chapter **II**

Linear transformations and matrices

In this chapter we define linear transformations and various operations: addition of two linear transformations, multiplication of two linear transformations, and multiplication of a linear transformation by a scalar. Linear transformations are functions of vectors in one vector space U with values which are vectors in the same or another vector space V which preserve linear combinations. They can be represented by matrices in the same sense that vectors can be represented by n-tuples. This representation requires that operations of addition, multiplication, and scalar multiplication of matrices be defined to correspond to these operations with linear transformations. Thus we establish an algebra of matrices by means of the conceptually simpler algebra of linear transformations.

The matrix representing a linear transformation of U into V depends on the choice of a basis in U and a basis in V. Our first problem, a recurrent problem whenever matrices are used to represent anything, is to see how a change in the choice of bases determines a corresponding change in the matrix representing the linear transformation. Two matrices which represent the same linear transformation with respect to different sets of bases must have some properties in common. This leads to the idea of equivalence relations among matrices. The exact nature of this equivalence relation depends on the bases which are permitted.

In this chapter no restriction is placed on the bases which are permitted and we obtain the widest kind of equivalence. In Chapter III we identify U and V and require that the same basis be used in both. This yields a more restricted kind of equivalence, and a study of this equivalence is both interesting and fruitful. In Chapter V we make further restrictions in the permissible bases and obtain an even more restricted equivalence.

When no restriction is placed on the bases which are permitted, the

equivalence is so broad that it is relatively uninteresting. Very useful results are obtained, however, when we are permitted to change basis only in the image space V. In every set of mutually equivalent matrices we select one, representative of all of them, which we call a normal form, in this case the Hermite normal form. The Hermite normal form is one of our most important and effective computational tools, far exceeding in utility its application to the study of this particular equivalence relation.

The pattern we have described is worth conscious notice since it is recurrent and the principal underlying theme in this exposition of matrix theory. We define a concept, find a representation suitable for effective computation, change bases to see how this change affects the representation, and then seek a normal form in each class of equivalent representations.

1 | Linear Transformations

Let U and V be vector spaces over the same field of scalars F.

Definition. A *linear transformation* σ of U into V is a single-valued mapping of U into V which associates to each element $\alpha \in U$ a unique element $\sigma(\alpha) \in V$ such that for all $\alpha, \beta \in U$ and all $a, b \in F$ we have

$$\sigma(a\alpha + b\beta) = a\sigma(\alpha) + b\sigma(\beta). \tag{1.1}$$

We call $\sigma(\alpha)$ the *image* of α under the linear transformation σ. If $\bar{\alpha} \in V$, then any vector $\alpha \in U$ such that $\sigma(\alpha) = \bar{\alpha}$ is called an *inverse image* of $\bar{\alpha}$. The set of *all* $\alpha \in U$ such that $\sigma(\alpha) = \bar{\alpha}$ is called the *complete inverse image* of $\bar{\alpha}$, and it is denoted by $\sigma^{-1}(\bar{\alpha})$. Generally, $\sigma^{-1}(\bar{\alpha})$ need not be a single element as there may be more than one $\alpha \in U$ such that $\sigma(\alpha) = \bar{\alpha}$.

By taking particular choices for a and b we see that for a linear transformation $\sigma(\alpha + \beta) = \sigma(\alpha) + \sigma(\beta)$ and $\sigma(a\alpha) = a\sigma(\alpha)$. Loosely speaking, the image of the sum is the sum of the images and the image of the product is the product of the images. This descriptive language has to be interpreted generously since the operations before and after applying the linear transformation may take place in different vector spaces. Furthermore, the remark about scalar multiplication is inexact since we do not apply the linear transformation to scalars; the linear transformation is defined only for vectors in U. Even so, the linear transformation does preserve the structural operations in a vector space and this is the reason for its importance. Generally, in algebra a structure-preserving mapping is called a *homomorphism*. To describe the special role of the elements of F in the condition, $\sigma(a\alpha) = a\sigma(\alpha)$, we say that a linear transformation is a homomorphism *over* F, or an *F-homomorphism*.

If for $\alpha \neq \beta$ it necessarily follows that $\sigma(\alpha) \neq \sigma(\beta)$, the homomorphism σ is said to be *one-to-one* and it is called a *monomorphism*. If A is any subset of

U, $\sigma(A)$ will denote the set of all images of elements of A; $\sigma(A) = \{\bar{\alpha} \mid \bar{\alpha} = \sigma(\alpha)$ for some $\alpha \in A\}$. $\sigma(A)$ is called the *image* of A. $\sigma(U)$ is often denoted by $\mathrm{Im}(\sigma)$ and is called the *image* of σ. If $\mathrm{Im}(\sigma) = V$ we shall say that the homomorphism is a mapping *onto* V and it is called an *epimorphism.*

We call the set U, on which the linear transformation σ is defined, the *domain* of σ. We call V, the set in which the images of σ are defined, the *codomain* of σ. Strictly speaking, a linear transformation must specify the domain and codomain as well as the mapping. For example, consider the linear transformation that maps every vector of U onto the zero vector of V. This mapping is called the *zero mapping*. If W is any subspace of V, there is also a zero mapping of U into W, and this mapping has the same effect on the elements of U as the zero mapping of U into V. However, they are different linear transformations since they have different codomains. This may seem like an unnecessarily fine distinction. Actually, for most of this book we could get along without this degree of precision. But the more deeply we go into linear algebra the more such precision is needed. In this book we need this much care when we discuss dual spaces and dual transformations in Chapter IV.

A homomorphism that is both an epimorphism and a monomorphism is called an *isomorphism*. If $\bar{\alpha} \in V$, the fact that σ is an epimorphism says that there is an $\alpha \in U$ such that $\sigma(\alpha) = \bar{\alpha}$. The fact that σ is a monomorphism says that this α is unique. Thus, for an isomorphism, we can define an inverse mapping σ^{-1} that maps $\bar{\alpha}$ onto α.

Theorem 1.1. *The inverse σ^{-1} of an isomorphism is also an isomorphism.*

PROOF. Since σ^{-1} is obviously one-to-one and onto, it is necessary only to show that it is linear. If $\bar{\alpha} = \sigma(\alpha)$ and $\bar{\beta} = \sigma(\beta)$, then $\sigma(a\alpha + b\beta) = a\bar{\alpha} + b\bar{\beta}$ so that $\sigma^{-1}(a\bar{\alpha} + b\bar{\beta}) = a\alpha + b\beta = a\sigma^{-1}(\bar{\alpha}) + b\sigma^{-1}(\bar{\beta})$. □

For the inverse isomorphism $\sigma^{-1}(\bar{\alpha})$ is an element of U. This conflicts with the previously given definition of $\sigma^{-1}(\bar{\alpha})$ as a complete inverse image in which $\sigma^{-1}(\bar{\alpha})$ is a subset of U. However, the symbol σ^{-1}, standing alone, will always be used to denote an isomorphism, and in this case there is no difficulty caused by the fact that $\sigma^{-1}(\bar{\alpha})$ might denote either an element or a one-element set.

Let us give some examples of linear transformations. Let $U = V = P$, the space of polynomials in x with coefficients in R. For $\alpha = \sum_{i=0}^{n} a_i x^i$, define $\sigma(\alpha) = \dfrac{d\alpha}{dx} = \sum_{i=0}^{n} i a_i x^{i-1}$. In calculus one of the very first things proved about the derivative is that it is linear, $\dfrac{d(\alpha + \beta)}{dx} = \dfrac{d\alpha}{dx} + \dfrac{d\beta}{dx}$ and $\dfrac{d(a\alpha)}{dx} = a\dfrac{d\alpha}{dx}$. The mapping $\tau(\alpha) = \sum_{i=0}^{n} \dfrac{a_i}{i+1} x^{i+1}$ is also linear. Notice that this is not the indefinite integral since we have specified that the constant

of integration shall be zero. Notice that σ is onto but not one-to-one and τ is one-to-one but not onto.

Let $U = R^n$ and $V = R^m$ with $m \leq n$. For each $\alpha = (a_1, \ldots, a_n) \in R^n$ define $\sigma(\alpha) = (a_1, \ldots, a_m) \in R^m$. It is clear that this linear transformation is one-to-one if and only if $m = n$, but it is onto. For each $\beta = (b_1, \ldots, b_m) \in R^m$ define $\tau(\beta) = (b_1, \ldots, b_m, 0, \ldots, 0) \in R^n$. This linear transformation is one-to-one, but it is onto if and only if $m = n$.

Let $U = V$. For a given scalar $a \in F$ the mapping of α onto $a\alpha$ is linear since

$$a(\alpha + \beta) = a\alpha + a\beta = a(\alpha) + a(\beta),$$

and

$$a(b\alpha) = (ab)\alpha = (ba)\alpha = b \cdot a(\alpha).$$

To simplify notation we also denote this linear-transformation by a. Linear transformations of this type are called *scalar transformations*, and there is a one-to-one correspondence between the field of scalars and the set of scalar transformations. In particular, the linear transformation that leaves every vector fixed is denoted by 1. It is called the *identity transformation* or *unit transformation*. If linear transformations in several vector spaces are being discussed at the same time, it may be desirable to identify the space on which the identity transformation is defined. Thus 1_U will denote the identity transformation on U.

When a basis of a finite dimensional vector space V is used to establish a correspondence between vectors in V and n-tuples in F^n, this correspondence is an isomorphism. The required arguments have already been given in Section I-3. Since V and F^n are isomorphic, it is theoretically possible to discuss the properties of V by examining the properties of F^n. However, there is much interest and importance attached to concepts that are independent of the choice of a basis. If a homomorphism or isomorphism can be defined uniquely by intrinsic properties independent of a choice of basis the mapping is said to be *natural* or *canonical*. In particular, any two vector spaces of dimension n over F are isomorphic. Such an isomorphism can be established by setting up an isomorphism between each one and F^n. This isomorphism will be dependent on a choice of a basis in each space. Such an isomorphism, dependent upon the arbitrary choice of bases, is not canonical.

Next, let us define the various operations between linear transformations. For each pair σ, τ of linear transformation of U into V, define $\sigma + \tau$ by the rule

$$(\sigma + \tau)(\alpha) = \sigma(\alpha) + \tau(\alpha) \qquad \text{for all} \qquad \alpha \in V.U$$

$\sigma + \tau$ is a linear transformation since

$$\begin{aligned}(\sigma + \tau)(a\alpha + b\beta) &= \sigma(a\alpha + b\beta) + \tau(a\alpha + b\beta) = a\sigma(\alpha) + b\sigma(\beta) \\ &\quad + a\tau(\alpha) + b\tau(\beta) = a[\sigma(\alpha) + \tau(\alpha)] + b[\sigma(\beta) + \tau(\beta)] \\ &= a(\sigma + \tau)(\alpha) + b(\sigma + \tau)(\beta).\end{aligned}$$

Observe that addition of linear transformation is commutative; $\sigma + \tau = \tau + \sigma$.

For each linear transformation σ and $a \in F$ define $a\sigma$ by the rule; $(a\sigma)(\alpha) = a[\sigma(\alpha)]$. $a\sigma$ is a linear transformation.

It is not difficult to show that with these two operations the set of all linear transformations of U into V is itself a vector space over F. This is a very important fact and we occasionally refer to it and make use of it. However, we wish to emphasize that we define the sum of two linear transformations if and only if they both have the same domain and the same codomain. It is neither necessary nor sufficient that they have the same image, or that the image of one be a subset of the image of the other. It is simply a question of being clear about the terminology and its meaning. The set of all linear transformations of U into V will be denoted by $\mathrm{Hom}(U, V)$.

There is another, entirely new, operation that we need to define. Let W be a third vector space over F. Let σ be a linear transformation of U into V and τ a linear transformation of V into W. By $\tau\sigma$ we denote the linear transformation of U into W defined by the rule: $(\tau\sigma)(\alpha) = \tau[\sigma(\alpha)]$. Notice that in this context $\sigma\tau$ has no meaning. We refer to this operation as either *iteration* or *multiplication* of linear transformation, and $\tau\sigma$ is called the *product* of τ and σ.

The operations between linear transformations are related by the following rules:

1. Multiplication is associative: $\pi(\tau\sigma) = (\pi\tau)\sigma$. Here π is a linear transformation of W into a fourth vector space X.
2. Multiplication is distributive with respect to addition:

$$(\tau_1 + \tau_2)\sigma = \tau_1\sigma + \tau_2\sigma \qquad \text{and} \qquad \tau(\sigma_1 + \sigma_2) = \tau\sigma_1 + \tau\sigma_2.$$

3. Scalar multiplication commutes with multiplication: $a(\tau\sigma) = \tau(a\sigma)$.

These properties are easily proved and are left to the reader.

Notice that if $W \neq U$, then $\tau\sigma$ is defined but $\sigma\tau$ is not. If all linear transformations under consideration are mappings of a vector space U into itself, then these linear transformations can be multiplied in any order. This means that $\tau\sigma$ and $\sigma\tau$ would both be defined, but it would not mean that $\tau\sigma = \sigma\tau$.

The set of linear transformation of a vector space into itself is a vector space, as we have already observed, and now we have defined a product which satisfies the three conditions given above. Such a space is called an *associative algebra*. In our case the algebra consists of linear transformation and it is known as a *linear algebra*. However, the use of terms is always in a state of flux, and today this term is used in a more inclusive sense. When referring to a particular set with an algebraic structure, "linear algebra" still denotes what we have just described. But when referring to an area of

study, the term "linear algebra includes virtually every concept in which linear transformations play a role, including linear transformations between different vector spaces (in which the linear transformations cannot always be multiplied), sequences of vector spaces, and even mappings of sets of linear transformations (since they also have the structure of a vector space).

Theorem 1.2. $\mathrm{Im}(\sigma)$ *is a subspace of* V.

PROOF. If $\bar{\alpha}$ and $\bar{\beta}$ are elements of $\mathrm{Im}(\sigma)$, there exist $\alpha, \beta \in U$ such that $\sigma(\alpha) = \bar{\alpha}$ and $\sigma(\beta) = \bar{\beta}$. For any $a, b \in F$, $\sigma(a\alpha + b\beta) = a\sigma(\alpha) + b\sigma(\beta) = a\bar{\alpha} + b\bar{\beta} \in \mathrm{Im}(\sigma)$. Thus $\mathrm{Im}(\sigma)$ is a subspace of V. □

Corollary 1.3. *If* U_1 *is a subspace of* U, *then* $\sigma(U_1)$ *is a subspace of* V. □

It follows from this corollary that $\sigma(0) = 0$ where 0 denotes the zero vector of U and the zero vector of V. It is even easier, however, to show it directly. Since $\sigma(0) = \sigma(0 + 0) = \sigma(0) + \sigma(0)$ it follows from the uniqueness of the zero vector that $\sigma(0) = 0$.

For the rest of this book, unless specific comment is made, we assume that all vector spaces under consideration are finite dimensional. Let $\dim U = n$ and $\dim V = m$.

The dimension of the subspace $\mathrm{Im}(\sigma)$ is called the *rank* of the linear transformation σ. The rank of σ is denoted by $\rho(\sigma)$.

Theorem 1.4. $\rho(\sigma) \leq \{m, n\}$.

PROOF. If $\{\alpha_1, \ldots, \alpha_s\}$ is linearly dependent in U, there exists a non-trivial relation of the form $\sum_i a_i\alpha_i = 0$. But then $\sum_i a_i\sigma(\alpha_i) = \sigma(0) = 0$; that is, $\{\sigma(\alpha_1), \ldots, \sigma(\alpha_s)\}$ is linearly dependent in V. A linear transformation preserves linear relations and transforms dependent sets into dependent sets. Thus, there can be no more than n linearly independent elements in $\mathrm{Im}(\sigma)$. In addition, $\mathrm{Im}(\sigma)$ is a subspace of V so that $\dim \mathrm{Im}(\sigma) \leq m$. Thus $\rho(\sigma) = \dim \mathrm{Im}(\sigma) \leq \min\{m, n\}$. □

Theorem 1.5. *If* W *is a subspace of* V, *the set* $\sigma^{-1}(W)$ *of all* $\alpha \in U$ *such that* $\sigma(\alpha) \in W$ *is a subspace of* U.

PROOF. If $\alpha, \beta \in \sigma^{-1}(W)$, then $\sigma(a\alpha + b\beta) = a\sigma(\alpha) + b\sigma(\beta) \in W$. Thus $a\alpha + b\beta \in \sigma^{-1}(W)$ and $\sigma^{-1}(W)$ is a subspace. □

The subspace $K(\sigma) = \sigma^{-1}(0)$ is called the *kernel* of the linear transformation σ. The dimension of $K(\sigma)$ is called the *nullity* of σ. The nullity of σ is denoted by $\nu(\sigma)$.

Theorem 1.6. $\rho(\sigma) + \nu(\sigma) = n$.

