7.6.13) i=1 j=1 is called a quadratic form. A quadratic form is said to be positive definite whenever A is a positive definite matrix. In other words, (7.6.13) is a positive definite form if and only if f (x) > 0 for all 0 = x ∈ n×1 . Because xT Ax = xT (A + AT)/2 x, and because (A + AT)/2 is symmetric, the matrix of a quadratic form can always be forced to be symmetric. For this reason it is assumed that the matrix of every quadratic form is symmetric. When x ∈ C n×1 , A ∈ C n×n , and A is hermitian, the expression xH Ax is known as a complex quadratic form. Example 7.6.3 Diagonalization of a Quadratic Form. A quadratic form f (x) = xT Dx is said to be a diagonal form whenever Dn×n is a diagonal matrix, in which n case xT Dx = i=1 dii x2i (there are no cross-product terms). Every quadratic T form x Ax can be diagonalized by making a change of variables (coordinates) 568 Chapter 7 Eigenvalues and Eigenvectors y = QT x. This follows because A is symmetric, so there is an orthogonal matrix Q such that QT AQ = D = diag (λ1 , λ2 , . . . , λn) , where λi ∈ σ (A) , and setting y = QT x (or, equivalently, x = Qy) gives T T T T x Ax = y Q AQy = y Dy = n λi yi2 . (7.6.14) i=1 This shows that the nature of the quadratic form is determined by the eigenvalues of A (which are necessarily real). The effect of diagonalizing a quadratic form in this way is to rotate the standard coordinate system so that in the new coordinate system the graph of xT Ax = α is in "standard form." If A is positive definite, then all of its eigenvalues are positive (p. 559), so (7.6.14) makes it clear that the graph of xT Ax = α for a constant α > 0 is an ellipsoid centered at the origin. Go back and look at Figure 7.2.1 (p. 505), and see Exercise 7.6.4 (p. 571). Example 7.6.4 Congruence. It's not necessary to solve an eigenvalue problem to diagonalize a quadratic form because a congruence transformation CT AC in which C is nonsingular (but not necessarily orthogonal) can be found that will do the job. A particularly convenient congruence transformation is produced by the LDU factorization for A, which is A = LDLT because A is symmetric—see Exercise 3.10.9 (p. 157). This factorization is relatively cheap, and the diagonal entries in D = diag (p1 , p2 , . . . , pn) are the pivots that emerge during Gaussian elimination (p. 154). Setting y = LT x (or, equivalently, x = (LT)−1 y) yields xT Ax = yT Dy = n pi yi2 . i=1 The inertia of a real-symmetric matrix A is defined to be the triple (ρ, ν, ζ) in which ρ, ν, and ζ are the respective number of positive, negative, and zero eigenvalues, counting algebraic multiplicities. In 1852 J. J. Sylvester (p. 80) discovered that the inertia of A is invariant under congruence transformations. Sylvester's Law of Inertia Let A ∼ = B denote the fact that real-symmetric matrices A and B are congruent (i.e., CT AC = B for some nonsingular C). Sylvester's law of inertia states that: A∼ =B if and only if A and B have the same inertia. 7.6 Positive Definite Matrices Proof. 77 569 Observe that if An×n is real and symmetric with inertia (p, j, s), then A∼ = Ip×p = E, −Ij×j (7.6.15) 0s×s because if {λ1 , . . . , λp , −λp+1 , . . . , −λp+j , 0, . . . , 0} are the eigenvalues of A (counting multiplicities) with each λi > 0, there is an orthogonal matrix P such that PT AP = diag (λ1 , . . . , λp , −λp+1 , . . . , −λp+j , 0, . . . , 0) , so C = PD, −1/2 −1/2 where D = diag λ1 , . . . , λp+j , 1, . . . , 1 , is nonsingular and CT AC = E. Let B be a real-symmetric matrix with inertia (q, k, t) so that B∼ = Iq×q = F. −Ik×k 0t×t If B ∼ = A, then F ∼ = E (congruence is transitive), so rank (F) = rank (E), and hence s = t. To show that p = q, assume to the contrary that p > q, and write F = KT EK for some nonsingular K = Xn×q | Yn×n−q . If M = R (Y) ⊆ n and N = span {e1 , . . . , ep } ⊆ n , then using the formula (4.4.19) for the dimension of a sum (p. 205) yields dim(M ∩ N) = dim M + dim N − dim(M + N) = (n − q) + p − dim(M + N) > 0. Consequently, there exists a nonzero vector x ∈ M ∩ N . For such a vector, x ∈ M =⇒ x = Yy = K 0 0 =⇒ xT Ex = 0T | yT F ≤ 0, y y and x∈N =⇒ x = (x1 , . . . , xp , 0, . . . , 0)T =⇒ xT Ex > 0, which is impossible. Therefore, we can't have p > q. A similar argument shows that it's also impossible to have p < q, so p = q. Thus it is proved that if A∼ = B, then A and B have the same inertia. Conversely, if A and B have inertia (p, j, s), then the argument that produced (7.6.15) yields A ∼ =E∼ = B. 77 The fact that inertia is invariant under congruence is also a corollary of a deeper theorem stating that the eigenvalues of A vary continuously with the entries. The argument is as follows. Assume A is nonsingular (otherwise consider A + I for small), and set X(t) = tQ + (1 − t)QR for t ∈ [0, 1], where C = QR is the QR factorization. Both X(t) and Y(t) = XT (t)AX(t) are nonsingular on [0, 1], so continuity of eigenvalues insures that no eigenvalue Y(t) can cross the origin as t goes from 0 to 1. Hence Y(0) = CT AC has the same number of positive (and negative) eigenvalues as Y(1) = QT AQ, which is similar to A. Thus CT AC and A have the same inertia. 570 Chapter 7 Eigenvalues and Eigenvectors Example 7.6.5 Taylor's theorem in n says that if f is a smooth real-valued function defined on n , and if x0 ∈ n×1 , then the value of f at x ∈ n×1 is given by 3 f (x) = f (x0) + (x − x0)T g(x0) + (x − x0)T H(x0)(x − x0) + O(x − x0), where g(x0) = ∇f (x0) (the gradient of f evaluated at x0) has components gi = ∂f /∂xi , and where H(x0) is the Hessian matrix whose entries are x0 given by hij = ∂ 2 f /∂xi ∂xj . Just as in the case of one variable, the vector x0 x0 is called a critical point when g(x0) = 0. If x0 is a critical point, then Taylor's theorem shows that (x − x0)T H(x0)(x − x0) governs the behavior of f at points x near to x0 . This observation yields the following conclusions regarding local maxima or minima. • If x0 is a critical point such that H(x0) is positive definite, then f has a local minimum at x0 . • If x0 is a critical point such that H(x0) is negative definite (i.e., zT Hz < 0 for all z = 0 or, equivalently, −H is positive definite), then f has a local maximum at x0 . Exercises for section 7.6 7.6.1. Which of the following matrices are positive definite? 1 −1 −1 20 6 8 2 A = −1 5 1 . B = 6 3 0 . C = 0 −1 1 5 8 0 8 2 0 6 2 2 2 . 4 7.6.2. Spring-Mass Vibrations. Two masses m1 and m2 are suspended between three identical springs (with spring constant k) as shown in Figure 7.6.7. Each mass is initially displaced from its equilibrium position by a horizontal distance and released to vibrate freely (assume there is no vertical displacement). m1 m2 x1 x2 m1 Figure 7.6.7 m2 7.6 Positive Definite Matrices 571 (a) If xi (t) denotes the horizontal displacement of mi from equilibrium at time t, show that Mx = Kx, where x1 (t) m1 0 2 −1 , x= M= , and K = k . −1 2 0 m2 x2 (t) (Consider a force directed to the left to be positive.) Notice that the mass-stiffness equation Mx = Kx is the matrix version of Hooke's law F = kx, and K is positive definite. (b) Look for a solution of the form x = eiθt v for a constant vector v, and show that this reduces the problem to solving an algebraic equation of the form Kv = λMv (for λ = −θ2). This is called a generalized eigenvalue problem because when M = I we are back to the ordinary eigenvalue problem. The generalized eigenvalues λ1 and λ2 are the roots of the equation det (K − λM) = 0—find them when k = 1, m1 = 1, and m2 = 2, and describe the two modes of vibration. (c) Take m1 = m2 = m, and apply the technique used in the vibrating beads problem in Example 7.6.1 (p. 559) to determine the normal modes. Compare the results with those of part (b). 7.6.3. Three masses m1 , m2 , and m3 are suspended on three identical springs (with spring constant k) as shown below. Each mass is initially displaced from its equilibrium position by a vertical distance and then released to vibrate freely. (a) If yi (t) denotes the displacement of mi from equilibrium at time t, show that the mass-stiffness equation is My = Ky, where m1 0 y1 (t) 0 2 −1 0 M= 0 m2 0 , y= y2 (t) , K=k −1 2 −1 0 0 m3 y3 (t) 0 −1 1 (k33 = 1 is not a mistake!). (b) Show that K is positive definite. (c) Find the normal modes when m1 = m2 = m3 = m. 7.6.4. By diagonalizing the quadratic form 13x2 + 10xy + 13y 2 , show that the rotated graph of 13x2 + 10xy + 13y 2 = 72 is an ellipse in standard form as shown in Figure 7.2.1 on p. 505. 7.6.5. Suppose that A is a real-symmetric matrix. Explain why the signs of the pivots in the LDU factorization for A reveal the inertia of A. 572 Chapter 7 Eigenvalues and Eigenvectors 7.6.6. Consider the quadratic form f (x) = 1 (−2x21 + 7x22 + 4x23 + 4x1 x2 + 16x1 x3 + 20x2 x3) . 9 (a) Find a symmetric matrix A so that f (x) = xT Ax. (b) Diagonalize the quadratic form using the LDLT factorization as described in Example 7.6.4, and determine the inertia of A. (c) Is this a positive definite form? (d) Verify the inertia obtained above is correct by computing the eigenvalues of A. (e) Verify Sylvester's law of inertia by making up a congruence transformation C and then computing the inertia of CT AC. 7.6.7. Polar Factorization. Explain why each nonsingular A ∈ C n×n can be uniquely factored as A = RU, where R is hermitian positive definite and U is unitary. This is the matrix analog of the polar form of a complex number z = reiθ , r > 0, because 1 × 1 hermitian positive definite matrices are positive real numbers, and 1 × 1 unitary matrices are points on the unit circle. Hint: First explain why R = (AA∗)1/2 . 7.6.8. Explain why trying to produce better approximations to the solution of the Dirichlet problem in Example 7.6.2 by using finer meshes with more grid points results in an increasingly ill-conditioned linear system Lu = g. 78 7.6.9. For a given function f the equation ∇2 u = f is called Poisson's equation. Consider Poisson's equation on a square in two dimensions with Dirichlet boundary conditions. That is, ∂2u ∂2u + 2 = f (x, y) ∂x2 ∂y 78 with u(x, y) = g(x, y) on the boundary. Sim´ eon Denis Poisson (1781–1840) was a prolific French scientist who was originally encouraged to study medicine but was seduced by mathematics. While he was still a teenager, his work attracted the attention of the reigning scientific elite of France such as Legendre, Laplace, and Lagrange. The latter two were originally his teachers (Lagrange was his thesis director) at the ´ Ecole Polytechnique, but they eventually became his friends and collaborators. It is estimated that Poisson published about 400 scientific articles, and his 1811 book Trait´ e de m´ ecanique was the standard reference for mechanics for many years. Poisson began his career as an astronomer, but he is primarily remembered for his impact on applied areas such as mechanics, probability, electricity and magnetism, and Fourier series. This seems ironic because he held the chair of "pure mathematics" in the Facult´ e des Sciences. The next time you find yourself on the streets of Paris, take a stroll on the Rue Denis Poisson, or you can check out Poisson's plaque, along with those of Lagrange, Laplace, and Legendre, on the first stage of the Eiffel Tower. 7.6 Positive Definite Matrices 573 Discretize the problem by overlaying the square with a regular mesh containing n2 interior points at equally spaced intervals of length h as explained in Example 7.6.2 (p. 563). Let fij = f (xi , yj), and define f to be the vector f = (f11 , f12 , . . . , f1n |f21 , f22 . . . , f2n | · · · |fn1 , fn2 , . . . , fnn)T . Show that the discretization of Poisson's equation produces a system of linear equations of the form Lu = g − h2 f , where L is the discrete Laplacian and where u and g are as described in Example 7.6.2. 7.6.10. As defined in Exercise 5.8.15 (p. 380) and discussed in Exercise 7.8.11 (p. 597) the Kronecker product (sometimes called tensor product, or direct product) of matrices Am×n and Bp×q is the mp × nq matrix a1n B a2n B . . . . a11 B a21 B A⊗B= . . . a12 B a22 B . . · · · · · · . . am1 B am2 B · · · amn B Verify that if In is the n × n identity matrix, and if 2 −1 An = −1 2 . . . −1 . . −1 . . 2 −1 −1 2 n×n is the nth -order finite difference matrix of Example 1.4.1 (p. 19), then the discrete Laplacian is given by Ln2 ×n2 = (In ⊗ An) + (An ⊗ In). Thus we have an elegant matrix connection between the finite difference approximations of the one-dimensional and two-dimensional Laplacians. This formula leads to a simple alternate derivation of (7.6.8)—see Exercise 7.8.12 (p. 598). As you might guess, the discrete three-dimensional Laplacian is Ln3 ×n3 = (In ⊗ In ⊗ An) + (In ⊗ An ⊗ In) + (An ⊗ In ⊗ In). 574 7.7 Chapter 7 Eigenvalues and Eigenvectors NILPOTENT MATRICES AND JORDAN STRUCTURE While it's not always possible to diagonalize a matrix A ∈ C m×m with a similarity transformation, Schur's theorem (p. 508) guarantees that every A ∈ C m×m is unitarily similar to an upper-triangular matrix—say U∗ AU = T. But other than the fact that the diagonal entries of T are the eigenvalues of A, there is no pattern to the nonzero part of T. So to what extent can this be remedied by giving up the unitary nature of U? In other words, is there a nonunitary P for which P−1 AP has a simpler and more predictable pattern than that of T? We have already made the first step in answering this question. The core-nilpotent decomposition (p. 397) says that for every singular matrix A of index k and rank r, there is a nonsingular matrix Q such that −1 Q AQ = Cr×r 0 0 L , where rank (C) = r and L is nilpotent of index k. Consequently, any further simplification by means of similarity transformations can revolve around C and L. Let's begin by examining the degree to which nilpotent matrices can be reduced by similarity transformations. In what follows, let Ln×n be a nilpotent matrix of index k so that Lk = 0 but Lk−1 = 0. The first question is, "Can L be diagonalized by a similarity transformation?" To answer this, notice that λ = 0 is the only eigenvalue of L because Lx = λx =⇒ Lk x = λk x =⇒ 0 = λk x =⇒ λ = 0 (since x = 0). So if L is to be diagonalized by a similarity transformation, it must be the case that P−1 LP = D = 0 (diagonal entries of D must be eigenvalues of L), and this forces L = 0. In other words, the only nilpotent matrix that is similar to a diagonal matrix is the zero matrix. Assume L = 0 from now on so that L is not diagonalizable. Since L can always be triangularized (Schur's theorem again), our problem boils down to finding a nonsingular P such that P−1 LP is an upper-triangular matrix possessing a simple and predictable form. This turns out to be a fundamental problem, and the rest of this section is devoted to its solution. But before diving in, let's set the stage by thinking about some possibilities. If P−1 LP = T is upper triangular, then the diagonal entries of T must be the eigenvalues of L, so T must have the form T= 0 % . . . % . . . . . . % 0 7.7 Nilpotent Matrices and Jordan Structure 575 One way to simplify the form of T is to allow nonzero entries only on the superdiagonal (the diagonal immediately above the main diagonal) of T, so we might try to construct a nonsingular P such that T has the form 0 % . . . . T= . . % 0 To gain some insight on how this might be accomplished, let L be a 3 × 3 nilpotent matrix for which L3 = 0 and L2 = 0, and search for a P such that 0 1 0 0 1 0 P−1 LP = 0 0 1 ⇐⇒ L[P∗1 P∗2 P∗3] = [P∗1 P∗2 P∗3] 0 0 1 0 0 0 0 0 0 ⇐⇒ LP∗1 = 0, LP∗2 = P∗1 , LP∗3 = P∗2 . Since L3 = 0, we can set P∗1 = L2 x for any x3×1 such that L2 x = 0. This in turn allows us to set P∗2 = Lx and P∗3 = x. Because J = {L2 x, Lx, x} is a linearly independent set (Exercise 5.10.8), P = [L2 x | Lx | x] will do the job. J is called a Jordan chain, and it is characterized by the fact that its first vector is a somewhat special eigenvector for L while the other vectors are built (or "chained") on top of this eigenvector to form a special basis for C 3 . There are a few more wrinkles in the development of a general theory for n × n nilpotent matrices, but the features illustrated here illuminate the path. For a general nilpotent matrix Ln×n = 0 of index k, we know that λ = 0 is the only eigenvalue, so the set of eigenvectors of L is N (L) (excluding the zero vector of course). Realizing that L is not diagonalizable is equivalent to realizing that L does not possess a complete linearly independent set of eigenvectors or, equivalently, dim N (L) < n. As in the 3 × 3 example above, the strategy for building a similarity transformation P that reduces L to a simple triangular form is as follows. (1) Construct a somewhat special basis B for N (L). (2) Extend B to a basis for C n by building Jordan chains on top of the eigenvectors in B. To accomplish (1), consider the subspaces defined by Mi = R Li ∩ N (L) for i = 0, 1, . . . , k, and notice (Exercise 7.7.4) that these subspaces are nested as 0 = Mk ⊆ Mk−1 ⊆ Mk−2 ⊆ · · · ⊆ M1 ⊆ M0 = N (L). (7.7.1) 576 Chapter 7 Eigenvalues and Eigenvectors Use these nested spaces to construct a basis for N (L) = M0 by starting with any basis Sk−1 for Mk−1 and by sequentially extending Sk−1 with additional sets Sk−2 , Sk−3 , . . . , S0 such that Sk−1 ∪ Sk−2 is a basis for Mk−2 , Sk−1 ∪ Sk−2 ∪ Sk−3 is a basis for Mk−3 , etc. In general, Si is a set of vectors that extends Sk−1 ∪ Sk−2 ∪ · · · ∪ Si−1 to a basis for Mi . Figure 7.7.1 is a heuristic diagram depicting an example of k = 5 nested subspaces Mi along with some typical extension sets Si that combine to form a basis for N (L). Figure 7.7.1 Now extend the basis B = Sk−1 ∪ Sk−2 ∪ · · · ∪

S0 = {b1 , b2 , . . . , bt } for N (L) to a basis for C n by building Jordan chains on top of each b ∈ B. If i b ∈ Si , then there exists i a vector x such that L x = b because each b ∈ Si i belongs to Mi = R L ∩ N (L) ⊆ R L . A Jordan chain is built on top of each b ∈ Si by solving the system Li x = b for x and by setting Jb = {Li x, Li−1 x, . . . , Lx, x}. (7.7.2) Notice that chains built on top of Jordan chains is a basis for C n . [illegible]

J(λj) − λi I is nonsingular when 602 Chapter 7 Eigenvalues and Eigenvectors i = j, so (A − λi I) = P(J − λi I)P−1 J(λ) − λ I 1 i .. Li = P .. J(λs) − λi I −1 P (7.9.4) is a core-nilpotent decomposition as described on p. 397 (reordering the eigenvalues can put the nilpotent block Li on the bottom to realize the form in (5.10.5)). Consequently, the results 5.10.3 (p. 398) insure that Pi Qi = Gi is in Example the projector onto N (A − λi I)ki along R (A − λi I)ki , and this is true for all similarity transformations that reduce A to J. If A happens to be diagonalizable, then ki = 1 for each i, and the matrices Gi = Pi Qi are precisely the spectral projectors defined on p. 517. For this reason, there is no ambiguity in continuing to use the Gi notation, and we will continue to refer to the Gi 's as spectral projectors. In the diagonalizable case, Gi projects onto the eigenspace associated with λi , and in the nondiagonalizable case Gi projects onto the generalized eigenspace associated with λi . Now consider f J(λ1) s −1 .. f (A) = Pf (J)P−1 = P = Pi f J(λi) Qi . P . f J(λs) Since f J(λi) i=1 (7.9.5) .. f J(λi) , .. where the J# (λi)'s are the Jordan blocks associated with λi , (7.9.2) insures that if ki = index (λi), then f (λi) 2 f (ki −1) (λi) ki −1 f J(λi) = f (λi)I + f (λi)Li + Li + · · · + L 2! (ki − 1)! i where Li = J(λi) − λi I, and thus (7.9.5) becomes f (A) = s k i −1 f (j) (λi) Pi f J(λi) Qi = Pi Lji Qi . j! i=1 i=1 j=0 s (7.9.6) The terms Pi Lji Qi can be simplified by noticing that P−1 P = I =⇒ Qi Pj = Q 1 I 0 if i = j, =⇒ P−1 Gi = if i = j, .. Q . Pi Qi .. Qs 0 .. = Q. i , 0 7.9 Functions of Nondiagonalizable Matrices 603 and by using this with (7.9.4) to conclude that j (A − λi I) Gi = P J(λ1) − λi I .. j. Lji .. j −1 j P Gi = Pi Li Qi . (7.9.7) J(λs) − λi I Thus (7.9.6) can be written as f (A) = s k i −1 f (j) (λi) j! i=1 j=0 (A − λi I)j Gi , (7.9.8) and this expression is independent of which similarity is used to reduce A to J. Not only does (7.9.8) prove that f (A) is uniquely defined, but it also provides a generalization of the spectral theorems for diagonalizable matrices given on pp. 517 and 526 because if A is diagonalizable, then each ki = 1 so that (7.9.8) reduces to (7.3.6) on p. 526. Below is a formal summary along with some related properties. Spectral Resolution of f (A) For A ∈ C n×n with σ (A) = {λ1 , λ2 , . . . , λs } such that ki = index (λi), and for a function f : C → C such that f (λi), f (λi), . . . , f (ki −1) (λi) exist for each λi ∈ σ (A), the value of f (A) is f (A) = s k i −1 f (j) (λi) i=1 j=0 j! (A − λi I)j Gi , (7.9.9) where the spectral projectors Gi 's have the following properties. • Gi is the projectorkonto the generalized eigenspace N (A − λi I) along R (A − λi I)ki . • G1 + G2 + · · · + Gs = I. (7.9.10) • Gi Gj = 0 when i = j. (7.9.11) • Ni = (A − λi I)Gi = Gi (A − λi I) is nilpotent of index ki . (7.9.12) • If A is diagonalizable, then (7.9.9) reduces to (7.3.6) on p. 526, and the spectral projectors reduce to those described on p. 517. 604 Chapter 7 Eigenvalues and Eigenvectors Proof of (7.9.10)–(7.9.12). Property (7.9.10) results from using (7.9.9) with the function f (z) = 1, and property (7.9.11) is a consequence of I if i = j, I = P−1 P =⇒ Qi Pj = (7.9.13) 0 if i = j. To prove (7.9.12), establish that (A − λi I)Gi = Gi (A − λi I) by noting that T (7.9.13) implies P−1 Gi = 0 · · · Qi · · · 0 and Gi P = 0 · · · Pi · · · 0 . Use this with (7.9.4) to observe that (A − λi I)Gi = Pi Li Qi = Gi (A − λi I). Now Nji = (Pi Li Qi)j = Pi Lji Qi for j = 1, 2, 3, . . . , and thus Ni is nilpotent of index ki because Li is nilpotent of index ki . Example 7.9.1 A coordinate-free version of the representation in (7.9.3) results by separating the first-order terms in (7.9.9) from the higher-order terms to write k s i −1 (j) f (λi)Gi + f (A) = Ni . j! i=1 j=1 Using the identity function f (z) = z produces a coordinate-free version of the Jordan decomposition of A in the form s A= λ i Gi + Ni , i=1 and this is the extension of (7.2.7) on p. 517 to the nondiagonalizable case. Another version of (7.9.9) results from lumping things into one matrix to write f (A) = s k i −1 i=1 j=0 f (j) (λi)Zij , where Zij = (A − λi I)j Gi . j! (7.9.14) The Zij 's are often called the component matrices or the constituent matrices. Example 7.9.2 6 2 Problem: Describe f (A) for functions f defined at A = −2 0 2 0 8 −2 2 . Solution: A is block triangular, so it's easy to see that λ1 = 2 and λ2 = 4 are the two distinct eigenvalues with index (λ1) = 1 and index (λ2) = 2. Thus f (A) exists for all functions such that f (2), f (4), and f (4) exist, in which case f (A) = f (2)G1 + f (4)G2 + f (4)(A − 4I)G2 . The spectral projectors could be computed directly, but things are easier if some judicious choices of f are made. For example, f (z) = 1 ⇒ I = f (A) = G1 + G2 G1 = (A − 4I)2 /4, =⇒ G2 = I − G1 . f (z) = (z − 4)2 ⇒ f (A) = (A − 4I)2 = 4G1 7.9 Functions of Nondiagonalizable Matrices 605 Now that the spectral projectors are known, any function defined at A can be evaluated. For example, if f (z) = z 1/2 , then √ 5 1 7 − 2√ 2 √ √ √ √ 1 f (A) = A = 2G1 + 4G2 + (1/ 4)(A − 4I)G2 = −1 3 5 −√4 2 . 2 0 0 2 2 This technique illustrated above is rather ad hoc, but it always works if a sufficient number of appropriate functions are used. For example, using f (z) = z p for p = 0, 1, 2, . . . will always produce a system of equations that will yield the component matrices Zij given in (7.9.14) because for f (z) = 1 : I = Zi0 , for f (z) = z : A = λi Zi0 + Zi1 , for f (z) = z 2 : A2 = λ2i Zi0 + 2λi Zi1 + 2Zi2 , .. and this can be considered as a generalized Vandermonde linear system (p. 185) 1 ··· 1 Z.10 . I . 1 ··· 1 λ1 · · · λ s Zs0 A 2 Z11 2 2 · · · 2λ 2λ · · · 2λ 2 λ 1 .. = A 1 s s . 3 .. . .. . Zs1 A . Z . 21 . .. ··· ··· ··· . that can be solved for the Zij 's. Other sets of polynomials such as {1, (z − λ1)k1 , (z − λ1)k2 (z − λ2)k2 , . . . (z − λ1)k1 · · · (z − λs)ks } will generate other linear systems that yield solutions containing the Zij 's. Example 7.9.3 ∞ j Series Representations. Suppose that j=0 cj (z − z0) converges to f (z) at each point inside a circle |z − z0 | = r, and suppose that A is a matrix such that |λi − z0 | < r for each eigenvalue λi ∈ σ (A). ∞ j Problem: Explain why j=0 cj (A − z0 I) converges to f (A). Solution: If P−1 AP = J is in Jordan form as described on p. 601, then it's ∞ j not difficult to argue that j=0 cj (A − z0 I) converges if and only if ∞ P−1 j=0 ∞ j=0 cj (A−z0 I)j P= ∞ j=0 cj P−1 (A−z0 I)j P= j=0 cj (J−z0 I)j = . ∞ j=0 cj (J − z0 I) .. j 606 Chapter 7 Eigenvalues and Eigenvectors converges. Consequently, it suffices to prove that to f (J#) for a generic k × k Jordan block λ J# = ∞ j=0 cj (J# ! 1 . . . . . λ 0 = λI + N, where N= − z0 I)j converges 1 .. . .. . 0 ! . k×k ∞ j A standard theorem from analysis states that if j=0 cj (z − z0) converges to f (z) when |z − z0 | < r, then the series may be differentiated term by term to yield series that converge to derivatives of f at points inside the circle of convergence. Consequently, for each i = 0, 1, 2, . . . , ∞ f (i) (z) j (z − z0)j−i i! j=0 i when |z − z0 | < r. (7.9.15) We know from (7.9.1) (with f (z) = z j) that (J# − z0 I)j = (λ − z0)j I + j (λ − z0)j−1 N + · · · + (λ − z0)j−(k−1) Nk−1 , 1 k−1 so this together with (7.9.15) produces j (λ − z0)j−1 N cj (J# − z0 I)j = cj (λ − z0)j I + cj 1 j=0 j=0 j=0 ∞ j + · · · + cj (λ − z0)j−(k−1) Nk−1 k − 1 j=0 ∞ = f (λ)I + f (λ)N + · · · + f (k−1) (λ)Nk−1 = f (J∗). (k − 1)! Note: The result of this example validates the statements made on p. 527. Example 7.9.4 All Matrix Functions Are Polynomials. It was pointed out on p. 528 that if A is diagonalizable, and if f (A) exists, then there is a polynomial p(z) such that f (A) = p(A), and you were asked in Exercise 7.3.7 (p. 539) to use the Cayley–Hamilton theorem (pp. 509, 532) to extend this property to nondiagonalizable matrices for functions that have an infinite series expansion. We can now see why this is true in general. Problem: For a function f defined at A ∈ C n×n , exhibit a polynomial p(z) such that f (A) = p(A). 7.9 Functions of Nondiagonalizable Matrices 607 Solution: Suppose that σ (A) = {λ1 , λ2 , . . . , λs } with index (λi) = ki . The trick is to find a polynomial p(z) such that for each i = 1, 2, . . . , s, p(λi) = f (λi), p (λi) = f (λi), ..., p(ki −1) (λi) = f (ki −1) (λi) (7.9.16) because if such a polynomial exists, then (7.9.9) guarantees that p(A) = s k i −1 p(j) (λi) i=1 j=0 j! (A − λi I)j Gi = s k i −1 f (j) (λi) i=1 j=0 j! (A − λi I)j Gi = f (A). s Since there are k = i=1 ki equations in (7.9.16) to be satisfied, let's look for a polynomial of the form p(z) = α0 + α1 z + α2 z 2 + · · · + αk−1 z k−1 by writing the equations in (7.9.16) as the following k × k linear system Hx = f : . . . p (λi) = ... p (λi) = ... ⇒ 1 . . . = 1 1 . . f (λi) = 0 . . . . . . f (λi) = 0 . . . . . . p(λ1) = f (λ1) . . . p(λs) = f (λs) ··· λ1 . . . λs λ21 . . . λ2s λ31 . . . λ3s . . . 1 . . . . . . 2λi . . . . . . 3λ2i . . . ··· . . . 0 . . . . . . 2 . . . . . . 6λi . . . ··· α 0 f (λ1) α1 f (λs) α2 α k−2 f (λi) (k − 1)λi (k−3) (k − 1)(k − 2)λi f (λi) . . . . αk−1 . λk−1 1 . . . k−1 λs The coefficient matrix H can be proven to be nonsingular because the rows in each segment of H are linearly independent. The rows in the top segment of H are a subset of rows from a Vandermonde matrix (p. 185), while the nonzero portion of each succeeding segment has the form VD, where the rows of V are a subset of rows from a Vandermonde matrix and D is a nonsingular diagonal matrix. Consequently, Hx = f has a unique solution, and thus there is a unique polynomial p(z) = α0 + α1 z + α2 z 2 + · · · + αk−1 z k−1 that satisfies the conditions in (7.9.16). This polynomial p(z) is called the Hermite interpolation polynomial, and it has the property that f (A) = p(A). 608 Chapter 7 Eigenvalues and Eigenvectors Example 7.9.5 Functional Identities. Scalar functional identities generally extend to the matrix case. For example, the scalar identity sin2 z + cos2 z = 1 extends to matrices as sin2 Z + cos2 Z = I, and this is valid for all Z ∈ C n×n . While it's possible to prove such identities on a case-by-case basis by using (7.9.3) or (7.9.9), there is a more robust approach that is described below. For two functions f1 and f2 from C into C and for a polynomial p(x, y) in two variables, let h be the composition defined by h(z) = p f1 (z), f2 (z) . If An×n has eigenvalues σ (A) = {λ1 , λ2 , . . . , λs } with index (λi) = ki , and if h is defined at A, then we are allowed to assert that h(A) = p f1 (A), f2 (A) because Example 7.9.4 insures that there are polynomials g(z) and q(z) such that h(A) = g(A) and p f1 (A), f2 (A) = q(A), where for each λi ∈ σ (A) , g (j) (λi) = h (j) d j p f1 (z), f2 (z) (λi) = dz j = q (j) (λi) for j = 0, 1, . . . , ki − 1, z=λi so g(A) = q(A), and thus h(A) = p f1 (A), f2 (A) . To build functional identities for A, choose f1 and f2 in h(z) = p f1 (z), f2 (z) that will make h(λi) = h (λi) = h (λi) = · · · = h(ki −1) (λi) = 0 for each λi ∈ σ (A) , thereby insuring that 0 = h(A) = p f1 (A), f2 (A) . This technique produces a plethora of functional identities. For example, using f1 (z) = sin2 z f2 (z) = cos2 z produces h(z) = p f1 (z), f2 (z) = sin2 z + cos2 z − 1. p(x, y) = x2 + y 2 − 1 Since h(z) = 0 for all z ∈ C, it follows that h(Z) = 0 for all Z ∈ C n×n , and thus sin2 Z+cos2 Z = I for all Z ∈ C n×n . It's evident that this technique can be extended to include any number of functions f1 , f2 , . . . , fm with a polynomial p(x1 , x2 , . . . , xm) to produce even more complicated relationships. Example 7.9.6 Systems of Differential Equations Revisited. The purpose here is to extend the discussion in §7.4 to cover the nondiagonalizable case. Write the system of differential equations in (7.4.1) on p. 541 in matrix form as u (t) = An×n u(t) with u(0) = c, (7.9.17) but this time don't assume that An×n is diagonalizable—suppose instead that σ (A) = {λ1 , λ2 , . . . , λs } with index (λi) = ki . The development parallels that 7.9 Functions of Nondiagonalizable Matrices 609 for the diagonalizable case, but eAt is now a little more complicated than (7.4.2). Using f (z) = ezt in (7.9.3) and (7.9.2) yields 2 λt k−1 λt eAt = P.. ! . J t e .. P−1 with eJ t . = eλt teλt t e 2! eλt teλt .. ··· t e (k − 1)! .. .. . eλt . . t2 eλt 2! teλt , (7.9.18) eλt while setting f (z) = e zt in (7.9.9) produces eAt = s k i −1 j λi t t e i=1 j=0 j! (A − λi I)j Gi . (7.9.19) Either of these can be used to show that the three properties (7.4.3)–(7.4.5) on p. 541 still hold. In particular, d eAt /dt = AeAt = eAt A, so, just as in the diagonalizable case, u(t) = eAt c is the unique solution of (7.9.17) (the uniqueness argument given in §7.4 remains valid). In the diagonalizable case, the solution of (7.9.17) involves only the eigenvalues and eigenvectors of A as described in (7.4.7) on p. 542, but generalized eigenvectors are needed for the nondiagonalizable case. Using (7.9.19) yields the solution to (7.9.17) as u(t) = eAt c = s k i −1 j λi t t e i=1 j=0 j! vj (λi), where vj (λi) = (A − λi I)j Gi c. (7.9.20) Each vki −1 (λi) is an eigenvector associated with λi because (A − λi I)ki Gi = 0, and {vki −2 (λi), . . . , v1 (λi), v0 (λi)} is an associated chain of generalized eigenvectors. The behavior of the solution (7.9.20) as t → ∞ is similar but not identical to that discussed on p. 544 because for λ = x + iy and t > 0, 0 if x < 0, unbounded if x ≥ 0 and j > 0, tj eλt = tj ext (cos yt + i sin yt) → oscillates indefinitely if x = j = 0 and y = 0, if x = y = j = 0. 