I. Eigenvectors and eigenvalues

This is the second half of a two quarter sequence in linear algebra. The default syllabus for the first course in this sequence is available online from the course directory

`www.math.ucr.edu/~res/math132`

in the files `math131.*`, where the $*$ is either `ps` in PostScript format or `pdf` in Acrobat **pdf** format.

There is another linear algebra course (Mathematics 113) that is an alternate prerequisite for this course. Given that 113 is a 5 credit hour course and 131 is a 4 credit hour course, it is probably not surprising that there is some overlap in material between 113 and 131. A default syllabus for this alternate prerequisite is available online from the course directory (see above) in the files `math113.*`, where the $*$ is either `ps` in PostScript format or `pdf` in Acrobat **pdf** format.

Set-theoretic background material

All upper level mathematics courses use the language of set theory in some form. The philosophy of the course and these notes is to use the minimum amount necessary to get things done efficiently, but some ideas simply cannot be avoided. A few sections at the beginning of the current text for the Department's course on Discrete Mathematics (Mathematics 112) should contain nearly everything that is needed. The name of the text is *Discrete Mathematics and its Applications, Fifth Edition*, by K. H. Rosen (ISBN 0-072-42434-6), and the relevant sections on logic, set theory and functions are 1.6–1.8.

Whenever one works with finite sets in the mathematical sciences it is almost certain that the proofs of some statements will use a logical idea known as (finite) *mathematical induction.* For example, this is precisely what one needs to prove that the sum of the first n odd positive integers is equal to n^2. Everything that we shall need about mathematical induction is contained in Appendix A of the text.

It is important to note that mathematical induction is much different from the notion of "inductive reasoning." The latter involves concluding that something is always true based upon experience with some number of specific cases, while the former is a method for proving a sequence of statements by showing that the first one is true and for all N the truth of Statement N implies the truth of Statement $N + 1$ (so since the first statement is true the second is also true, and the latter in turn implies that the third statement is true, *etc.*).

Overview and starting point for the course

Most if not all the material in this course involves the following two basic types of questions:

1. *Questions about the fundamental properties and some basic uses of dot products (= inner products) and similar operations on vectors.*

2. *Questions about the extent to which one can simplify matrices, linear transformations and closely related objects by changing coordinates or bases.*

This first unit will begin with a discussion of some material from the preceding courses for review and transitional purposes, and it will proceed to discuss some key points involving questions of the second type. One obvious motivation of studying such questions is to make such objects easier to work with, both conceptually and computationally.

I.A : Review topics on matrices

(Fraleigh and Beauregard, §§1.5, 1.6, 4.2, 4.3)

Since this is a review of prerequisite material rather than a discussion of new material, for the most part we shall simply state the main points and give references to the course text. However, at certain points we shall also reformulate some of the basic results from the prerequisites in a manner that is necessary or useful for our purposes.

Given a matrix A with entries in, say, the real numbers, perhaps the most fundamental information about A concerns the numbers m and n of rows and columns respectively. The square matrices, for which $m = n$, are an important special class of examples because the product of two $n \times n$ matrices is again an $n \times n$ matrix. The $n \times n$ identity matrix (see page 41 of the text) is a particularly significant $n \times n$ matrix because the matrix product satisfies $A\,I = A = I\,A$ for all $n \times n$ matrices A. Although $A\,I = I\,A$, the product of two matrices often depends upon the order of multiplication, so that $A\,B$ is not necessarily equal to $B\,A$ (see page 40 of the text).

First courses in linear algebra devote much attention to a special class of $n \times n$ matrices known as *invertible matrices*. On page 75 of the text, a square matrix A is defined to be invertible if and only if there is an $n \times n$ matrix B such that $A\,B = I = B\,A$, and on page 74–75 of the text there is a proof that at most one such matrix B exists. This matrix is called the *inverse* of A; as noted on page 75 of the text, mathematicians and others frequently denote this inverse by A^{-1} but almost never by $1/A$ when working with square matrices having at least two rows and columns.

The following result appears on page 81 of the text.

THEOREM. *An $n \times n$ matrix A is invertible if and only if for every $n \times 1$ matrix (or column vector) $\mathbf{b}$ the matrix equation $A\,\mathbf{x} = \mathbf{b}$ has a solution.*■

In fact, by Exercise 27 on page 85 the equation has a *unique* solution if A is invertible.■

The uniqueness statement in the preceding sentence reflects another important characterization of invertible matrices. Since the proof uses a special case of a basic result contained in Theorem 2.3 on page 133 of the text, we shall first state the portion of the latter that we need.

PROPOSITION. *If W is a k-dimensional subspace of, say, $\mathbf{R}^n$ and S is a set of k linearly independent vectors in W, then S is a basis for W.*■

We can now formulate and prove the desired characterization of invertible matrices.

THEOREM. *An $n \times n$ matrix A is invertible if the only solution of the matrix equation $A \times \mathbf{x} = \mathbf{0}$ is the trivial solution $\mathbf{x} = \mathbf{0}$.*

Proof. If A is invertible, then the previously cited exercise shows that the matrix equation has only one solution. Since $\mathbf{x} = \mathbf{0}$ is readily seen to be a solution, it follows that it is the only solution.

Conversely, suppose that there are no nontrivial solutions to the the homogeneous matrix equation $A \times \mathbf{x} = \mathbf{0}$. We claim that for each (or column vector) $\mathbf{b}$ the matrix equation $A\mathbf{x} = \mathbf{b}$ has at most one solution. To see this, suppose that $\mathbf{y}_1$ and $\mathbf{y}_2$ solve such an equation. Then their difference $\mathbf{y}_1 - \mathbf{y}_2$ solves the corresponding reduced equation $A \times \mathbf{x} = \mathbf{0}$, so by the assumption on A we must have $\mathbf{y}_1 - \mathbf{y}_2 = \mathbf{0}$ or equivalently $\mathbf{y}_1 = \mathbf{y}_2$. Theorem 1.17 on page 96 of the text now implies that the columns of A are linearly independent, and since there are n of these columns they must form a basis for the space of all $n \times 1$ column vectors (more discussion of this point appears below). By Theorem 1.12 on pages 81–82, it follows that A is invertible.■

I.B : Review topics on linear transformations

(Fraleigh and Beauregard, §§2.3, 3.4, 4.4)

Vectors provide a means for studying geometric and physical problems using algebraic methods, so it should not be surprising that many constructions in matrix algebra have geometric interpretations. In particular, if we are given an $m \times n$ matrix A, then A may be viewed as defining a linear transformation from the space of $n \times 1$ column vectors to the space of $m \times 1$ column vectors by the formula

$$\mathcal{L}_A(\mathbf{x}) \quad = \quad A\mathbf{x}$$

(in words, $\mathcal{L}_A$ is meant to signify left multiplication by A). This mapping has two fundamental properties:

(*i*) For all vectors $\mathbf{x}$ and $\mathbf{y}$, we have $\mathcal{L}_A(\mathbf{x}+\mathbf{y}) = \mathcal{L}_A(\mathbf{x}) + \mathcal{L}_A(\mathbf{y})$.

(*ii*) For all scalars c and vectors $\mathbf{x}$, we have $\mathcal{L}_A(c\,\mathbf{x}) = c\,\mathcal{L}_A(\mathbf{x})$.

More generally, given two vector spaces V and W, a function T from V to W is said to be a *linear transformation* if the analogs of (*i*) and (*ii*) hold with T replacing $\mathcal{L}_A$ (see Definition 2.3 on page 142 of the text for the definition when V and W are $\mathbf{R}^n$ and $\mathbf{R}^m$, and see Definition 3.9 on page 213 of the text for the general case). Pages 144-145 of the text describe some examples of linear transformations as well as some examples of functions from one vector space to another that are **NOT** linear transformations.

Under the standard identifications of $\mathbf{R}^n$ and $\mathbf{R}^m$ with the spaces of $n \times 1$ and $m \times 1$ column vectors, a basic result states that **every** *linear transformation from* $\mathbf{R}^n$ *to* $\mathbf{R}^m$ *has the form* $\mathcal{L}_A$ *for some unique* $m \times n$ *matrix* A. A reference for this in the text is the Corollary on page 146. One basic identity involving the correspondence between linear transformations and matrices from the Corollary on page 146 is given in the boxed statement on page 150 of the text. In order to state two others we need to define some fundamental concepts that are not presented in the text.

Definition. Let V and W be vector spaces. Given two linear transformations S and T from V to W, their *sum* $S+T$ is defined by the formula

$$[S+T]\,(\mathbf{v}) \quad = \quad S(\mathbf{v}) \ + \ T(\mathbf{v}) \ .$$

Given a scalar c and T as above, their *scalar product* cT is defined by the formula

$$[cT]\,(\mathbf{v}) \quad = \quad c\,T(\mathbf{v}) \ .$$

Before proceeding, we should verify that these constructions yield linear transformations.

PROPOSITION. *If S, T and c are given as above, then $S+T$ and cT define linear transformations from V to W.*

Proof. We first verify that

$$\left[S+T\right](\mathbf{x} + \mathbf{y}) = \left[S+T\right](\mathbf{x}) + \left[S+T\right](\mathbf{y})$$

by noting that the left hand side is equal to

$$\left[S+T\right](\mathbf{x} + \mathbf{y}) = \Big(S(\mathbf{x} + \mathbf{y})\Big) + \Big(T(\mathbf{x} + \mathbf{y})\Big)$$

which is equal to

$$S(\mathbf{x}) + S(\mathbf{y}) + T(\mathbf{x}) + T(\mathbf{y})$$

because S and T are linear. The latter in turn is equal to

$$S(\mathbf{x}) + T(\mathbf{x}) + S(\mathbf{y}) + T(\mathbf{y})$$

which is equal to

$$\left[S+T\right](\mathbf{x}) + \left[S+T\right](\mathbf{y})$$

and hence $S+T$ satisfies the first property for a linear transformation.

We now show that $S+T$ satisfies the second property. By definition we have

$$\left[S+T\right](c\,\mathbf{x}) = S(c\,\mathbf{x}) + T(c\,\mathbf{x})$$

and by the linearity conditions the right hand side is equal to

$$c\,S(\mathbf{x}) + c\,T(\mathbf{x}) = c\big(S(\mathbf{x}) + T(\mathbf{x})\big) = \big[c\,(S+T)\big](\mathbf{x})\ .$$

This completes the proof that $S+T$ is a linear transformation.

Turning to cT, we first show that

$$\left[cT\right](\mathbf{x} + \mathbf{y}) = \left[cT\right](\mathbf{x}) + \left[cT\right](\mathbf{y})$$

by noting that the left hand side is equal to

$$c\Big(T(\mathbf{x} + \mathbf{y})\Big)$$

which is equal to

$$c\Big(T(\mathbf{x}) + T(\mathbf{y})\Big) = c\,T(\mathbf{x}) + c\,T(\mathbf{y})$$

because T is linear. The latter in turn is equal to

$$\left[cT\right](\mathbf{x}) + \left[cT\right](\mathbf{y})$$

and therefore we have verified the first of the conditions for cT to be a linear transformation.

To see that cT satisfies the second property, note that

$$\big[cT\big](a\,\mathbf{x}) \quad = \quad c\,\big(T(a\,\mathbf{x})\big) \quad = \quad ca\,T(\mathbf{x} \quad =$$

$$ac\,T(\mathbf{x} \quad = \quad a\,\big(c\,T(\mathbf{x})\big) \quad = \quad a\,[c\,T]\,(\mathbf{x})$$

and therefore cT satisfies the seond defining property for a linear transformation.■

We shall next verify that our definitions of addition and scalar multiplication for linear transformations are compatible with the definitions for matrices.

THEOREM. *Let $V = \mathbf{R}^n$ viewed as the space of $n \times 1$ column vectors, Let $W = \mathbf{R}^m$ viewed as the space of $m \times 1$ row vectors, let S and T be linear transformations from V to W, and let c be a scalar. If A and B are the $m \times n$ matrices corresponding to T and S as in the text, then the matrices corresponding to $S + T$ and cT are equal to $A + B$ and $c\,A$ respectively.*

Proof. Let $\{\,\mathbf{e}_j\,\}$ and $\{\,\mathbf{e}'_i\,\}$ denote the standard unit vector bases for $\mathbf{R}^n$ and $\mathbf{R}^m$ respectively. According to the description of associated matrices in the Corollary on page 146 of the text, the entries $A_{i,j}$ of the matrix A associated to the linear transformation T are given by the following formula:

$$T(\mathbf{e}_j) \quad = \quad \sum_{i=1}^{m} a_{i,j}\,\mathbf{e}'_i$$

A similar formula holds for the matrix B associated to S:

$$S(\mathbf{e}_j) \quad = \quad \sum_{i=1}^{m} b_{i,j}\,\mathbf{e}'_i$$

Of course, the same sort of identity holds for the matrix C associated to the sum $S + T$, and the problem is to express each entry of C in terms of the entries of A and B. But the definition of the sum of two linear transformations yields the following equations:

$$\big[S+T\big](\mathbf{e}_j)) \quad = \quad S(\mathbf{e}_j) \;+\; T(\mathbf{e}_j) \quad =$$

$$\sum_{i=1}^{m} b_{i,j}\,\mathbf{e}'_i \;+\; \sum_{i=1}^{m} a_{i,j}\,\mathbf{e}'_i \quad = \quad \sum_{i=1}^{m} (a_{i,j} + b_{i,j})\;\mathbf{e}'_i$$

It follows that the matrix corresponding to $S + T$ is $A + B$. Similarly, the definition of the scalar product of a linear transformation and a scalar yields the following equations:

$$\big[cT\big](\mathbf{e}_j)) \quad = \quad c\,T(\mathbf{e}_j) \quad = \quad c\left(\sum_{i=1}^{m} a_{i,j}\,\mathbf{e}'_i\right) \quad + \quad \sum_{i=1}^{m} c\,a_{i,j}\,\mathbf{e}'_i$$

These imply that the matrix corresponding to cT is $c\,A$.■

Invertibility and linear transformations

Given a function f defined on a set X and taking values in another set Y, there is an inverse function f^{-1} going from Y to X defined by the rule

$$x \quad = \quad f^{-1}(y) \quad \Longleftrightarrow \quad y \quad = \quad f(x)$$

if and only if f maps X onto Y in a one-to-one fashion. In the case of linear transformations, we have the following important fact that appears as Theorem 3.8 on page 220 of the text:

THEOREM. *If T is a linear transformation from a vector space V to a vector space W that is $1-1$ and onto, the the inverse map T^{-1} from W to V is also linear.*■

The following important observation appears on page 151 of the text:

CORRESPONDENCE OF INVERSES. *Under the $1-1$ correspondence between $m \times n$ matrices and linear transformations from $\mathbf{R}^n$ to $\mathbf{R}^m$, a linear transformation T is invertible if and only if its associated matrix A is invertible, and in this case the matrix associated to T^{-1} is equal to A^{-1}.*■

Additional examples of linear transformations

Various examples of linear transformations arising in calculus are presented in Section 3.4 of the text. Here are a few more:

(1) Let $\mathcal{F}$ be the set of all functions defined on the real line $\mathbf{R}$ and taking values in $\mathbf{R}$, and let $g \in \mathcal{F}$. Then the map T sending a function f to the product $g \cdot f$ is a linear transformation from F to itself. Verification of this will be left to the reader as an exercise. In fact, using calculus one can say more: If g is continuous, then T maps the subspace $\mathcal{C}$ of continuous functions into itself, and if g is infinitely differentiable, then T maps the subspace $\mathcal{D}_\infty$ of infinitely differentiable functions into itself. Therefore these multiplication maps can also be viewed as linear transformations from $\mathcal{C}$ or $\mathcal{D}_\infty$ to itself provided g belongs to the corresponding subspace. Verification that the map T is a linear transformation is left to the reader as an exercise, and solutions will be given in a solutions file to be posted in the course directory.

(2) Given a function G defined on $\mathbf{R}$ and taking values in $\mathbf{R}$, a second linear transformation from $\mathcal{F}$ to itself is defined by the map T sending f to the composite $f \circ g$, one particular example of this type would be given by the formula

$$\big[Tf\big](x) \quad = \quad f\left(x^3 - x + 1\right) \ .$$

Once again, standard results from calculus imply that T maps $\mathcal{C}$ into itself if g is continuous and T maps $\mathcal{D}_\infty$ to itself if g is infinitely differentiable (this argument is harder). As for the preceding example, detailed justifications appear in the solutions file for this unit in the course directory).

(3) Let $k(u,v)$ be a function with continuous partial derivatives everywhere on the plane $\mathbf{R}^2$, and consider the transformation from $\mathcal{C}$ to $\mathcal{F}$ defined by the formula

$$\big[Tf\big](x) \quad = \quad \int_0^1 \ k(x,y)\, f(y)\, dy \ .$$

This can be viewed as a continuous analog of the usual formula for linear transformations on $\mathbf{R}^n$ in terms of matrices. in fact, the methods of calculus imply that the image of this linear transformation lies inside $\mathcal{C}$, but this is not needed for our purposes (however, the details appear in the solutions file for this unit in the course directory).

I.1 : Basic definitions

(Fraleigh and Beauregard, §5.1)

Given an $n \times n$ matrix A, it is often important to understand the behavior of the linear transformation $\mathcal{L}_A$ in the clearest possible conceptual terms. One simple but far-reaching question is to analyze the proper subspaces W of $\mathbf{R}^n$ such that $\mathcal{L}_A(W) \subset W$. One aspect ot this involves the existence of such subspaces, and another is the behavior of $\mathcal{L}_A$ on such A-invariant subspaces. The study of eigenvectors and eigenvalues deals with the special case of !-dimensional subspaces.

It will be useful to formulate everything in terms of vector spaces and linear transformations.

Definition. Let V be a vector space and let T be a linear transformation from V to itself. A scalar c is called an *eigenvalue* for T if there is a nonzero vector $\mathbf{x} \in V$ such that $T(\mathbf{x}) = c\,\mathbf{x}$. Every vector $\mathbf{x}$ satisfying this equation is called an *eigenvector* for T associated to the eigenvalue c.

The terms "eigenvector" and "eigenvalue" are half German and half English (sort of like the word "cheeseburger" — the city name "Anaheim" is a similar example, being half German and half Spanish). Alternate terms are described in Definition 5.1 on page 289 of the text, but the "eigen" terminology is pretty standard in mathematics as well as other disciplines.

EXAMPLES. (1) Suppose that A is an $n \times n$ diagonal matrix with diagonal entries $d_1, \cdots, d_n$. Then each unit vector $\mathbf{e}_i$ is an eigenvector, and the associated eigenvalue is d_i. If the no two diagonal entries are equal, then every eigenvector is a scalar multiple of a unit vector. In general, given a scalar λ the set of eigenvectors associated to λ is the span of the set of all unit vectors $\mathbf{e}_i$ such that $d_i = \lambda$.

(2) There are plenty of other easy examples. In particular, if A is the 2×2 matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

then 2 is an eigenvalue and [the transpose of] $(1,1)$ is an associated eigenvector, and 0 is also an eigenvalue with associated eigenvector equal to [the transpose of] $(1,-1)$. More generally, note that *if A is a square matrix that is not invertible, then 0 is an eigenvalue and every nontrivial solution to the homogeneous system of linear equations $A\,\mathbf{x} = \mathbf{0}$ is an eigenvector.* Yet another example is given at the top of page 289 of the text.

(3) In contrast, the matrix

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

which induces the linear transformation on $\mathbf{R}^2$ corresponding to multiplication by 90 degrees, does not have any real eigenvectors or eigenvalues. Geometrically this is clear because each line through the origin is sent to its perpendicular at the origin. To see this algebraically, we need to think about possible solutions to the system of equations given in vector form by

$$(\lambda\,x, \lambda\,y) \quad = \quad T\,(x,y) \quad = \quad (-y, x)\ .$$

If we eliminate y from this system we obtain an equation of the form $x = -\lambda^2\,x$ which has a solution if and only if either $x = 0$ or $\lambda^2 = -1$. Since the latter cannot happen over the real numbers, we must have $x = 0$, and substituting back for y we find that $y = -\lambda\,x = -\lambda\,0 = 0$. Thus A has no eigenvalues. Of course, if we are working over the complex numbers, then the equation $\lambda^2 = -1$ does have a solution, and when we discuss vector spaces over the complex numbers in Section IV.3 of these notes this will be extremely significant.

(3) There are numerous examples from calculus and differential equations involving eigenvalues and eigenvectors. In particular, if T is the differentiation operator on the space $\mathcal{D}_\infty$ of infinitely differentiable functions then every real number $\lambda \neq 0$ is an eigenvalue for T, and the set of eigenvectors consists of all scalar multiples of the function $\exp(\lambda\, x)$. Since the derivative of a constant function is zero, it also follows that 0 is an eigenvector, and basic results of calculus imply that the associated eigenvectors are all (nonzero) constant functions. Similarly, if S denotes double differentiation on the same space $\mathcal{D}_\infty$ then every real number α is also an eigenvalue for S. If $\alpha = 0$ then the associated eigenvectors are the nonzero first degree polynomial functions, and if $\alpha > 0$ and we write $\alpha = \lambda^2$ then the eigenvectors are the nonzero linear combinations of $\exp(\lambda\, x)$ and $\exp(-\lambda\, x)$. On the other hand, if $\alpha < 0$ and we write $\alpha = -\lambda^2$, then the eigenvectors are the nonzero linear combinations of $\sin \lambda x$ and $\cos \lambda x$.

Finding eigenvectors and eigenvalues of matrices

In order to be able to work effectively with eigenvectors and eigenvalues, one needs more than trial and error techniques or ingenuity at making educated guesses. The following result is a fundamentally important tool for finding eigenvalues and eigenvectors, and for small matrices it is also computationally effective.

THEOREM. *Given an $n \times n$ matrix A, define its* **characteristic polynomial** *by $\chi_A(t) = \det(A - tI)$. Then λ is an eigenvalue of A if and only if λ is a root of $\chi_A(t)$.*

The derivation of this is given on pages 290–291 of the text.■

In theory, the theorem reduces the computation of eigenvalues to computing the roots of the characteristic polynomial and the computation of eigenvectors to finding the solutions to the homogeneous linear systems of equations $(A - \lambda I)\mathbf{x} = \mathbf{0}$ where λ runs through all the eigenvalues of A. For 2×2 matrices with integer entries, it is easy to compute the roots of the characteristic polynomial and the solutions to the associated systems of homogeneous linear equations, and even in the 3×3 case one can carry out the computations without too much trouble if the characteristic polynomial is well behaved (in particular, if one of its roots is an integer). However, in more complicated situations one needs better techniques. For the most part these are outside the scope of this course (and are treated in courses on numerical methods), but later on we shall give a reasonably efficient method for computing $\chi_A(t)$. In this course most of the emphasis will be on conceptual issues related to eigenvalues and eigenvectors.

We have already noted that a 2×2 matrix need not have any real eigenvalues. However, it turns out that 3×3 matrices always do, and this is a special case of the following more general result:

PROPOSITION. *Let A be an $n \times n$ matrix (with real entries) where n is odd. Then A has a real eigenvalue.*

Proof. In view of the previous theorem, it is enough to show that $\chi_A(t)$ has a real root. By construction this polynomial has the form $(-1)^n t^n + g(t)$ where g is a polynomial of lower degree. Since

$$\lim_{t \to \infty} \frac{g(t)}{t^n} \quad = \quad 0$$

it follows that the limits of $\chi_A(t)$ as t approaches $\pm\infty$ are $\mp\infty$, and accordingly it follows that χ_A is negative for all sufficiently large values of t and χ_A is positive for all sufficiently small values of t. By the Intermediate Value Theorem from calculus, it follows that there is some real number λ such that $\chi_A(\lambda) = 0$.∎

In contrast to the preceding result, the characteristic polynomial of a square matrix with an even number of rows and columns can always be positive (look at the previous example of a 2×2 matrix with no real eigenvalues, where the characteristic polynomial is $t^1 + 1$).

Several important facts, examples and illustrations appear on pages 297–299 of the text.

I.2 : Diagonalization

(Fraleigh and Beauregard, §5.2)

We have already noted that a diagonal matrix has a basis of eigenvectors, generally associated to different eigenvalues. In fact, such bases often exist for matrices that are not necessarily diagonal. The following simple but far-reaching observation will be very useful in illustrating this fact.

PROPOSITION. *Suppose that $\lambda_1, \cdots, \lambda_k$ are distinct eigenvalues for the linear transformation T from V to itself, and let $\mathbf{x}_1, \cdots, \mathbf{x}_k$ be associated eigenvectors for the respective eigenvalues. Then the set $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ is linearly independent. In particular, if $\dim V = n$, then $k \leq n$.*

Proof. The final sentence follows from the rest of the result because a linearly independent subset of and n-dimensional vector space has at most n elements. To prove the linear independence of the eigenvalues, we shall proceed by induction on k. If $k = 1$ the result is true because a subset consisting of one nonzero vector is always linearly independent. Suppose the result is true for subsets of $(k-1)$ vectors, and suppose that in the equation

$$\sum_{i=1}^{k} c_i\,\mathbf{x}_i \quad = \quad \mathbf{0}$$

there is some nonzero coefficient c_ℓ. Then we may solve for $\mathbf{x}_\ell$ in terms of the remaining vectors and conclude that

$$\mathbf{x}_\ell \quad = \quad \sum_{i\neq\ell} b_i\,\mathbf{x}_i$$

where $b_i = -c_i/c_\ell$. If we apply T to both sides of the formula for $\mathbf{x}_\ell$ we obtain the equation

$$\lambda_\ell\,\mathbf{x}_\ell \quad = \quad T(\mathbf{x}_\ell) \quad = \quad T\left(\sum_{i\neq\ell} b_i\,T(\mathbf{x}_i)\right) \quad =$$

$$\sum_{i\neq\ell} b_i\,T(\mathbf{x}_i) \quad = \quad \sum_{i\neq\ell} b_i\,\lambda_i\,\mathbf{x}_i$$

and using the expression for $\mathbf{x}_\ell$ as a linear combination of the other eigenvectors we see that the left hand expression in the preceding chain of equations is equal to

$$\sum_{i\neq\ell} b_i\,\lambda_\ell\mathbf{x}_i\ .$$

Since the induction hypothesis implies that the eigenvectors aside from $\mathbf{x}_\ell$ are linearly independent, we may equate the coefficients of these vectors in both expressions for $T(\mathbf{x}_\ell)$ and conclude that

$$b_i\,\lambda_\ell \quad = \quad b_i\,\lambda_i$$

for all $i \neq \ell$. Since $\lambda_i \neq \lambda_\ell$ if $i \neq \ell$ the only way one can have such equations is if $b_i = 0$ for all i. But this means that $\mathbf{x}_\ell = \mathbf{0}$, contradicting our original assumption that $\mathbf{x}_\ell$ was a (nonzero) eigenvector. The contradiction arises from the assumption that the set of eigenvectors is linearly dependent, and therefore it follows that this set must be linearly independent.■

Here is another important observation.

THEOREM. *If A is an $n \times n$ matrix and $\chi_A(t)$ has n distinct real roots, then A has a basis of eigenvectors.*

Proof. Let $\lambda_1, \cdots, \lambda_n$ be the distinct roots of the characteristic polynomial. Then for each λ_i there is an associated nonzero eigenvector $\mathbf{x}_i$, and by the previous result the set of all these vectors is linearly independent. Since there are n such vectors and they lie in an n-dimensional vector space, this set of eigenvectors must form a basis for the space of $n \times 1$ column vectors.■

APPLICATION TO THE CASE $n = 2$.

Suppose that A is the 2×2 matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

so that $\chi_A(t) = t^2 - (a+d)t + (ad - bc)$. Then the quadratic formula implies that this polynomial has two distinct real roots if and only if $(a+d)^2 - 4(ad - bc) > 0$, while it has a double real root if and only if $(a+d)^2 - 4(ad-bc) = 0$, and it no real roots at all if and only if $(a+d)^2 - 4(ad-bc) < 0$. This should suggest that a real 2×2 matrix has distinct real roots in roughly half of all cases, and accordingly such matrices have bases of real eigenvectors about half of the time. Of course, if one also considers complex roots, then in the cases where $(a+d)^2 - 4(ad - bc) < 0$ one has two unequal complex roots, and hence the characteristic polynomial has two roots in almost every case. Similarly, for higher values of n a real polynomial of degree n has no repeated roots for "almost all" choices of coefficients. However, proving this rigorously would require a lengthy digression and therefore we shall not attempt to make this intuitive statement mathematically precise.■

EXAMPLES. If the characteristic polynomial has a repeated real root, a basis of eigenvectors does not necessarily exist. This can already be seen from the 2×2 matrices

$$A(\lambda) \quad = \quad \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

where λ is an arbitrary real number. The characteristic polynomial of $A(\lambda)$ is equal to $(t - \lambda)^2$, so that λ is the only eigenvalue of the matrix, but the associated eigenvectors are given by all nonzero multiples of the transpose of $\mathbf{e}_1$, and therefore there is no basis of eigenvectors. For larger matrices one has similar examples where the characteristic polynomial is $\lambda - t)^n$ but the associated eigenvectors are the multiples of the transpose of $\mathbf{e}_1$; in particular, if $n = 3$ one has the following basic analog of the 2×2 matrix displayed above:

$$\begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$$

A matrix is said to be *diagonalizable* if it has a basis of eigenvectors. The discussion on pages 306–312 of the text contains several important results about diagonalizable matrices. Two points that will be particularly significant in this course are the applications of diagonalization to computing powers of matrices (Corollary 2 on page 307 of the text) and the concept of similarity defined on page 310 of the text. Theorem 3.5 on page 314 of the text states a fundamental result that will be proved and applied later in this course.

I.3 : Differential and difference equations

(Fraleigh and Beauregard, §5.3)

As noted on page 289 of the text, if an $n \times n$ matrix has a basis of eigenvectors, then one can use this basis to describe vector sequences of the form $A^k\,\mathbf{x}$ without large amounts of explicit computation. Specifically, if we have a basis of eigenvectors $\mathbf{v}_i$ with assopciated eigenvalues λ_i and we write an arbitrary vector $\mathbf{x}$ as a linear combination $\sum_i\ c_i\,\mathbf{v}_i$, then we have the formula

$$A^k\,\mathbf{x} \;=\; \sum_{i=1}^{n} \lambda_i^k\, c_i\,\mathbf{v}_i\ .$$

The text continues with one reason for interest in such sequences; namely, questions about the limiting behavior of the sequence $\{\, A^k\mathbf{x}\,\}$ as $k \to \infty$. In particular, the text notes that

$$\lim_{k\to\infty}\ \left|A^k\mathbf{x}\right| \;=\; \infty$$

if $\mathbf{x}$ is a nontrivial linear combination of eigenvectors such that at least one has an eigenvalue λ_i satisfying $|\lambda_i| > 1$, while the limit of the sequence is zero if $\mathbf{x}$ is a linear combination of eigenvectors whose associated eigenvalues all have absolute values less than 1. Section 1.7 of the text discusses the relevance of analyzing sequences of the form $A^k\,\mathbf{x}$ to problems about the evolution of population distributions over long periods of time. For the "good" examples of this type, there is a nonzero vector $\mathbf{v}$, that is unique up to multiplication by a nonzero scalar, such that for each $\mathbf{x}$ the limit of the sequence of vectors $A^k\,\mathbf{x}$ is a scalar multiple of A, and the vector $\mathbf{v}$ is an eigenvector associated to the eigenvalue $+1$. A derivation of this fact will be given later in the course.

Applications to homogeneous finite difference equations

On pages 286–288 of the text an analysis of the sequence of vectors $A^k\,\mathbf{x}$ in a special case is used to give a closed formula for the *Fibonacci numbers* that are defined recursively by the rules $F_0 = F_1 = 1$ and $F_{n+2} = F_{n+1} + F_n$ for $n \geq 0$. The reformulation in terms of sequences having the form $A^k\,\mathbf{x}$ arises by making the definitions

$$\mathbf{v}_0 \;=\; \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad , \quad \mathbf{v}_n \;=\; \begin{pmatrix} F_{n+1} \\ F_n \end{pmatrix} \quad , \quad A \;=\; \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

so that the recursive equation is equivalent to the matrix equation $\mathbf{v}_{n+1} = A\,\mathbf{v}_n$, which in turn yields the matrix formula $\mathbf{v}_n = A^n\,\mathbf{v}_0$. Without repeating the relatively messy details of the computation for the explicit formula, we shall note the main steps:

[1] Show that A has two distinct real eigenvalues λ_1 and λ_2.

[2] Find the associated eigenvectors $\mathbf{x}_1$ and $\mathbf{x}_2$.

[3] Express $\mathbf{v}_0$ as a linear combination of these vectors, say $\mathbf{x} = c_1\,\mathbf{x}_1 + c_2\,\mathbf{x}_2$.

[4] Substitute all the computations from the previous steps into the formula $\mathbf{v}_n = A^n\,\mathbf{v}_0 = \lambda_1^n\,c_1\mathbf{x}_1 + \lambda_2^n\,c_2\mathbf{x}_2$.

[5] Take the second coordinate of $\mathbf{v}_n$ to obtain the formula for F_n.

The explicit computation is carried out for this example on pages 319–320 of the text.

This five step process applies equally well to many other finite difference equations of the form

$$x_{n+2} \quad = \quad b\,x_{n+1} \; + \; c\,\mathbf{x}_n$$

provided we are given the values for $\mathbf{x}_0$ and $\mathbf{x}_1$. The main changes are (i) the obvious need to change the initial vector $\mathbf{v}_0$, and (ii) the matrix A is replaced by the matrix

$$\begin{pmatrix} b & c \\ 1 & 0 \end{pmatrix}$$

and in order to have two distinct real eigenvalues we need to have $b^2 + 4c > 0$.

An important nonhomogeneous difference equation

One can also use matrices to analyze the nonhomogeneous difference equation

$$x_{n+1} \quad = \quad a\,x_n \; + \; b$$

where a and b are constants and $x_0 = C$. This equation is particularly noteworthy because of its relation to amortized loans. The idea for such a loan is that an amount S of money is lent at time 0, the periodic interest rate between payment periods is $r \times 100\%$ (for monthly payments, this is **1/12** of the annual interest rate!), and the periodic payment is P. If x_n denotes the balance after the n^{th} periodic payment, then these balances satisfy the difference equation

$$x_{n+1} \quad = \quad (1+r)\,x_n \; - \; P\ .$$

We can translate the general nonhomogeneous difference equation into a matrix equation by setting

$$\mathbf{y}_0 \quad = \quad \begin{pmatrix} C \\ 1 \end{pmatrix} \quad , \quad \mathbf{y}_n \quad = \quad \begin{pmatrix} x_n \\ 1 \end{pmatrix} \quad , \quad A \quad = \quad \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$$

and as before the solution of the difference equation is given by the formula $\mathbf{y}_n = A^n\,\mathbf{y}_0$. If $a \neq 1$ then the characteristic polynomial has distinct real roots (namely, a and 1), so we can use the previous five step method to find an explicit formula for x_n.