PROOF. Let $\{\alpha_1, \ldots, \alpha_\nu, \beta_1, \ldots, \beta_k\}$ be a basis of U such that $\{\alpha_1, \ldots, \alpha_\nu\}$ is a basis of $K(\sigma)$. For $\alpha = \sum_i a_i\alpha_i + \sum_j b_j\beta_j \in U$ we see that $\sigma(\alpha) = \sum_i a_i\sigma(\alpha_i) + \sum_j b_j\sigma(\beta_j) = \sum_j b_j\sigma(\beta_j)$. Thus $\{\sigma(\beta_1), \ldots, \sigma(\beta_k)\}$ spans $\mathrm{Im}(\sigma)$. On the other hand if $\sum_j c_j\sigma(\beta_j) = 0$, then $\sigma(\sum_j c_j\beta_j) = \sum_j c_j\sigma(\beta_j) = 0$; that

is, $\sum_j c_j\beta_j \in K(\sigma)$. In this case there exist coefficients d_i such that $\sum_j c_j\beta_j = \sum_i d_i\alpha_i$. If any of these coefficients were non-zero we would have a non-trivial relation among the elements of $\{\alpha_1, \ldots, \alpha_\nu, \beta_1, \ldots, \beta_k\}$. Hence, all $c_j = 0$ and $\{\sigma(\beta_1), \ldots, \sigma(\beta_k)\}$ is linearly independent. But then it is a basis of $\mathrm{Im}(\sigma)$ so that $k = \rho(\sigma)$. Thus $\rho(\sigma) + \nu(\sigma) = n$. □

Theorem 1.6 has an important geometric interpretation. Suppose that a 3-dimensional vector space R^3 were mapped onto a 2-dimensional vector space R^2. In this case, it is simplest and sufficiently accurate to think of σ as the linear transformation which maps $(a_1, a_2, a_3) \in R^3$ onto $(a_1, a_2) \in R^2$ which we can identify with $(a_1, a_2, 0) \in R^3$. Since $\rho(\sigma) = 2$, $\nu(\sigma) = 1$. Clearly, every point $(0, 0, a_3)$ on the x_3-axis is mapped onto the origin. Thus $K(\sigma)$ is the x_3-axis, the line through the origin in the direction of the projection, and $\{(0, 0, 1) = \alpha_1\}$ is a basis of $K(\sigma)$. It should be evident that any plane through the origin not containing $K(\sigma)$ will be projected onto the x_1x_2-plane and that this mapping is one-to-one and onto. Thus the complementary subspace $\langle\beta_1, \beta_2\rangle$ can be taken to be any plane through the origin not containing the x_3-axis. This illustrates the wide latitude of choice possible for the complementary subspace $\langle\beta_1, \ldots, \beta_\rho\rangle$.

Theorem 1.7. *A linear transformation σ of U into V is a monomorphism if and only if $\nu(\sigma) = 0$, and it is an epimorphism if and only if $\rho(\sigma) = \dim V$.*

PROOF. $K(\sigma) = \{0\}$ if and only if $\nu(\sigma) = 0$. If σ is a monomorphism, then certainly $K(\sigma) = \{0\}$ and $\nu(\sigma) = 0$. On the other hand, if $\nu(\sigma) = 0$ and $\sigma(\alpha) = \sigma(\beta)$, then $\sigma(\alpha - \beta) = 0$ so that $\alpha - \beta \in K(\sigma) = \{0\}$. Thus, if $\nu(\sigma) = 0$, σ is a monomorphism.

It is but a matter of reading the definitions to see that σ is an epimorphism if and only if $\rho(\sigma) = \dim V$. □

If $\dim U = n < \dim V = m$, then $\rho(\sigma) = n - \nu(\sigma) \leq n < m$ so that σ cannot be an epimorphism. If $n > m$, then $\nu(\sigma) = n - \rho(\sigma) \geq n - m > 0$, so that σ cannot be a monomorphism. Any linear transformation from a vector space into a vector space of higher dimension must fail to be an epimorphism. Any linear transformation from a vector space into a vector space of lower dimension must fail to be a monomorphism.

Theorem 1.8. *Let U and V have the same finite dimension n. A linear transformation σ of U into V is an isomorphism if and only if it is an epimorphism. σ is an isomorphism if and only if it is a monomorphism.*

PROOF. It is part of the definition of an isomorphism that it is both an epimorphism and a monomorphism. Suppose σ is an epimorphism. $\rho(\sigma) = n$ and $\nu(\sigma) = 0$ by Theorem 1.6. Hence, σ is a monomorphism. Conversely if σ is a monomorphism, then $\nu(\sigma) = 0$ and, by Theorem 1.6, $\rho(\sigma) = n$. Hence, σ is an epimorphism. □

Thus a linear transformation σ of U into V is an isomorphism if two of the following three conditions are satisfied: (1) $\dim U = \dim V$, (2) σ is an epimorphism, (3) σ is a monomorphism.

Theorem 1.9. $\rho(\tau) = \rho(\tau\sigma) + \dim \{\mathrm{Im}(\sigma) \cap K(\tau)\}$.

PROOF. Let τ' be a new linear transformation defined on $\mathrm{Im}(\sigma)$ mapping $\mathrm{Im}(\sigma)$ into W so that for all $\alpha \in \mathrm{Im}(\sigma)$, $\tau'(\alpha) = \tau(\alpha)$. Then $K(\tau') = \mathrm{Im}(\sigma) \cap K(\tau)$ and $\rho(\tau') = \dim \tau[\mathrm{Im}(\sigma)] = \dim \tau\sigma(U) = \rho(\tau\sigma)$. Then Theorem 1.6 takes the form

$$\rho(\tau') + \nu(\tau') = \dim \mathrm{Im}(\sigma),$$

or

$$\rho(\tau\sigma) + \dim \{\mathrm{Im}(\sigma) \cap K(\tau)\} = \rho(\sigma). \square$$

Corollary 1.10. $\rho(\tau\sigma) = \dim \{\mathrm{Im}(\sigma) + K(\tau)\} - \nu(\tau)$.

PROOF. This follows from Theorem 1.9 by application of Theorem 4.8 of Chapter I. $\square$

Corollary 1.11. *If* $K(\tau) \subset \mathrm{Im}(\sigma)$, *then* $\rho(\sigma) = \rho(\tau\sigma) + \nu(\tau)$. $\square$

Theorem 1.12. *The rank of a product of linear transformations is less than or equal to the rank of either factor:* $\rho(\tau\sigma) \leq \min \{\rho(\tau), \rho(\sigma)\}$.

PROOF. The rank of $\tau\sigma$ is the dimension of $\tau[\sigma(U)] \subset \tau(V)$. Thus considering $\dim \sigma(U)$ as the "n" and $\dim \tau(V)$ as the "m" of Theorem 1.3 we see that $\dim \tau\sigma(U) = \rho(\tau\sigma) \leq \min \{\dim \sigma(V), \dim \tau(V)\} = \min \{\rho(\sigma), \rho(\tau)\}$. $\square$

Theorem 1.13. *If* σ *is an epimorphism, then* $\rho(\tau\sigma) = \rho(\tau)$. *If* τ *is a monomorphism, then* $\rho(\tau\sigma) = \rho(\sigma)$.

PROOF. If σ is an epimorphism, then $K(\tau) \subset \mathrm{Im}(\sigma) = V$ and Corollary 1.11 applies. Thus $\rho(\tau\sigma) = \rho(\sigma) - \nu(\tau) = m - \nu(\tau) = \rho(\tau)$. If τ is a monomorphism, then $K(\tau) = \{0\} \subset \mathrm{Im}(\sigma)$ and Corollary 1.11 applies. Thus $\rho(\tau\sigma) = \rho(\sigma) - \nu(\tau) = \rho(\sigma)$. $\square$

Corollary 1.14. *The rank of a linear transformation is not changed by multiplication by an isomorphism (on either side).* $\square$

Theorem 1.15. σ *is an epimorphism if and only if* $\tau\sigma = 0$ *implies* $\tau = 0$. τ *is a monomorphism if and only if* $\tau\sigma = 0$ *implies* $\sigma = 0$.

PROOF. Suppose σ is an epimorphism. Assume $\tau\sigma$ is defined and $\tau\sigma = 0$. If $\tau \neq 0$, there is a $\beta \in V$ such that $\tau(\beta) \neq 0$. Since σ is an epimorphism, there is an $\alpha \in U$ such that $\sigma(\alpha) = \beta$. Then $\tau\sigma(\alpha) = \tau(\beta) \neq 0$. This is a contradiction and hence $\tau = 0$. Now, suppose $\tau\sigma = 0$ implies $\tau = 0$. If σ is not an epimorphism then $\mathrm{Im}(\sigma)$ is a subspace of V but $\mathrm{Im}(\sigma) \neq V$. Let $\{\beta_1, \ldots, \beta_r\}$ be a basis of $\mathrm{Im}(\sigma)$, and extend this independent set to a basis $\{\beta_1, \ldots, \beta_r, \ldots, \beta_m\}$ of V. Define $\tau(\beta_i) = \beta_i$ for $i > r$ and $\tau(\beta_i) = 0$ for

$i \leq r$. Then $\tau\sigma = 0$ and $\tau \neq 0$. This is a contradiction and, hence, σ is an epimorphism.

Now, assume $\tau\sigma$ is defined and $\tau\sigma = 0$. Suppose τ is a monomorphism. If $\sigma \neq 0$, there is an $\alpha \in U$ such that $\sigma(\alpha) \neq 0$. Since τ is a monomorphism, $\tau\sigma(\alpha) \neq 0$. This is a contradiction and, hence, $\sigma = 0$. Now assume $\tau\sigma = 0$ implies $\sigma = 0$. If τ is not a monomorphism there is an $\alpha \in U$ such that $\alpha \neq 0$ and $\tau(\alpha) = 0$. Let $\{\alpha_1, \ldots, \alpha_n\}$ be any basis of U. Define $\sigma(\alpha_i) = \alpha$ for each i. Then $\tau\sigma(\alpha_i) = \tau(\alpha) = 0$ for all i and $\tau\sigma = 0$. This is a contradiction and, hence, τ is a monomorphism. □

Corollary 1.16. *σ is an epimorphism if and only if $\tau_1\sigma = \tau_2\sigma$ implies $\tau_1 = \tau_2$. τ is a monomorphism if and only if $\tau\sigma_1 = \tau\sigma_2$ implies $\sigma_1 = \sigma_2$.*

The statement that $\tau_1\sigma = \tau_2\sigma$ implies $\tau_1 = \tau_2$ is called a *right-cancellation*, and the statement that $\tau\sigma_1 = \tau\sigma_2$ implies $\sigma_1 = \sigma_2$ is called a *left-cancellation*. Thus, an epimorphism is a linear transformation that can be cancelled on the right, and a monomorphism is a linear transformation that can be cancelled on the left.

Theorem 1.17. *Let $A = \{\alpha_1, \ldots, \alpha_n\}$ be any basis of U. Let $B = \{\beta_1, \ldots, \beta_n\}$ be any n vectors in V* (not necessarily linearly independent). *There exists a uniquely determined linear transformation σ of U into V such that $\sigma(\alpha_i) = \beta_i$ for $i = 1, 2, \ldots, n$.*

PROOF. Since A is a basis of U, any vector $\alpha \in U$ can be expressed uniquely in the form $\alpha = \sum_{i=1}^{n} a_i\alpha_i$. If σ is to be linear we must have

$$\sigma(\alpha) = \sum_{i=1}^{n} a_i\sigma(\alpha)_i = \sum_{i=1}^{n} a_i\beta_i \in U.$$

It is a simple matter to verify that the mapping so defined is linear. □

Corollary 1.18. *Let $C = \{\gamma_1, \ldots, \gamma_r\}$ be any linearly independent set in U, where U is finite dimensional. Let $D = \{\delta_1, \ldots, \delta_r\}$ be any r vectors in V. There exists a linear transformation σ of U into V such that $\sigma(\gamma_i) = \delta_i$ for $i = 1, \ldots, r$.*

PROOF. Extend C to a basis of U. Define $\sigma(\gamma_i) = \delta_i$ for $i = 1, \ldots, r$, and define the values of σ on the other elements of the basis arbitrarily. This will yield a linear transformation σ with the desired properties. □

It should be clear that, if C is not already a basis, there are many ways to define σ. It is worth pointing out that the independence of the set C is crucial to proving the existence of the linear transformation with the desired properties. Otherwise, a linear relation among the elements of C would impose a corresponding linear relation among the elements of D, which would mean that D could not be arbitrary.

Theorem 1.17 establishes, for one thing, that linear transformations really do exist. Moreover, they exist in abundance. The real utility of this theorem and its corollary is that it enables us to establish the existence of a linear transformation with some desirable property with great convenience. All we have to do is to define this function on an independent set.

Definition. A linear transformation π of V into itself with the property that $\pi^2 = \pi$ is called a *projection.*

Theorem 1.19. *If π is a projection of V into itself, then $V = \mathrm{Im}(\pi) \oplus K(\pi)$ and π acts like the identity on* $\mathrm{Im}(\pi)$.

PROOF. For $\alpha \in V$, let $\alpha_1 = \pi(\alpha)$. Then $\pi(\alpha_1) = \pi^2(\alpha) = \pi(\alpha) = \alpha_1$. This shows that π acts like the identity on $\mathrm{Im}(\pi)$. Let $\alpha_2 = \alpha - \alpha_1$. Then $\pi(\alpha_2) = \pi(\alpha) - \pi(\alpha_1) = \alpha_1 - \alpha_1 = 0$. Thus $\alpha = \alpha_1 + \alpha_2$ where $\alpha_1 \in \mathrm{Im}(\pi)$ and $\alpha_2 \in K(\pi)$. Clearly, $\mathrm{Im}(\pi) \cap K(\pi) = \{0\}$. □

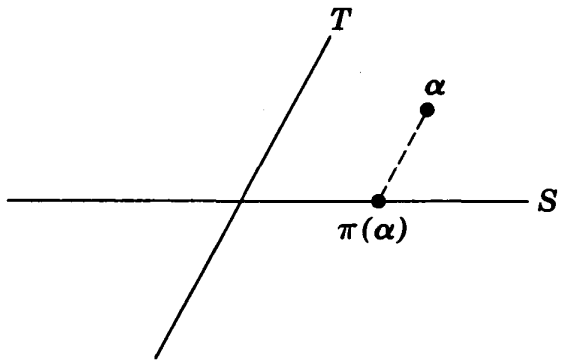

Fig. 1

If $S = \mathrm{Im}(\pi)$ and $T = K(\pi)$, we say that π is a projection of V onto S along T. In the case where V is the real plane, Fig. 1 indicates the interpretation of these words. α is projected onto a point of S in a direction parallel to T.

EXERCISES

1. Show that $\sigma((x_1, x_2)) = (x_2, x_1)$ defines a linear transformation of R^2 into itself.

2. Let $\sigma_1((x_1, x_2)) = (x_2, -x_1)$ and $\sigma_2((x_1, x_2)) = (x_1, -x_2)$. Determine $\sigma_1 + \sigma_2$, $\sigma_1\sigma_2$ and $\sigma_2\sigma_1$.

3. Let $U = V = R^n$ and let $\sigma((x_1, x_2, \ldots, x_n)) = (x_1, x_2, \ldots, x_k, 0, \ldots, 0)$ where $k < n$. Describe $\mathrm{Im}(\sigma)$ and $K(\sigma)$.

4. Let $\sigma((x_1, x_2, x_3, x_4)) = (3x_1 - 2x_2 - x_3 - 4x_4, x_1 + x_2 - 2x_3 - 3x_4)$. Show that σ is a linear transformation. Determine the kernel of σ.

5. Let $\sigma((x_1, x_2, x_3)) = (2x_1 + x_2 + 3x_3, 3x_1 - x_2 + x_3, -4x_1 + 3x_2 + x_3)$. Find

a basis of $\sigma(U)$. (*Hint:* Take particular values of the x_i to find a spanning set for $\sigma(U)$.) Find a basis of $K(\sigma)$.

6. Let D denote the operator of differentiation,

$$D(y) = \frac{dy}{dx}, \quad D^2(y) = D[D(y)] = \frac{d^2y}{dx^2}, \text{ etc.}$$

Show that D^n is a linear transformation, and also that $p(D)$ is a linear transformation if $p(D)$ is a polynomial in D with constant coefficients. (Here we must assume that the space of functions on which D is defined contains only functions differentiable at least as often as the degree of $p(D)$.)

7. Let $U = V$ and let σ and τ be linear transformations of U into itself. In this case $\sigma\tau$ and $\tau\sigma$ are both defined. Construct an example to show that it is not always true that $\sigma\tau = \tau\sigma$.

8. Let $U = V = P$, the space of polynomials in x with coefficients in R. For $\alpha = \sum_{i=0}^{n} a_i x^i$ let

$$\sigma(\alpha) = \sum_{i=0}^{n} i a_i x^{i-1}$$

and

$$\tau(\alpha) = \sum_{i=0}^{n} \frac{a_i}{i+1} x^{i+1}.$$

Show that $\sigma\tau = 1$, but that $\tau\sigma \neq 1$.

9. Show that if two scalar transformations coincide on U then the defining scalars are equal.

10. Let σ be a linear transformation of U into V and let $A = \{\alpha_1, \ldots, \alpha_n\}$ be a basis of U. Show that if the values $\{\sigma(\alpha_1), \ldots, \sigma(\alpha_n)\}$ are known, then the value of $\sigma(\alpha)$ can be computed for each $\alpha \in U$.

11. Let U and V be vector spaces of dimensions n and m, respectively, over the same field F. We have already commented that the set of all linear transformations of U into V forms a vector space. Give the details of the proof of this assertion. Let $A = \{\alpha_1, \ldots, \alpha_n\}$ be a basis of U and $B = \{\beta_1, \ldots, \beta_m\}$ be a basis of V. Let σ_{ij} be the linear transformation of U into V such that

$$\sigma_{ij}(\alpha_k) = \begin{cases} 0 & \text{if } k \neq j, \\ \beta_i & \text{if } k = j. \end{cases}$$

Show that $\{\sigma_{ij} \mid i = 1, \ldots, m; j = 1, \ldots, n\}$ is a basis of this vector space.