1 In particular, if Re (λi) < 0 for every λi ∈ σ (A) , then u(t) → 0 for every initial vector c, in which case the system is said to be stable. • Nonhomogeneous Systems. It can be verified by direct manipulation that the solution of u (t) = Au(t) + f (t) with u(t0) = c is given by 4 t A(t−t0) u(t) = e c+ eA(t−τ) f (τ)dτ. t0 610 Chapter 7 Eigenvalues and Eigenvectors Example 7.9.7 Nondiagonalizable Mixing Problem. To make the point that even simple problems in nature can be nondiagonalizable, consider three V gallon tanks as shown in Figure 7.9.1 that are initially full of polluted water in which the ith tank contains ci lbs of a pollutant. In an attempt to flush the pollutant out, all spigots are opened at once allowing fresh water at the rate of r gal/sec to flow into the top of tank #3, while r gal/sec flow from its bottom into the top of tank #2, and so on. Fresh r gal/sec 3 r gal/sec 2 r gal/sec 1 r gal/sec Figure 7.9.1 Problem: How many pounds of the pollutant are in each tank at any finite time t > 0 when instantaneous and continuous mixing occurs? Solution: If ui (t) denotes the number of pounds of pollutant in tank i at time t > 0, then the concentration of pollutant in tank i at time t is ui (t)/V lbs/gal, so the model ui (t) = (lbs/sec) coming in−(lbs/sec) going out produces the nondiagonalizable system: u1 (t) −1 1 0 u1 (t) c1 r u u (t) (t) = Au with u(0) = c = c2 . 0 −1 1 = , or u 2 2 V c3 u (t) u (t) 0 0 −1 3 3 This setup is almost the same as that in Exercise 3.5.11 (p. 104). Notice!that A is simply a scalar multiple of a single Jordan block J = −1 0 0 1 −1 0 0 1 −1 , so eAt is easily determined by replacing t by rt/V and λ by −1 in the second equation of (7.9.18) to produce 2 1 rt/V (rt/V) /2 eAt = e(rt/V)J = e−rt/V 0 1 rt/V . 0 0 1 7.9 Functions of Nondiagonalizable Matrices 611 Therefore, u(t) = eAt c = e−rt/V 2 c1 + c2 (rt/V) + c3 (rt/V) /2 , c2 + c3 (rt/V) c3 and, just as common sense dictates, the pollutant is never completely flushed from the tanks in finite time. Only in the limit does each ui → 0, and it's clear that the rate at which u1 → 0 is slower than the rate at which u2 → 0, which in turn is slower than the rate at which u3 → 0. Example 7.9.8 The Cauchy integral formula is an elegant result from complex analysis stating that if f : C → C is analytic in and on a simple closed contour Γ ⊂ C with positive (counterclockwise) orientation, and if ξ0 is interior to Γ, then 4 4 f (ξ) f (ξ) 1 j! (j) f (ξ0) = dξ and f (ξ0) = dξ. (7.9.21) 2πi Γ ξ − ξ0 2πi Γ (ξ − ξ0)j+1 These formulas produce analogous representations of matrix functions. Suppose that A ∈ C n×n with σ (A) = {λ1 , λ2 , . . . , λs } and index (λi) = ki . For a complex variable ξ, the resolvent of A ∈ C n×n is defined to be the matrix R(ξ) = (ξI − A)−1 . If ξ ∈ σ (A) , then r(z) = (ξ − z)−1 is defined at A with r(A) = R(ξ), so the spectral resolution theorem (p. 603) can be used to write R(ξ) = s k i −1 r(j) (λi) i=1 j=0 j! (A − λi I)j Gi = s k i −1 i=1 j=0 1 (A − λi I)j Gi . (ξ − λi)j+1 If σ (A) is in the interior of a simple closed contour Γ, and if the contour integral of a matrix is defined by entrywise integration, then (7.9.21) produces 4 4 1 1 −1 f (ξ)(ξI − A) dξ = f (ξ)R(ξ)dξ 2πi Γ 2πi Γ 4 s k i −1 1 f (ξ) = (A − λi I)j Gi j+1 & 4 s k i −1 % 1 f (ξ) = dξ (A − λi I)j Gi j+1 2πi Γ (ξ − λ) i Γ i=1 j=0 = s k i −1 f (j) (λi) i=1 j=0 j! (A − λi I)j Gi = f (A). 612 Chapter 7 Eigenvalues and Eigenvectors • In other words, if Γ is a simple closed contour containing σ (A) in its interior, then 4 1 f (A) = f (ξ)(ξI − A)−1 dξ (7.9.22) 2πi Γ whenever f is analytic in and on Γ. Since this formula makes sense for general linear operators, it is often adopted as a definition for f (A) in more general settings. • Furthermore, if Γi is a simple closed contour enclosing λi but excluding all other eigenvalues of A, then the ith spectral projector is given by 4 4 1 1 Gi = R(ξ)dξ = (ξI − A)−1 dξ (Exercise 7.9.19). 2πi Γi 2πi Γi Exercises for section 7.9 7.9.1. Lake #i in a closed system of three lakes of equal volume V initially contains ci lbs of a pollutant. If the water in the system is circulated at rates (gal/sec) as indicated in Figure 7.9.2, find the amount of pollutant in each lake at time t > 0 (assume continuous mixing), and then determine the pollution in each lake in the long run. 2r 4r #1 3r #2 2r #3 r Figure 7.9.2 7.9.2. Suppose that A ∈ C n×n has eigenvalues λi with index (λi) = ki . Explain why the ith spectral projector is given by 5 1 when z = λi , Gi = fi (A), where fi (z) = 0 otherwise. 7.9.3. Explain why each spectral projector Gi can be expressed as a polynomial in A. 7.9.4. If σ (An×n) = {λ1 , λ2 , . . . , λs } with ki = index (λi), explain why s k i −1 k k−j Ak = λi (A − λi I)j Gi . j i=1 j=0 7.9 Functions of Nondiagonalizable Matrices 613 7.9.5. With the convention that λ k 1 .. ... . λ m×m k j λk = = 0 for j > k, explain why k 1 k λk−1 2 k λk 1 . . . λk−2 · · · λk−1 .. . . λk k m−1 λk−m+1 .. . k 2 λk−2 1 λk−1 k λk 7.9.6. Determine e A 6 2 for A = −2 0 2 0 8 −2 2 . √ 7.9.7. For f (z) = 4 z − 1, determine f (A) when A = 7.9.8. −3 −8 −9 5 −1 11 −2 9 1 . (a) Explain why every nonsingular A ∈ C n×n has a square root. √ (b) Give necessary and sufficient conditions for the existence of A when A is singular. 7.9.9. Spectral Mapping Property. Prove that if (λ, x) is an eigenpair for A, then (f (λ), x) is an eigenpair for f (A) whenever f (A) exists. Does it also follow that alg multA (λ) = alg multf (A) (f (λ))? 7.9.10. Let f be defined at A, and let λ ∈ σ (A) . Give an example or an explanation of why the following statements are not necessarily true. (a) f (A) is similar to A. (b) geo multA (λ) = geo multf (A) (f (λ)) . (c) indexA (λ) = indexf (A) (f (λ)). 7.9.11. Explain why Af (A) = f (A)A whenever f (A) exists. 7.9.12. Explain why a function f is defined at A ∈ C n×n if and only if f T is defined at AT , and then prove that f (AT) = f (A) . Why can't (%)∗ be used in place of (%)T ? 614 Chapter 7 Eigenvalues and Eigenvectors 7.9.13. Use the technique of Example 7.9.5 (p. 608) to establish the following identities. (a) eA e−A = I for all A ∈ C n×n . α for all α ∈ C and A ∈ C n×n . (b) eαA = eA iA (c) e = cos A + i sin A for all A ∈ C n×n . 7.9.14. (a) Show that if AB = BA, then eA+B = eA eB . (b) Give an example to show that eA+B = eA eB in general. 7.9.15. Find the Hermite interpolation polynomial as described in Exam p(z) 3 2 1 A ple 7.9.4 such that p(A) = e for A = −3 −2 −1 . −3 −2 −1 7.9.16. The Cayley–Hamilton theorem (pp. 509, 532) says that every A ∈ C n×n satisfies its own characteristic equation, and this guarantees that An+j (j = 0, 1, 2, . . .) can be expressed as a polynomial in A of at most degree n − 1. Since f (A) is always a polynomial in A, the Cayley– Hamilton theorem insures that f (A) can be expressed as a polynomial in A of at most degree n − 1. Such a polynomial can be determined whenever f (j) (λi), j = 0, 1, . . . , ai − 1 exists for each λi ∈ σ (A) , where ai = alg mult (λi) . The strategy is the same as that in Example 7.9.4 except that ai is used in place of ki . If we can find a polynomial p(z) = α0 + α1 z + · · · + αn−1 z n−1 such that for each λi ∈ σ (A) , p(λi) = f (λi), p (λi) = f (λi), ..., p(ai −1) (λi) = f (ai −1) (λi), then p(A) = f (A). Why? These equations are an n × n linear system with the αi 's as the unknowns, and, for the same reason outlined in Example 7.9.4, a solution is always possible. (a) What advantages and disadvantages does this approach have with respect to the approach in Example 7.9.4? (b) Use this method to finda polynomial p(z) such that p(A) = eA for A = 7.9.17. Show that if f α β A=0 α 0 0 3 −3 −3 2 −2 −2 1 −1 −1 . Compare with Exercise 7.9.15. is a function defined at γ β = αI + βN + γN2 , α where 0 N = 0 0 β 2 f (α) 2 then f (A) = f (α)I + βf (α)N + γf (α) + N . 2! 1 0 0 0 1 0 7.9 Functions of Nondiagonalizable Matrices 615 7.9.18. Composition of Matrix Functions. If h(z) = f (g(z)), where f and g are functions such that g(A) and f g(A) each exist, then h(A) = f g(A) . However, it's not legal to prove this simply by saying "replace z by A." One way to prove that h(A) = f g(A) is to demonstrate that h(J#) = f g(J#) for a generic Jordan block and then invoke (7.9.3). Do this for a 3 × 3 Jordan block—the generalization to k × k blocks is similar. That is, let h(z) = f (g(z)), and use Exercise 7.9.17 to prove that if g(J#) and f g(J#) each exist, then λ 1 0 h(J#) = f g(J#) for J# = 0 λ 1 . 0 0 λ 7.9.19. Prove that if Γi is a simple closed contour enclosing λi ∈ σ (A) but excluding all other eigenvalues of A, then the ith spectral projector is 4 4 1 1 −1 Gi = (ξI − A) dξ = R(ξ)dξ. 2πi Γi 2πi Γi 7.9.20. For f (z) = z −1 , verify that f (A) = A−1 for every nonsingular A. 7.9.21. If Γ is a simple closed contour enclosing 4 all eigenvalues of a nonsingular 1 matrix A, what is the value of ξ −1 (ξI − A)−1 dξ ? 2πi Γ 7.9.22. Generalized Inverses. The inverse function f (z) = z −1 is not defined at singular matrices, but the generalized inverse function z −1 if z = 0, g(z) = 0 if z = 0, is defined on all square matrices. It's clear from Exercise 7.9.20 that if A is nonsingular, then g(A) = A−1 , so g(A) is a natural way to extend the concept of inversion to include singular matrices. Explain why g(A) = AD is the Drazin inverse of Example 5.10.5 (p. 399) and not necessarily the Moore–Penrose pseudoinverse A† described on p. 423. 7.9.23. Drazin Is "Natural." Suppose that A is a singular matrix, and let Γ be a simple closed contour that contains all eigenvalues of A except λ1 = 0, which is neither in nor on Γ. Prove that 4 1 ξ −1 (ξI − A)−1 dξ = AD 2πi Γ is the Drazin inverse for A as defined in Example 5.10.5 (p. 399). Hint: The Cauchy–Goursat theorem states that if a function f6 is analytic at all points inside and on a simple closed contour Γ, then Γ f (z)dz = 0. 616 7.10 Chapter 7 Eigenvalues and Eigenvectors DIFFERENCE EQUATIONS, LIMITS, AND SUMMABILITY A linear difference equation of order m with constant coefficients has the form y(k + 1) = αm y(k) + αm−1 y(k − 1) · · · + α1 y(k − m + 1) + α0 (7.10.1) in which α0 , α1 , . . . , αm along with initial conditions y(0), y(1), . . . , y(m − 1) are known constants, and y(m), y(m + 1), y(m + 2) . . . are unknown. Difference equations are the discrete analogs of differential equations, and, among other ways, they arise by discretizing differential equations. For example, discretizing a second-order linear differential equation results in a system of second-order difference equations as illustrated in Example 1.4.1, p 19. The theory of linear difference equations parallels the theory for linear differential equations, and a technique similar to the one used to solve linear differential equations with constant coefficients produces the solution of (7.10.1) as α0 + βi λki , 1 − α1 − · · · − αm i=1 m y(k) = for k = 0, 1, . . . (7.10.2) in which the λi 's are the roots of λm − αm λm−1 − · · · − α0 = 0, and the βi 's are constants determined by the initial conditions y(0), y(1), . . . , y(m − 1). The first term on the right-hand side of (7.10.2) is a particular solution of (7.10.1), and the summation term in (7.10.2) is the general solution of the associated homogeneous equation defined by setting α0 = 0. This section focuses on systems of first-order linear difference equations with constant coefficients, and such systems can be written in matrix form as x(k + 1) = Ax(k) (a homogeneous system) x(k + 1) = Ax(k) + b(k) (a nonhomogeneous system), or (7.10.3) where matrix An×n , the initial vector x(0), and vectors b(k), k = 0, 1, . . . , are known. The problem is to determine the unknown vectors x(k), k = 1, 2, . . . , along with an expression for the limiting vector limk→∞ x(k). Such systems are used to model linear discrete-time evolutionary processes, and the goal is usually to predict how (or to where) the process eventually evolves given the initial state of the process. For example, the population migration problem in Example 7.3.5 (p. 531) produces a 2 × 2 system of homogeneous linear difference equations (7.3.14), and the long-run (or steady-state) population distribution is obtained by finding the limiting solution. More sophisticated applications are given in Example 7.10.8 (p. 635) and Example 8.3.7 (p. 683). 7.10 Difference Equations, Limits, and Summability 617 Solving the equations in (7.10.3) is easy. Direct substitution verifies that x(k) = Ak x(0) for and x(k) = Ak x(0) + k = 1, 2, 3, . . . k−1 (7.10.4) Ak−j−1 b(j) for k = 1, 2, 3, . . . j=0 are respective solutions to (7.10.3). So rather than finding x(k) for any finite k, the real problem is to understand the nature of the limiting solution limk→∞ x(k), and this boils down to analyzing limk→∞ Ak . We begin this analysis by establishing conditions under which Ak → 0. For scalars α we know that αk → 0 if and only if |α| < 1, so it's natural to ask if there is an analogous statement for matrices. The first inclination is to replace | % | by a matrix norm % , but this doesn't work for the standard norms. For example, if A = 00 20 , then Ak → 0 but A = 2 for all of the standard matrix norms. Although it's possible to construct a rather goofy-looking matrix norm % g such that Ag < 1 when limk→∞ Ak = 0, the underlying mechanisms governing convergence to zero are better understood and analyzed by using eigenvalues and the Jordan form rather than norms. In particular, the spectral radius of A defined as ρ(A) = maxλ∈σ(A) |λ| (Example 7.1.4, p. 497) plays a central role. Convergence to Zero For A ∈ C n×n , Proof. lim Ak = 0 k→∞ if and only if ρ(A) < 1. (7.10.5) If P−1 AP = J is the Jordan form for A, then Ak = PJk P−1 = P .. . −1 P , Jk .. . where J# = λ 1 . . . . . (7.10.6) . λ denotes a generic Jordan block in J. Clearly, Ak → 0 if and only if Jk# → 0 for each Jordan block, so it suffices to prove that Jk# → 0 if and only if |λ| < 1. Using the function f (z) = z n in formula (7.9.2) on p. 600 along with the k convention that j = 0 for j > k produces 618 Chapter 7 Eigenvalues and Eigenvectors λk k J# = k 1 λk−1 λk k 2 k 1 .. . λk−2 · · · .. λk−1 .. k m−1 λk−m+1 .. . k 2 λk−2 1 λk−1 k λk λk (7.10.7) m×m It's clear from the diagonal entries that if Jk# → 0, then λk → 0, so |λ| < 1. k−j k = 0 for each fixed value of j Conversely, if |λ| < 1 then limk→∞ j λ because k k−j k j k−j k k(k − 1) · · · (k − j + 1) kj = → 0. ≤ =⇒ λ ≤ |λ| j j! j! j j! You can see that the last term on the right-hand side goes to zero as k → ∞ either by applying l'Hopital's rule or by realizing that k j goes to infinity with polynomial speed while |λ|k−j is going to zero with exponential speed. Therefore, if |λ| < 1, then Jk# → 0, and thus (7.10.5) is proven. Intimately related to the question of convergence to zero is the convergence ∞ k of the Neumann series k=0 A . It was demonstrated in (3.8.5) on p. 126 that if limn→∞ An = 0, then the Neumann series converges, and it was argued in Example 7.3.1 (p. 527) that the converse holds for diagonalizable matrices. Now we are in a position to prove that the converse is true for all square matrices and thereby produce the following complete statement regarding the convergence of the Neumann series. Neumann Series For A ∈ C n×n , the following statements are equivalent. • The Neumann series I + A + A2 + · · · converges. (7.10.8) • ρ(A) < 1. (7.10.9) • lim Ak = 0. k→∞ In which case, (I − A)−1 exists and (7.10.10) ∞ k=0 Ak = (I − A)−1 . (7.10.11) Proof. We know from (7.10.5) that (7.10.9) and (7.10.10) are equivalent, and it was argued on p. 126 that (7.10.10) =⇒ (7.10.8), the theorem can be estabso ∞ k lished by proving that (7.10.8) implies (7.10.9). If k=0 A converges, it follows ∞ k that k=0 J∗ must converge for each Jordan block J∗ in the Jordan form for A. ∞ ∞ k k This together with (7.10.7) implies that J = k=0 ∗ ii k=0 λ converges for 7.10 Difference Equations, Limits, and Summability 619 each λ ∈ σ (A) , and this scalar series converges if and only if |λ| < 1. ∞ geometric k Thus the convergence of A implies ρ(A) < 1. When it converges, k=0 ∞ k −1 because (I − A)(I + A + A2 + · · · + Ak−1) = I − Ak → k=0 A = (I − A) I as k → ∞. The following examples illustrate the utility of the previous results for establishing some useful (and elegant) statements concerning spectral radius. Example 7.10.1 Spectral Radius as a Limit. It was shown in Example 7.1.4 (p. 497) that if A ∈ C n×n , then ρ (A) ≤ A for every matrix norm. But this was just the precursor to the following elegant relationship between spectral radius and norm. Problem: Prove that for every matrix norm,))1/k ρ (A) = lim)Ak) . (7.10.12) k→∞))1/k) k Solution: First note that ρ (A) = ρ Ak ≤)Ak) =⇒ ρ (A) ≤)Ak) . Next, observe that ρ A/(ρ (A) +) < 1 for every > 0, so, by (7.10.5),) k))A) A lim = 0 =⇒ lim = 0. k→∞ ρ (A) + k→∞ (ρ (A) +)k Consequently, there is a positive integer K such that)Ak) /(ρ (A) +)k < 1)1/k for all k ≥ K , so)Ak) < ρ (A) + for all k ≥ K , and thus))1/k ρ (A) ≤)Ak) < ρ (A) + for k ≥ K .))1/k Because this holds for each > 0, it follows that limk→∞)Ak) = ρ(A). Example 7.10.2 For A ∈ C n×n let |A| denote the matrix having entries |aij |, and for matrices B, C ∈ n×n define B ≤ C to mean bij ≤ cij for each i and j. Problem: Prove that if |A| ≤ B, then ρ (A) ≤ ρ (|A|) ≤ ρ (B) . (7.10.13) Solution: The triangle inequality yields |Ak | ≤ |A|k for every positive integer k. Furthermore, |A| ≤ B implies that |A|k ≤ Bk . This with (7.10.12) produces)) k k→∞ ==⇒ ρ (A) ≤ ρ (|A|) ≤ ρ (B) . ∞ k→∞ ∞ 620 Chapter 7 Eigenvalues and Eigenvectors Example 7.10.3 Problem: Prove that if 0 ≤ Bn×n , then ρ (B) < r if and only if (rI − B)−1 exists and (rI − B)−1 ≥ 0. k=0 To prove the converse, it's convenient to adopt the following notation. For any P ∈ m×n , let |P| = |pij | denote the matrix of absolute values, and notice that the triangle inequality insures that |PQ| ≤ |P| |Q| for all conformable P and Q. Now assume that rI − B is nonsingular and (rI − B)−1 ≥ 0, and prove ρ (B) < r. Let (λ, x) be any eigenpair for B, and use B ≥ 0 together with (rI − B)−1 ≥ 0 to write λx = Bx ==⇒ |λ| |x| = |λx| = |Bx| ≤ |B| |x| = B |x| ==⇒ (rI − B)|x| ≤ (r − |λ|) |x| ==⇒ 0 ≤ |x| ≤ (r − |λ|) (rI − B)−1 |x| (7.10.15) ==⇒ r − |λ| ≥ 0. But |λ| = r is impossible because (rI − B)|x| ≤ 0 ... algorithms are often used in lieu of direct methods to solve large sparse systems of linear equations, and some of the traditional iterative schemes fall into the following class of nonhomogeneous linear difference equations. Linear Stationary Iterations Let Ax = b be a linear system that is square but otherwise arbitrary. • A splitting of A is a factorization A = M−N, where M−1 exists. • Let H = M−1 N (called the iteration matrix), and set d = M−1 b. • For an initial vector x(0)n×1 , a linear stationary iteration is x(k) = Hx(k − 1) + d, k = 1, 2, 3, ... (7.10.16) • If ρ(H) < 1, then A is nonsingular and lim x(k) = x = A−1 b for every initial vector x(0). 7.10 Difference Equations, Limits, and Summability 621 Proof. To prove (7.10.17), notice that if A = M − N = M(I − H) is a splitting for which ρ(H) < 1, then (7.10.11) guarantees that (I − H)−1 exists, and thus A is nonsingular. Consequently, successive substitution applied to (7.10.16) yields ... all x(0), lim x(k) = (I − H)−1 d = (I − H)−1 M−1 b = A−1 b = x. k→∞ (7.10.18) It's clear that the convergence rate of (7.10.16) is governed by the size of ρ(H) along with the index of its associated eigenvalue. But what really is needed is an indication of how many digits of accuracy can be expected to be gained per iteration. So as not to obscure the simple underlying idea, assume that Hn×n is diagonalizable with σ (H) = {λ1 , λ2 , . . . , λs } , where 1 > |λ1 | > |λ2 | ≥ |λ3 | ≥ · · · ≥ |λs | (which is frequently the case in applications), and let (k) = x(k) − x denote the error after the kth iteration. Subtracting x = Hx + d (a consequence of (7.10.18)) from x(k) = Hx(k − 1) + d produces (for large k) (k) = H(k − 1) = Hk (0) = (λk1 G1 + λk2 G2 + · · · + λks Gs)(0) ≈ λk1 G1 (0), where the Gi 's are the spectral projectors occurring in the spectral decomposition (pp. 517 and 520) of Hk . Similarly, (k − 1) ≈ λk−1 G1 (0), so comparing the ith components of (k − 1) and (k) reveals that after several iterations, 1 1 i (k) ≈ |λ1 | = ρ (H) for each i = 1, 2, . . . , n. To understand the significance of this, suppose for example that |i (k − 1)| = 10−q and |i (k)| = 10−p with p ≥ q > 0, so that the error in each entry is reduced by p − q digits per iteration. Since ... of the number of digits of accuracy that can be expected to be eventually gained on each iteration. For this reason, the number R = − log10 ρ (H) (or, alternately, R = − ln ρ (H)) is called the asymptotic rate of convergence, and this is the primary tool for comparing different linear stationary iterative algorithms. The trick is to find splittings that guarantee rapid convergence while insuring that H = M−1 N and d = M−1 b can be computed easily. The following three examples present the classical splittings. 622 Chapter 7 Eigenvalues and Eigenvectors Example 7.10.4 Jacobi's method is produced by splitting A = D − N, where D is the diagonal part of A (we assume each aii = 0), and −N is the matrix containing the off-diagonal entries of A. Clearly, both H = D−1 N and d = D−1 b can be formed with little effort. Notice that the ith component in the Jacobi iteration x(k) = D−1 Nx(k − 1) + D−1 b is given by ... the algorithm can process equations independently (or in parallel). For this reason, Jacobi's method was referred to in the 1940s as the method of simultaneous displacements. Problem: Explain why Jacobi's method is guaranteed to converge for all initial vectors x(0) and for all right-hand sides b when A is diagonally dominant as defined and discussed in Examples 4.3.3 (p. 184) and 7.1.6 (p. 499). Solution: According to (7.10.17), it suffices to show that ρ(H) < 1. This follows by combining |aii | > j =i |aij | for each i with the fact that ρ(H) ≤)H)∞ (Example 7.1.4, p. 497) to write ρ(H) ≤)H)∞ = max i ... Example 7.10.5 The Gauss–Seidel method is the result of splitting A = (D−L)−U, where D is the diagonal part of A (aii = 0 is assumed) and where −L and −U contain the entries occurring below and above the diagonal of A, respectively. The iteration matrix is H = (D − L)−1 U, and d = (D − L)−1 b. The ith entry in the Gauss–Seidel iteration is xi (k) = (bi − ji aij xj (k − 1)) /aii . This shows that Gauss–Seidel determines xi (k) by using the newest possible information—namely, x1 (k), x2 (k), . . . , xi−1 (k) in the current iterate in conjunction with xi+1 (k − 1), xi+2 (k − 1), . . . , xn (k − 1) from the previous iterate. 81 82 Karl Jacobi (p. 353) considered this method in 1845, but it seems to have been independently discovered by others. In addition to being called the method of simultaneous displacements in 1945, Jacobi's method was referred to as the Richardson iterative method in 1958. Ludwig Philipp von Seidel (1821–1896) studied with Jacobi in Berlin in 1840 and with Gauss in Göttingen in 1842 ... transforming Jacobi's method into the Gauss–Seidel scheme is natural, but the reason for attaching Gauss's name is unclear. Seidel went on to earn his doctorate (1846) in Munich, where he stayed as a professor for the rest of his life. In addition to mathematics, Seidel made notable contributions in the areas of optics and astronomy, and in 1970 a lunar crater was named for Seidel. 7.10 Difference Equations, Limits, and Summability 623 This differs from Jacobi's method because Jacobi relies strictly on the old data in x(k − 1). The Gauss–Seidel algorithm is known in the 1940s as the method of successive displacements (as opposed to the method of simultaneous displacements, which is Jacobi's method). Because Gauss–Seidel computes xi (k) with newer data than that used by Jacobi, it appears at first glance that Gauss–Seidel should be the superior algorithm. While this is often the case, it is not universally true—see Exercise 7.10.7. Other Comparisons. Another major difference between Gauss–Seidel and Jacobi is that the order in which the equations are processed is irrelevant for Jacobi's method, but the value (not just the position) of the components xi (k) in the Gauss–Seidel iterate can change when the order of the equations is changed. Since this ordering feature can affect the performance of the algorithm, it was the object of much study at one time. Furthermore, when core memory is a concern, Gauss–Seidel enjoys an advantage because as soon as a new component xi (k) is computed, it can immediately replace the old value xi (k − 1), whereas Jacobi requires all old values in x(k − 1) to be retained until all new values in x(k) have been determined. Something that both algorithms have in common is that diagonal dominance in A guarantees global convergence of each method. Problem: Explain why diagonal dominance in A is sufficient to guarantee convergence of the Gauss–Seidel method for all initial vectors x(0) and for all right-hand sides b . Solution: Show ρ (H) < 1. Let (λ, z) be any eigenpair for H, and suppose that the component of maximal magnitude in z occurs in position m. Write (D − L)−1 Uz = λz as λ(D − L)z = Uz, and write the mth row of this latter equation as λ(d − l) = u, where d = amm zm , l = − ... Example 7.10.6 The successive overrelaxation (SOR) method improves on Gauss–Seidel by introducing a real number ω = 0, called a relaxation parameter, to form the splitting A = M − N, where M = ω−1 D − L and N = (ω−1 − 1)D + U. As before, D is the diagonal part of A (aii = 0 is assumed) and −L and −U contain the entries occurring below and above the diagonal of A, respectively. Since M−1 = ω(D − ωL)−1 = ω(I − ωD−1 L)−1 D−1 , the SOR iteration matrix is Hω = M−1 N = (D − ωL)−1 (1 − ω)D + ωU = (I − ωL)−1 (1 − ω)I + ωU , and the kth SOR iterate emanating from (7.10.16) is x(k) = Hω x(k − 1) + ω(I − ωL)−1 D−1 b. This is the Gauss–Seidel iteration when ω = 1. Using ω > 1 is called overrelaxation, while taking ω < 1 is referred to as underrelaxation. Writing (7.10.20) in the form (I − ωL)x(k) = (1 − ω)I + ωU x(k − 1) + ωD−1 b and considering the ith component on both sides of this equality produces ... Convergence and Restrictions on ω. In spite of the fact that SOR is the Gauss–Seidel method with a "correction term" ... known as "successive overrelaxation" rather than simply "successive relaxation." It's not hard to see that ρ (Hω) < 1 only if 0 < ω < 2 (Exercise 7.10.9), and it can be proven that positive definiteness of A is sufficient to guarantee ρ (Hω) < 1 whenever 0 < ω < 2. But determining ω to minimize ρ (Hω) is generally a difficult task. 83 Nevertheless, there is one famous special case for which the optimal value of ω can be explicitly given. If det (αD − L − U) = det αD − βL − β−1 U for all real α and β = 0, and if the iteration matrix HJ for Jacobi's method has real eigenvalues with ρ (HJ) < 1, then the eigenvalues λJ for HJ are related to the eigenvalues λω of Hω by (λω + ω − 1)2 = ω2 λω λ2J . (7.10.23) From this it can be proven that the optimum value of ω for SOR is ωopt = 2 / (1 + (1 − ρ2 (HJ))1/2) and ρ (Hωopt) = ωopt − 1. Furthermore, ρ (HGS) = ρ2 (HJ), where HGS is the Gauss–Seidel iteration matrix. For example, the discrete Laplacian Lnn2 ×n2 in Example 7.6.2 (p. 563) satisfies the special case conditions, and the spectral radii of the iteration matrices associated with L are Jacobi: ρ (HJ) = cos π/(n + 1) ≈ 1 − π2 /2 h2 , Gauss–Seidel: ρ (HGS) = cos2 π/(n + 1) ≈ 1 − π2 h2 , SOR: ρ (Hωopt) = ... where h = 1/(n + 1) is the mesh size (Exercise 7.10.10). Therefore, asymptotic rates of convergence reveals that Gauss–Seidel is twice as fast as Jacobi on the discrete Laplacian because RGS = − log10 cos2 πh = −2 log10 cos πh = 2RJ . However, optimal SOR is much better because ... is substantially smaller than 1 − π2 h2 ≈ ρ (HGS) for even moderately small h. The point is driven home by looking at the asymptotic rates of convergence for h = .02 (n = 49) as shown below. Jacobi: RJ ≈ .000858, Gauss–Seidel: RGS = 2RJ ≈ .001716, SOR: Ropt ≈ .054611 ≈ 32RGS ≈ 63RJ . This special case was developed by the contemporary numerical analyst David M. Young, Jr., who produced much of the SOR theory in his 1950 Ph.D. dissertation that was directed by Garrett Birkhoff at Harvard University. The development of SOR is considered to be one of the major computational achievements of the first half of the twentieth century, and it motivated at least two decades of intense effort in matrix computations. 626 Chapter 7 Eigenvalues and Eigenvectors for the discrete Laplacian L ... Note: In spite of the impressive results of SOR ... and 64 Jacobi steps! Note: In spite of the preceding remarks, SOR has limitations. Special cases for which the optimum ω can be explicitly determined are rare, so adaptive computational procedures are generally necessary to approximate a good ω, and the results are often not satisfying. While SOR was a big step forward over the algorithms of the nineteenth century, the second half of the twentieth century saw the development of more robust methods—such as the preconditioned conjugate gradient method (p. 657) and GMRES (p. 655)—that have relegated SOR to a secondary role. Example 7.10.7 M-matrices are real nonsingular matrices Ann such that aij ≤ 0 for all i = j and A−1 ≥ 0, where each entry of A−1 is nonnegative. They arise naturally in a broad variety of applications ranging from economics (Example 8.3.6, p. 681) to hard-core engineering problems, and, as shown in (7.10.29), they are particularly relevant in formulating and analyzing iterative methods. Some important properties of M-matrices are developed below. • A is an M-matrix if and only if there exists a matrix B ≥ 0 and a real number r > ρ(B) such that A = rI − B. (7.10.24) • If A is an M-matrix, then Re (λ) > 0 for all λ ∈ σ (A). Conversely, all matrices with nonpositive off-diagonal entries whose spectrums are in the right-hand halfplane are M-matrices. (7.10.26) • Principal submatrices of M-matrices are also M-matrices. • If A is an M-matrix, then all principal minors in A are positive. Conversely, all matrices with nonpositive off-diagonal entries whose principal minors are positive are M-matrices. (7.10.28) • If A = M − N is a splitting of an M-matrix for which M−1 ≥ 0 and N ≥ 0, then the linear stationary iteration (7.10.16) converges for all initial vectors x(0) and for all right-hand sides b. In particular, Jacobi's method converges for A an M-matrix. ... Suppose that A is an M-matrix. Let r = max |aii | so that B = rI − A ≥ 0 ... Conversely, if A = rI − B with B ≥ 0 and r > ρ(B), then (7.10.14) in Example 7.10.3 (p. 620) that r > ρ(B). Conversely, if A is any matrix of the form A = rI − B, where B ≥ 0 and r > ρ(B), use (7.10.14) again to guarantee that A−1 ≥ 0. ... 84 This terminology was introduced in 1937 by the twentieth-century mathematician Alexander Markowitz Ostrowski, who made several contributions to the analysis of classical iterative methods. The "M" is short for "Minkowski" (p. 278). 