APPLICATION TO AMORTIZED LOANS. Usually one wants to compute the periodic payment necessary to pay off the loan after some fixed period of time consisting of N payment periods; in other words, one wants to choose P so that $x_N = 0$. We shall work this out explicitly by computing the explicit formula for x_n and then solving the equation $x_N = 0$ to find P.

In this case we have $x_0 = S$, $a = 1 + r$ and $b = -P$. Since $r > 0$ we also know that $a \neq 1$ so the eigenvalues of the matrix are $1 + r$ and 1. Direct computation shows that the associated eigenvectors are

$$\mathbf{x}_1 \quad = \quad \begin{pmatrix} P \\ r \end{pmatrix} \quad , \quad \mathbf{x}_{1+r} \quad = \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}\ .$$

Straightforward computation then shows that

$$\mathbf{y}_0 \quad = \quad \frac{1}{r}\,\mathbf{x}_1 \;+\; \left[S - \frac{P}{r}\right]\,\mathbf{x}_{1+r}$$

so that we have

$$\mathbf{y_n} \quad = \quad \frac{1}{r}\begin{pmatrix} P \\ r \end{pmatrix} \;+\; (1+r)^n \left[S - \frac{P}{r}\right]\begin{pmatrix} 1 \\ 0 \end{pmatrix} \; .$$

The equation $x_N = 0$ then implies that

$$\frac{P}{r} \;+\; (1+r)^N \left[S - \frac{P}{r}\right] \quad = \quad 0$$

and if we solve this for P we obtain the formula

$$P \quad = \quad \frac{r\,S\,(1+r)^N}{(1+r)^N - 1} \; .$$

When using this formula in practice, it is important to remember that N represents the number of **months** and r corresponds to the monthly interest rate.

There is a homework exercise on comparing the payments with a fixed rate of interest over 15, 20, 25 and 30 years.

Systems of linear differential equations

Calculus courses and other lower level undergraduate science courses emphasize the importance of the first order linear differential equation $y' = \lambda\,y$ and its relevance to a wide range of problems. In more complicated situations one also confronts corresponding systems of linear differential equations having the form

$$y_i'(t) \quad = \quad \sum_{i=0}^{n} \; a_{i,j}\,y_j(t)$$

where 1 ranges from 1 to n for some n. If the matrix of coefficients $A = (\,a_{i,j}\,)$ is diagonalizable, then we shall indicate a describe a simple method for writing down the solution to such a system. Further examples will be considered later in the course.

It is generally convenient to write systems as above in the matrix form

$$Y'(x) \quad = \quad A\,Y(x)$$

which looks very much like the differential equation $y' = \lambda\,y$. The main result of this section states that solutions of the system also resemble solutions of the single equation very closely.

THEOREM. *Suppose that the matrix A has a basis of eigenvectors $\mathbf{v}_i$ with corresponding eigenvalues λ_i, let $\mathbf{w}$ be an $n \times 1$ column vector, and suppose that $\mathbf{w} = \sum_i \; c_i\,\mathbf{v}_i$. Then a solution of the system of differential equations $Y'(x) = A\,Y(x)$ with $Y(0) = \mathbf{w}$ is given by the function*

$$Y(x) \quad = \quad \sum_{i=1}^{n} \exp(\lambda_i\,x)\,c_i\,\mathbf{v}_i$$

Sketch of proof. This is simply a routine verification that $Y'(x)$ is equal to $A\,Y(x)$ and $Y(0) = \mathbf{w}$. In fact, it turns out that this gives the **unique** solution whose value at 0 is $\mathbf{w}$; a proof of this is given in the supplementary file(s) `expmatrix.*` in the course directory.■

EXAMPLE. Consider Example 1 on pages 308–309 of the text. The matrix A in this case is

$$\begin{pmatrix} -3 & 5 \\ -2 & 4 \end{pmatrix}$$

and it has eigenvalues $+2$ and -1 with associated eigenvectors

$$\mathbf{v}_{+2} \;=\; \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{v}_{-1} \;=\; \begin{pmatrix} 5 \\ 2 \end{pmatrix} .$$

If we want to find the solution to this system of differential equations with

$$Y(0) \;=\; \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

the first step is to check that $Y(0) = \mathbf{v}_{-1} - 2\,\mathbf{v}_2$, and knowing this we can write down the general solution as

$$Y(x) \;=\; e^{2x}\,\mathbf{v}_{-1} \;-\; e^{-x}\,\mathbf{v}_2 \;.$$

Physical applications of first order linear systems

It is natural to ask for examples where systems of linear differential equations arise in physics and engineering. One of the simplest illustrations come from population biology and *competing species* problems. In this model, we may suppose that two species are competing for the same food and living space, so that population increases in either species interfere with population increases in the other. If one tries to model the populations of both species as functions of time (say measured in years) with an equation of the form $Y' = A\,Y$, then one expects that the off-diagonal entries $a_{2,1}$ and $a_{1,2}$ will be negative. To illustrate this further we shall consider the example with

$$A \;=\; \begin{pmatrix} 3 & -1 \\ -2 & 2 \end{pmatrix}$$

for which the general solution of the system turns out to be as follows:

$$Y(t) \;=\; y_1(0)\,e^t \begin{pmatrix} 1 \\ 2 \end{pmatrix} \;+\; y_2(0)\,e^{4t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Since a negative population of a species is not possible, it is reasonable to require that the coordinates of $Y(0)$ should both be positive. Furthermore, this model will break down for a particular choice of $Y(0)$ if and when one reaches a time t_0 such that one of the coordinates of $Y(t_0)$ is zero. For example, if $y_1(0) = 90$ and $y_2(0) = 150$, the second coordinate will vanish when $t = \frac{1}{3}\log(16) \approx 0.92$, and this means that the second species will die out after about 11 months even though there were more of this species at the beginning. On the other hand, if there are more than twice as many of the second species than the first at the beginning, then the first species will eventually be eliminated, and if there are exactly twice as many of the second species as the first when $t = 0$ then neither species will be eliminated.

Similarly, one can also consider *predator-prey* situations. In such cases one species eats the other, and for such a model the matrix A should have positive diagonal entries and also satisfy $a_{1,2} < 0 < a_{2,1}$. Since it is more difficult to construct good but simple examples with the mathematical tools developed to this point, we shall not give examples here (however, one will be given in Section IV.5 of these notes).

Physical applications and higher order systems

It turns out that some of the most easily explained examples of systems of linear differential equations in the sciences and engineering involve systems of higher order linear differential equations, so we shall begin with some comments on how such systems can be analyzed using the ideas of this section. *This material will not be used subsequently and may be omitted without loss of logical continuity.*

There is a standard trick for turning a system of second or higher order differential equations into an equivalent system of first order equations. Specifically, if we are given a system of n^{th} order differential equations of the form

$$y_i^{(n)} \quad = \quad F_i(x, \; \cdots \; y_j^{(k)}, \; \cdots \;)$$

where $1 \leq i, j \leq m$ and $0 < k < n$, then this system is equivalent to the system of $m\,n$ first order differential equations

$$z'_{i,j} \quad = \quad z_{i,j+1} \qquad\qquad (0 \leq j < n-1)$$

$$z'_{i,n-1} \quad = \quad F_i(x, \; \cdots \; z_{j,k}, \; \cdots \;)$$

where j and k are as before in equations of the second type. If in addition each F_i is a linear function of the terms $y_j^{(k)}$ and does not depend upon the x variable (in the language of differential equations, the system is *autonomous*), then the new system of first order equations is also linear.

EXAMPLE. We shall illustrate how this applies to the standard second order linear differential equation $y'' + by' + cy = 0$. Since this is a particularly simple situation, instead of working with multiply subscripted functions we shall merely set y equal to itself and p equal to y'. Then the original second order equation is equivalent to the system of two first order equations given by $y' = p$ and $p' = -b\,p + c\,y$.

It is instructive to compare the solutions to this system with the known solutions of the original second order equation. The matrix form of the system of equations is given by

$$\begin{pmatrix} y' \\ p' \end{pmatrix} \quad = \quad \begin{pmatrix} 0 & 1 \\ -b & -c \end{pmatrix} \begin{pmatrix} y \\ p \end{pmatrix}$$

and the characteristic polynomial of the 2×2 matrix A in this expression is given by $t^2 + b\,t + c = 0$, which is the same polynomial that arises in the usual approach to solving the original second order differential equation, To stay within the setting of this section, let us assume that this polynomial has two distinct real roots, say λ_1 and λ_2; then the identity $t^2 + b\,t + c = (t - \lambda_1) \cdot (t - \lambda_2)$ implies that $-b = \lambda_1 + \lambda_2$ and $-c = -\lambda_1\lambda_2$. If we write out $A - \lambda_1 I$ and $A - \lambda_2 I$ using these identities, we see that eigenvectors for λ_1 and λ_2 are given respectively by

$$\mathbf{v}_1 \quad = \quad \begin{pmatrix} 1 \\ \lambda_1 \end{pmatrix} \qquad , \qquad \mathbf{v}_2 \quad = \quad \begin{pmatrix} 1 \\ \lambda_2 \end{pmatrix}$$

and since linear combinations of these vectors are solutions for the system of equations, we get solutions to the original equation of the form

$$y(x) = c_1 \, \exp(\lambda_1 \, x) + c_2 \, \exp(\lambda_2 \, x)$$

exactly as in a differential equations course. This is by no means intended to recommend the use of a first order linear system to solve the original equation, but instead it is simply meant to show that one obtains the same solutions using the associated system of two first order equations.

MULTIPLE SPRING MASS PROBLEMS. The reason for digressing to discuss higher order equations is that one of the easiest examples to discuss involves second order linear differential equations of the form $Y''(t) = A \cdot Y(t)$. One of the simplest applications of second order differential equations is to the describe the up and down motion of a mass suspended from a spring, and this is discussed in Illustration 3 on pages 298–299 of the text. The key conclusion is that the motion of the mass is given by a differential equation of the form $y'' = -\omega^2 y$ for some $\omega \neq 0$. If one considers mechanical systems that have several masses and several springs coupling them to each other and to some external framework, then one obtains a system of differential equations $Y'' = A\,Y$ as above, and by the standard trick for converting this into a system of first order equations we an rewrite this as a system of first order equations using matrices in block form as follows:

$$\begin{pmatrix} Y' \\ P' \end{pmatrix} \quad = \quad \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix} \begin{pmatrix} Y \\ P \end{pmatrix}$$

We now come to a temporary complication. Just as the auxiliary equation $m^2 + \omega^2 = 0$ for the single mass problem $y'' = -\omega^2 \, y$ does not have any real roots, in the more complicated case the large square matrix will not have any real eigenvalues in the cases that are relevant to physics and engineering. However, when we extend linear algebra to include complex scalars in Unit IV we shall find ways of overcoming this and other problems of a similar type. At this point, the purpose of the discussion is simply to illustrate one way in which systems of linear differential equations are needed and used outside of mathematics itself.

Here are some links to sites that describe the multiple spring mass problems in more detail and provide some helpful illustrations:

`http://www.efunda.com/formulae/vibrations/mdof_eom.cfm`

`http://links.math.rpi.edu/devmodules/multism/index.html`

The second site has some excellent pictures that might be particularly helpful.

II. Perpendicularity (Orthogonality)

Vectors can and should be viewed both geometrically and algebraically. One important tool for passing between these viewpoints is the *dot product* — also known as the *inner product* or *scalar product* — which provides algebraic means for working with geometric notions such as lengths or angle measurements. One particularly important fact is that two vectors are perpendicular (or *orthogonal*) if and only if their dot product vanishes. Geometric applications of dot products appear frequently in single variable and multivariable calculus courses, and much of the material of this unit may be viewed as a continuation of the study of geometrical questions using inner or dot products. It seems worth mentioning that the terms "orthogonal" and "inner product" are used far more often in mathematics than the more familiar "perpendicular" and "dot product" beyond the level of lower division undergraduate texts.

Aside from providing a means for working with lengths and angle measurements in the standard n-dimensional space $\mathbf{R}^n$, inner products are also important in connection with many topics in calculus and beyond calculus. One particular example involves the infinite series expansions of periodic functions satisfying identities like

$$f(x + 2\,k\pi) \quad = \quad f(x) \qquad \text{(for all integers } k\text{)}$$

in terms of the standard periodic trigonometric functions of the form $\sin n\,x$ and $\cos m\,x$ where m and n are nonnegative integers (by convention, $\cos mx$ is the constant function whose value is 1 at each point). Some of these implications will also be discussed in this unit.

In the 3-dimensional vector space $\mathbf{R}^3$ one also has the **cross product** which assigns to an ordered pair of vectors $(\mathbf{a}, \mathbf{b})$ a third vector $\mathbf{a} \times \mathbf{b} \in \mathbf{R}^3$. Given the relative ease in defining generalizations of the inner product and the usefulness of the cross product in mathematics and physics, it is natural to ask whether there are also generalizations of the cross product. However, it is rarely possible to define good generalizations of the cross product that satisfy most of the latter's good properties. Partial but significantly more complicated generalizations can be constructed using relatively sophisticated techniques (for example, from tensor algebra or Lie algebras), but such material goes far beyond the scope of this course. Here are two online references containing further information:

`http://www.math.niu.edu/~rusin/known-math/95/prods`

`http://www.math.niu.edu/~rusin/known-math/96/octonionic`

Needless to say, we shall not use the material in these reference subsequently.

II.A : Review topics

(Fraleigh and Beauregard, §§1.2, 3.5)

The standard inner product on $\mathbf{R}^n$ is defined in Definition 1.6 on page 24 of the text, and its basic properties are stated and discussed on pages 24–26 of the text (in particular, see Theorem

1.3 on pages 24–25); proofs of all the important properties are either included on these pages or in Theorems 1.4 and 1.5 on pages 29–30 of the text.

The abstract notion of an *inner product space* is discussed in Section 3.5 of the text. Specifically, the formal definition appears in Definition 3.12 on page 230, the fundamentally important **Schwarz inequality** (also known as the Cauchy inequality, the Buniakovsky inequality, or some hyphenated combination) appears as Theorem 3.11 on page 335, and the **Triangle inequaltiy** is stated on the same page. As noted in the text, the proof of the latter for $\mathbf{R}^n$ in Theorem 1.5 on page 30 of the text goes through unchanged for abstract inner product spaces.

A crucial reason for introducing abstract innner product spaces is the existence of many such objects besides the usual ones. Perhaps the most noteworthy is the inner product structure in Example 3 on pages 231–233 of the text. This is the example that plays an important role in studying the infinite series expansions for periodic functions that were mentioned at the beginning of this unit. Exercises 10–13 on page 236 of the text discuss some basic points regarding this inner product on the space of continuous functions. Significant variants of this inner product (so-called *weighted inner products*) are discussed in Exercises 27 on page 237 of the text.

Following standard terminology, we shall say that a set S of nonzero vectors is *orthogonal* if each pair of distinct vectors in S is orthogonal, and we shall say that S is *orthonormal* if in addition $|\mathbf{v}| = 1$ for all $\mathbf{v} \in S$.

Some additional examples

In order to keep the discussion brief we shall only consider nonstandard inner products on $\mathbf{R}^2$; there are also nonstandard inner products on $\mathbf{R}^n$ for each $n \geq 3$. Given two vectors $\mathbf{x} = (x_1 . \mathbf{x}_2)$ and $\mathbf{y} = (y_1 . \mathbf{y}_2)$ in $\mathbf{R}^2$ and a real number $r > 0$, consider the homogeneous quadratic polynomial

$$\Phi_r(\mathbf{x}, \mathbf{y}) \quad = \quad (1+r)x_1y_1 \ + \ x_2y_1 \ + \ x_1y_2 \ + \ x_2y_2 \ .$$

It is a routine exercise in algebra to check that Φ_r satisfies the first three properties of an inner product as given on page 230 of the text. Perhaps the quickest way to see positivity is to rewrite $\Phi_r(\mathbf{x}, \mathbf{x})$ as

$$r\, x_1^2 + (x_1 + x_2)^2$$

and noting that the latter is always nonnegative, with equality if and only if $x_1 = x_1 + x_2 = 0$, which is equivalent to saying that $x_1 = x_2 = 0$.

More generally, given any symmetric $n \times n$ matrix G (recall this means that G is equal to its own transpose), one can ask for conditions under which the function

$$\varphi_G(\mathbf{x}, \mathbf{y}) \quad = \quad \sum_{i,j} g_{i,j}\, x_i\, y_j$$

defines an inner product on $\mathbf{R}^n$. We shall give some easily stated and relatively computable conditions for this in the final unit of the course, and we shall apply these conditions to some basic questions in multivariable calculus.

II.1 : Orthogonal bases

(Fraleigh and Beauregard, §6.2)

Based upon our experience with two and three dimensions, it seems reasonable to expect that n-dimensional space is characterized by the existence of n separate directions, each of which is perpendicular to every other direction. For example, when one passes from two to three dimensions, one has a third direction that is perpendicular to a standard pair of perpendicular directions in the plane. Certainly $\mathbf{R}^n$ has n mutually perpendicular directions given by the n standard unit vectors, and one can ask whether this property holds in greater generality. One aspect of this was answered in Exercise 19 on page 237 of the text (from Section 3.5): *If we are given a set S of vectors such that every vector in S is nonzero and perpendicular to every other vector in S, the S is linearly independent* (in principle, the proof is contained in the proof of Theorem 6.2 on page 228 of the text). Consequently, if V is an n-dimensional inner product space and S as in the previous sentence is contained in V, then S contains at most n elements. The main result of this section will imply that one can always find such a subset with exactly n elements. In fact, there is a recursive process for constructing such a basis that is given as follows:

GRAM-SCHMIDT ORTHONORMALIZATION PROCESS. *Let $\mathcal{A} = \{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ be a set of linearly independent vectors in the inner product space V. Then there is an orthonormal set of vectors $\mathcal{U} = \{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ such that for each $k \le m$ the vector $\mathbf{u}_k$ is a linear combination of $\mathbf{v}_1, \cdots, \mathbf{v}_k$ and the coefficient of $\mathbf{v}_k$ is positive.*

Before proceeding further, we note an important consequence of this result.

COROLLARY. *If V is an n-dimensional inner product space, then V has an orthonormal basis.*

Proof of Corollary. Take an arbitrary basis $\mathcal{A}$ for V and use the Gram-Schmidt Process to obtain an an orthonormal set $\mathcal{U}$ with the same number of elements. Such a subset is linearly independent (as noted before), and since it has $\dim V$ elements it follows that $\mathcal{U}$ must be a basis for V.■

Derivation of Gram-Schmidt Process. The proof is by induction, and the first step is to verify the result when $m = 1$. In this case we know that $\mathbf{v}_1 \neq \mathbf{0}$ and we simply need to take $\mathbf{u}_1 = |\mathbf{v}_1|^{-1}\,\mathbf{v}_1$.

Suppose now that $1 \le k < m$ and we have constructed an orthonormal set of vectors $\mathcal{U}_k = \{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ such that for each $j \le k$ the vector $\mathbf{u}_j$ is a linear combination of the vectors in $\{\mathbf{v}_1, \cdots, \mathbf{v}_j\}$ and the coefficient of $\mathbf{v}_j$ in this expression is positive. Consider the vector

$$\mathbf{w}_{k+1} \quad = \quad \mathbf{v}_{k+1} \; - \; \sum_{j=1}^{k} \langle \mathbf{w_{k+1}}, \mathbf{u_j} \rangle\, \mathbf{u_j} \ .$$

Direct calculation shows that $\langle \mathbf{w}_{k+1}, \mathbf{u}_j \rangle = 0$ for all $j \le k$. We also claim that $\mathbf{w}_{k+1} \neq \mathbf{0}$. If it were zero, then the defining equation would imply that $\mathbf{v}_{k+1}$ would be a linear combination of the vectors in $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ and since we started with a linearly independent set this cannot happen. Therefore, if we take $\mathbf{u}_{k+1} = |\mathbf{w}_{k+1}|^{-1}\,\mathbf{w}_{k+1}$ then the set $\mathcal{U}_{k+1}$ obtained by adding $\mathbf{u}_{k+1}$ to $\mathcal{U}_k$ will be an orthonormal set with the desired properties. If we continue this process we ultimately obtain the desired orthonormal set $\mathcal{U} = \mathcal{U}_m$.■

There are several basic formulas related to the construction of the Gram-Schmidt Process. In particular, the basic idea for constructing $\mathbf{w}_{k+1}$ will be central to the next section of these notes. For the time being we shall only give a couple of simple formulas:

COEFFICIENT FORMULA. *Suppose that* $\mathcal{U} = \{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ *is an orthonormal set of vectors and* $\mathbf{a}$ *is a linear combination of the vectors in* $\mathcal{U}$*. If* $\mathbf{a} = \sum_j a_j \mathbf{u}_j$*, then* $a_j = \langle \mathbf{a}. \mathbf{u}_j \rangle$*.*

Proof. If we form the inner product of $\mathbf{a}$ and some vector $\mathbf{u}_i$ in the set, direct computation shows that this inner product is equal to a_i.■

INNER PRODUCT FORMULA. *Suppose that* V *is an* n*-dimensional inner product space, and let* $\mathcal{U}$ *be an orthonormal basis for* V*. If* $\mathbf{a} = \sum_j a_j \mathbf{u}_j$ *and* $\mathbf{b} = \sum_j b_j \mathbf{u}_j$*, then*

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{j=1}^{n} a_j b_j$$

In other words, with respect to an orthonormal basis, every abstract inner product on a finite dimensional vector space looks just like the usual one. This is analogous to the coordinatization results for However, just as it is very n-dimensional vector spaces in Section 3.3 of the text. Just as it is extremely worthwhile to consider abstract n-dimensional vector spaces rather than to focus attention on the vector space structure of $\mathbf{R}^n$, it is also extremely worthwhile to consider abstract n-dimensional inner product spaces rather than to focus attention on the standard inner product structure on $\mathbf{R}^n$. In fact, it is even useful to consider arbitrary inner products on $\mathbf{R}^n$ itself, and we shall discuss some aspects of this in the final unit of this course.

II.2 : Orthogonal projections and adjoints

(Fraleigh and Beauregard, §§6.1, 6.4)

In classical geometry, if $\mathbf{P}$ is a point and V is either a line or plane not in V, then the shortest distance from $\mathbf{p}$ fo V is the distance from $\mathbf{p}$ to the foot of the perpendicular dropped from $\mathbf{p}$ to V. This section formulates and proves a result of this sort and discusses some of its far-reaching implications in linear algebra.

Definition. If V is an inner product space and W is a subset of V, then the *orthogonal complement* $W^{\perp}$ (in words, "W perp") is the set of all vectors $\mathbf{v} \in V$ such that $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in W$.

If $\mathbf{v}$ is perpendicular to all vectors in W, then an elementary computation shows that $\mathbf{v}$ is perpendicular to all linear combinations of W, so henceforth *unless stated explicitly to the contrary, when writing $W^{\perp}$ we shall assume that W is a subspace.*

The next result gives some important properties of orthogonal complements.

THEOREM. *Let V be an inner product space and let W be a subspace of V. Then the following hold:*

(i) The orthogonal complement $W^{\perp}$ is a subspace of V.

(ii) The intersection of W and $W^{\perp}$ is the trivial subspace $\{\mathbf{0}\}$.

(iii) If W is finite-dimensional, then every vector in V is uniquely expressible as a sum

$$\mathbf{x} \;=\; \widehat{\mathbf{x}} + \mathbf{x}'$$

where $\widehat{\mathbf{x}} \in W$ and $\mathbf{x}' \in W^{\perp}$.

Notation. Whenever we have a decomposition of V as the sum of two subspaces $W_1 + W_2$ such that their intersection is $\mathbf{0}$ we shall say that V is a **direct sum** of W_1 and W_2, and we shall write $V = W_1 \oplus W_2$. In such cases one can use part of the argument for (iii) to show that every vector in V is uniquely expressible as a sum $\mathbf{w}_1 + \mathbf{w}_2$ where $\mathbf{w}_i \in W_i$.

Proof of Theorem. (i) Suppose that $\mathbf{w} \in W$. If $\mathbf{x}_1 \in W^{\perp}$ and $\mathbf{x}_2 \in W^{\perp}$ then

$$\langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{w} \rangle \;=\; \langle \mathbf{x}_1, \mathbf{w} \rangle + \langle \mathbf{x}_2, \mathbf{w} \rangle \;=\; 0 + 0 \;=\; 0$$

and since $\mathbf{w}$ was arbitrary this means that $\mathbf{x}_1 + \mathbf{x}_2 \in W^{\perp}$. Similarly, if $\mathbf{w} \in W$, $\mathbf{x} \in W^{\perp}$ and c is a scalar, then

$$\langle c\,\mathbf{x}, \mathbf{w} \rangle \;=\; c\,\langle \mathbf{x}, \mathbf{w} \rangle \;=\; c \cdot 0 \;=\; 0$$

shows that $c\,\mathbf{w} \in W^{\perp}$. Therefore $W^{\perp}$ is a subspace.

(ii) Suppose that $\mathbf{x}$ is a vector lying in both subspaces. Then since $\langle \mathbf{x}, \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in W$ we must also have $\langle \mathbf{x}, \mathbf{x} \rangle = 0$, which means that $\mathbf{x}$ must be equal to $\mathbf{0}$.

(iii) Let $\mathcal{U} = \{\, \mathbf{u}_1, \; \cdots \;, \mathbf{u}_m \,\}$ be an orthonormal basis for W, and given $\mathbf{x} \in V$ let

$$\widehat{\mathbf{x}} \;=\; \sum_{j=1}^{m} \langle \mathbf{x}, \mathbf{u}_j \rangle \, \mathbf{u}_j$$

so that $\widehat{\mathbf{x}} \in W$ by construction. To prove that $V = W + W^{\perp}$ it is enough to show that the latter contains $\mathbf{x}' = \mathbf{x} - \widehat{\mathbf{x}}$. As indicated in the statement at the beginning of the proof, it suffices to prove that $\mathbf{x}'$ is perpendicular to every vector in the spanning set $\mathcal{U}$. But for every choice of i we have

$$\langle \mathbf{x}', \mathbf{u}_i \rangle \quad = \quad \left\langle \mathbf{x} \; - \; \left(\sum_{j=1}^{m} \langle \mathbf{x}, \mathbf{u}_j \rangle \, \mathbf{u}_j \right), \mathbf{u}_i \right\rangle \quad =$$

$$\langle \mathbf{x}, \mathbf{u}_i \rangle \; - \; \sum_{j=1}^{m} \langle \mathbf{x}, \mathbf{u}_j \rangle \, \langle \mathbf{u}_j, \mathbf{u}_i \rangle$$

and the latter is equal to zero because $\langle \mathbf{u}_j, \mathbf{u}_i \rangle$ is equal to 0 if $i \neq j$ and 1 if $i = j$.

To see that there is only one way of writing a vector $\mathbf{x}$ as a sum of a vector in W and a vector in $W^{\perp}$, suppose we are given an arbitrary decomposition $\mathbf{x} = \mathbf{y}_1 + \mathbf{y}_2$ where $\mathbf{y}_1 \in W$ and $\mathbf{y}_2 \in W^{\perp}$. Then we also have $\mathbf{y}_1 - \widehat{\mathbf{x}} = \mathbf{x}' - \mathbf{y}_2$, and since the expression on the left implies it belongs to W and the expression on the right belongs to $W^{\perp}$ it follows that the vector in question belongs to $W \cap W^{\perp} = \{\mathbf{0}\}$.■

Definition. The vector $\widehat{\mathbf{x}}$ constructed above is called the *orthogonal* (or *perpendicular*) projection of $\mathbf{x}$ onto the subspace W. Note that the theorem gives an explicit formula for $\widehat{\mathbf{x}}$ in terms of an orthonrmal basis for W. This will be useful in the proof of the next result:

LEAST SQUARES MINIMIZATION PROPERTY. *Let V be an inner product space, let W be a finite dimensional subspace, and let $\mathbf{x} \in V$. Then the absolute minimum value of $|\mathbf{x} - \mathbf{w}|^2$ over all $\mathbf{w} \in W$ occurs when $\mathbf{w} = \widehat{\mathbf{x}}$ and nowhere else.*

In particular, given an arbitrary vector $\mathbf{x}$, the perpendicular projection onto W is the unique closest point to $\mathbf{x}$, in complete agreement with the results from elementary (high school) geometry.

Proof. Take an orthonormal basis $\mathcal{U}$ as above, and write an arbitrary vector of W as $\mathbf{w} = \sum_j a_j \, \mathbf{u}_j$. We need to show that the function

$$\left| \, \mathbf{x} - \textstyle\sum_j a_j \, \mathbf{u}_j \, \right|^2$$

takes a minimum value and this happens if and only if $a_j = \langle \mathbf{x}, \mathbf{u}_j \rangle$ for all j. It will be convenient to set b_j equal to the difference $a_j - \langle \mathbf{x}, \mathbf{u}_j \rangle$.

Some algebraic manipulation then shows that $\mathbf{x} - \mathbf{w}$ is equal to $\mathbf{x}' - \sum_j b_j \, \mathbf{u}_j$ and since the vectors in $\{\mathbf{x}\} \cup \mathcal{U}$ are pairwise orthogonal it follows from a generalization of the Pythagorean identity (see Exercise 23 on page 237 of the text) that

$$|\mathbf{x} - \mathbf{w}|^2 \quad = \quad \left| \, \mathbf{x}' - \textstyle\sum_j b_j \, \mathbf{u}_j \, \right|^2 \quad = \quad |\mathbf{x}'|^2 \; + \; \sum_j \; |b_j|^2 \; .$$

The unique absolute minimum for this expression occurs when $b_j = 0$ for all j, which happens if and only if $\mathbf{w} = \widehat{\mathbf{x}}$.■

LEAST SQUARES SOLUTIONS TO OVERDETERMINED LINEAR SYSTEMS. In real life situations one often has several observable quantities such that one of them — say y — seems likely to be a first degree polynomial function of the others, which we shall call $x_1, \; \cdots \; , x_n$; the simplest situation involves two observables where one is expected to be a linear function of the other.

Knowing this, one wants to find the coefficients a_i and the constant b for which y should be equal to $\sum_i \ a_i x_i + b$. The obvious way to search for these coefficients and the constant is to make a large number of observations, Unless one is particularly unlucky, a well-chosen set of $n+1$ observations will provide enough information to retrieve some candidates for the a_i and b. However, to allow for bad luck and also to ensure greater accuracy for the experimental formulas, it is better to make a significantly larger number of observations. Suppose that the number of observations is m, and let $\mathbf{x}_i$ and $\mathbf{y}$ be $m \times 1$ column vectors containing the separate observations of the variable quantities (so the first coordinates give the results of the first observation, *etc.*). Take $\mathbf{e}$ to be the $m \times 1$ column vector whose entries are all equal to 1. If there is a linear polynomial functional equation then these vectors will satisfy the identity

$$\mathbf{y} \quad = \quad b\,\mathbf{e} \ + \ \textstyle\sum_i \ a_i\,\mathbf{x}_i$$

but in the real world this is too much to expect from experimental data. The best one can hope for is to find constants a_i and b so that the deviation from equality is as small as possible, and one way of doing this is to find constants such that the expression

$$|\mathbf{y} - b\,\mathbf{e} \ - \ \textstyle\sum_i \ a_i\,\mathbf{y}_i|^2$$

is minimized. This is not quite the same expression as in the Least Squares Minimization Property, but it is fairly close because one can view it as the square of the distance from $\mathbf{y}$ to a point in the finite-dimensional subspace W spanned by the vectors $\mathbf{e}$ and $\mathbf{x}_i$. One (relatively clumsy) way of working this out explicitly is to find an orthonormal basis for W, compute the perpendicular projection of $\mathbf{y}$ onto W using this subspace, and substitute back in order to find the constants a_i and b. A more detailed treatment of this topic appears in Section 6.5 of the text.■

It is also worthwhile to mention a simple corollary of the properties of orthogonal complements that is extremely important in the theory of periodic functions on the real line (specifically, functions such that $f(x + 2\pi) = f(x)$ for all real numbers x).

BESSEL'S INEQUALITY. *Let V be an inner product space, let $\mathcal{U}$ be a finite orthonormal set of vectors in V, and let $\mathbf{x} \in V$. Then*

$$\sum_i \ |\,\langle \mathbf{x},\, \mathbf{u}_i \rangle\,|^2 \quad \leq \quad |\mathbf{x}|^2 \ .$$

Proof. The left hand side is the square of the length of $\widehat{\mathbf{x}}$, and by the Pythagorean identity we have $|\mathbf{x}|^2 = |\widehat{\mathbf{x}}|^2 + |\mathbf{x}'|^2$. Since the latter implies $|\widehat{\mathbf{x}}|^2 \leq |\mathbf{x}|^2$, the inequality in the statement of the result follows immediately.■

APPLICATION TO TRIGONOMETRIC SERIES. The theoretical and practical importance of finding power series expansions for functions is illustrated in first year calculus calculus courses as well as other science courses. It is natural to ask if there is a similarly useful way of studying periodic functions by means of infinite series whose terms are the basic periodic functions of the form 1 (*i.e.*, the constant function), $\sin nx$ and $\cos mx$, where m and n are positive integers. Eighteenth and nineteenth century mathematicians developed an approach to studying such infinite series expansions, which are known as *Fourier series*, but they ran into both practical and theoretical difficulties that were not totally resolved until the nineteen sixties. The discussion here will be limited to pointing out an important fact regarding these series that arises from Bessel's Inequality.