For the following sequence of problems let $\dim U = n$ and $\dim V = m$. Let σ be a linear transformation of U into V and τ a linear transformation of V into W.

12. Show that $\rho(\sigma) \leq \rho(\tau\sigma) + \nu(\tau)$. (*Hint:* Let $V' = \sigma(U)$ and apply Theorem 1.6 to τ defined on V'.)

13. Show that $\max\{0, \rho(\sigma) + \rho(\tau) - m\} \leq \rho(\tau\sigma) \leq \min\{\rho(\tau), \rho(\sigma)\}$.

14. Show that $\max\{n - m + \nu(\tau), \nu(\sigma)\} \leq \nu(\tau\sigma) \leq \min\{n, \nu(\sigma) + \nu(\tau)\}$. (For $m = n$ this inequality is known as Sylvester's law of nullity.)

15. Show that if $\nu(\tau) = 0$, then $\rho(\tau\sigma) = \rho(\sigma)$.

16. It is not generally true that $\nu(\sigma) = 0$ implies $\rho(\tau\sigma) = \rho(\tau)$. Construct an example to illustrate this fact. (*Hint:* Let m be very large.)

17. Show that if $m = n$ and $\nu(\sigma) = 0$, then $\rho(\tau\sigma) = \rho(\tau)$.

18. Show that if σ_1 and σ_2 are linear transformations of U into V, then

$$\rho(\sigma_1 + \sigma_2) \leq \min\{m, n, \rho(\sigma_1) + \rho(\sigma_2)\}.$$

19. Show that $|\rho(\sigma_1) - \rho(\sigma_2)| \leq \rho(\sigma_1 + \sigma_2)$.

20. If S is any subspace of V there is a subspace T such that $V = S \oplus T$. Then every $x \in V$ can be represented uniquely in the form $\alpha = \alpha_1 + \alpha_2$ where $\alpha_1 \in S$ and $\alpha_2 \in T$. Show that the mapping π which maps α onto α_1 is a linear transformation. Show that T is the kernel of π. Show that $\pi^2 = \pi$. The mapping π is called a *projection of* V onto S along T.

21. (Continuation) Let π be a projection. Show that $1 - \pi$ is also a projection. What is the kernel of $1 - \pi$? Onto what subspace is $1 - \pi$ a projection? Show that $\pi(1 - \pi) = 0$.

2 | Matrices

Definition. A *matrix* over a field F is a rectangular array of scalars. The array will be written in the form

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{2.1}$$

whenever we wish to display all the elements in the array or show the form of the array. A matrix with m rows and n columns is called an $m \times n$ matrix. An $n \times n$ matrix is said to be of *order* n.

We often abbreviate a matrix written in the form above to $[a_{ij}]$ where the first index denotes the number of the row and the second index denotes the number of the column. The particular letter appearing in each index position is immaterial; it is the position that is important. With this convention a_{ij} is a scalar and $[a_{ij}]$ is a matrix. Whereas the elements a_{ij} and a_{kl} need not be equal, we consider the matrices $[a_{ij}]$ and $[a_{kl}]$ to be identical since both $[a_{ij}]$ and $[a_{kl}]$ stand for the entire matrix. As a further convenience we often use upper case Latin italic letters to denote matrices; $A = [a_{ij}]$. Whenever we use lower case Latin italic letters to denote the scalars appearing

in the matrix, we use the corresponding upper case Latin italic letter to denote the matrix. The matrix in which all scalars are zero is denoted by 0 (the third use of this symbol!). The a_{ij} appearing in the array $[a_{ij}]$ are called the *elements* of $[a_{ij}]$. Two matrices are equal if and only if they have exactly the same elements. The *main diagonal* of the matrix $[a_{ij}]$ is the set of elements $\{a_{11}, \ldots, a_{tt}\}$ where $t = \min\{m, n\}$. A *diagonal matrix* is a square matrix in which the elements not in the main diagonal are zero.

Matrices can be used to represent a variety of different mathematical concepts. The way matrices are manipulated depends on the objects which they represent. Considering the wide variety of situations in which matrices have found application, there is a remarkable similarity in the operations performed on matrices in these situations. There are differences too, however, and to understand these differences we must understand the object represented and what information can be expected by manipulating with the matrices. We first investigate the properties of matrices as representations of linear transformations. Not only do the matrices provide us with a convenient means of doing whatever computation is necessary with linear transformations, but the theory of vector spaces and linear transformations also proves to be a powerful tool in developing the properties of matrices.

Let U be a vector space of dimension n and V a vector space of dimension m, both over the same field F. Let $\mathsf{A} = \{\alpha_1, \ldots, \alpha_n\}$ be an arbitrary but fixed basis of U, and let $\mathsf{B} = \{\beta_1, \ldots, \beta_m\}$ be an arbitrary but fixed basis of V. Let σ be a linear transformation of U into V. Since $\sigma(\alpha_j) \in V$, $\sigma(\alpha_j)$ can be expressed uniquely as a linear combination of the elements of B;

$$\sigma(\alpha_j) = \sum_{i=1}^{m} a_{ij}\beta_i. \tag{2.2}$$

We define the *matrix representing σ with respect to the bases A and B* to be the matrix $A = [a_{ij}]$.

The correspondence between linear transformations and matrices is actually one-to-one and onto. Given the linear transformation σ, the a_{ij} exist because B spans V, and they are unique because B is linearly independent. On the other hand, let $A = [a_{ij}]$ be any $m \times n$ matrix. We can define $\sigma(\alpha_j) = \sum_{i=1}^{m} a_{ij}\beta_i$ for each $\alpha_j \in \mathsf{A}$, and then we can extend the proposed linear transformation to all of U by the condition that it be linear. Thus, if $\xi = \sum_{j=1}^{n} x_j\alpha_j$, we define

$$\begin{aligned}\sigma(\xi) &= \sum_{j=1}^{n} x_j\sigma(\alpha_j) \\ &= \sum_{j=1}^{n} x_j\left(\sum_{i=1}^{m} a_{ij}\beta_i\right) \\ &= \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{ij}x_j\right)\beta_i. \end{aligned} \tag{2.3}$$

σ can be extended to all of U because A spans U, and the result is well defined (unique) because A is linearly independent.

Here are some examples of linear transformations and the matrices which represent them. Consider the real plane $R^2 = U = V$. Let $A = B = \{(1, 0), (0, 1)\}$. A 90° rotation counterclockwise would send $(1, 0)$ onto $(0, 1)$ and it would send $(0, 1)$ onto $(-1, 0)$. Since $\sigma((1, 0)) = 0 \cdot (1, 0) + 1 \cdot (0, 1)$ and $\sigma((0, 1)) = (-1) \cdot (1, 0) + 0 \cdot (0, 1)$, σ is represented by the matrix

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

The elements appearing in a column are the coordinates of each image of a basis vector under a transformation.

In general, a rotation counterclockwise through an angle of θ will send $(1, 0)$ onto $(\cos\theta, \sin\theta)$ and $(0, 1)$ onto $(-\sin\theta, \cos\theta)$. Thus this rotation is represented by

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{2.4}$$

Suppose now that τ is another linear transformation of U into V represented by the matrix $B = [b_{ij}]$. Then for the sum $\sigma + \tau$ we have

$$\begin{aligned}(\sigma + \tau)(\alpha_j) = \sigma(\alpha_j) + \tau(\alpha_j) &= \sum_{i=1}^{m} a_{ij}\beta_i + \sum_{i=1}^{m} b_{ij}\beta_i \\ &= \sum_{i=1}^{m} (a_{ij} + b_{ij})\beta_i. \end{aligned} \tag{2.5}$$

Thus $\sigma + \tau$ is represented by the matrix $[a_{ij} + b_{ij}]$. Accordingly, we define the *sum* of two matrices to be that matrix obtained by the addition of the corresponding elements in the two arrays; $A + B = [a_{ij} + b_{ij}]$ is the matrix corresponding to $\sigma + \tau$. The sum of two matrices is defined if and only if the two matrices have the same number of rows and the same number of columns.

If a is any scalar, for the linear transformation $a\sigma$ we have

$$(a\sigma)(\alpha_j) = a\sum_{i=1}^{m} a_{ij}\beta_i = \sum_{i=1}^{m} (aa_{ij})\beta_i. \tag{2.6}$$

Thus $a\sigma$ is represented by the matrix $[aa_{ij}]$. We therefore define *scalar multiplication* by the rule $aA = [aa_{ij}]$.

Let W be a third vector space of dimension r over the field F, and let $C = \{\gamma_1, \ldots, \gamma_r\}$ be an arbitrary but fixed basis of W. If the linear transformation σ of U into V is represented by the $m \times n$ matrix $A = [a_{ij}]$ and the

linear transformation τ of V into W is represented by the $r \times m$ matrix $B = [b_{ki}]$, what matrix represents the linear transformation $\tau\sigma$ of U into W?

$$\begin{aligned}(\tau\sigma)(\alpha_j) &= \tau(\sigma(\alpha_j)) = \tau\left(\sum_{i=1}^{m} a_{ij}\beta_i\right) \\ &= \sum_{i=1}^{m} a_{ij}\tau(\beta_i) \\ &= \sum_{i=1}^{m} a_{ij}\left(\sum_{k=1}^{r} b_{ki}\gamma_k\right) \\ &= \sum_{k=1}^{r}\left(\sum_{i=1}^{m} b_{ki}a_{ij}\right)\gamma_k. \qquad (2.7)\end{aligned}$$

Thus, if we define $c_{kj} = \sum_{i=1}^{m} b_{ki}a_{ij}$, then $C = [c_{kj}]$ is the matrix representing the product transformation $\tau\sigma$. Accordingly, we call C the *matrix product* of B and A, in that order: $C = BA$.

For computational purposes it is customary to write the arrays of B and A side by side. The element c_{kj} of the product is then obtained by multiplying the corresponding elements of row k of B and column j of A and adding. We can trace the elements of row k of B with a finger of the left hand while at the same time tracing the elements of column j of A with a finger of the right hand. At each step we compute the product of the corresponding elements and accumulate the sum as we go along. Using this simple rule we can, with practice, become quite proficient, even to the point of doing "without hands."

Check the process in the following examples:

$$\begin{bmatrix} 1 & 4 & -1 & 2 \\ 0 & 2 & 1 & 3 \\ -2 & 1 & -2 & 2 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 0 & 2 \\ 2 & 1 \\ 3 & -2 \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ 11 & -1 \\ 0 & -2 \end{bmatrix}.$$

All definitions and properties we have established for linear transformations can be carried over immediately for matrices. For example, we have:

1. $0 \cdot A = 0$. (The "0" on the left is a scalar, the "0" on the right is a matrix with the same number of rows and columns as A.)
2. $1 \cdot A = A$.
3. $A(B + C) = AB + AC$.
4. $(A + B)C = AC + BC$.
5. $A(BC) = (AB)C$.

Of course, in each of the above statements we must assume the operations proposed are well defined. For example, in 3, B and C must be the same

size and A must have the same number of columns as B and C have rows.

The *rank* and *nullity* of a matrix A are the rank and nullity of the associated linear transformation, respectively.

Theorem 2.1. *For an $m \times n$ matrix A, the rank of A plus the nullity of A is equal to n. The rank of a product BA is less than or equal to the rank of either factor.*

These statements have been established for linear transformations and therefore hold for their corresponding matrices. □

The rank of σ is the dimension of the subspace $\mathrm{Im}(\sigma)$ of V. Since $\mathrm{Im}(\sigma)$ is spanned by $\{\sigma(\alpha_1), \ldots, \sigma(\alpha_n)\}$, $\rho(\sigma)$ is the number of elements in a maximal linearly independent subset of $\{\sigma(\alpha_1), \ldots, \sigma(\alpha_n)\}$. Expressed in terms of coordinates, $\sigma(\alpha_j) = \sum_{i=1}^{m} a_{ij}\beta_i$ is represented by the m-tuple $(a_{1j}, a_{2j}, \ldots, a_{mj})$, which is the m-tuple in column j of the matrix $[a_{ij}]$. Thus $\rho(\sigma) = \rho(A)$ is also equal to the maximum number of linearly independent columns of A. This is usually called the column rank of a matrix A, and the maximum number of linearly independent rows of A is called the row rank of A. We, however, show before long that the number of linearly independent rows in a matrix is equal to the number of linearly independent columns. Until that time we consider "rank" and "column rank" as synonymous.

Returning to Equation (2.3), we see that, if $\xi \in U$ is represented by $(x_1, \ldots, x_n)$ and the linear transformation σ of U into V is represented by the matrix $A = [a_{ij}]$, then $\sigma(\xi) \in V$ is represented by $(y_1, \ldots, y_m)$ where

$$y_i = \sum_{j=1}^{n} a_{ij}x_j \qquad (i = 1, \ldots, m). \tag{2.8}$$

In view of the definition of matrix multiplication given by Equation (2.7) we can interpret Equations (2.8) as a matrix product of the form

$$Y = AX \tag{2.9}$$

where

$$Y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}.$$

This single matric equation contains the m equations in (2.8).

We have already used the n-tuple $(x_1, \ldots, x_n)$ to represent the vector $\xi = \sum_{i=1}^{n} x_i\alpha_i$. Because of the usefulness of equation (2.9) we also find it convenient to represent ξ by the one-column matrix X. In fact, since it is

somewhat wasteful of space and otherwise awkward to display one-column matrices we use the n-tuple $(x_1, \ldots, x_n)$ to represent not only the vector ξ but also the column matrix X. With this convention $[x_1 \cdots x_n]$ is a one-row matrix and $(x_1, \ldots, x_n)$ is a one-column matrix.

Notice that we have now used matrices for two different purposes, (1) to represent linear transformations, and (2) to represent vectors. The single matric equation $Y = AX$ contains some matrices used in each way.

EXERCISES

1. Verify the matrix multiplication in the following examples:

(*a*)
$$\begin{bmatrix} 3 & 1 & -2 \\ -5 & 2 & 3 \end{bmatrix}\begin{bmatrix} 2 & 1 & -3 \\ -1 & 6 & 1 \\ 1 & 0 & -2 \end{bmatrix} = \begin{bmatrix} 3 & 9 & -4 \\ -9 & 7 & 11 \end{bmatrix}.$$

(*b*)
$$\begin{bmatrix} 2 & 1 & -3 \\ -1 & 6 & 1 \\ 1 & 0 & -2 \end{bmatrix}\begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 10 \\ 15 \\ 4 \end{bmatrix}.$$

(*c*)
$$\begin{bmatrix} 3 & 1 & -2 \\ -5 & 2 & 3 \end{bmatrix}\begin{bmatrix} 10 \\ 15 \\ 4 \end{bmatrix} = \begin{bmatrix} 37 \\ -8 \end{bmatrix}.$$

2. Compute
$$\begin{bmatrix} 3 & 9 & -4 \\ -9 & 7 & 11 \end{bmatrix}\begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}.$$

Interpret the answer to this problem in terms of the computations in Exercise 1.

3. Find AB and BA if
$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ -1 & -2 & -3 & -4 \\ -5 & -6 & -7 & -8 \end{bmatrix}.$$

4. Let σ be a linear transformation of R^2 into itself that maps $(1, 0)$ onto $(3, -1)$ and $(0, 1)$ onto $(-1, 2)$. Determine the matrix representing σ with respect to the bases $A = B = \{(1, 0), (0, 1)\}$.

5. Let σ be a linear transformation of R^2 into itself that maps $(1, 1)$ onto $(2, -3)$ and $(1, -1)$ onto $(4, -7)$. Determine the matrix representing σ with respect to the bases $A = B = \{(1, 0), (0, 1)\}$. (*Hint:* We must determine the effect of σ when it is applied to $(1, 0)$ and $(0, 1)$. Use the fact that $(1, 0) = \frac{1}{2}(1, 1) + \frac{1}{2}(1, -1)$ and the linearity of σ.)

6. It happens that the linear transformation defined in Exercise 4 is one-to-one, that is, σ does not map two different vectors onto the same vector. Thus, there is a linear transformation that maps $(3, -1)$ onto $(1, 0)$ and $(-1, 2)$ onto $(0, 1)$. This linear transformation reverses the mapping given by σ. Determine the matrix representing it with respect to the same bases.

7. Let us consider the geometric meaning of linear transformations. A linear transformation of R^2 into itself leaves the origin fixed (why?) and maps straight lines into straight lines. (The word "into" is required here because the image of a straight line may be another straight line or it may be a single point.) Prove that the image of a straight line is a subset of a straight line. (*Hint:* Let σ be represented by the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Then σ maps (x, y) onto $(a_{11}x + a_{12}y, a_{21}x + a_{22}y)$. Now show that if (x, y) satisfies the equation $ax + by = c$ its image satisfies the equation

$$(aa_{22} - ba_{21})x + (a_{11}b - a_{12}a)y = (a_{11}a_{22} - a_{12}a_{21})c.)$$

8. (Continuation) We say that a straight line is mapped onto itself if every point on the line is mapped onto a point on the line (but not all onto the same point) even though the points on the line may be moved around.