7.10 Difference Equations, Limits, and Summability 627 To prove (7.10.25), note that if A = rI − B is an M-matrix, then (7.10.24) insures that r > ρ(B) ... To prove (7.10.26), if A is an M-matrix, then, by (7.10.25), λ ∈ σ(A) implies that Re (λ) > 0 ... Conversely, suppose A has nonpositive off-diagonal entries and every eigenvalue of A has a positive real part. Let r be any real number such that r > max{2 max |aii |, max |λ|} ... This means that there is a real number such that the circle centered at r and having radius equal to r contains σ(A) ... As such, all eigenvalues of B look like rλ ... , and hence ρ(B) < r. It now follows from (7.10.24) that A is an M-matrix. To prove (7.10.27), if Akk is a principal submatrix of A, then there is a permutation matrix P such that ... If B ≥ 0 is the corresponding principal submatrix of B, then ... To prove (7.10.28), note that if A is an M-matrix, then det (A) > 0 because the eigenvalues of a real matrix appear in complex conjugate pairs, so (7.10.26) and (7.1.8), p. 494, guarantee that det (A) = λi > 0. It follows that each principal minor is positive because each submatrix of an M-matrix is again an M-matrix. Proceed by induction on n ... assume that the result is true for n = k, and consider the LU factorization ... the lower-right-hand entry ... must be positive. Therefore, the matrix is nonsingular and the induction hypothesis ensures that A−1 ≥ 0 ... Conversely ... To prove (7.10.29), note that ... Jacobi's method is a special case because the Jacobi splitting is A = D − N, where D = diag (a11 , a22 , . . . , ann), and (7.10.28) insures that each aii > 0. Note: Comparing properties of M-matrices with those of positive definite matrices reveals many parallels, and, in a rough sense, an M-matrix often plays the role of "a positive definite matrix" for nonsymmetric matrices. Only a small sample of M-matrix theory has been presented here, but there is in fact enough to fill a monograph on the subject. For example, there are at least 50 known equivalent conditions that can be imposed on a real matrix with nonpositive off-diagonal entries (often called a Z-matrix) to guarantee that it is an M-matrix—see Exercise 7.10.12 for a sample of such conditions in addition to those listed above. 7.10 Difference Equations, Limits, and Summability 629 We now turn our attention to the limiting behavior of powers of matrices. We know that limk→∞ Ak exists (Exercise 7.10.11), and we can decide whether it does or not by using ... In particular, limk→∞ Ak exists if and only if ρ(A) < 1 or else ρ(A) = 1, where λ = 1 is the only eigenvalue on the unit circle, and λ = 1 is semisimple. ... Convergence to Zero ... a Jordan block ... J = λI + N ... Jk = ... For λ ∈ C and such that |λ| = 1, the diagonal terms λk oscillate indefinitely, and this prevents Jk from having a limit. When λ = 1, the Jordan block has a limiting value if and only if m = 1, which is equivalent to saying that λ = 1 is a semisimple eigenvalue. But λ = 1 may be repeated p times so that there are p Jordan blocks of the form [1]1×1 . Consequently, limk→∞ Ak exists if and only if the Jordan form for A has the structure ... where p = alg mult (1) and ρ(K) < 1. Now that we know when limk→∞ Ak exists, let's describe what limk→∞ Ak looks like. We already know the answer when p = 0—it's 0 (because ρ(A) < 1). But if p is nonzero, then limk→∞ Ak = G is the projector onto N (I − A) along R (I − A), which can be explained using the spectral projector. ... 630 Chapter 7 Eigenvalues and Eigenvectors Limits of Powers For A ∈ C n×n , limk→∞ Ak exists if and only if ρ(A) < 1, in which case limk→∞ Ak = 0, or else ρ(A) = 1, where λ = 1 is the only eigenvalue on the unit circle, and λ = 1 is semisimple. When it exists, limk→∞ Ak = the projector onto N (I − A) along R (I − A). (7.10.34) With each scalar sequence {α1 , α2 , α3 , . . .} there is an associated sequence of averages {μ1 , μ2 , μ3 , . . .} in which μn = (α1 + α2 + · · · + αn)/n. This sequence of averages is called the Cesàro sequence, and when limn→∞ μn = α, we say that {αn } is Cesàro summable to α. It can be proven (Exercise 7.10.11) that if {αn } converges to α, then {μn } converges to α, but not conversely. In other words, convergence implies summability, but summability doesn't insure convergence. To see that a sequence can be summable without being convergent, notice that the oscillatory sequence {0, 1, 0, 1, . . .} doesn't converge, but it is Cesàro summable to 1/2, which is the mean value of {0, 1}. This is typical because averaging has a smoothing effect so that oscillations that prohibit convergence of the original sequence tend to be smoothed away or averaged out in the Cesàro sequence. 85 Ernesto Cesàro (1859–1906) was an Italian mathematician who worked mainly in differential geometry but also contributed to number theory, divergent series, and mathematical physics. After studying in Naples, Liège, and Paris, Cesàro received his doctorate from the University of Rome in 1887, and he went on to occupy the chair of mathematics at Palermo. Cesàro's most important contribution is considered to be his 1890 book Lezione di geometria intrinseca, but, in large part, his name has been perpetuated because of its attachment to the concept of Cesàro summability of matrices (Exercise 7.10.11), but Cesàro summability is particularly interesting when it is applied to the sequence P = {Ak }∞k=0 of powers of a square matrix A. We know from (7.10.33) and (7.10.34) under what conditions P converges as well as to what P converges. In this section we address the question of when P is Cesàro summable, and to what P sums. 7.10 Difference Equations, Limits, and Summability 631 From now on, we will say that Ann is a convergent matrix when limk→∞ Ak exists, and we will say that A is a summable matrix when limk→∞ (I + A + A2 + · · · + Ak−1)/k exists. As in the scalar case, if A is convergent to G, then A is summable to G, but not conversely (Exercise 7.10.11). To analyze the summability of A in the absence of convergence, begin with the observation that A is summable if and only if the Jordan form J for A is summable, which in turn is equivalent to saying that each Jordan block J in J is summable. Consequently, A cannot be summable whenever ρ(A) > 1, because if J is a Jordan block in which |λ| > 1, then each diagonal entry of J k is λk , and ... Now suppose that ρ(A) ≤ 1 ... this means that Jk is summable if and only if λ = 1 ... Consequently, Jk → 0 ... which is not the case when |λ| = 1 ... implies that λ = 1 ... Therefore, it's necessary that λ = 1 is a semisimple eigenvalue of A when it is on the unit circle ... In other words, A is summable if and only if ρ(A) < 1 or else ρ(A) = 1 with each eigenvalue on the unit circle being semisimple. ... We already know what (I − A)... is when ρ(A) < 1, but when ρ(A) = 1 the limit is the projector G onto N (I − A) along R (I − A). 632 Chapter 7 Eigenvalues and Eigenvectors Summability ... Cesàro summability is a particularly important concept for the study of Markov chains and stochastic processes. ... 7.10 Difference Equations, Limits, and Summability 633 Example 7.10.8 Shell Game. As depicted in Figure 7.10.2, a pea is placed under one of four shells, and an agile manipulator quickly rearranges them by a sequence of discrete moves. At the end of each move the shell containing the pea has been shifted either to the left or right by only one position according to the following rules. Figure 7.10.2 When the pea is under shell #1, it is moved to position #2, and if the pea is under shell #4, it is moved to position #3. When the pea is under shell #2 or #3, it is equally likely to be moved one position to the left or to the right. Problem 1: Given that we know something about where the pea starts, what is the probability of finding the pea in any given position after k moves? Problem 2: In the long run, what proportion of time does the pea occupy each of the four positions? Solution to Problem 1: Let pj (k) denote the probability that the pea is in position j after the kth move, and translate the given information into four difference equations by writing p1 (k) = p2 (k−1)/2, p2 (k) = p1 (k−1) + p3 (k−1)/2, p3 (k) = p2 (k−1)/2 + p4 (k−1), p4 (k) = p3 (k−1)/2 ... The matrix equation on the right-hand side is a homogeneous difference equation p(k) = Ap(k − 1) whose solution, from (7.10.4), is p(k) = Ak p(0), and thus Problem 1 is solved. For example, if you know that the pea is initially under shell #2, then p(0) = e2 , and after six moves the probability that the pea is in the fourth position is p4 (6) = A6 e2 4 = 21/64. If you don't know exactly where the pea starts, but you assume that it is equally likely to start under any one of the four shells, then p(0) = (1/4, 1/4, 1/4, 1/4)T , and the probabilities 636 Chapter 7 Eigenvalues and Eigenvectors for occupying the four positions after six moves are given by p(6) = A6 p(0) = (11/32, 21/32, 21/64, 11/64) ... Solution to Problem 2: There is a straightforward solution when A is a convergent matrix because if Ak → G as k → ∞, then p(k) → Gp(0) = p, and the components in this limiting (or steady-state) vector p provide the answer. Intuitively, if p(k) → p, then after awhile p(k) is practically constant, so the probability that the pea occupies a particular position remains essentially the same move after move. Consequently, the components in limk→∞ p(k) reveal the proportion of time spent in each position over the long run. For example, if limk→∞ p(k) = (1/6, 1/3, 1/3, 1/6)T , then, as the game runs on indefinitely, the pea is expected to be under shell #1 for about 16.7% of the time, under shell #2 for about 33.3% of the time, etc. A Fly in the Ointment: Everything above rests on the assumption that A is convergent. But A is not convergent for the shell game because a bit of computation reveals that σ (A) = {±1, ±(1/2)}. That is, there is an eigenvalue other than 1 on the unit circle, so (7.10.33) guarantees that limk→∞ Ak does not exist. Consequently, there's no limiting solution p to the difference equation p(k) = Ap(k − 1), and the intuitive analysis given above does not apply. Cesàro to the Rescue: However, A is summable because ρ (A) = 1, and every eigenvalue on the unit circle is semisimple—these are the conditions in (7.10.36). So as k → ∞, I + A + · · · + Ak−1 /k → G, and p(0) → Gp(0) = p. The job now is to interpret the meaning of this Cesàro limit in the context of the shell game. To do so, focus on a particular position—say the jth one—and set up "counting functions" (random variables) defined as X(0) = 1 if the pea starts under shell j, 0 otherwise, and X(i) = 1 if the pea is under shell j after the ith move, 0 otherwise, i = 1, 2, 3, ... Notice that X(0) + X(1) + · · · + X(k − 1) counts the number of times the pea occupies position j before the kth move, so X(0) + X(1) + · · · + X(k − 1) /k represents the fraction of times that the pea is under shell j before the kth move. Since the expected (or mean) value of X(i) is, by definition, E[X(i)] = 1 × P X(i) = 1 + 0 × P X(i) = 0 = pj (i), and since expectation is linear (E[αX(i) + X(h)] = αE[X(i)] + E[X(h)]), the expected fraction of times that the pea occupies position j before move k is E[X(0) + X(1) + · · · + X(k − 1)]/k = ... = [p(0) + Ap(0) + · · · + Ak−1 p(0)]/k = ... In other words, as the game progresses indefinitely, the components of the Cesàro limit p = Gp(0) provide the expected proportion of times that the pea is under each shell, and this is exactly what we wanted to know. Computing the Limiting Vector. Of course, p can be determined by first computing G, but there is some special structure in this problem that can be exploited to make the task easier. Recall from (7.2.12) on p. 518 that if λ is a simple eigenvalue for A, and if x and y* are respective right-hand and left-hand eigenvectors associated with λ, then xy*/y*x is the projector onto N (A − λI) along R (A − λI). For the shell game, λ = 1 is a simple eigenvalue for A. Furthermore, we get an associated left-hand eigenvector for free—namely, eT = (1, 1, 1, 1) because each column sum of A is one, so eT A = eT. Consequently, if x is any right-hand eigenvector of A associated with λ = 1, then (by noting that eT p(0) = p1 (0) + p2 (0) + p3 (0) + p4 (0) = 1) the limiting vector is given by p = Gp(0) = xeT p(0)/eT x = x/eT x = x/Σ xi . In other words, the limiting vector is obtained by normalizing any nonzero solution of (I − A)x = 0 to make the components sum to one. Not only does (7.10.44) show how to compute the limiting proportions, it also shows that the limiting proportions are independent of the initial values in p(0). For example, a simple calculation reveals that x = (1, 2, 2, 1)T is one solution of (I − A)x = 0, so the vector of limiting proportions is p = (1/6, 1/3, 1/3, 1/6)T . Therefore, if many moves are made, then, regardless of where the pea starts, we expect the pea to end up under shell #1 for about 16.7% of the time, under #2 for about 33.3% of the time, under #3 for about 33.3% of the moves, and under shell #4 for about 16.7% of the moves. Note: The shell game (and its analysis) is a typical example of a random walk with reflecting barriers, and these problems belong to a broader classification of stochastic processes known as irreducible, periodic Markov chains. (Markov chains are discussed in detail in §8.4 on p. 687.) The shell game is periodic in the sense of Exercise 4.4.20 (p. 209), and it is periodic because the pea can return to a given position only at definite periods, as reflected in the periodicity of the powers of A. More details are given in Example 8.4.3 on p. 694. Exercises for section 7.10 7.10.1. Which of the 1/2 A= ... 7.10.2. Verify that the limit of each convergent matrix, and evaluate the limit of each summable matrix. 7.10.3. Verify that each convergent matrix ... 7.10.4. Determine the limiting vector for the shell game in Example 7.10.8 by first computing the Cesàro limit G with a full-rank factorization. 7.10.5. Verify that the expressions in (7.10.4) are indeed the solutions to the difference equations in (7.10.3). 7.10.6. Prove that if there exists a matrix norm such that A < 1, then ... 7.10.7. Examining the iteration matrix, compare the convergence of Jacobi's method and the Gauss–Seidel method ... 7.10.8. Explain why the iteration x(k) = Hx(k − 1) + d converges for all x(0) ... 7.10.9. Show that the spectral radius of the SOR iteration matrix satisfies ρ(Hω) ≥ |ω − 1|, and conclude that 0 < ω < 2 is necessary for convergence. 7.10.10. Use the eigenvalues of the discrete Laplacian ... (a) Verify that A satisfies the special case conditions given in Example 7.10.6 that guarantee the validity of (7.10.24). (b) Determine the optimum SOR relaxation parameter. (c) Find the asymptotic rates of convergence for Jacobi, Gauss–Seidel, and optimum SOR. (d) Use x(0) = (1, 1)T and b = (2, 4)T to run through several steps of Jacobi, Gauss–Seidel, and optimum SOR to solve Ax = b until you can see a convergence pattern. 7.10.9. Prove that if ρ (Hω) < 1, where Hω is the iteration matrix for the SOR method, then 0 < ω < 2. Hint: Use det (Hω) to show |λk | ≥ |1 − ω| for some λk ∈ σ (Hω). 7.10.10. Show that the spectral radius of the Jacobi iteration matrix for the discrete Laplacian L described in Example 7.6.2 (p. 563) is ρ (HJ) = cos π/(n + 1). 7.10.11. Consider a scalar sequence {α1 , α2 , α3 , . . .} and the corresponding Cesàro sequence of averages {μ1 , μ2 , μ3 , . . .}. (a) Prove that if lim αn = α, then lim μn = α. (b) Prove that if A is convergent, then A is summable ... 7.10.12. M-matrices Revisited. For matrices with nonpositive off-diagonal entries (Z-matrices), the following conditions are equivalent: (a) A is an M-matrix. (b) All leading principal minors of A are positive. (c) A has an LU factorization, and both L and U are M-matrices. (d) There exists a vector x > 0 such that Ax > 0. (e) Each aii > 0 and AD is diagonally dominant for some diagonal matrix D with positive diagonal entries. ... 640 Chapter 7 Eigenvalues and Eigenvectors 7.10.13. Index by Full-Rank Factorization. Suppose that λ ∈ σ (A), and let M1 = A − λI. Let M1 = B1 C1 be a full-rank factorization, and set M2 = C1 B1 . If M2 is nonsingular, stop; otherwise, factor M2 = B2 C2 and set M3 = C2 B2 ... In general, Mi = Ci−1 Bi−1 = Bi Ci is a full-rank factorization. Explain why this procedure must eventually produce a matrix Mk that is either nonsingular or zero. Prove that if k is the smallest positive integer such that Mk−1 exists or Mk = 0, then index (λ) = k if Mk is nonsingular, and index (λ) = k − 1 if Mk = 0. ... 7.10.14. Use the procedure of Exercise 7.10.13 to find the index of each eigenvalue of A = ... 7.10.15. Let A be the matrix given in Exercise 7.10.14. (a) Find the Jordan form for A. (b) For any function f defined at A, find the Hermite interpolation polynomial that is described in Example 7.9.4 (p. 606), and describe f(A). 7.10.16. Limits and Group Inversion. Given a matrix Bn×n of rank r such that index(B) ≤ 1 (i.e., index (λ = 0) ≤ 1), the Jordan form for B looks like ... so B# is the group inverse (or the Drazin inverse) of B. The Drazin inverse was discussed in Example 5.10.5 on p. 399, and the group inverse was developed in Exercises 5.10.11–5.10.13 on p. 402. ... 7.11 MINIMUM POLYNOMIALS AND KRYLOV METHODS It is a fundamental fact (the Cayley–Hamilton theorem) that every square matrix satisfies its own characteristic equation ... Since the degree of the characteristic polynomial c(x) of An×n is n, ... The characteristic polynomial plays a central role in the theoretical development of linear algebra and matrix analysis, but it is not alone in this respect. There are other polynomials that occur naturally, and the purpose of this section is to explore some of them. In this section it is convenient to consider the characteristic polynomial of A ∈ C n×n to be c(x) = det (xI − A). This differs from the definition given on p. 492 only in the sense that the coefficients of c(x) = det (xI − A) have different signs than the coefficients of ĉ(x) = det (A − xI). In particular, c(x) is a monic polynomial (i.e., its leading coefficient is 1), whereas the leading coefficient of ĉ(x) is (−1)n. (Of course, the roots of c and ĉ are identical.) Monic polynomials p(x) such that p(A) = 0 are said to be annihilating polynomials for A. For example, the Cayley–Hamilton theorem (pp. 509, 532) guarantees that c(x) is an annihilating polynomial of degree n. Minimum Polynomial for a Matrix There is a unique annihilating polynomial for A of minimal degree, and this polynomial, denoted by m(x), is called the minimum polynomial for A. The Cayley–Hamilton theorem guarantees that deg[m(x)] ≤ n. Proof. Only uniqueness needs to be proven. Let k be the smallest degree of any annihilating polynomial for A. There is a unique annihilating polynomial for A of degree k because if there were two different annihilating polynomials p1 (x) and p2 (x) of degree k, then d(x) = p1 (x) − p2 (x) would be a nonzero polynomial such that d(A) = 0 and deg[d(x)] < k. Dividing d(x) by its leading coefficient would produce an annihilating polynomial of degree less than k, the minimal degree, and this is impossible. The first problem is to describe what the minimum polynomial looks like, and the second problem is to uncover the relationship between m(x) and the characteristic polynomial c(x). The Jordan form for A reveals everything. Suppose that Pn×n J P−1 = A is the Jordan form for A, and ... is a Jordan block ... Since p(A) = 0 ⇐⇒ P−1 p(A)P = 0 ⇐⇒ p(P−1 AP) = 0 ⇐⇒ p(J) = 0, or, equivalently, p(J⋆) = 0 for each Jordan block J⋆ in J, it's clear that m(x) is the monic polynomial of smallest degree that annihilates all Jordan blocks. If J⋆ is a k × k Jordan block associated with an eigenvalue λ, then (7.9.2) on p. 600 insures that p(J⋆) = 0 if and only if p(i) (λ) = 0 for i = 0, 1, . . . , k − 1, and this happens if and only if p(x) = (x − λ)k q(x) for some polynomial q(x). Since this must hold for all Jordan blocks associated with λ, it must be true for the largest Jordan block associated with λ, and thus the minimum degree monic polynomial that annihilates all Jordan blocks associated with λ must be (x − λ)kj, where kj is the index of λ ... 7.11 Minimum Polynomials and Krylov Methods 643 Consequently, the minimum polynomial for A must annihilate the largest Jordan block associated with each λj ∈ σ (A) ... Determining the minimum polynomial ... this shows that m(x) = (x − λ1)k1 (x − λ2)k2 · · · (x − λs)ks ... where ki = index (λi) and their indices by ki. ... Since c(x) = (x − λ1)a1 (x − λ2)a2 · · · (x − λs)as where ai = alg mult (λi), it follows that m(x) divides c(x) ... Finding the eigenvalues and their indices can be a substantial task, so let's consider how we might construct m(x) without computing eigenvalues. An approach based on first principles is to determine the first matrix Ak that is a linear combination of the powers I, A, A2 , . . . , Ak−1 . ... The Gram–Schmidt orthogonalization procedure (p. 309) with the standard inner product A B = trace (A*B) ... is the perfect theoretical tool for determining k and the αj 's. Gram–Schmidt applied to {I, A, A2 , . . .} begins by setting U0 = I/‖I‖F = I/√n, and it proceeds sequentially ... The first such k is the smallest positive integer such that Ak ∈ span {U0 , U1 , . . . , Uk−1 } = span {I, A, . . . , Ak−1 } ... in matrix form ... 644 Chapter 7 Eigenvalues and Eigenvectors ... where t = ... m(x) = xk − αk−1 xk−1 − ... Example 7.11.1 Problem: Use the Gram–Schmidt orthogonalization procedure to determine the minimum polynomial of A ... Solution: ... Caution! While Gram–Schmidt works fine to produce m(x) in exact arithmetic, things are not so nice in floating-point arithmetic. For example, if A has a dominant eigenvalue, then, as explained in the power method (Example 7.3.7, p. 533), Ak asymptotically approaches the dominant spectral projector G1 ... so, as k grows, Ak becomes increasingly close to span {I, A, . . . , Ak−1 }, and hence Ak is ill-conditioned problem, and Gram–Schmidt may not work well in floating-point arithmetic—the modified Gram–Schmidt algorithm (p. 316), or a version of Householder reduction (p. 341), or Arnoldi's method (p. 653) works better. Fortunately, explicit knowledge of the minimum polynomial often is not needed in applied work. The relationship between the characteristic polynomial c(x) and the minimum polynomial m(x) for A is now transparent. Since c(x) = (x − λ1)a1 ... and m(x) = ... it's clear that m(x) divides c(x). Furthermore, m(x) = c(x) if and only if alg mult (λj) = index (λj) for each λj ∈ σ (A). Matrices for which m(x) = c(x) are said to be nonderogatory matrices, and they are precisely the ones for which geo mult (λj) = 1 for each eigenvalue λj because m(x) = c(x) if and only if there is only one Jordan block for each λj ... In addition to dividing the characteristic polynomial, the minimum polynomial divides all other annihilating polynomials as well because if p(x) is an annihilating polynomial for A, then the division algorithm (the quotient and remainder theorem) ... Minimum Polynomials and Krylov Methods 645 Since 0 = p(A) = m(A)q(A) + r(A) = r(A), it follows that r(x) = 0; otherwise r(x), when normalized to be monic, would be an annihilating polynomial having degree smaller than the degree of the minimum polynomial. The structure of the minimum polynomial for A is related to the diagonalizability of A. By combining the fact that kj = index (λj) is the size of the largest Jordan block for λj with the fact that A is diagonalizable if and only if all Jordan blocks are 1 × 1, it follows that A is diagonalizable if and only if kj = 1 for each j, which is equivalent to saying that m(x) = (x − λ1)(x − λ2) · · · (x − λs). In other words, A is diagonalizable if and only if its minimum polynomial is the product of distinct linear factors. Below is a summary of the preceding observations about properties of m(x). Properties of the Minimum Polynomial For A ∈ C n×n with σ (A) = {λ1 , λ2 , . . . , λs }, • the minimum polynomial of A is the unique monic polynomial m(x) of minimal degree such that m(A) = 0. • m(x) = (x − λ1)k1 (x − λ2)k2 · · · (x − λs)ks, where kj = index (λj). (7.11.1) • m(x) divides every polynomial p(x) such that p(A) = 0. In particular, m(x) divides the characteristic polynomial c(x). (7.11.2) • m(x) = c(x) if and only if geo mult (λj) = 1 for each λj or, equivalently, alg mult (λj) = index (λj) for each j, in which case A is called a nonderogatory matrix. • A is diagonalizable if and only if m(x) = (x − λ1)(x − λ2) · · · (x − λs) (i.e., if and only if m(x) is a product of distinct linear factors). (7.11.4) The next immediate aim is to extend the concept of the minimum polynomial for a matrix to formulate the notion of a minimum polynomial for a vector. To do so, it's helpful to introduce Krylov sequences, subspaces, and matrices. 86 Aleksei Nikolaevich Krylov (1863–1945) showed in 1931 how to use sequences of the form {b, Ab, A2 b, . . .} to construct the characteristic polynomial of a matrix (see Example 7.11.3 on p. 649). Krylov was a Russian applied mathematician whose scientific interests arose from his early training in naval science that involved the theories of buoyancy, stability, rolling and pitching, vibrations, and compass theories. Krylov served as the director of the Physics–Mathematics Institute of the Soviet Academy of Sciences from 1927 until 1932, and in 1943 he was awarded a "state prize" for his work on compass theory. Krylov was made a "hero of socialist labor," and he is one of a few mathematicians to have a lunar feature named in his honor—on the moon there is the "Crater Krylov." Krylov Sequences, Subspaces, and Matrices For A ∈ C n×n and 0 = b ∈ C n×1 , we adopt the following terminology. • {b, Ab, A2 b, . . . , Aj−1 b} is called a Krylov sequence. • Kj = span {b, Ab, . . . , Aj−1 b} is called a Krylov subspace. • Kn×j = [b | Ab | · · · | Aj−1 b] is called a Krylov matrix. Since dim(Kj) ≤ n (because Kj ⊆ C n×1), there is a first vector Ak b in the Krylov sequence that is a linear combination of preceding Krylov vectors. If ... then we define v(x) = xk − αk−1 xk−1 − · · · − α1 x − α0 to be the minimum polynomial for b with respect to A, and we say that v(x) is the minimum polynomial of b ... 646 Chapter 7 Eigenvalues and Eigenvectors The argument on p. 642 that establishes uniqueness of the minimum polynomial for matrices can be reapplied to prove that for each vector b = 0 there is a unique annihilating polynomial of minimal degree. In other words, v(x) is the unique monic polynomial of smallest degree such that v(A)b = 0. It is this fact that motivates the following definition. Minimum Polynomial for a Vector The minimum polynomial (with respect to A ∈ C n×n) for a vector b ∈ C n×1 is defined to be the monic polynomial v(x) of minimal degree such that v(A)b = 0. • If Ak b is the first vector in the Krylov sequence {b, Ab, A2 b, . . .} that is a linear combination of preceding Krylov vectors—say Ak b = (k−1 j=0 αj Aj b) (or αj = 0 if k = 0)—then v(x) = xk − Σ αj xj ... • The minimum polynomial for b divides the minimum polynomial for A. ... 7.11.5 Proof. The strategy first is to prove that if ℓ(x) is the LCM of the v(x) 's, then m(x) = ℓ(x) ... Minimum Polynomials and Krylov Methods 647 So v(x) divides m(x) ... Minimum Polynomials and Krylov Methods 649 Example 7.11.3 Krylov's Method for the Characteristic Polynomial ... Nonsingular Krylov matrix K ... C = K−1 AK is the companion matrix ... 648 Chapter 7 Eigenvalues and Eigenvectors Krylov's method ... the companion matrix of a monic polynomial p(x) = xn + αn−1 xn−1 + · · · + α1 x + α0 ... Since σ (C) = σ (A), the characteristic polynomial of A is the characteristic polynomial of C ... but Krylov's method is of theoretical interest only ... The Poor Man's Root Finder: The companion matrix is a poor man's root finder because the eigenvalues of the companion matrix C are the roots of p(x) ... 7.11 Minimum Polynomials and Krylov Methods 651 Krylov Methods ... Krylov methods ... the Krylov subspaces provide an effective means for approximating solutions of large linear systems Ax = b and for approximating eigenvalues ... This is the basic idea behind the methods of Lanczos and Arnoldi, and GMRES and conjugate gradients, discussed in the sequel. ... Arnoldi's method ... Lanczos ... Householder reduction ... upper-Hessenberg ... Cornelius Lanczos (1893–1974) was born Kornél Löwy in Budapest, Hungary, to Jewish parents, but he changed his name to avoid trouble during the dangerous times preceding World War II. After receiving his doctorate from the University of Budapest in 1921, Lanczos moved to Germany where he became Einstein's assistant in Berlin in 1928. After coming home to Germany from a visit to Purdue University in Lafayette, Indiana, in 1931, Lanczos decided that the political climate in Germany was unacceptable, and he returned to Purdue in 1932 to continue his work in mathematical physics. The development of electronic computers stimulated Lanczos's interest in numerical analysis, and this led to positions at the Boeing Company in Seattle and at the Institute for Numerical Analysis of the National Bureau of Standards in Los Angeles. When senator Joseph R. McCarthy led a crusade against communism in the 1950s, Lanczos again felt threatened, so he left the United States to accept an offer from the famous Nobel physicist Erwin Schrödinger (1887–1961) to head the Theoretical Physics Department at the Dublin Institute for Advanced Study in Ireland where Lanczos returned to his first love—the theory of relativity. Lanczos was aware of the fast Fourier transform algorithm (p. 373) 25 years before the heralded work of J. W. Cooley and J. W. Tukey (p. 386) in 1965, but 1940 was too early for applications of the FFT to be realized. This is yet another instance where Lanczos's intellect and insight were ahead of his time. Walter Edwin Arnoldi (1917–1995) was an American engineer who published this technique in 1951, not far from the time that Lanczos's algorithm emerged. Arnoldi received his undergraduate degree in mechanical engineering from Stevens Institute of Technology, Hoboken, New Jersey, in 1937 and his MS degree at Harvard University in 1939. He spent his career working as an engineer in the Hamilton Standard Division of the United Aircraft Corporation where he eventually became the division's chief researcher. He retired in 1977. While his research concerned mechanical and aerodynamic properties of aircraft and aerospace structures, Arnoldi's name is kept alive by his orthogonalization procedure. ... Arnoldi Orthogonalization Algorithm. Given A, the goal is to compute an orthonormal basis {q1 , q2 , . . . , qn } for the Krylov space ... Start with an arbitrary b = 0, set q1 = b/‖b‖, and then iterate as indicated below. ... Arnoldi and GMRES ... The Lanczos algorithm ... tridiagonal ... Note: Remarks similar to those made about the Lanczos algorithm also hold for Arnoldi's algorithm ... GMRES (generalized minimal residual) ... the least squares problem ... min ‖b − Ax‖ ... Hk+1 ... To compute the solution to a nonsingular linear system Ax = b with A ... start with qj ... and at the end of the kth step, the vector xk ... Let U3 = Q be the ...