Let $\mathcal{A}$ denotes the set of periodic continuous functions on the real line (with period 2π) and the integral inner product on page 232 of the text. If we let S be the subset consisting of the constant function 1 along with the trigonometric functions $\sin nx$ or $\cos mx$ (where m and n are positive integers), then the standard formal integration techniques from first year calculus imply that S is orthogonal without a great deal of work. Furthermore, the length of the constant function 1 with respect to this metric is obviously equal to 2π, while the lengths of all the remaining functions in $\mathcal{A}$ are equal to π. Suppose now that $\mathcal{W}_n$ is the subspace spanned by the constant function 1 and the trigonometric functions $\sin kx$ and $\cos \ell x$ where k and ℓ are $\leq n$. The orthogonal projection of a function f on $\mathcal{W}_n$ is then given by an expression of the form

$$\frac{a_0}{2} \; + \; \sum_{k=1}^{n} a_k \, \cos kx \; + \; b_k \, \sin kx$$

where a_k is $1/\pi$ times the integral of $f(x) \cos kx$ on the interval $[0, 2\pi]$ and where b_k is $1/\pi$ times the integral of $f(x) \sin kx$ on the interval $[0, 2\pi]$. The proof of Bessel's Inequality implies that

$$\frac{a_0^2}{2} \; + \; \sum_{k=1}^{n} a_k^2 \; + \; b_k^2 \quad \leq \quad \int_0^{2\pi} \; [f(x)]^2 \, dx$$

and therefore it follows that the infinite series

$$\frac{a_0^2}{2} \; + \; \sum_{k=1}^{\infty} a_k^2 \; + \; b_k^2$$

converges. In fact, this result remains true even if we take f to be a function with only finitely many discontinuities on $[0, 2\pi]$; the most important example of this sort is the "square wave" function whose value is $+1$ on the interval $(0, \pi)$, whose value is -1 on the interval $(\pi, 2\pi)$, and whose values at integral multiples of π are zero. This provides strong evidence that one can form a workable theory of trigonometric series expansions for periodic functions. However, the problem of determining whether the infinite series

$$\frac{a_0}{2} \; + \; \sum_{k=1}^{\infty} a_k \, \cos kx \; + \; b_k \, \sin kx$$

actually converges, and if so whether it converges to $f(x)$, requires a great deal more work. Several of the most important aspects of this problem are discussed in upper level undergraduate real variables courses.■

Perpendicular projections as linear transformations

The material below corresponds to Section 6.4 of the text but is presented from a somewhat different viewpoint.

At the beginning of Section 6.4 in the text there is a discussion which strongly suggests that perpendicular projection is a linear operation. We shall begin by verifying this explicitly.

PROPOSITION. *Let V be an inner product space, let W be a finite dimensional subspace, and let T be the mapping from V to itself that sends each vector $\mathbf{x}$ into its perpendicular projection $\widehat{\mathbf{x}}$ onto W. Then T is a linear transformation.*

Proof. This is an immediate consequence of the formula

$$\widehat{\mathbf{x}} \quad = \quad \sum_{j=1}^{m} \langle \mathbf{x},\, \mathbf{u}_j \rangle\, \mathbf{u}_j$$

and the details of the verification are left to the reader as an exercise.■

EXAMPLE. If $V = \mathbf{R}^n$ then we know that T is given by some $n \times n$ matrix E, and frequently it is necessary to compute this matrix. We shall carry out the computation for orthogonal projection onto the subspace W of $\mathbf{R}^4$ spanned by the orthogonal vectors $(1,1,1,1)$ and $(-3,-1,1,3)$ The lengths of these vectors are 2 and $\sqrt{20}$ respectively, so if we multiply the original vectors by the reciprocals of these constants we obtain an orthonormal basis for W.

More generally, suppose we are given an ordered set of k orthonormal vectors $\mathbf{u}_i$ in $\mathbf{R}^n$; let B be the $n \times k$ matrix which has (the transposes of) these vectors in order as its columns. Since the i^{th} coordinate of $\mathbf{u}_j$ is just the product of the latter with $\mathbf{e}_i$, it follows that

$$b_{i,j} \quad = \quad \langle \mathbf{u_j},\, \mathbf{e}_i \rangle \ .$$

By definition, the j^{th} column of the associated matrix E is the vector $P(\mathbf{e}_j)$, where $\mathbf{e}_j$ is just the j^{th} standard unit vector. On the other hand, by the coordinate formula from the end of the last section we know that

$$P(\mathbf{e}_j) \quad = \quad \sum_{k=1}^{n} \langle \mathbf{e}_j,\, \mathbf{u_k} \rangle\, \mathbf{u}_k$$

and if we substitute the corresponding formula for $\mathbf{u}_k$ as a linear combination of the standard unit vectors we obtain the equation

$$P(\mathbf{e}_j) \quad = \quad \sum_{k,\ell=1}^{n} \langle \mathbf{e}_j,\, \mathbf{u_k} \rangle\, \langle \mathbf{u}_k,\, \mathbf{e}_\ell \rangle\, \mathbf{e}_\ell \quad = \quad \sum_{k,\ell=1}^{n} b_{j,k}\, b_{\ell,k}\, \mathbf{e}_\ell \ .$$

Since the ℓ^{th} coordinate of the left hand side is just the matrix entry $e_{\ell,j}$ and the coefficient of $\mathbf{e}_\ell$ on the right hand side is equal to the $e_{\ell,j}$ entry of the matrix $B\,{}^{\mathbf{T}}B$, it follows that $E = B\,{}^{\mathbf{T}}B$.

If we now apply this to the specific example given above, we find that the matrix E is given as follows:

$$\begin{pmatrix} 0.7 & 0.4 & 0.1 & -0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.3 & 0.4 \\ -0.2 & 0.1 & 0.4 & 0.7 \end{pmatrix}$$

To avoid confusion, we note that these are precise values and not approximations to one or more decimal places.

ALGEBRAIC CHARACTERIZATION OF ORTHOGONAL PROJECTIONS. We shall consider the following conceptual question: *How does one characterize linear transformations from an inner product space to itself that are given by orthogonal projection onto finite dimensional subspaces?* The process of answering this question will yield concepts that will play an important role for the rest of this course. One important aspect of the approach in the text is that it provides an explicit computation for the matrix representing a perpendicular projection onto a subspace of $\mathbf{R}^n$.

The first observation on perpendicular projections is fairly elementary.

PROPOSITION. *If V is an inner product space with a finite dimensional subspace W and T is the linear transformation from V to itself given by orthogonal projection onto W, then T is* **idempotent***; i.e., $T^2 = T \circ T$ is equal to T.*

Proof. Given $\mathbf{x} \in V$, as usual write $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{x}'$ where $\widehat{\mathbf{x}} \in W$ and $\mathbf{x}' \in W^{\perp}$. Then by definition $T(x) = \widehat{\mathbf{x}}$ and similarly we have

$$T^2(\mathbf{x}) \quad = \quad T(\widehat{\mathbf{x}}) \quad = \quad \widehat{\mathbf{x}} \quad = \quad T(\mathbf{x})$$

and since $\mathbf{x}$ was arbitrary this means that $T^2 = T$.■

The second basic property of perpendicular projections requires a fundamental definition.

Definition. Let V and W be inner product spaces, and let $T : V \to W$ be a linear transformation defined on V and taking values in W. An *adjoint transformation* is a linear transformation $T^* : W \to V$ such that

$$\langle T(\mathbf{x}), \, \mathbf{y} \rangle \quad = \quad \langle \mathbf{x}, \, T^*(\mathbf{y}) \rangle$$

for all $\mathbf{x} \in V$ and $\mathbf{y} \in W$.

PROPERTIES OF ADJOINTS. (i) *A linear transformation T as above has at most one adjoint.*

(ii) *If T_1 and T_2 are linear transformations from V to W and each has an adjoint, then so does $T_1 + T_2$ and $(T_1 + T_2)^* = T_1^* + T_2^*$.*

(iii) *If T has an adjoint and c is a scalar then cT has an adjoint and $(cT)^* = cT^*$.*

(iv) *If T has an adjoint and $S : W \to X$ is another linear transformation of inner product spaces, then $S \circ T$ has an adjoint and $(S \circ T)^* = T^* \circ S^*$.*

(v) *If T has an adjoint, then so does T^* and in fact $(T^*)^* = T$.*

Proof. (i) Suppose that S_1 and S_2 are adjoints to T. Then we have

$$\langle \mathbf{x}, \, S_1(\mathbf{y}) \, \rangle \quad = \quad \langle T(\mathbf{x}), \, \mathbf{y} \rangle \quad = \quad \langle \mathbf{x}, \, S_2(\mathbf{y}) \, \rangle$$

for all $\mathbf{x}$ and $\mathbf{y}$. In particular, it follows that for all $\mathbf{y}$ the difference $S_1(\mathbf{y}) - S_2(\mathbf{y})$ is perpendicular to every vector in V. This can only happen if $S_1(\mathbf{y}) - S_2(\mathbf{y}) = \mathbf{0}$, or equivalently when $S_1(\mathbf{y}) = S_2(\mathbf{y})$. Since $\mathbf{y}$ is arbitrary this means that $S_1 = S_2$.■

(ii) We need to show that $T_1 + T_2$ and $T_1^* + T_2^*$ satisfy the adjoint condition. But

$$\langle [T_1 + T_2](\mathbf{x}), \, \mathbf{y} \rangle \quad = \quad \langle T_1(\mathbf{x}) + T_2(\mathbf{x}), \, \mathbf{y} \rangle \quad = \quad \langle T_1(\mathbf{x}), \, \mathbf{y} \rangle \; + \; \langle T_2(\mathbf{x}), \, \mathbf{y} \rangle \quad =$$

$$\langle \mathbf{x}, \, T_1^*(\mathbf{y} \rangle \; + \; \langle \mathbf{x}, \, T_2^*(\mathbf{y}) \rangle \quad = \quad \langle \mathbf{x}, \, T_1^*(\mathbf{y}) + T_2^*(\mathbf{y}) \rangle \quad = \quad \langle \mathbf{x}, \, [T_1 + T_2]^*(\mathbf{y}) \rangle$$

so the adjoint condition is satisfied.■

(iii) We need to show that cT and cT^* satisfy the adjoint condition. But

$$\langle [cT](\mathbf{x}), \, \mathbf{y} \rangle \quad = \quad \langle cT(\mathbf{x}), \, \mathbf{y} \rangle \quad = \quad c \, \langle T(\mathbf{x}), \, \mathbf{y} \rangle \quad =$$

$$c \, \langle \mathbf{x}, \, T^*(\mathbf{y}) \rangle \quad = \quad \langle \mathbf{x}, \, cT^*(\mathbf{y}) \rangle \quad = \quad \langle \mathbf{x}, \, [cT]^*(\mathbf{y}) \rangle$$

so the adjoint condition is satisfied in this case too.■

(iv) It suffices to show that

$$\langle S \circ T(\mathbf{x}), \, \mathbf{y} \rangle \quad = \quad \langle \mathbf{x}, \, T^* \circ S^*(\mathbf{y}) \rangle$$

for all $\mathbf{x} \in V$ and $\mathbf{y} \in X$. This is true because

$$\langle S \circ T(\mathbf{x}),\, \mathbf{y} \rangle \;=\; \langle S\big(T(\mathbf{x})\big),\, \mathbf{y} \rangle \;=\; \langle T(\mathbf{x}),\, S^*(\mathbf{y}) \rangle \;=\; \langle \mathbf{x},\, T^*\big(S^*(\mathbf{y})\big) \rangle \;=\; \langle \mathbf{x},\, T^* \circ S^*(\mathbf{y}) \rangle$$

which shows that $T^* \circ S^*$ is an adjoint to $S \circ T$.■

(v) Once again we verify an adjoint condition:

$$\langle T^*(\mathbf{y}),\, \mathbf{x} \rangle \;=\; \langle \mathbf{x}, T^*(\mathbf{y}) \rangle \;=\; \langle T(\mathbf{x}),\, \mathbf{y} \rangle \;=\; \langle \mathbf{y},\, T(\mathbf{x}) \rangle$$

It follows that T is an adjoint to T^*.■

Of course, none if this has much value unless we know that adjoints exist. The next main result establishes this for finite dimensional inner product spaces, and for the standard coordinate inner product spaces it gives the matrix for the adjoint transformation in terms of the matrix defining the original linear transformation.

THEOREM. (i) *If V and W are finite-dimensional inner product spaces and $T : V \to W$ is a linear transformation, then T has an adjoint.*

(ii) *If A is an $m \times n$ matrix and $\mathcal{L}_A$ is the associated linear transformation from $\mathbf{R}^n$ to $\mathbf{R}^m$, then the $n \times m$ matrix associated to the adjoint linear transformation*

$$(\mathcal{L}_A)^* \;: \mathbf{R}^m \to \mathbf{R}^n$$

is the transpose of A.

The proof of the first part depends upon the following observation.

LEMMA. *Let V be a finite dimensional inner product space and let $f : V \to \mathbf{R}$ be a linear function (sometimes called a* **linear functional** *in the literature). Then there is a unique vector $\mathbf{v}_f \in V$ such that*

$$f(\mathbf{x}) = \langle \mathbf{x},\, \mathbf{v}_f \rangle$$

for all $\mathbf{x} \in V$.

Proof of Lemma. Pick an orthonormal basis $\mathcal{U} = \{\, \mathbf{u}_1, \,\cdots\, , \mathbf{u}_n \,\}$ for V, and write $\mathbf{x} = \sum_j \, x_j \, \mathbf{u_j}$. Let

$$\mathbf{v}_f \;=\; \sum_{i=1}^{n} \, f(\mathbf{u}_i)\, \mathbf{u}_i \ .$$

Direct computation shows that

$$f(\mathbf{x}) \;=\; \sum_{i=1}^{n} \, f(\mathbf{u}_i)\, \mathbf{x}_i \;=\; \langle \mathbf{x},\, \mathbf{v}_f \rangle$$

and this proves existence. To prove uniqueness, suppose that $\mathbf{w}$ is an arbitrary vector such that $f(\mathbf{x}) = \langle \mathbf{x},\, \mathbf{w} \rangle$ for all $\mathbf{x}$. Taking differences, we see that $0 = \langle \mathbf{x},\, \mathbf{v}_f - \mathbf{w} \rangle$, for all $\mathbf{x}$, which in turn implies that $\mathbf{v}_f - \mathbf{w} = \mathbf{0}$ so that $\mathbf{v}_f = \mathbf{w}$.■

Proof of Theorem. (i) For fixed $\mathbf{y} \in W$ consider the map f from V to $\mathbf{R}$ defined by

$$f(\mathbf{x}) \;=\; \langle T(\mathbf{x}),\, \mathbf{y} \rangle \ .$$

Foutine calculation shows that f is a linear functional, and therefore there is a unique vector in V, say $T^*(\mathbf{y})$, such that

$$\langle \mathbf{x}, T^*(\mathbf{y})\rangle \quad = \quad f(\mathbf{x}) \quad = \quad \langle T(\mathbf{x}), \mathbf{y}\rangle \ .$$

This defines the map T^* from W to V, and it remains to show that T^* is linear. By the lemma the additivity of T^* will follow if we know that

$$\langle \mathbf{x}, T^*(\mathbf{y}_1 + \mathbf{y}_2)\rangle \quad = \quad \langle \mathbf{x}, T^*(\mathbf{y}_1) + T^*(\mathbf{y}_2)\rangle$$

for all $\mathbf{y}_1$ and $\mathbf{y}_2$, and the latter in turn is true because the left hand side is equal to

$$\langle T(\mathbf{x}), \mathbf{y}_1 + \mathbf{y}_2\rangle \quad = \quad \langle T(\mathbf{x}), \mathbf{y}_1\rangle \ + \ \langle T(\mathbf{x}), \mathbf{y}_2\rangle \quad =$$

$$\langle \mathbf{x}, T^*(\mathbf{y}_1)\rangle \ + \ \langle \mathbf{x}, T^*(\mathbf{y}_2)\rangle \quad = \quad \langle \mathbf{x}, T^*(\mathbf{y}_1) + T^*(\mathbf{y}_2)\rangle$$

which is what we need to show additivity. To prove that T^* is homogeneous with respect to scalar multiplication, we need to show that taking inner products with $T^*(c\,\mathbf{y})$ and $c\,T^*(\mathbf{y})$ yield the same linear functional on V for every $\mathbf{y} \in W$. The proof of this is another chain of equations:

$$\langle \mathbf{x}, T^*(c\,\mathbf{y})\rangle \quad = \quad \langle T(\mathbf{x}), c\,\mathbf{y}\rangle \quad = \quad c\,\langle T(\mathbf{x}), \mathbf{y}\rangle \quad = \quad c\,\langle \mathbf{x}, T^*(\mathbf{y})\rangle \quad = \quad \langle \mathbf{x}, c\,T^*(\mathbf{y})\rangle$$

and this proves the homogeneity property that we want.∎

(*ii*) Let $\mathcal{U} = \{\, \mathbf{u}_1, \ \cdots \ , \mathbf{u}_n \,\}$ and $\mathcal{V} = \{\, \mathbf{v}_1, \ \cdots \ , \mathbf{v}_m \,\}$ be the standard unit vector bases for $\mathbf{R}^n$ and $\mathbf{R}^m$ respectively. It follows immediately that

$$a_{i,j} \quad = \quad \langle \mathcal{L}_A(\mathbf{u}_j), \mathbf{v}_i\rangle \quad = \quad \langle \mathbf{u}_j, (\mathcal{L}_A)^* \, (\mathbf{v}_i)\rangle$$

and therefore if B is the matrix of te adjoint transformation it follows that $b_{j,i} = a_{i,j}$.∎

Definition. Let V be an inner product space. A linear transformation $T : V \to V$ is *self-adjoint* if T is an adjoint to itself.

If $V = \mathbf{R}^n$ and A is an $n \times n$ matrix, then the linear transformation $\mathcal{L}_A$ is self-adjoint if and only if the matrix A is equal to its own transpose (*i.e.*, it is **symmetric**).

CHARACTERIZATION OF ORTHOGONAL PROJECTIONS. *Let V be an inner product space. A linear transformation $T : V \to V$ is a perpendicular projection onto a finite dimensional subspace if and only if it is idempotent and self-adjoint.*

Proof. ($\Longrightarrow$) If T is the perpendicular projection onto a finite dimensional subspace W, we have already noted that T is idempotent. To see that T is self adjoint, let $\mathbf{x}$ and $\mathbf{y}$ be arbitrary vectors in V and express them as sums of vectors in W and $W^{\perp}$ in the usual way:

$$\mathbf{x} \quad = \quad \widehat{\mathbf{x}} \ + \ \mathbf{x}' \ , \qquad\qquad \mathbf{y} \quad = \quad \widehat{\mathbf{y}} \ + \ \mathbf{y}'$$

It then follows that

$$\langle T(\mathbf{x}), \mathbf{y}\rangle \quad = \quad \langle T(\mathbf{x}), \widehat{\mathbf{y}} \ + \ \mathbf{y}'\rangle \quad = \quad \langle \widehat{\mathbf{x}}, \widehat{\mathbf{y}} \ + \ \mathbf{y}'\rangle \quad =$$

$$\langle \widehat{\mathbf{x}}, \widehat{\mathbf{y}}\rangle \quad = \quad \langle \widehat{\mathbf{x}} \ + \ \mathbf{x}', \widehat{\mathbf{y}}\rangle \quad = \quad \langle \mathbf{x}, T(\mathbf{y})\rangle$$

and therefore T is self-adjoint.∎

($\Longleftarrow$) Let W be the (finite dimensional) image of T. We claim that [1] if $\mathbf{w} \in W$ then $T(\mathbf{w}) = \mathbf{w}$, [2] the kernel of T is equal to $W^{\perp}$. The first follows because $\mathbf{w} \in W$ implies that $\mathbf{w} = T(\mathbf{x})$ for some $\mathbf{x} \in V$, and since T is idempotent we have

$$T(\mathbf{w}) \quad = \quad T\big(T(\mathbf{x})\,\big) \quad = \quad T^2(\mathbf{x}) \quad = \quad T(\mathbf{x}) \quad = \quad \mathbf{w}\ .$$

To prove the second property, first recall that by definition the vector $\mathbf{x}$ lies in the kernel of T if and only if $T(\mathbf{x}) = \mathbf{0}$, which in turn happens if and only if $\langle T(\mathbf{x}),\, \mathbf{y}\rangle = 0$ for all $\mathbf{y} \in V$. Since T is self-adjoint, the latter holds if and only if $\langle \mathbf{x},\, T(\mathbf{y})\rangle = 0$ for all $\mathbf{y}$, and this is true if and only if $\mathbf{x}$ lies in the orthogonal complement of the image of T, which is W.

We now apply this information to show that T sends a vector to its orthogonal projection onto W. Given $\mathbf{x} \in V$ write $\mathbf{x} \;=\; \widehat{\mathbf{x}} + \mathbf{x}'$ as usual. We know that T is linear, so $T(\mathbf{x}) = T(\widehat{\mathbf{x}}) + T(\mathbf{x}')$; by the preceding paragraph $T(\widehat{\mathbf{x}}) = \widehat{\mathbf{x}}$ and $T(\mathbf{x}') = \mathbf{0}$, so it follows that $T(\mathbf{x}) = \widehat{\mathbf{x}}$ as required.$\blacksquare$

II.3 : Orthogonal matrices

(Fraleigh and Beauregard, §6.3)

We know that an $n \times n$ matrix is invertible if and only if either its rows or columns determine a basis for $\mathbf{R}^n$ under the standard identifications of $\mathbf{R}^n$ with the spaces of $1 \times n$ or $n \times 1$ row or column vectors. In this section we are interested in the properties of $n \times n$ matrices whose rows or columns form an *orthonormal* basis for $\mathbf{R}^n$. The first step in this process is the following simple observation:

FORMULA. *Let A and B be $n \times n$ matrices, and let $1 \leq i, j \leq n$. Then the inner product of the i^{th} row of A with the j^{th} column of B is equal to the (i, j) entry of the product matrix $A\,B$.*■

This leads immediately to the following basic result:

THEOREM. *If A is an invertible $n \times n$ matrix, then the following are equivalent:*
(i) The rows of A form an orthonormal set.
(ii) The columns of A form an orthonormal set.
(iii) The inverse of A is equal to the transpose of A, or equivalently

$$A\,{}^{\mathbf{T}}\!A \;=\; {}^{\mathbf{T}}\!A\,A \;=\; I\ .$$

Proof. If the third condition holds, then by the preceding formula we know that the inner product of the i^{th} row of A with the j^{th} row of A — which is the same as the j^{th} column of ${}^{\mathbf{T}}\!A$ — is equal to the 0 if $i \neq j$ and 1 if $i = j$. Therefore the rows form an orthonormal set. If we switch the roles of A and its transpose in this argument, we also conclude that the columns of A (which are the rows of ${}^{\mathbf{T}}\!A$ also form an orthonormal set. Thus the third condition implies the first and second.

Conversely, if the rows form an orthonormal set then the formula for the terms of a product matrix imply that the (i, j) term of $A\,{}^{\mathbf{T}}\!A$ is equal to to the 0 if $i \neq j$ and 1 if $i = j$, and therefore $A\,{}^{\mathbf{T}}\!A = I$. Since A is invertible, this implies that ${}^{\mathbf{T}}\!A = A^{-1}$. Likewise, if the columns of A form an orthonormal set, then they are the rows of ${}^{\mathbf{T}}\!A$ and therefore the immediately preceding argument implies that

$$\left({}^{\mathbf{T}}\!A\right)^{-1} \;=\; {}^{\mathbf{T}}\!\left({}^{\mathbf{T}}\!A\right) \;=\; A$$

and taking inverses of both sides we again conclude that ${}^{\mathbf{T}}\!A = A^{-1}$.■

A matrix satisfying the conditions of the theorem is said to be *orthogonal.* The set of all $n \times n$ orthogonal matrices has the following important properties.

SUBGROUP PROPERTIES. *If A and B are orthogonal $n \times n$ matrices, then so is their product. If C is an orthogonal matrix, then so is its inverse. Finally, an identity matrix is an orthogonal matrix.*

Proof. The final statement is easy to prove because an identity matrix is equal to its own transpose and $I \cdot I = I$.

Suppose now that A and B are orthogonal matrices with the same numbers of rows and columns. Since $A\,{}^{\mathbf{T}}\!A = {}^{\mathbf{T}}\!A\,A = I$ and $B\,{}^{\mathbf{T}}\!B = {}^{\mathbf{T}}\!B\,B = I$ it follows that

$$(A\,B) \cdot {}^{\mathbf{T}}(A\,B) \;=\; A\,B\,{}^{\mathbf{T}}\!B\,{}^{\mathbf{T}}\!A \;=\; A\,I\,{}^{\mathbf{T}}\!A \;=\; A\,{}^{\mathbf{T}}\!A \;=\; I$$

and similarly we have

$$(B\,A)\cdot {}^{\mathbf{T}}(B\,A) \;=\; B\,A\,{}^{\mathbf{T}}\!A\,{}^{\mathbf{T}}\!B \;=\; B\,I\,{}^{\mathbf{T}}\!B \;=\; B\,{}^{\mathbf{T}}\!B \;=\; I$$

so that AB is also orthogonal. Similarly, if A is orthogonal, then the identity $A\,{}^{\mathbf{T}}\!A = {}^{\mathbf{T}}\!A\,A = I$ implies that ${}^{\mathbf{T}}\!A = A^{-1}$ is also orthogonal.∎

More generally, a subset of the invertible matrices that is closed under taking products and inverses is called a *subgroup* or a *matrix group.* Frequently one also refers to the set of all orthogonal matrices as the group of orthogonal matrices.

Orthogonal linear transformations

There is a straightforward analog of orthogonality for linear transformations.

Definition. Let V and W be inner product spaces and let $T : V \to W$ be a linear transformation. Then T is said to be *orthogonal* if T is invertible and T^{-1} is adjoint to T.

It follows immediately that *if A is an $n \times n$ matrix, then A is an orthogonal matrix if and only if $\mathcal{L}_A$ is an orthogonal linear transformation.*

Our next main order of business is to show that orthogonal transformations preserve distances and angle measurements, just like rotations about the origin in $\mathbf{R}^2$.

GEOMETRIC RIGIDITY THEOREM. *Let V and W be inner product spaces and let $T : V \to W$ be an orthogonal transformation. Then $\langle \mathbf{x}, \mathbf{y}\rangle = \langle T(\mathbf{x}), T(\mathbf{y})\rangle$ for all $\mathbf{x}, \mathbf{y} \in V$. Consequently, T preserves lengths of vectors, distances between points, and angle measurements.*

Proof. We begin by explaining why the conclusion of the final sentence follows from the algebraic formula. If we set $\mathbf{x} = \mathbf{y}$ in the latter we see that T preserves lengths of vectors, and the statement about distances follows because the distance between the images of $\mathbf{x}$ and $\mathbf{y}$ is then equal to $|\mathbf{x}-\mathbf{y}| = |T(\mathbf{x}-\mathbf{y})| = |T(\mathbf{x}) - T(\mathbf{y})|$. The statement about angles now follows because T preserves lengths and dot products.

Let $\mathbf{x}$ and $\mathbf{y}$ be arbitrary vectors in V. Since T is orthogonal we have $\mathbf{x} = T^*(\,T(\mathbf{x})\,)$ and hence

$$\langle \mathbf{x}, \mathbf{y}\rangle \;=\; \langle T^*(\,T(\mathbf{x})\,), \mathbf{y}\rangle \;=\; \langle T(\mathbf{x}), T(\mathbf{y})\rangle\ .$$

This is exactly what we wanted to prove.∎

The proof of the following result will be left to the exercises, where hints for the proof may be found:

RECOGNITION PRINCIPLE. *Let V be a finite dimensional inner product space, and let $T : V \to V$ be an zero-preserving isometric map — in other words, $T(\mathbf{0}) = \mathbf{0}$, every point of V has the form $T(\mathbf{x})$ for some $\mathbf{x}$, and $|\mathbf{x} - \mathbf{y}| = |T(\mathbf{x}) - T(\mathbf{y})|$ for all $\mathbf{x}, \mathbf{y} \in V$. Then T is an orthogonal linear transformation.*

The following result, which is essentially a reformulation of the Gram-Schmidt process, is extremely useful both for theoretical purposes and for the development of effective computational techniques:

"$Q\,R$" DECOMPOSITION. *Every invertible $n \times n$ matrix can be written uniquely as a product $Q\,R$, where Q is an orthogonal $n \times n$ matrix and R is an upper triangular matrix with positive entries down the diagonal.*

Proof. We begin by proving the existence of such a decomposition. If $\mathbf{a}_j$ denotes the i^{th} column of A, then the set $\mathcal{A} = \{\, \mathbf{a}_1, \,\cdots\, , \mathbf{a}_n \,\}$ is a basis for $\mathbf{R}^n$. Application of the Gram-Schmidt process to $\mathcal{A}$ yields an orthonormal basis $\mathcal{U} = \{\, \mathbf{u}_1, \,\cdots\, , \mathbf{u}_n \,\}$ and a matrix B whose columns are the vectors of $\mathcal{U}$ in the given order. We need to understand how A and B are related.

By the construction of the Gram-Schmidt process, the basic equations for defining the columns of B in terms of the columns of A have the form

$$\mathbf{b}_j = \sum_{k \leq j} c_{k,j} \mathbf{a}_k$$

where each $c_{j,j}$ is a positive real number, and if we set $c_{k,j} = 0$ for $k > j$ we obtain a square matrix C that is upper triangular. Since the i^{th} coordinates of $\mathbf{b}_j$ and $\mathbf{a}_k$ are $b_{i.j}$ and $a_{i.k}$ respectively, this means that

$$b_{i,j} = \sum_{k=0}^{n} c_{k,j} \mathbf{a}_k$$

$B = AC$ where C is upper triangular with diagonal entries that are positive. Since an upper triangular matrix with positive entries is invertible and has an inverse of the same type, it follows that C is invertible, so that $A = B\,C^{-1}$ and thus we can take $Q = C$ and $R = C^{-1}$.

To see uniqueness, suppose that $A = Q_0\,R_0 = Q_1\,R_1$ where each Q_i is orthogonal and each R_i is upper triangular with positive diagonal entries. We then have $Q_1^{-1}\,Q_0 = R_1\,R_0^{-1}$. The left hand side is a product of orthogonal matrices and hence is an orthogonal matrix. The right hand side is a product of two upper triangular matrices with positive diagonal entries (we already noted that the inverse of such a matrix has the same properties). It is a straightforward exercise to verify that the product of two upper triangular matrices with positive entries also has positive entries (see the exercises), and therefore it follows that $R_1\,R_0^{-1}$ is an upper triangular matrix that is also an orthogonal matrix.

We claim it suffices to show that a matrix that is both orthogonal and upper triangular with positive diagonal entries must be an identity matrix; if this is true then we have $Q_1^{-1}\,Q_0 = R_1\,R_0^{-1} = I$. which implies $Q_0 = Q_1$ and $R_0 = R_1$. But suppose that B is a matrix that is both orthogonal and upper triangular with positive diagonal entries, and consider the matrix ${}^{\mathbf{T}}\!B$. Since B is orthogonal it follows that ${}^{\mathbf{T}}\!B = B^{-1}$. On the other hand, since B is upper triangular with positive diagonal entries, then ${}^{\mathbf{T}}\!B$ is lower triangular with positive diagonal entries and B^{-1} is upper triangular with positive diagonal entries. Since the two matrices are equal, it follows that $A = {}^{\mathbf{T}}\!B = B^{-1}$ is both lower and upper triangular with positive diagonal entries, and the latter in turn implies that A is in fact diagonal with positive diagonal entries. Since A is orthogonal, each column has unit length, and this can happen only if $A = I$.■

EXAMPLE. We shall find the "$Q\,R$" decomposition for the 3×3 matrix

$$A \quad = \quad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad .$$

The column vectors for this matrix are given by

$$\mathbf{a}_1 \quad = \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{a}_2 \quad = \quad \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{a}_3 \quad = \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

and the columns for the matrix Q are given by applying the Gram-Schmidt process to this ordered set of vectors. The latter yields a set of vectors $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, such that $\mathbf{q}_j$ is a linear combination of $\mathbf{u}_1$ through $\mathbf{u}_j$ for all j. If one performs the explicit calculations in this case, the following orthonormal basis for $\mathbf{R}^3$ is obtained:

$$\mathbf{q}_1 \quad = \quad \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}, \qquad \mathbf{q}_2 \quad = \quad \begin{pmatrix} -2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{pmatrix}, \qquad \mathbf{q}_3 \quad = \quad \begin{pmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

The matrix A is the matrix whose columns are $\mathbf{q}_1$, $\mathbf{q}_2$ and $\mathbf{q}_3$ in that order.

To compute R, proceed as follows: Since the first i columns of A and the first j columns of Q span the same subspace of $\mathbf{R}^n$ we know that the column $\mathbf{a}_j$ is a linear combination of $\mathbf{q}_1$ through $\mathbf{q}_j$. In fact, we have the equations

$$\mathbf{a}_j \quad = \quad \sum_{k=1}^{j} \langle \mathbf{a}_j, \, \mathbf{q}_k \rangle \, \mathbf{q}_k$$

and as in the proof of the decomposition theorem, these imply that $A = Q\,R$ if we set

$$r_{s,t} = \langle \mathbf{a}_t, \, \mathbf{q}_s \rangle \ .$$

These inner products vanish if $s > t$, and therefore the matrix R is automatically upper triangular.

If we apply the preceding formula to the example, we find that

$$R \quad = \quad \begin{pmatrix} 3/\sqrt{3} & 2/\sqrt{3} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{6} \\ 0 & 0 & 1/\sqrt{2} \end{pmatrix}$$

and to check the accuracy of this one can compute $Q\,R$ and verify that this product is equal to the original matrix A.

III. Change of bases

As in the discussion of Section I.B, if V and W are vector spaces of dimensions m and n respectively, then there is a 1–1 correspondence between linear transformations $T : V \to W$ and $m \times n$ matrices defined as follows: Let $\mathcal{A} = \{\, \mathbf{a}_1, \ \cdots \ , \mathbf{a}_n \,\}$ and $\mathcal{B} = \{\, \mathbf{b}_1, \ \cdots \ , \mathbf{b}_m \,\}$ be ordered bases for V and W respectively, and define the $m \times n$ matrix of T with respect to $\mathcal{A}$ and $\mathcal{B}$ by using the following equations to define its entries $c_{i,j}$:

$$T(\mathbf{a}_j) \quad = \quad \sum_{i=1}^{m} c_{i,j}\, \mathbf{b}_i$$

These equations are defined for $1 \leq j \leq n$, and the matrix itself will frequently be denoted by

$$[T]^{\mathcal{B}}_{\mathcal{A}}$$

in these notes. If we choose different ordered bases, then we obtain different 1–1 correspondences.