(*a*) A linear transformation maps $(1, 0)$ onto $(-1, 0)$ and $(0, 1)$ onto $(0, -1)$. Show that every line through the origin is mapped onto itself. Show that each such line is mapped onto itself with the sense of direction inverted. This linear transformation is called an *inversion* with respect to the origin. Find the matrix representing this linear transformation with respect to the basis $\{(1, 0), (0, 1)\}$.

(*b*) A linear transformation maps $(1, 1)$ onto $(-1, -1)$ and leaves $(1, -1)$ fixed. Show that every line perpendicular to the line $x_1 + x_2 = 0$ is mapped onto itself with the sense of direction inverted. Show that every point on the line $x_1 + x_2 = 0$ is left fixed. Which lines through the origin are mapped onto themselves? This linear transformation is called a *reflection* about the line $x_1 + x_2 = 0$. Find the matrix representing this linear transformation with respect to the basis $\{(1, 0), (0, 1)\}$. Find the matrix representing this linear transformation with respect to the basis $\{(1, 1), (1, -1)\}$.

(*c*) A liner transformation maps $(1, 1)$ onto $(2, 2)$ and $(1, -1)$ onto $(3, -3)$. Show that the lines through the origin and passing through the points $(1, 1)$ and $(1, -1)$ are mapped onto themselves and that no other lines are mapped onto themselves. Find the matrices representing this linear transformation with respect to the bases $\{(1, 0), (0, 1)\}$ and $\{(1, 1), (1, -1)\}$.

(*d*) A linear transformation leaves (1, 0) fixed and maps (0, 1) onto (1, 1). Show that each line $x_2 = c$ is mapped onto itself and translated within itself a distance equal to c. This linear transformation is called a *shear*. Which lines through the origin are mapped onto themselves? Find the matrix representing this linear transformation with respect to the basis $\{(1, 0), (0, 1)\}$.

(*e*) A linear transformation maps (1, 0) onto $(\frac{5}{13}, \frac{12}{13})$ and (0, 1) onto $(-\frac{12}{13}, \frac{5}{13})$. Show that every line through the origin is rotated counterclockwise through the angle $\theta = \arccos \frac{5}{13}$. This linear transformation is called a *rotation*. Find the matrix representing this linear transformation with respect to the basis $\{(1, 0), (0, 1)\}$.

(*f*) A linear transformation maps (1, 0) onto $(\frac{2}{3}, \frac{2}{3})$ and (0, 1) onto $(\frac{1}{3}, \frac{1}{3})$. Show that each point on the line $2x_1 + x_2 = 3c$ is mapped onto the single point (c, c). The line $x_1 - x_2 = 0$ is left fixed. The only other line through the origin which is mapped into itself is the line $2x_1 + x_2 = 0$. This linear transformation is called a *projection* onto the line $x_1 - x_2 = 0$ parallel to the line $2x_1 + x_2 = 0$. Find the matrices representing this linear transformation with respect to the bases $\{(1, 0), (0, 1)\}$ and $\{(1, 1), (1, -2)\}$.

9. (Continuation) Describe the geometric effect of each of the linear transformations of R^2 into itself represented by the matrices

$$(a)\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \quad (b)\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix} \quad (c)\begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix}$$

$$(d)\begin{bmatrix}1 & 0\\ a & 1\end{bmatrix} \quad (e)\begin{bmatrix}b & 0\\ 0 & c\end{bmatrix} \quad (f)\begin{bmatrix}\frac{3}{5} & -\frac{4}{5}\\ \frac{4}{5} & \frac{3}{5}\end{bmatrix}.$$

(*Hint:* In Exercise 7 we have shown that straight lines are mapped into straight lines. We already know that linear transformations map the origin onto the origin. Thus it is relatively easy to determine what happens to straight lines passing through the origin. For example, to see what happens to the x_1-axis it is sufficient to see what happens to the point (1, 0). Among the transformations given appear a rotation, a reflection, two projections, and one shear.)

10. (Continuation) For the linear transformations given in Exercise 9 find all lines through the origin which are mapped onto or into themselves.

11. Let $U = R^2$ and $V = R^3$ and σ be a linear transformation of U into V that maps (1, 1) onto (0, 1, 2) and (−1, 1) onto (2, 1, 0). Determine the matrix that represents σ with respect to the bases $A = \{(1, 0), (0, 1)\}$ in $B = \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ in R^3. (*Hint:* $\frac{1}{2}(1, 1) - \frac{1}{2}(-1, 1) = (1, 0)$.)

12. What is the effect of multiplying an $n \times n$ matrix A by an $n \times n$ diagonal matrix D? What is the difference between AD and DA?

13. Let a and b be two numbers such that $a \neq b$. Find all 2×2 matrices A such that

$$A\begin{bmatrix}a & 0\\ 0 & b\end{bmatrix} = \begin{bmatrix}a & 0\\ 0 & b\end{bmatrix}A.$$

14. Show that the matrix $C = [a_i b_j]$ has rank one if not all a_i and not all b_j are zero. (*Hint:* Use Theorem 1.12.)

15. Let a, b, c, and d be given numbers (real or complex) and consider the function

$$f(x) = \frac{ax + b}{cx + d}.$$

Let g be another function of the same form. Show that gf where $gf(x) = g(f(x))$ is a function that can also be written in the same form. Show that each of these functions can be represented by a matrix in such a way that the matrix representing gf is the product of the matrices representing g and f. Show that the inverse function exists if and only if $ad - bc \neq 0$. To what does the function reduce if $ad - bc = 0$?

16. Consider complex numbers of the form $x + yi$ (where x and y are real numbers and $i^2 = -1$) and represent such a complex number by the duple (x, y) in R^2. Let $a + bi$ be a fixed complex number. Consider the function f defined by the rule

$$f(x + yi) = (a + bi)(x + yi) = u + vi.$$

(*a*) Show that this function is a linear transformation of R^2 into itself mapping (x, y) onto (u, v).

(*b*) Find the matrix representing this linear transformation with respect to the basis $\{(1, 0), (0, 1)\}$.

(*c*) Find the matrix which represents the linear transformation obtained by using $c + di$ in place of $a + bi$. Compute the product of these two matrices. Do they commute?

(*d*) Determine the complex number which can be used in place of $a + bi$ to obtain a transformation represented by this matrix product. How is this complex number related to $a + bi$ and $c + di$?

17. Show by example that it is possible for two matrices A and B to have the same rank while A^2 and B^2 have different ranks.

3 | Non-singular Matrices

Let us consider the case where $U = V$, that is, we are considering transformations of V into itself. Generally, a homomorphism of a set into itself is called an *endomorphism.* We consider a fixed basis in V and represent the linear transformation of V into itself with respect to that basis. In this case the matrices are square or $n \times n$ matrices. Since the transformations we are considering map V into itself any finite number of them can be iterated in any order. The commutative law does not hold, however. The same remarks hold for square matrices. They can be multiplied in any order but

the commutative law does not hold. For example

$$\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix},$$
$$\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

The linear transformation that leaves every element of V fixed is the identity transformation. We denote the identity transformation by 1, the scalar identity. Clearly, the identity transformation is represented by the matrix $I = [\delta_{ij}]$ for any choice of the basis. Notice that $IA = AI = A$ for any $n \times n$ matrix A. I is called the *identity matrix*, or *unit matrix*, of order n. If we wish to point out the dimension of the space we write I_n for the identity matrix of order n. The scalar transformation a is represented by the matrix aI. Matrices of the form aI are called *scalar matrices.*

Definition. A one-to-one linear transformation σ of a vector space onto itself is called an *automorphism.* An automorphism is only a special kind of isomorphism for which the domain and codomain are the same space. If $\sigma(\alpha) = \bar{\alpha}$, the mapping $\sigma^{-1}(\bar{\alpha}) = \alpha$ is called the *inverse transformation* of σ. The rotations represented in Section 2 are examples of automorphisms.

Theorem 3.1. *The inverse σ^{-1} of an automorphism σ is an automorphism.*

Theorem 3.2 *A linear transformation τ of an n-dimensional vector space into itself is an automorphism if and only if it is of rank n; that is, if and only if it is an epimorphism.*

Theorem 3.3. *A linear transformation σ of an n-dimensional vector space into itself is an automorphism if and only if its nullity is 0, that is, if and only if it is a monomorphism.*

PROOF (of Theorems 3.1, 3.2, and 3.3). These properties have already been established for isomorphisms. □

Since it is clear that transformations of rank less than n do not have inverses because they are not onto, we see that automorphisms are the only linear transformations which have inverses. A linear transformation that has an inverse is said to be *non-singular* or *invertible*; otherwise it is said to be *singular.* Let A be the matrix representing the automorphism σ, and let A^{-1} be the matrix representing the inverse transformation σ^{-1}. The matrix $A^{-1}A$ represents the transformation $\sigma^{-1}\sigma$. Since $\sigma^{-1}\sigma$ is the identity transformation, we must have $A^{-1}A = I$. But σ is also the inverse transformation of σ^{-1} so that $\sigma\sigma^{-1} = 1$ and $AA^{-1} = I$. We shall refer to A^{-1} as the *inverse* of A. A matrix that has an inverse is said to be *non-singular* or *invertible.* Only a square matrix can have an inverse.

On the other hand suppose that for the matrix A there exists a matrix B such that $BA = I$. Since I is of rank n, A must also be of rank n and, therefore, A represents an automorphism σ. Furthermore, the linear transformation which B represents is necessarily the inverse transformation σ^{-1} since the product with σ must yield the identity transformation. Thus $B = A^{-1}$. The same kind of argument shows that if C is a matrix such that $AC = I$, then $C = A^{-1}$. Thus we have shown:

Theorem 3.4. *If A and B are square matrices such that $BA = I$, then $AB = I$. If A and B are square matrices such that $AB = I$, then $BA = I$. In either case B is the unique inverse of A.* □

Theorem 3.5. *If A and B are non-singular, then* (1) *AB is non-singular and $(AB)^{-1} = B^{-1}A^{-1}$,* (2) *A^{-1} is non-singular and $(A^{-1})^{-1} = A$,* (3) *for $a \neq 0$, aA is non-singular and $(aA)^{-1} = a^{-1}A^{-1}$.*

PROOF. In view of the remarks preceding Theorem 3.4 it is sufficient in each case to produce a matrix which will act as a left inverse.

(1) $(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}IB = B^{-1}B = I$.

(2) $AA^{-1} = I$.

(3) $(a^{-1}A^{-1})(aA) = (a^{-1}a)(A^{-1}A) = I$. □

Theorem 3.6. *If A is non-singular, we can solve uniquely the equations $XA = B$ and $AY = B$ for any matrix B of the proper size, but the two solutions need not be equal.*

PROOF. Solutions exist since $(BA^{-1})A = B(A^{-1}A) = B$ and $A(A^{-1}B) = (AA^{-1})B = B$. The solutions are unique since for any C having the property that $CA = B$ we have $C = CAA^{-1} = BA^{-1}$, and similarly with any solution of $AY = B$. □

As an example illustrating the last statement of the theorem, let

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}.$$

Then

$$X = BA^{-1} = \begin{bmatrix} 1 & -2 \\ 2 & -3 \end{bmatrix}, \quad \text{and} \quad Y = A^{-1}B = \begin{bmatrix} -3 & -2 \\ 2 & 1 \end{bmatrix}.$$

We add the remark that for non-singular A, the solution of $XA = B$ exists and is unique if B has n columns, and the solution of $AY = B$ exists and is unique if B has n rows. The proof given for Theorem 3.6 applies without change.

Theorem 3.7. *The rank of a (not necessarily square) matrix is not changed by multiplication by a non-singular matrix.*

PROOF. Let A be non-singular and let B be of rank ρ. Then by Theorem 2.1 AB is of rank $r \leq \rho$, and $A^{-1}(AB) = B$ is of rank $\rho \leq r$. Thus $r = \rho$. The proof that BA is of rank ρ is similar. □

Theorem 1.14 states the corresponding property for linear transformations.

The existence or non-existence of the inverse of a square matrix depends on the matrix itself and not on whether it represents a linear transformation of a vector space into itself or a linear transformation of one vector space into another. Thus it is convenient and consistent to extend our usage of the term "non-singular" to include isomorphisms. Accordingly any square matrix with an inverse is *non-singular*.

Let U and V be vector spaces of dimension n over the field F. Let $A = \{\alpha_1, \ldots, \alpha_n\}$ be a basis of U and $B = \{\beta_1, \ldots, \beta_n\}$ be a basis of V. If $\xi = \sum_{i=1}^{n} x_i\alpha_i$ is any vector in U we can define $\sigma(\xi)$ to be $\sum_{i=1}^{n} x_i\beta_i$. It is easily seen that σ is an isomorphism and that ξ and $\sigma(\xi)$ are both represented by $(x_1, \ldots, x_n) \in F^n$. Thus any two vector spaces of the same dimension over F are isomorphic. As far as their internal structure is concerned they are indistinguishable. Whatever properties may serve to distinguish them are, by definition, not vector space properties.

EXERCISES

1. Show that the inverse of

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 6 \end{bmatrix} \quad \text{is} \quad A^{-1} = \begin{bmatrix} -2 & 0 & 1 \\ 0 & 3 & -2 \\ 1 & -2 & 1 \end{bmatrix}.$$

2. Find the square of the matrix

$$A = \tfrac{1}{3}\begin{bmatrix} 1 & 2 & 2 \\ 2 & -2 & 1 \\ 2 & 1 & -2 \end{bmatrix}.$$

What is the inverse of A? (Geometrically, this matrix represents a 180° rotation about the line containing the vector (2, 1, 1). The inverse obtained is therefore not surprising.)

3. Compute the image of the vector (1, −2, 1) under the linear transformation represented by the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 0 & 1 & 2 \end{bmatrix}.$$

Show that A cannot have an inverse.

4. Since

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix} = \begin{bmatrix} 3x_{11} - 5x_{12} & -x_{11} + 2x_{12} \\ 3x_{21} - 5x_{22} & -x_{21} + 2x_{22} \end{bmatrix}$$

we can find the inverse of $\begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix}$ by solving the equations

$$\begin{aligned} 3x_{11} - 5x_{12} &= 1 \\ -x_{11} + 2x_{12} &= 0 \\ 3x_{21} - 5x_{22} &= 0 \\ -x_{21} + 2x_{22} &= 1. \end{aligned}$$

Solve these equations and check your answer by showing that this gives the inverse matrix.

We have not as yet developed convenient and effective methods for obtaining the inverse of a given matrix. Such methods are developed later in this chapter and in the following chapter. If we know the geometric meaning of the matrix, however, it is often possible to obtain the inverse with very little work.

5. The matrix $\begin{bmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix}$ represents a rotation about the origin through the angle $\theta = \arccos \frac{3}{5}$. What rotation would be the inverse of this rotation? What matrix would represent this inverse rotation? Show that this matrix is the inverse of the given matrix.

6. The matrix $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ represents a reflection about the line $x_1 + x_2 = 0$. What operation is the inverse of this reflection? What matrix represents the inverse operation? Show that this matrix is the inverse of the given matrix.

7. The matrix $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ represents a shear. The inverse transformation is also a shear. Which one? What matrix represents the inverse shear? Show that this matrix is the inverse of the given matrix.

8. Show that the transformation that maps (x_1, x_2, x_3) onto $(x_3, -x_1, x_2)$ is an automorphism of F^3. Find the matrix representing this automorphism and its inverse with respect to the basis $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$.

9. Show that an automorphism of a vector space maps every subspace onto a subspace of the same dimension.

10. Find an example to show that there exist non-square matrices A and B such that $AB = I$. Specifically, show that there is an $m \times n$ matrix A and an $n \times m$ matrix B such that AB is the $m \times m$ identity. Show that BA is not the $n \times n$ identity. Prove in general that if $m \neq n$, then AB and BA cannot both be identity matrices.

4 | Change of Basis

We have represented vectors and linear transformations as n-tuples and matrices with respect to arbitrary but fixed bases. A very natural question arises: What changes occur in these representations if other choices for bases are made? The vectors and linear transformations have meaning independent of any particular choice of bases, independent of any coordinate systems, but their representations are entirely dependent on the bases chosen.

Definition. Let $A = \{\alpha_1, \ldots, \alpha_n\}$ and $A' = \{\alpha'_1, \ldots, \alpha'_n\}$ be bases of the vector space U. In a typical "change of basis" situation the representations of various vectors and linear transformations are known in terms of the basis A, and we wish to determine their representations in terms of the basis A'. In this connection, we refer to A as the "old" basis and to A' as the "new" basis. Each α'_j is expressible as a linear combination of the elements of A; that is,

$$\alpha'_j = \sum_{i=1}^{n} p_{ij}\alpha_i. \tag{4.1}$$

The associated matrix $P = [p_{ij}]$ is called the *matrix of transition* from the basis A to the basis A'.

The columns of P are the n-tuples representing the new basis vectors in terms of the old basis. This simple observation is worth remembering as it is usually the key to determining P when a change of basis is made. Since the columns of P are the representations of the basis A' they are linearly independent and P has rank n. Thus P is non-singular.