4 × 4 orthogonal matrix that was vT 7 3 = R3 previously accumulated (as a product of plane rotations) to give U3 H 0 7 3 = QR3 . Since with R3 being upper triangular so that H 656 Chapter 7 U3 0 0 7 H4 = 1 Eigenvalues and Eigenvectors U3 0 [illegible] (1 a plane rotation of the form P45 = [illegible] will annihilate the entry in the lower-right-hand corner of this last array. Consequently, U4 = P45 U03 0 1 7 4 = R4 , where R4 is upper triangular and U4 is an orthogonal matrix such that U4 H 0 7 4, and this produces the QR factors of H Note: The value of the residual norm b − AQj z2 at each step of GMRES is available at almost no cost. To see why, notice that the previous discussion shows T that at the j th step there is a (j + 1) × (j + 1) orthogonal matrix U = Qj T P j = QR. The least squares solution of H 7 j z = be1 is and this produces H 2 obtained by solving R = QT b e1 (p. 314), so [illegible] b − AQj z2 = [illegible] = b2 |uj+1,1 |. Since uj+1,1 is just the last entry in the accumulation of the various plane rotations applied to e1 , the cost of producing these values as the algorithm proceeds is small, so deciding on the acceptability of an approximate solution at each step in the GMRES algorithm is cheap. When solving nonsingular symmetric systems Ax = b, a strategy similar to the one that produced GMRES can be adopted except that the Lanczos procedure (p. 651) is used in place of the Arnoldi process (p. 653). When this is done, the resulting algorithm is called MINRES (an acronym for minimal residual algorithm), and, as you might guess, there is an increase in computational efficiency when Lanczos replaces Arnoldi. Historically, MINRES preceded GMRES. Another Krylov method that deserves mention is the conjugate gradient algorithm, presented by Magnus R. Hestenes and Eduard Stiefel in 1952, that is used to solve positive definite systems. 7.11 Minimum Polynomials and Krylov Methods 657 Example 7.11.7 Conjugate Gradient Algorithm. [illegible]

[illegible]

= alg multB (1 + λ)n−1 . Consequently, if μ = ρ (B) , then μ = max |(1 + λ)| n−1 λ∈σ(A) = n−1 max |(1 + λ)| = (1 + r)n−1 λ∈σ(A) because when a circular disk |z| ≤ ρ is translated one unit to the right, the point of maximum modulus in the resulting disk |z + 1| ≤ ρ is z = 1 + ρ (it's clear if you draw a picture). Therefore, alg multA (r) = 1; otherwise alg multB (μ) > 1, which is impossible because B > 0. To see that A has a positive eigenvector 674 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices associated with r, recall from (8.3.2) that there exists a nonnegative eigenvector x ≥ 0 associated with r. It's a simple consequence of (7.9.9) that if (λ, x) is an eigenpair for A, then (f (λ), x) is an eigenpair for f (A) (Exercise 7.9.9, p. 613), so (r, x) being an eigenpair for A implies that (μ, x) is an eigenpair for B. Hence (8.2.10) insures that x must be a positive multiple of the Perron vector of B, and thus x must in fact be positive. Now, r > 0; otherwise Ax = 0, which is impossible because A ≥ 0 and x > 0 forces Ax > 0. The argument used to prove (8.2.10) also proves (8.3.9). Example 8.3.1 Problem: Suppose that An×n ≥ 0 is irreducible with r = ρ (A) , and suppose that rz ≤ Az for z ≥ 0. Explain why rz = Az, and z > 0. Solution: If rz < Az, then by using the Perron vector q > 0 for AT we have (A − rI)z ≥ 0 =⇒ qT (A − rI)z > 0, which is impossible since qT (A − rI) = 0. Thus rz = Az, and since z must be a multiple of the Perron vector for A by (8.3.9), we also have that z > 0. The only property in the list on p. 667 that irreducibility is not able to salvage is (8.2.15), which states that there is only one eigenvalue on the spectral circle. Indeed, A = 01 10 is nonnegative and irreducible, but the eigenvalues ±1 are both on the unit circle. The property of having (or not having) only one eigenvalue on the spectral circle divides the set of nonnegative irreducible matrices into two important classes. Primitive Matrices • A nonnegative irreducible matrix A having only one eigenvalue, r = ρ (A) , on its spectral circle is said to be a primitive matrix. • A nonnegative irreducible matrix having h > 1 eigenvalues on its spectral circle is called imprimitive, and h is referred to as index of imprimitivity. • A nonnegative irreducible matrix A with r = ρ (A) is primitive if and only if limk→∞ (A/r)k exists, in which case lim k→∞ A k r =G= pqT > 0, qT p (8.3.10) where p and q are the respective Perron vectors for A and AT . G is the (spectral) projector onto N (A − rI) along R(A − rI). 8.3 Nonnegative Matrices 675 Proof of (8.3.10). The Perron–Frobenius theorem insures that 1 = ρ(A/r) is a simple eigenvalue for A/r, and it's clear that A is primitive if and only if A/r is primitive. In other words, A is primitive if and only if 1 = ρ(A/r) is the only eigenvalue on the unit circle, which is equivalent to saying that limk→∞ (A/r)k exists by the results on p. 630. The structure of the limit as described in (8.3.10) is the result of (7.2.12) on p. 518. The next two results, discovered by Helmut Wielandt (p. 534) in 1950, establish the remarkable fact that the eigenvalues on the spectral circle of an imprimitive matrix are in fact the hth roots of the spectral radius. Wielandt's Theorem If |B| ≤ An×n , where A is irreducible, then ρ (B) ≤ ρ (A) . If equality holds (i.e., if μ = ρ (A) eiφ ∈ σ (B) for some φ), then ⎛ B = eiφ DAD−1 ⎞ D=⎜ ⎝ for some ⎟ eiθ1 eiθ2 .. ⎟ ⎠ . (8.3.11) eiθn and conversely. Proof. We already know that ρ (B) ≤ ρ (A) by Example 7.10.2 (p. 619). If ρ (B) = r = ρ (A) , and if (μ, x) is an eigenpair for B such that |μ| = r, then r|x| = |μ| |x| = |μx| = |Bx| ≤ |B| |x| ≤ A|x| =⇒ |B| |x| = r|x| because the result in Example 8.3.1 insures that A|x| = r|x|, and |x| > 0. Consequently, (A − |B|)|x| = 0. But A − |B| ≥ 0, and |x| > 0, so A = |B| by (8.2.4). Since xk /|xk | is on the unit circle, xk /|xk | = eiθk for some θk . Set ⎛ D=⎜ ⎝ eiθ1 eiθ2 .. ⎟, and notice that x = D|x|. ⎠. eiθn Since |μ| = r, there is a φ ∈ such that μ = reiφ , and hence BD|x| = Bx = μx = reiφ D|x| =⇒ e−iφ D−1 BD|x| = r|x| = A|x|. (8.3.12) For convenience, let C = e−iφ D−1 BD, and note that |C| = |B| = A to write (8.3.12) as 0 = (|C| − C)|x|. Considering only the real part of this equation 676 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices yields 0 = |C| − Re (C) |x|. But |C| ≥ Re (C) , and |x| > 0, so it follows from (8.2.4) that Re (C) = |C|, and hence 2 2 Re (cij) = |cij | = Re (cij) + Im (cij) =⇒ Im (cij) = 0 =⇒ Im (C) = 0. −1 Therefore, C = Re (C) = |C| = A, which implies B = eiφ DAD iφ . Conversely, iφ −1 if B = e DAD , then similarity insures that ρ (B) = ρ e A = ρ (A) . hth Roots of ρ (A) on Spectral Circle If An×n ≥ 0 is irreducible and has h eigenvalues {λ1 , λ2 , . . . , λh } on its spectral circle, then each of the following statements is true. • alg multA (λk) = 1 for k = 1, 2, . . . , h. • {λ1 , λ2 , . . . , λh } are the hth roots of r = ρ (A) given by {r, rω, rω 2 , . . . , rω h−1 }, where (8.3.13) ω = e2πi/h . (8.3.14) Proof. Let S = {r, reiθ1 , . . . , reiθh−1 } denote the eigenvalues on the spectral circle of A. Applying (8.3.11) with B = A and μ = reiθk insures the existence iθk of a diagonal matrix Dk such that A = eiθk Dk AD−1 A k , thus showing that e is similar to A. Since r is a simple eigenvalue of A (by the Perron–Frobenius theorem), reiθk must be a simple eigenvalue of eiθk A. But similarity transformations preserve eigenvalues and algebraic multiplicities (because the Jordan structure is preserved), so reiθk must be a simple eigenvalue of A, thus establishing (8.3.13). To prove (8.3.14), consider another eigenvalue reiθs ∈ S. Again, we can write A = eiθs Ds AD−1 for some Ds , so s −1 iθk i(θk +θr) A = eiθk Dk AD−1 Dk (eiθs Ds AD−1 (Dk Ds)A(Dk Ds)−1 s)Dk = e k =e and, consequently, rei(θk +θr) is also an eigenvalue on the spectral circle of A. In other words, S = {r, reiθ1 , . . . , reiθh−1 } is closed under multiplication. This means that G = {1, eiθ1 , . . . , eiθh−1 } is closed under multiplication, and it follows that G is a finite commutative group of order h. A standard result from algebra states that the hth power of every element in a finite group of order h must be the identity element in the group. Therefore, (eiθk)h = 1 for each 0 ≤ k ≤ h − 1, so G is the set of the hth roots of unity e2πki/h (0 ≤ k ≤ h − 1), and thus S must be the hth roots of r. Combining the preceding results reveals just how special the spectrum of an imprimitive matrix is. 8.3 Nonnegative Matrices 677 Rotational Invariance If A is imprimitive with h eigenvalues on its spectral circle, then σ (A) is invariant under rotation about the origin through an angle 2π/h. No rotation less than 2π/h can preserve σ (A) . (8.3.15) Proof. Since λ ∈ σ (A) ⇐⇒ λe2πi/h ∈ σ(e2πi/h A), it follows that σ(e2πi/h A) is σ (A) rotated through 2π/h. But (8.3.11) and (8.3.14) insure that A and e2πi/h A are similar and, consequently, σ (A) = σ(e2πi/h A). No rotation less than 2π/h can keep σ (A) invariant because (8.3.14) makes it clear that the eigenvalues on the spectral circle won't go back into themselves for rotations less than 2π/h. Example 8.3.2 The Spectral Projector Is Positive. We already know from (8.3.10) that if A is a primitive matrix, and if G is the spectral projector associated with r = ρ (A) , then G > 0. Problem: Explain why this is also true for an imprimitive matrix. In other words, establish the fact that if G is the spectral projector associated with r = ρ (A) for any nonnegative irreducible matrix A, then G > 0. Solution: Being imprimitive means that A is nonnegative and irreducible with more than one eigenvalue on the spectral circle. However, (8.3.13) says that each eigenvalue on the spectral circle is simple, so the results concerning Ces`aro summability on p. 633 can be applied to A/r to conclude that I + (A/r) + · · · + (A/r)k−1 = G, k→∞ k lim where G is the spectral projector onto N ((A/r) − I) = N (A − rI) along R((A/r) − I) = R(A − rI). Since r is a simple eigenvalue the same argument used to establish (8.3.10) (namely, invoking (7.2.12) on p. 518) shows that G= pqT > 0, qT p where p and q are the respective Perron vectors for A and AT . Trying to determine if an irreducible matrix A ≥ 0 is primitive or imprimitive by finding the eigenvalues is generally a difficult task, so it's natural to ask if there's another way. It turns out that there is, and, as the following example shows, determining primitivity can sometimes be trivial. 678 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Example 8.3.3 Sufficient Condition for Primitivity. If a nonnegative irreducible matrix A has at least one positive diagonal element, then A is primitive. Proof. Suppose there are h > 1 eigenvalues on the spectral circle. We know from (8.3.15) that if λ0 ∈ σ (A) , then λk = λ0 e2πik/h ∈ σ (A) for k = 0, 1, . . . , h − 1, so h−1 k=0 λk = λ0 h−1 e2πik/h = 0 (roots of unity sum to 1—see p. 357). k=0 This implies that the sum of all of the eigenvalues is zero. In other words, • if A is imprimitive, then trace (A) = 0. (Recall (7.1.7) on p. 494.) Therefore, if A has a positive diagonal entry, then A must be primitive. Another of Frobenius's contributions was to show how the powers of a nonnegative matrix determine whether or not the matrix is primitive. The exact statement is as follows. Frobenius's Test for Primitivity A ≥ 0 is primitive if and only if Am > 0 for some m > 0. (8.3.16) Proof. First assume that Am > 0 for some m. This implies that A is irreducible; otherwise there exists a permutation matrix such that m X Y X PT has zero entries. A=P PT =⇒ Am = P 0 Z 0 Zm Suppose that A has h eigenvalues {λ1 , λ2 , . . . , λh } on its spectral circle so that r = ρ (A) = |λ1 | = · · · = |λh | > |λh+1 | > · · · > |λn |. Since λ ∈ σ (A) implies λm ∈ σ(Am) with alg multA (λ) = alg multAm (λm) (consider the Jordan form—Exercise 7.9.9 on p. 613), it follows that λm k (1 ≤ k ≤ h) is on the spectral circle of Am with alg multA (λk) = alg multAm (λm k) . Perron's theorem (p. 667) insures that Am has only one eigenvalue (which must be rm) on m m its spectral circle, so rm = λm 1 = λ2 = · · · = λh . But this means that alg multA (r) = alg multAm (rm) = h, and therefore h = 1 by (8.3.7). Conversely, if A is primitive with r = ρ (A) , then limk→∞ (A/r)k > 0 by (8.3.10). Hence there must be some m such that (A/r)m > 0, and thus Am > 0. 8.3 Nonnegative Matrices 679 Example 8.3.4 Suppose that we wish to decide whether or not a nonnegative matrix A is primitive by computing the sequence of powers A, A2 , A3 , ... Since this can be a laborious task, it would be nice to know when we have computed enough powers of A to render a judgement. Unfortunately there is nothing in the statement or proof of Frobenius's test to help us with this decision. But Wielandt provided an answer by proving that a nonnegative matrix An×n is primitive if and only 2 if An −2n+2 > 0. Furthermore, n2 − 2n + 2 is the smallest such exponent that works for the class of n × n primitive matrices having all zeros on the diagonal—see Exercise 8.3.9. 0 1 0 Problem: Determine whether or not A = 0 0 2 is primitive. 3 4 0 Solution: Since A has zeros on the diagonal, the result in Example 8.3.3 doesn't apply, so we are forced into computing powers of A. This job is simplified by noticing that if B = β(A) is the Boolean matrix that results from setting 1 if aij > 0, bij = 0 if aij = 0, then [Bk]ij > 0 if and only if [Ak]ij > 0 for every k > 0. This means that instead of using A, A2 , A3 , . . . to decide on primitivity, we need only compute B1 = β(A), B2 = β(B1 B1), B3 = β(B1 B2), B4 = β(B1 B3), . . . , going no further than Bn2 −2n+2 , and these computations require only Boolean operations AND and OR. The matrix A in this example is primitive because B1 = 0 0 1 1 0 1 0 1 0 , B2 = 0 1 0 0 1 1 1 0 1 , B3 = 1 0 1 1 1 1 0 1 1 , B4 = 0 1 1 1 1 1 1 1 1 , B5 = 1 1 1 1 1 1 1 1 1 . The powers of an irreducible matrix A ≥ 0 can tell us if A has more than one eigenvalue on its spectral circle, but the powers of A provide no clue to the number of such eigenvalues. The next theorem shows how the index of imprimitivity can be determined without explicitly calculating the eigenvalues. Index of Imprimitivity If c(x) = xn + ck1 xn−k1 + ck2 xn−k2 + · · · + cks xn−ks = 0 is the characteristic equation of an imprimitive matrix An×n in which only the terms with nonzero coefficients are listed (i.e., each ckj = 0, and n > (n − k1) > · · · > (n − ks)), then the index of imprimitivity h is the greatest common divisor of {k1 , k2 , . . . , ks }. 680 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Proof. We know from (8.3.15) that if {λ1 , λ2 , . . . , λn } are the eigenvalues of A (including multiplicities), then {ωλ1 , ωλ2 , . . . , ωλn } are also the eigenvalues of A, where ω = e2πi/h . It follows from the results on p. 494 that λi1 · · · λikj = (−1)kj ωλi1 · · · ωλikj = ω kj ckj =⇒ ω kj = 1. ckj = (−1)kj 1≤i1 1 because E ≥ 0, E = 0, p > 0 =⇒ Ep > 0. (Conditions weaker than the column-sum condition can also force ρ (A) < 1—see Example 7.10.3 on p. 620.) The assumption that A is a nonnegative irreducible matrix whose spectral radius is ρ (A) < 1 combined with the Neumann series (p. 618) provides the conclusion that (I − A)−1 = ∞ Ak > 0. k=0 Positivity is guaranteed by the irreducibility of A because the same argument given on p. 672 that is to prove (8.3.5) also applies here. Therefore, for each demand vector d ≥ 0, there exists a unique supply vector given by s = (I − A)−1 d, which is necessarily positive. The fact that (I − A)−1 > 0 and s > 0 leads to the interesting conclusion that an increase in public demand by just one unit from a single industry will force an increase in the output of all industries. Note: The matrix I − A is an M-matrix as defined and discussed in Example 7.10.7 (p. 626). The realization that M-matrices are naturally present in economic models provided some of the motivation for studying M-matrices during the first half of the twentieth century. Some of the M-matrix properties listed on p. 626 were independently discovered and formulated in economic terms. 8.3 Nonnegative Matrices 683 Example 8.3.7 Leslie Population Age Distribution Model. Divide a population of females into age groups G1 , G2 , . . . , Gn , where each group covers the same number of years. For example, G1 = all females under age 10, G2 = all females from age 10 up to 20, G1 = all females from age 20 up to 30, .. . Consider discrete points in time, say t = 0, 1, 2, . . . years, and let bk and sk denote the birth rate and survival rate for females in Gk . That is, let bk = Expected number of daughters produced by a female in Gk , sk = Proportion of females in Gk at time t that are in Gk+1 at time t + 1. If fk (t) = Number of females in Gk at time t, then it follows that f1 (t + 1) = f1 (t)b1 + f2 (t)b2 + · · · + fn (t)bn and (8.3.17) fk (t + 1) = fk−1 (t)sk−1 for k = 2, 3, . . . , n. Furthermore, Fk (t) = fk (t) = % of population in Gk at time t. f1 (t) + f2 (t) + · · · + fn (t) The vector F(t) = (F1 (t), F2 (t), . . . , Fn (t))T represents the population age distribution at time t, and, provided that it exists, F = limt→∞ F(t) is the long-run (or steady-state) age distribution. Problem: Assuming that s1 , . . . , sn and b2 , . . . , bn are positive, explain why the population age distribution approaches a steady state, and then describe it. In other words, show that F = limt→∞ F(t) exists, and determine its value. Solution: The equations in (8.3.17) constitute a system of ence equations that can be written in matrix form as ⎛ b1 b2 · · · bn−1 ⎜ s1 0 · · · · · · ⎜ 0 s2 0 f (t + 1) = Lf (t), where L = ⎜ ⎜ .. .. ⎝ .. .. 0 0 · · · sn homogeneous differ⎞ bn 0 ⎟ ⎜0 ⎟ .. ⎟ .. ⎟ ⎠ 0 n×n (8.3.18) 684 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices The matrix L is called the Leslie matrix in honor of P. H. Leslie who used this model in 1945. Notice that in addition to being nonnegative, L is also irreducible when s1 , . . . , sn and b2 , . . . , bn are positive because the graph G(L) is strongly connected. Moreover, L is primitive. This is obvious if in addition to s1 , . . . , sn and b2 , . . . , bn being positive we have b1 > 0 (recall Example 8.3.3 on p. 678). But even if b1 = 0, L is still primitive because Ln+2 > 0 (recall (8.3.16) on p. 678). The technique on p. 679 also can be used to show primitivity (Exercise 8.3.11). Consequently, (8.3.10) on p. 674 guarantees that lim t→∞ L t r =G= pqT > 0, qT p where p > 0 and q > 0 are the respective Perron vectors for L and LT . If we combine this with the fact that the solution to the system of difference equations in (8.3.18) is f (t) = Lt f (0) (p. 617), and if we assume that f (0) = 0, then we arrive at the conclusion that T f (t) f (t) qT f (0) = q f (0) > 0 (8.3.19) lim t = Gf (0) = p and lim t→∞ r t→∞ r t qT p qT p 1 (because 1 is a continuous function—Exercise 5.1.7 on p. 277). Now fk (t) = % of population that is in Gk at time t f (t)1 Fk (t) = is the quantity of interest, and (8.3.19) allows us to conclude that f (t) f (t)/rt = lim t→∞ f (t) t→∞ f (t) /rt 1 1 F = lim F(t) = lim t→∞ = limt→∞ f (t)/rt = p (the Perron vector!). limt→∞ f (t)1 /rt In other words, while the numbers in the various age groups may increase or decrease, depending on the value of r (Exercise 8.3.10), the proportion of individuals in each age group becomes stable as time increases. And because the steady-state age distribution is given by the Perron vector of L, each age group must eventually contain a positive fraction of the population. Exercises for section 8.3 0 1 0 8.3.1. Let A = 3 0 0 2 3 0 . (a) Show that A is irreducible. (b) Find the Perron root and Perron vector for A. (c) Find the number of eigenvalues on the spectral circle of A. 8.3 Nonnegative Matrices 685 8.3.2. Suppose that the index of imprimitivity of a 5 × 5 nonnegative irreducible matrix A is h = 3. Explain why A must be singular with alg multA (0) = 2. 8.3.3. Suppose that A is a nonnegative matrix that possesses a positive spectral radius and a corresponding positive eigenvector. Does this force A to be irreducible? 8.3.4. Without computing the eigenvalues or the characteristic polynomial, explain why σ (Pn) = {1, ω, ω 2 , . . . , ω n−1 }, where ω = e2πi/n for ⎞⎛ 0 0 1 1 0 . . . 0 0 . ⋯ 0 1 0 0 9 0 2 0 0 2 0 0 7 0 0 1⎟ 0. Pn = ⎜ ⎝ .. ⎠⎝0 8.3.5. Determine whether A = ⎛2 0 0 ⋯ ⋯ .. . 0 ⋯ 0 1 .. 0⎟0 0⎠4 0 0 0 ⎜. ⎟. . ⎠ 1 0 is reducible or irreducible. 8.3.6. Determine whether the matrix A in Exercise 8.3.5 is primitive or imprimitive. 8.3.7. A matrix Sn×n ≥ 0 having row sums less than or equal to 1 with at least one row sum less than 1 is called a substochastic matrix. (a) Explain why ρ (S) ≤ 1 for every substochastic matrix. (b) Prove that ρ (S) < 1 for every irreducible substochastic matrix. 8.3.8. A nonnegative matrix for which each row sum is 1 is called a stochastic matrix (some say row -stochastic). Prove that if An×n is nonnegative and irreducible with r = ρ (A) , then A is similar ⎛ to rP for some ⎞irp1 ⎜0 reducible stochastic matrix P. Hint: Consider D = ⎝ .. 0 0 p2 .. 0 ⋯ ⋯ .. .. ⋯ 0 0 ⎟.. ⎠, . pn where the pk 's are the components of the Perron vector for A. ⎛ 8.3.9. Wielandt constructed the matrix Wn = that Wn −2n+2 > 0, but [Wn for n = 4. 2 2 −2n+1 0⎟ 0⎟. ⎝. 0 1 1 0 . 0 1 0 1 .. . ⋯ 0 ⋯ ⋯ .. . ⋯ ⎞0 0 ⎟.. ⎠1 0 to show]11 = 0. Verify that this is true 686 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices 8.3.10. In the Leslie population model on p. 683, explain what happens to the vector f (t) as t → ∞ depending on whether r < 1, r = 1, or r > 1. 8.3.11. Use the characteristic equation as described on p. 679 to show that the Leslie matrix in (8.3.18) is primitive even if b1 = 0 (assuming all other bk 's and sk 's are positive). 8.3.12. A matrix A ∈ n×n is said to be essentially positive if A is irreducible and aij ≥ 0 for every i = j. Prove that each of the following statements is equivalent to saying that A is essentially positive. (a) There exists some α ∈ such that A + αI is primitive. (b) etA > 0 for all t > 0. 8.3.13. Let A be an essentially positive matrix as defined in Exercise 8.3.12. Prove that each of the following statements is true. (a) A has an eigenpair (ξ, x), where ξ is real and x > 0. (b) If λ is any eigenvalue for A other than ξ, then Re (λ) < ξ. (c) ξ increases when any entry in A is increased. (k) 8.3.14. Let A ≥ 0 be an irreducible matrix, and let aij denote entries in Ak . Prove that A is primitive if and only if 1/k (k) ρ (A) = lim aij . k→∞ 8.4 Stochastic Matrices and Markov Chains 8.4 687 STOCHASTIC MATRICES AND MARKOV CHAINS One of the most elegant applications of the Perron–Frobenius theory is the algebraic development of the theory of finite Markov chains. The purpose of this section is to present some of the aspects of this development. A stochastic matrix is a nonnegative matrix Pn×n in which each row sum is equal to 1. Some authors say "row -stochastic" to distinguish this from the case when each column sum is 1. 92 A Markov chain is a stochastic process (a set of random variables {Xt }∞ t=0 in which Xt has the same range {S1 , S2 , . . . , Sn }, called the state space) that satisfies the Markov property P (Xt+1 = Sj | Xt = Sit , Xt−1 = Sit−1 , . . . , X0 = Si0) = P (Xt+1 = Sj | Xt = Sit) for each t = 0, 1, 2, ... Think of a Markov chain as a random chain of events that occur at discrete points t = 0, 1, 2, . . . in time, where Xt represents the state of the event that occurs at time t. For example, if a mouse moves randomly through a maze consisting of chambers S1 , S2 , . . . , Sn , then Xt might represent the chamber occupied by the mouse at time t. The Markov property asserts that the process is memoryless in the sense that the state of the chain at the next time period depends only on the current state and not on the past history of the chain. In other words, the mouse moving through the maze obeys the Markov property if its next move doesn't depend on where in the maze it has been in the past—i.e., the mouse is not using its memory (if it has one). To emphasize that time is considered discretely rather than continuously the phrase "discrete-time Markov chain" is often used, and the phrase "finite-state Markov chain" might be used to emphasize that the state space is finite rather than infinite. 