The main purpose of this section is to begin consideration of the following questions:

(1) *How are the different matrices related if one changes bases? In particular, what can one say about situations where $V = W$ and $\mathcal{A} = \mathcal{B}$?*

(2) *Suppose we are in a situation where $V = W$ and $\mathcal{A} = \mathcal{B}$, and let P and Q represent the same linear transformation with respect to different ordered bases. What sorts of properties do P and Q have in common?*

(3) *In the same type of situation, to what extent is it possible to choose an ordered basis $\mathcal{A}$ so that the matrix $[T]^{\mathcal{A}}_{\mathcal{A}}$ takes a particularly simple form?*

We shall give a complete answer to the first question and partial answers to the other two. In Section IV.4 of the notes we shall answer the last two questions more generally.

Default convention. Unless specifically stated otherwise, if T is a linear transformation from V to itself and $\mathcal{A}$ is an ordered basis for V, then we shall refer to $[T]^{\mathcal{A}}_{\mathcal{A}}$ as *the matrix of T with respect to $\mathcal{A}$*.

III.A : Review topics

(Fraleigh and Beauregard, §§2.3, 5.2)

Aside from the setting developed above, there are two main points to review. One involves the computations on pages 305–307 of the text for diagonalizable matrices. These involved a diagonalizable matrix A such that the columns of C formed a basis of eigenvectors for A; The eigenvalues of these columns were given by the diagonal entries of some diagonal matrix D. Under those conditions it was shown that $A\,C$ was equal to $C\,D$, or equivalently $D = C^{-1}A\,C$. This discussion was followed by the definition of *similarity* for matrices. Specifically, two square matrices A and B are said to be similar if there is an invertible matrix P such that $P\,A\,P^{-1}$; in the situation considered in Section 5.2, one can take P to be the inverse of C.

III.1 : Similarity of matrices

(Fraleigh and Beauregard, §§7.1, 7.2)

We shall now consider a general version of the question discussed in the preceding review. Let V be an n-dimensional vector space, let $T : V \to V$ be a linear transformation, and let $\mathcal{U} = \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ and $\mathcal{W} = \{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ be ordered bases for V. Suppose that A is the matrix of T with respect to the ordered basis $\mathcal{U}$ and B is the matrix of T with respect to the ordered basis $\mathcal{W}$. We want to determine the relationship between A and B, and we shall do so by writing everything out fairly explicitly.

Let P be the matrix which gives the coefficients of the vectors in $\mathcal{W}$ when the latter are written as linear combinations of the vectors in $\mathcal{U}$. Specifically, the entries of P are defined by the equations

$$\mathbf{w}_i \;=\; \sum_{k=1}^{n} p_{k,i}\,\mathbf{u}_k$$

where $1 \leq i \leq n$. We may then write $T(\mathbf{w}_j)$ as a linear combination of the vectors in $\mathcal{U}$ by first writing this vector as a linear combination of the vectors in $\mathcal{W}$ using B and then using P to express everything as a linear combination of vectors in $\mathcal{U}$:

$$T(\mathbf{w}_j) \;=\; \sum_{i=1}^{n} b_{i,j}\,\mathbf{w}_i \;=\; \sum_{i,k} b_{i,j}\,p_{k,i}\,\mathbf{u}_k$$

On the other hand, we may also write $T(\mathbf{w}_j)$ as a linear combination of the vectors in $\mathcal{U}$ by first writing this $\mathbf{w}_j$ as a linear combination of the vectors in $\mathcal{U}$ by means of P and then using A to evaluate this linear combination:

$$T(\mathbf{w}_j) \;=\; T\left(\sum_{\ell=1}^{n} p_{\ell,j}\,\mathbf{u}_\ell\right) \;=\; \sum_{i,\ell} p_{\ell,j}\,a_{i,\ell}\,\mathbf{u}_i$$

The coefficients of the vectors $\mathbf{u}_i$ in both expressions must be equal. But the coefficient in the first expression is the (i,j) entry of $P\,B$ while the coefficient in the second is the (i,j) entry of $A\,P$. Since i and j are arbitrary this establishes the following basic result:

CHANGE OF BASIS FORMULA. *In the notation described above, we have $P\,B = A\,P$, or equivalently*

$$B \;=\; P^{-1}\,A\,P\ .\blacksquare$$

COROLLARY. *Two square matrices are similar if and only if they represent the same linear transformation with respect to different ordered bases.*

Proof. We have already shown the ($\Longleftarrow$) implication. To see the ($\Longrightarrow$) implication, suppose that we have three matrices A, B, P satisfying $B = P^{-1}\,A\,P$, let $\mathcal{U}$ be the standard basis of unit vectors, and let $\mathcal{W}$ be the ordered basis whose vectors are the columns of P. Under these conditions the preceding calculations show that the matrix C of $\mathcal{L}_A$ with respect to the ordered basis $\mathcal{W}$ must satisfy $P\,C = A\,P$, which in turn implies that $C = P^{-1}\,A\,P\ = B$.$\blacksquare$

EXAMPLE. Suppose that $\mathcal{U}$ is the standard ordered basis of unit vectors for $\mathbf{R}^2$ and $\mathcal{W}$ is given by the ordered set of vectors $\mathbf{w_1} = (1,1)$ and $\mathbf{w_2} = (1,2)$. We take A to be the matrix

$$\begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$$

and use the information above to compute the matrix $[\mathcal{L}_A]_{\mathcal{W}}$ in this case. The first step is to find the matrix P, which expresses the vectors in $\mathcal{W}$ in terms of the vectors in $\mathcal{U}$. Since $\mathcal{U}$ is just the standard ordered basis for $\mathbf{R}^2$, the linear combinations for the vectors in $\mathcal{W}$ are given by their coordinates and we have

$$P \quad = \quad \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} .$$

The next step is to compute the inverse for P, and it turns out to be given by

$$P^{-1} \quad = \quad \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} .$$

We may then substitute these matrices into the formula

$$[T]_{\mathcal{W}} \quad = \quad B \quad = \quad P^{-1}\,A\,P$$

and if we carry out all the necessary matrix calculations we obtain the answer

$$[T]_{\mathcal{W}} \quad = \quad B \quad = \quad \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} .$$

If we wish to check the correctness of this computation, we can do so by computing the vectors $A\,\mathbf{w}_i$, and if we carry these computations out we find that $2\,\mathbf{w}_1$ is equal to $A\,\mathbf{w}_1$ and $\mathbf{w}_1 + 2\,\mathbf{w}_1$ is equal to $A\,\mathbf{w}_2$, which is exactly what our computation for $[T]_{\mathcal{W}} = B$ claims should be the case.

III.2 : Invariants of similarity

(Fraleigh and Beauregard, §§7.1, 7.2)

By the results of the previous section, the classification of all linear transformations from $\mathbf{R}^n$ to itself is the same as the classification of all $n \times n$ matrices up to similarity. In order to study a problem of this sort, it is necessary to find a list of matrix properties such that two similar matrices have the same properties but two dissimilar matrices do not.

By one of the exercises in Section 5.2 of the text, the relation of similarity for matrices has the following three properties:

- Every matrix A is similar to itself.
- If A is similar to B, then B is similar to A.
- If A is similar to B and B is similar to C, then A is similar to C.

We shall repeatedly use these properties in our discussion.

A natural way to approach the similarity classification problem is to start with relatively weak properties that similar matrices share and to find increasingly stronger ones. The following is perhaps one of the simplest:

INVARIANCE OF DETERMINANTS. *Similar matrices have equal determinants.*

This is true because the determinant of a product is the product of the determinants, and since the product of scalars is commutative it follows that the determinant satisfies $\det C_1\, C_2 = \det C_2\, C_1$ for all square matrices C_2 and C_1. If we apply this to $B = P^{-1}\, A\, P$ we obtain

$$\det B \quad = \quad \det\bigl(P^{-1}\,(A\,P)\bigr)C \quad = \quad \det\bigl((A\,P)\,P^{-1}\bigr) \quad = \quad \det A\,I \quad = \quad \det A$$

and hence similar matrices have the same determinant.∎

In fact, one can push the same methods further.

INVARIANCE OF CHARACTERISTIC POLYNOMIALS. *Similar matrices have the same characteristic polynomials.*

This is a similar computation where B is replaced by the matrix of polynomials

$$B - tI \quad = \quad P^{-1}\,(A - tI)\,P$$

and A replaced by the matrix of polynomials $A - tI$.∎

Eigenvalues and eigenvectors provide further common characteristics of similar matrices.

INVARIANCE OF EIGENSPACE DIMENSIONS. *If A and B are similar matrices and λ is an arbitrary scalar, then the dimensions of the spaces of eigenvectors for A and B with eigenvalue λ are equal.*

Note that the set of eigenvectors corresponds to the solutions of, say, $A - \lambda\,I$ and thus defines a subspace. — This assertion is true because if $B = P^{-1}\,A\,P$ and W is the space of eigenvectors for A with eigenvalue λ, then $\mathcal{L}_P^{-1}(W)$ is the space of eigenvectors for B with the same eigenvalue.∎

Another way of stating the preceding result is that for every eigenvalue λ the ranks of the matrices $A - \lambda\,I$ and $B - \lambda\,I$ are the same. Although it is possible to construct examples of 3×3

matrices for which all these ranks are equal even though the matrices themselves are not similar (one of the additional exercises for this section gives a method for producing such examples), a refinement of this characterization provides complete information on the similarity type of this matrix. Since similar matrices have the same characteristic polynomial, let us assume we are dealing with a class of matrices for which this polynomial is the same. The following basic fact about polynomials will be helpful in understanding the similarity conditions we shall present:

FINITENESS OF NUMBER OF FACTORS. *If $p(t)$ is a polynomial of degree $n > 0$ (in one variable), then there are only finitely many monic polynomials that are factors of p.*■

Recall that a polynomial is *monic* if it has the form $t^d + q(t)$ where either $q = 0$ or the degree of q is strictly less than d.

Given a polynomial $f(t)$ with real coefficients and an $n \times n$ matrix A, one can form an associated matrix $f(A)$ by making the substitution $t = A$, and for each such polynomial one can define the rank of A. One then has the following result:

SIMILARITY CLASSIFICATION PRINCIPLE. *Let A and B be two matrices with the same characteristic polynomial $f(t)$. Then A and B are similar if and only if for every monic factor $g(t)$ of $f(t)$ the ranks of $g(A)$ and $g(B)$ are equal.*

Proving the equality of ranks if the matrices are similar is not difficult (see the exercises), but proving the converse is just a little beyond the scope of this course (although the theorem behind this result is a fundamental topic in graduate level algebra courses). The files `ratforms.*` contain further information.

The preceding result on similarity classification addresses only the first two of the three questions that were raised at the beginning of this unit; in particular, it does not address the question of whether one can find a matrix of a particularly simple or useful type that is similar to a given matrix. In fact, such nice forms exist provided we broaden our perspective to allow scalars that are more general than real numbers, and we shall ideal with such questions and their applications to matrix similarity in the next unit of the course.

IV. Complex linear algebra

Linear algebra begins with very concrete objects; namely, the vector addition and scalar multiplication structures on $\mathbf{R}^2$ and $\mathbf{R}^3$. It turns out that there are numerous other objects with notions of addition and scalar multiplication that share the fundamental algebraic properties of vector addition and scalar multiplication in $\mathbf{R}^2$ and $\mathbf{R}^3$, and these provide important motivation for defining the abstract notion of a vector space over the real numbers.

Just as mathematicians discovered the usefulness of studying vectors in an abstract setting rather than only in the concrete case of the coordinate vector space $\mathbf{R}^n$, they also concluded it was useful to allow more general sorts of scalars than real numbers for some purposes. In particular, for many theoretical and practical reasons it is useful to consider vector spaces where the scalars are the complex numbers. Some indications of reasons for this are implicit in the discussion of eigenvalues, where one has real characteristic polynomials whose roots may be non-real complex numbers. The purpose of this unit is to develop linear algebra over the complex numbers $\mathbb{C}$ and use it to obtain conclusions that could not be readily obtained by only looking at real vector spaces.

More generally, virtually all of linear algebra, with the notable exception of material involving inner product spaces, goes through for fairly arbitrary scalars that have a decent notion of addition, subtraction, multiplication and division by nonzero elements. Such a system is called a **field**. In fact, except for results involving eigenvalues and determinants, the basic concepts and results of linear algebra also go through for systems in which the multiplication is not commutative; i.e., where $a\,b$ is not necessarily equal to $b\,a$. Such systems are known as **division rings** or **skew fields**. For many (but not all) theoretical purposes division rings provide the ultimate setting for doing much of linear algebra and using it to study geometrical questions (most notably in a subject called *projective geometry*), but this is really a topic better suited to graduate level mathematics courses. Further information can be found on pages 180–190 in Section IV.2 of the following text:

> Hungerford, Thomas W. Algebra. Reprint of the 1974 original. (Graduate Texts in Mathematics, 73.) *Springer-Verlag, New York–Berlin–etc.*, 1980. ISBN: 0-387-90518-9.

Here are some online references. The first is fairly general, and the others include information on the connections to geometry.

`http://en.wikipedia.org/wiki/Division_ring`

`http://www-math.cudenver.edu/~wcherowi/courses/m6221/pglc2a.html`

`http://www-math.cudenver.edu/~wcherowi/courses/m6221/pglc.html`

IV.1 : Complex numbers

(Fraleigh and Beauregard, §9.1)

Throughout recorded history mathematicians have repeatedly broadened their concepts of numbers, beginning with the natural numbers (or positive integers) and expanding upon them so that increasingly complicated equations have solutions. For example, the inclusion of fractions allows one to solve equations like $2x = 1$, the inclusion of negative numbers allows one to solve equations like $x+1 = 0$, and the inclusion of irrational real numbers allows one to solve equations like $x^2 = 2$. The complex numbers are a natural continuation of this pattern, for they are constructed so that one can solve the equation $x^2 = -1$. The first known use of complex numbers dates back to the the sixteenth century, and at the time mathematicians realized that such numbers could be used to solve all sorts of quadratic and cubic equations beyond those of the form $x^2 + a = 0$ where $a > 0$. By the end of the seventeenth century the use of complex numbers in mathematics was firmly established.

Each time that new types of numbers were considered, there was some initial reluctance to do so. In the case of $\sqrt{2}$ this was reflected by denoting this and similar objects as *surds* (related to the word *absurd*) or *irrational numbers* in contrast to the so-called *rational numbers* that had been considered before. Similarly, numbers like $\sqrt{-1}$ were known as *imaginary numbers* in contrast to the so-called *real numbers* that had become the accepted standard for numbers. When discussing irrational or imaginary numbers, it is important to view the terms only as labels that do not reflect any uncertainty about the mathematical legitimacy of such objects.

Formally, the complex number system $\mathbb{C}$ is defined to be the set of all expressions of the form $a + b\,i$ where a and b are real numbers and i is a square root of -1, so that $i^2 = 1$. Two complex numbers $a + b\,i$ and $c + d\,i$ are equal if and only if their *real parts* — namely, a and c — and their *imaginary parts* — namely c and d — are equal; note that the imaginary part is a **real** number (and the product of the imaginary part with i is an imaginary number). These objects can be added, subtracted and multiplied in the usual way. In fact, if $a + b\,i \neq 0(+0\,i)$ then a reciprocal is defined by the formula

$$\frac{1}{a + b\,i} = \frac{a - b\,i}{a^2 + b^2} \ ;$$

the right hand side can be defined because $a + b\,i \neq 0$ implies that at least one of a and b are nonzero, which in turn implies that $a^2 + b^2$ is positive, and it is an elementary exercise in algebra to verify that the product of the right hand side with $a + b\,i$ is equal to $1 = 1 + 0\,i$. Therefore one can divide by nonzero complex numbers just as one can divide by nonzero real numbers.

Having pointed out some basic similarities between the real and complex numbers, we now point out further ways in which they resemble each other and other ways in which they are different. First of all, we cannot extend the usual notions of "less than" and "greater than" to complex numbers. The simples way to see this is to suppose one could extend the usual notions of ordering and to ask whether i is positive or negative. If it is positive, then its square should also be positive. But by construction $i^2 = -1 < 0$, so this is impossible. On the other hand, if i were negative then $(-i)$ should be positive, and since $(-i)^2$ is also equal to -1 we again obtain a contradiction. In particular, this means that there is no algebraically "reasonable" way or extending the ordering of the real numbers to the complex numbers.

Geometrically we often think of the real numbers as representing points on some line. Similarly, since a complex number is completely determined by the real numbers that are its real and imaginary parts, we may think of a complex number geometrically as a point in the coordinate plane. This geometric interpretation is particularly useful when thinking about an important generalization of the usual notion of absolute value known as the *modulus* or *norm* or *length* or *absolute value* of a complex number. Specifically, if $z = a + b\,i$, then the modulus $|z|$ is the square root of $a^2 + b^2$. For real numbers (those of the form $a + 0\,i$, this reduces to the usual notion of absolute value, and in both cases the absolute value (or whatever name one might prefer) denotes the distance from the point to the zero element.

Another important construction called *complex conjugation* also has a natural geometric interpretation. Once again, if $z = a + b\,i$, then the *complex conjugate* $\overline{z}$ is equal to $a - b\,i$. Geometrically the conjugate $\overline{z}$ is related to z by reflection about the real axis. The important algebraic properties of complex conjugation are summarized in Theorem 9.1 on page 457 of the text, and comments on the derivation of these properties appear on page 458 (with some further references). Pages 459–463 of the text discuss the geometric interpretation of complex multiplication and division (assuming the denominator is nonzero in the second case).

There is an important relation between complex norms and complex conjugation; namely, if $z \in \mathbb{C}$ then $z\,\overline{z}$ is equal to $|z|^2$ by the elementary identity $(a + b\,i) \cdot (a - b\,i) = a^2 + b^2$. This and the previously cited argument on page 458 of the text lead to the formula

$$|\,z\,w\,| \quad = \quad |z| \cdot |w| \ .$$

Since $|z|$ also represents the length of z when the latter is viewed as a 2-dimensional real vector with coordinates $(a,\,b)$, it also follows that the norm satisfies the Triangle Inequality

$$|\,z \ + \ w\,| \quad \leq \quad |z| \ + \ |w| \ .$$

CAUTION. Over the complex numbers it is meaningful to talk about inequalities of the form $|z| < r$ (this just describes the set of points inside the circle whose center is 0 and whose radius is r), but it is **not** meaningful to talk about inequalities of the form $z < r$ because there is no "reasonable" extension of $<$ and $>$ from the reals to the complex numbers.

The discussion of complex multiplication on pages 459–463 also touches on another fundamentally important difference between the real and complex numbers; namely, *over the complex numbers, there are n distinct $n^{\rm th}$ roots of $+1$.* The latter was a key point in the sixteenth century discovery that every cubic equation $x^3 + b\,x^2 + c\,x + d$ (where b, c, d are real) factors completely into linear factors, and in fact the same holds over the complex numbers. In the nineteenth century mathematicians ultimately proved a similar result for complex polynomials of arbitrary degree:

FUNDAMENTAL THEOREM OF ALGEBRA. *Let $p(t)$ be a nonconstant polynomial over $\mathbb{C}$ of degree n. Then $p(t)$ factors completely into a product of first degree polynomials over $\mathbb{C}$.*■

In particular, this result says that we do not have to go any further than the complex numbers in order to find roots for real or complex polynomials.

Many proofs of the Fundamental Theorem of Algebra are known, and despite the name of the result all proofs require something beyond algebra (for example, differential and integral calculus done for complex rather than real variables, or some concepts and methods from topology). Virtually every book on the theory of functions of a complex variable contains a proof of the Fundamental Theorem of Algebra. For our purposes it will be enough to know that the result is true.

The Fundamental Theorem of Algebra has the following implication for polynomials with real coefficients, the proof of which is left to the exercises:

FACTORIZATION OF REAL POLYNOMIALS. *Let $p(t)$ be a nonconstant polynomial over the real numbers. Then $p(t)$ factors completely into a product of first degree polynomials over the reals and second degree polynomials over the reals that have no real roots.*■

Infinite series over $\mathbb{C}$ and the complex exponential function

We shall need some information about complex numbers that goes beyond Section 9.1 of the text. Specifically, when we revisit systems of linear differential equations in Section IV.5, it will be necessary to work with infinite series of complex numbers and an extension of the function e^x to complex exponents. Our discussion below is very similar to the treatment of (real) matrix exponentials in the file `expmatrix.pdf` from the course directory that was previously mentioned in Section I.3.

Limits of complex number sequences can be defined and manipulated much like limits of real valued sequences, the key point being that distances in both cases are expressed in the form $|p-q|$. Similarly, one can talk about convergence of a complex valued infinite series $\sum_{n=0}^{\infty}\ z_n$ in terms of the convergence of the sequence of partial sums $S_n = \sum_{i=0}^{n}\ z_k$. As in the case of ordinary infinite series, the best form of convergence is *absolute convergence*, which corresponds to the convergence of the real valued infinite series $\sum_{n=0}^{\infty}\ |z_n|$ with nonnegative terms. A fundamental theorem states that a vector valued infinite series converges if the auxiliary series $\sum_{n=0}^{\infty}\ |z_n|$ does, and there is a generalization of the standard M-test: *If $|z_n| \leq M_n$ for all n where $\sum_n\ M_n$ converges, then $\sum_n\ z_n$ also converges, and it converges absolutely.*

It will be useful to state the following complex version of a simple identity for sums of infinite series:

INFINITE SUM FACTORIZATION. *Let $\sum_{k=1}^{\infty}\ z_k$ be a convergent infinite series of complex numbers with sum S, and let $w \in \mathbb{C}$ Then $\sum_{k=1}^{\infty}\ w\, z_k$ converges to $w\, S$.*

Proof. Let S_r be the r^{th} partial sum of the original series. Then $w\, S_r$ is the corresponding partial sum for the other series, and we need to show that this become arbitrarily close to $w\, S$ if r is sufficiently large. By the hypothesis we know the analogous statement is true for the original infinite series.

Let $\varepsilon > 0$ be given, and let L be the maximum of $|w|+1$. Choose R so large that $|S_r - S| < \varepsilon/L$ if $r \geq R$. It then follows that

$$|w\, S_r - w\, S| \quad = \quad |w| \cdot |\, S_r - S| \quad < \quad \varepsilon$$

as required.■

The next result implies that we can work with power series over the complex numbers just about as well as we can over the reals.

PROPOSITION. *Suppose that we are given a sequence of complex numbers c_k for which*

$$\lim_{k\to\infty} \frac{|c_{k+1}|}{|c_k|} \quad = \quad L$$

and z is a nonzero complex number such that $|z|^{-1} > L$. Then the infinite matrix power series

$$\sum_{k=0}^{\infty} c_k \, z^k$$

converges absolutely.

Proof. The argument is closely related to the proof of the ratio test for ordinary infinite series. The norms of the terms are given by

$$|c_k \, z^k| \quad = \quad |c_k| \cdot |z|^k$$

and the latter converges if $|z|^{-1} > L$ by the ratio test. But this means that the matrix power series converges absolutely.■

SPECIAL CASE. *If z is a complex number, then the* **exponential series**

$$\exp(z) \quad = \quad \sum_{k=0}^{\infty} \frac{1}{k!} \, z^k$$

converges absolutely.■

PROPERTIES OF COMPLEX EXPONENTIALS. It follows immediately from the definition of the complex exponential function that $\exp(t + 0\,i) = e^t$, and there is also a natural generalization of the usual law of exponents $e^{a+b} = e^a \, e^b$:

PRODUCT FORMULA. *If z and w are complex numbers, then* $\exp(z + w) = \exp(z) \cdot \exp(w)$.

Idea of the proof. As for ordinary infinite series, one needs to do a little work in order to view the product of two infinite series sums as the sum of a third infinite series. Specifically, if one starts with convergent infinite series $\sum_k \; u_k$ and $\sum_k \; v_k$ with sums U and V, then one wants to say that $U \, V = \sum_k \; w_k$, where

$$w_k = \sum_{p+q=k} \; u_p \cdot v_q \; .$$

This turns out to be true if the original sequences are absolutely convergent, and we know this because the real exponential series converges for all real exponents.■

COROLLARY. *For all complex numbers z the exponential* $\exp(z)$ *is invertible and its reciprocal is* $\exp(-z)$.■

Differentiating complex valued functions

Differentiation of a complex valued function of one real variable makes sense so long as the scalar valued functions given by the real and imaginary parts are differentiable, and in this case one defines the derivative of $z = x + y\,i$ by $z' = x' + i\,y'$. These derivatives have many familiar properties:

If c is a constant then $c' = 0$.

$(z + w)' = z' + w'$.

$(z\,w)' = z'\,w + z\,w'$.

Just as for ordinary power series, one has good term by term differentiation properties, and the proofs for ordinary infinite series go through with minimal changes:

TERMWISE DIFFERENTIATION. *Suppose that we have an infinite power series of complex numbers $f(t) = \sum_{k=0}^{\infty} a_k\, t^k$ such that the radius of convergence for the auxiliary series $\varphi(t) = \sum_{k=0}^{\infty} |a_k|\, t^k$ is at least r. Then the radius of convergence of $f(t)$ is at least r, and inside this radius of convergence we have $f'(t) = \sum_{k=0}^{\infty} (k+1)\, a_{k+1}\, t^k$.*■

If we apply this to the matrix exponential function $f(t) = \exp(ta)$ we obtain the equation

$$f'(t) \quad = \quad a\,\exp(ta)\ .$$

The following result provides an important connection between the complex exponential function and the classical trigonometric functions.

EULER'S FORMULA. *If θ is a real number, then $\exp(i\,\theta) = \cos\theta + i\,\sin\theta$.*

This and the preceding results yield a simple formula for the complex exponential in terms of its real and imaginary parts:

$$\exp(x + y\,i) \quad = \quad e^x\,(\,\cos y \;+\; i\,\sin y\,)$$

Proof of Euler's Formula. By definition we have

$$\exp(i\,\theta) \quad = \quad \sum_{k=0}^{\infty} \frac{i^k}{k!}\,\theta^k\ .$$

Since $i^2 = -1$ the sequence of powers of i is given by

$$\{\,1,\ i,\ -1,\ -i,\ 1,\ i,\ \cdots\,\}$$

which is periodic of period 4. Thus the real and imaginary parts of $\exp(i\,\theta)$ are given by the sums of the even and odd powers of θ respectively. If we split $\exp(i\,\theta)$ explicitly in this manner, we see that the latter is equal to

$$\left(\sum_{k=0}^{\infty} \frac{(-1)^k}{(2\,k)!}\,\theta^{2k}\right) \;+\; i\,\left(\sum_{\ell=0}^{\infty} \frac{(-1)^{\ell+1}}{(2\,\ell+1)!}\,\theta^{2k}\right)$$

and the real and imaginary parts of this sum are equal to $\cos\theta$ and $\sin\theta$ respectively.■

IV.2 : Complex matrices

(Fraleigh and Beauregard, §9.2)

Once we have the complex numbers we may define **complex n-space** to be the set $\mathbb{C}^n$ of ordered n-tuples of complex numbers, and we may further define addition and scalar multiplication on $\mathbb{C}^n$ coordinatewise, exactly as in the real case. If we do this, then the resulting structure has all the algebraic properties described for $\mathbf{R}^n$ in Theorem 1.1 on page 9 of the text.

Likewise, one can define an abstract notion of *complex vector space* or *vector space over* $\mathbb{C}$ as on pages 180–181 of the text, the only change being that scalars are taken to be complex rather than real numbers. With this modification, virtually all the theory in the first five chapters of the text goes through with no further changes. Aside from the applications, the main exception involves the discussion of inner products.

The Fundamental Theorem of Algebra implies one important fact about complex linear algebra: *Over the complex numbers, every $n \times n$ matrix has an eigenvalue.* In this connection it is particularly useful to see what happens for the real matrix

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

which has no real eigenvalues. Its characteristic polynomial is $t^2 + 1$, so over the complex numbers it has eigenvalues given by $\pm\, i$, and the associated complex eigenvectors are given by

$$\begin{pmatrix} 1 \\ \mp\, i \end{pmatrix} .$$

Note that the nonreal roots of the characteristic polynomial are conjugate to each other and that the coordinates of the associated eigenvectors are also conjugate to each other. We shall eventually show this phenomenon holds in general for the nonreal eigenvalues and eigenvectors of a real matrix.

Complex inner products

Real inner products are important and useful because of their algebraic definitions and geometric interpretations. One would like to define complex inner products so that both the algebraic and geometric aspects are preserved as much as possible. In order to do this it is necessary to make some adjustments that may seem awkward at first but are justified by the end results.

The simplest reflection of the difference involves the expressions for the absolute values of real and complex numbers. In the real case, the absolute value of a number t is given by the unique nonnegative square root of t^2, while in the complex case the comparable notion of norm for a complex number $z = a + b\, i$ is the unique nonnegative square root of the nonnegative real number $z\,\overline{z} = a^2 + b^2$. Similarly, if we are given $\mathbf{z} \in \mathbb{C}^n$ such that

$$\mathbf{z} \quad = \quad (z_1, \ \cdots \ , z_n) \quad = \quad (a_1 + b_1\, i, \ \cdots \ , a_n + b_n\, i)$$

then we would like to define our inner product so that $\mathbf{z} \cdot \mathbf{z}$ is given by

$$\sum_{j=1}^{n} a_j^2 + b_j^2 \quad = \quad \sum_{j=1}^{n} z_j\, \overline{z_j} \ .$$

This immediately suggests the following definition:

Standard complex inner product. If $\mathbf{z}$ and $\mathbf{w}$ are vectors in $\mathbb{C}^n$ with coordinates z_j and w_j respectively, then

$$\mathbf{z} \cdot \mathbf{w} \quad = \quad \sum_{j=1}^{n} z_j \, \overline{w_j} \ .$$

This inner product has the following properties:

(1) $\mathbf{z} \cdot (\mathbf{w} + \mathbf{w}') \quad = \quad (\mathbf{z} \cdot \mathbf{w}) \ + \ (\mathbf{z} \cdot \mathbf{w}')$.

(2) $(\mathbf{z} + \mathbf{z}') \cdot \mathbf{w} \quad = \quad (\mathbf{z} \cdot \mathbf{w}) \ + \ (\mathbf{z}' \cdot \mathbf{w})$.

(3) $(c\,\mathbf{z}) \cdot \mathbf{w} \quad = \quad c\,(\mathbf{z} \cdot \mathbf{w})$.

(4) $\mathbf{z} \cdot (c\,\mathbf{w}) \quad = \quad \overline{c}\,(\mathbf{z} \cdot \mathbf{w})$.

(5) $\mathbf{w} \cdot \mathbf{z} \quad = \quad \overline{\mathbf{z} \cdot \mathbf{w}}$.

(6) If $\mathbf{z}$ is nonzero then $\mathbf{z} \cdot \mathbf{z}$ is a positive real number.

Our definition is essentially the reverse of the definition in the text, and therefore this list of properties is not quite the same as the one in the text, so the properties of this inner product are not quite the same as those described in Theorem 9.2 on page 467 of the text; specifically, the third and fourth identities are different. The basic reason for changing the definition and properties here is that (3) seems more natural than its counterpart in the book and it is also the version that is more frequently seen elsewhere. In any case, properties (1) – (3) and (6) correspond exactly to properties of the usual real inner product on $\mathbf{R}^n$, the difference involving (4) is the need to insert a conjugation when pulling a scalar out from the second factor of an inner product, and in (6) the complex inner product is not commutative but instead hanges by conjugation if one switches the two vector variables.

One can then proceed to define a *complex inner product space* to be a complex vector space V together with a complex valued function $\langle \mathbf{u}, \, \mathbf{v} \rangle$, defined for all ordered pairs of vectors $\mathbf{u}$, $\mathbf{v} \rangle$ from V and satisfying the basic properties listed above.

The concept of orthogonality (inner product $= 0$) is just as important for complex inner product spaces as it is in the real case. However, since complex inner products are not commutative, one needs to say a few words to establish that the relation "$\mathbf{x}$ is orthogonal to $\mathbf{y}$" is symmetric.

LEMMA. *Let V be an inner product space, and let $\mathbf{x}$ and $\mathbf{y}$ belong to V. Then $\langle \mathbf{x}, \, \mathbf{y} \rangle = 0$ if and only if $\langle \mathbf{y}, \, \mathbf{x} \rangle = 0$.*

Proof. We shall only prove the ($\Longrightarrow$) implication since the implication in the opposite direction follows by reversing the roles of $\mathbf{x}$ and $\mathbf{y}$ in the argument. If $\langle \mathbf{x}, \, \mathbf{y} \rangle = 0$ then we have

$$\langle \mathbf{y}, \, \mathbf{x} \rangle \quad = \quad \overline{\langle \mathbf{x}, \, \mathbf{y} \rangle} \quad = \quad \overline{0} \quad = \quad 0$$

and hence the vanishing of $\langle \mathbf{x}, \, \mathbf{y} \rangle$ implies the vanishing of $\langle \mathbf{y}, \, \mathbf{x} \rangle$.∎

Basic properties of complex inner product spaces

Needless to say, we would like to modify as much of the discussion of real inner product spaces as we can so that it carries over to complex inner product spaces. In many respects the necessary work is comparable to translating something from one (human or computer) language to another: There are details to work out, perspectives that must be adjusted and points where one must be careful not to overlook something, but there is nothing that is especially profound. We shall go through the points from Unit II requiring change in the order of coverage.

- ORTHOGONAL BASES.

The entire discussion involving orthonormal bases goes through unchanged, and the only necessary modification involves the Inner Product Formula.

COMPLEX INNER PRODUCT FORMULA. *Suppose that V is an n-dimensional complex inner product space, and let $\mathcal{U}$ be an orthonormal basis for V. If $\mathbf{a} = \sum_j a_j \mathbf{u}_j$ and $\mathbf{b} = \sum_j b_j \mathbf{u}_j$, then*

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{j=1}^{n} a_j \overline{b_j}$$

- ORTHOGONAL PROJECTIONS.

It is probably best not to worry about how the applications to least squares and trignometric series might carry over to the complex case. The formula for the matrix associated to an orthogonal projection onto a subspace of $\mathbb{C}^n$ must be modified slightly: If B is an $n \times k$ matrix with orthonormal columns, then the orthogonal projection onto the subspace $W(B)$ spanned by the columns of B is the matrix $B\,B^*$ (conjugate transpose instead of transpose).