Now let $\xi = \sum_{i=1}^{n} x_i\alpha_i$ be an arbitrary vector of U and let $\xi = \sum_{i=1}^{n} x'_i\alpha'_i$ be the representation of ξ in terms of the basis A'. Then

$$\begin{aligned}\xi &= \sum_{j=1}^{n} x'_j\alpha'_j = \sum_{j=1}^{n} x'_j\left(\sum_{i=1}^{n} p_{ij}\alpha_i\right)\\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} p_{ij}x'_j\right)\alpha_i.\end{aligned} \tag{4.2}$$

Since the representation of ξ with respect to the basis A is unique we see that $x_i = \sum_{j=1}^{n} p_{ij}x'_j$. Notice that the rows of P are used to express the old coordinates of ξ in terms of the new coordinates. For emphasis and contradistinction, we repeat that the columns of P are used to express the new basis vectors in terms of the old basis vectors.

Let $X = (x_1, \ldots, x_n)$ and $X' = (x'_1, \ldots, x'_n)$ be $n \times 1$ matrices representing the vector ξ with respect to the bases A and A'. Then the set of relations $\{x_i = \sum_{j=1}^{n} p_{ij}x'_j\}$ can be written as the single matric equation

$$X = PX'. \tag{4.3}$$

Now suppose that we have a linear transformation σ of U into V and that $A = [a_{ij}]$ is the matrix representing σ with respect to the bases A in U and $\mathsf{B} = \{\beta_1, \ldots, \beta_m\}$ in V. We shall now determine the representation of σ with respect to the bases A' and B.

$$\begin{aligned}\sigma(\alpha'_j) &= \sum_{k=1}^{n} p_{kj}\sigma(\alpha_k) = \sum_{k=1}^{n} p_{kj}\left(\sum_{i=1}^{m} a_{ik}\beta_i\right) \\ &= \sum_{i=1}^{m}\left(\sum_{k=1}^{n} a_{ik}p_{kj}\right)\beta_i \\ &= \sum_{i=1}^{m} a'_{ij}\beta_i. \end{aligned} \tag{4.4}$$

Since B is a basis, $a'_{ij} = \sum_{k=1}^{n} a_{ik}p_{kj}$ and the matrix $A' = [a'_{ij}]$ representing σ with respect to the bases A' and B is related to A by the matric equation

$$A' = AP. \tag{4.5}$$

This relation can also be demonstrated in a slightly different way. For an arbitrary $\xi = \sum_{j=1}^{n} x_j\alpha_j \in U$ let $\sigma(\xi) = \sum_{i=1}^{m} y_i\beta_i$. Then we have

$$Y = AX = A(PX') = (AP)X'. \tag{4.6}$$

Thus AP is a matrix representing σ with respect to the bases A' and B. Since the matrix representing σ is uniquely determined by the choice of bases we have $A' = AP$.

Now consider the effect of a change of basis in the image space V. Thus let B be replaced by the basis $\mathsf{B}' = \{\beta'_1, \ldots, \beta'_m\}$. Let $Q = [q_{ij}]$ be the matrix of transition from B to B', that is, $\beta'_j = \sum_{i=1}^{m} q_{ij}\beta_i$. Then if $A'' = [a''_{ij}]$ represents σ with respect to the bases A and B' we have

$$\begin{aligned}\sigma(\alpha_j) &= \sum_{k=1}^{m} a''_{kj}\beta'_k = \sum_{k=1}^{m} a''_{kj}\left(\sum_{i=1}^{m} q_{ik}\beta_i\right) \\ &= \sum_{i=1}^{m}\left(\sum_{k=1}^{m} q_{ik}a''_{kj}\right)\beta_i = \sum_{i=1}^{m} a_{ij}\beta_i. \end{aligned} \tag{4.7}$$

Since the representation of $\sigma(\alpha_j)$ in terms of the basis B is unique we see that $A = QA''$, or

$$A'' = Q^{-1}A. \tag{4.8}$$

Combining these results, we see that, if both changes of bases are made at once, the new matrix representing σ is $Q^{-1}AP$.

As in the proof of Theorem 1.6 we can choose a new basis $\mathsf{A}' = \{\alpha'_1, \ldots, \alpha'_n\}$ of U such that the last $\nu = n - \rho$ basis elements form a basis of $K(\sigma)$. Since $\{\sigma(\alpha'_1), \ldots, \sigma(\alpha'_\rho)\}$ is a basis of $\sigma(U)$ and is linearly independent in V, it can

be extended to a basis B' of V. With respect to the bases A' and B' we have $\sigma(\alpha'_j) = \beta'_j$ for $j \leq \rho$ while $\sigma(\alpha'_j) = 0$ for $j > \rho$. Thus the new matrix $Q^{-1}AP$ representing σ is of the form

$$\begin{array}{r} \\ \rho \text{ rows} \\ \\ \\ \\ \\ m - \rho \text{ rows} \end{array} \begin{array}{c} \rho \text{ columns} \\ \left[\begin{array}{ccccc|c} 1 & 0 & 0 & \cdots & 0 & \\ 0 & 1 & 0 & \cdots & 0 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & & \\ 0 & 0 & 0 & & 1 & \\ \hline & & 0 & & & 0 \end{array}\right] \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \end{array} \nu \text{ columns}$$

Thus we have

Theorem 4.1. *If A is any $m \times n$ matrix of rank ρ, there exist a non-singular $n \times n$ matrix P and a non-singular $m \times m$ matrix Q such that $A' = Q^{-1}AP$ has the first ρ elements of the main diagonal equal to* 1, *and all other elements equal to zero.* □

When A and B are unrestricted we can always obtain this relatively simple representation of a linear transformation by a proper choice of bases. More interesting situations occur when A and B are restricted. Suppose, for example, that we take $U = V$ and $A = B$. In this case there is but one basis to change and but one matrix of transition, that is, $P = Q$. In this case it is not possible to obtain a form of the matrix representing σ as simple as that obtained in Theorem 4.1. We say that any two matrices representing the same linear transformation σ of a vector space V into itself are *similar*. This is equivalent to saying that two matrices A and A' are similar if and only if there exists a non-singular matrix of transition P such that $A' = P^{-1}AP$. This case occupies much of our attention in Chapters III and V.

EXERCISES

1. In P_3, the space of polynomials of degree 2 or smaller with coefficients in F, let $A = \{1, x, x^2\}$.

$$A' = \{p_1(x) = x^2 + x + 1, p_2(x) = x^2 - x - 2, p_3(x) = x^2 + x - 1\}$$

is also a basis. Find the matrix of transition from A to A'.

2. In many of the uses of the concepts of this section it is customary to take $\mathsf{A} = \{\alpha_i \mid \alpha_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{in})\}$ as the old basis in R^n. Thus, in R^2 let $\mathsf{A} = \{(1, 0), (0, 1)\}$ and $\mathsf{A}' = \{(\frac{1}{2}, \sqrt{3}/2), (-\sqrt{3}/2, \frac{1}{2})\}$. Show that

$$P = \begin{bmatrix} \frac{1}{2} & -\sqrt{3}/2 \\ \sqrt{3}/2 & \frac{1}{2} \end{bmatrix}$$

is the matrix of transition from A to A'.

3. (Continuation) With A' and A as in Exercise 2, find the matrix of transition R from A' to A. (Notice, in particular, that in Exercise 2 the columns of P are the components of the vectors in A' expressed in terms of basis A, whereas in this exercise the columns of R are the components of the vectors in A expressed in terms of the basis A'. Thus these two matrices of transition are determined relative to different bases.) Show that $RP = I$.

4. (Continuation) Consider the linear transformation of σ of R^2 into itself which maps

$$\begin{array}{lll} (1, 0) & \text{onto} & (\frac{1}{2}, \sqrt{3}/2) \\ (0, 1) & \text{onto} & (-\sqrt{3}/2, \frac{1}{2}). \end{array}$$

Find the matrix A that represents σ with respect to the basis A.

You should obtain $A = P$. However, A and P do not represent the same thing. To see this, let $\xi = (x_1, x_2)$ be an arbitrary vector in R^2 and compute $\sigma(\xi)$ by means of formula (2.9) and the new coordinates of ξ by means of formula (4.3).

A little reflection will show that the results obtained are entirely reasonable. The matrix A represents a rotation of the real plane counterclockwise through an angle of $\pi/3$. The matrix P represents a rotation of the coordinate axes counterclockwise through an angle of $\pi/3$. In the latter case the motion of the plane relative to the coordinate axes is clockwise through an angle of $\pi/3$.

5. In R^3 let $\mathsf{A} = \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ and let $\mathsf{A}' = \{(0, 1, 1), (1, 0, 1), (1, 1, 0)\}$. Find the matrix of transition P from A to A' and the matrix of transition P^{-1} from A' to A.

6. Let A, B, and C be three bases of V. Let P be the matrix of transition from A to B and let Q be the matrix of transition from B to C. Is PQ or QP the matrix of transition from A to C? Compare the order of multiplication of matrices of transition and matrices representing linear transformation.

7. Use the results of Exercise 6 to resolve the question raised in the parenthetical remark of Exercise 3, and implicitly assumed in Exercise 5. If P is the matrix of transition from A to A' and Q is the matrix of transition from A' to A, show that $PQ = I$.

5 | Hermite Normal Form

We may also ask how much simplification of the matrix representing a linear transformation σ of U into V can be effected by a change of basis in

V alone. Let $A = \{\alpha_1, \ldots, \alpha_n\}$ be the given basis in U and let $U_k = \langle \alpha_1, \ldots, \alpha_k \rangle$. The subspaces $\sigma(U_k)$ of V form a non-decreasing chain of subspaces with $\sigma(U_{k-1}) \subset \sigma(U_k)$ and $\sigma(U_n) = \sigma(U)$. Since $\sigma(U_k) = \sigma(U_{k-1}) + \langle \sigma(\alpha_k) \rangle$ we see from Theorem 4.8 of Chapter I that $\dim \sigma(U_k) \leq \dim \sigma(U_{k-1}) + 1$; that is, the dimensions of the $\sigma(U_k)$ do not increase by more than 1 at a time as k increases. Since $\dim \sigma(U_n) = \rho$, the rank of σ, an increase of exactly 1 must occur ρ times. For the other times, if any, we must have $\dim \sigma(U_k) = \dim \sigma(U_{k-1})$ and hence $\sigma(U_k) = \sigma(U_{k-1})$. We have an increase by 1 when $\sigma(\alpha_k) \notin \sigma(U_{k-1})$ and no increase when $\sigma(\alpha_k) \in \sigma(U_{k-1})$.

Let $k_1, k_2, \ldots, k_\rho$ be those indices for which $\sigma(\alpha_{k_i}) \notin \sigma(U_{k_i-1})$. Let $\beta'_i = \sigma(\alpha_{k_i})$. Since $\beta'_i \notin \sigma(U_{k_i-1}) = \langle \beta'_1, \ldots, \beta'_{i-1} \rangle$, the set $\{\beta'_1, \ldots, \beta'_\rho\}$ is linearly independent (see Theorm 2.3, Chapter I-2). Since $\{\beta'_1, \ldots, \beta'_\rho\} \subset \sigma(U)$ and $\sigma(U)$ is of dimension ρ, $\{\beta'_1, \ldots, \beta'_\rho\}$ is a basis of $\sigma(U)$. This set can be extended to a basis B' of V. Let us now determine the form of the matrix A' representing σ with respect to the bases A and B'.

Since $\sigma(\alpha_{k_i}) = \beta'_i$, column k_i has a 1 in row i and all other elements of this column are 0's. For $k_i < j < k_{i+1}$, $\sigma(\alpha_j) \in \sigma(U_{k_i})$ so that column j has 0's below row i. In general, there is no restriction on the elements of column j in the first i rows. A' thus has the form

$$\begin{array}{cccccccccc} & & \text{column} & & & & \text{column} & & & \\ & & k_1 & & & & k_2 & & & \end{array}$$

$$\begin{bmatrix} 0 & \cdots & 0 & 1 & a'_{1,k_1+1} & \cdots & 0 & a'_{1,k_2+1} & \cdots & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & a'_{2,k_2+1} & \cdots & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \\ \cdot & & \cdot & \cdot & \cdot & & \cdot & \cdot & & \\ \cdot & & \cdot & \cdot & \cdot & & \cdot & \cdot & & \\ \cdot & & \cdot & \cdot & \cdot & & \cdot & \cdot & & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & \end{bmatrix}. \tag{5.1}$$

Once A and σ are given, the k_i and the set $\{\beta'_1, \ldots, \beta'_\rho\}$ are uniquely determined. There may be many ways to extend this set to the basis B', but the additional basis vectors do not affect the determination of A' since every element of $\sigma(U)$ can be expressed in terms of $\{\beta'_1, \ldots, \beta'_\rho\}$ alone. Thus A' is uniquely determined by A and σ.

Theorem 5.1. *Given any $m \times n$ matrix A of rank ρ, there exists a non-singular $m \times m$ matrix Q such that $A' = Q^{-1}A$ has the following form:*

(1) *There is at least one non-zero element in each of the first ρ rows of A', and the elements in all remaining rows are zero.*

(2) *The first non-zero element appearing in row i ($i \leq \rho$) is a* 1 *appearing in column k_i, where $k_1 < k_2 < \cdots < k_\rho$.*

(3) *In column k_i the only non-zero element is the* 1 *in row i.*

The form A' is uniquely determined by A.

PROOF. In the applications of this theorem that we wish to make A is usually given alone without reference to any bases A and B, and often without reference to any linear transformation σ. We can, however, introduce any two vector spaces U and V of dimensions n and m over F and let A be any basis of U and B be any basis of V. We can consider A as defining a linear transformation σ of U into V with respect to the bases A and B. The discussion preceding Theorem 5.1 shows that there is at least one non-singular matrix Q such that $Q^{-1}A$ satisfies conditions (1), (2), and (3).

Now suppose there are two non-singular matrices Q_1 and Q_2 such that $Q_1^{-1}A = A_1'$ and $Q_2^{-1}A = A_2'$ both satisfy the conditions of the theorem. We wish to conclude that $A_1' = A_2'$. No matter how the vector spaces U and V are introduced and how the bases A and B are chosen we can regard Q_1 and Q_2 as matrices of transition in V. Thus A_1' represents σ with respect to bases A and B_1' and A_2' represents σ with respect to bases A and B_2'. But condition (3) says that for $i \leq \rho$ the ith basis element in both B_1' and B_2' is $\sigma(\alpha_{k_i})$. Thus the first ρ elements of B_1' and B_2' are identical. Condition (1) says that the remaining basis elements have nothing to do with determining the coefficients in A_1' and A_2'. Thus $A_1' = A_2'$. □

We say that a matrix satisfying the conditions of Theorem 5.1 is in *Hermite normal form.* Often this form is called a *row-echelon form.* And sometimes the term, Hermite normal form, is reserved for a square matrix containing exactly the numbers that appear in the form we obtained in Theorem 5.1 with the change that row i beginning with a 1 in column k_i is moved down to row k_i. Thus each non-zero row begins on the main diagonal and each column with a 1 on the main diagonal is otherwise zero. In this text we have no particular need for this special form while the form described in Theorem 5.1 is one of the most useful tools at our disposal.

The usefulness of the Hermite normal form depends on its form, and the uniqueness of that form will enable us to develop effective and convenient short cuts for determining that form.

Definition. Given the matrix A, the matrix A^T obtained from A by interchanging rows and columns in A is called the *transpose* of A. If $A^T = [a_{ij}']$, the element a_{ij}' appearing in row i column j of A^T is the element a_{ji} appearing in row j column i of A. It is easy to show that $(AB)^T = B^T A^T$. (See Exercise 4.)

Proposition 5.2. *The number of linearly independent rows in a matrix is equal to the number of linearly independent columns.*

PROOF. The number of linearly independent columns in a matrix A is its rank ρ. The Hermite normal form $A' = Q^{-1}A$ corresponding to A is also of rank ρ. For A' it is obvious that the number of linearly independent rows in A' is also equal to ρ, that is, the rank of $(A')^T$ is ρ. Since Q^T is non-singular, the rank of $A^T = (QA')^T = (A')^T Q^T$ is also ρ. Thus the number of linearly independent rows in A is ρ. □

EXERCISES

1. Which of the following matrices are in Hermite normal form?

(*a*) $\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$,

(*b*) $\begin{bmatrix} 0 & 0 & 2 & 0 & 4 \\ 0 & 1 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$,

(*c*) $\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$,

(*d*) $\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$,

(*e*) $\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.

2. Determine the rank of each of the matrices given in Exercise 1.

3. Let σ and τ be linear transformations mapping R^3 into R^2. Suppose that for a given pair of bases A for R^3 and B for R^2, σ and τ are represented by

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

respectively. Show that there is no basis B' of R^2 such that B is the matrix representing σ with respect to A and B'.

4. Show that
   (*a*) $(A + B)^T = A^T + B^T$,
   (*b*) $(AB)^T = B^T A^T$,
   (*c*) $(A^{-1})^T = (A^T)^{-1}$.

6 | Elementary Operations and Elementary Matrices

Our purpose in this section is to develop convenient computational methods. We have been concerned with the representations of linear transformations by matrices and the changes these matrices undergo when a basis is changed. We now show that these changes can be effected by elementary operations on the rows and columns of the matrices.

We define three types of *elementary operations* on the rows of a matrix A.

Type I: Multiply a row of A by a non-zero scalar.
Type II: Add a multiple of one row to another row.
Type III: Interchange two rows.

Elementary column operations are defined in an analogous way.