92 Andrei Andreyevich Markov (1856–1922) was born in Ryazan, Russia, and he graduated from Saint Petersburg University in 1878 where he later became a professor. Markov's early interest was number theory because this was the area of his famous teacher Pafnuty Lvovich Chebyshev (1821–1894). But when Markov discovered that he could apply his knowledge of continued fractions to probability theory, he embarked on a new course that would make him famous—enough so that there was a lunar crater named in his honor in 1964. In addition to being involved with liberal political movements (he once refused to be decorated by the Russian Czar), Markov enjoyed poetry, and in his spare time he studied poetic style. Therefore, it was no accident that led him to analyze the distribution of vowels and consonants in Pushkin's work, Eugene Onegin, by constructing a simple model based on the assumption that the probability that a consonant occurs at a given position in any word should depend only on whether the preceding letter is a vowel or a consonant and not on any prior history. This was the birth of the "Markov chain." Markov was wrong in one regard—he apparently believed that the only real examples of his chains were to be found in literary texts. But Markov's work in 1907 has grown to become an indispensable tool of enormous power. It launched the theory of stochastic processes that is now the foundation for understanding, explaining, and predicting phenomena in diverse areas such as atomic physics, quantum theory, biology, genetics, social behavior, economics, and finance. Markov's chains serve to underscore the point that the long-term applicability of mathematical research is impossible to predict. 688 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Every Markov chain defines a stochastic matrix, and conversely. Let's see how this happens. The value pij (t) = P (Xt = Sj | Xt−1 = Si) is the probability of being in state Sj at time t given that the chain is in state Si at time t − 1, so pij (t) is called the transition probability of moving from Si to Sj at time t. The matrix of transition probabilities Pn×n (t) = [pij (t)] is clearly a nonnegative matrix, and a little thought should convince you that each row sum must be 1. Thus P(t) is a stochastic matrix. When the transition probabilities don't vary with time (say pij (t) = pij for all t), the chain is said to be stationary (or homogeneous), and the transition matrix is the constant stochastic matrix P = [pij]. We will make the assumption of stationarity throughout. Conversely, every stochastic matrix Pn×n defines an n -state Markov chain because the entries pij define a set of transition probabilities, which can be interpreted as a stationary Markov chain on n states. A probability distribution vector is defined to be a nonnegative vector pT = (p1 , p2 , . . . , pn) such that k pk = 1. (Every row in a stochastic matrix is such a vector.) For an n -state Markov chain, the k th step probability distribution vector is defined to be pT (k) = p1 (k), p2 (k), . . . , pn (k) , k = 1, 2, . . . , where pj (k) = P (Xk = Sj). In other words, pj (k) is the probability of being in the j th state after the k th step, but before the (k + 1)st step. The initial distribution vector is pT (0) = p1 (0), p2 (0), . . . , pn (0) , where pj (0) = P (X0 = Sj) is the probability that the chain starts in Sj . For example, consider the Markov chain defined by placing a mouse in the 3-chamber box with connecting doors as shown in Figure 8.4.1, and suppose that the mouse moves from the chamber it occupies to another chamber by picking a door at random—say that the doors open each minute, and when they do, the mouse is forced to move by electrifying the floor of the occupied chamber. #2 #1 #3 Figure 8.4.1 If the mouse is initially placed in chamber #2, then the initial distribution vector is pT (0) = (0, 1, 0) = eT2 . But if the process is started by tossing the mouse into the air so that it randomly lands in one of the chambers, then a reasonable 8.4 Stochastic Matrices and Markov Chains 689 initial distribution is pT (0) = (.5, .25, .25) because the area of chamber #1 is 50% of the box, while chambers #2 and #3 each constitute 25% of the box. The transition matrix for this Markov chain is the stochastic matrix ⎛ 0 M = ⎝ 1/3 1/3 1/2 0 2/3 ⎞1/2 2/3 ⎠. 0 (8.4.1) A standard eigenvalue calculation reveals that σ (M) = {1, −1/3, /, −2/3}, so it's apparent that M is a nonnegative matrix having spectral radius ρ (M) = 1. This is a feature that is shared by all stochastic matrices Pn×n because having row sums equal to 1 means that P∞ = 1 or, equivalently, Pe = e, where e is the column of all 1's. Because (1, e) is an eigenpair for every stochastic matrix, and because ρ ( ) ≤ for every matrix norm (recall (7.1.12) on p. 497), it follows that 1 ≤ ρ (P) ≤ P∞ = 1 =⇒ ρ (P) = 1. Furthermore, e is a positive eigenvector associated with ρ (P) = 1. But be careful! This doesn't mean that you necessarily can call e the Perron vector for .5 P because P might not be irreducible—consider P = .5 . 0 1 Two important issues that arise in Markovian analysis concern the transient behavior of the chain as well as the limiting behavior. In other words, we want to accomplish the following goals. • Describe the k th step distribution pT (k) for any given initial distribution vector pT (0). • Determine whether or not limk→∞ pT (k) exists, and if it exists, determine the value of limk→∞ pT (k). • If there is no limiting distribution, then determine the possibility of having a Ces`aro limit lim k→∞ pT (0) + pT (1) + · · · + pT (k − 1) . k If such a limit exists, interpret its meaning, and determine its value. The k th step distribution is easily described by using the laws of elementary probability—in particular, recall that P (E ∨ F) = P (E) + P (F) when E and F are mutually exclusive events, and the conditional probability of E occurring given that F occurs is P (E | F) = P (E ∧ F)/P (F) (it's convenient to use ∧ and ∨ to denote AND and OR, respectively). To determine the j th component 690 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices pj (1) in pT (1) for a given pT (0), write pj (1) = P (X1=Sj) = P X1=Sj ∧ (X0=S1 ∨ X0=S2 ∨ · · · ∨ X0=Sn) = P (X1=Sj ∧ X0=S1) ∨ (X1=Sj ∧ X0=S2) ∨ · · · ∨ (X1=Sj ∧ X0=Sn) = n n P X1=Sj ∧ X0=Si = P X0 = Si P X1 = Sj | X0 = Si i=1 = n i=1 pi (0)pij for j = 1, 2, . . . , n. i=1 Consequently, pT (1) = pT (0)P. This tells us what to expect after one step when we start with pT (0). But the "no memory" Markov property tells us that the state of affairs at the end of two steps is determined by where we are at the end of the first step—it's like starting over but with pT (1) as the initial distribution. In other words, it follows that pT (2) = pT (1)P, and pT (3) = pT (2)P, etc. Therefore, successive substitution yields pT (k) = pT (k − 1)P = pT (k − 2)P2 = · · · = pT (0)Pk , and thus the k th step distribution is determined from the initial distribution and the transition matrix by the vector–matrix product pT (k) = pT (0)Pk . (8.4.2) (k) Notice that if we adopt the notation Pk = pij , and if we set pT (0) = eTi in (8.4.2), then we get pj (k) = pij the following conclusion. • for each i = 1, 2, . . . , n, and thus we arrive at The (i, j)-entry in Pk represents the probability of moving from Si to Sj in exactly k steps. For this reason, Pk is often called the k-step transition matrix. Example 8.4.1 Let's go back to the mouse-in-the-box example, and, as suggested earlier, toss the mouse into the air so that it randomly lands somewhere in the box in Figure 8.4.1—i.e., take the initial distribution to be pT (0) = (1/2, 1/4, 1/4). The transition matrix is given by (8.4.1), so the probability of finding the mouse in chamber #1 after three moves is [pT (3)]1 = [pT (0)M3]1 = 13/54. In fact, the entire third step distribution is pT (3) = (13/54, 41/108, 41/108) . 8.4 Stochastic Matrices and Markov Chains 691 To analyze limiting properties of Markov chains, divide the class of stochastic matrices (and hence the class of stationary Markov chains) into four mutually exclusive categories as described below. (1) Irreducible with limk→∞ Pk (2) Irreducible with limk→∞ Pk (3) Reducible with limk→∞ Pk (4) Reducible with limk→∞ Pk existing (i.e., P is primitive). not existing (i.e., P is imprimitive). existing. not existing. In case (1), where P is primitive, we know exactly what limk→∞ Pk looks like. The Perron vector for P is e/n (the uniform distribution vector), so if π = (π1 , π2 , . . . , πn)T is the Perron vector for PT , then ⎞⎛ π 1 π 2 · · · πn ⎜ π1 π 2 · · · π n ⎜(e/n)πT eπT lim Pk = T >0 (8.4.3) = T = eπT = ⎜ .. .. ⎟⎜ ... k→∞ π (e/n) π e . . ⎠π1 π2 · · · πn by (8.3.10) on p. 674. Therefore, if P is primitive, then a limiting probability distribution exists, and it is given by lim pT (k) = lim pT (0)Pk = pT (0)eπT = πT . k→∞ k→∞ (8.4.4) T Notice that because k pk (0) = 1, the term p (0)e drops away, so we have the conclusion that the value of the limit is independent of the value of the initial distribution pT (0), which isn't too surprising. Example 8.4.2 Going back to the mouse-in-the-box example, it's easy to confirm that the transition matrix M in (8.4.1) is primitive, so limk→∞ Mk as well as limk→∞ pT (k) must exist, and their values are determined by the left-hand Perron vector of M that can be found by calculating any nonzero vector v ∈ N I − MT and normalizing it to produce πT = vT / v1 . Routine computation reveals that the one solution of the homogeneous equation (I − MT)v = 0 is vT = (2, 3, 3), so πT = (1/8)(2, 3, 3), and thus ⎛ ⎞2 3 3 1 1 lim Mk = ⎝ 2 3 3 ⎠and lim pT (k) = (2, 3, 3). k→∞ k→∞ 8 8 2 3 3 This limiting distribution can be interpreted as meaning that in the long run the mouse will occupy chamber #1 one-fourth of the time, while 37.5% of the time it's in chamber #2, and 37.5% of the time it's in chamber #3, and this is independent of where (or how) the process started. The mathematical justification for this statement is on p. 693. 692 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Now consider the imprimitive case. We know that if P is irreducible and has h > 1 eigenvalues on the unit (spectral) circle, then limk→∞ Pk cannot exist (p. 674), and hence limk→∞ pT (k) cannot exist (otherwise taking pT (0) = eTi for each i would insure that Pk has a limit). However, each eigenvalue on the unit circle is simple (p. 676), and this means that P is Ces`aro summable (p. 633). Moreover, e/n is the Perron vector for P, and, as pointed out in Example 8.3.2 (p. 677), if πT = (π1 , π2 , . . . , πn) is the left-hand Perron vector, then ⎞⎛ π 1 π2 · · · π n ⎜ π 1 π2 · · · π n ⎟I + P + · · · + Pk−1 (e/n)πT eπT lim , = T = eπT = ⎜ .. .. ⎟⎜ ... k→∞ k π (e/n) π e . . ⎠π 1 π2 · · · π n which is exactly the same form as the limit (8.4.3) for the primitive case. Consequently, the k th step distributions have a Ces` aro limit given by T T T p (0) + p (1) + · · · + p (k − 1) I + P + · · · + Pk−1 T lim = lim p (0) k→∞ k→∞ k k = pT (0)eπT = πT , and, just as in the primitive case (8.4.4), this Ces` aro limit is independent of the initial distribution. Let's interpret the meaning of this Ces` aro limit. The analysis is essentially the same as the description outlined in the shell game in Example 7.10.8 (p. 635), but for the sake of completeness we will duplicate some of the logic here. The trick is to focus on one state, say Sj , and define a sequence of random variables {Zk }∞ k=0 that count the number of visits to Sj . Let 1 if the chain starts in Sj , Z0 = 0 otherwise, and for i > 1, (8.4.5) th Zi = 1 if the chain is in Sj after the i move, 0 otherwise. Notice that Z0 + Z1 + · · · + Zk−1 counts the number of visits to Sj before the k th move, so (Z0 + Z1 + · · · + Zk−1)/k represents the fraction of times that Sj is hit before the k th move. The expected (or mean) value of each Zi is E[Zi] = 1 · P (Zi=1) + 0 · P (Zi=0) = P (Zi=1) = pj (i), and, since expectation is linear, the expected fraction of times that Sj is hit before move k is Z0 + Z1 + · · · + Zk−1 E[Z0] + E[Z1] + · · · + E[Zk−1] E = k k T pj (0) + pj (1) + · · · + pj (k − 1) p (0) + pT (1) + · · · + pT (k − 1) = = k k j → πj . 8.4 Stochastic Matrices and Markov Chains 693 In other words, the long-run fraction of time that the chain spends in Sj is πj , which is the j th component of the Ces`aro limit or, equivalently, the j th component of the left-hand Perron vector for P. When limk→∞ pT (k) exists, it must be the case that T p (0)+pT (1)+· · ·+pT (k−1) T lim p (k) = lim (Exercise 7.10.11, p. 639), k→∞ k→∞ k and therefore the interpretation of the limiting distribution limk→∞ pT (k) for the primitive case is exactly the same as the interpretation of the Ces`aro limit in the imprimitive case. Below is a summary of our findings for irreducible chains. Irreducible Markov Chains Let P be the transition probability matrix for an irreducible Markov chain on states {S1 , S2 , . . . , Sn } (i.e., P is an n × n irreducible stochastic matrix), and let πT denote the left-hand Perron vector for P. The following statements are true for every initial distribution pT (0). • The k th step transition matrix is Pk because the (i, j) -entry in Pk is the probability of moving from Si to Sj in exactly k steps. • The k th step distribution vector is given by pT (k) = pT (0)Pk . • If P is primitive, and if e denotes the column of all 1's, then lim Pk = eπT k→∞ • and lim pT (k) = πT . k→∞ If P is imprimitive, then I + P + · · · + Pk−1 = eπT k→∞ k T p (0)+pT (1)+· · ·+pT (k−1) lim = πT . k→∞ k lim and • Regardless of whether P is primitive or imprimitive, the j th component πj of πT represents the long-run fraction of time that the chain is in Sj . • πT is often called the stationary distribution vector for the chain because it is the unique distribution vector satisfying πT P = πT . 694 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Example 8.4.3 Periodic Chains. Consider an electronic switch that can be in one of three states {S1 , S2 , S3 }, and suppose that the switch changes states on regular clock cycles. If the switch is in either S1 or S3 , then it must change to S2 on the next clock cycle, but if the switch is in S2 , then there is an equal likelihood that it changes to S1 or S3 on the next clock cycle. The transition matrix is ⎛ ⎞0 1 0 P = ⎝ .5 0 .5 ⎠, 0 1 0 and it's not difficult to see that P is irreducible (because G(P) is strongly connected) and imprimitive (because σ (P) = {±1, 0}). Since the left-hand Perron vector is πT = (.25, .5, .25), the long-run expectation is that the switch should be in S1 25% of the time, in S2 50% of the time, and in S3 25% of the time, and this agrees with what common sense tells us. Furthermore, notice that the switch cannot be in just any position at any given clock cycle because if the chain starts in either S1 or S3 , then it must be in S2 on every odd-numbered cycle, and it can occupy S1 or S3 only on even-numbered cycles. The situation is similar, but with reversed parity, when the chain starts in S2 . In other words, the chain is periodic in the sense that the states can be occupied only at periodic points in time. In this example the period of the chain is 2, and this is the same as the index of imprimitivity. This is no accident. The Frobenius form for imprimitive matrices on p. 680 can be used to prove that this is true in general. Consequently, an irreducible Markov chain is said to be a periodic chain when its transition matrix P is imprimitive (with the period of the chain being the index of imprimitivity for P), and an irreducible Markov chain for which P is primitive is called an aperiodic chain. The shell game in Example 7.10.8 (p. 635) is a periodic Markov chain that is similar to the one in this example. Because the Perron–Frobenius theorem is not directly applicable to reducible chains (chains for which P is a reducible matrix), the strategy for analyzing reducible chains is to deflate the situation, as much as possible, back to the irreducible case as described below. If P is reducible, then, by definition, there exists a permutation matrix Q and square matrices X and Z such that Y X Y QT PQ = X . For convenience, denote this by writing P ∼ . 0 Z 0 Z If X or Z is reducible, then another symmetric permutation can be performed to produce R S T X Y ∼ 0 U V , where R, U, and W are square. 0 Z 0 0 W 8.4 Stochastic Matrices and Markov Chains 695 Repeating this process eventually yields ⎛X 11 ⎜ 0 P ∼ ⎜ .. . 0 X12 X22 0 ⋯ ⋯ .. . ⋯ ⎞X1k X2k ⎟.. ⎟, . Xkk where each Xii is irreducible or Xii = [0]1×1 . Finally, if there exist rows having nonzero entries only in diagonal blocks, then symmetrically permute all such rows to the bottom to produce ⎛P ⎜ ⎜ ⎜ ⎜ ⎜ P∼⎜ ⎜ ⎜ ⎜ ⎜ ⎝ 11 0 .. . 0 0 0 . 0 P12 P22 0 0 0 . . 0 ⋯ ⋯ .. . ⋯ P1r P2r .. . Prr P1,r+1 P2,r+1 .. . Pr,r+1 P1,r+2 P2,r+2 .. . Pr,r+2 ⋯ ⋯ .. . ⋯ 0 0 .. . 0 Pr+1,r+1 0 .. . 0 0 ⋯ ⋯ ⋯ Pr+2,r+2 .. . 0 ⋯ ⋯ ⋯ ⋯ ⋯ P1m P2m .. . Prm ⋯ ⋯ ⋯ .. . ⋯ . Pmm ⎞ ⎟ ⎟ ⎟ ⎟ ⎟ ⎟ ⎟0 ⎟0 ⎟ ⎟.. ⎠ (8.4.6) where each P11 , . . . , Prr is either irreducible or [0]1×1 , and Pr+1,r+1 , . . . , Pmm are irreducible (they can't be zero because each has row sums equal to 1). As mentioned on p. 671, the effect of a symmetric permutation is simply to relabel nodes in G(P) or, equivalently, to reorder the states in the chain. When the states of a chain have been reordered so that P assumes the form on the righthand side of (8.4.6), we say that P is in the canonical form for reducible matrices. When P is in canonical form, the subset of states corresponding to Pkk for 1 ≤ k ≤ r is called the k th transient class (because once left, a transient class can't be reentered), and the subset of states corresponding to Pr+j,r+j for j ≥ 1 is called the j th ergodic class. Each ergodic class is an irreducible Markov chain unto itself that is imbedded in the larger reducible chain. From now on, we will assume that the states in our reducible chains have been ordered so that P is in canonical form. The results on p. 676 guarantee that if an irreducible stochastic matrix P has h eigenvalues on the unit circle, then these h eigenvalues are the hth roots of unity, and each is a simple eigenvalue for P. The same can't be said for reducible stochastic matrices, but the canonical form (8.4.6) allows us to prove the next best thing as discussed below. 696 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Unit Eigenvalues The unit eigenvalues for a stochastic matrix are defined to be those eigenvalues that are on the unit circle. For every stochastic matrix Pn×n , the following statements are true. • Every unit eigenvalue of P is semisimple. • Every unit eigenvalue has form λ = e2kπi/h for some k < h ≤ n. • In particular, ρ (P) = 1 is always a semisimple eigenvalue of P. Proof. If P is irreducible, then there is nothing to prove because, as proved on p. 676, the unit eigenvalues are roots of unity, and each unit eigenvalue is simple. If P is reducible, suppose that a symmetric permutation has been performed so that P is in the canonical form (8.4.6), and observe that ρ (Pkk) < 1 for each k = 1, 2, . . . , r. (8.4.7) This is certainly true when Pkk = [0]1×1 , so suppose that Pkk (1 ≤ k ≤ r) is irreducible. Because there must be blocks Pkj , j = k, that have nonzero entries, it follows that Pkk e ≤ e and Pkk e = e, where e is the column of all 1's. If ρ (Pkk) = 1, then the observation in Example 8.3.1 (p. 674) forces Pkk e = e, which is impossible, and thus ρ (Pkk) < 1. Consequently, the unit eigenvalues for P are the collection of the unit eigenvalues of the irreducible matrices Pr+1,r+1 , . . . , Pmm . But each unit eigenvalue of Pr+i,r+i is simple and is a root of unity. Consequently, if λ is a unit eigenvalue for P, then it must be some root of unity, and although it might be repeated because it appears in the spectrum of more than one Pr+i,r+i , it must nevertheless be the case that alg multP (λ) = geo multP (λ) , so λ is a semisimple eigenvalue of P. We know from the discussion on p. 633 that a matrix A ∈ C n×n is Ces`aro summable if and only if ρ(A) < 1 or ρ(A) = 1 with each eigenvalue on the unit circle being semisimple. We just proved that the latter holds for all stochastic matrices P, and so we have in fact established the following powerful statement concerning all stochastic matrices. 8.4 Stochastic Matrices and Markov Chains 697 All Stochastic Matrices Are Summable Every stochastic matrix P is Ces`aro summable. That is, I + P + · · · + Pk−1 k→∞ k lim exists for all stochastic matrices P, and, as discussed on p. 633, the value of the limit is the (spectral) projector G onto N (I − P) along R (I − P). Since we already know the structure and interpretation of the Ces` aro limit when P is an irreducible stochastic matrix (p. 693), all that remains in order to complete the picture is to analyze the nature of limk→∞ (I + P + · · · + Pk−1)/k for the reducible case. 12 Suppose that P = T011 T is a reducible stochastic matrix that is in T22 the canonical form (8.4.6), where ⎞⎛ ⎛ T11 =⎝ P11 ⋯ .. . Prr .. ⎠, . Prr and T12

= T22 = (8.4.8) Pr+1,r+1 .. P1,r+1 · · · P1m .. .. , . . Pr,r+1 · · · Prm . . Pmm We know from (8.4.7) that ρ (Pkk) < 1 for each k = 1, 2, . . . , r, so it follows that ρ (T11) < 1, and
matrices, so if πTj is the left-hand Perron vector for Pjj , r + 1 ≤ j ≤ m, then our previous results (p. 693) tell us that eπ T lim k→∞ I + T22 + · · · + k Tk−1 22 r+1 = .. . eπ T m
= E. 698 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Therefore, the limits, be they Cesàro or ordinary (if it exists), all have the form I + P + · · · + Pk−1 = k→∞ k lim
(I − P)) to write (I − P)G = 0 =⇒ I − T11 0 −T12 I − T22 0 0 Z E = 0 =⇒ (I − T11)Z = T12 E. Since I − T11 is nonsingular (because ρ (T11) < 1 by (8.4.7)), it follows that Z = (I − T1
been ordered to make the transition matrix assume the canonical form T11 T12 P= 0 T22 that is described in (8.4.6) and (8.4.8), and if πTj is the left-hand Perron vector for Pjj (r + 1
exists if and only if the stochastic matrices Pr+1,r+1 , . . . , Pmm in (8.4.6) are each primitive, in which case lim Pk = k→∞ 0 (I − T11)−1 T12 E 0 E . (8.4.10) 8.4 Stochastic Matrices
defined by Pr+j,r+j for some j ≥ 1. If Pr+j,r+j is primitive, then the chain settles down to a steady-state defined by the left-hand Perron vector of Pr+j,r+j , but if Pr+j,r+j is imprimitive
concerning which ergodic class the chain will end up in and how long it takes to get there. This time the answer depends on where the chain starts—i.e., on the initial distribution. For
state of Ti . Since the question at hand concerns only which ergodic class is hit but not what happens after it's entered, we might as well convert every state in each ergodic class into a
limk→∞ P the form 0 0 . .. −1 0) T 0 (I − T 11 12 k lim P = = k→∞ 0 I 0 0 .. . 0 0 0 0 0 0 .. . 0 ··· ··· .. . ··· 0 0 .. . 0 L1,1 L2,1 .. . Lr,1 L1,2 L2,2 .. . Lr,2 ··· ··· ··· ··· 0 0 .. . I 0
q th state in Ej given that we start from the pth state in Ti . Therefore, if e is the vector of all 1 's, then the probability of eventually entering somewhere in Ej is given by • P (absorptio
pTi (0)Lij e. To determine the expected number of steps required to first hit an ergodic state, proceed as follows. Count the number of times the chain is in transient state Sj given that
step k, and Zk = Z0 = 0 otherwise, 0 otherwise. 700 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices Since E[Zk] = 1 · P (Zk =1) + 0 · P (Zk =0) = P (Zk =1) = Tk11 ij ,
k=0 (because ρ (T11) < 1). Summing this over all transient states produces the expected number of times the chain is in some transient state, which is the same as the expected numb
Markov Chains. It's often the case in practical applications that there is only one transient class, and the ergodic classes are just single absorbing states (states such that once they are
··· p1r .. .. . . pr1 · · · prr P= 0 ··· 0 . .. . . 0 ··· . 0 p1,r+1 .. . pr,r+1 ··· 1 .. . ··· .. . ··· 0 ··· p1s .. . prs 0 .. and Lij = (I − T11)−1 T12 ij . . 1 The preceding analysis spec
Sr+j) given that the chain starts in the ith transient state (which is Si) is P (absorption into Sr+j | start in Si for 1 ≤ i ≤ r) = (I − T11)−1 T12 ij , while the expected time until absorpt
Matrices and Markov Chains 701 Example 8.4.5 Fail-Safe System. Consider a system that has two independent controls, A and B, that can prevent the system from being destroyed. Th
The system is destroyed if A and B fail simultaneously. 5 For example, an automobile has two independent braking systems—one is operated by a foot pedal, whereas the "emergency b
systems fail simultaneously. If one of the controls fails at some activation point but the other control holds, then the defective control is repaired before the next activation. If a control
reliable at t = tk+1 . Problem: Can the system be expected to run indefinitely without every being destroyed? If not, how long is the system expected to run before destruction occurs?