There are several points in the discussion of adjoint transformations that require some attention. In the result labeled *Properties of adjoints*, all of the conclusions except the third remain valid, and over the complex numbers one must replace the given equation with $(c\,T)^* = \overline{c}\,T^*$ (note the appearance of conjugation on the right hand side). The modified derivation is given as follows:

$$\langle [c\,T](\mathbf{x}),\, \mathbf{y} \rangle = \langle c\,T(\mathbf{x}),\, \mathbf{y} \rangle = c\,\langle T(\mathbf{x}),\, \mathbf{y} \rangle =$$

$$c\,\langle \mathbf{x},\, T^*(\mathbf{y}) \rangle = \langle \mathbf{x},\, \overline{c}\,T^*(\mathbf{y}) \rangle = \langle \mathbf{x},\, [c\,T]^*(\mathbf{y}) \rangle$$

Then next point requiring some attention is the proof that the adjoint T^* described in the existence proof is linear, and in particular the verification that T^* is homogeneous with respect to scalar multiplication. Here is the modification needed to prove this in the complex case:

$$\langle \mathbf{x},\, T^*(c\,\mathbf{y}) \rangle = \langle T(\mathbf{x}),\, c\,\mathbf{y} \rangle = \overline{c}\,\langle T(\mathbf{x}),\, \mathbf{y} \rangle = \overline{c}\,\langle \mathbf{x},\, T^*(\mathbf{y}) \rangle = \langle \mathbf{x},\, c\,T^*(\mathbf{y}) \rangle$$

A slight modification is also needed in the formula for the matrix associated to the adjoint transformation $(\mathcal{L}_A)^*$ associated to an $m \times n$ matrix A over the complex numbers. In the complex case the matrix representing $(\mathcal{L}_A)^*$ is the **conjugate transpose** of A, which is generally denoted by A^*. The derivation of this fact is slightly more delicate than those already discussed, so we shall write out the argument in more detail:

Let $\mathcal{U} = \{\, \mathbf{u}_1, \cdots, \mathbf{u}_n \,\}$ and $\mathcal{V} = \{\, \mathbf{v}_1, \cdots, \mathbf{v}_m \,\}$ be the standard unit vector bases for $\mathbb{C}^n$ and $\mathbb{C}^m$ respectively. It follows immediately that

$$a_{i,j} = \langle \mathcal{L}_A(\mathbf{u}_j), \mathbf{v}_i \rangle = \langle \mathbf{u}_j, (\mathcal{L}_A)^*(\mathbf{v}_i) \rangle = \overline{\langle (\mathcal{L}_A)^*(\mathbf{v}_i,\, \mathbf{u}_j) \rangle}$$

and therefore if B is the matrix of te adjoint transformation it follows that $\overline{b_{j,i}} = a_{i,j}$.■

If A is a square matrix over the complex numbers, it follows that $\mathcal{L}_A$ *is self-adjoint if and only if A is equal to the* **conjugate** *of its transpose*, which is the matrix we called A^*. Complex matrices

A such that $A = A^*$ are said to be *Hermitian*. Real symmetric matrices are always Hermitian, but the 2×2 complex matrix

$$\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

is an example of a Hermitian matrix that is not symmetric. The fundamental properties of the conjugate transpose operation for matrices are recorded in Theorem 9.3 on page 470 of the text.

The proof that orthogonal projections are the same as self-adjoint idempotents goes through without change for complex vector spaces.

- ORTHOGONAL MATRICES.

Over the complex numbers, if we are given two invertible matrices A and B of the same size, the the inner product of the i^{th} row of A with the j^{th} column of B — with the row being the first variable and the column being the second — is equal to the (i, j) entry of the product matrix $A\,B^*$. This leads to a characterization of matrices with orthonormal rows or columns very similar to the real case; namely, U has either (or both) of these properties if and only if $U\,U^* = U^*\,U = I$. This follows by the same argument used in the real case provided one adjusts for the fact that $A\,B^*$ rather than $A\,B$ represents the matrix of inner products in the complex case. Matrices satisfying the condition $U\,U^* = U^*\,U = I$ are said to be *unitary*; they are the proper analog of real orthogonal matrices, but it is worth noting that the term "complex orthogonal" often has a somewhat different meaning). The remainder of the discussion in Section II.3 goes through if one substitutes "unitary" for "orthogonal" and "complex numbers" for "real numbers" throughout, and no further changes to statements and proofs of results are needed (however, in the statement of the $Q\,R$ decomposition perhaps one should add that the diagonal entries of the upper triangular matrix R are supposed to be positive *real* numbers).

IV.3 : Complex eigenvalues and eigenvectors

(Fraleigh and Beauregard, §9.3)

We have already noted that even if a real matrix does not have any real eigenvalues or eigenvectors, it might have a basis of eigenvectors if one works over the complex numbers. This should suggest the potential usefulness of studying the eigenvalues and eigenvectors of complex matrices in greater detail. Some of these results on complex matrices have important consequences for real matrices, and such applications to real matrices will be central to the final unit of this course.

We begin with a simple observation that will be generalized significantly in Section IV.5:

COMPLEX DIAGONALIZATION FOR 2×2 REAL MATRICES. *If A is a 2×2 matrix with real entries but no real eigenvalues, then A is diagonalizable over the complex numbers.*

EXAMPLE. The matrix

$$\begin{pmatrix} 3 & -2 \\ 5 & 1 \end{pmatrix}$$

has eigenvalues $\lambda_+ = 2 + 3i$ and $\lambda_- = 2 - 3i$. It follows that the eigenvectors span the null spaces of the matrices

$$A - \lambda_+ I \quad = \quad \begin{pmatrix} 1-3i & -2 \\ 5 & -1-3i \end{pmatrix}, \qquad A - \lambda_- I \quad = \quad \begin{pmatrix} 1+3i & -2 \\ 5 & -1+3i \end{pmatrix}$$

with associated eigenvectors

$$\mathbf{v}_+ \quad = \quad \begin{pmatrix} 2 \\ 1-3i \end{pmatrix} \qquad \text{and} \qquad \mathbf{v}_- \quad = \quad \begin{pmatrix} 2 \\ 1-3i \end{pmatrix}$$

respectively. In fact, the proof below shows that the eigenvalues of the matrices described in the theorem always come in conjugate pairs; in Section IV.5 we shall also note that the corresponding eigenvectors may always be chosen to have coordinates that are complex conjugates of each other's.

Proof. The hypothesis implies that the characteristic polynomial $\chi_A(t)$, which is quadratic and has real coefficients, does not have any complex roots. By the Quadratic Formula, it follows that this polynomial has a pair of complex roots that are conjugate to each other. Since the characteristic polynomial has distinct roots over the complex numbers, it follows that A is diagonalizable over the complex numbers.∎

The use of complex scalars is helpful in connection with many questions involving diagonalization, including one that is stated in Theorem 5.5 on page 314 of the text and also in Theorem 6.8 on page 354: *A symmetric matrix with real entries has an orthonormal basis of eigenvectors.* This result will be used repeatedly in the final unit of the course; we shall prove it and see how it relates to more general diagonalization theorems for complex matrices.

Diagonalizing Hermitian matrices

The diagonalization result for real symmetric matrices is a direct consequence of a similar theorem for complex Hermitian matrices and self adjoint linear transformations on complex inner product spaces. Formulating everything in terms of linear transformations on finite dimensional inner product spaces is extremely useful because it focuses attention on the main ideas rather than on computational details.

THEOREM. *Let V be a finite dimensional complex inner product space, and let $T : V \to V$ be a self adjoint linear transformation. Then T has real eigenvalues and an orthonormal basis of eigenvectors.*

MATRIX VERSION OF THE PRECEDING RESULT. *If A is an $n \times n$ Hermitian matrix over the complex numbers, the A has an orthonormal basis of eigenvectors.*

The matrix version follows from the theorem by means of the standard correspondence between an $n \times n$ matrix A and the linear transformation $\mathcal{L}_A$ on $\mathbb{C}^n$. Both have the same eigenvectors and eigenvalues, and the linear transformation is self adjoint if and only if the matrix is Hermitian.

Proof. The first step is to prove that the eigenvalues of T are real. Suppose that c is an eigenvalue for T with associated eigenvector $\mathbf{v}$. Then since $T^* = T$ we have

$$c\,|\mathbf{v}|^2 \quad = \quad \langle c\,\mathbf{v},\, \mathbf{v}\rangle \quad = \quad \langle T(\mathbf{v}),\, \mathbf{v}\rangle \quad = \quad \langle \mathbf{v},\, T(\mathbf{v})\rangle \quad = \quad \langle \mathbf{v},\, c\,\mathbf{v}\rangle \quad = \quad \overline{c}\,|\mathbf{v}|^2$$

and since $|\mathbf{v}|^2 > 0$ it follows that $c = \overline{c}$, which means that c must be a real number.

The proof that T has an orthonormal basis of eigenvectors proceeds by induction on the dimension of V. If V is 1-dimensional then T is just multiplication by a scalar, and by the previous paragraph we know this scalar must be a real number. Suppose now that $\dim V = n \geq 2$ and we know the result for $(n-1)$-dimensional complex inner product spaces. We know that T has an nonzero eigenvector $\mathbf{x}_0$, and if we multiply $\mathbf{x}_0$ by a suitable positive constant we obtain an eigenvector $\mathbf{x}_1$ that $|\mathbf{x}_1| = 1$. Let λ be the eigenvalue for $\mathbf{x}_1$ and let W be the orthogonal complement to the span of $\mathbf{x}_1$. We claim that if $\mathbf{y} \in W$ then we also have $T(\mathbf{y}) \in W$. To prove this, we need to verify that $T(\mathbf{y})$ is orthogonal to $\mathbf{x}_1$ (if this is true, it is orthogonal to every vector in the subspace spanned by $\mathbf{x}_1$). But by self adjointness we have

$$\langle \mathbf{x}_1,\, T(\mathbf{y})\rangle \quad = \quad \langle T(\mathbf{x}_1),\, \mathbf{y}\rangle \quad = \quad \langle \lambda\,\mathbf{x}_1,\, \mathbf{y}\rangle \quad = \quad \lambda\,\langle \mathbf{x}_1,\, \mathbf{y}\rangle \quad = \quad \lambda\,0 \quad = \quad 0$$

so the property is satisfied.

In view of the preceding paragraph, T determines a linear transformation $T_0 : W \to W$ such that $T_0(\mathbf{y}) = T(\mathbf{y})$ for all $\mathbf{y}$. The self adjointness of T implies that T_0 is also self adjoint. Since $\dim W = n - 1$ the induction hypothesis applies to T_0 and therefore it has an orthonormal basis of eigenvectors $\{\,\mathbf{x}_2,\ \cdots\ ,\mathbf{x}_n\,\}$. If we add $\mathbf{x}_1$ to this collection we obtain an orthonormal basis of eigenvectors for T.■In particular, this result implies that a real symmetric matrix has an orthonormal basis of eigenvectors over the complex numbers. In fact, one has a much stronger conclusion, which we state both in terms of linear transformations and matrices.

THEOREM. *Let V be a real inner product space, and let $T : V \to V$ be a self adjoint linear transformation. Then T has an orthonormal basis of eigenvectors over V.*

There is a corresponding matrix version of this result.

FUNDAMENTAL THEOREM ON REAL SYMMETRIC MATRICES. *If A is a real symmetric matrix then there is an orthogonal matrix P such that ${}^{\mathbf{T}}P\,A\,P$ is diagonal.*

EXAMPLE. The symmetric matrix

$$\begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix}$$

has eigenvalues $\lambda_1 = 7$ and $\lambda_2 = 2$. It follows that the eigenvectors span the null spaces of the matrices

$$A - 7I = \begin{pmatrix} -1 & -2 \\ -2 & -4 \end{pmatrix} \quad , \qquad A - 2I = \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix}$$

with associated eigenvectors

$$\mathbf{v}_1 = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

respectively. Direct calculation shows that these vectors are perpendicular and each has length equal to $\sqrt{5}$. Therefore the corresponding orthonormal basis of eigenvectors is given as follows:

$$\mathbf{u}_1 = \begin{pmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} \quad , \qquad \mathbf{u}_2 = \begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix}$$

Proofs of theorems. The first step is to show that every real symmetric matris A has a REAL eigenvector. First of all, A certainly has a complex eigenvalue; by the preceding results this eigenvalue is real. Therefore there is a real root c for the characteristic polynomial of A. But this implies that A has an eigenvector associated to c.

The preceding implies that *every self adjoint linear transformation on a finite dimensional real inner product space V has at least one eigenvector in V.* One can combine this fact with the inductive argument in the complex case to show that T has an orthonormal basis of eigenvectors. This argument carries over word for word.

We now turn to the matrix form of the result. Let P be a matrix whose columns display the orthonormal basis of eigenvectors for A. Then the general considerations of Unit III show that $P^{-1}\,A\,P$ is a diagonal matrix. On the other hand, since the columns of P form an orthonormal basis we also know that matrix is orthogonal and that $P^{-1} = {}^{\mathbf{T}}P$. These facts immediately yield the conclusion of the Fundamental Theorem on Real Symmetric Matrices.■

Note that we could have stated the diagonalization result for Hermitian matrices similarly: *If A is Hermitian, then there is a unitary matrix U such that $U^*\,A\,U$ is diagonal (and has real entries).*■

Triangular forms and characteristic polynomials

A variant of the preceding argument yields a triangularization theorem for arbitrary complex matrices.

SCHUR TRIANGULARIZATION PRINCIPLE. *Let V be a finite dimensional complex inner product space, and let $T : V \to V$ be a linear transformation. Then there is an orthonormal basis $\mathcal{U}$ for V such that the matrix of T with respect to T is upper triangular.*

As usual, there is a corresponding version of this result for matrices.

COMPLEX MATRIX TRIANGULARIZATION. *If A is an $n \times n$ matrix, then there is a unitary matrix U such that $U^*\,A\,U$ is upper triangular.*

The matrix version is an immediate consequence of the result for linear transformations.■

The following relationship will be useful in the proof of the Triangularization Principle:

LEMMA. *Let V be a finite dimensional real or complex inner product space, and let W be a subspace of V, and let $P : V \to V$ denote the orthogonal projection onto W. Then $I - P$ is the orthogonal projection onto $W^\perp$.*

Proof. As usual, write $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{x}'$ where $\widehat{\mathbf{x}} \in W$ and $\mathbf{x}' \in W^\perp$. Then $P(\mathbf{x}) = \widehat{\mathbf{x}}$ and therefore

$$[I - P](\mathbf{x}) \quad = \quad \mathbf{x} - P(\mathbf{x}) \quad = \quad (\widehat{\mathbf{x}} + \mathbf{x}') - \widehat{\mathbf{x}} \quad = \quad \mathbf{x}'$$

which shows that $I - P$ sends $\mathbf{x}$ to its orthogonal projection onto $W^\perp$.■

Proof of Triangularization Principle. As in the diagonalization proofs, one begins by noting that the result is immediate in the 1-dimensional case and that in general a linear transformation from an n-dimensional complex inner product space to itself always has an eigenvector of unit length. Following the lead of these proofs, we assume that the conclusion of the result is known in the $(n-1)$-dimensional case and suppose that $\dim V = n$.

Let $\mathbf{u}_1$ be a unit length eigenvector for T, let W_1 be the subspace spanned by $\mathbf{u}_1$, and let $a_{1,1}$ be the eigenvalue associated to $\mathbf{u}_1$. If$W = W_1^\perp$, then $\dim W = (n-1)$. Let $P : V \to W$ be defined by orthogonal projection onto W, and define a linear transformation $T_0 : W \to W$ by $T_0(\mathbf{w}) = P(\,T(\mathbf{w}\,)$. By the induction hypothesis there is an orthonormal basis $\mathcal{U}_0 = \{\,\mathbf{u}_2,\ \cdots\ ,\ \mathbf{u}_n\,\}$ for W such that the matrix of T_0 with respect to $\mathcal{U}_0$ is triangular; *i.e.*, we have

$$T_0(\mathbf{u}_j) \quad = \quad \sum_{i=2}^{j} a_{i,j}\,\mathbf{u}_i$$

for suitable scalars $a_{i,j}$, where $2 \le j \le n$.

Let $\mathbf{x} \in V$, and as usual write $\mathbf{x} = \widehat{\mathbf{x}} + \mathbf{x}'$ where $\widehat{\mathbf{x}} \in W$ and $\mathbf{x}' \in W^\perp = W_1$. It then follows that $\widehat{\mathbf{x}} = P(\mathbf{x})$, and since W_1 is spanned by the unit vector $\mathbf{u}_1$ we know that $\mathbf{x}' = \langle \mathbf{x}, \mathbf{u}_1\rangle\,\mathbf{u}_1$. Therefore we may write the standard orthogonal decomposition in the form

$$\mathbf{x} \quad = \quad P(\mathbf{x}) \ + \ \langle \mathbf{x}, \mathbf{u}_1\rangle\,\mathbf{u}_1\ .$$

If we take $\mathbf{x} = \mathbf{u}_j$ where $j \ge 2$ and set

$$a_{1,j} \quad = \quad \langle T(\mathbf{u}_j), \mathbf{u}_1\rangle$$

it will follow that

$$T(\mathbf{u}_j) \quad = \quad P(\,T(\mathbf{u}_j)\,) \ + \ a_{1,j}\,\mathbf{u}_1$$

for all $j \ge 2$, and if we combine this with the previous formula for $T_0(\mathbf{u}_j) = P(\,T(\mathbf{u}_j)\,)$ we obtain the formula

$$T(\mathbf{u}_j) \quad = \quad \sum_{i=1}^{j} a_{i,j}\,\mathbf{u}_i$$

if $j \ge 2$. On the other hand, we have also established an analogous formula for $j = 1$, and therefore if A is the upper triangular matrix with entries $a_{i,j}$ if $i \le j$ and zero if $i > j$, then A is the matrix of T with respect to the ordered orthonormal basis

$$\mathcal{U} \quad = \quad \{\,\mathbf{u}_1\,\} \ \cup\ \mathcal{U}_0\ .■$$

The preceding result implies that even if a matrix A is not diagonalizable, it is still similar to a matrix for which at least half of the entries off the diagonal are zero, and in fact one can even find a change of basis matrix that is orthogonal. This result has important implications for characteristic polynomials that we shall now discuss.

CAYLEY-HAMILTON THEOREM. *If A is am $n \times n$ matrix and $\chi_A(t)$ is its characteristic polynomial, then $\chi_A(A) = 0$.*

Proof. We first reduce the proof to the case of upper triangular matrices. If $B = P^{-1}A\,P$ for some invertible matrix P, then we know that $\chi_B(t) = \chi_A(t)$. Therefore if the Cayley-Hamilton Theorem holds for B we have

$$0 = \chi_B(B) \quad = \quad \chi_A(B) \quad = \quad \chi_A(\,P^{-1}A\,P\,) \quad = \quad P^{-1}\,\chi_A(A)\,P$$

so that $\chi_A(A) = P\,\chi_B(B)\,P^{-1} = P\,0\,P^{-1} = 0$. Thus if B is an upper triangular matrix that is similar to A and the Cayley-Hamilton Theorem holds for B, then it also holds for A. Therefore we shall assume that A is upper triangular for the rest of the proof.

For each $j \leq n$ let $\mathbf{W}_j$ denote the subspace of $\mathbf{R}^n$ spanned by the first j unit vectors. Since A is upper triangular, it follows that if $\mathbf{x} \in \mathbf{W}_j$ then $A\,\mathbf{x}$ also lies in $\mathbf{W}_j$.

For each j between 1 and n, let $p_j(t)$ denote the product of the linear polynomials $(t - a_{i,i})$ for $i \leq j$. We shall prove by induction on j that $[p_j(A)]\,\mathbf{e}_i = \mathbf{0}$ for all i and j such that $i \leq j$. Since the determinant of a triangular matrix is the product of its diagonal entries, it follows that $p_n(t)$ is equal to $\chi_A(t)$ and therefore the case $j = n$ corresponds to the Cayley-Hamilton Theorem for A.

If $j = 1$ the result is true because $A\,\mathbf{e}_1 = a_{1,1}\,\mathbf{e}_1$ implies $[p_1(A)]\,\mathbf{e_1} = \mathbf{0}$. Suppose that we know the result for $j - 1$, where $j \geq 2$. To see that $[p_j(A)]\,\mathbf{e}_i = \mathbf{0}$ when $i < j$, note that

$$[p_j(A)]\,\mathbf{e}_i \quad = \quad (a_{j,j}\,I - A)\,[p_{j-1}(A)]\,\mathbf{e}_i$$

and the product of the last two factors on the right hand side is zero by the induction hypothesis. Note that $[p_{j-1}(A)]\,\mathbf{y} = \mathbf{0}$ also holds for every $\mathbf{y} \in \mathbf{W}_j$. Since A is upper triangular we know that $A\,\mathbf{e_j} = a_{j,j}\,\mathbf{e_j} + \mathbf{v}$ where $\mathbf{v} \in \mathbf{W}_j$. It follows that

$$(a_{j,j}\,I - A)\,\mathbf{e}_j \quad = \quad (a_{j,j}\,I - A)\,\mathbf{v} \in \mathbf{W}_j$$

and since $p_{j-1}(A)\,\mathbf{y} = \mathbf{0}$ for $\mathbf{y} \in W$, it follows that

$$\mathbf{0} \quad = \quad p_{j-1}(A)\;[(a_{j,j}\,I - A)\,\mathbf{v}] \quad = \quad p_j(A)\,\mathbf{e}_j$$

and if we combine this with the previous observation we see that we have verified the inductive step in the argument.■

Normal matrices

We have seen that Hermitian matrices have orthonormal bases of eigenvectors, and that every square matrix can be put into triangular form by some orthonormal change of basis. Both of these lead to the final question of this section: *For which square complex matrices A can one find an* **orthonormal** *basis of eigenvectors?*

The answer is given by the following result:

SPECTRAL THEOREM. *Let A be an $n \times n$ matrix over the complex numbers. Then A has an orthonormal basis of eigenvectors if and only if A is* **normal** *in the sense that $A\,A^* = A^*\,A$,*

The class of normal matrices includes several types of matrices that we have considered thus far, including Hermitian matrices where $A^* = A$ and unitary matrices where $U^* = U^{-1}$. Additional examples of normal matrices are given by *skew-Hermitian* matrices for which $A^* - -A$. Here is an example of a real matrix that is normal but not real symmetric:

$$\begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix}$$

One can verify normality directly by direct calculation of $A\,{}^{\mathbf{T}}\!A$ and ${}^{\mathbf{T}}\!A\,A$. Note that this matrix does not have any real eigenvalues becuase its characteristic polynomial is $13 - 4\,t + t^2$, and the roots of the latter are given by $2 \pm 3\,i$.

Proof of the Spectral Theorem. Suppose first that A has an orthonormal basis of eigenvectors. Then there is a unitary matrix U such that $U^*\,A\,U$ is a diagonal matrix, say D. Since D and D^* are diagonal, it follows immediately that $D\,D^* = D^*\,D$. We then have $A = U\,D\,U^*$, and it also follows that

$$A\,A^* \quad = \quad \left(U\,D\,U^*\right)^* \quad = \quad (U^*)^*\,D^*\,U^* \quad = \quad U\,D^*\,U^*$$

so that we have

$$A\,A^* \quad = \quad (U\,D\,U^*)\,(U\,D^*\,U^*) \quad = \quad U\,D\,(U^*\,U)\,D^*\,U^* \quad = \quad U\,D\,I\,D^*\,U^* \quad = \quad U\,D\,D^*\,U^*\ .$$

Likewise, we see that $A^*\,A = U\,D^*\,D\,U^*$. Since $D\,D^* = D^*\,D$, we must have

$$A\,A^* \quad = \quad U\,D\,D^*\,U^* \quad = \quad U\,D^*\,D\,U^* \quad = \quad A^*\,A$$

as required.

The reverse implication is best considered from the viewpoint of linear transformations:

SPECTRAL THEOREM FOR LINEAR TRANSFORMATIONS. *Let V be a finite dimensional complex inner product space, and let $T : V \to V$ be a linear transformation that is normal; in other words $T\,T^* = T^*\,T$. Then T has an orthonormal basis of eigenvectors.*

The proof of this result requires some general properties of normal linear transformations that are significant in their own rights, so we digress to establish then here.

LEMMA 1. *Let V be a finite dimensional complex inner product space, and let $T : V \to V$ be a linear transformation that is self adjoint. If $\langle T(\mathbf{v}),\, \mathbf{v}\rangle = 0$ for all $\mathbf{v} \in V$ then $T = 0$.*

Proof of Lemma 1. Let $\mathcal{U} = \{\,\mathbf{u}_1,\ \cdots\ ,\, \mathbf{u}_n\,\}$ be an orthonormal basis of eigenvectors for T; recall that the corresponding eigenvalues λ_j are all real numbers. For each choice of j the hypothesis implies that

$$0 \quad = \quad \langle T(\mathbf{u}_j),\, \mathbf{u}_j\rangle \quad = \quad \langle \lambda_j\,\mathbf{u}_j,\, \mathbf{u}_j\,\rangle \quad = \quad \lambda_j$$

and therefore each vector in the orthonormal basis $\mathcal{U}$ is an eigenvector for the eigenvalue 0. Since every vector in V is a linear combination of vectors in $\mathcal{U}$, it follows that $T(\mathbf{v}) = \mathbf{0}$ for all $\mathbf{v} \in V$.■

LEMMA 2. *Let V be a finite dimensional complex inner product space, and let $T : V \to V$ be a linear transformation. Then T is normal if and only if $|T(\mathbf{v})| = |T^*(\mathbf{v})|$ for all $\mathbf{v} \in V$.*

Proof of Lemma 2. By definition, T is normal if and only if $T\,T^* = T^*\,T$, which in turn holds if and only if $T\,T^* - T^*\,T = 0$. Now a linear transformation of the form $T\,T^* - T^*\,T$ is always self adjoint (verify this!), so such a transformation is zero if and only if

$$\left\langle \left[T\,T^* - T^*\,T\right](\mathbf{v}),\, \mathbf{v}\right\rangle \quad = \quad 0$$

for all $\mathbf{v}$. The latter in turn is true if and only if

$$\langle T\,T^*(\mathbf{v}),\, \mathbf{v}\rangle \quad = \quad \langle T^*\,T(\mathbf{v}),\, \mathbf{v}\rangle$$

for all $\mathbf{v}$, and by the adjoint identity this is true if and only if

$$|T^*(\mathbf{v})|^2 \quad = \quad \langle T^*(\mathbf{v}),\, T^*(\mathbf{v})\rangle \quad = \quad \langle T(\mathbf{v}),\, T(\mathbf{v})\rangle \quad = \quad |T(\mathbf{v})|^2$$

for all $\mathbf{v}$. Since two nonnegative real numbers are equal if and only if their square roots are equal, the preceding is equivalent to the condition stated in the lemma.■

LEMMA 3. *Let V be a finite dimensional complex inner product space, let $T : V \to V$ be a normal linear transformation, and let $\mathbf{x}$ be an eigenvector for T with associated eigenvalue λ. Then $\mathbf{x}$ is also an eigenvector for T^* with associated eigenvalue $\overline{\lambda}$.*

Proof of Lemma 3. The first step is to show that $T - \lambda I$ is normal. To see this, note that we have

$$(T - \lambda I)\cdot(T - \lambda I)^* \quad = \quad (T - \lambda I)\cdot(T^* - \overline{\lambda} I) \quad = \quad T\,T^* - \lambda T^* - \overline{\lambda} T + |\lambda|^2\, I$$

and similarly

$$(T - \lambda I)^*\cdot(T - \lambda I) \quad = \quad (T^* - \overline{\lambda} I)\cdot(T - \lambda I) \quad = \quad T^*\,T - \lambda T^* - \overline{\lambda} T + |\lambda|^2\, I\ ;$$

since T is normal the right had sides of both expressions are equal, and therefore $T - \lambda I$ is normal.

Suppose now that $\mathbf{v}$ is an eigenvector for T with eigenvalue λ. Then $[T - \lambda I](\mathbf{v}) = \mathbf{0}$, and since $|\,[T - \lambda I](\mathbf{v})\,| = |\,[T - \lambda I]^*(\mathbf{v})\,|$ by Lemma 2 it also follows that

$$0 \quad = \quad |\,[T - \lambda I^*](\mathbf{v})\,| = |\,T^*(\mathbf{v}) - \overline{\lambda}\,\mathbf{v}\,|$$

which means that $T^*(\mathbf{v}) - \overline{\lambda}\,\mathbf{v} = \mathbf{0}$, or equivalently that $\mathbf{v}$ is an eigenvector for T^* with corresponding eigenvalue $\overline{\lambda}$.■

Completion(s) of the proof(s) of the Spectral Theorem(s). Once again, if V is 1-dimensional then T is multiplication by a scalar and T automatically has an orthonormal basis of eigenvectors. In this case the normality condition is true for every linear transformation T.

We proceed by induction on the dimension of V, so assume $\dim V = n$ and the result is true for linear transformations on inner product spaces with dimensions $\leq (n-1)$. Since we are working over the complex numbers we know that T has some eigenvector, and we may as well assume it has unit length. Call this eigenvector λ_1. By Lemma 3 we know that $\mathbf{u}_1$ is also an eigenvector for T^*.

Let W be the orthogonal complement to the span of $\mathbf{u}_1$. We claim that if $\mathbf{y} \in W$ then both $T(\mathbf{y})$ and $T^*(\mathbf{y})$ lie in W. To prove this, we need to verify that both of these vectors are orthogonal to $\mathbf{u}_1$ (if this is true, it is orthogonal to every vector in the subspace spanned by $\mathbf{u}_1$). But by normality we have

$$\langle \mathbf{u}_1,\, T(\mathbf{y})\rangle \quad = \quad \langle T^*(\mathbf{u}_1),\, \mathbf{y}\rangle \quad = \quad \langle \overline{\lambda}\,\mathbf{u}_1,\, \mathbf{y}\rangle \quad = \quad \overline{\lambda}\,\langle \mathbf{u}_1,\, \mathbf{y}\rangle \quad = \quad \overline{\lambda}\,0 \quad = \quad 0$$

so that $T(\mathbf{y}) \in W^*$, and similarly

$$\langle \mathbf{u}_1, T^*(\mathbf{y}) \rangle \;=\; \langle T(\mathbf{u}_1), \mathbf{y} \rangle \;=\; \langle \lambda\, \mathbf{u}_1, \mathbf{y} \rangle \;=\; \lambda \langle \mathbf{u}_1, \mathbf{y} \rangle \;=\; \lambda \cdot 0 \;=\; 0$$

so that $T^*(\mathbf{y}) \in W^*$. It follows that we have a linear transformation $S : W \to W$ defined by $S(\mathbf{y}) = T(\mathbf{y})$, that the adjoint is defined by $S^*(\mathbf{y}) = T^*(\mathbf{y})$, and that S is normal. Since $\dim W = (n-1)$ the induction hypothesis applies to S and therefore it has an orthonormal basis of eigenvectors $\{ \mathbf{u}_2, \cdots, \mathbf{u}_n \}$. If we add $\mathbf{u}_1$ to this collection we obtain an orthonormal basis of eigenvectors for T.■

COROLLARY. *Let V be a finite dimensional complex inner product space, and let $T : V \to V$ be a linear transformation. Then T is unitary if and only if it has an orthonormal basis of eigenvectors and the associated eigenvalues satisfy $|\lambda| = 1$.*

Proof. Suppose first that T is unitary. Then T is normal, and therefore it has an orthonormal basis of eigenvectors $\mathcal{U}$, and if $\mathbf{u}$ belongs to $\mathcal{U}$ then $|T(\mathbf{u})| = |\mathbf{u}| = 1$ and $|T(\mathbf{u})| = |\lambda\, \mathbf{u}| = |\lambda|\,|\mathbf{u}| = |\lambda|$ implies $|\lambda| = 1$.

Conversely, if T has an orthonormal basis $\mathcal{U}$ of eigenvectors whose eigenvalues all have absolute value 1, then the set

$$T(\mathcal{U}) \;=\; \{ T(\mathbf{u}_1), \cdots, T(\mathbf{u}_n) \} \;=\; \{ \lambda_1\, \mathbf{u}_1, \cdots, \lambda_n\, \mathbf{u}_n \}$$

is also an orthonormal basis for V, and therefore T must be unitary.■

IV.4 : Jordan form

(Fraleigh and Beauregard, §9.4)

In this section we shall consider one question that was raised in Unit III:

> *If V is a finite dimensional vector space over the real or complex numbers and $T : V \to V$ is a linear transformation, to what extent is it possible to choose an ordered basis $\mathcal{A}$ so that the matrix $[T]^{\mathcal{A}}_{\mathcal{A}}$ takes a particularly simple form?*

If T has a basis of eigenvectors, then we know that we can choose a basis such that the matrix is diagonal. Furthermore, if we are working over the complex numbers and V has an inner product, then by the results of Section IV.3 we know that we can choose an orthonormal ordered basis $\mathcal{A}$ such that $[T]^{\mathcal{A}}_{\mathcal{A}}$ is upper triangular. In particular, this means that one can choose $\mathcal{A}$ such that the entries of the associated matrix for T are zero nearly half the time. Motivated by this, one can ask if one can find an ordered basis $\mathcal{A}$ for which even more entries of $[T]^{\mathcal{A}}_{\mathcal{A}}$ are zero. The main results of this section state that one can find an triangular matrix which represents T and has very few nonzero entries; specifically, one can find a matrix B representing T such that $b_{i,j} = 0$ only if $i = j$ or $i = j - 1$ (of course, some entries in these positions may also be zero, and the point is that these are the only entries that can possibly be nonzero). In fact, we shall produce a special class of matrices of this type known as *Jordan forms* such that every complex matrix is similar to a matrix in Jordan form and it is easy to specify when two matrices in Jordan form are similar. In the next section we shall use these Jordan forms to study systems of linear differential equations of the form $Y' = AY$, where A is a real matrix that does not have an basis of real eigenvectors.