From a logical point of view these operations are redundant. An operation of type III can be accomplished by a combination of operations of types I and II. It would, however, require four such operations to take the place of one operation of type III. Since we wish to develop convenient computational methods, it would not suit our purpose to reduce the number of operations at our disposal. On the other hand, it would not be of much help to extend the list of operations at this point. The student will find that, with practice, he can combine several elementary operations into one step. For example, such a combined operation would be the replacing of a row by a linear combination of rows, provided that the row replaced appeared in the linear combination with a non-zero coefficient. We leave such short cuts to the student.

An elementary operation can also be accomplished by multiplying A on the left by a matrix. Thus, for example, multiplying the second row by the scalar c can be effected by the matrix

$$E_2(c) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & c & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{6.1}$$

The addition of k times the third row to the first row can be effected by the matrix

$$E_{31}(k) = \begin{bmatrix} 1 & 0 & k & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{6.2}$$

The interchange of the first and second rows can be effected by the matrix

$$E_{12} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{6.3}$$

These matrices corresponding to the elementary operations are called *elementary matrices.* These matrices are all non-singular and their inverses are also elementary matrices. For example, the inverses of $E_2(c)$, $E_{31}(k)$, and E_{12} are respectively $E_2(c^{-1})$, $E_{31}(-k)$, and E_{12}.

Notice that the elementary matrix representing an elementary operation is the matrix obtained by applying the elementary operation to the unit matrix.

Theorem 6.1. *Any non-singular matrix A can be written as a product of elementary matrices.*

PROOF. At least one element in the first column is non-zero or else A would be singular. Our first goal is to apply elementary operations, if necessary, to obtain a 1 in the upper left-hand corner. If $a_{11} = 0$, we can interchange rows to bring a non-zero element into that position. Thus we may as well suppose that $a_{11} \neq 0$. We can then multiply the first row by a_{11}^{-1}. Thus, to simplify notation, we may as well assume that $a_{11} = 1$. We now add $-a_{i1}$ times the first row to the ith row to make every other element in the first column equal to zero.

The resulting matrix is still non-singular since the elementary operations applied were non-singular. We now wish to obtain a 1 in the position of element a_{22}. At least one element in the second column other than a_{12}

is non-zero for otherwise the first two columns would be dependent. Thus by a possible interchange of rows, not including row 1, and multiplying the second row by a non-zero scalar we can obtain $a_{22} = 1$. We now add $-a_{i2}$ times the second row to the ith row to make every other element in the second column equal to zero. Notice that we also obtain a 0 in the position of a_{12} without affecting the 1 in the upper left-hand corner.

We continue in this way until we obtain the identity matrix. Thus if $E_1, E_2, \ldots, E_r$ are elementary matrices representing the successive elementary operations, we have

$$I = E_r \cdots E_2 E_1 A, \tag{6.4}$$

or

$$A = E_1^{-1} E_2^{-1} \cdots E_r^{-1}. \quad \square$$

In Theorem 5.1 we obtained the Hermite normal form A' from the matrix A by multiplying on the left by the non-singular matrix Q^{-1}. We see now that Q^{-1} is a product of elementary matrices, and therefore that A can be transformed into Hermite normal form by a succession of elementary row operations. It is most efficient to use the elementary row operations directly without obtaining the matrix Q^{-1}.

We could have shown directly that a matrix could be transformed into Hermite normal form by means of elementary row operations. We would then be faced with the necessity of showing that the Hermite normal form obtained is unique and not dependent on the particular sequence of operations used. While this is not particularly difficult, the demonstration is uninteresting and unilluminating and so tedious that it is usually left as an "exercise for the reader." Uniqueness, however, is a part of Theorem 5.1, and we are assured that the Hermite normal form will be independent of the particular sequence of operations chosen. This is important as many possible operations are available at each step of the work, and we are free to choose those that are most convenient.

Basically, the instructions for reducing a matrix to Hermite normal form are contained in the proof of Theorem 6.1. In that theorem, however, we were dealing with a non-singular matrix and thus assured that we could at certain steps obtain a non-zero element on the main diagonal. For a singular matrix, this is not the case. When a non-zero element cannot be obtained with the instructions given we must move our consideration to the next column.

In the following example we perform several operations at each step to conserve space. When several operations are performed at once, some care must be exercised to avoid reducing the rank. This may occur, for example, if we subtract a row from itself in some hidden fashion. In this example we avoid this pitfall, which can occur when several operations of

type III are combined, by considering one row as an operator row and adding multiples of it to several others.

Consider the matrix

$$\begin{bmatrix} 4 & 3 & 2 & -1 & 4 \\ 5 & 4 & 3 & -1 & 4 \\ -2 & -2 & -1 & 2 & -3 \\ 11 & 6 & 4 & 1 & 11 \end{bmatrix}$$

as an example.

According to the instructions for performing the elementary row operations we should multiply the first row by $\frac{1}{4}$. To illustrate another possible way to obtain the "1" in the upper left corner, multiply row 1 by -1 and add row 2 to row 1. Multiples of row 1 can now be added to the other rows to obtain

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 4 \\ 0 & 0 & 1 & 2 & -3 \\ 0 & -5 & -7 & 1 & 11 \end{bmatrix}.$$

Now, multiply row 2 by -1 and add appropriate multiples to the other rows to obtain

$$\begin{bmatrix} 1 & 0 & -1 & -1 & 4 \\ 0 & 1 & 2 & 1 & -4 \\ 0 & 0 & 1 & 2 & -3 \\ 0 & 0 & 3 & 6 & -9 \end{bmatrix}.$$

Finally, we obtain

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & -3 & 2 \\ 0 & 0 & 1 & 2 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

which is the Hermite normal form described in Theorem 5.1. If desired, Q^{-1} can be obtained by applying the same sequence of elementary row operations to the unit matrix. However, while the Hermite normal form is necessarily unique, the matrix Q^{-1} need not be unique, as the proof of Theorem 5.1 should show.

Rather than trying to remember the sequence of elementary operations used to reduce A to Hermite normal form, it is more efficient to perform these operations on the unit matrix at the same time we are operating on A. It is suggested that we arrange the work in the following way:

$$\begin{bmatrix} 4 & 3 & 2 & -1 & 4 & 1 & 0 & 0 & 0 \\ 5 & 4 & 3 & -1 & 4 & 0 & 1 & 0 & 0 \\ -2 & -2 & -1 & 2 & -3 & 0 & 0 & 1 & 0 \\ 11 & 6 & 4 & 1 & 11 & 0 & 0 & 0 & 1 \end{bmatrix} = [A, I]$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 4 & 5 & -4 & 0 & 0 \\ 0 & 0 & 1 & 2 & -3 & -2 & 2 & 1 & 0 \\ 0 & -5 & -7 & 1 & 11 & 11 & -11 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & -1 & -1 & 4 & 4 & -3 & 0 & 0 \\ 0 & 1 & 2 & 1 & -4 & -5 & 4 & 0 & 0 \\ 0 & 0 & 1 & 2 & -3 & -2 & 2 & 1 & 0 \\ 0 & 0 & 3 & 6 & -9 & -14 & 9 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 2 & -1 & 1 & 0 \\ 0 & 1 & 0 & -3 & 2 & -1 & 0 & -2 & 0 \\ 0 & 0 & 1 & 2 & -3 & -2 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -8 & 3 & -3 & 1 \end{bmatrix}.$$

In the end we obtain

$$Q^{-1} = \begin{bmatrix} 2 & -1 & 1 & 0 \\ -1 & 0 & -2 & 0 \\ -2 & 2 & 1 & 0 \\ -8 & 3 & -3 & 1 \end{bmatrix},$$

Verify directly that $Q^{-1}A$ is in Hermite normal form.

If A were non-singular, the Hermite normal form obtained would be the identity matrix. In this case Q^{-1} would be the inverse of A. This method of finding the inverse of a matrix is one of the easiest available for hand computation. It is the recommended technique.

EXERCISES

1. Elementary operations provide the easiest methods for determining the rank of a matrix. Proceed as if reducing to Hermite normal form. Actually, it is not necessary to carry out all the steps as the rank is usually evident long before the Hermite normal form is obtained. Find the ranks of the following matrices:

(*a*) $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$,

(*b*) $\begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 3 \\ -2 & -3 & 0 \end{bmatrix}$,

(*c*) $\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \end{bmatrix}$.

2. Identify the elementary operations represented by the following elementary matrices:

(*a*) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}$,

(*b*) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$,

(*c*) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

3. Show that the product

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

is an elementary matrix. Identify the elementary operations represented by each matrix in the product.

4. Show by an example that the product of elementary matrices is not necessarily an elementary matrix.

5. Reduce each of the following matrices to Hermite normal form.

(*a*) $\begin{bmatrix} 2 & 1 & 3 & -2 \\ 2 & -1 & 5 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}$,

(*b*) $\begin{bmatrix} 1 & 2 & 3 & 3 & 10 & 6 \\ 2 & 1 & 0 & 0 & 2 & 3 \\ 2 & 2 & 2 & 1 & 5 & 5 \\ -1 & 1 & 3 & 2 & 5 & 2 \end{bmatrix}$.

6. Use elementary row operations to obtain the inverses of

(*a*) $\begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix}$, and

(*b*) $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 6 \end{bmatrix}$.

7. (*a*) Show that, by using a sequence of elementary operations of type II only, any two rows of a matrix can be interchanged with one of the two rows multiplied by -1. (In fact, the type II operations involve no scalars other than ± 1.)

(*b*) Using the results of part (*a*), show that a type III operation can be obtained by a sequence of type II operations and a single type I operation.

(*c*) Show that the sign of any row can be changed by a sequence of type II operations and a single type III operation.

8. Show that any matrix A can be reduced to the form described in Theorem 4.1 by a sequence of elementary row operations and a sequence of elementary column operations.

7 | Linear Problems and Linear Equations

For a given linear transformation σ of U into V and a given $\beta \in V$ the problem of finding any or all $\xi \in U$ for which $\sigma(\xi) = \beta$ is called a *linear problem.* Before providing any specific methods for solving such problems, let us see what the set of solutions should look like.

If $\beta \notin \sigma(U)$, then the problem has no solution.

If $\beta \in \sigma(U)$, the problem has at least one solution. Let ξ_0 be one such solution. We call any such ξ_0 a *particular solution.* If ξ is any other solution, then $\sigma(\xi - \xi_0) = \sigma(\xi) - \sigma(\xi_0) = \beta - \beta = 0$ so that $\xi - \xi_0$ is in the kernel of σ. Conversely, if $\xi - \xi_0$ is in the kernel of σ then $\sigma(\xi) = \sigma(\xi_0 + \xi - \xi_0) = \sigma(\xi_0) + \sigma(\xi - \xi_0) = \beta + 0 = \beta$ so that ξ is a solution. Thus the set of all solutions of $\sigma(\xi) = \beta$ is of the form

$$\{\xi_0\} + K(\sigma). \tag{7.1}$$

Since $\{\xi_0\}$ contains just one element, there is a one-to-one correspondence between the elements of $K(\sigma)$ and the elements of $\{\xi_0\} + K(\sigma)$. Thus the size of the set of solutions can be described by giving the dimension of $K(\sigma)$. The set of all solutions of the problem $\sigma(\xi) = \beta$ is not a subspace of U unless $\beta = 0$. Nevertheless, it is convenient to say that the set is of dimension ν, the nullity of σ.

Given the linear problem $\sigma(\xi) = \beta$, the problem $\sigma(\xi) = 0$ is called the *associated homogeneous problem.* The *general solution* is then any particular solution plus the solution of the associated homogeneous problem. The solution of the associated homogeneous problem is the kernel of σ.

Now let σ be represented by the $m \times n$ matrix $A = [a_{ij}]$, β be represented by $B = (b_1, \ldots, b_m)$, and ξ by $X = (x_1, \ldots, x_n)$. Then the linear problem $\sigma(\xi) = \beta$ becomes

$$AX = B \tag{7.2}$$

in matrix form, or

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \qquad (i = 1, \ldots, m) \tag{7.3}$$

in the form of a system of linear equations.

Given A and B, the *augmented matrix* $[A, B]$ of the system of linear equations is defined to be

$$[A, B] = \begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{bmatrix}. \tag{7.4}$$

Theorem 7.1. *The system of simultaneous linear equations $AX = B$ has a solution if and only if the rank of A is equal to the rank of the augmented matrix $[A, B]$. Whenever a solution exists, all solutions can be expressed in terms of $\nu = n - \rho$ independent parameters, where ρ is the rank of A.*

PROOF. We have already seen that the linear problem $\sigma(\xi) = \beta$ has a solution if and only if $\beta \in \sigma(U)$. This is the case if and only if β is linearly dependent on $\{\sigma(\alpha_1), \ldots, \sigma(\alpha_n)\}$. But this is equivalent to the condition that B be linearly dependent on the columns of A. Thus adjoining the column of b_i's to form the augmented matrix must not increase the rank. Since the rank of the augmented matrix cannot be less than the rank of A we see that the system has a solution if and only if these two ranks are equal.

Now let Q be a non-singular matrix such that $Q^{-1}A = A'$ is in Hermite normal form. Any solution of $AX = B$ is also a solution of $A'X = Q^{-1}AX = Q^{-1}B = B'$. Conversely, any solution of $A'X = B'$ is also a solution of $AX = QA'X = QB' = B$. Thus the two systems of equations are equivalent.

Now the system $A'X = B'$ is particularly easy to solve since the variable x_{k_i} appears only in the ith equation. Furthermore, non-zero coefficients appear only in the first ρ equations. The condition that $\beta \in \sigma(U)$ also takes on a form that is easily recognizable. The condition that B' be expressible as a linear combination of the columns of A' is simply that the elements of B' below row ρ be zero. The system $A'X = B'$ has the form

$$\begin{aligned} x_{k_1} + a'_{1,k_1+1}x_{k_1+1} + \cdots 0 + a'_{1,k_2+1}x_{k_2+1} + \cdots &= b'_1 \\ x_{k_2} + a'_{2,k_2+1}x_{k_2+1} &= b'_2 \\ \cdots & \end{aligned} \tag{7.5}$$

Since each x_{k_i} appears in but one equation with unit coefficient, the remaining $n - \rho$ unknowns can be given values arbitrarily and the corresponding values of the x_{k_i} computed. The $n - \rho$ unknowns with indices not the k_i are the $n - \rho$ parameters mentioned in the theorem. □

As an example, consider the system of equations:

$$\begin{aligned} 4x_1 + 3x_2 + 2x_3 - x_4 &= 4 \\ 5x_1 + 4x_2 + 3x_3 - x_4 &= 4 \\ -2x_1 - 2x_2 - x_3 + 2x_4 &= -3 \\ 11x_1 + 6x_2 + 4x_3 + x_4 &= 11. \end{aligned}$$

The augmented matrix is

$$\begin{bmatrix} 4 & 3 & 2 & -1 & 4 \\ 5 & 4 & 3 & -1 & 4 \\ -2 & -2 & -1 & 2 & -3 \\ 11 & 6 & 4 & 1 & 11 \end{bmatrix}.$$

This is the matrix we chose for an example in the previous section. There we obtained the Hermite normal form

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & -3 & 2 \\ 0 & 0 & 1 & 2 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus the system of equations $A'X = B'$ corresponding to this augmented matrix is

$$\begin{aligned} x_1 + x_4 &= 1 \\ x_2 - 3x_4 &= 2 \\ x_3 + 2x_4 &= -3. \end{aligned}$$

It is clear that this system is very easy to solve. We can take any value whatever for x_4 and compute the corresponding values for x_1, x_2, and x_3. A particular solution, obtained by taking $x_4 = 0$, is $X_0 = (1, 2, -3, 0)$. It is more instructive to write the new system of equations in the form

$$\begin{aligned} x_1 &= 1 - x_4 \\ x_2 &= 2 + 3x_4 \\ x_3 &= -3 - 2x_4 \\ x_4 &= x_4 \end{aligned}$$

In vector form this becomes

$$(x_1, x_2, x_3, x_4) = (1, 2, -3, 0) + x_4(-1, 3, -2, 1).$$

We can easily verify that $(-1, 3, -2, 1)$ is a solution of the associated homogeneous problem. In fact, $\{(-1, 3, -2, 1)\}$ is a basis for the kernel, and $x_4(-1, 3, -2, 1)$, for an arbitrary x_4, is a general element of the kernel. We have, therefore, expressed the general solution as a particular solution plus the kernel.

The elementary row operations provide us with the recommended technique for solving simultaneous linear equations by hand. This application is the principal reason for introducing elementary row operations rather than column operations.

Theorem 7.2. *The equation $AX = B$ fails to have a solution if and only if there exists a one-row matrix C such that $CA = 0$ and $CB = 1$.*

PROOF. Suppose the equation $AX = B$ has a solution and a C exists such that $CA = 0$ and $CB = 1$. Then we would have $0 = (CA)X = C(AX) = CB = 1$, which is a contradiction.

On the other hand, suppose the equation $AX = B$ has no solution. By Theorem 7.1 this implies that the rank of the augmented matrix $[A, B]$ is greater than the rank of A. Let Q be a non-singular matrix such that $Q^{-1}[A, B]$ is in Hermite normal form. Then if ρ is the rank of A, the $(\rho + 1)$st row of $Q^{-1}[A, B]$ must be all zeros except for a 1 in the last column. If C is the $(\rho + 1)$st row of Q^{-1} this means that

$$C[A, B] = [0 \quad 0 \quad \cdots \quad 0 \quad 1],$$

or

$$CA = 0 \qquad \text{and} \qquad CB = 1. \ \square$$

This theorem is important because it provides a positive condition for a negative conclusion. Theorem 7.1 also provides such a positive condition and it is to be preferred when dealing with a particular system of equations. But Theorem 7.2 provides a more convenient condition when dealing with systems of equations in general.