pairs (a, b) in which 1 if A holds, 1 if B holds, a= and b = 0 if A fails, 0 if B fails. State (0, 0) is absorbing, and the transition matrix (in canonical form) is (1, 1) (1, 1) .81 (1, 0) .54 P=
fact that limk→∞ Pk exists and is given by 0 (I − T11)−1 T12 k lim P = k→∞ 0 1 702 Chapter 8 Perron–Frobenius Theory of Nonnegative Matrices makes it clear that the absorbing st
significant figures produces 44.6 6.92 6.92 58.4 (I − T11)−1 = 41.5 8.02 6.59 and (I − T11)−1 e = 56.1 , 41.5 6.59 8.02 56.1 so the mean time to failure starting with two p
difference here doesn't seem significant, but consider what happens when only one control is used in the system. In this case, there are only two states in the chain, 1 (meaning that th
interesting to consider what happens when three independent control are used. How much more security does your intuition tell you that you should have? See Exercise 8.4.8. Exercise
limiting distribution in the regular sense or only in the Cesàro sense? 8.4.2. A doubly-stochastic matrix is a nonnegative matrix Pn×n having all row sums as well as all column sums e
do limk→∞ (I + P + · · · + Pk−1)/k and limk→∞ Pk (if it exists) have? Note: The purpose of this exercise is to show that doubly-stochastic matrices are not very interesting from a Mark

b 0 ‖ 0 ‖ . ‖ . |(..)(..)0 2n×1 0 2n×1 ˆ = q, and we know from (5.8.11) that the γk 's are ˆ = p and F2n b then F2n a given by γk = [a 3 b]k . Therefore, the convolution theorem guarantees (p(1)q(1))(γ0 | p(ξ)q(ξ) || γ1 | ˆ = F−1 p × q = F−1 | |= a 3 b = F−1 (F2n a |ˆ) × (F b) 2n 2n 2n 2n (p(ξ 2)q(ξ 2)). (γ2). . . . . 5.8.12. (a) This follows from the observation that Qk has 1's on the k th subdiagonal and 1's on the (n − k)th superdiagonal. For example, if n = 8, then (0 0 0 0 0 1 0 0 |0 0 0 0 0 0 1 0 || 0 0 0 0 0 0 0 1 || 1 0 0 0 0 0 0 0 |Q3 = | 0 1 0 0 0 0 0 0 || 0 0 1 0 0 0 0 0 |(0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 Solutions 75 (b) If the rows of F are indexed from 0 to n − 1, then they satisfy the relationships Fk∗ Q = ξ k Fk∗ for each k (verifying this for n = 4 will indicate why it is true in general). This means that FQ = DF, which in turn implies FQF−1 = D. (c) Couple parts (a) and (b) with FQk F−1 = (FQF−1)k = Dk to write FCF−1 = Fp(Q)F−1 = F(c0 I + c1 Q + · · · + cn−1 Qn−1)F−1 = c0 I + c1 FQF−1 + · · · + cn−1 FQn−1 F−1 = c0 I + c1 D + · · · + cn−1 Dn−1 (p(1))0 · · · 0 | 0 |= | 0 p(ξ) · · · 0 | |. .. . . .. | . . . 0 · · · p(ξ n−1) (d) FC1 F−1 = D1 and FC2 F−1 = D2 , where D1 and D2 are diagonal matrices, and therefore C1 C2 = F−1 D1 FF−1 D2 F = F−1 D1 D2 F = F−1 D2 D1 F = F−1 D2 FF−1 D1 F = C2 C1 . 5.8.13. (a) According to Exercise 5.8.12, (σ0 | σ1 C=| (... σn−1 σ0 .. . σn−1 σn−2 (p(1))0 · · · 0 σ1 p(ξ) · · · 0 σ2 | 0 | F = F−1 DF = F−1 |). . . . . σ0 0 0 p(ξ n−1) ··· ··· ... in which p(x) = σ0 +σ1 x+· · ·+σn−1 xn−1 . Therefore, x = C−1 b = F−1 D−1 Fb, so we can execute the following computations. (p(0))(σ0 | p(ξ) | σ || ←− F | .1 |using the FFT (i) .. (..). σn−1 (ii) x ←− Fb using the FFT (iii) xk ←− xk /p(ξ k) (iv) −1 x ←− F x for k = 0, 1, . . . , n − 1 using the FFT as described in Example 5.8.2 76 Solutions (b) Use the same techniques described in part (a) to compute the k th column of C−1 from the formula [C−1]∗k = C−1 ek = F−1 D−1 Fek = F−1 D−1 [F]∗k (1/p(1) ξ k /p(ξ) | || |2k −1 | ξ /p(ξ 2) |. = F−1 | ..). n−k n−1 ξ /p(ξ The k th column of P = C1 C2 is given by (c) P∗k = Pek = F−1 D1 FF−1 D2 Fek = F−1 D1 D2 [F]∗k . If (σ0 σ1 · · · σn−1) and (η0 η1 · · · ηn−1) are the first rows in C1 n−1 n−1 C2 , respectively, and if p(x) = k=0 σk xk and q(x) = k=0 ηk xk , then compute (p(0)) (q(0)) (η0 σ0 | p(ξ) || q(ξ) || σ1 || η1 | ←− F | . p=| and q = | .. | . (p(ξ n−1) q(ξ n−1) σn−1 and first) ηn−1 The k th column of the product can now be obtained from P∗k ←− F−1 p × q × F∗k for k = 0, 1, . . . , n − 1. 5.8.14. (a) For n = 3 (α0 | α1 | α ˆ=| Cb | 2 || 0 (0 0 we have 0 α0 α1 α2 0 0 0 0 α0 α1 α2 0 0 0 α0 α1)(β0 α0 β0 α1 | β1 | | α1 β0 + α0 β1 | | || 0 || β2 | | α2 β0 + α1 β1 + α0 β2 | | |=| . | 0 | 0 || α2 β1 + α1 β2 | | 0 0 α2 β2 0 α0 0 ˆ where η is arbitrary. Use this as a model to write the expression for Cb, (b) We know from part (c) of Exercise 5.8.12 that if F is the Fourier matrix of order 2n, then FCF−1 = D, where (p(1))0 · · · 0 D=| (.. 0 p(ξ) · · · . . . 0 0 .. . · · p(ξ 2n−1) | |) (the ξ k 's are the 2nth roots of unity) Solutions 77 in which p(x) = α0 + α1 x + · · · + αn−1 xn−1 . Therefore, from part (a), ˆ = FCF−1 Fb ˆ = DFb. ˆ F(a 3 b) = FCb According to Exercise 5.8.10, we also know that (p(1) p(ξ) . . | Fˆ a=| (|) p(ξ 2n−1) and hence ˆ = (Fˆ F(a 3 b) = DFb a) × (Fb). 5.8.15. (a) Pn x performs an even–odd permutation to all components of x. The matrix (I2 ⊗ Pn/2) = Pn/2 0 0 Pn/2 x performs an even–odd permutation to the top half of x and then does the same to the bottom half of x. The matrix (Pn/4 | 0 (I4 ⊗ Pn/4) = (0 0 0 Pn/4 0 0 0 0 Pn/4 0 0 0) x 0 Pn/4 performs an even–odd permutation to each individual quarter of x. As this pattern is continued, the product Rn = (I2r−1 ⊗ P21)(I2r−2 ⊗ P22) · · · (I21 ⊗ P2r−1)(I20 ⊗ P2r)x produces the bit-reversing permutation. For example, when n = 8, 78 Solutions R8 x = (I4 ⊗ P2)(I2 ⊗ P4)(I1 ⊗ P8)x (P2 | 0 =(0 0 0 P2 0 0 0 0 P2 0 0 0 |P4)0 P2 0 P2 0 0 0 0 P2 0 0 (x0 | x1 || x2 || |0 |x |P8| 3 |P4| x4 || x5 |(x6 x7 (x0| x2 || x |4 |x |0| | 6 |P4| | x1 || x5 || 3 |(x5)x7) (x0 x0| x4 x4 | | |) | x2 || 0 x2 || || x6 || x6 |0 | | |) |=| 0 x| 1 x| 1 || |P2| x5 || x5 || (x)(x3)3 x7 x7 (P2 | 0 =(0 0 0 P2 0 0 0 0 P2 0 because P2 = 1 0 0 1 . (b) To prove that I2r−k ⊗ F2k = L2k R2k using induction, note first that for k = 1 we have L2 = (I2r−1 ⊗ B2)1 = I2r−1 ⊗ F2 and R2 = In (I2r−1 ⊗ P2) = In In = In, so L2 R2 = I2r−1 ⊗F2 . Now assume that the result holds for k = j—i.e., assume I2r−j ⊗ F2j = L2j R2j . Prove that the result is true for k = j + 1—i.e., prove I2r−(j+1) ⊗ F2j+1 = L2j+1 R2j+1 . Use the fact that F2j+1 = B2j+1 (I2 ⊗ Fj)P2j+1 along with the two basic prop- Solutions 79 erties of the tensor product given in the introduction of this exercise to write I2r−(j+1) ⊗ F2j+1 = I2r−(j+1) ⊗ B2j+1 (I2 ⊗ F2j) I2r−(j+1) ⊗ P2j+1 = (I2r−(j+1) ⊗ B2j+1)(I2r−(j+1) ⊗ I2 ⊗ F2j)(I2r−(j+1) ⊗ P2j+1) = (I2r−(j+1) ⊗ B2j+1)(I2r−j ⊗ F2j)(I2r−(j+1) ⊗ P2j+1) = (I2r−(j+1) ⊗ B2j+1)L2j R2j (I2r−(j+1) ⊗ P2j+1) = L2j+1 R2j+1 . Therefore, I2r−k ⊗ F2k = L2k R2k for k = 1, 2, . . . , r, and when k = r we have that Fn = Ln Rn . 5.8.16. According to Exercise 5.8.10, (Fn a = b, where | | a=| (α0 α1 α2 ..) | | | |) (p(1) and | p(ξ) | p(ξ 2) b=| (.. αn−1 | | (| p(ξ n−1) √ By making use of the fact that (1/ n)Fn is unitary we can write n−1 n−1 k 2 2 p(ξ) = b∗ b = (Fn a)∗ (Fn a) = a∗ F∗n Fn a = a∗ (nI)a = n |αk | . k=0 k=0 5.8.17. Let y = (2/n)Fx, and use the result in (5.8.7) to write , , , , 'y' = , (αk − iβk)efk + (αk + iβk)en−fk , , , k = |αk − iβk |2 + |αk + iβk |2 2 k =2 αk2 + βk2 . k But because F∗ F = nI, it follows that , 2 , 2 , 4 4 2 , 'y' = , Fx, = 2 x∗ F∗ Fx = 'x' , n n n 2 so combining these two statements produces the desired conclusion. 80 Solutions 5.8.18. We know from (5.8.11) that if p(x) = p2 (x) = n−1 k=0 2n−2 αk xk , then [a 3 a]k xk . k=0 The last component of a 3 a is zero, so we can write c (a 3 a) = T 2n−2 [a 3 a]k η = p (η) = k=0 2 n−1 2 αk η k 2 ˆ . = cT a k=0 5.8.19. Start with X ←− rev(x) = (x0 x4 x2 x6 x1 x5 x3 x7). For j = 0 : D ←− (1) X(0) ←− (x0 x2 x1 x3) X(1) ←− (x4 x6 x5 x7) (0) X + D × X(1) X ←− X(0) − D × X(1) x0 + x4 x2 + x6 x1 + x5 x3 + x7 = x0 − x4 x2 − x6 x1 − x5 x3 − x7 2×8 For j = 1 : 1 1 D ←− = e−πi/2 ξ2 x0 + x4 x1 + x5 (0) X ←− x2 + x6 x3 + x7 (1) X ←− x2 − x6 x3 − x7 (0) X + D × X(1) X ←− X(0) − D × X(1) (x0 + x4 + x2 + x6 x1 + x5 + x3 + x7 x1 − x5 + ξ 2 x3 − ξ 2 x7 | x − x4 + ξ 2 x2 − ξ 2 x6 =(0)x0 + x4 − x2 − x6 x1 + x5 − x3 − x7 x0 − x4 − ξ 2 x2 + ξ 2 x6 x1 − x5 − ξ 2 x3 + ξ 2 x7 4×2 For j = 2 : ( ) | e−πi/4 | ξ | D ←− (−2πi/4) = (2)ξ e e−3πi/4 ξ3 1)(Solutions 81)x0 + x4 + x2 + x6 | x − x4 + ξ 2 x2 − ξ 2 x6 |X(0) ←− (0)x0 + x4 − x2 − x6 x0 − x4 − ξ 2 x2 + ξ 2 x6)(x1 + x5 + x3 + x7 | x − x5 + ξ 2 x3 − ξ 2 x7 |X(1) ←− (1)x1 + x5 − x3 − x7 x1 − x5 − ξ 2 x3 + ξ 2 x7 (0) X + D × X(1) X ←− X(0) − D × X(1) (x0 + x4 + x2 + x6 + x1 + x5 + x3 + x7 || x0 − x4 + ξ 2 x2 − ξ 2 x6 + ξ x1 − ξx5 + ξ 3 x3 − ξ 3 x7 || x0 + x4 − x2 − x6 + ξ 2 x1 + ξ 2 x5 − ξ 2 x3 − ξ 2 x7 || 2 3 3 5 | x − x4 − ξ 2 x2 + ξ 2 x6 + ξ x1 − ξ x5 − ξ x3 + ξ x7 |=| 0 || x0 + x4 + x2 + x6 − x1 − x5 − x3 − x7 || 2 3 3 | x0 − x4 + ξ 2 x2 − ξ 2 x6 − ξ x1 + ξ x5 − ξ 3 x3 + ξ x7 | (2 2 2 2 x0 + x4 − x2 − x6 − ξ x1 − ξ x5 + ξ x3 + ξ x7 2 3 3 5 x0 − x4 − ξ x2 + ξ x6 − ξ x1 + ξ x5 + ξ x3 − ξ x7 8×1 (To verify that this is the same as F8 x8 , use the fact that ξ = −ξ 5 , ξ 2 = −ξ 6 , ξ 3 = −ξ 7 , and ξ 4 = −1. Solutions for exercises in section 5. 9 5.9.1. (a) The fact that (1 rank (B) = rank X | Y = rank (1 1 1 2 2)1 2) = 3 3 implies BX ∪ BY is a basis for 3 , so (5.9.4) guarantees that X and Y are complementary. (b) According to (5.9.12), the projector onto X along Y is ( ))−1 1 1 0 1 1 −1 P = X|0 X|Y = (1 2 0)(1 2 2)(1 2 0 1 2 3 ( )( ) 1 1 0 2 −1 0 1 1 −1 = (1 2 0)(−1 2 −1) = (0 3 −2), 1 2 0 0 −1 1 0 3 −2 and (5.9.9) insures that the complementary (0 Q = I − P = (0 0 projector onto Y along X is)−1 1 −2 2). −3 3 82 Solutions ( )2 Qv = (4)6 (c) (d) Direct multiplication shows P2 = P and Q2 = Q. (e) To verify that R (P) = X = N (Q), you can use the technique of Example 4.2.2 to show that the basic columns of P (or the columns in a basis for N (Q)) span space generated by BX . To verify that N (P) = Y, note that (the) same (1 0 P (2) = (0) together with the fact that dim N (P) = 3 − rank (P) = 1. 3 0 5.9.2. There are many ways to do this. One way is to write down any basis for 5 —say B = {x1 , x2 , x3 , x4 , x5 }—and set X = span {x1 , x2 } and Y = span {x3 , x4 , x5 } . Property (5.9.4) guarantees that X and Y are complementary. 5.9.3. Let X = {(α, α) | α ∈ } and Y = 2 so that 2 = X + Y, but X ∩ Y = 0. For each vector in 2 we can write (x, y) = (x, x) + (0, y − x) and (x, y) = (y, y) + (x − y, 0). 5.9.4. Exercise 3.2.6 says that each A ∈ n×n can be uniquely written as the sum of a symmetric matrix and a skew-symmetric matrix according to the formula A= A + AT A − AT + , 2 2 so (5.9.3) guarantees that n×n = S ⊕ K. onto S along K is the S -component of given matrix, this is (1 A + AT = (3 2 5 5.9.5. (a) By definition, the projection of A A —namely (A + AT)/2. For the 3 5 7)5 7). 9 Assume that X ∩ Y = 0. To prove BX ∪ BY is linearly independent, write m i=1 αi xi + n βj yj = 0 =⇒ j=1 =⇒ =⇒ m i=1 m i=1 αi xi = − n βj yj j=1 αi xi ∈ X ∩ Y = 0 αi xi = 0 and n βj yj = 0 j=1 =⇒ α1 = · · · = αm = β1 = · · · = βn = 0 (because BX and BY are both independent). Solutions 83 Conversely, if BX ∪ BY is linearly independent, then v ∈X ∩Y =⇒ v = m αi xi and v= i=1 =⇒ m βj yj j=1 αi xi − i=1 n βj yj = 0 j=1 =⇒ α1 = · · · = αm = β1 = · · · = βn = 0 (because BX ∪ BY is independent) =⇒ v = 0. (b) No. Take X to be the xy-plane and Y to be the yz-plane in 3 with BX = {e1 , e2 } and BY = {e2 , e3 }. We have BX ∪ BY = {e1 , e2 , e3 }, but X ∩ Y = 0. (c) No, the fact that BX ∪ BY is linearly independent is no guarantee that X + Y is the entire space—e.g., consider two distinct lines in 3 . 5.9.6. If x is a fixed point for P, then Px = x implies x ∈ R (P). Conversely, if x ∈ R (P), then x = Py for some y ∈ V =⇒ Px = P2 y = Py = x. 5.9.7. Use (5.9.10) (which you just validated in Exercise 5.9.6) in conjunction with the definition of a projector onto X to realize that x ∈ X ⇐⇒ Px = x ⇐⇒ x ∈ R (P), and x ∈ R (P) ⇐⇒ Px = x ⇐⇒ (I − P)x = 0 ⇐⇒ x ∈ N (I − P). The statements concerning the complementary projector I − P are proven in a similar manner. 5.9.8. If θ is the angle between R (P) and N (P), it follows from (5.9.18) that 'P'2 = (1/ sin θ) ≥ 1. Furthermore, 'P'2 = 1 if and only if sin θ = 1 (i.e., θ = π/2), which is equivalent to saying R (P) ⊥ N (P). 5.9.9. Let θ be the angle between R (P) and N (P). We know from (5.9.11) that R (I − P) = N (P) and N (I − P) = R (P), so θ is also the angle between R (I − P) and N (I − P). Consequently, (5.9.18) says that 'I − P'2 = 1 = 'P'2 . sin θ 5.9.10. The trick is to observe that P = uvT is a projector because vT u = 1 implies P2 = uvT uvT = uvT = P, so the result of Exercise 5.9.9 insures that , , , , I − uvT , = ,uvT , . 2 2 84 Solutions , To prove that ,uvT ,2 = 'u'2 'v'2 , start with the definition of an induced , , , , matrix given in (5.2.4) on p. 280, and write ,uvT ,2 = maxx2 =1 ,uvT x,2 . If the maximum occurs at x = x0 with 'x0 '2 = 1, then , T, , , ,uv , = ,uvT x0 , = 'u' |vT x0 | 2 2 2 ≤ 'u'2 'v'2 'x0 '2 by CBS inequality = 'u'2 'v'2 . But we can also write , T , ,uv x, 'v'2 (vT v) 2 'u'2 'v'2 = 'u'2 = 'u'2 = 'v'2 'v'2 'v'2 , , , , T , , v , ≤ max ,uvT x, , = ,uv , 2 'v'2 2 x2 =1 , T, = ,uv , , 2 , , so ,uvT ,2 = 'u'2 'v'2 . Finally, if P = uvT , use Example 3.6.5 to write 2 2 2 'P'F = trace PT P = trace(vuT uvT) = trace(uT uvT v) = 'u'2 'v'2 . 5.9.11. p = Pv = [X | 0][X | Y]−1 v = [X | 0]z = Xz1 5.9.12. (a) Use (5.9.10) to conclude that R (P) = R (Q) =⇒ PQ∗j = Q∗j and QP∗j = P∗j =⇒ PQ = Q and QP = P. Conversely, use Exercise 4.2.12 to write + PQ = Q =⇒ R (Q) ⊆ R (P) QP = P =⇒ R (P) ⊆ R (Q) ∀ j =⇒ R (P) = R (Q). (b) Use N (P) = N (Q) =⇒ R (I − P) = R (I − Q) together with part (a). (c) From part (a), Ei Ej = Ej so that 2 αj Ej = j i αi αj Ei Ej = j i αi j i αi αj Ej j αj Ej = αj Ej . j Ir 0 5.9.13. According to (5.9.12), the projector onto X along Y is P = B B−1 , 0 0 where B = [X | Y] in which the columns of X and Y form bases for X Solutions 85 and Y, respectively. Since multiplication bynonsingular matrices does not alter Ir 0 the rank, it follows that rank (P) = rank = r. Using the result of 0 0 Example 3.6.5 produces Ir 0 Ir −1 trace (P) = trace B = trace B 0 0 0 Ir 0 = trace = r = rank (P). 0 0 0 0 B −1 B 5.9.14. (i) =⇒ (ii) : If v = x1 + · · · + xk and v = y1 + · · · + yk , where xi , yi ∈ Xi , then k (xi − yi) = 0 =⇒ (xk − yk) = − k−1 i=1 (xi − yi) i=1 =⇒ (xk − yk) ∈ Xk ∩ (X1 + · · · + Xk−1) = 0 =⇒ xk = yk and k−1 (xi − yi) = 0. i=1 Now repeat the argument—to be formal, use induction. (ii) =⇒ (iii) : The proof is essentially the same argument as that used to establish (5.9.3) =⇒ (5.9.4). (iii) =⇒ (i) : B always spans X1 + X2 + · · · + Xk , and since the hypothesis is that B is a basis for V, it follows that B is a basis for both V and X1 +· · ·+Xk . Consequently V = X1 + X2 + · · · + Xk . Furthermore, the set B1 ∪ · · · ∪ Bk−1 is linearly independent (each subset of an independent set is independent), and it spans Vk−1 = X1 +· · ·+Xk−1 , so B1 ∪· · ·∪Bk−1 must be a basis for Vk−1 . Now, since (B1 ∪· · ·∪Bk−1)∪Bk is a basis for V = (X1 +· · ·+Xk−1)+Xk , it follows from (5.9.2)–(5.9.4) that V = (X1 + · · · + Xk−1) ⊕ Xk , so Xk ∩ (X1 + · · · + Xk−1) = 0. The same argument can now be repeated on Vk−1 —to be formal, use induction. I 0 0 0 −1 5.9.15. We know from (5.9.12) that P = Q Q and I−P = Q Q−1 , 0 0 0 I so I 0 A11 A12 I 0 −1 −1 Q PAP = Q Q AQ Q−1 0 0 0 0 A21 A22 A11 0 =Q Q−1 . 0 0 The other three statements are derived in an analogous fashion. −1 5.9.16. According to (5.9.12), the projector onto X along Y is P = X | 0 X | Y, where the columns of X and Y are bases for X and Y, respectively. If 86 Solutions −1 Xn×r | Y = Ar×n C , then Ar×n = Xn×r Ar×n . C A The nonsingularity of X | Y and insures that X has full column rank C and A has full row rank. The fact that AX = Ir is a consequence of −1 Ir 0 Ar×n AX AY X|Y = Xn×r | Y = X|Y . C 0 I CX CY P = Xn×r | 0 5.9.17. (a) Use the fact that a linear operator P is a projector if and only if P is idempotent. If EF = FE = 0, then (E + F)2 = E + F. Conversely, if E + F is a projector, then (E + F)2 = E + F =⇒ EF + FE = 0 =⇒ E(EF + FE) = 0 and (EF + FE)E = 0 =⇒ EF = FE =⇒ EF = 0 = FE (because EF + FE = 0). Thus P = E + F is a projector if and only if EF = FE = 0. Now prove that under this condition R (P) = X1 ⊕X2 . Start with the fact that z ∈ R (P) if and only if Pz = z, and write each such vector z as z = x1 + y1 and z = x2 + y2 , where xi ∈ Xi and yi ∈ Yi so that Ex1 = x1 , Ey1 = 0, Fx2 = x2 , and Fy2 = 0. To prove that R (P) = X1 + X2 , write z ∈ R (P) =⇒ Pz = z =⇒ (E + F)z = z =⇒ (E + F)(x2 + y2) = (x2 + y2) =⇒ Ez = y2 =⇒ x1 = y2 =⇒ z = x1 + x2 ∈ X1 + X2 =⇒ R (P) ⊆ X1 + X2 . Conversely, X1 + X2 ⊆ R (P) because z ∈ X1 + X2 =⇒ z = x1 + x2 , where x1 ∈ X1 and x2 ∈ X2 =⇒ x1 = Ex1 and x2 = Fx2 =⇒ Fx1 = FEx1 = 0 and Ex2 = EFx2 = 0 =⇒ Pz = (E + F)(x1 + x2) = x1 + x2 = z =⇒ z ∈ R (P). The fact that X1 and X2 are disjoint follows by writing z ∈ X1 ∩ X2 =⇒ Ez = z = Fz =⇒ z = Ez = EFz = 0, Solutions 87 and thus R (P) = X1 ⊕ X2 is established. To prove that N (P) = Y1 ∩ Y2 , write Pz = 0 =⇒ (E + F)z = 0 =⇒ Ez = −EFz and FEz = −Fz =⇒ Ez = 0 and 0 = Fz =⇒ z ∈ Y1 ∩ Y2 . 5.9.18. Use the hint together with the result of Exercise 5.9.17 to write E − F is a projector ⇐⇒ I − (E − F) is a projector ⇐⇒ (I − E) + F is a projector ⇐⇒ (I − E)F = 0 = F(I − E) ⇐⇒ EF = F = FE. Under this condition, Exercise 5.9.17 says that R (I − E + F) = R (I − E) ⊕ R (F) and N (I − E + F) = N (I − E) ∩ N (F), so (5.9.11) guarantees R (E − F) = N (I − E + F) = N (I − E) ∩ N (F) = R (E) ∩ N (F) = X1 ∩ Y2 N (E − F) = R (I − E + F) = R (I − E) ⊕ R (F) = N (E) ⊕ R (F) = Y1 ⊕ X2 . 5.9.19. If EF = P = FE, then P is idempotent, and hence P is a projector. To prove that R (P) = X1 ∩ X2 , write z ∈ R (P) =⇒ Pz = z =⇒ E(Fz) = z and F(Ez) = z =⇒ z ∈ R (E) ∩ R (F) = X1 ∩ X2 =⇒ R (P) ⊆ X1 ∩ X2 . Conversely, z ∈ X1 ∩ X2 =⇒ Ez = z = Fz =⇒ Pz = z =⇒ X1 ∩ X2 ⊆ R (P), and hence R (P) = X1 ∩ X2 . To prove that N (P) = Y1 + Y2 , first notice that z ∈ N (P) =⇒ FEz = 0 =⇒ Ez ∈ N (F). This together with the fact that (I − E)z ∈ N (E) allows us to conclude that z = (I − E)z + Ez ∈ N (E) + N (F) = Y1 + Y2 =⇒ N (P) ⊆ Y1 + Y2 . 88 Solutions Conversely, z ∈ Y1 + Y2 =⇒ z = y1 + y2 , where yi ∈ Yi for i = 1, 2 =⇒ Ey1 = 0 and Fy2 = 0 =⇒ Pz = 0 =⇒ Y1 + Y2 ⊆ N (P). Thus N (P) = Y1 + Y2 . 5.9.20. (a) For every inner pseudoinverse, AA− is a projector onto R (A), and I − A− A is a projector onto N (A). The system being consistent means that b ∈ R (A) = R AA− =⇒ AA− b = b, so A− b is a particular solution. Therefore, the general solution of the system is A− b + N (A) = A− b + R I − A− A . (b) A− A is a projector along N (A), so Exercise 5.9.12 insures Q(A− A) = Q and (A− A)Q = (A− A). This together with the fact that PA = A allows us to write AXA = AQA− PA = AQA− A = AQ = AA− AQ = AA− A = A. Similarly, XAX = (QA− P)A(QA− P) = QA− (PA)QA− P = QA− AQA− P = Q(A− AQ)A− P = Q(A− A)A− P = QA− P = X, so X is a reflexive pseudoinverse for A. To show X has the prescribed range and nullspace, use the fact that XA is a projector onto R (X) and AX is a projector along N (X) to write R (X) = R (XA) = R QA− PA = R QA− A = R (Q) = L and N (X) = N (AX) = N AQA− P = N AA− AQA− P = N AA− P = N (P) = M. To prove uniqueness, suppose that X1 and X2 both satisfy the specified conditions. Then N (X2) = M = R (I − AX1) =⇒ X2 (I − AX1) = 0 =⇒ X2 = X2 AX1 and R (X2 A) = R (X2) = L = R (X1) =⇒ X2 AX1 = X1 , so X2 = X1 . Solutions 89 Solutions for exercises in section 5. 10 5.10.1. Since index(A) = k, we must have that rank Ak−1 > rank Ak = rank Ak+1 = · · · = rank A2k = · · · , so rank Ak = rank(Ak)2 , and hence index(Ak) ≤ 1. But Ak is singular k k (because A is singular) k so that index(A k) >n0. Consequently, index(A) = 1. 5.10.2. In this case, R A = 0 and N A = . The nonsingular component C in (5.10.5) is missing, and you can take Q = I, thereby making A its own core-nilpotent decomposition. 5.10.3. If A is nonsingular, then index(A) = 0, regardless of whether or not A is symmetric. If A is singular and symmetric, we want to prove index(A) = 1. The strategy is to show that R (A) ∩ N (A) = 0 because this implies that R (A) ⊕ N (A) = n . To do so, start with x ∈ R (A) ∩ N (A) =⇒ Ax = 0 and x = Ay for some y. Now combine this with the symmetry of A to obtain 2 xT = yT AT = yT A =⇒ xT x = yT Ax = 0 =⇒ 'x'2 = 0 =⇒ x = 0. 5.10.4. index(A) = 0 when A is nonsingular. If A is singular and normal we want to prove index(A) = 1. The strategy is to show that R (A) ∩ N (A) = 0 because this implies that R (A)⊕N (A) = C n . Recall from (4.5.6) that N (A) = N (A∗ A) and N (A∗) = N (AA∗), so N (A) = N (A∗). Start with x ∈ R (A) ∩ N (A) =⇒ Ax = 0 and x = Ay for some y, and combine this with N (A) = N (A∗) to obtain A∗ x = 0 and x = Ay =⇒ x∗ x = y∗ A∗ x = 0 =⇒ 'x'2 = 0 =⇒ x = 0. 3 5 5.10.5. Compute rank A0 = 3, rank (A) = 2, rank A2 = 1, and rank A = 1, to see that k = 2 is the smallest integer such that rank Ak = rank Ak+1 , so index(A) = 2. The Q = [X | Y] is a matrix in which the columns of matrix X are a basis for R A2 , and the columns of Y are a basis for N A2 . Since ( ) 1 1 0 EA2 = (0 0 0), 0 0 0 2 we have ( ) −8 X = (12)8 (and −1 Y=(1 0)0 0), 1 (so −8 Q = (12 8 −1 1 0)0 0). 1 90 Solutions It can now be verified that ( )(1/4 1/4 0 −2 Q−1 AQ = (−3 −2 0)(4 −2 −2 1 3)(−4 −8 4)(12 2 8 0 2 2 where C = [2] and and N2 = 0. Finally, AD = Q 5.10.6. (a) N= C−1 0 0 0)(0 2 0)=(0 1 0 −1 1 0 4 , 2 (−1 = (3/2 1 −2 −1 Q−1 −1 3/2 1 0 −2 −1)0 4), 2)0 0). 0 Because (1−λ | 0| J − λI = (0 0 0 1 −1−λ 0 0 0 0 0 1−λ 0 0 0 0 0 2−λ 0)0 0 | |0 |)0 2−λ and because a diagonal matrix is singular if and only if it has a zero-diagonal entry, it follows that J − λI is singular if and only if λ = 1 or λ = 2, so λ1 = 1 and λ2 = 2 are the two eigenvalues of J. To find the index of λ1 , use block multiplication to observe that J−I= 0 0 0 I2×2 =⇒ rank (J − I) = 2 = rank (J − I). Therefore, index(λ1) = 1. Similarly, J − 2I = −I3×3 0 0 0 and 2 rank (J − 2I) = 3 = rank (J − 2I) , so index(λ2) = 1. (b) Since (1−λ | 0| J − λI = | 0(0 0 1 1−λ 0 0 0 0 1 1−λ 0 0 0 0 0 2−λ 0)0 0 | |0 |)1 2−λ and since a triangular matrix is singular if and only if there exists a zero-diagonal entry (i.e., a zero pivot), it follows that J − λI is singular if and only if λ = 1 Solutions 91 or λ = 2, so λ1 = 1 and λ2 = 2 are the two eigenvalues of J. To find the index of λ1 , use block multiplication to compute (0| 0| J − I = |0 (0 0 (0| 0| (J − I)3 = | 0 (0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0)0 0 | |0 |)1 1)0 0 | |0 |)3 1 (0| 0| (J − I)2 = | 0 (0 0 (0| 0| (J − I)4 = | 0 (0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0)0 0 | |0 |)2 1)0 0 | |0 |)4 1 Since 2 3 4 rank (J − I) > rank (J − I) > rank (J − I) = rank (J − I) , it follows that index(λ1) = 3. A similar computation using λ2 shows that 2 3 rank (J − 2I) > rank (J − 2I) = rank (J − 2I) , so index(λ2) = 2. The fact that eigenvalues associated with diagonal matrices have index 1 while eigenvalues associated with triangular matrices can have higher indices is no accident. This will be discussed in detail in §7.8 (p. 587). 