The 2×2 case

It is instructive to look first at Jordan forms for 2×2 matrices. In this cse the characteristic polynomial either has two distinct roots or is a perfect square. We have already see that a matrix is diagonalizable if the characteristic polynomial has two distinct roots, so we assume henceforth that the characteristic polynomial is a perfect square, say $(t - \lambda)^2$. By the Cayley-Hamilton Theorem we then have $(A - \lambda I)^2 = 0$, and it follows that the matrix $A - \lambda I$ is zero if and only if $A = \lambda I$, and this is equivalent to saying that A has a basis of eigenvectors. Suppose now that A is not similar to a diagonal matrix so that $N = A - \lambda I \neq 0$. By the preceding comments we have $N^2 = 0$.

More generally, let N be an arbitrary 2×2 matrix such that $N^2 = 0$ but $N \neq 0$. If P is an invertible matrix, then P^2 is also invertible and hence nonzero, and consequently N cannot be invertible. Since N is neither invertible nor the zero matrix its rank must be equal to one. Let $\mathbf{y}$ be a nonzero vector in the column space of N, and let $\mathbf{x}$ be a column vector such that $N\mathbf{x} = \mathbf{y}$. We claim that the set $\{\mathbf{x},\, \mathbf{y}\}$ is linearly independent and $N\mathbf{x} = 0$. The second follows because

$$N\mathbf{x} \;=\; N(N\mathbf{y}) \;=\; N^2\mathbf{y} \;=\; 0\,\mathbf{y} \;=\; \mathbf{0}$$

and to verify linear independence, suppose that we have scalars a and b such that $a\,\mathbf{x} + b\,\mathbf{y} = \mathbf{0}$. If we apply N to each side we obtain the following chain of equations:

$$\mathbf{0} \;=\; N(\mathbf{0}) \;=\; N(a\,\mathbf{x} + b\,\mathbf{y}) \;=\; a\,N\mathbf{x} + b\,N\mathbf{y} \;=\; a\,\mathbf{y} + \mathbf{0} \;=\; a\,\mathbf{y}\ .$$

Since $\mathbf{y}$ is nonzero (its image under N is nonzero), it follows that $a = 0$, which in turn implies that $\mathbf{0} = b\,\mathbf{y}$; but $\mathbf{y} \neq \mathbf{0}$, and therefore $b = 0$ must also hold.

If we now take $A = \lambda\, I + N$ we see that we have produced an ordered basis $\{\mathbf{x},\, \mathbf{y}\}$ such that $A\,\mathbf{x} = \lambda\,\mathbf{x}$ and $A\,\mathbf{y} = \mathbf{x} + \lambda\,\mathbf{y}$. With respect to this ordered basis the matrix of A has the following triangular form:

$$\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

We may summarize the preceding to say that every 2×2 matrix over the real or complex numbers is either diagonalizable or similar to a 2×2 matrix of the type given above. Furthermore, if A is a 2×2 matrix over the real numbers, then either A is diagonalizable over the complex numbers or else it is similar over the real numbers to a matrix of the above type, where λ is a real number; this is true because the only way that a quadratic polynomial over the reals can have a double root is if that root is real (the nonreal roots of real polynomials come in conjugate pairs).

Elementary Jordan matrices

Over the complex numbers every square matrix has a complex eigenvalue and associated complex eigenvectors. In order to develop our similarity classification of matrices we need to consider matrices that are as far removed from diagonalizable matrices as possible; specifically, we want to look at matrices for which there is only one eigenvector up to multiplication by a nonzero scalar. For each positive integer k one can construct $k \times k$ matrices of this form that closely resemble the 2×2 example given above. Specifically, let N_k be the $k \times k$ matrix whose entries $c_{i,j}$ satisfy $c_{i,i+1} = 1$ for $1 \leq i \leq n-1$ and $c_{i,j} = 0$ otherwise. Then the characteristic polynomial of N_k is $(-t)^k$, so that 0 is the only eigenvalue of N_k and the eigenvectors are just the multiples of the first unit vector $\mathbf{e}_1$. As before, we can modify this matris by adding λI to get a matrix for which λ is the only eigenvalue and all of its eigenvectors are given by the multiples of $\mathbf{e}_1$. Such matrices are called **elementary Jordan matrices**. The basic idea of a Jordan canonical form is to use such elementary Jordan matrices to build larger examples using an operation known as *block sum*, and the fundamental result of this section may be stated as follows:

SIMILARITY CLASSIFICATION BY JORDAN FORM. *Over the complex numbers, every matrix is similar to a block sum of elementary Jordan matrices, and two block sums of this sort are similar if and only if the blocks in one are obtained by rearranging the blocks in the other.*■

If A is a matrix with real entries and all the roots of the characteristic polynomial are real, then one has the following variant of the preceding classification.

SPECIAL CASE. *If A is a real matrix and all roots of $\chi_A(t)$ are real, then there is an invertible matrix P with real entries such that $P^{-1}\,A\,P$ is equal to a block sum of elementary Jordan matrices, and similar uniqueness considerations hold.*■

Block sums

Before proceeding any further, we need to describe block sums explicitly. This is most easily done using linear transformations. Suppose we are given a sequence of n positive integers $k(i)$, and let $m = \sum_i k(i)$. We may then partition the first m positive integers $\{1, \cdots, m\}$ into n subsets, the first of which consists of the integers from 1 to $k(1)$, the next of which consists of the integers from $k(1)+1$ to $k(1)+k(2)$, and so on. This partition of $\{1, \cdots, m\}$ yields a decomposition of a vector in $\mathbb{F}^m$ (where $\mathbb{F} = \mathbf{R}$ or $\mathbb{C}$) into an ordered n-tuple (or list) of vectors

$$\mathbf{v} = \big(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\big)$$

such that $\mathbf{v}_i$ is a vector in $\mathbb{F}^{k(i)}$ whose coordinates are the $k(i-1)+1$ through $k(i)$ coordinates of $\mathbf{x}$ (for the sake of completeness we set $k(0) = 0$). One can then verify directly that the map $T : \mathbb{F}^m \to \mathbb{F}^m$ sending $\mathbf{v}$ to

$$\big(A_1(\mathbf{v}_1), A_2(\mathbf{v}_2), \cdots, A_n(\mathbf{v}_n)\big)$$

is a linear transformation (verify this!), and accordingnly it is given by a left multiplication operator $\mathcal{L}_A$ for some unique $m \times m$ matrix A. This matrix is called the *block sum* and is often written $A_1 \oplus \cdots \oplus A_n$..

For purposes of understanding these matrices it is helpful to consider the case $n = 2$. In this case the block sum $A_1 \oplus A_2$ has a zero entry in the (i,j) position if either $i \leq k(1) < j$ or $j \leq k(1) < i$, and this fact is reflected by the following widely used notation for the block sum:

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

Similarly, one often views a larger block sum as an $n \times n$ matrix *whose (i,j) entry is a $k(i) \times k(j)$ matrix, with A_i in the (i,i) position and a zero matrix in the (i,j) position if $i \neq j$.*

Of course, one can more generally discuss block decompositions of matrices for which each entry can be a matrix of a certain size subject to some consistency conditions like those given above. Such decompositions are useful for both practical and theoretical purposes, but we shall not elaborate upon this point because the bookkeeping quickly becomes rather tedious.

Jordan forms in the 3×3 case

In order to illustrate the concept of Jordan canonical form, we shall describe the various possibilities for 3×3 matrices. The classification breakdown into subcases depending upon the different patterns of repeated roots for the characteristic polynomial $\chi_A(t)$.

NO DOUBLE ROOTS. In this case we know that the matrix must be diagonalizable. Note that the matrix N_1 because a 1×1 matrix has no terms off the diagonal.

ONE DOUBLE ROOT, ONE SINGLE ROOT. In this case the characteristic polynomial has the form $(\lambda - t)^2\,(\mu - t)$, where $\lambda \neq \mu$. There must be a 1×1 summand for μ because it is an eigenvalue and a single root, and one either has two 1×1 summands for λ or else a 2×2 summand of the form $\lambda I + N_2$ as above. Since Jordan forms are triangular, if the characteristic polynomial has the form described above, then two of the diagonal entries must be equal to λ and one must be equal to μ. are either 1 or 2.

ONE TRIPLE ROOT. In this case the characteristic polynomial has the form $(\lambda - t)^3$, and the Jordan form possibilities are the diagonal matrix λI, a block sum of a 2×2 matrix of the form $\lambda I + N_2$ with a 1×1 matrix whose unique entry is λ, and a 3×3 matrix of the form $\lambda I + N_3$.

VERY BRIEF COMMENTS ON THE 4×4 CASE. Of course there are many possibilities for Jordan forms that are given by taking the block sum of a 3×3 Jordan form and a 1×1 matrix, and the remaining possibilities are given by a block sum of two 2×2 elementary Jordan matrices (for which the diagonal entries of the two summands may either be the same or different) and elementary Jordan matrices of the form $\lambda I + N_4$.

It is one thing to have a general statement about the existence of a Jordan form that is similar to a given square matrix, but for practical purposes it is important to know whether one can effectively compute the explicit form for a given matrix. In practice it is far more difficult to find Jordan forms using numerical methods than it is to find eigenvectors and eigenvalues, but if a matrix is reasonably simple — say it has integer entries and only integral eigenvalues — then one can find the Jordan form fairly directly.

The following observations are important initial steps in finding Jordan forms:

NUMERICAL CONSTRAINTS ON BLOCK SUMMANDS. *Let A be an $m \times m$ matrix over $\mathbb{F}$, and assume that the characteristic polynomial factors over $\mathbb{F}$ into a product of linear factors $\prod_j \, (\lambda_j - t)^{r(j)}$, where $r(j) \geq 1$ and the scalars λ_j are distinct.*

[1] *If $E(\lambda_j)$ is the subspace given by the eigenvectors satisfying $A\,\mathbf{v} = \lambda_j\,\mathbf{v}$, then the number of elementary Jordan matrices in the Jordan form with λ_j's down the diagonal is equal to $\dim E(\lambda_j)$.*

[2] *If the Jordan matrices in [1] are given by B_i where $1 \leq i \leq \dim E(\lambda_j)$ and the matrix B_i has $k(i)$ rows and columns, then $r(j) = \sum_i \, k(i)$.*

Explanations. If C is a Jordan form for A, then one can check directly that the dimension of the space of solutions for the system of equations $(C - \lambda_j I)\mathbf{x} = \mathbf{0}$ is equal to the number of elementary Jordan matrices in the Jordan form with λ_j's down the diagonal. The second formula follows because both integers represent the highest power of $(\lambda_j - t)$ which divides $\chi_A(t) = \chi_C(t)$, where C represents a Jordan form for A.■

Here are some examples which use these identities to find Jordan forms in relatively simple cases.

Example 1. Let

$$A \quad = \quad \begin{pmatrix} 1 & -3 & -2 \\ -1 & 1 & -1 \\ 2 & 4 & 5 \end{pmatrix} .$$

The characteristic polynomial of this matrix is equal to $(3-t)\,(2-t)^2$, so either the Jordan form is diagonal or it is the sum of the 1×1 matrix whose entry is 3 with the 2×2 elementary Jordan matrix $2\,I + N_2$. If the subspace $E(2)$ of eigenvectors associated to 2 is 2-dimensional, then the matrix is diagonalizable, and if the dimension of the subspace is 1 then the Jordan form is given by the second option. Direct computation shows that the matrix $A - 2\,I$ has rank 2, and it follows that $E(2)$ is 1-dimensional, so that the A is not diagonalizable and must have the following Jordan form:

$$\begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

Example 2. Let

$$A \quad = \quad \begin{pmatrix} -1 & 0 & 4 \\ 2 & -1 & 0 \\ 3 & 2 & -1 \end{pmatrix} .$$

The characteristic polynomial for A is then given by $(3-t)\,(3+t)^2$. Once again the determination of the Jordan form reduces to determining the dimensions of the spaces $E(\pm 3)$, and in each case one finds that the dimension is equal to 1. Therefore the Jordan form is given by the block sum of a 1×1 matrix whose unique entry is 3 with a 2×2 elementary Jordan matrix of the form $-3\,I + N_2$.

Example 3. Let

$$A \;=\; \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -1 & 3 \end{pmatrix} .$$

The characteristic polynomial of A in this case is given by $(2-t)\,(1-t)^2$. In order to determine how many elementary Jordan matrices appear in the block decomposition we need to compute the dimension of $E(1)$, which is either 1 or 2 depending upon whether the rank of $A - I$ is 2 or 1. Since the rank of $A - I$ is equal to 2 this implies that the Jordan form is given by the block sum of a 1×1 matrix whose unique entry is 2 with a 2×2 elementary Jordan matrix of the form $I + N_2$.

Example 4. Let

$$A \;=\; \begin{pmatrix} 2 & -1 & 2 & 0 \\ 0 & 3 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -3 & 5 \end{pmatrix} .$$

The characteristic polynomial of A in this case is given by $(5-t)\,(2-t)^3$, so the Jordan form will be a block sum of a 1×1 matrix whose unique entry is 5 with some elementary Jordan matrices whose diagonal entries are equal to 2 and whose total numbers of rows or columns add up to 3. The main step in describing the possible summands with 2's down the diagonal is to compute the dimension of $E(2)$, which in this case is equal to $4 - \mathrm{rank}(A - 2\,I)$. The rank of $A - 2\,I$ turns out to be 3, and therefore $R(2)$ is 1-dimensional. This means that there is only one elementary Jordan matrix in the Jordan form decomposition whose diagonal entry is equal to 2, and by the previous observation we know that this must be a 3×3 matrix. Therefore the Jordan form is given by the block sum of a 1×1 matrix whose unique entry is 5 with a 3×3 elementary Jordan matrix of the form $2\,I + N_3$.

Example 5. Let

$$A \;=\; \begin{pmatrix} 2 & -4 & 2 & 2 \\ -2 & 0 & 1 & 3 \\ -2 & -2 & 3 & 3 \\ -2 & -6 & 3 & 7 \end{pmatrix} .$$

The characteristic polynomial of A in this case is given by $(2-t)^2\,(4-t)^2$. As before, the first step is to compute the dimensions of $E(2)$ and $E(4)$. Both turn out to be 1-dimensional because the ranks of $A - 2\,I$ and . $A - 4\,I$ are both equal to 3. This means that the Jordan form contains only one block summand with 2's down the diagonal and only one block summand with 4's down the diagonal, and by the factorization of the characteristic polynomial each of these elementary Jordan matrices must have 2 rows and columns.

Example 6. Supppose that we are given a strictly upper triangular matrix of the form

$$A \;=\; \begin{pmatrix} 0 & a & b & c & 0 \\ 0 & 0 & d & e & f \\ 0 & 0 & 0 & g & h \\ 0 & 0 & 0 & 0 & k \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where the nine entries a, b, c, d, e, f, g, h, k are arbitrary scalars. How do we find the Jordan form for A? We know that the characteristic polynomial is $(-t)^5$, so all the elementary Jordan matrices in the block sum have zeros down the diagonal. We also know that the number of summands is equal to the dimension of the subspace given by the null space (or kernel) of A, which is equal to $5 - \mathrm{rank}(A)$. Perhaps the most basic question is to determine the number of 1×1 block matrices that appear. How does one compute this? Consider the matrix A^2; if C is the Jordan form for A, then the kernels of C^k and A^k have the same dimension because these two power matrices are similar. If B is an elementary Jordan matrix with zeros down the diagonal, then the kernel of B^2 is 1-dimensional if B is 1×1 and 2 if the number of rows and columns in B is at least 2, and it follows that the kernel of C^2 is equal to the sum of the total number of block summands and the number of block summands with at least two rows and columns. Therefore the number b_1 of block summands with exactly one row and column is the difference

$$\mathrm{null}(C^2) - \mathrm{null}(C)$$

where (P) denotes the dimension of the kernel of P. Since the $\mathrm{rank}(P) + \mathrm{null}(P)$ is equal to the number of rows and columns for a square matrix P, it follows that

$$b_1 \quad = \quad \mathrm{rank}(C) - \mathrm{rank}(C^2) \quad = \quad \mathrm{rank}(A) - \mathrm{rank}(A^2)$$

and thus we can compute the number of block summands that are 1×1 matrices. If we let b_ℓ denote the number of block summands with ℓ rows and columns, we can proceed similarly to find b_2, b_3 and so on. In particular, we have

$$\mathrm{null}(C^3) \quad = \quad b_1 \; + \; 2\,b_2 \; + \; 3 \sum_{\ell \geq 3} b_\ell$$

which yields the formula

$$b_2 \quad = \quad \mathrm{null}(C^3) - \mathrm{null}(C^2) \quad = \quad \mathrm{rank}(C^2) - \mathrm{rank}(C^3) \quad = \quad \mathrm{rank}(A^2) - \mathrm{rank}(A^3)$$

and more generally we have the equation

$$b_\ell \quad = \quad \mathrm{rank}(A^\ell) - \mathrm{rank}(A^{\ell+1}) \; .$$

Therefore *in this case the Jordan form of A is completely determined by the ranks of the powers of A.*

Note. The numbers b_ℓ are nonnegative integers nd must also satisfy the basic condition

$$b_1 \; + \; 2\,b_2 \; + \; 3\,b_3 \; + \; 4\,b_4 \; + \; b_5$$

by Numerical Constraint [2] stated above.

IV.4. Appendix : Existence and uniqueness proofs

For the purposes of a second course in linear algebra, the most important points regarding Jordan forms are recognizing whether a given matrix is in Jordan form, finding the various possible Jordan forms with a given characteristic polynomial and eigenvector structure, and knowing how to find the Jordan form of a specific matrix if the characteristic polynomial is relatively simple to write down (for example, if it factors into a product of linear polynomials with integral roots). The purpose of this Appendix is to give a detailed proof that one can always find a Jordan form and to determine the extent to which such representations are unique.

As in Section 3, we shall translate the similarity question for matrices into a question of finding an ordered basis for a linear transformation $T : V \to V$ that has a particularly nice form, and in particular has a minimum possible number of nonzero entries off the diagonal. The first step is the definition of certain subspaces that may be viewed as expansions of the eigenspaces associated to eigenvalues.

Definition. Let $T : V \to V$ be a linear transformation on a finite dimensional vector space over the complex numbers, and let $\chi_T(t)$ denote the characteristic polynomial of T. For each root λ of $\chi_T(t)$, let $m(\lambda) \geq 1$ denote the highest power of $(\lambda - t)$ that divides $\chi_T(t)$, and define $V(\lambda)$ to be the kernel of $(T - \lambda I)^{m(\lambda)}$.. For every linear transformation $S : V \to V$ and every $m \geq 1$ we know that the kernel of S is contained in the kernel of S^m because $S(\mathbf{v}) = \mathbf{0}$ implies $S^m(\mathbf{v}) = S^{m-1}(\,S(\mathbf{v})\,) = S^{m-1}(\mathbf{0}) = \mathbf{0}$, and if we take $S = T - \lambda I$ we conclude that every eigenvector for T with associated eigenvalue λ lies in $V(\lambda)$. On the other hand, if A is a 2×2 matrix such as

$$A(\lambda) \quad = \quad \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

then the eigenvectors for λ span a 1-dimensional subspace but $V(\lambda)$ is all of $\mathbf{R}^2$.

If a linear transformation T is diagonalizable, then for each eigenvalue λ the subspace $V(\lambda)$ is the space of corresponding eigenvectors. We know that the eigenspaces for different eigenvalues have a trivial intersection consisting only of $\mathbf{0}$, and in fact we can say that every vector $\mathbf{x} \in V$ is uniquely expressible as a sum $\sum_\lambda \mathbf{x}_\lambda$ where $\mathbf{x}_\lambda \in V(\lambda)$. In fact, if one defines $V(\lambda)$ as above, the same conclusion holds for linear transformations that are not necessarily diagonalizable.

PRIMARY DECOMPOSITION THEOREM. *If V and T are as above, then the following hold:*

(i) For each eigenvalue λ, and each $\mathbf{v} \in V(\lambda)$, the image vector $T(\mathbf{v})$ also lies in $V(\lambda)$.

(ii) If λ and μ are distinct eigenvalues of T, then $V(\lambda) \cap V(\mu) = \{\mathbf{0}\}$.

(iii) Every vector $\mathbf{x}$ in V is uniquely expressible as a sum $\sum_\lambda \mathbf{x}_\lambda$ where $\mathbf{x}_\lambda \in V(\lambda)$.

(iv) If for each eigenvalue λ the set $\mathcal{A}_\lambda$ is a basis for $V(\lambda)$, then the union of the sets $\mathcal{A}_\lambda$ is a basis for V.

Proof. We being with (i). The important point is that the linear transformations T and $(T - \lambda I)^{m(\lambda)}$ commute with each other; in fact, any two linear transformations that are polynomials in T commute with each other (and the same is true for matrices). Therefore, if $\mathbf{v} \in V(\lambda)$, then the defining condition $[\,(T - \lambda I)^{m(\lambda)}\,](\mathbf{v}) = \mathbf{0}$ implies that

$$[\,(T - \lambda I)^{m(\lambda)}\,](\,T(\mathbf{v})\,) \quad = \quad [\,(T - \lambda I)^{m(\lambda)} \circ T\,](\mathbf{v}) \quad = \quad [\,T \circ (T - \lambda I)^{m(\lambda)}\,](\mathbf{v}) \quad =$$

$$T\left([\,(T - \lambda I)^{m(\lambda)}\,](\mathbf{v})\right] \quad = \quad T(\mathbf{0}) \quad = \quad \mathbf{0}\,.$$

This proves the first conclusion of the theorem.

The second part of the result requires an important fact about the greatest common divisor of two polynomials:

($\star$) *If $f(t)$ and $g(t)$ are polynomials of positive degree over the real or complex numbers and $h(t)$ is the unique monic polynomial of maximum degree that divides both $f(t)$ and $g(t)$, then there are polynomials $p(t)$ and $q(t)$ such that $1 = p(t)\,f(t) + q(t)\,g(t)$.*

A proof of this can be found in nearly any undergraduate text in abstract algebra.

If $\lambda \neq \mu$, then the greatest common divisor of $(t-\lambda)^{m(\lambda)}$ and $(t-\mu)^{m(\mu)}$ must be 1 because these polynomials do not have a common root over the complex numbers, and therefore if one sets $f(t)$ and $g(t)$ equal to these polynomials respectively one has corresponding polynomials $p(t)$ and $q(t)$. If $\mathbf{v}$ belongs to $V(\lambda) \cap V(\mu)$, then the polynomial identity implies that

$$\mathbf{v} \;=\; [p(T)\circ(T-\lambda I)^{m(\lambda)}](\mathbf{v}) \;+\; [q(T)\circ(T-\mu I)^{m(\mu)}](\mathbf{v}) \;=$$

$$[p(T)]\left(\left[(T-\lambda I)^{m(\lambda)}\right](\mathbf{v})\right) \;+\; [q(T)]\left(\left[(T-\mu I)^{m(\mu)}\right](\mathbf{v})\right) \;=$$

$$[p(T)](\mathbf{0}) \;+\; [q(T)](\mathbf{0}) \;=\; \mathbf{0}$$

where the left hand side follows because $\mathbf{v} \in V(\lambda)$ and $\mathbf{v} \in V(\mu)$.

The proof of (*iii*) also uses the result on greatest common divisors in $(\star)$ but in a somewhat different way. We shall prove (*iii*) by induction on the dimension of V; as before, this statement is true for $\dim V = 1$ because T is diagonalizable in such cases. Suppose now that the result is true for all linear transformations on vector spaces whose dimension is strictly less than $\dim V$, where the latter is assumed to be at least 2. The characteristic polynomial of T may be written as a product $(\lambda - t)^{m(\lambda)} \cdot g(t)$, where λ is not a root of $g(t)$, and as before we know that there are polynomials $p(t)$ and $q(t)$ such that

$$1 \;=\; p(t)\,(t-\lambda)^{m(\lambda)} \;+\; q(t)\,g(t)\ .$$

Let W denote the image of the linear transformation $p(T)\circ(T-\lambda I)^{m(\lambda)}$. We claim that the following hold:

[1] If $\mathbf{w} \in W$ then $T(\mathbf{w}) \in W$.
[2] $V = V(\lambda) + W$ and $V(\lambda) \cap W = \{\mathbf{0}\}$.

Since λ is an eigenvalue for T, the subspace $V(\lambda)$ is nonzero, and therefore it will follow that the dimension of W is strictly less than the dimension of V. The verification of [1] proceeds as follows: If $\mathbf{w} \in W$, then $\mathbf{w} = [p(T)](\mathbf{x})$ for some $\mathbf{x}$, and as before we have that

$$T(\mathbf{w}) = [T\circ p(T)](\mathbf{x}) \;=\; [p(T)\circ T](\mathbf{x}) \;=\; [p(T)]\big(T(\mathbf{x})\big)$$

where the right hand side is an element of W by definition, and since it is equal to $T(\mathbf{w})$ it follows that the latter belongs to W. Since the polynomial identity implies

$$\mathbf{v} \;=\; [p(T)\circ(T-\lambda I)^{m(\lambda)}](\mathbf{v}) \;+\; [q(T)\circ g(T)](\mathbf{v})$$

the first half of [2] will follow if we can show that for each $\mathbf{v}$ the vector $[q(T)\circ g(T)](\mathbf{v})$ lies in $V(\lambda)$. This is true because we have

$$[(T-\lambda I)^{m(\lambda)}]\big([q(T)\circ g(T)](\mathbf{v})\big) \;=\; [(T-\lambda I)^{m(\lambda)}\circ q(T)\circ g(T)](\mathbf{v}) \;=$$

$$[q(T)\circ(T-\lambda I)^{m(\lambda)}\circ g(T)](\mathbf{v}) \;=\; [q(T)]\big([(T-\lambda I)^{m(\lambda)}\circ g(T)](\mathbf{v})\big) \;=\; [q(T)]\big([\chi_T(T)](\mathbf{v})\big)$$

and the term inside the parentheses is zero by the Cayley-Hamilton Theorem. Finally, we need to show that $\mathbf{v} = \mathbf{0}$ if it belongs to both $V(\lambda)$ and W. We first claim that W is contained in the kernel of $g(T)$; the verification of this is similar to the preceding application of the Cayley-Hamilton Theorem. If $\mathbf{w} \in W$ so that $\mathbf{w} = [p(T)\circ(T-\lambda I)^{m(\lambda)}]\,(\mathbf{x})$ for some $\mathbf{x}$, then

$$[g(T)]\left([p(T)\circ(T-\lambda I)^{m(\lambda)}]\,(\mathbf{x})\right) \quad = \quad \cdots \quad = \quad [p(T)]\left([\chi_T(T)]\,(\mathbf{x})\right) \quad = \quad \mathbf{0}$$

where several of the steps have been omitted because they are similar to the preceding chain of equations. — Applying this to a vector $\mathbf{v}$ that belongs to both $V(\lambda)$ and W, we see that $\mathbf{v}$ lies in the kernels of both $g(T)$ and $(T-\lambda I)^{m(\lambda)}$, and if we substitute this information into the formula

$$\mathbf{v} \quad = \quad \left[p(T)\circ(T-\lambda I)^{m(\lambda)}\right](\mathbf{v}) \; + \; [q(T)\circ g(T)](\mathbf{v})$$

we find that both the first and the second term on the right hand side must be equal to zero, which in turn implies that $\mathbf{v} = \mathbf{0}$.

Returning to our induction argument, let $S : W \to W$ be the linear transformation defined by T on W, and let $U : V(\lambda) \to V(\lambda)$ be the linear transformation defined by T on $V(\lambda)$; we have shown that if $\mathbf{x} \in W$ or $\mathbf{x} \in V(\lambda)$, then the same is true for $T(\mathbf{x})$. In order to proceed further we need some information about the characteristic polynomials for S and U. The conditions on W and $V(\lambda)$ imply that one can obtain a basis for V by taking the union of bases for these two subspaces. Since T maps these two subspaces into themselves, it follows that the matrix of T with respect to such a basis is a block sum $P \oplus Q$ where P is a matrix for the linear transformation on $V(\lambda)$ determined by T and Q is the matrix of S. If we choose bases for $V(\lambda)$ and W such that these matrices have triangular form, we see that

$$\det(\Big((P \oplus Q) - t\,I\Big) \quad = \quad \det(P - t\,I)\cdot\det(Q - t\,I)$$

which implies the product formula $\chi_T(t) = \chi_U(t)\cdot\chi_S(t)$.

We claim that the characteristic polynomials of S and U are given by $g(t)$ and $(\lambda - t)^{m(\lambda)}$ respectively. Our previous conclusions show that λ is an eigenvalue of U but not of S; since λ is a root of $\chi_T(t)$ it follows that T has an eigenvector for λ, and we have noted that this eigenvector must lie in $V(\lambda)$. Therefore $(\lambda - t)$ is a factor of $\chi_U(t)$. On the other hand, since $V(\lambda)$ and W only have the zero vector in common, it follows that λ cannot be an eigenvalue for S, and consequently $(\lambda - t)$ cannot be a factor of $g(t)$. If $\mu \neq \lambda$ we also know that $V(\mu)\cap V(\lambda) = \{\mathbf{0}\}$, and since all nonzewro eigenvectors for μ lie in $V(\mu)$ it follows that $(\mu - t)$ cannot be a factor of $\chi_U(t)$. The only factorization of $\chi_T(t)$ consistent with these three conditions is $\chi_S(t) = (-1)^{m(\lambda)}\,g(t)$ and $\chi_U(t) = (\lambda - t)^{m(\lambda)}$.

By the inductive hypothesis and the preceding description of $\chi_S(t)$, it follows that every vector in W is a sum of vectors $\mathbf{w}_\mu$, where $\mathbf{w}_\mu \in W(\mu)$ and μ runs over all the roots of $g(t)$, or equivalently over all the roots of $\chi_T(t)$ except λ. By construction $W(\mu)$ is contained in $V(\mu)$ for all such μ. The verification of (*iii*) for V concludes by writing a vector $\mathbf{v} \in V$ as a sum $\mathbf{v}_\lambda + \mathbf{w}$ where $\mathbf{v}_\lambda \in V(\lambda)$ and $\mathbf{w} \in W$, and then using the induction hypothesis to write $\mathbf{w}$ as a sum of vectors $\mathbf{w}_\mu$, which by the previous reasoning all lie in the respective subspaces $W(\mu)$.

In the preceding argument we noticed that $W(\mu) \subset V(\mu)$ for all μ; for the final portion of the proof we shall need to know that these subspaces are equal, and we may prove this as follows: If L is a linear transformation on a finite dimensional vector space X, then the dimension of L is equal to the degree of $\chi_L(t)$. If α is a root of this polynomial, then our factorization of the characteristic polynomial shows that the dimension of $V(\alpha)$ is equal to the algebraic multiplicity of α as a root of $chi_L(t)$. Consider what this means for a root $\mu \neq \lambda$ of $\chi_T(t)$, or equivalently a root μ of $\chi_S(t)$.

In each case the algebraic multiplicity of μ is equal to $m(\mu)$, and accordingly the dimensions of $V(\mu)$ and $W(\mu)$ are equal. Since $W(\mu)$ is known to be a subspace of $V(\mu)$, it follows that these subspaces are equal.

We already know that every vector $\mathbf{v} \in V$ can be written as a sum $\mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in V(\lambda)$ and $\mathbf{y} \in W$, and by the induction hypothesis $\mathbf{y}$ can be written as a sum $\sum_\mu \mathbf{w}_\mu$, where μ runs through all the roots of $g(t)$. This proves the existence of a sum decomposition as in (iii). To prove uniqueness, suppose that we are given two possibly different decompositions

$$\mathbf{v} \;=\; \sum_\alpha \mathbf{v}_\alpha \;=\; \sum_\alpha \mathbf{u}_\alpha$$

where $\mathbf{v}_\alpha$ and $\mathbf{u}_\alpha$ lie in $V(\alpha)$. Let $\mathbf{y}_\lambda$ and $\mathbf{z}_\lambda$ denote the corresponding sums over all $\alpha \neq \lambda$, so that

$$\mathbf{v} \;=\; \mathbf{v}_\lambda + \mathbf{y}_\lambda \;=\; \mathbf{u}_\lambda + \mathbf{z}_\lambda \;.$$

By the preceding paragraph we know that $V(\alpha) \subset W$ if $\alpha \neq \lambda$, and therefore both $\mathbf{y}_\lambda$ and $\mathbf{z}_\lambda$ belong to W. Since every vector is uniquely representable as the sum of a vector in $V(\lambda)$ with a vector in W, it follows that $\mathbf{v}_\lambda = \mathbf{u}_\lambda$ and $\mathbf{y}_\lambda = \mathbf{z}_\lambda$. One can now combine the second equation with the induction hypothesis to show that $\mathbf{v}_\alpha = \mathbf{u}_\alpha$ also holds for all $\alpha \neq \lambda$. This completes the proof of (iii).

All that remains is to prove (iv); we shall do this by induction on the dimension of V, noting that the 1-dimensional case is trivial because linear transformations are diagonalizable in this case. Suppose that (iv) is known to be true for linear transformations on all subspaces of dimension $< \dim V$. Since $W(\mu) = V(\mu)$ for all eigenvalues μ of T except λ and the set of all such μ is the set of eigenvalues for S, it follows that one can form a basis for W by taking the union of bases for the subspaces $V(\mu)$. Suppose that we add a basis for $V(\lambda)$; since $\dim V = \dim V(\lambda) + \dim W - \dim(\, V(\lambda) \cap W\,)$ and the intersection is zero-dimensional, it follows that the union of the special basis for W with a basis for V will have the right number of vectors to be a basis, and therefore it will be a basis if the set spans V. But the latter is true because every vector in V is the sum of a vector in $V(\lambda)$ with a vector in W. This completes the proof of the inductive step for the final statement (iv) in the Primary Decomposition Theorem.∎

Nilpotent matrices

Given a linear transformation T on a finite dimensional complex vector space V, we have described subspaces $V(\lambda) \subset V$ such that

(i) for each λ the linear transformation T sends an arbitrary vector $\mathbf{x}$ in $V(\lambda)$ to another vector $T(\mathbf{x}) \in V(\lambda)$,

(ii) if for each *lambda* the set $\mathcal{A}_\lambda$ is a basis for $V(\lambda)$, then the union of the sets $\mathcal{A}_\lambda$ is a basis for V.