Although the sytems of linear equations in the exercises that follow are written in expanded form, they are equivalent in form to the matric equation

$AX = B$. From any linear problem in this set, or those that will occur later, it is possible to obtain an extensive list of closely related linear problems that appear to be different. For example, if $AX = B$ is the given linear problem with A an $m \times n$ matrix and Q is any non-singular $m \times m$ matrix, then $A'X = B'$ with $A' = QA$ and $B' = QB$ is a problem with the same set of solutions. If P is a non-singular $n \times n$ matrix, then $A''X'' = B$ where $A'' = AP$ is a problem whose solution X'' is related to the solution X of the original problem by the condition $X'' = P^{-1}X$.

For the purpose of constructing related exercises of the type mentioned, it is desirable to use matrices P and Q that do not introduce tedious numerical calculations. It is very easy to obtain a non-singular matrix P that has only integral elements and such that its inverse also has only integral elements. Start with an identity matrix of the desired order and perform a sequence of elementary operations of types II and III. As long as an operation of type I is avoided, no fractions will be introduced. Furthermore, the inverse operations will be of types II and III so the inverse matrix will also have only integral elements.

For convenience, some matrices with integral elements and inverses with integral elements are listed in an appendix. For some of the exercises that are given later in this book, matrices of transition that satisfy special conditions are also needed. These matrices, known as orthogonal and unitary matrices, usually do not have integral elements. Simple matrices of these types are somewhat harder to obtain. Some matrices of these types are also listed in the appendix.

EXERCISES

1. Show that $\{(1, 1, 1, 0), (2, 1, 0, 1)\}$ spans the subspace of all solutions of the system of linear equations

$$\begin{aligned} 3x_1 - 2x_2 - x_3 - 4x_4 &= 0 \\ x_1 + x_2 - 2x_3 - 3x_4 &= 0. \end{aligned}$$

2. Find the subspace of all solutions of the system of linear equations

$$\begin{aligned} x_1 + 2x_2 - 3x_3 + x_4 &= 0 \\ 3x_1 - x_2 + 5x_3 - x_4 &= 0 \\ 2x_1 + x_2 \qquad\quad + x_4 &= 0. \end{aligned}$$

3. Find all solutions of the following two systems of non-homogeneous linear equations.

(*a*)
$$\begin{aligned} x_1 + 3x_2 + 5x_3 - 2x_4 &= 11 \\ 3x_1 - 2x_2 - 7x_3 + 5x_4 &= 0 \\ 2x_1 + x_2 \qquad + x_4 &= 7, \end{aligned}$$

(*b*)
$$\begin{aligned} x_1 + 3x_2 + 2x_3 + 5x_4 &= 10 \\ 3x_1 - 2x_2 - 5x_3 + 4x_4 &= -5 \\ 2x_1 + x_2 - x_3 + 5x_4 &= 5. \end{aligned}$$

4. Find all solutions of the following system of non-homogeneous linear equations

$$\begin{aligned} 2x_1 - x_2 - 3x_3 &= 1 \\ x_1 - x_2 + 2x_3 &= -2 \\ 4x_1 - 3x_2 + x_3 &= -3 \\ x_1 - 5x_3 &= 3. \end{aligned}$$

5. Find all solutions of the system of equations,

$$\begin{aligned} 7x_1 + 3x_2 + 21x_3 - 13x_4 + x_5 &= -14 \\ 10x_1 + 3x_2 + 30x_3 - 16x_4 + x_5 &= -23 \\ 7x_1 + 2x_2 + 21x_3 - 11x_4 + x_5 &= -16 \\ 9x_1 + 3x_2 + 27x_3 - 15x_4 + x_5 &= -20. \end{aligned}$$

6. Theorem 7.1 states that a necessary and sufficient condition for the existence of a solution of a system of simultaneous linear equations is that the rank of the augmented matrix be equal to the rank of the coefficient matrix. The most efficient way to determine the rank of each of these matrices is to reduce each to Hermite normal form. The reduction of the augmented matrix to normal form, however, automatically produces the reduced form of the coefficient matrix. How, and where? How is the comparison of the ranks of the coefficient matrix and the augmented matrix evident from the appearance of the reduced form of the augmented matrix?

7. The differential equation $d^2y/dx^2 + 4y = \sin x$ has the general solution $y = C_1 \sin 2x + C_2 \cos 2x + \frac{1}{3} \sin x$. Identify the associated homogeneous problem, the solution of the associated homogeneous problem, and the particular solution.

8 | Other Applications of the Hermite Normal Form

The Hermite normal form and the elementary row operations provide techniques for dealing with problems we have already encountered and handled rather awkwardly.

A Standard Basis for a Subspace

Let $\mathsf{A} = \{\alpha_1, \ldots, \alpha_n\}$ be a basis of U and let W be a subspace of U spanned by the set $\mathsf{B} = \{\beta_1, \ldots, \beta_r\}$. Since every subspace of U is spanned by a finite set, it is no restriction to assume that B is finite. Let $\beta_i = \sum_{j=1}^{n} b_{ij}\alpha_j$ so that $(b_{i1}, \ldots, b_{in})$ is the n-tuple representing β_i. Then in the matrix $B = [b_{ij}]$ each row is the representation of a vector in B. Now suppose an elementary row operation is applied to B to obtain B'. Every row of B' is a linear combination of the rows of B and, since an elementary row operation has an inverse, every row of B is a linear combination of the rows of B'. Thus the rows of B and the rows of B' represent sets spanning the same subspace W. We can therefore reduce B to Hermite normal form and obtain a particular set spanning W. Since the non-zero rows of the Hermite normal form are linearly independent, they form a basis of W.

Now let C be another set spanning W. In a similar fashion we can construct a matrix C whose rows represent the vectors in C and reduce this matrix to Hermite normal form. Let C' be the Hermite normal form obtained from C, and let B' be the Hermite normal form obtained from B. We do not assume that B and C have the same number of elements, and therefore B' and C' do not necessarily have the same number of rows. However, in each the number of non-zero rows must be equal to the dimension of W. We claim that the non-zero rows in these two normal forms are identical.

To see this, construct a new matrix with the non-zero rows of C' written beneath the non-zero rows of B' and reduce this matrix to Hermite normal form. Since the rows of C' are dependent on the rows of B', the rows of C' can be removed by elementary operations, leaving the rows of B'. Further reduction is not possible since B' is already in normal form. But by interchanging rows, which are elementary operations, we can obtain a matrix in which the non-zero rows of B' are beneath the non-zero rows of C'. As before, we can remove the rows of B' leaving the non-zero rows of C' as the normal form. Since the Hermite normal form is unique, we see that the non-zero rows of B' and C' are identical. The basis that we obtain from the non-zero rows of the Hermite normal form is the *standard basis* with respect to A for the subspace W.

This gives us an effective method for deciding when two sets span the same subspace. For example, in Chapter I-4, Exercise 5, we were asked to show that $\{(1, 1, 0, 0), (1, 0, 1, 1)\}$ and $\{(2, -1, 3, 3), (0, 1, -1, -1)\}$ span the same space. In either case we obtain $\{(1, 0, 1, 1), (0, 1, -1, -1)\}$ as the standard basis.

The Sum of Two Subspaces

If A_1 is a subset spanning W_1 and A_2 is a subset spanning W_2, then $A_1 \cup A_2$ spans $W_1 + W_2$ (Chapter I, Proposition 4.4). Thus we can find a basis for $W_1 + W_2$ by constructing a large matrix whose rows are the representations of the vectors in $A_1 \cup A_2$ and reducing it to Hermite normal form by elementary row operations.

The Characterization of a Subspace by a Set of Homogeneous Linear Equations

We have already seen that the set of all solutions of a system of homogeneous linear equations is a subspace, the kernel of the linear transformation represented by the matrix of coefficients. The method for solving such a system which we described in Section 7 amounts to passing from a characterization of a subspace as the set of all solutions of a system of equations to its description as the set of all linear combinations of a basis. The question

naturally arises: If we are given a spanning set for a subspace W, how can we find a system of simultaneous homogeneous linear equations for which W is exactly the set of solutions?

This is not at all difficult and no new procedures are required. All that is needed is a new look at what we have already done. Consider the homogeneous linear equation $a_1x_1 + \cdots + a_nx_n = 0$. There is no significant difference between the a_i's and the x_i's in this equation; they appear symmetrically. Let us exploit this symmetry systematically.

If $a_1x_1 + \cdots + a_nx_n = 0$ and $b_1x_1 + \cdots + b_nx_n = 0$ are two homogeneous linear equations then $(a_1 + b_1)x_1 + \cdots + (a_n + b_n)x_n = 0$ is a homogeneous linear equation as also is $aa_1x_1 + \cdots + aa_nx_n = 0$ where $a \in F$. Thus we can consider the set of all homogeneous linear equations in n unknowns as a vector space over F. The equation $a_1x_1 + \cdots + a_nx_n = 0$ is represented by the n-tuple $(a_1, \ldots, a_n)$.

When we write a matrix to represent a system of equations and reduce that matrix to Hermite normal form we are finding a standard basis for the subspace of the vector space of all homogeneous linear equations in $x_1, \ldots, x_n$ spanned by this system of equations just as we did in the first part of this section for a set of vectors spanning a subspace. The rank of the system of equations is the dimension of the subspace of equations spanned by the given system.

Now let W be a subspace given by a spanning set and solve for the subspace E of all equations satisfied by W. Then solve for the subspace of solutions of the system of equations E. W must be a subspace of the set of all solutions. Let W be of dimension v. By Theorem 7.1 the dimension of E is $n - v$. Then, in turn, the dimension of the set of all solutions of E is $n - (n - v) = v$. Thus W must be exactly the space of all solutions. Thus W and E characterize each other.

If we start with a system of equations and solve it by means of the Hermite normal form, as described in Section 7, we obtain in a natural way a basis for the subspace of solutions. This basis, however, will not be the standard basis. We can obtain full symmetry between the standard system of equations and the standard basis by changing the definition of the standard basis. Instead of applying the elementary row operations by starting with the left-hand column, start with the right-hand column. If the basis obtained in this way is called the standard basis, the equations obtained will be the standard equations, and the solution of the standard equations will be the standard basis. In the following example the computations will be carried out in this way to illustrate this idea. It is not recommended, however, that this be generally done since accuracy with one definite routine is more important.

Let

$$W = \langle (1, 0, -3, 11, -5), (3, 2, 5, -5, 3), (1, 1, 2, -4, 2), (7, 2, 12, 1, 2) \rangle.$$

We now find a standard basis by reducing

$$\begin{bmatrix} 1 & 0 & -3 & 11 & -5 \\ 3 & 2 & 5 & -5 & 3 \\ 1 & 1 & 2 & -4 & 2 \\ 7 & 2 & 12 & 1 & 2 \end{bmatrix}$$

to the form

$$\begin{bmatrix} 2 & 0 & 5 & 0 & 1 \\ 1 & 0 & 2 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

From this we see that the coefficients of our systems of equations satisfy the conditions

$$\begin{aligned} 2a_1 \quad &+ 5a_3 \quad + a_5 = 0 \\ a_1 \quad &+ 2a_3 + a_4 \quad = 0 \\ a_1 + a_2 \quad & \quad = 0. \end{aligned}$$

The coefficients a_1 and a_3 can be selected arbitrarily and the others computed from them. In particular, we have

$$(a_1, a_2, a_3, a_4, a_5) = a_1(1, -1, 0, -1, -2) + a_3(0, 0, 1, -2, -5).$$

The 5-tuples $(1, -1, 0, -1, -2)$ and $(0, 0, 1, -2, -5)$ represent the two standard linear equations

$$\begin{aligned} x_1 - x_2 \quad & \quad -x_4 - 2x_5 = 0 \\ & x_3 - 2x_4 - 5x_5 = 0. \end{aligned}$$

The reader should check that the vectors in W actually satisfy these equations and that the standard basis for W is obtained.

The Intersection of Two Subspaces

Let W_1 and W_2 be subspaces of U of dimensions ν_1 and ν_2, respectively, and let $W_1 \cap W_2$ be of dimension ν. Then $W_1 + W_2$ is of dimension $\nu_1 + \nu_2 - \nu$. Let E_1 and E_2 be the spaces of equations characterizing W_1 and W_2. As we have seen E_1 is of dimension $n - \nu_1$ and E_2 is of dimension $n - \nu_2$. Let the dimension of $E_1 + E_2$ be ρ. Then $E_1 \cap E_2$ is of dimension $(n - \nu_1) + (n - \nu_2) - \rho = 2n - \nu_1 - \nu_2 - \rho$.

Since the vectors in $W_1 \cap W_2$ satisfy the equations in both E_1 and E_2, they satisfy the equations in $E_1 + E_2$. Thus $\nu \leq n - \rho$. On the other hand,

W_1 and W_2 both satisfy the equations in $E_1 \cap E_2$ so that $W_1 + W_2$ satisfies the equations in $E_1 \cap E_2$. Thus $\nu_1 + \nu_2 - \nu \leq n - \{2n - \nu_1 - \nu_2 - \rho\} = \nu_1 + \nu_2 + \rho - n$. A comparison of these two inequalities shows that $\nu = n - \rho$ and hence that $W_1 \cap W_2$ is characterized by $E_1 + E_2$.

Given W_1 and W_2, the easiest way to find $W_1 \cap W_2$ is to determine E_1 and E_2 and then $E_1 + E_2$. From $E_1 + E_2$ we can then find $W_1 \cap W_2$. In effect, this involves solving three systems of equations, and reducing to Hermit normal form three times, but it is still easier than a direct assault on the problem.

As an example consider Exercise 8 of Chapter I-4. Let $W_1 = \langle(1, 2, 3, 6), (4, -1, 3, 6), (5, 1, 6, 12)\rangle$ and $W_2 = \langle(1, -1, 1, 1), (2, -1, 4, 5)\rangle$. Using the Hermite normal form, we find that $E_1 = \langle(-2, -2, 0, 1), (-1, -1, 1, 0)\rangle$ and $E_2 = \langle(-4, -3, 0, 1), (-3, -2, 1, 0)\rangle$. Again, using the Hermite normal form we find that the standard basis for $E_1 + E_2$ is $\{(1, 0, 0, \frac{1}{2}), (0, 1, 0, -1), (0, 0, 1, -\frac{1}{2})\}$. And from this we find quite easily that, $W_1 \cap W_2 = \langle(-\frac{1}{2}, 1, \frac{1}{2}, 1)\rangle$.

Let $B = \{\beta_1, \beta_2, \ldots, \beta_n\}$ be a given finite set of vectors. We wish to solve the problem posed in Theorem 2.2 of Chapter I. How do we show that some β_k is a linear combination of the β_i with $i < k$; or how do we show that no β_k can be so represented?

We are looking for a relation of the form

$$\beta_k = \sum_{i=1}^{k-1} x_{ik}\beta_i. \tag{8.1}$$

This is not a meaningful numerical problem unless β is a given specific set. This usually means that the β_i are given in terms of some coordinate system, relative to some given basis. But the relation (8.1) is independent of any coordinate system so we are free to choose a different coordinate system if this will make the solution any easier. It turns out that the tools to solve this problem are available.

Let $A = \{\alpha_1, \ldots, \alpha_m\}$ be the given basis and let

$$\beta_j = \sum_{i=1}^{m} a_{ij}\alpha_i, \qquad j = 1, \ldots, n. \tag{8.2}$$

If $A' = \{\alpha_1', \ldots, \alpha_m'\}$ is the new basis (which we have not specified yet), we would have

$$\beta_j = \sum_{i=1}^{m} a_{ij}'\alpha_i', \qquad j = 1, \ldots, n. \tag{8.3}$$

What is the relation between $A = [a_{ij}]$ and $A' = [a_{ij}']$? If P is the matrix of transition from the basis A to the basis A', by formula (4.3) we see that

$$A = PA'. \tag{8.4}$$

Since P is non-singular, it can be represented as a product of elementary matrices. This means A' can be obtained from A by a sequence of elementary row operations.

The solution to (8.1) is now most conveniently obtained if we take A' to be in Hermite normal form. Suppose that A' is in Hermite normal form and use the notation given in Theorem 5.1. Then, for β_{k_i} we would have

$$\beta_{k_i} = \alpha'_i, \tag{8.5}$$

and for j between k_r and k_{r+1} we would have

$$\begin{aligned} \beta_j &= \sum_{i=1}^{r} a'_{ij}\alpha'_i \\ &= \sum_{i=1}^{r} a'_{ij}\beta_{k_i} \end{aligned} \tag{8.6}$$

Since $k_i \leq k_r < j$, this last expression is a relation of the required form. (Actually, every linear relation that exists among the β_i can be obtained from those in (8.6). This assertion will not be used later in the book so we will not take space to prove it. Consider it "an exercise for the reader.")

Since the columns of A and A' represent the vectors in B, the rank of A is equal to the number of vectors in a maximal linearly independent subset of B. Thus, if B is linearly independent the rank of A will be n, this means that the Hermite normal form of A will either show that B is linearly independent or reveal a linear relation in B if it is dependent.