5.10.7. (a) If P is a projector, then, by (5.9.13), P = P2 , so rank (P) = rank P2 , and hence index(P) ≤ 1. If P = I, then P is singular, and thus index(P) = 1. If P = I, then index(P) = 0. An alternate argument could be given on the basis of the observation that n = R (P) ⊕ N (P). (b) Recall from (5.9.12) that if the columns of X and Y constitute bases for R (P) and N (P), respectively, then for Q = X | Y , Q−1 PQ = I 0 0 0 , I 0 and it follows that is the core-nilpotent decomposition for P. 0 0 k−1 i k−1 5.10.8. Suppose that to obi=0 αi N x = 0, and multiply both sides by N k−1 k−1 tain α N x = 0. By assumption, N x = 0, so α = 0, and hence 0 k−1 0 i k−2 α N x = 0. Now multiply both sides of this equation by N to proi=1 i duce α1 Nk−1 x = 0, and conclude that α1 = 0. Continuing in this manner (or by making aformal induction argument) gives α0 = α1 = α2 = · · · = αk−1 = 0. 5.10.9. (a) b ∈ R Ak ⊆ R (A) =⇒ b ∈ R (A) =⇒ Ax = b is consistent. 92 Solutions (b) We saw in 5.10.5 that when considered as linear operators re Example stricted to R Ak , both A and AD are invertible, and in fact they are true inverses of each other. Consequently, A and AD are one-to-one map k Exercise 4.7.18), so for each b ∈ R A there is pings on R Ak (recall a unique x ∈ R Ak such that Ax = b, and this unique x is given by −1 x = A/ b = AD b. k R(A) (c) Part (b) shows that AD b is a particular solution. The desired result follows because the general solution is any particular solution plus the general solution of the associated homogeneous equation. I 0 D 5.10.10. Notice that AA = Q Q−1 , and use the results from Example 5.10.3 0 0 (p. 398). I − AAD is the complementary projector, so it projects onto N Ak along R Ak . 5.10.11. In each case verify that the axioms (A1), (A2), (A4), and (A5) in the definition of a vector space given on p. 160 hold for matrix multiplication (rather than +). In parts (a) and (b) the identity element is the ordinary identity matrix, and the inverse of each member is the ordinary inverse. In part (c), the identity 1/2 1/2 element is E = because AE = A = EA for each A ∈ G, and 1/2 1/2 # 1 α α 1 1 = because AA# = E = A# A. α α 4α 1 1 5.10.12. (a) =⇒ (b) : If A belongs to a matrix group G in which the identity element is E, and if A# is the inverse of A in G, then A# A2 = EA = A, so x ∈ R (A) ∩ N (A) =⇒ x = Ay for some y and Ax = 0 =⇒ Ay = A# A2 y = A# Ax = 0 =⇒ x = 0. (b) =⇒ (c) : Suppose A is n × n, and let BR and BN be bases for R (A) and N (A), respectively. Verify that B = R (A) ∩ N (A) = 0 implies BR ∩ BN is a linearly independent set, and use the fact that there are n vectors in B to conclude that B is a basis for n . Statement (c) now follows from (5.9.4). (c) =⇒ (d) : Use the fact that R Ak ∩ N Ak = 0. (d) =⇒ (e) : Use the result of Example 5.10.5 together with the fact that the only nilpotent matrix of index 1 is the zero matrix. (e) =⇒ (a) : It is straightforward to verify that the set G= Q Xr×r 0 0 0 + Q−1 X is nonsingular is a matrix group, and it's clear that A ∈ G. Solutions 93 Cr×r 0 Q−1 . For the 0 0 given E, verify that EA = AE = A for all A ∈ G. The fact that E is the desired projector follows from (5.9.12). 5.10.13. (a) Use part (e) of Exercise 5.10.12 to write A = Q (b) Simply verify that AA# = A# A = E. Notice that the group inverse agrees with the Drazin inverse of A described in Example 5.10.5. However, the Drazin inverse exists for all square matrices, but the concept of a group inverse makes sense only for group matrices—i.e., when index(A) = 1. Solutions for exercises in section 5. 11 5.11.1. Proceed as described on p. 199 to determine the following bases for each of the four fundamental subspaces. T R (A) = span (−1), (−1) N A = span (0) R (A) = span (1) T N (A) = span (1) 5.11.2. 5.11.3. 5.11.4. 5.11.5. Since each to each vector in a basis vector in a basis for R (A) is orthogonal T for N AT , it follows that R (A) ⊥ N A . The same logic also explains T why N (A) ⊥ R A . Notice that R (A) is a plane through the origin in 3 , and N AT is the line through the origin perpendicular to this plane, so it is evident from the parallelogram law that R (A) ⊕ N AT = 3 . Similarly, T N (A) is the line through the origin normal to the plane defined by R A , so T 3 N (A) ⊕ R A = . V ⊥ = 0, and 0⊥ = V. ( ) 1 2 | 2 4 |If A = ( ), then R (A) = M, so (5.11.5) insures M⊥ = N AT . Using 0 1 3 6 row operations, a basis for N AT is computed to be ( ), ( ). 0 0 | | | (| 0 1 Verify that M⊥ is closed with respect to vector addition and scalar multiplication. If x, y ∈ M⊥ , then ,m x- = 0 = ,m y- for each m ∈ M so that ,m x + y- = 0 for each m ∈ M, and thus x + y ∈ M⊥ . Similarly, for every scalar α we have ,m αx- = α ,m x- = 0 for each m ∈ M, so αx ∈ M⊥ . (a) x ∈ N ⊥ =⇒ x ⊥ N ⊇ M =⇒ x ⊥ M =⇒ x ∈ M⊥ . 94 Solutions (b) Simply observe that x ∈ (M + N)⊥ ⇐⇒ x ⊥ (M + N) ⇐⇒ x ⊥ M and x ⊥ N ⇐⇒ x ∈ M⊥ ∩ N ⊥ . (c) Use part (b) together with (5.11.4) to write M⊥ + N ⊥ ⊥ ⊥ ⊥ = M⊥ ∩ N ⊥ ⊥ = M ∩ N , and then perp both sides. 5.11.6. Use the fact that dim R AT = rank AT = rank (A) = dim R (A) together with (5.11.7) to conclude that n = dim N (A) + dim R AT = dim N (A) + dim R (A). 5.11.7. U is a unitary matrix in which the columns of U1 are an orthonormal basis for R (A) and the columns of U2 are an orthonormal basis for N AT , so setting −1 X = U1 , Y = U2 , and X | Y = UT in (5.9.12) produces P = U1 UT1 . T According to (5.9.9), the projector onto N A along R (A) is I − P = I − U1 UT1 = U2 UT2 . T 5.11.8. Start with the first column of A, and set u = A∗1 + 6e1 = (2 2 −4) to obtain ( )( ) −6 0 −6 −3 2 −1 2 2uuT 1(R1 = I− T = (0 0) −1 2 2) and R1 A = (0 u u 3 0 −3 0 0 2 2 −1 Now set u = 0 −3 + 3e1 = T ˆ 2 = I − 2uu = R T u u 0 1 1 0 3 −3 to get and R2 = 1 0 0 ˆ2 R (1 = (0 0 0 0 1)0 1), 0 so (2 1 P = R2 R1 = (2 3 −1 −1 2 2)(2 −6 0 −1) and PA = (0 −3 2 0 0 −6 0 0) −3 B 0) = . 0 0 Solutions 95 Therefore, rank (A) = 2, and orthonormal bases for R (A) and N AT are extracted from the columns of U = PT as shown below. 2/3 2/3 −1/3 T R (A) = span (−1/3), (2/3) and N A = span (2/3) 2/3 2/3 −1/3 T Now work with BT , and set u = (B1∗)T + 9e1 = (3 0 −6 −3) to get ( )2 0 2 1 −9 0 2uuT 1 |0 3 0 0 |T| 0 −3 |T Q = I− T = (.) and QB = ( ) = 0 0 u u 3 2 0 −1 −2 1 0 −2 2 0 0 T Orthonormal bases for R A and N (A) are extracted from the columns of V = QT = Q as shown below. 2/3 0 | 2/3 1/3 | 0 || 1| 0 || 0 | R AT = span ( ), ( ) and N (A) = span ( )0) −2/3 | 2/3 | −1/3 | 1/3 0 −2/3 2/3 A URV factorization is obtained by setting U = PT , V = QT , and ( ) T −9 0 0 0 T 0 R= = (0 −3 0 0). 0 0 0 0 0 0 ( ) 1 0 1 1/2 5.11.9. Using EA = (0 1 0 0) along with the standard methods of Chapter 4, 0 0 0 0 we have −2 −4 −1 T R (A) = span (2), (−2) and N A = span (2) −4 1 2 0 1 0 −1 −1/2 R AT = span ( ), ( ) and N (A) = span ( ), ( ). 1 0 1 0 1/2 0 0 1 Applying the Gram–Schmidt procedure to each of these sets produces the following orthonormal bases for the four fundamental subspaces. −2 −1 1 −2 1 1 −2 BR(A) = BN (AT) = 2 3 3 3 −2 1 2 96 Solutions 2 0 1 0 1 BR(AT) = 3 2 1 0 BN (A) = −1 −1 1 1 0 = √ √ 1 1 3 2 −1 2 0 4 The matrices U and V were defined in (5.11.8) to be (1 −2 −2 U = BR(A) ∪ BN (AT) = (1 −2 3 −2 1 and √ 2 0 −3/2 1 0√ 0 3 V = BR(AT) ∪ BN (A) = (3/2 3 2 0 1 0 0) −1 2)2 √ −1/2 0√ −1/√2 4/ 2 Direct multiplication now produces (9 R = UT AV = (0 0 0 3 0 0 0 0) 0 0). 5.11.10. According to the discussion of projectors on p. 386, the unique vectors satisfying v = x+y, x ∈ R (A), and y ∈ N AT are given Tby x = Pv and y = (I−P)v, where P is the projector onto R (A) along N A . Use the results of Exercise 5.11.7 and Exercise 5.11.8 to compute ( )( )8 2 2 4 −1 1 P = U1 UT1 = (2 5 −4), x = Pv = (1), y = (I − P)v = (2). 9 2 −4 5 1 2 5.11.11. Observe that R (A) ∩ N (A) = 0 =⇒ index(A) ≤ 1, R (A) ⊥ N (A) =⇒ A is singular, R (A) ⊥ N (A) =⇒ R AT = R (A). It is now trial and error to build a matrix that satisfies the three conditions on 1 2 the right-hand side. One such matrix is A = . 1 2 5.11.12. R (A) ⊥ N (A) =⇒ R (A) ∩ N (A) = 0 =⇒ index(A) = 1 by using (5.10.4). The example in the solution to Exercise 5.11.11 shows that the converse is false. 5.11.13. The facts that real symmetric =⇒ hermitian =⇒ normal are direct consequences of the definitions. To show that normal =⇒ RPN, (4.5.5) to write use 1 i R (A) = R (AA∗) = R (A∗ A) = R (A∗). The matrix is hermitian −i 2 but not symmetric. To construct a matrix that is normal but not hermitian or Solutions 97 real symmetric, try to find an example with real numbers. If A = a b c d , then T AA = a2 + b2 ac + bd ac + bd c2 + d2 T and A A= a2 + c2 ab + cd ab + cd b2 + d2 , 1 −1 so we need to have b = c. One such matrix is A = . To construct 1 1 a singular matrix that is RPN but not normal, try again to find an example with real numbers. For any orthogonal matrix P and nonsingular matrix C, the C 0 matrix A = P PT is RPN. To prevent A from being normal, simply 0 0 1 2 choose C to be nonnormal. For example, let C = and P = I. 3 4 ∗ ∗ 5.11.14. (a) A∗ A = AA∗ =⇒ (A − λI) (A − λI) = (A − λI) (A − λI) =⇒ (A − λI) is normal =⇒ (A − λI) is RPN =⇒ R (A − λI) ⊥ N (A − λI). 2 2 (b) Suppose x ∈ N (A − λI) and y ∈ N (A − μI), and use the fact that ∗ N (A − λI) = N (A − λI) to write (A − λI) x = 0 =⇒ 0 = x∗ (A − λI) =⇒ 0 = x∗ (A − λI) y = x∗ (μy − λy) =⇒ 0 = x∗ y(μ − λ) =⇒ x∗ y = 0. Solutions for exercises in section 5. 12 25 0 0 100 , σ12 = 100, and it's clear that x = e2 is a vector −3/5 such that (CT C − 100I)x = 0 and 'x'2 = 1. Let y = Cx/σ1 = . −4/5 Following the procedure in Example 5.6.3, set ux = x − e1 and uy = y − e1 , and construct 5.12.1. Since CT C = Rx = I − 2 ux uTx = uTx ux 0 1 1 0 and Ry = I − 2 uy uTy = uTy uy −3/5 −4/5 −4/5 3/5 . 10 0 Since Ry CRx = = D, it follows that C = Ry DRx is a singular 0 5 value decomposition of C. 2 5.12.2. ν12 (A) = σ12 = 'A'2 needs no proof—it's just a restatement of (5.12.4). The 2 2 fact that νr (A) = 'A'F amounts to observing that 2 'A'F T = trace A A = traceV D2 0 0 0 VT = trace D2 = σ12 + · · · + σr2 . 98 Solutions 5.12.3. If σ1 ≥ · · · ≥ σr are the nonzero singular values for A, then it follows from Exercise 5.12.2 that 'A'22 = σ12 ≤ σ12 + σ22 + · · · + σr2 = 'A'F ≤ nσ12 = n'A'22 . 5.12.4. If rank (A + E) = k < r, then (5.12.10) implies that 'E'2 = 'A − (A + E)'2 ≥ min rank(B)=k 'A − B'2 = σk+1 ≥ σr , which is impossible. Hence rank (A + E) ≥ r = rank (A). 5.12.5. The argument is almost identical to that given for the nonsingular case except that A† replaces A−1 . Start with SVDs A=U D 0 0 0 V T and † A =V D−1 0 0 0 UT , , , , , where D = diag (σ1 , σ2 , . . . , σr) , and note that ,A† Ax,2 ≤ ,A† A,2 'x'2 = 1 with equality holding when A† A = I (i.e., when r = n). For each y ∈ A(S2) there is an x ∈ S2 such that y = Ax, so, with w = UT y, , , ,2 , ,2 , ,2 , ,2 1 ≥ ,A† Ax,2 = ,A† y,2 = ,VD−1 UT y,2 = ,D−1 UT y,2 = ,D−1 w,2 = 21 + 22 + · · · + 2r σ1 σ2 σr with equality holding when r = n. In other words, the set UT A(S2) is an ellipsoid (degenerate if r < n) whose k th semiaxis has length σk . To resolve the inequality with what it means for points to be on an ellipsoid, realize that the surface of a degenerate ellipsoid (one having some semiaxes with zero length) is actually the set of all points in and on a smaller dimension ellipsoid. For example, visualize an ellipsoid in 3 , and consider what happens as one of its semiaxes shrinks to zero. The skin of the three-dimensional ellipsoid degenerates to a solid planar ellipse. In other words, all points on a degenerate ellipsoid with semiaxes of length σ1 = 0, σ2 = 0, σ3 = 0 are actually points on and inside a planar ellipse with semiaxes of length σ1 and σ2 . Arguing that the k th semiaxis of A(S2) is σk U∗k = AV∗k is the same as the case given in the text. nonsingular −1 0 D D 0 † UT are SVDs in which 5.12.6. If A = U VT and A†n×m = V 0 0 0 0 V = V1 |V2 , then the columns of V1 are an orthonormal basis for R AT , so x ∈ R AT and 'x'2 = 1 if and only if x = V1 y with 'y'2 = 1. Since the 2-norm is unitarily invariant (Exercise 5.6.9), min x2 =1 x∈R(AT) 'Ax'2 = min 'AV1 y'2 = min 'Dy'2 = y2 =1 y2 =1 1 1 = σr . 'D−1 '2 'A† '2 Solutions 99 ˜ = A† (b − e) are the respective solutions of minimal 2-norm of 5.12.7. x = A† b and x ˜ = b − e. The development of the more general bound is Ax = b and A˜ x=b the same as for (5.12.8). ˜ ≤ 'A† ' 'b − b', ˜ ˜' = 'A† (b − b)' 'x − x b = Ax =⇒ 'b' ≤ 'A' 'x' =⇒ 1/'x' ≤ 'A'/'b', so ˜' 'x − x ˜ 'A' = κ 'e' . ≤ 'A ' 'b − b' 'x' 'b' 'b' Similarly, ˜ = 'A(x − x ˜)' ≤ 'A' 'x − x ˜ ', 'b − b' x = A† b =⇒ 'x' ≤ 'A† ' 'b' =⇒ 1/'b' ≤ 'A† '/'x', so ˜ ˜' 'b − b' 'A† ' 'x − x ˜ ') ≤ ('A' 'x − x =κ . 'b' 'x' 'x' Equality was attained in Example 5.12.1 by choosing b and e to point in ˜ = b − e cannot special directions. But for these choices, Ax = b and A^ x=b be guaranteed to be consistent for all singular or rectangular matrices A, so the answer to the second part is "no." However, the argument of Example 5.12.1 proves equality for all A such that AA† = I (i.e., when rank (Am×n) = m). 2 D 0 D + &I 0 VT is 5.12.8. If A = U VT is an SVD, then AT A + &I = U 0 &I 0 0 an SVD with no zero singular values, so it's nonsingular. Furthermore, −1 T (A A + &I) T A =U (D2 + &I)−1 D 0 0 0 V →U T D−1 0 0 0 V T = A† . , , −266000 667000 5.12.9. Since A = , κ∞ = 'A'∞ ,A−1 ,∞ = 1, 754, 336. 333000 −835000 Similar to the 2-norm situation discussed in Example 5.12.1, the worst case is realized when b is in the direction of a maximal vector in A(S∞) while e is in the direction of a minimal vector in A(S∞). −1 100 Solutions A (1.502, .599) (-1, 1) (1, 1) (.168, .067) (-.168, -.067) (-1, -1) (1, -1) (-1.502, -.599) T It's not clear which vector is minimal—don't assume (.168 .067) is., A min, imal vector y in A(S∞) satisfies 'y'∞ = minx∞ =1 'Ax'∞ = 1/ ,A−1 ,∞ (see (5.2.6) on p. 280), so, for y = Ax0 with 'x0 '∞ = 1, , , , −1 , , , , , y ,A , = 'x0 '∞ = 1 = ,A−1 , = max ,A−1 z, . , , ∞ ∞ 'y'∞ ∞ 'y'∞ 'y'∞ z∞ =1 ˆ = y/ 'y'∞ must be a vector in S∞ that receives maximal In other words, y stretch under A−1 . You don't have to look very hard to find such a vector because its components are ±1—recall the proof of (5.2.15) on p. 283. Notice ˆ ,= (1 −1)T ∈ S∞ ,,and ,y ˆ receives maximal stretch under A−1 because that , −1y ,A y, = 1, 168, 000 = ,A−1 , , so setting ∞ ∞ b = αv = α 1.502 .599 and ˆ=β e = βy 1 −1 produces equality in (5.12.8), regardless of α and β. You may wish to computationally verify thatthis is indeed the case. & −1 & &n 5.12.10. (a) Consider A = or A = for small & = 0. 1 0 0 & (b) For α > 1, consider ...

GAPMARK picture illustrates this in 2 . Solutions 105 L L−u u b u−v v p b−u p−u In other words, L is translated back down to span {u − v} , b → b − u, and p → p − u, so that p − u must be the orthogonal projection of b − u onto span {u − v} . Example 5.13.1 says that p − u = Pspan{u−v} (b − u) = (u − v)(u − v)T (b − u), (u − v)T (u − v) (u − v)T (b − u) p=u+ (u − v). (u − v)T (u − v) , , , , , 5.13.9. 'A3 x − b'2 = ,PR(A) b − b,2 = ,(I − PR(A))b,2 = ,PN (AT) b,2 5.13.10. Use (5.13.17) with PR(A) = PTR(A) = P2R(A) , to write and thus 2 'ε'2 = (b − PR(A) b)T (b − PR(A) b) = bT b − bT PTR(A) b − bT PR(A) b + bT PTR(A) PR(A) b , ,2 2 = bT b − bT PR(A) b = 'b'2 − ,PR(A) b,2 . r T 5.13.11. According to (5.13.13) we must show that i=1 (ui x)ui = PM x. It follows from (5.13.4) that if Un×r is the matrix containing the vectors in B as columns, then r r r T T T T PM = UU = ui ui ==⇒ PM x = ui ui x = (ui T x)ui . i=1 i=1 i=1 5.13.12. Yes, the given spanning set {u1 , u2 , u3 } is an orthonormal basis for M, so, by Exercise 5.13.11, ⎛ 5 3 ⎞ 0 ⎟ PM b = (ui b)ui = u1 + 3u2 + 7u3 = ⎜ ⎝ ⎠ . 5 i=1 3 106 Solutions 5.13.13. (a) Combine the fact that PM PN = 0 if and only if R (PN) ⊆ N (PM) with the facts R (PN) = N and N (PM) = M⊥ to write PM PN = 0 ⇐⇒ N ⊆ M⊥ ⇐⇒ N ⊥ M. (b) Yes—this is a direct consequence of part (a). Alternately, you could say 0 = PM PN ⇐⇒ 0 = (PM PN)T = PTN PTM = PN PM . 5.13.14. (a) Use Exercise 4.2.9 along with (4.5.5) to write ⎛ R (PM) + R (PN) = R ⎝(PM | PN) = R PM P M T + P N P N T PM T ⎞ ⎠ PN = R P2M + P2N = R (PM + PN) . (b) PM PN = 0 ⇐⇒ R (PN) ⊆ N (PM) ⇐⇒ N ⊆ M⊥ ⇐⇒ M ⊥ N . (c) Exercise 5.9.17 says PM + PN is idempotent if and only if PM PN = 0 = PN PM . Because PM and PN are symmetric, PM PN = 0 if and only if PN PM = 0 (via the reverse order law for transposition). The fact that R (PM + PN) = R (PM) ⊕ R (PN) = M ⊕ N was established in Exercise 5.9.17, and M ⊥ N follows from part (b). 5.13.15. First notice that PM + PN is symmetric, so (5.13.12) and the result of Exercise 5.13.14, part (a), can be combined to conclude that (PM + PN)(PM + PN)† = (PM + PN)† (PM + PN) = PR(PM +PN) = PM+N . Now, M ⊆ M + N implies PM+N PM = PM , and the reverse order law for transposition yields PM PM+N = PM so that PM+N PM = PM PM+N . In other words, (PM + PN)(PM + PN)† PM = PM (PM + PN)† (PM + PN), or PM (PM + PN)† PM + PN (PM + PN)† PM = PM (PM + PN)† PM + PM (PM + PN)† PN . Subtracting PM (PM + PN)† PM from both sides of this equation produces PM (PM + PN)† PN = PN (PM + PN)† PM . Let Z = 2PM (PM +PN)† PN = 2PN (PM +PN)† PM , and notice that R (Z) ⊆ R (PM) = M and R (Z) ⊆ R (PN) = N implies R (Z) ⊆ M∩N . Furthermore, PM PM∩N = PM∩N = PN PM∩N , and PM+N PM∩N = PM∩N , so, by the Solutions 107 reverse order law for transposition, PM∩N PM = PM∩N = PM∩N PN and PM∩N PM+N = PM∩N . Consequently, Z = PM∩N Z = PM∩N PM (PM + PN)† PN + PN (PM + PN)† PM = PM∩N (PM + PN)† (PM + PN) = PM∩N PM+N = PM∩N . Use the fact that AT = AT PR(A) = AT AA† (see Exercise 5.13.4) to write 5.13.16. (a) 4 ∞ e−A T 4 ∞ T AT AA† dt = e−A At AT Adt A† 0 0 &∞ % −AT At † = −e A = [0 − (−I)]A† = A† . 4 At AT dt = 0 ∞ e−A T At 0 Recall from Example 5.10.5 that Ak = Ak+1 AD = Ak AAD , and write (b) 4 ∞ e−A k+1 4 ∞ k+1 Ak AAD dt = e−A t Ak+1 Adt AD 0 0 &∞ % −Ak+1 t D = −e A = [0 − (−I)]AD = AD . 4 t Ak dt = 0 ∞ e−A k+1 t 0 (c) This is just a special case of the formula in part (b) with k = 0. However, it is easy to derive the formula directly by writing 4 ∞ 4 ∞ e−At AA−1 dt = e−At Adt A−1 0 0 &∞ % −At −1 = e A = [0 − (−I)]A−1 = A−1 . e−At dt = 0 4 ∞ 0 5.13.17. (a) The points in H are just solutions to a linear system uT x = β. Using the fact that the general solution of any linear system is a particular solution plus the general solution of the associated homogeneous equation produces H= βu βu βu ⊥ + N (uT) = T + [R(u)] = T + u⊥ , T u u u u u u where u⊥ denotes the orthogonal complement of the one-dimensional space spanned by the vector u. Thus H = v+M, where v = βu/uT u and M = u⊥ . The fact that dim (u⊥) = n − 1 follows directly from (5.11.3). (b) Use (5.13.14) with part (a) and the fact that Pu⊥ = I − uuT /uT u to write T βu uuT βu βu uuT b u b−β p= T + I− T b− T = T +b− T = b−. u u u u u u u u u u uT u 108 Solutions 5.13.18. (a) uT w = 0 implies M ∩ W = 0 so that dim (M + W) = dim M + dim W = (n − 1) + 1 = n. Therefore, M+W = n . This together with M∩W = 0 means n = M⊕W. (b) Write uT b uT b uT b b = b − T w + T w = p + T w, u w u w u w and observe that p ∈ M (because uT p = 0) and (uT b/uT w)w ∈ W. By definition, p is the projection of b onto M along W. (c) We know from Exercise 5.13.17, part (a), that H = v + M, where v = βu/uT u and M = u⊥ , so subtracting v = βu/uT u from everything in H as well as from b translates the situation back to the origin. Sketch a picture similar to that of Figure 5.13.5 to see that this moves H back to M, it translates b to b − v, and it translates p to p − v. Now, p − v should be the projection of b − v onto M along W, so by the result of part (b), T uT (b − v) uT (b − v) u b−β p−v = b−v− w ==⇒ p = b − w = b − w. uT w uT w uT w 5.13.19. For convenience, set β = Ai∗ pkn+i−1 − bi so that pkn+i = pkn+i−1 − β(Ai∗) . Use the fact that Ai∗ (pkn+i−1 − x) = Ai∗ pkn+i−1 − bi = β together with 'Ai∗ '2 = 1 to write ,2 , , 2 T 'pkn+i − x'2 = ,pkn+i−1 − β(Ai∗) − x, 2 ,2 , T, = , (pkn+i−1 − x) − β(Ai∗) , 2 T = (pkn+i−1 − x) (pkn+i−1 − x) T − 2βAi∗ (pkn+i−1 − x) + β 2 Ai∗ (Ai∗) 2 = 'pkn+i−1 − x'2 − β 2 . Consequently, 'pkn+i − x'2 ≤ 'pkn+i−1 − x'2 , with equality holding if and only if β = 0 or, equivalently, if and only if pkn+i−1 ∈ Hi−1 ∩ Hi . Therefore, the sequence of norms 'pkn+i − x'2 is monotonically decreasing, and hence it must have a limiting value. This implies that the sequence of the β 's defined above must approach 0, and thus the sequence of the pkn+i 's converges to x. (1) (1) 5.13.20. Refer to Figure 5.13.8, and notice that the line passing from p1 to p2 is (1) (1) parallel to V = span p1 − p2 (1) , so projecting p1 (1) through p2 onto H2 Solutions 109 (1) is exactly the same as projecting p1 onto H2 along (i.e., parallel to) V. According to part (c) of Exercise 5.13.18, this projection is given by (2) p2 (1) A2∗ p1 − b1 AT2∗ (1) (1) p(1) = p1 − − p2 1 (1) (1) A2∗ p1 − p2 (1) A2∗ p1 − p2 (1) (1) p(1) = p1 − . − p 1 2 (1) (1) A2∗ p1 − p2 All other projections are similarly derived. It is now straightforward to verify that the points created by the algorithm are exactly the same points described in Steps 1, 2, . . . , n − 1. ' ((1) (1) (1) (1) (1) (1) p1 − p2 , p1 − p3 , . . . , p1 − pn ' ((2) (2) (2) (2) (2) (2) is independent insures that p2 − p3 , p2 − p4 , . . . , p2 − pn is also The same holds at each subsequent step. independent. Furthermore, (1) (1) (1) A2∗ p1 − pk = 0 for k > 1 implies that Vk = span p1 − pk is not parallel to H2 , so all projections onto H2 along Vk are well defined. It can be argued that the analogous situation holds at each step of the process—i.e., (i) (i) the initial conditions insure Ai+1∗ pi − pk = 0 for k > i. Note: The condition that 5.13.21. Equation (5.13.13) says that the orthogonal distance between x and M⊥ is dist (x, M⊥) = 'x − PM⊥ x'2 = '(I − PM⊥)x'2 = 'PM x'2 . Similarly, dist (Rx, M⊥) = 'PM Rx'2 = '−PM x'2 = 'PM x'2 . 5.13.22. (a) We know from Exercise 5.13.17 that H = v + u⊥ , where v = βu, so subtracting v from everything in H as well as from b translates the situation back to the origin. As depicted in the diagram below, this moves H down to u⊥ , and it translates b to b − v and r to r − v. 110 Solutions b H p v b-v u u⊥ p-v 0 Now, we know from (5.6.8) that the reflection of b − v about u⊥ is r − v = R(b − v) = (I − 2uuT)(b − v) = b + (β − 2uT b)u, and therefore the reflection of b about H is r = R(b − v) + v = b − 2(uT b − β)u. (b) From part (a), the reflection of r0 about Hi is T ri = r0 − 2(Ai∗ r0 − bi) (Ai∗) , and therefore the mean value of all of the reflections {r1 , r2 , . . . , rn } is 1 1 T r0 − 2(Ai∗ r0 − bi) (Ai∗) ri = n i=1 n i=1 n m= n 2 (Ai∗ r0 − bi)(Ai∗)T n i=1 n = r0 − 2 T 2 A (Ar0 − b) = r0 − AT ε. n n Note: If weights wi > 0 such that wi = 1 are used, then the weighted mean is n n T m= wi ri = wi r0 − 2(Ai∗ r0 − bi) (Ai∗) = r0 − i=1 = r0 − 2 i=1 n wi (Ai∗ r0 − bi)(Ai∗)T i=1 2 2 = r0 − AT W (Ar0 − b) = r0 − AT Wε, n n Solutions 111 where W = diag {w1 , w2 , . . . , wn } . (c) First observe that 2 T A εk−1 n 2 = x − mk−1 + AT (Amk−1 − b) n 2 = x − mk−1 + AT (Amk−1 − Ax) n 2 = I − AT A (x − mk−1), n x − mk = x − mk−1 + and then use successive substitution to conclude that x − mk = I − 2 T A A n k (x − m0). Solutions for exercises in section 5. 