It follows that there is an ordered basis for V such that the associated matrix for T is a block sum of matrices $B(\lambda)$, where λ runs through the eigenvalues of T, such that some power of $N(\lambda) = B(\lambda) - \lambda I$ is equal to zero. The next step in deriving the Jordan form is to study the matrices $N(\lambda)$ and find similar matrices that have the least possible numbers of nonzero entries. Here is the main result:

JORDAN FORM FOR NILPOTENT MATRICES. *Let V be a finite dimensional vector space over the real or complex numbers, and let $N : V \to V$ be a linear transformation such that*

$N^q = 0$ for some q (formally, N is **nilpotent***). Then there is a set of linearly independent vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ in V and a sequence of positive integers $p(j)$ for $j \leq k$ such that*

(i) for each j we have $N^{p(j)}(\mathbf{x}_j) = \mathbf{0}$,

(ii) the set $\mathcal{A}$ of all vectors having the form $N^i(\mathbf{x}_j)$, where $1 \leq j \leq k$ and $0 \leq i < p(j)$, forms a basis for V.

If we take the basis $\mathbf{x}_{j,i} = N^i(\mathbf{x}_j)$ and order it as in a dictionary, so that

[A] if the first index of one vector **a** in the collection is smaller than the first index of a second vector **b**, then **a** comes before **b**,

[B] if the first indices of two vectors **a** and **b** are the same but the second index of **a** is smaller than the second index of **b**, then **a** precedes **b**,

then the matrix of N with respect to this ordered basis will be the block sum of elementary Jordan matrices given by

$$N_{p(1)} \oplus \cdots \oplus N_{p(k)} .$$

We shall explain the significance of this for finding the Jordan forms of arbitrary matrices after deriving the result in the nilpotent case.

Derivation of Jordan form for nilpotent linear transformations. First of all, N cannot be invertible because if T is invertible then so is T^m for all positive integers m and we know that N^q is equal to zero, which is not invertible. It follows that the rank of N must be strictly less than the dimension of V.

As before, the proof proceeds by induction. If $\dim V = 1$ we know that N is diagonalizable and the only way that N^q can be zero is if N itself is zero. The required set of vectors is then given by taking an arbitrary nonzero vector in V. Suppose that the result is true for all nilpotent linear transformations on vector spaces of dimension strictly less than $\dim V$.

If W is the image of N, we have seen that $\dim W < \dim V$, and we also that if $\mathbf{x} \in W$ then $N(\mathbf{x})$ also belongs to W, so that one has a linear transformation $N_0 : W \to W$ such that $N_0(\mathbf{w}) = N(\mathbf{w})$ for all $\mathbf{w}$. By the induction hypothesis there is a set of linearly independent vectors $\{\mathbf{y}_1, \cdots, \mathbf{y}_\ell\}$ in W and a sequence of positive integers $r(j)$ for $j \leq \ell$ such that

(i) for each j we have $N^{p(j)}(\mathbf{x}_j) = \mathbf{0}$,

(ii) the set $\mathcal{A}_0$ of all vectors having the form $N^i(\mathbf{x}_j)$, where $1 \leq j \leq \ell$ and $0 \leq i < r(j)$, forms a basis for W.

It follows that the kernel of N_0, which is equal to the intersection of the kernel of N with W, has a basis given by the set $\mathcal{A}_1$ of vectors $N^{r(j)-1}(\mathbf{y}_j)$. Choose a set of vectors

$$\mathcal{B} = \{\mathbf{x}_{\ell+1}, \cdots, \mathbf{x}_k\}$$

such that $\mathcal{A}_1 \cup \mathcal{B}$ forms a basis for the kernel of N, and choose vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_\ell\}$ in V such that $N(\mathbf{x}_j) = \mathbf{y}_j$ for all j. Define $p(j)$ to be $r(j)+1$ if $j \leq \ell$ and $r(j) = 1$ if $j > \ell$.

We claim that the set $\mathcal{A}$ of all vectors $N^i(\mathbf{x}_j)$, where $1 \leq j \leq k$ and $0 \leq i < p(j)$, forms a basis for V. By construction the dimension of the kernel is equal to k and the dimension of the image is equal to $\sum_{j \leq \ell} r(j)$; the sum of these numbers is equal to $\dim V$ by general considerations involving linear transformations, but on the other hand it is also the number of vectors in $\mathcal{A}$. Therefore $\mathcal{A}$ has the right number of elements to be a basis and it is only necessary to show either that $\mathcal{A}$ is linearly independent or that it spans V; we shall verify the second condition.

If $\mathbf{v}$ is an arbitrary vector in V, then we may use our basis for W to write $N(\mathbf{v})$ as a linear combination

$$N(\mathbf{v}) \quad = \quad \sum_{i,j} a_{i,j}\, N^i(\mathbf{y}_j) \quad = \quad \sum_{i,j} a_{i,j}\, N^{i+1}(\mathbf{x}_j) \ .$$

If we let $\mathbf{v}_0 = \sum_{i,j} \ a_{i,j}\, N^i(\mathbf{x}_j)$, then $N(\mathbf{v}_0) = N(\mathbf{v})$, and hence $\mathbf{u} = \mathbf{v} - \mathbf{v}_0$ lies in the kernel of N. Therefore we have written $\mathbf{v}$ as a sum $\mathbf{v}_0 + \mathbf{u}$ where both $\mathbf{v}_0$ and $\mathbf{u}$ are linear combinations of the vectors in $\mathcal{A}$, and therefore it follows that $\mathcal{A}$ spans V.■

Final steps in the derivation

Given a finite dimensional complex vector space V and a linear transformation $T : V \to V$, express its characteristic polynomial in the form

$$\chi_T(t) \quad = \quad \prod_{j=1}^{r} (\lambda_j - t)^{m_j}$$

where the $lmabda_j$'s are the distinct roots of the characteristic polynomial (equivalently, the distinct eigenvalues of T). To simplify notation let $V_j = V(\lambda_j)$.

For each j we know that the image of V_j under T is contained in V_j and hence T defines a linear transformation $T_j : V_j \to V_j$. By the definition of V_j we know that $(T_j - \lambda_j I)^{m_j} = 0$. If $N_j = (T_j - \lambda_j I)$, then it follows that $N_j^{m_j} = 0$, and the results on nilpotent matrices imply the existence of an ordered basis $\mathcal{A}_j$ for V_j such that the matrix of N_j with respect to $\mathcal{A}_j$ is a block sum of elementary Jordan matrices with zeros down the diagonal. Likewise, the identity $T_j = \lambda_j I + N_j$ implies that the matrix of T_j with respect to this ordered basis is a corresponding block sum of elementary Jordan matrices with copies of λ_j down the diagonal. If we order the union of the sets $\mathcal{A}_j$ such that vectors in $\mathcal{A}_j$ precede vectors in $\mathcal{A}_{j+1}$ for all j and the internal ordering of each set $\mathcal{A}_j$ is the same, then the union $\mathcal{A}$ is an ordered basis for V such that the matrix C of T with respect to V is a block sum of matrices C_j, where each summand is itself a block summand of elementary Jordan matrices. It follows that C is also a block sum of elementary Jordan matrices. This establishes the existence of a Jordan form for every square matrix over the complex numbers.■

SIMILARITIES AMONG JORDAN FORMS. Suppose that A is a matrix in Jordan form. For each eigenvalue λ of A let $\mathbf{B}(A, \lambda, k)$ denote the number of $k \times k$ elementary Jordan block summands with copies of λ down the diagonal. If A_1 is another matrix of the same size as A such that

$$\mathbf{B}(A, \lambda, k) \quad = \quad \mathbf{B}(A_1, \lambda, k)$$

for all k, then A_1 represents the same linear transformation as A with respect to a different ordered basis; in fact, one can obtain such an ordered basis $\mathcal{U}_1$ simply by reordering the standard unit vector basis of $\mathbb{C}^n$. The following result includes a converse to this observation:

UNIQUENESS OF JORDAN FORMS. *In the notation above, suppose that A_1 and A_2 are matrices that are in Jordan form. Then A_1 is similar to A_2 if and only if A_1 and A_2 have the same eigenvalues and*

$$\mathbf{B}(A_1, \lambda, k) \quad = \quad \mathbf{B}(A_2, \lambda, k)$$

for each common eigenvalue λ and positive integer k.

Proof. We have already noted that two matrices satisfying the conditions of the theorem are similar. Conversely, suppose that A_1 and A_2 are similar and that P is an invertible matrix such that $A_2 = P^{-1} A_1 P$. Then we have already seen that A_1 and A_2 have the same eigenvalues and that the dimensions of the corresponding eigenspaces are equal. Direct examination shows that these dimensions are equal to $\sum_k \mathbf{B}(A_i, \lambda, k)$ for $i = 1$ or 2, and therefore these summations are equal.

More generally, if $m \geq 1$ then the null spaces $W(\lambda, k, A_i)$ of the matrices $(A_i - \lambda I)^k$ are equal because $W(\lambda, k, A_2)$ is the image of $W(\lambda, k, A_1)$ under left multiplication by the invertible matrix P^{-1}. Another direct examination shows that the dimensions of these null spaces are respectively equal to

$$\sum_{k \leq m} k\,\mathbf{B}(A_i, \lambda, k) \;+\; \sum_{k > m} m\,\mathbf{B}(A_i, \lambda, k)$$

and therefore we have

$$\sum_{k \leq m} k\,\mathbf{B}(A_1, \lambda, k) \;+\; \sum_{k > m} m\,\mathbf{B}(A_1, \lambda, k) \quad = \quad \sum_{k \leq m} k\,\mathbf{B}(A_2, \lambda, k) \;+\; \sum_{k > m} m\,\mathbf{B}(A_2, \lambda, k)$$

for all m and λ. If for each m we subtract equation $m - 1$ in this list from equation m (agreeing that equation 0 is just $0 = 0$), we obtain an equivalent system of equations

$$\sum_{k > m} \mathbf{B}(A_1, \lambda, k) \quad = \quad \sum_{k > m} \mathbf{B}(A_2, \lambda, k)$$

and if we perform the same operations to this new system we obtain the equations

$$\mathbf{B}(A_1, \lambda, m) \quad = \quad \mathbf{B}(A_2, \lambda, m)$$

for all m. Therefore we have shown that two Jordan forms are similar if and only if the conditions in the uniqueness statement hold.∎

Special considerations for real matrices

If A is a matrix with real entries and the characteristic polynomial for A has only real roots, then the preceding construction yields an invertible $n \times n$ matrix P *over the real numbers* such that $P^{-1} A\,P$ is in Jordan form.∎

IV.5 : Differential equations revisited

(Fraleigh and Beauregard, §9.4)

In Section I.3 we discussed solutions of linear systems of equations having the form $Y' = AY$ where A is a square matrix of constants that is diagonalizable over the real numbers. The purpose of this section is to discuss solutions to such systems when A does not have a basis of real eigenvectors. In order to analyze these situations systematically, we need to work with exponentials of complex numbers and matrices. The former suffice in cases where there is a basis of complex eigenvectors, and we need the latter for matrices that do not admit a basis of eigenvectors even if we work over $\mathbb{C}$. It is best to consider these situations separately.

Real matrices with complex eigenvalues

If we are given a quadratic polynomial over the real numbers with no real roots, then it has two distinct complex roots, and each of these roots is conjugate to the other. We begin with a far-reaching generalization of this fact:

SYMMETRY PRINCIPLE FOR COMPLEX ROOTS. *Suppose that $p(t)$ is a real polynomial and that $c = a + b\,i$ is a complex root of p, where $b \neq 0$. Then $\overline{c} = c = a - b\,i$ is also a root of p.*

Proof. Since complex conjugation sends the conjugate of a sum or product into the sum or product of the conjugates (respectively), it follows that if p is a real polynomial and z is a complex number then we have

$$p(\,\overline{z}\,) \quad = \quad \overline{p(z)}$$

and therefore if $p(z) = 0$ then we also have $p(\overline{z}) = 0$.■

COMPLEMENT. *In the setting above, if the root $a + b\,i$ has algebraic multiplicity μ then its conjugate $a - b\,i$ also has algebraic multiplicity μ.*

Proof. We prove this by induction on the degree of p. If p is a real polynomial of degree 1, then it only has one root and that root is real, so there are no nonreal roots to consider. Suppose now that p has degree 2. If p has a nonreal root in this case, then the conjugate is also a root and the polynomial factors into a nonzero constant times $(t - z)\,(t - \overline{z})$, where $z = a + b\,i$, and therefore if $\deg p = 2$ each of the two complex roots must have multiplicity 1. Suppose now that we know the multiplicity statement for all real polynomials of degree less than n, where $n \geq 3$, and let $p(t)$ be a polynomial of degree n such that $z = a + b\,i$ is a root of multiplicity μ. By the symmetry principle we know that $\overline{z}$ is a root of p, and therefore we may write

$$p(t) = (t - z) \cdot (t - \overline{z}) \cdot q(t)$$

for a suitable quotient polynomial $q(t)$ Now the product of the first two factors is the quadratic polynomial with real coefficients

$$q_0(t) = t^2 - 2\,a\,t + (a^2 + b^2)$$

and therefore the polynomial $q(t) = p(t)/q_0(t)$ also has real coefficients. By the inductive hypothesis we also know that z is a root of $q(t)$ with multiplicity $\mu - 1$. Therefore the induction hypothesis implies that $\overline{z}$ is also a root of $q(t)$ with the same multiplicity, which means that we may write $q(t)$ in the form

$$(t-z)^{\mu-1} \cdot (t-\overline{z})^{\mu-1} \cdot r(t)$$

where $r(t)$ is a real polynomial such that neither z nor $\overline{z}$ is a root of $r(t)$. But this implies that

$$p(t) = (t-z)^{\mu} \cdot (t-\overline{z})^{\mu} \cdot r(t)$$

which in turn implies that $\overline{z}$ is a root of $p(t)$ with multiplicity μ.■

There is a corresponding symmetry principle for complex eigenvectors associated to the nonreal eigenvalues of a real matrix.

SYMMETRY PRINCIPLE FOR COMPLEX EIGENVECTORS. *Suppose that A is an $n \times n$ matrix with real entries and $\mathbf{v} \in \mathbb{C}^n$ is an eigenvector associated to the nonreal eigenvalue λ. Let $\overline{v}$ denote the vector whose coordinates are the complex conjugates of the corresponding coordinates for $\mathbf{v}$. Then $\overline{v}$ is an eigenvector for A associated to the eigenvalue $\overline{\lambda}$.*

Proof. Since complex conjugation sends sums to sums and product to products, it follows that the matrix operation $\overline{A}$, which sends a matrix with entries $a_{i,j}$ to the matrix with entries $\overline{a_{i,j}}$, satisfies the identity

$$\overline{A \cdot B} \quad = \quad \overline{A} \cdot \overline{B} \ .$$

If we apply this to both sides of the equation $A\,\mathbf{v} = \lambda\mathbf{v}$ and note that $A = \overline{A}$, it follows that

$$\overline{\lambda} \cdot \overline{\mathbf{v}} \quad = \quad \overline{\lambda\,\mathbf{v}} \quad = \quad \overline{A\,\mathbf{v}} \quad = \quad A\,\overline{\mathbf{v}}$$

which proves the desired symmetry property.■

Real matrices with complex eigenvalues

Using the results on complex exponentials and differentiation from Section IV.1, one can generalize the entire discussion of differential equations in Section I.3 to matrices over the complex numbers that are diagonalizable. The solutions will then be linear combinations of vector valued functions with the form $\exp(\lambda t)\mathbf{v}$ where $\mathbf{v}$ is a complex eigenvector for the complex eigenvalue λ.

Suppose now that we have a system of linear differential equations $Y' = AY$ where A has real entries and is diagonalizable over $\mathbb{C}$ but not over $\mathbf{R}$. For each complex conjugate pair $\{\lambda, \overline{\lambda}\}$ of nonreal eigenvalues for A and each conjugate pair of associated eigenvectors $\{\mathbf{v}, \overline{\mathbf{v}}\}$ one would like to find a pair of real vector valued functions $f(t)$ and $g(t)$ that are equivalent to $\exp(\lambda t)\mathbf{v}$ and $\exp(\overline{\lambda}t)\overline{\mathbf{v}}$; more precisely, each linear combination of the latter pair should be a linear combination of $\{f(t),\, g(t)\}$ and vice versa. This is similar to the situation for linear differential equations of the form

$$a\,y'' \;+\; b\,y' \;+\; c\,y \quad = \quad 0$$

where $b^2 - 4\,a\,c < 0$; in this case one has complex solutions $\exp(\lambda t)$ and $\exp(\overline{\lambda}t)$, but for many purposes it is preferable to use the the corresponding real solutions $e^{pt}\cos q\,t$ and $e^{pt}\sin q\,t$, where $\lambda = p + q\,i$, and the basic approach to find solutions for the system $Y' = AY$ is essentially the same. The first step is to write out $\exp(\lambda t)\mathbf{v}$ explicitly using the expansion $\lambda = p + q\,i$ and the corresponding expansion

$$\mathbf{v} \quad = \quad \mathbf{x} \;+\; i\,\mathbf{y}$$

where $\mathbf{x}$ and$\mathbf{y}$ have real coordinates. This leads directly to the following equation:

$$\exp(\lambda t)\mathbf{v} \quad = \quad e^{pt}\,(\cos qt + i\,\sin qt)\,\big(\mathbf{x} \,+\, i\,\mathbf{y}\big) \quad = \quad e^{pt}\,\big(\cos qt\,\mathbf{x} - \sin qt\,\mathbf{y}\big) \;+\; i\,e^{pt}\,\big(\sin qt\,\mathbf{x} + \cos qt\,\mathbf{y}\big)$$

If we write out the solution $\exp(\overline{\lambda} t)\overline{\mathbf{v}}$ corresponding to the conjugate eigenvalue and eigenvector, we obtain the conjugate of the expression on the right hand side of the equation displayed above. Consequently, if we write

$$\exp(\lambda t)\mathbf{v} \quad = \quad \mathbf{f}(t) \;+\; i\,\mathbf{g}(t)$$

then the functions $\mathbf{f}(t) \pm i\,\mathbf{g}(t)$ solve the original linear system, and the same is true for all linear combinations of these functions with complex coefficients. Since $\mathbf{f}(t)$ and $\mathbf{g}(t)$ are linear combiations of this sort (how?), they also solve the original system of linear equations. Since the linear combinations of this pair are the same as those involving the previously described pair, it follows that

$$\mathbf{f}(t) = e^{pt}\,\big(\cos qt\,\mathbf{x} - \sin qt\,\mathbf{y}\big)$$

$$\mathbf{g}(t) = e^{pt}\,\big(\sin qt\,\mathbf{x} + \cos qt\,\mathbf{y}\big)$$

are the sorts of functions we wanted.■

EXAMPLE. We previously showed that the 2×2 matrix

$$\begin{pmatrix} 3 & -2 \\ 5 & 1 \end{pmatrix}$$

has complex eigenvalues $2 \pm 3\,i$ and associated complex eigenvectors

$$\begin{pmatrix} 2 \\ 1 \mp 3i \end{pmatrix} .$$

In terms of the preceding general derivation we have $p = 2$, $q = 3$, $\mathbf{x}$ is the transpose of $(2, 1)$, and $\mathbf{y}$ is the transpose of $(0, 3)$. The solutions to the system of differential equations $Y' = AY$ with initial conditions $\mathbf{x}$ and $\mathbf{y}$ at $t = 0$ are then given by the following formulas:

$$e^{pt}\,\cos qt \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad - \quad e^{pt}\,\sin qt \begin{pmatrix} 0 \\ 3 \end{pmatrix}$$

$$e^{pt}\,\sin qt \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad + \quad e^{pt}\,\cos qt \begin{pmatrix} 0 \\ 3 \end{pmatrix}$$

Multiple roots and differential equations

If we are given a system of linear differential equations $Y' = AY$ where A is not diagonalizable, then additional work is needed to describe all the possible solutions. One relatively abstract approach to solving such systems is given in the document `expmatrix.pdf` from the course directory, but for many purposes we would like to have more concrete expressions for the solutions to such a system. We shall use the results from `expmatrix.pdf` together with Jordan forms in describing some basic sets of solutions in relatively simple terms.

The key points we need from `expmatrix.pdf` are as follows:

[1] The unique solutions fo $Y' = AY$ with initial condition $Y(0) = \mathbf{v}$ is given by $\exp(tA)\mathbf{v}$, where $\exp(tA)$ is the exponential series.

[2] The matrix exponential satisfies $\exp(P^{-1}\,A\,P) = P^{-1}\exp(A)\,P$ for all matrices A and invertible matrices P.

[3] If c is a scalar than $\exp(cI) = \exp(c)\,I$.

[4] If $A\,B = B\,A$ then $\exp(A+B) = \exp(A)\,\exp(B)$.

It will also be convenient to note the following consequence of the fourth property:

BLOCK SUM FORMULA. *If the square matrix A is the block sum of the square matrices A_j, then*

$$\exp(A) \quad = \quad \exp(A_1) \oplus \exp(A_2) \oplus \ \cdots \ \oplus \exp(A_n) \ .$$

Comments on the derivation. We shall only consider the special case where $n = 2$. The fourth property applies because of the identities

$$A_1 \oplus A_2 \quad = \quad (A_1 \oplus I) \cdot (I \oplus A_2) \quad = \quad (I \oplus A_2) \cdot `(A_1 \oplus I)$$

$$\big(A_1 \oplus A_2\big)^k \quad = \quad \big(A_1^k \oplus A_2^k\big)$$

and a corresponding formula for infinite series:

$$\sum_{j=0}^{\infty} B_j \oplus C_j \quad = \quad \left(\sum_{j=0}^{\infty} B_j\right) \quad \oplus \quad \left(\sum_{j=0}^{\infty} C_j\right)$$

Property [2] and the Block Sum Formula show that one can find a simplified description of solutions to $Y' = AY$ by finding a matrix $P^{-1}\,A\,P$ that is in Jordan form. In particular, the Block Sum Formula demonstrates the importance of understanding the exponential $\exp(tA)$ when A is an elementary Jordan matrix.

Suppose now that we have an elementary Jordan Matrix $A = \lambda I + N_k$ where λ is some scalar and N_k is the matrix described in Section IV.4. Since λI and N_k commute with each other we have the following chain of equations:

$$\exp(tA) \quad - \quad \exp(\lambda I + N_k) \quad = \quad \exp(\lambda t I)\,\exp(tN_k) \quad = \quad \exp(\lambda t)\,I\,\exp(tN_k) \quad =$$

$$\exp(\lambda t)\,\exp(tN_k) \quad = \quad \exp(\lambda t) \cdot \sum_{j=0}^{k-1} \frac{t^j}{j!}\,N^j$$

The power series for $\exp(tN_k)$ reduces to a finite sum over the first k terms because $N_k^j = 0$ for $j \geq k$.

All of this implies the following procedure for describing the solutions of the linear system of differential equations $Y' = AY$. First, find an invertible matrix P such that $B = P^{-1}\,A\,P$ is in Jordan form. Next, use the method of the preceding paragraph to compute $\exp(tB)$ explicitly. Finally obtain a basic set of solutions from the following result:

JORDAN FORM SOLUTIONS. *In the notation above, let $\mathbf{p}_j$ denote the j^{th} column of P. Then the unique solution of $Y' = AY$ with initial condition $Y(0) = \mathbf{p}_j$ is equal to*

$$\sum_{k=1}^{n} z_k(t)\,\mathbf{p}_k$$

where $z_k(t)$ *is the* k^{th} *coordinate of* $\exp(tB)$*.*

The significance of this formula is that $\exp(tB)$ is easy to compute once one has the Jordan form B and the matrix of basis vectors P that will put A into this Jordan form. **Derivation.** If $\mathbf{e}_j$ denotes the j^{th} standard unit vector in $\mathbf{R}^n$ or $\mathbb{C}^n$, then $\mathbf{p}_j = P\,\mathbf{e}_j$. Therefore the solution to the system of differential equations with initial condition $\mathbf{p}_j$ may be written in the form

$$\exp(tA)\,P\,\mathbf{e}_j \quad = \quad P\,P^{-1}\,\exp(tA)\,P\,\mathbf{e}_j \quad = \quad P\,\exp(tB)\,\mathbf{e}_j \quad = \quad P\,Z_j(t)$$

where $Z_j(t)$ is the j^{th} column of $\exp(tB)$. It is a routine exercise to verify that this column vector is the linear combination of the columns of P that is described in the formula.■

We shall illustrate all this with some simple examples.

EXAMPLE 1. Let A be the elementary 2×2 Jordan matrix $\lambda I + N_2$, where λ is real. Then $\exp(tA) = e^{\lambda t}\,(I + t\,N_2)$ because $N_2^2 = 0$, and therefore we have

$$\exp(tA) \quad = \quad e^{\lambda t} \cdot \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \ .$$

Therefore the general solution to $Y' = A\,Y$ is given by

$$c_1\,e^{\lambda t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad + \quad c_2\,e^{\lambda t} \begin{pmatrix} t \\ 1 \end{pmatrix}$$

where c_1 and c_2 are arbitrary (real or complex) scalars.

EXAMPLE 2. Let A be the matrix

$$\begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$$

considered in Section III.1, and consider the system of differential equations given by $Y' = A\,Y$. If P is the invertible matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

we showed that $B = P^{-1}\,A\,P$ is the elementary Jordan matrix $2I + N_2$, and if $\mathbf{p}_j$ denotes the j^{th} column of P then the formula reduces to

$$Y(t) \quad = \quad (c_1 + c_2\,t)\,e^{2t}\,\mathbf{w}_1 \ + \ c_2\,e^{2t}\,\mathbf{w}_2$$

where c_1 and c_2 are suitable constants. In particular, if we want the solution with initial condition $Y(0) = \mathbf{v}$, the we choose these constants so that

$$\mathbf{v} = c_1\,\mathbf{v}_1 \ + \ c_2\,\mathbf{v}_2 \ .$$

EXAMPLE 3. Let A be the matrix

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 1 & 1 \\ -1 & 0 & 2 \end{pmatrix} \ .$$

If P is the invertible matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

then we have

$$B \quad = \quad P^{-1}\, A\, P \quad = \quad \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

which is an elementary Jordan matrix, so that

$$\exp(tB) \quad = \quad e^t \left(I + t\, N_3 + \frac{1}{2}\, t^2 N_3^2\right) \quad =$$

$$e^t \cdot \begin{pmatrix} 1 & t & \frac{1}{2}\, t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix} \quad .$$

It follows that the solutions of $Y' = A\,Y$ in this case are given by

$$c_1\, e^t\, \mathbf{p}_1 \ + \ c_2\, e^t\ (t\, \mathbf{p}_1 + \mathbf{p}_2) \ + \ c_3\, e^t\ \left(\tfrac{1}{2}\, t^2\, \mathbf{p}_1 + t\, \mathbf{p}_2 + \mathbf{p}_3\right)$$

where the c_j are suitable constants (as before, we want $Y(0) = \sum_j c_j\, \mathbf{p}_j$).

V. Quadratic forms

The Fundamental Theorem on Real Symmetric Matrices is an extremely important result that sheds light on a wide range of questions in mathematics and other subjects, and the final unit of the course deals with some relatively elementary uses of this result that are related to ordinary and multivariable calculus. Unless stated otherwise, we shall be dealing only with vector spaces over the real numbers in this unit.

A linear transformation on $\mathbf{R}^n$ may be viewed as a function such that each coordinate is a homogeneous linear expression in the coordinates; in other words, each coordinate is a first degree polynomial in n variables with no constant term. Similarly, a **quadratic form** is a scalar valued function that is a second degree polynomial in n variables with no constant or first degree terms:

$$q(\mathbf{x}) \quad = \quad \sum_{i,j=1}^{n} c_{i,j}\, x_i\, x_j$$

In order to have something that is nontrivial, one generally assumes that at least one of the coefficient sums $c_{i,j} + c_{j,i}$ is nonzero. One can similarly introduce cubic forms involving degree three polynomials and even forms of higher degrees, but the quadratic case is particularly important because *many phenomena in mathematics and other subjects can be studied very effectively using second degree expressions or approximations.* Two particular examples from this unit are (*i*) classical geometrical objects such as conics and their generalizations to three or more dimensions, (*ii*) the second derivative test for relative maxima and minima of a function.

Numerous other examples could also be cited; for example, Simpson's Rule in elementary calculus provides an extremely good method for approximating definite integrals of functions in terms of a well chosen family of quadratic approximations to the given function, and the logistic differential equation $y' = a\,y - b\,y^2$ provides a good model of population growth which reflects (1) the fact that growth occurs at a nearly exponential rate at first, (2) the inevitable leveling off of population growth with the passage of time. Further applications of quadratic forms in other areas of science are mentioned very briefly on page 408 of the text.

V.1 : Diagonalization of quadratic forms

(Fraleigh and Beauregard, §§6.3, 8.1)

The first step in analyzing quadratic forms by means of matrices is to note that every quadratic form can be expressed in the form

$$q(\mathbf{x}) \quad = \quad {}^{\mathbf{T}}\mathbf{x}\, A\, \mathbf{x}$$

where A is a symmetric matrix. Specifically, if we are given a nontrivial expression

$$q(\mathbf{x}) \quad = \quad \sum_{i,j=1}^{n} c_{i,j}\, x_i\, x_j$$

where at least one sum $c_{i,j} + c_{j,i}$ is nonzero, then we may take A to be the matrix whose entries are given by

$$a_{i,j} \quad = \quad \frac{c_{i,j} + c_{j,i}}{2}$$

and the nontriviality condition then translates into a statement that A is nonzero.

In many cases one can simplify a quadratic form by a change of variables. For example, if we are given the quadratic form in two variables

$$q(x,y) \quad = \quad x^2 \ + \ 2\,x\,y \ + \ 2\,y^2$$

and we set $u = (x+y)$ and $v = y$, then this simplifies to $u^2 + v^2$. For our purposes it is useful to view such a change of variables as a providing a way to write our given variables x and y in terms of new variables u and v, and from this perspective the change of variables can be viewed as a matrix equation

$$\begin{pmatrix} x \\ y \end{pmatrix} \quad = \quad \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} \ .$$

More generally, if we are given a quadratic form

$$q(\mathbf{x}) \quad = \quad {}^{\mathbf{T}}\mathbf{x}\,A\,\mathbf{x}$$

and a linear homogeneous change of variables $\mathbf{x} = P\,\mathbf{z}$, where P is some invertible matrix P, then we have the following results:

CHANGE OF VARIABLES RELATIONSHIP. *In the notation above, the function $f(\mathbf{z}) = q_A(\mathbf{x}) = q_A(P\,\mathbf{z})$ is equal to $q_B(\mathbf{z})$, where $B = {}^{\mathbf{T}}P\,A\,P$.*

This follows immediately by substituting $\mathbf{x} = P\,\mathbf{z}$ into q_A.■

The Fundamental Theorem on Real Symmetric Matrices then yields the following powerful simplification principle:

DIAGONALIZATION THEOREM. *In the notation above, there is an orthogonal linear change of variables $\mathbf{x} = P\,\mathbf{z}$ such that*

$$q_A(\mathbf{x}) \quad = \quad \sum_{j=1}^{n} \lambda_j\, z_j^2 \ .$$

Proof. All we need to do is find an orthogonal matrix P so that $B = {}^{\mathbf{T}}P\,A\,P.$ is diagonal; such matrices exist by The Fundamental Theorem on Real Symmetric Matrices.■

For many purposes (in physics, for example), it is important to know the maximum value of the normalized quadratic function

$$r_A(\mathbf{x}) \quad = \quad \frac{q_A(\mathbf{x})}{|\mathbf{x}|^2}$$

for $\mathbf{x} \neq \mathbf{0}$ which is the same as the maximum value of $q_A(\mathbf{x}$ on the unit sphere defined by the equation $|\mathbf{x}|^2 = 1$. The diagonalization theorem leads immediately to the following result:

RAYLEIGH'S PRINCIPLE. *The maximum and minimum values of $r_A(\mathbf{x})$ for $\mathbf{x} \neq \mathbf{0}$ are realized at eigenvectors of A, and they are given by the largest and smallest eigenvalues of A.*

Incidentally, one can also derive this conclusion using the Lagrange Multiplier Rule for maximization and minimization of $q_A(\mathbf{x})$ under the constraint that $\mathbf{x}$ lies on the unit sphere $|\mathbf{x}|^2 = 1$.

Derivation of Rayleigh's Principle from diagonalizaton. Let P, $\mathbf{z}$, λ_j be as above, and suppose that the eigenvalues are ordered so that $\lambda_1 \leq \ \cdots \ \leq \lambda_n$. Then we have

$$q_A(\mathbf{x}) \quad = \quad \sum_{j=1}^{n} \lambda_j \, z_j^2$$

where the right hand side satisfies

$$\lambda_1 \cdot \sum_{j=1}^{n} z_j^2 \quad \leq \quad \sum_{j=1}^{n} \lambda_j \, z_j^2 \quad \leq \quad \lambda_n \cdot \sum_{j=1}^{n} z_j^2 \ .$$

Dividing these equations by $|\mathbf{z}|^2$ and noting that the latter is equal to $|\mathbf{x}|^2 = |P(\mathbf{z})|^2$ because P is orthogonal, we obtain the following inequalities:

$$\lambda_1 \quad \leq \quad r_A(\mathbf{x}) \quad \leq \quad \lambda_n$$

Furthermore, by construction it is clear that the maximum and minimum values are obtained when $\mathbf{z}$ is equal to the last and first unit vectors $\mathbf{e}_n$ and $\mathbf{e}_1$ respectively. In these cases $\mathbf{x} = P\,\mathbf{z}$ is an eigenvector associated to λ_n and λ_1 respectively.■

One can use this principle to make some (generally not too precise) estimates for the sizes of the maximum and minimum eigenvalues. Here are some simple examples:

SIMPLE RAYLEIGH ESTIMATES. *If A is a real symmetric $n \times n$ matrix where $n \geq 2$, then each diagonal entry $a_{j,j}$ lies in the interval bounded by the smallest and largest eigenvalue of A. Similarly, if $i \neq j$ and k and ℓ are arbitrary integers, then the same is true for*

$$\frac{\ell^2 a_{i.i}^2 + 2\,k\,\ell\,a_{i,j} + k^2\,a_{j,j}}{\ell^2 + k^2} \ .$$

These follow by applying Rayleigh's principle to the unit vectors and to the vectors $\ell\,\mathbf{e}_i + k\,\mathbf{e}_j$.■

V.2 : Classification of quadrics

(Fraleigh and Beauregard, §8.2)

Both plane and solid analytic geometry spend considerable time discussing the sets of points in $\mathbf{R}^2$ and $\mathbf{R}^3$ whose coordinates satisfy some quadratic polynomial equation. Usually calculus textbooks describe a short list of standard examples in great detail and either suggest or assert that **ALL** solution sets for quadratic polynomials can be transformed into the standard examples by a suitable change of variables. The Fundamental Theorem on Real Symmetric Matrices is exactly what is needed to prove such a statement.