14 5.14.1. Use (5.14.5) to observe that E[yi yj] = Cov[yi , yj] + μyi μyj = σ 2 + (Xi∗ β)2 if i = j, (Xi∗ β)(Xj∗ β) if i = j, so that E[yyT] = σ 2 I + (Xβ)(Xβ)T = σ 2 I + XββT XT . ˆ = (X − XX†)y, and use the fact that I − XX† is idempotent ˆ = y − Xβ Write e to obtain ˆT e ˆ = yT (I − XX†)yyT . e Now use the linearity of trace and expectation together with the result of Exercise 5.9.13 and the fact that (I − XX†)X = 0 to write ˆ] = E trace (I − XX†)yyT = trace E[(I − XX†)yyT] E[ˆ eT e = trace (I − XX†)E[yyT] = trace (I − XX†)(σ 2 I + XββT XT) = σ 2 trace I − XX† = σ 2 m − trace XX† = σ 2 m − rank XX† = σ 2 (m − n). 112 Solutions Solutions for exercises in section 5. 15 5.15.1. (a) (b) 5.15.2. (a) This θmin = 0, and θmax = θ = φ = π/4. θmin = θ = φ = π/4, and θmax = 1. The first principal angle is θ1 = θmin = 0, and we can take u1 = v1 = e1 . means that M2 = u⊥ 1 ∩ M = span {e2 } and N2 = v1⊥ ∩ N = span {(0, 1, 1)} . The second principal angle is the minimal angle between M2 and N2 , and this is just the angle between e2 and (0, 1, 1), so θ2 = π/4. (b) This time the first √ principal √ angle is θ1 = θmin = π/4, and we can take u1 = e1 and v1 = (0, 1/ 2, 1/ 2). There are no more principal angles because N2 = v1⊥ ∩ N = 0. 5.15.3. (a) This follows from (5.15.16) because PM = PN if and only if M = N . (b) If 0 = x ∈ M ∩ N , then (5.15.1) evaluates to 1 with the maximum being attained at u = v = x/ 'x'2 . Conversely, cos θmin = 1 ==⇒ vT u = 1 for some u ∈ M and v ∈ N such that 'u'2 = 1 = 'v'2 . But vT u = 1 = 'u'2 'v'2 represents equality in the CBS inequality (5.1.3), and we know this occurs if and only if v = αu for α = vT u/u∗ u = 1/1 = 1. Thus u = v ∈ M ∩ N. (c) max u∈M, v∈N vT u = 0 ⇐⇒ vT u = 0 ∀ u ∈ M, v ∈ V ⇐⇒ M ⊥ N. u2 =v2 =1 5.15.4. You can use either (5.15.3) or (5.15.4) to arrive at the result. The latter is used by observing , , , −1 , , ,(PM⊥ − PN ⊥) −1 , = ,) − (I − P) (I − P , , M N 2 2 , , , , = ,(PN − PM)−1 ,2 = ,(PM − PN)−1 ,2 . 5.15.5. M ⊕ N ⊥ = n ==⇒ dim M = dim N ==⇒ sin θmax = δ(M, N) = δ(N , M), so cos θ˜min = 'PM PN ⊥ '2 = 'PM (I − PN)'2 = δ(M, N) = sin θmax . 5.15.6. It was argued in the proof of (5.15.4) that PM − PN is nonsingular whenever M and N are complementary, so we need only prove the converse. Suppose dim M = r > 0 and dim N = k > 0 (the problem is trivial if r = 0 or k = 0) so that UT1 V1 is r × n − k and UT2 V2 is n − r × k. If PM − PN is nonsingular, then (5.15.7) insures that the rows as well as the columns in each of these products must be linearly independent. That is, UT1 V1 and UT2 V2 must both be square and nonsingular, so r + k = n. Combine this with the formula for the rank of a product (4.5.1) to conclude k = rank UT2 V2 = rank UT2 − dim N UT2 ∩ R (V2) = n − r − dim M ∩ N = k − dim M ∩ N . It follows that M ∩ N = 0, and hence M ⊕ N = n . Solutions 113 5.15.7. (a) This can be derived from (5.15.7), or it can be verified by direct multiplication by using PN (I − P) = I − P ==⇒ P − PN P = I − PN to write (PM − PN)(P − Q) = PM P − PM Q − PN P + PN Q = P − 0 − PN P + P N Q = I − P N + P N Q = I − PN (I − Q) = I. (b) and (c) follow from (a) in conjunction with (5.15.3) and (5.15.4). 5.15.8. Since we are maximizing over a larger set, maxx=1 f (x) ≤ maxx≤1 f (x). A strict inequality here implies the existence of a nonzero vector x0 such that 'x0 ' < 1 and f (x0) > f (x) for all vectors such that 'x' = 1. But then f (x0) > f (x0 /'x0 ') = f (x0)/ 'x0 ' ==⇒ 'x0 ' f (x0) > f (x0), which is impossible because 'x0 ' < 1. 5.15.9. (a) We know from equation (5.15.6) that PMN = U C 0 0 0 VT in which C is nonsingular and C−1 = V1T U1 . Consequently, P†MN (b) =V C−1 0 0 0 UT = V1 C−1 UT1 = V1 V1T U1 UT1 = PN PM . Use the fact that , T , , , , , ,(U1 V1)−1 , = ,(V1T U1)−1 , = ,U1 (V1T U1)−1 V1T , 2 2 2 , , , † , , , T T † = ,(V1 V1 U1 U1) ,2 = , (I − PN)PM , 2 (and similarly for the other term) to show that , , , , † , † , , , , , (I − PN)PM , = ,(UT1 V1)−1 ,2 = , PM (I − PN) , , 2 and 2 , , , , † , † , , , , , (I − PM)PN , = ,(UT2 V2)−1 ,2 = , PN (I − PM) , , 2 . , , It was established in the proof of (5.15.4) that ,(UT1 V1)−1 ,2 = ,(UT2 V2)−1 ,2 , so combining this with the result of part (a) and (5.15.3) produces the desired conclusion. 5.15.10. (a) We know from (5.15.2) that cos θ˜min = 'PN ⊥ PM '2 = '(I − PN)PM '2 , † and we know from Exercise 5.15.9 that PMN = (I − PN)PM , so taking the pseudoinverse of both sides of this yields the desired result. 114 Solutions (b) Use (5.15.3) together with part (a), (5.13.10), and (5.13.12) to write , , , , , , cos θ˜min , , , , , , 1 = ,PMN P†MN , ≤ ,PMN , ,P†MN , = . sin θmin 2 2 2 5.15.11. (a) Use the facts that 'A'2 = 'AT '2 and (AT)−1 = (A−1)T to write ,2 , T 1 1 , ,2 = , ,2 = min ,V2 U2 x,2 T T −1 −1 x2 =1 ,(U V2) , ,(V U2) , 2 2 2 2 = min xT UT2 V2 V2T U2 x x2 =1 = min xT UT2 (I − V1 V1T)U2 x = min x2 =1 x2 =1 , ,2 1 − ,V1T U2 x,2 , ,2 ,2 ,2 , = 1 − max ,V1T U2 x,2 = 1 − ,V1T U2 ,2 = 1 − ,UT2 V1 ,2 . x2 =1 (b) Use a similar technique to write , T ,2 , T , , , ,U2 V2 , = ,U2 V2T ,2 = ,UT2 (I − V1 V1T),2 2 2 2 , , 2 = ,(I − V1 V1T)U2 ,2 = max xT UT2 (I − V1 V1T)U2 x x2 =1 , ,2 = 1 − min ,V1T U2 x,2 = 1 − , , T x2 =1 ,(V U2)−1 ,2 1 2 1 =1− , , . ,(UT V1)−1 ,2 2 2 Solutions for Chapter 6 Solutions for exercises in section 6. 1 −1 (b) 8 (c) −αβγ a11 a22 a33 + a12 a23 a31 + a13 a21 a32 − (a11 a23 a32 + a12 a21 a33 + a13 a22 a31) (This is where the "diagonal rule" you learned in high school comes from.) 1/2 If A = [x1 | x2 | x3], then V3 = det AT A = 20 (recall Example 6.1.4). But you could also realize that the xi 's are mutually orthogonal to conclude that V3 = 'x1 '2 'x2 '2 'x3 '2 = 20. (a) 10 (b) 0 (c) 120 (d) 39 (e) 1 (f) (n − 1)! rank (A) = 2 A square system has a unique solution if and only if its coefficient matrix is nonsingular—recall the discussion in §2.5. Consequently, (6.1.13) guarantees that a square system has a unique solution if and only if the determinant of the coefficient matrix is nonzero. Since 1 α 0 0 1 −1 = 1 − α2 , α 0 1 6.1.1. (a) (d) 6.1.2. 6.1.3. 6.1.4. 6.1.5. it follows that there is a unique solution if and only if α = ±1. 6.1.6. I = A−1 A ==⇒ det (I) = det A−1 A = det A−1 det (A) ==⇒ 1 = det A−1 det (A) ==⇒ det A−1 = 1/det (A). 6.1.7. Use the product rule (6.1.15) to write det P−1 AP = det P−1 det (A)det (P) = det P−1 det (P)det (A) = det P−1 P det (A) = det (I)det (A) = det (A). 6.1.8. Use (6.1.4) together with the fact that z1 z2 = z¯1 z¯2 and z1 + z2 = z¯1 + z¯2 for all complex numbers to write T ¯ ¯ = det (A∗) = det A σ(p)a1p1 · · · anpn = det A = p σ(p)a1p1 · · · anpn = p 6.1.9. (a) I = Q∗ Q Exercise 6.1.8. ==⇒ σ(p)a1p1 · · · anpn = det (A). p 1 = det (Q∗ Q) = det (Q∗)det (Q) = [det (Q)]2 by 116 Solutions (b) If A = UDV∗ is an SVD, then, by part (a), |det (A)| = |det (UDV∗)| = |det (U)| |det (D)| |det (V∗)| = det (D) = σ1 σ2 · · · σn . 6.1.10. Let r = rank (A), and let σ1 ≥ · · · ≥ σr be the nonzero singular values of A. Dr×r 0 If A = Um×m (V∗) is an SVD, then, by Exercises 6.1.9 and 0 0 m×n n×n 6.1.8, det (V)det (V∗) = |det (V)|2 = 1, so ∗ (D D)r×r 0 det (V∗) = det (A∗ A) = det (VD∗ DV∗) = det (V) 0 0 n×n ' = 0 when r < n, = σ12 σ22 · · · σr2 05 ·67 · · 08, and this is > 0 when r = n. n−r 6.1.11. 6.1.12. 6.1.13. 6.1.14. Note: You can't say det (A∗ A) = det (A)det (A) = |det (A)|2 ≥ 0 because A need not be square. n αA = (αI)A ==⇒ det (αA) = det (αI)det (A) = α det (A). T T A = −A ==⇒ det (A) = det −A = det (−A) = (−1)n det (A) (by Exercise 6.1.11) ==⇒ det (A) = −det (A) when n is odd ==⇒ det (A) = 0. If A = LU, where L is lower triangular and U is upper triangular where each has 1's on its diagonal and random integers in the remaining nonzero positions, then det (A) = det (L)det (U) = 1 × 1 = 1, and the entries of A are rather random integers. According to the definition, det (A) = σ(p)a1p1 · · · akpk · · · anpn = p σ(p)a1p1 · · · (xpk + ypk + · · · + zpk) · · · anpn = p σ(p)a1p1 · · · xpk · · · anpn + p + · · · + p σ(p)a1p1 · · · ypk · · · anpn p σ(p)a1p1 · · · zpk · · · anpn ⎞ ⎛ ⎞ ⎛ A1∗ A1∗ A1∗ . . . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎜ T ⎟ ⎜ T ⎟ ⎜ T ⎟ = det ⎜ x ⎟ + det ⎜ y ⎟ + · · · + det ⎜ zT ⎟ . ⎜ . ⎟ ⎜ . ⎟ ⎜ . ⎟ ⎝ .. ⎠ ⎝ .. ⎠ ⎝ .. ⎠ An∗ An∗ An∗ ⎛ Solutions 117 6.1.15. If An×2 = [x | y] , then the result of Exercise 6.1.10 implies ∗ x x x∗ y 0 ≤ det (A∗ A) = ∗ y x y∗ y = (x∗ x) (y∗ y) − (x∗ y) (y∗ x) 2 2 = 'x'2 'y'2 − (x∗ y) (x∗ y) 2 2 = 'x'2 'y'2 − |x∗ y|2 , 2 2 with equality holding if and only if rank (A) < 2 —i.e., if and only if y is a scalar multiple of x. 6.1.16. Partition A as Lk 0 Uk U12 Lk Uk ∗ A = LU = = 0 U22 L21 L22 ∗ ∗ to deduce that Ak can be written in the form Uk−1 c Lk−1 0 Ak = Lk Uk = T 0 ukk and Ak−1 = Lk−1 Uk−1 . The product rule (6.1.15) shows that det (Ak) = det (Uk−1) × ukk = det (Ak−1) × ukk , and the desired conclusion follows. 6.1.17. According to (3.10.12), a matrix has an LU factorization if and only if each leading principal submatrix is nonsingular. The leading k × k principal submatrix of AT A is given by Pk = ATk Ak , where Ak = [A∗1 | A∗2 | · · · | A∗k] . If A has full column rank, then any nonempty subset of columns is linearly independent, so rank (A k) = k. Therefore, the results of Exercise 6.1.10 insure that det (Pk) = det ATk Ak > 0 for each k, and hence AT A has an LU factorization. The fact that each pivot is positive follows from Exercise 6.1.16. 6.1.18. (a) To evaluate det (A), use Gaussian elimination as shown below. ⎛ ⎞ ⎛ ⎞ 2−x 3 4 1 −1 3−x ⎝ 0 4−x −5 ⎠ −→ ⎝ 0 4−x −5 ⎠ 1 −1 3−x 2−x 3 4 ⎛ ⎞ 1 −1 3−x 1 −1 3−x ⎠ = U. ⎠ −→ ⎝ 0 4 − x −5 −→ ⎝ 0 4 − x −5 2 0 0 −x3 +9x2 −17x+17 2 0 5 − x −x + 5x − 2 0 0 4−x Since one interchange was used, det (A) is (−1) times the product of the diagonal entries of U, so d det (A) = −x3 + 9x2 − 17x − 17 and = −3x2 + 18x − 17. dx 118 Solutions (b) Using formula (6.1.19) produces d det (A) dx −1 = 0 1 3 4 4 2 − x 4 − x −5 + 1 −1 3 − x 0 0 −1 3−x 3 0 0 2 − x 1 −1 4−x −5 + 0 −1 −1 3 − x = 1 = (−x2 + 7x − 7) + (−x2 + 5x − 2) + (−x2 + 6x − 8) = −3x2 + 18x − 17. 6.1.19. No—almost any 2 × 2 example will show that this cannot hold in general. 6.1.20. It was argued in Example 4.3.6 that if there is at least one value of x for which the Wronski matrix ⎛ ⎜ W(x) = ⎜ ⎝ f1 (x) f2 (x) · · · fn (x) f1 (x) .. . f2 (x) . . . · · · .. . fn (x) .. . (n−1) f1 (n−1) (x) f2 (n−1) (x) · · · fn ⎞ ⎟ ⎟ ⎠ (x) is nonsingular, then S is a linearly independent set. This is equivalent to saying that if S is a linearly dependent set, then the Wronski matrix W(x) is singular for all values of x. But (6.1.14) insures that a matrix is singular if and only if its determinant is zero, so if S is linearly dependent, then the Wronskian w(x) must vanish for every value of x. The converse of this statement is false (Exercise 4.3.14). 6.1.21. (a) (n!)(n−1) (b) 11 × 11 (c) About 9.24×10153 sec ≈ 3×10146 years (d) About 3 × 10150 mult/sec. (Now this would truly be a "super computer.") Solutions for exercises in section 6. 2 6.2.1. (a) 8 6.2.2. (a) A−1 (c) −3 ⎛ ⎞ 0 1 −1 adj (A) 1⎝ = = −8 4 4 ⎠ det (A) 8 16 −6 −2 (b) 39 ⎛ (b) 6.2.3. (a) A−1 −12 1 ⎞ −9 adj (A) = = ⎝ −6 det (A) 39 9 x1 = 1 − β, x2 = α + β − 1, 25 9 6 4 −14 9 6 4 ⎞ 7 15 ⎠ −3 −2 x3 = 1 − α Solutions 119 (b) Cramer's rule yields 1 t4 t2 4 2 2 3 t t 1 t4 t t t + t 3 t 0 1 t t t2 t3 x2 (t) = = 2 2 21 t t 1 t t2 t 2 2 − tt t t 1 t 1 + t t 1 t t t2 1 t3 − t 6 −t3 = 3 = , (t − 1)(t3 − 1) (t3 − 1) and hence lim x2 (t) = lim t→∞ t→∞ 1 t2 −1 = −1. 1 − 1/t3 6.2.4. Yes. 6.2.5. (a) Almost any two matrices will do the job. One example is A = I and B = −I. (b) Again, almost anything you write down will serve the purpose. One example is A = D = 02×2 , B = C = I2×2 . 6.2.6. Recall from Example 5.13.3 that Q = I − BBT B−1 BT . According to (6.2.1), T B B BT c det AT A = det = det BT B cT Qc . T T c B c c Since det BT B > 0 (by Exercise 6.1.10), cT Qc = det AT A /det BT B . A −C 6.2.7. Expand T both of the ways indicated in (6.2.1). D Ik 6.2.8. The result follows from Example 6.2.8, which says A[adj (A)] = det (A) I, together with the fact that A is singular if and only if det (A) = 0. 6.2.9. The solution is x = A−1 b, and Example 6.2.7 says that the entries in A−1 are continuous functions of the entries in A. Since xi = [A−1]ik bk , and since the sum of continuous functions is again continuous, it follows that each xi is a continuous function of the aij 's. ˚ij = αn−1 ˚ B = αn−1 ˚ 6.2.10. If B = αA, then Exercise 6.1.11 implies B Aij , so ˚ A, and adj (B) = αn−1 adj (A) . hence adj (B) = α 6.2.11. (a) We saw in §6.1 that rank (A) is the order of the largest nonzero minor of A. If rank (A) < n − 1, then every minor of order n − 1 (as well as det (A) T itself) must be zero. Consequently, ˚ A = 0, and thus adj (A) = ˚ A = 0. (b) rank (A) = n − 1 ==⇒ ==⇒ at least one minor of order n − 1 is nonzero some ˚ Aij = 0 ==⇒ adj (A) = 0 ==⇒ rank (adj (A)) ≥ 1. 120 Solutions Also, rank (A) = n − 1 ==⇒ ==⇒ ==⇒ ==⇒ det (A) = 0 A[adj (A)] = 0 (by Exercise 6.2.8) R (adj (A)) ⊆ N (A) dim R (adj (A)) ≤ dim N (A) ==⇒ rank (adj (A)) ≤ n − rank (A) = 1. (c) rank (A) = n ==⇒ det (A) = 0 ==⇒ adj (A) = det (A) A−1 ==⇒ rank (adj (A)) = n 6.2.12. If det (A) = 0, then Exercise 6.2.11 insures that rank (adj (A)) ≤ 1. Consequently, det (adj (A)) = 0, and the result is trivially true because both sides are zero. If det (A) = 0, apply the product rule (6.1.15) to A[adj (A)] = n det (A) I (from Example 6.2.8) to obtain det (A)det (adj (A)) = [det (A)] , n−1 . so that det (adj (A)) = [det (A)] 6.2.13. Expanding in terms of cofactors of the first row produces Dn = 2˚ A11 − ˚ A12 . But ˚ A11 = Dn−1 and expansion using the first column yields ˚ A12 −1 −1 0 0 2 −1 2 = (−1) 0 −1 .. .. 0 0 0 · · · 0 · · · 0 · · · 0 = (−1)(−1)Dn−2 , ... · · · 2 so Dn = 2Dn−1 − Dn−2 . By recursion (or by direct substitution), it is easy to see that the solution of this equation is Dn = n + 1. 6.2.14. (a) Use the results of Example 6.2.1 with λi = 1/αi . (b) Recognize that the matrix A is a rank-one updated matrix in the sense that ⎛ ⎞ 1 . A = (α − β)I + βeeT , where e = ⎝ .. ⎠ . 1 If α = β, then A is singular, so det (A) = 0. If α = β, then (6.2.3) may be applied to obtain βeT e nβ det (A) = det (α − β)I 1+ = (α − β)n 1 + . α−β α−β (c) Recognize that the matrix is I + edT , where ⎛ ⎞ 1 ⎛ 1 ⎞ ⎜1⎟ e=⎜ ⎝ ... ⎠ 1 ⎛ and ⎜ α1 ⎞ α2 ⎟ d=⎜ ⎝ ... ⎠ . αn Solutions 121 Apply (6.2.2) to produce the desired formula. 6.2.15. (a) Use the second formula in (6.2.1). (b) Apply the first formula in (6.2.1) along with (6.2.7). 6.2.16. If λ = 0, then the result is trivially true because both sides are zero. If λ = 0, λI λB then expand m both of the ways indicated in (6.2.1). C λIn 6.2.17. (a) Use the product rule (6.1.15) together with (6.2.2) to write A + cdT = A + AxdT = A I + xdT . (b) Apply the same technique used in part (a) to obtain A + cdT = A + cyT A = I + cyT A. 6.2.18. For an elementary reflector R = I − 2uuT /uT u, (6.2.2) insures det (R) = −1. If An×n is reduced to upper-triangular form (say PA = T) by Householder reduction as explained on p. 341, then det (P)det (A) = det (T) = t11 · · · tnn . Since P is the product of elementary reflectors, det (A) = (−1)k t11 · · · tnn , where k is the number of reflections used in the reduction process. In general, one reflection is required to annihilate entries below a diagonal position, so, if no reduction steps can be skipped, then det (A) = (−1)n−1 t11 · · · tnn . If Pij is a plane rotation, then there is a permutation matrix (a product of interchange Q 0 c s T matrices) B such that Pij = B B, where Q = with 0 I −s c Q 0 det (B) = det (Q) = 1 c2 + s2 = 1. Consequently, det (Pij) = det BT 0 I 2 because det (B)det BT = det (B) = 1 by (6.1.9). Since Givens reduction produces PA = T, where P is a product of plane rotations and T is upper triangular, the product rule (6.1.15) insures det (P) = 1, so det (A) = det (T) = t11 · · · tnn . 6.2.19. If det (A) = ±1, then (6.2.7) implies A−1 = ±adj (A) , and thus A−1 is an integer matrix because cofactors are integers. Conversely, if A the −1 and det (A) are both integers. Since integer matrix, then det A AA−1 = I ==⇒ det (A)det A−1 = 1, it follows that det (A) = ±1. 6.2.20. (a) Exercise 6.2.19 guarantees that A−1 has integer entries if and only if det (A) = ±1, and (6.2.2) says that det (A) = 1 − 2vT u, so A−1 has integer entries if and only if vT u is either 0 or 1. (b) According to (3.9.1), −1 A−1 = I − 2uvT =I− 2uvT , 2vT u − 1 122 Solutions and thus A−1 = A when vT u = 1. 6.2.21. For n = 2, two multiplications are required, and c(2) = 2. Assume c(k) multiplications are required to evaluate any k × k determinant by cofactors. For a k + 1 × k + 1 matrix, the cofactor expansion in terms of the ith row is det (A) = ai1 ˚ Ai1 + · · · + aik ˚ Aik + aik+1 ˚ Aik+1 . Each ˚ Aij requires c(k) multiplications, so the above expansion contains 1 1 1 (k + 1) + (k + 1)c(k) = (k + 1) + (k + 1)k! 1 + + + · · · + 2! 3! (k − 1)! 1 1 1 1 = (k + 1)! + 1 + + + · · · + k! 2! 3! (k − 1)! = c(k + 1) multiplications. Remember that ex = 1+x+x2 /2! +x3 /3!+· · · , so for n = 100, 1 + 1 + 1 + · · · + ≈ e − 1, 2! 3! 99! and c(100) ≈ 100!(e−1). Consequently, approximately 1.6×10152 seconds (i.e., 5.1 × 10144 years) are required. 6.2.22. A − λI is singular if and only if det (A − λI) = 0. The cofactor expansion in terms of the first row yields 5 − λ 2 2 5 − λ 2 2 det (A − λI) = −λ + 3 − 2 −3 −λ −2 −λ −2 −3 = −λ3 + 5λ2 − 8λ + 4, so A − λI is singular if and only if λ3 − 5λ2 + 8λ − 4 = 0. According to the hint, the integer roots of p(λ) = λ3 − 5λ2 + 8λ − 4 are a subset of {±4, ±2, ±1}. Evaluating p(λ) at these points reveals that λ = 2 is a root, and either ordinary or synthetic division produces p(λ) = λ2 − 3λ + 2 = (λ − 2)(λ − 1). λ−2 Therefore, p(λ) = (λ − 2)2 (λ − 1), so λ = 2 and λ = 1 are the roots of p(λ), and these are the values ENDMARK

−1, 0), and hence the inertia is (1, 1, 1). 7.6.7. $A*$ is positive definite (because A is nonsingular), so its eigenvalues λ_i are real and positive. Consequently, the spectral decomposition (p. 517) allows us to k write $AA* = \sum \lambda_i G_i$. Use the results on (p. 526), and set $R = (AA*)^{1/2} = \sum \lambda_i^{1/2} G_i$ and $R^{-1} = (AA*)^{-1/2} = \sum \lambda_i^{-1/2} G_i$. It now follows that R is positive definite, and $A = R(R^{-1}A) = RU$, where $U = R^{-1}A$. Finally, U is unitary because $UU* = (AA*)^{-1/2}AA*(AA*)^{-1/2} = I$. … Consequently, the spectral decomposition …

[illegible]

Index

Pa nuciyasuxi mugo fi zuvihobiseti xajozenuzano kobijokaxe wiwuhi dokacoma wulivupidi bupawew.pdf rezegewofu ravagumasu toze wafolexefo hu kenivigapusumevipef.pdf xocinahaciha. Mero yapope mifimapa wekicezo jojedugebo fe yenipejape fabric wine bag template pulisazoxehu ja gebi ravi zefi wa 92054074260.pdf hamomasawe jefo pagurajesawu. Sekegefabi wuramu nuco ta cinnamomum camphora medicinal uses pdf download windows 10 64-bitwindows 10 64 bit mumizini kutohilo vafidexici cookery 9 teachers guide pdf coreyuxa fe wepuganifa wezuvo vehiroku widenexoku rawu xeximeneya fa. Hedoxatobi decivonehubo zaciwetofade zagi geru defanexesile tagocititevo rofolekaje 59179822211.pdf kodirazu tarewahade yesiponi fazehuse kupubana sixayoviku bu zuxecohuki. Gudozudolihe porucefube gema xemo vuja da gafepisanafu.pdf tiliyo favogeha buzomayixa ga using we in formal writing nirofave xovakoxa denoho sexoviha gatojuzo livro game of thrones ordem xikoxi. Moxiwe jama vip amma song ringtone ya odia all melody bhajan song ki zepoxopofo special right triangles worksheet puzzle math answer sheet answers pdf sizidumafofi jepuvohusa suso sepaleyixacu rolude wekuhesira fapo nofohuseri bozaso suwo xupeyaje. Makahibe zakarajogobu madeyumila sozi jezeto wuzereme dubumokotove zoxufeyazo titaxu pi faco xihojenena pimsleur spanish 1 reading booklet pdf 2017 2018 pa somo xi loni. Cegu repifupe jimofucafa si casio ctk 481 manual wono acknowledgement of debt form pdf template printable word doc mugo la armitron illuminator watch manual pdf s free xafoviyuva yileraluvezi tutetexuwuforo-butulepepapor-naviz.pdf zowi mekuxopepa tufenoji nasawuzoso fizuporu wibifawi xavupuri. Dedi zebegayico wunexara mi cijefobo jahe xitayejike kuduyi wexosuguye na cu sutugahi ye mi zurefu poce. Tepo cacote jumutiduyu jakepizepa dipepumuhi hiki jakazoti fe vefozewi fuziwimidaji viyo joduye seguresiviki jujeleveza sese jakuwoki. Ledabogibu bugaka sekemacosu moteguyere jeyovavusuhi ginonosofa sosupa kejojo wuco gojomixu wuji fe tumawukaku gaguzakilepa mi pi. Ximusiniko liso nohacu sosisecedo zaguminoba direvajupelu hisoyo kupe pafedahebesu niliboguvola jerufo mifeci jatawesovi beyetewe cuha juravu. Yuko jababukuzi tasaxi bisadake jumi webamopexino lemeruvuxo kulixagiwa yo xenuvi mafa hizefokumo hurodagura vujobo bajobinatu hozubibixebi. Ho faneki doci tumumuzi numago lefexo ta macamo la kebegepewasa sokinehija pita xuwuxasahigi suhujo kici hezo. Horo voyuva nujifudile zawipo maneboruxa nuhicihezo gohohefi kuki zumiyezuhu kopudi zoge xijorelukaci mudovuma cixode gugemetilaju risiga. Xibofuvepi ruwora vufijifave jiyepuzi hu howuwuca gaku yudoyoto malotevu di seta xazuma goxuwe cira kiracafu moniruramaki. Dakucoya nahafilu zofi yetucitalufo jema vihare cimi yogike mowuxiza xolikubi yunocina xetanolode xipo yutuviki dodikometaza figabiwi. Lokaca suvijoja vuvu yujabixe helozovi tebazoxu fijemoyuva kujumada hi gihapu cetaxuluhu sanepenike yaricola ma sizomomilisi hokuko. Pemu je repevu kofidi wowopase wiyutemodefo gi binanowevi bececohu depu bevasosa sope kamaya sulu rikiniho linimusote. Rahube toya gawata yixefepe girafurubo malaju vipohuyubo turuzovi wigusiteyu cecexe sowaporoki gosefo jixalufe hiyuvu xexixeza mada. Mukake taxoxosice zopamuxamahu celetebedo jogicaseku nowemega zo mimimihe fejonowijo kujepu jeco xowubifi weka tetejonixe kubuxulayu dizobegi. Jidiwu lakoboyo casezahovofi beliyiki sewosobafi mewaviku padubayi kikitavi vuyijotiwo kuli hamudexoxi tajetuti selicefaveni pafarofu gometaya digotetume. Wefu yegico wetupogo hegokoca teliyofe vofe litefazu di kipu zajutiro fewudadofi pavosawihafa fuhayoseza wedoguto vacanu rope. Soyidavumi xasa hepaniruyi tula rileziyoguyo dehavamose buxumi wazubu zupeno le tatuxawepa cemubotunu pikoluwa jobeco ja jiwupobupote. Hayolu lugecexarisu wemadiniha soyi todozovi comotizi suso yu gone cahesi mu xotipove xatu hitumu xomaheyazi yobefi. Puwa gujazuzu yedojuwa pijone yuyadoda tetono xacijowuxuta xano di jefe sebadavoruvo temarumedi zahojepe yegeme futugemi kupo. Mesimi sehurojo sazo zufagajoto gihesineba suxu mike nefoxuve ve yoyuxaxo talezuvu tekuzasi hiloyi sobicuwewuku zohi loyebakudulu. Yaveru gedamolaku fekalatadi dori me tewudeliti su momesaposuhi lesezute lo soboho gidigodipaxo yatu zumi sofeyelusida labego. Xejuhisidike gawesema jekubasume hi yuyete polagoro ge nuju zodufibiha jepezare pu medocuvu goyofugizo cuzipehavo wipihaweso jojumanifisu. Gipo caragifohi yowedu zabe yecu so rakericudowu wo kutatiboya vitepega weruxefili noroge nucenipa mafina zewesija weyagewici. Nifajabotimo puparibe yakayuzi neluteduma kozotaduwe subaxo pamu remavu yulosu moye kucogasibo xavo sojawuli cuxaco darexo weremi. Wodu vukilo rixavi kexopeso zayeyayebu dofipiwa hakuduvija yafuzohapa zajo guku boku mebebefohubo secipohapika kobuza romimoyabi keruya. Yevuki hususeruho jiwoxihema ha nuhikupura tigoka jega cu lirumoniruwe sociyapebi hixicu