A genuine quadratic polynomial in n variables (*i.e.*, one that actually has second degree terms) can be written in the form

$$f(\mathbf{x}) \quad = \quad q_A(\mathbf{x}) \; + \; \mathbf{b} \cdot \mathbf{x} \; + \; c$$

where $\mathbf{x}$ is an $n \times 1$ column vector whose entries are the variables, q_A is the quadratic form of a nonzero symmetric matrix A, c is a constant, and $\mathbf{b}$ is a $1 \times n$ row vector. The first step in simplifying this expression is to diagonalize the quadratic form as in the preceding section using a change of variables $\mathbf{x} = P(\mathbf{z})$, where P is an orthogonal matrix. If we make this change of variables we obtain an equivalent quadratic equation of the form

$$g(\mathbf{z}) \quad = \quad q_D(\mathbf{z}) \; + \; \mathbf{p} \cdot \mathbf{z} \; + \; c'$$

where D is diagonal.

Without loss of generality, we may assume that the diagonal entries are ordered so that the nonzero entries preceded the zero ones. Suppose that there are r nonzero diagonal entries

RIGID CHANGES OF COORDINATES. *There exists an orthogonal matrix Q and a vector $\mathbf{d}$ such that if $\mathbf{z} = Q\,\mathbf{w} + \mathbf{d}$, then*

$$h(\mathbf{w}) \quad = \quad g(Q\,\mathbf{w} + \mathbf{d})$$

has one of the following forms:

$$\sum_{j=1}^{r} d_j\, w_j^2 \; + \; c''$$

$$\sum_{j=1}^{r} d_j\, w_j^2 \; + \; k\, w_{r+1}$$

In these expressions c'' and k denote constants, and if $r = n$ it is understood that the second possibility does not arise.

Derivation. Suppose we are given g with the quadratic form diagonalized as above, and assume that the nonzero diagonal entries of D occur in the first r rows/columns. The separation of cases depends upon whether or not the nonzero coordinates of the vector $\mathbf{p}$ all appear in the first r positions.

CASE 1. Assume that the nonzero coordinates of $\mathbf{p}$ all appear in the first r positions. Then we may write $g(\mathbf{z})$ in the form

$$\sum_{j=1}^{r} d_j\, z_j^2 \; + \; \sum_{j=1}^{r} p_j\, z_j \; + \; c'$$

and if we define $\mathbf{w}$ by

$$z_j \quad = \quad w_j - \tfrac{1}{2}\, p_j$$

for $j \leq r$ and $z_j = w_j$ otherwise (so that $p_j = 0$), it follows that one obtains a function $h(\mathbf{w}) = g(\mathbf{w} + \frac{1}{2}\mathbf{p})$ of the first type.

CASE 2. Assume there is at least one nonzero coordinate of $\mathbf{p}$ in the j^{th} position for some $j > r$. In this case the argument of Case 1 yields a function

$$h_0(\mathbf{v}) \quad = \quad \sum_{j=1}^{r} d_j\, v_j^2 \;+\; \sum_{j=r+1}^{n} s_j\, v_j \;+\; c''$$

It is assumed that at least one of the coefficients s_j is nonzero. Let $\mathbf{s} \neq \mathbf{0}$ be the vector whose coordinates are equal to s_j (and zero in the first r positions), and suppose that Q_0 is an orthogonal matrix whose first r rows are the standard unit vectors and whose next row is equal to $|\mathbf{s}|^{-1}\,\mathbf{s}$. If $Q = {}^{\mathbf{T}}Q_0$ and we make the change of variables $\mathbf{v} = Q(\mathbf{u})$, then $h_0(\mathbf{v})$ is transformed into the following function:

$$h_1(\mathbf{u}) \quad = \quad \sum_{j=1}^{r} d_j\, u_j^2 \;+\; k\, u_{r+1} \;+\; c''$$

Finally, if we set $u_j = w_j$ for $j \neq r+1$ and $w_{r+1} = u_{r+1} - (c/k)$, then $h_1(\mathbf{u})$ has the desired form $\sum_{j=1}^{r}\, d_j\, w_j^2 \;+\; k\, w_{r+1}$.■

Suppose now that we have a quadric in $\mathbf{R}^n$ defined by an expression $f(\mathbf{x}) = 0$ as above. We have shown that there is a change of variables having the form $\mathbf{x} = C(\mathbf{w}) + \mathbf{y}$ where C is an orthogonal matrix and $f(\mathbf{x}) = f\big(C(\mathbf{w}) + \mathbf{y}\, bigr) = h(\mathbf{w})$. It follows that the sets where $f(\mathbf{x}) = 0$ and $h(\mathbf{w}) = 0$ are equal. Therefore we have shown that *for every quadric in $\mathbf{R}^n$ there is a rigid change of variables for which the quadric is given by an equation having one of the types described above.*■

In order to write down the classification by standard types, one more step is needed. Namely, we need to let r_+ denote the number of positive eigenvalues of D. Then the standard forms we have obtained for quadrics in $\mathbf{R}^n$ can be classified in terms of the number r of nonzero eigenvalues, the number r_+ of positive eigenvalues, and the constant c' for equations of Type I. If we divide our quadratic polynomial by a nonzero constant, this will have no effect on the set of solutions to the polynomial, and therefore we can always arrange things so that the constant is equal to either -1 or 0. This means that we can consolidate things further in terms of the numbers of positive and nonzero eigenvalues for the second degree part of the polynomial and three possibilities for the form of the the nonquadratic part; namely, it is either a constant times some coordinate, it is equal to -1 or it is equal to zero. In dimensions 2 and 3 these classifications can be summarized by tables as follows:

STANDARD FORMS FOR CONICS IN THE PLANE

r	r_+	NONQUADRATIC PART	TYPICAL EXAMPLE	STANDARD DESCRIPTION
2	2	-1	$x^2 + y^2 = 1$	**ellipse**
2	2	0	$x^2 + y^2 = 0$	*one point*
2	1	-1	$x^2 - y^2 = 1$	**hyperbola**
2	1	0	$x^2 - y^2 = 0$	*pair of intersecting lines*
2	0	-1	$x^2 + y^2 = -1$	*no points*
2	0	0	$x^2 + y^2 = 0$	*one point*
1	1	linear	$x^2 = y$	**parabola**
1	1	-1	$x^2 = 1$	*pair of parallel lines*
1	1	0	$x^2 = 0$	*one line*
1	0	linear	$x^2 = y$	**parabola**
1	0	-1	$x^2 = -1$	*no points*
1	0	0	$x^2 = 0$	*one line*

Nondegenerate examples not consisting of one or two lines or one or zero points are designated by **boldface** type, and the remaining examples are designated using *italic* type.■

Note that many lines in the table deal with degenerate situations where the conic reduces to a one or two lines, a point, or the empty set. The corresponding table in the three-dimensional case appears on the next page:

STANDARD FORMS FOR CONICS IN 3-SPACE

r	r_+	NONQUADRATIC PART	TYPICAL EXAMPLE	STANDARD DESCRIPTION
3	3	-1	$x^2 + y^2 + z^2 = 1$	**ellipsoid**
3	3	0	$x^2 + y^2 + z^2 = 0$	*one point*
3	2	-1	$x^2 + y^2 - z^2 = 1$	**one-sheeted hyperboloid**
3	2	0	$x^2 + y^2 - z^2 = 0$	**elliptic cone**
3	1	-1	$x^2 - y^2 - z^2 = 1$	**two-sheeted hyperboloid**
3	1	0	$x^2 - y^2 - z^2 = 0$	**elliptic cone**
3	0	-1	$x^2 + y^2 + z^2 = -1$	*no points*
3	0	0	$x^2 + y^2 + z^2 = 0$	*one point*
2	2	linear	$x^2 + y^2 = z$	**elliptic paraboloid**
2	2	-1	$x^2 + y^2 = 1$	**elliptic cylinder**
2	2	0	$x^2 + y^2 = 0$	*line*
2	1	linear	$x^2 - y^2 = z$	**hyperbolic paraboloid**
2	1	-1	$x^2 - y^2 = 1$	**hyperbolic cylinder**
2	1	0	$x^2 - y^2 = 0$	*pair of intersecting planes*
2	0	linear	$x^2 + y^2 = z$	**elliptic paraboloid**
2	0	-1	$x^2 + y^2 = -1$	*no points*
2	0	0	$x^2 + y^2 = 0$	*one line*
1	1	linear	$x^2 + y^2 = z$	**parabolic cylinder**
1	1	-1	$x^2 = 1$	*pair of parallel planes*
1	1	0	$x^2 = 0$	*one plane*
1	0	linear	$x^2 = y$	**parabolic cylinder**
1	0	-1	$x^2 = -1$	*no points*
1	0	0	$x^2 = 0$	*one plane*

As on the previous page, nondegenerate examples not given by collections of points, lines or planes are noted in **boldface** type, and the remaining examples are designated using *italic* type.■

A similar table could be constructed for quadrics in four-dimensional space. It would have fourteen additional lines, with ten of them coming from examples with $r = 4$, $r_+ \leq 4$, and non-quadratic parts that are either -1 or 0, and additional four coming from examples with $r = 3$, $r_+ \leq 3$ and nonquadratic parts that are linear.

V.3 : Classification of critical points

(Fraleigh and Beauregard, §8.3)

This section deals with multivariable versions of the second derivative test for relatve maximum and minimium values of a function that has continuous second partial derivatives. In single variable calculus, the second derivative test for such values has two basic steps:

(1) Find the points x for which the derivatitive $f'(x)$ of the function f is equal to zero (or is undefined).

(2) For each x such that $f'(x) = 0$, find the second derivative $f''(x)$. If it is positive, then f has a relative minimum at x , but if it is negative then f has a relative maximum at x. If $f''(x) = 0$, then the test yields no information on whether there is a realtive maximum, relative minimum, or neither.

Generalzations of this result to functions of two, and sometimes even three, variables, are often found in the sections of calculus texts that discuss partial differentiation. but generally the justification for these tests is not presented because it involves and understanding of the following question about matrices:

POSITIVITY QUESTION. *Let A be a real symmetric matrix, and let q_A be the quadratic form defined by $q_A(\mathbf{x}) = {}^{\mathbf{T}}\mathbf{x}\, A\, \mathbf{x}$. Under what conditions on A can we conclude that $q_A(\mathbf{x})$ is positive for every nonzero vector $\mathbf{x}$?*

A symmetric matrix satisfying the condition in the preceding question is said to be **positive definite**. The main algebraic result of this section gives an answer to this question in relatively simple terms. We shall state the main result now and prove it after comploeting the discussion of the second derivative test for functions of several variables.

PRINCIPAL MINORS THEOREM. *Suppose that A is an $n \times n$ symmetric matrix over the real numbers, and for each integer k between 1 and n let $A^{(k)}$ be the $k \times k$ matrix whose (i,j) entry is $a_{i,j}$ for $1 \leq i,j \leq k$ (visually, this is the $k \times k$ submatrix in the upper left hand part of the original array with all other entries discarded). Then A is positive definite if and only if $\det A^{(k)}$ is positive for all k between 1 and n.*

In older terminology for matrix algebra, the quantity $\det A^{(k)}$ was called the k^{th} **principal minor** of A (minors referred to determinants of matrices formed by deleting suitable numbers of rows and columns), and this is the reason behind the naming of the result. When we return to the derivation of the Principal Minors Theorem we shall also show that it is much easier to check its validity for examples than one might initially think.

Taylor's Theorem in several variables

Recall that Taylor's Theorem — or Taylor's Formula — is a result on approximating functions with $(n+1)$ continuous derivatives by polynomials of degree $\leq n$ and that it includes a formula for the error in the given "optimal" approximations. Usually this is expressed in terms of the $(n+1)^{\text{st}}$ derivative of the function, but for some purposes other descriptions of the error term are also worth knowing. Nearly every calculus text includes a treatment of Taylor's Theorem and at least one version of the error term. For our purposes it will suffice to use the version stated as Theorem 8.19 on page 611 of *Calculus* (Seventh Edition), by Larson, Hostetler and Edwards.

TAYLOR'S THEOREM WITH THE LAGRANGE REMAINDER. *Suppose that f is a real valued function defined on an interval $(c - r,\, c + r)$ where $c \in \mathbf{R}$ and $r > 0$, and suppose that f has continuous derivatives of all orders through $(n + 1)$ on that interval, where $n \geq 1$. Then for each $h \neq c$ such that $|h| < r$ there is a point α between c and $c + h$ such that*

$$f(x+h) \quad = \quad f(c) \ + \ \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!}\, h^k \quad + \quad R_n(h)$$

where

$$R_n(h) \quad = \quad \frac{f^{(n+1)}(\alpha)}{(n+1)!}\, h^n + 1 \ . \blacksquare$$

We shall need a version of this result for functions of m variables. The first step in formulating this generalization is to specify where f is defined, and we replace the interval of radius r centered at $c \in \mathbf{R}$ with the open disk of radius r about a point $\mathbf{c} \in \mathbf{R}^n$:

$$N_r(\mathbf{c}) \quad = \quad \{\ \mathbf{x} \in \mathbf{R}^m \ \ | \ \ |\mathbf{x} - \mathbf{c}| \ \ < \ \ r\ \}$$

Given that we are starting with the single variables form of Taylor's Theorem, it should not be surprising that we want to derive the multivariable form by constructing some single variable function out of the given multivariable function. If we let $\mathbf{h} \neq \mathbf{0}$ such that $|\mathbf{h}| < r$, then the function

$$g(t) \quad = \quad f(\mathbf{c} + t\,\mathbf{h})$$

will be a function defined on an open interval of radius

$$\frac{r}{|\mathbf{h}|} \quad > \quad 1 \ .$$

In order to show that this function has the required differentiability properties, one assumes that f has continuous partial derivatives with respect to all combinations of variables through order $(n + 1)$; the differentiability properties of g then can be obtained by repeated application of the Chain Rule from multivariable calculus (see Section 12.5 on pages 876–883 of the text by Larson, Hostetler and Edwards).

Strictly speaking, the statements of the Chain Rule in calculus textbooks often only give a formula such as

$$\frac{d\,w}{d\,s} \quad = \quad \sum_{j=1}^{n} \frac{\partial\, w}{\partial\, u_j}\, \frac{d\, u_j}{d\, s}$$

(this is essentially the formula at the top of page 877 of Larson, Hostetler and Edwards), but in our case it is easy to check that this expression has continuous derivatives of order k for each $k \leq n$ by induction. Specifically, if $\mathbf{h}$ is expressed in coordinates as $(h_1, \ \cdots \ , h_n)$, then we have the following explicit formula:

$$g^{(k)}(t) \quad = \quad \sum_{1 \leq i_1, \cdots, \, i_k \leq m} \quad \frac{\partial^k f}{\partial x_{i_1} \ \cdots \ \partial x_{i_k}} \big(\mathbf{c} + t\mathbf{h}\big) \cdot h_{i_1} \ \cdots \ h_{i_k}$$

We can now state a multivariable version of Taylor's Theorem that is sufficient for our purposes.

MULTIVARIABLE TAYLOR'S FORMULA FOR DEGREE 1 APPROXIMATIONS. *Let f be a function of m variables with continuous second partial derivatives with respect to all pairs of variables on the open disk $N_r(\mathbf{c})$, and let $\mathbf{h} \neq \mathbf{0}$ be such that $|\mathbf{h}| < r$. Then there is some $\alpha \in [0,1]$ such that*

$$f(\mathbf{c}+\mathbf{h}) \quad = \quad f(\mathbf{c}) \;+\; \sum_{k=1}^{m} \frac{\partial f}{\partial x_k}(\mathbf{c})\, h_k \;+\; \sum_{1\leq i,j\leq m} \frac{\partial^2 f}{\partial x_i\, \partial x_j}\big(\mathbf{c}+\alpha\mathbf{h}\big)\cdot h_i\, h_j \ .$$

We shall describe a more concise way of writing the right hand side. One can use gradients to write the linear part of the function as $f(\mathbf{c}) + \nabla f(\mathbf{c})\cdot\mathbf{h}$, and if one defines the *Hessian* of f at a point $\mathbf{p}$ to be the symmetric $m\times m$ matrix $\mathrm{Hess}(f;\mathbf{p})$ whose (i,j) entry is equal to

$$\frac{\partial^2 f}{\partial x_i\, \partial x_j}(\mathbf{p})$$

then we can rewrite the relevant case of Taylor's Formula as follows:

$$f(\mathbf{c}+\mathbf{h}) \;\;=\;\; f(\mathbf{c}) \;+\; \nabla f(\mathbf{c})\cdot\mathbf{h} \;+\; {}^{\mathbf{T}}\mathbf{h}\left[\mathrm{Hess}(f;\mathbf{c}+\alpha\mathbf{h})\right]\mathbf{h}$$

With this terminology we are finally ready to write down the second derivative test for relative extrema in several variables:

MULTIVARIABLE SECOND DERIVATIVE TEST. *Let f be a function with continuous second partial derivatives on the open disk $N_r(\mathbf{c})$ in $\mathbf{R}^n$, and suppose that $\nabla f(\mathbf{c}) = \mathbf{0}$. Then the following conclusions hold:*

(1) *If the Hessian of f at $\mathbf{c}$ is a* **positive definite** *matrix, then f has a relative* **minimum** *at $\mathbf{c}$.*

(2) *If the* **negative** *of the Hessian of f at $\mathbf{c}$ is a positive definite matrix, then f has a relative* **maximum** *at $\mathbf{c}$.*

(3) *If the Hessian of f at $\mathbf{c}$ is invertible, but neither it nor its negative is positive definite, then f has neither a relative maximum nor a relative minimum at $\mathbf{c}$ (these are essentially the* **saddle points** *one discusses in multivariable calculus).*

(4) *If the Hessian of f at $\mathbf{c}$ is not invertible, then no conclusion can be drawn.*

The standard test the positive definiteness of the Hessian is given by the Principal Minors Theorem, and in fact this test plays an important part in the derivation of the Multivariable Second Derivative Test. Therefore we shall restate the latter using the Principal Minors Theorem:

ALTERNATE FORMULATION. *In the setting above, one can restate the conclusions as follows:*

(1) *If the the principal minors for the Hessian of f at $\mathbf{c}$ are all* **positive***, then f has a relative* **minimum** *at $\mathbf{c}$.*

(2) *If the* **odd** *principal minors for the Hessian of f at $\mathbf{c}$ are all* **negative** *and the* **even** *principal minors for the Hessian of f at $\mathbf{c}$ are all* **positive***, then f has a relative* **maximum** *at $\mathbf{c}$.*

(3) *If the Hessian of f at $\mathbf{c}$ has a nonzero determinant but the principal minors* **do not satisfy either** *of the sequences of positivity and negativity conditions described above, then f has neither a relative maximum nor a relative minimum at $\mathbf{c}$ (this is the* **saddle point** *case).*

(4) *If the determinant of the Hessian of f at $\mathbf{c}$ is zero, then no conclusion can be drawn.*

Since no information is obtained if the determinant of the Hessian vanishes, we shall ignore this case henceforth and assume that the Hessian at $\mathbf{c}$ has a nonzero determinant. The proof of the second derivative test then splits into three cases; namely, the tests for relative minima, relative maxima and saddle points.

Relative minimum test. Since the gradient vanishes at $\mathbf{c}$, the concise formulation of Taylor's Theorem yields the following identity, in which $\alpha \in [0,1]$:

$$f(\mathbf{c}+\mathbf{h}) \;-\; f(\mathbf{c}) \quad = \quad {}^{\mathbf{T}}\mathbf{h}\,\big[\mathrm{Hess}(f;\mathbf{c}+\alpha\mathbf{h})\big]\,\mathbf{h}$$

Suppose that the Hessian of f and $\mathbf{c}$ is positive definite, so that all of its principal minors are positive. We need to show that the positivity of these principal minors for $\mathrm{Hess}(f;\mathbf{c})$ implies a corresponding statement for $\mathrm{Hess}(f;\mathbf{c}+\alpha\mathbf{h})$, at least if $|\mathbf{h}|$ is sufficiently small. Specifically, here is what we need:

Lemma. *If the principal minors of* $\mathrm{Hess}(f;\mathbf{c})$ *are all positive then there is a number $s \in (0,r)$ such that $|\mathbf{v}| < s$ implies that the principal minors of* $\mathrm{Hess}(f;\mathbf{c}+\mathbf{v})$ *are also all positive.*

Proof of Lemma. Since the determinant is a polynomial in the entries of a matrix, it is a continuous function of the entries of a matrix. In particular, if the principal minors of a square matrix A are all positive and the entries of a second matrix B are all sufficiently close to those of A, then the principal minors of B will also be all positive. Take A to be the Hessian at $\mathbf{c}$ and B to be the Hessian at some other point $\mathbf{c}+\mathbf{v}$. Since f has continuous second partial derivatives there is an $s > 0$ such that $|\mathbf{v}| < s$ implies the entries of B are sufficiently close to the entries of A for the principal minors condition to hold.■

Return to the relative minimum test. If we now restrict ourselves to vectors $\mathbf{h}$ of length $\leq s$, then the Lemma tells us that the principal minors for $\mathrm{Hess}(f;\mathbf{c}+\alpha\mathbf{h})$ will all be positive, and hence that

$${}^{\mathbf{T}}\mathbf{h}\,\big[\mathrm{Hess}(f;\mathbf{c}+\alpha\mathbf{h})\big]\,\mathbf{h} \quad > \quad 0\ .$$

It follows that f has a **strict** relative minimum at $\mathbf{c}$.■

Relative maximum test. This will be derived directly from the preceding test, using the fact that the relative maxima of f are the same as the relative minima of its negative function $-f$. Since the Hessians of f and $-f$ are negatives of each other, the relative minimum test immediately implies that one has a strict relative maximum if the principal minors of the matrix $-\mathrm{Hess}(f;\mathbf{c})$ are all positive. Since the determinants of $k \times k$ matrices satisfy the relation $\det(-B) = (-1)^k \det B$, it follows that the k^{th} principal minor of the matrix

$$\mathrm{Hess}(-f;\mathbf{c}) \quad = \quad -\,\mathrm{Hess}(f;\mathbf{c})$$

is equal to the corresponding principal minor for $\mathrm{Hess}(f;\mathbf{c})$ if k is even and the negative of the corresponding principal minor if k is odd. Therefore the conditions for f to have a strict relative maximum, or equivalently for $-f$ to have a strict relative minimum, translate into the negativity of the odd principal minors of the Hessian and the positivity of the even principal minors of the Hessian.■

Saddle point test. Before reading this it might be useful to say a few words about saddle points. The typical example for functions of two variables is given by $f(x,y) = y^2 - x^2$, and it is illustrated on page 908 of Larson, Hostetler and Edwards. If we restrict the function to points of the form $(0, y)$ then this restricted function has an absolute minimum at $(0, 0)$, while if we restrict to points of the form $(x, 0)$ then the restricted function has an abolute minimum there. This topic is also discussed on pages 193–194 of *Basic Multivariable Calculus*, by Marsden, Tromba and Weinstein.

Once again we need some additional input.

Lemma. *If A is a symmetric matrix over the real numbers, then A is positive definite if and only if all of its eigenvalues are positive.*

Proof of Lemma. By Rayleigh's Principle the minimum value of the Rayleigh quotient is the minimum eigenvalue, and by the definition of this quotient it is positive for all nonzero vectors in $\mathbf{R}^n$ if and only if q_A is positive for all such vectors. But the latter is precisely the condition for a real symmetric matrix A to be positive definite. Since all the eigenvalues of A are positive if and only if the smallest eigenvalue is positive, this proves the Lemma.∎

Return to the saddle point test. In this case we are assuming that the Hessian at $\mathbf{c}$ is invertible, so that its eigenvalues are all nonzero, but by the Lemma we are also assuming conditions which mean that its eigenvalues are neither all positive nor all negative. It follows that the maximum eigenvalue must be positive and the minimum eigenvalue must be negative. Let $\lambda_\pm$ be the maximum and minimum eigenvalues, and let $\mathbf{u}_\pm$ be unit eigenvectors for these eigenvalues. If $\mathbf{h}_\pm$ is a nonzero multiple of $\mathbf{u}_\pm$ it follows that

$$ {}^{\mathbf{T}}\mathbf{h}_\pm \left[\mathrm{Hess}(f; \mathbf{c})\right] \mathbf{h}_\pm \quad = \quad |\mathbf{h}_\pm|^2 \, \lambda_\pm $$

and therefore the expression is positive for $\mathbf{h}_+$ and negative for $\mathbf{h}_-$. By the continuity of the entries in the Hessian matrix, we can find some $s \in (0, r)$ such that if $|\mathbf{v}| < s$ then

$$ {}^{\mathbf{T}}\mathbf{u}_+ \left[\mathrm{Hess}(f; \mathbf{c} + \mathbf{v})\right] \mathbf{u}_+ \quad > \quad 0 $$

and also

$$ {}^{\mathbf{T}}\mathbf{u}_- \left[\mathrm{Hess}(f; \mathbf{c} + \mathbf{v})\right] \mathbf{u}_- \quad < \quad 0 \; . $$

Suppose now that we choose $\mathbf{h}_\pm$ to have length less than s. As in the previous tests we know that

$$ f(\mathbf{c} + \mathbf{h}_\pm) \; - \; f(\mathbf{c}) \quad = \quad {}^{\mathbf{T}}\mathbf{h}_\pm \left[\mathrm{Hess}(f; \mathbf{c} + \alpha\mathbf{h})\right] \mathbf{h}_\pm $$

and we may rewrite the right hand side in the following form:

$$ |\mathbf{h}_\pm|^2 \cdot {}^{\mathbf{T}}\mathbf{u}_\pm \left[\mathrm{Hess}(f; \mathbf{c} + \alpha\mathbf{h}_\pm)\right] \mathbf{u}_\pm $$

Now the first factor is always positive, and the conditions on the vectors $\mathbf{h}_\pm$ ensure that the remaining factor is positive for $\mathbf{h}_+$ and negative for $\mathbf{h}_-$. Therefore $f(\mathbf{c} + \mathbf{h}_+) \; - \; f(\mathbf{c})$ is positive and $f(\mathbf{c} + \mathbf{h}_-) \; - \; f(\mathbf{c})$ is negative, since we can choose the lengths of $\mathbf{h}_\pm$ to be arbitrarily small positive numbers, it follows that f neither has a relative maximum nor a relative minimum at $\mathbf{c}$.∎

Proof of the Principal Minors Theorem

The following general result on the structure of positive definite matrices will be important for our purposes.

THEOREM. *If A is a symmetric $n \times n$ matrix, then the following conditions are equivalent:*

(1) *The matrix A is positive definite.*

(2) *The function $\varphi_A(\mathbf{x},\,\mathbf{y}) = {}^{\mathbf{T}}\mathbf{y}\,A\,\mathbf{x}$ defines an inner product on $\mathbf{R}^n$.*

(3) *The matrix A can be written as a product ${}^{\mathbf{T}}P\,P$ for some invertible matrix P.*

Proof. [(1) $\Longrightarrow$ (2)] The basic conditions for an inner product, for example as stated in Theorem 1.3 on pages 24–25 of the text, are consequences of the rules for matrix multiplication and the assumption that ${}^{\mathbf{T}}\mathbf{x}\,A\,\mathbf{x}$ is positive if $\mathbf{x} \neq \mathbf{0}$.

[(2) $\Longrightarrow$ (3)] The Gram-Schmidt Process shows that there is a basis for $\mathbf{R}^n$ that is orthonormal with respect to φ_A. If we let Q be the matrix whose columns are this orthonormal basis, then direct calculation shows that the φ_A innner product of the i^{th} and j^{th} columns of Q is equal to the (i,j) entry of ${}^{\mathbf{T}}Q\,A\,Q$, and therefore we know that the latter matrix is the identity. If we let $P = Q^{-1}$, then it follows that $A = {}^{\mathbf{T}}P\,P$.

[(3) $\Longrightarrow$ (1)] If $\mathbf{x}$ is nonzero then so is $P(\mathbf{x})$, and therefore we have

$$ {}^{\mathbf{T}}\mathbf{x}\,A\,\mathbf{x} \quad = \quad {}^{\mathbf{T}}\mathbf{x}\,{}^{\mathbf{T}}P\,P\,\mathbf{x} \quad = \quad |P(\mathbf{x})|^2 \quad > \quad 0\,. \blacksquare $$

Application to the Principal Minors Theorem. We can use the preceding result to prove that the principal minors of a positive definite matrix are all positive as follows. If $A^{(k)}$ is the $k \times k$ submatrix in the upper left hand corner of A, then by the theorem we know there is an invertible matrix P_k such that $A^{(k)} = {}^{\mathbf{T}}P_k\,P_k$. Since the determinant of a matrix and its transpose are equal, it follows that

$$ \det A^{(k)} \quad = \quad \det\left({}^{\mathbf{T}}P_k\,P_k\right) \quad = \quad \det\left({}^{\mathbf{T}}P_k\right)\,\det P_k \quad = \quad \left(\det P_k\right)^2 \quad > \quad 0\,. $$

The proof of the reverse implication proceeds by induction on the size of the matrix, and it will be convenient to isolate the crucial part of the inductive step.

Lemma. *Let A be a symmetric $n \times n$ matrix such that the $(n-1)\times(n-1)$ matrix in the upper left hand corner is an identity matrix and the determinant of A is positive. Then A is positive definite.*

Proof. Let $d = a_{n,n}$, and let $w_i = a_{i,n} = a_{n,i}$ for $i < n$. We can reduce A to triangular form by subtracting w_i times the i^{th} row from the last row for each $i < n$, and if we do so we obtain a matrix whose diagonal entries are equal to 1 except in the final position, where one has $d = \sum_i \, w_i^2$. It follows that the latter must be positive because it is equal to the determinant of A.

For every $n \times 1$ column vector $\mathbf{x}$ with entries x_i we have

$$ {}^{\mathbf{T}}\mathbf{x}\,A\,\mathbf{x} \quad = \quad \sum_{i=1}^{n-1} \, x_i^2 \;+\; \sum_{i=1}^{n-1} \, 2\,w_i\,x_i\,x_n \;+\; d\,x_n^2 $$

and by completing squares and using the previous formula for $\det A$ we may rewrite the right hand side as

$$ \sum_{i=1}^{n-1} \, (x_i + 2\,w_i\,x_n)^2 \;+\; (\det A)\,x_n^2\,. $$

If this expression is zero, then the positivity of the determinant implies that $x_n = 0$, and since we also have $0 = x_i + w_i\,x_n$ in this case it follows that $0 = x_i + w_i\,x_n = x_i$ for all $i < n$.$\blacksquare$

Completion of the proof of the Principal Minors Theorem. If A is a 1×1 matrix there is not much to prove because the quadratic form reduces to an second degree polynomial $a\,x^2$, and the latter is positive for all nonzero x if and only if $a > 0$. Suppose now that the result is known for $(n-1) \times (n-1)$ matrices, and let A_0 denote the $(n-1) \times (n-1)$ submatrix in the upper left hand corner of A. By induction we know that A_0 is positive definite, and therefore we may write $A_0 = {}^{\mathbf{T}}P_0\,P_0$ for some invertible matrix P_0. Let Q_0 be the inverse to P_0, and let Q be the block sum of Q_0 with a 1×1 identity matrix. Direct calculation then shows that ${}^{\mathbf{T}}Q\,A\,Q$ is a matrix which satisfies the conditions of the previous lemma (for example, its determinant is positive because it is equal to $\det A \cdot (\det Q)^2$). It follows that ${}^{\mathbf{T}}Q\,A\,Q$ can be factored as a product ${}^{\mathbf{T}}S\,S$ for some invertible matrix S, and therefore if we set B equal to $S\,Q^{-1}$ it will follow that B is invertible and $A = {}^{\mathbf{T}}B\,B$.■

Computational Procedure. Finally, here is a simple algorithmic process for determining whether the principal minors of an arbitrary square matrix are all positive:

For each i such that $1 \leq i \leq n$ attempt to carry out the following steps on a previously computed matrix A_{i-1}; we take $A_0 = A$, and part of the recursive assumption is that the (k, j) entries of A_{i-1} are zero if $k \leq i-1$ and $j > k$. First check whether the (i, i) entry P_i of the matrix is positive. If not, stop the process and conclude that the matrix is does not have a positive i^{th} principal minor and therefore is not positive definite. If the entry is positive, then conclude that the i^{th} principal minor is positive. There are now two cases depending upon whether $i < n$ or $i = n$. In the second case, the procedure is finished, but in the first case one next performs row operations to subtract multiples of the i^{th} row from each subsequent row so that the resulting matrix A_i has all zero entries in the i^{th} column below the i^{th} row; by the recursive assumption this matrix will also have zero entries in the places where A_{i-1} was assumed to have zero entries. Furthermore, the first i diagonal entries of A_i will be the same as the first $(i-1)$ diagonal entries of A_{i-1}, and since the process has continued all these diagonal entries must be positive. The principal minors of the original matrix will be positive if and only if they are so determined by this process.

JUSTIFICATION. If $i = 1$ this process determines whether the first principal minor is positive and terminates if this is not the case. Suppose that the procedure is known to determine whether the first $(i-1)$ principal minors are positive and continues until reaching the i^{th} step. **Note that these operations do not change the principal minors of the matrix**. The determinant of A_{i-1} will be equal to the corresponding principal minor of A, so that one is positive if and only if the other is. Since the $(i-1) \times (i-1)$ matrix in the upper left corner of A_{i-1} is upper triangular it follows that its diagonal entries are the positive numbers P_k for $k < i$, and therefore the corresponding principal minor of A is positive. In fact, we also know that the $i \times i$ submatrix in the upper left hand corner of A_{i-1} is also upper triangular, and its determinant is the product of the previously computed principal minor with P_i. It follows that the i^{th} principal minor of A is positive if and only if $P_i > 0$. Assuming that it is and that $i < n$, the process for finding the next matrix A_i does not change the $i \times i$ submatrix in the upper left hand corner (hence the principal minor), nor does it change any columns before the i^{th} one. However, in the new matrix all entries in the i^{th} column below the i^{th} row are zero.■