For example, consider the set $\{(1, 0, -3, 11, -5), (3, 2, 5, -5, 3), (1, 1, 2, -4, 2), (7, 2, 12, 1, 2)\}$. The implied context is that a basis $\mathsf{A} = \{\alpha_1, \ldots, \alpha_5\}$ is considered to be given and that $\beta_1 = \alpha_1 - 3\alpha_3 + 11\alpha_4 - 5\alpha_5$ etc. According to (8.2) the appropriate matrix is

$$\begin{bmatrix} 1 & 3 & 1 & 7 \\ 0 & 2 & 1 & 2 \\ -3 & 5 & 2 & 12 \\ 11 & -5 & -4 & 1 \\ -5 & 3 & 2 & 2 \end{bmatrix}$$

which reduces to the Hermite normal form

$$\begin{bmatrix} 1 & 0 & 0 & -\frac{3}{4} \\ 0 & 1 & 0 & \frac{23}{4} \\ 0 & 0 & 1 & -\frac{19}{2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is easily checked that $-\frac{3}{4}(1, 0, -3, 11, -5) + \frac{23}{4}(3, 2, 5, -5, 3) - \frac{19}{2}(1, 1, 2, -4, 2) = (7, 2, 12, 1, 2)$.

EXERCISES

1. Determine which of the following set in R^4 are linearly independent over R.

   (*a*) $\{(1, 1, 0, 1), (1, -1, 1, 1), (2, 2, 1, 2), (0, 1, 0, 0)\}$.
   (*b*) $\{(1, 0, 0, 1), (0, 1, 1, 0), (1, 0, 1, 0), (0, 1, 0, 1)\}$.
   (*c*) $\{(1, 0, 0, 1), (0, 1, 0, 1), (0, 0, 1, 1), (1, 1, 1, 1)\}$.

This problem is identical to Exercise 8, Chapter I-2.

2. Let W be the subspace of R^5 spanned by $\{(1, 1, 1, 1, 1), (1, 0, 1, 0, 1), (0, 1, 1, 1, 0), (2, 0, 0, 1, 1), (2, 1, 1, 2, 1), (1, -1, -1, -2, 2), (1, 2, 3, 4, -1)\}$. Find a standard basis for W and the dimension of W. This problem is identical to Exercise 6, Chapter I-4.

3. Show that $\{(1, -1, 2, -3), (1, 1, 2, 0), (3, -1, 6, -6)\}$ and $\{(1, 0, 1, 0), (0, 2, 0, 3)\}$ do not span the same subspace. This problem is identical to Exercise 7, Chapter I-4.

4. If $W_1 = \langle(1, 1, 3, -1), (1, 0, -2, 0), (3, 2, 4, -2)\rangle$ and $W_2 = \langle(1, 0, 0, 1), (1, 1, 7, 1)\rangle$ determine the dimension of $W_1 + W_2$.

5. Let $W = \langle(1, -1, -3, 0, 1), (2, 1, 0, -1, 4), (3, 1, -1, 1, 8), (1, 2, 3, 2, 6)\rangle$. Determine the standard basis for W. Find a set of linear equations which characterize W.

6. Let $W_1 = \langle(1, 2, 3, 6), (4, -1, 3, 6), (5, 1, 6, 12)\rangle$ and $W_2 = \langle(1, -1, 1, 1), (2, -1, 4, 5)\rangle$ be subspaces of R^4. Find bases for $W_1 \cap W_2$ and $W_1 + W_2$. Extend the basis of $W_1 \cap W_2$ to a basis of W_1 and extend the basis of $W_1 \cap W_2$ to a basis of W_2. From these bases obtain a basis of $W_1 + W_2$. This problem is identical to Exercise 8, Chapter I-4.

9 | Normal Forms

To understand fully what a normal form is, we must first introduce the concept of an equivalence relation. We say that a *relation* is defined in a set if, for each pair (a, b) of elements in this set, it is decided that "a is related to b" or "a is not related to b." If a is related to b, we write $a \sim b$. An *equivalence relation* in a set S is a relation in S satisfying the following laws:

Reflexive law: $a \sim a$,
Symmetric law: If $a \sim b$, then $b \sim a$.
Transitive law: If $a \sim b$ and $b \sim c$, then $a \sim c$.

If for an equivalence relation we have $a \sim b$, we say that a is *equivalent* to b.

Examples. Among rational fractions we can define $a/b \sim c/d$ (for a, b, c, d integers) if and only if $ad = bc$. This is the ordinary definition of equality in rational numbers, and this relation satisfies the three conditions of an equivalence relation.

In geometry we do not ordinarily say that a straight line is parallel to itself. But if we agree to say that a straight line is parallel to itself, the concept of parallelism is an equivalence relation among the straight lines in the plane or in space.

Geometry has many equivalence relations: congruence of triangles, similarity of triangles, the concept of projectivity in projective geometry, etc. In dealing with time we use many equivalence relations: same hour of the day, same day of the week, etc. An equivalence relation is like a generalized equality. Elements which are equivalent share some common or underlying property. As an example of this idea, consider a collection of sets. We say that two sets are equivalent if their elements can be put into a one-to-one correspondence; for example, a set of three battleships and a set of three cigars are equivalent. Any set of three objects shares with any other set of three objects a concept which we have abstracted and called "three." All other qualities which these sets may have are ignored.

It is most natural, therefore, to group mutually equivalent elements together into classes which we call *equivalence classes.* Let us be specific about how this is done. For each $a \in S$, let S_a be the set of all elements in S equivalent to a; that is, $b \in S_a$ if and only if $b \sim a$. We wish to show that the various sets we have thus defined are either disjoint or identical.

Suppose $S_a \cap S_b$ is not empty; that is, there exists a $c \in S_a \cap S_b$ such that $c \sim a$ and $c \sim b$. By symmetry $b \sim c$, and by transitivity $b \sim a$. If d is any element of S_b, $d \sim b$ and hence $d \sim a$. Thus $d \in S_a$ and $S_b \subset S_a$. Since the relation between S_a and S_b is symmetric we also have $S_a \subset S_b$ and hence $S_a = S_b$. Since $a \in S_a$ we have shown, in effect, that a proposed equivalence class can be identified by any element in it. An element selected from an equivalence class will be called a *representative* of that class.

An equivalence relation in a set S defines a partition of that set into equivalence classes in the following sense: (1) Every element of S is in some equivalence class, namely, $a \in S_a$. (2) Two elements are in the same equivalence class if and only if they are equivalent. (3) Non-identical equivalence classes are disjoint. On the other hand, a partition of a set into disjoint subsets can be used to define an equivalence relation; two elements are equivalent if and only if they are in the same subset.

The notions of equivalence relations and equivalence classes are not nearly so novel as they may seem at first. Most students have encountered these ideas before, although sometimes in hidden forms. For example, we may say that two differentiable functions are equivalent if and only if

they have the same derivative. In calculus we use the letter "C" in describing the equivalence classes; for example, $x^3 + x^2 + 2x + C$ is the set (equivalence class) of all functions whose derivative is $3x^2 + 2x + 2$.

In our study of matrices we have so far encountered four different equivalence relations:

I. The matrices A and B are said to be *left associate* if there exists a non-singular matrix Q such that $B = Q^{-1}A$. Multiplication by Q^{-1} corresponds to performing a sequence of elementary row operations. If A represents a linear transformation σ of U into V with respect to a basis A in U and a basis B in V, the matrix B represents σ with respect to A and a new basis in V.

II. The matrices A and B are said to be *right associate* if there exists a non-singular matrix P such that $B = AP$.

III. The matrices A and B are said to be *associate* if there exist non-singular matrixes P and Q such that $B = Q^{-1}AP$. The term "associate" is not a standard term for this equivalence relation, the term most frequently used being "equivalent." It seems unnecessarily confusing to use the same term for one particular relation and for a whole class of relations. Moreover, this equivalence relation is perhaps the least interesting of the equivalence relations we shall study.

IV. The matrices A and B are said to be *similar* if there exists a non-singular matrix P such that $B = P^{-1}AP$. As we have seen (Section 4) similar matrices are representations of a single linear transformation of a vector space into itself. This is one of the most interesting of the equivalence relations, and Chapter III is devoted to a study of it.

Let us show in detail that the reation we have defined as left associate is an equivalence relation. The matrix Q^{-1} appears in the definition because Q represents the matrix of transition. However, Q^{-1} is just another non-singular matrix, so it is clearly the same thing to say that A and B are left associate if and only if there exists a non-singular matrix Q such that $B = QA$.

(1) $A \sim A$ since $IA = A$.

(2) If $A \sim B$, there is a non-singular matrix Q such that $B = QA$. But then $A = Q^{-1}B$ so that $B \sim A$.

(3) If $A \sim B$ and $B \sim C$, there exist non-singular matrices Q and P such that $B = QA$ and $C = PB$. But then $PQA = PB = C$ and PQ is non-singular so that $A \sim C$.

For a given type of equivalence relation among matrices a *normal form* is a particular matrix chosen from each equivalence class. It is a representative of the entire class of equivalent matrices. In mathematics the terms "normal" and "canonical" are frequently used to mean "standard" in some particular sense. A normal form or canonical form is a standard

form selected to represent a class of equivalent elements. A normal form should be selected to have the following two properties: Given any matrix A, (1) it should be possible by fairly direct and convenient methods to find the normal form of the equivalence class containing A, and (2) the method should lead to a unique normal form.

Often the definition of a normal form is compromised with respect to the second of these desirable properties. For example, if the normal form were a matrix with complex numbers in the main diagonal and zeros elsewhere, to make the normal form unique it would be necessary to specify the order of the numbers in the main diagonal. But it is usually sufficient to know the numbers in the main diagonal without regard to their order, so it would be an awkward complication to have to specify their order.

Normal forms have several uses. Perhaps the most important use is that the normal form should yield important or useful information about the concept that the matrix represents. This should be amply illustrated in the case of the concept of left associate and the Hermite normal form. We introduced the Hermite normal form through linear transformations, but we found that it yielded very useful information when the matrix was used to represent linear equations or bases of subspaces.

Given two matrices, we can use the normal form to tell whether they are equivalent. It is often easier to reduce each to normal form and compare the normal forms than it is to transform one into the other. This is the case, for example, in the application described in the first part of Section 8.

Sometimes, knowing the general appearance of the normal form, we can find all the information we need without actually obtaining the normal form. This is the case for the equivalence relation we have called associate. The normal form for this equivalence relation is described in Theorem 4.1. There is just one normal form for each possible value of the rank. The number of different equivalence classes is $\min\{m, n\} + 1$. With this notion of equivalence the rank of a matrix is the only property of importance. Any two matrices of the same rank are associate. In practice we can find the rank without actually computing the normal form of Theorem 4.1. And knowing the rank we know the normal form.

We encounter several more equivalence relations among matrices. The type of equivalence introduced will depend entirely on the underlying concepts the matrices are used to represent. It is worth mentioning that for the equivalence relations we introduce there is no necessity to prove, as we did for an example above, that each is an equivalence relation. An underlying concept will be defined without reference to any coordinate system or choice of basis. The matrices representing this concept will transform according to certain rules when the basis is changed. Since a given basis can be retained the relation defined is reflexive. Since a basis changed can be changed back

to the original basis, the relation defined is symmetric. A basis changed once and then changed again depends only on the final choice so that the relation is transitive.

For a fixed basis A in U and B in V two different linear transformations σ and τ of U into V are represented by different matrices. If it is possible, however, to choose bases A' in U and B' in V such that the matrix representing τ with respect to A' and B' is the same as the matrix representing σ with respect to A and B, then it is certainly clear that σ and τ share important geometric properties.

For a fixed σ two matrices A and A' representing σ with respect to different bases are related by a matrix equation of the form $A' = Q^{-1}AP$. Since A and A' represent the same linear transformation we feel that they should have some properties in common, those dependent upon σ.

These two points of view are really slightly different views of the same kind of relationship. In the second case, we can consider A and A' as representing two linear transformations with respect to the same basis, instead of the same linear transformation with respect to different bases.

For example, in R^2 the matrix $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ represents a reflection about the x_1-axis and $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ represents a reflection about the x_2-axis. When both linear transformations are referred to the same coordinate system they are different. However, for the purpose of discussing properties independent of a coordinate system they are essentially alike. The study of equivalence relations is motivated by such considerations, and the study of normal forms is aimed at determining just what these common properties are that are shared by equivalent linear transformations or equivalent matrices.

To make these ideas precise, let σ and τ be linear transformations of V into itself. We say that σ and τ are *similar* if there exist bases A and B of V such that the matrix representing σ with respect to A is the same as the matrix representing τ with respect to B. If A and B are the matrices representing σ and τ with respect to A and P is the matrix of transition from A to B, then $P^{-1}BP$ is the matrix representing τ with respect to B. Thus σ and τ are similar if $P^{-1}BP = A$.

In a similar way we can define the concepts of left associate, right associate, and associate for linear transformations.

*10 | Quotient Sets, Quotient Spaces

Definition. If S is any set on which an equivalence relation is defined, the collection of equivalence classes is called the *quotient* or *factor set*. Let $\bar{S}$ denote the quotient set. An element of $\bar{S}$ is an equivalence class. If a is an

element of S and $\bar{a}$ is the equivalence class containing a, the mapping η that maps a onto $\bar{a}$ is well defined. This mapping is called the *canonical mapping*.

Although the concept of a quotient set might appear new to some, it is certain that almost everyone has encountered the idea before, perhaps in one guise or another. One example occurs in arithmetic. In this setting, let S be the set of all formal fractions of the form a/b where a and b are integers and $b \neq 0$. Two such fractions, a/b and c/d, are equivalent if and only if $ad = bc$. Each equivalence class corresponds to a single rational number. The rules of arithmetic provide methods of computing with rational numbers by performing appropriate operations with formal fractions selected from the corresponding equivalence classes.

Let U be a vector space over F and let K be a subspace of U. We shall call two vectors $\alpha, \beta \in U$ equivalent modulo K if and only if their difference lies in K. Thus $\alpha \sim \beta$ if and only if $\alpha - \beta \in K$. We must first show this defines an equivalence relation. (1) $\alpha \sim \alpha$ because $\alpha - \alpha = 0 \in K$. (2) $\alpha \sim \beta \Rightarrow \alpha - \beta \in K \Rightarrow \beta - \alpha \in K \Rightarrow \beta \sim \alpha$. (3) $\{\alpha \sim \beta \text{ and } \beta \sim \gamma\} \Rightarrow \{\alpha - \beta \in K \text{ and } \beta - \gamma \in K\}$. Since K is a subspace $\alpha - \gamma = (\alpha - \beta) + (\beta - \gamma) \in K$ and, hence, $\alpha \sim \gamma$. Thus "$\sim$" is an equivalence relation.

We wish to define vector addition and scalar multiplication in $\bar{U}$. For $\alpha \in U$, let $\bar{\alpha} \in \bar{U}$ denote the equivalence class containing α. α is called a *representative* of $\bar{\alpha}$. Since $\bar{\alpha}$ may contain other elements besides α, it may happen that $\alpha \neq \alpha'$ and yet $\bar{\alpha} = \bar{\alpha}'$. Let $\bar{\alpha}$ and $\bar{\beta}$ be two elements in $\bar{U}$. Since $\alpha, \beta \in U$, $\alpha + \beta$ is defined. We wish to define $\overline{\alpha + \beta}$ to be the sum of $\bar{\alpha}$ and $\bar{\beta}$. In order for this to be well defined we must end up with the same equivalence class as the sum if different representatives are chosen from $\bar{\alpha}$ and $\bar{\beta}$. Suppose $\bar{\alpha} = \bar{\alpha}'$ and $\bar{\beta} = \bar{\beta}'$. Then $\alpha - \alpha' \in K$, $\beta - \beta' \in K$, and $(\alpha + \beta) - (\alpha' + \beta') \in K$. Thus $\overline{\alpha + \beta} = \overline{\alpha' + \beta'}$ and the sum is well defined. Scalar multiplication is defined similarly. For $a \in F$, $a\bar{\alpha}$ is thus defined to be the equivalence class containing $a\alpha$; that is, $a\bar{\alpha} = \overline{a\alpha}$ These operations in $\bar{U}$ are said to be *induced* by the corresponding operation in U.

Theorem 10.1. *If U is a vector space over F, and K is a subspace of U, the quotient set $\bar{U}$ with vector addition and scalar multiplication defined as above is a vector space over F.*

PROOF. We leave this as an exercise. □

For any $\alpha \in U$, the symbol $\alpha + K$ is used to denote the set of all elements in U that can be written in the form $\alpha + \gamma$ where $\gamma \in K$. (Strictly speaking, we should denote the set by $\{\alpha\} + K$ so that the plus sign combines two objects of the same type. The notation introduced here is traditional and simpler.) The set $\alpha + K$ is called a *coset* of K. If $\beta \in \alpha + K$, then $\beta - \alpha \in K$ and

$\beta \sim \alpha$. Conversely, if $\beta \sim \alpha$, then $\beta - \alpha = \gamma \in K$ so $\beta \in \alpha + K$. Thus $\alpha + K$ is simply the equivalence class $\bar{\alpha}$ containing α. Thus $\alpha + K = \beta + K$ if and only if $\alpha \in \bar{\beta} = \beta + K$ or $\beta \in \bar{\alpha} = \alpha + K$.

The notation $\alpha + K$ to denote $\bar{\alpha}$ is convenient to use in some calculations. For example, $\bar{\alpha} + \bar{\beta} = (\alpha + K) + (\beta + K) = \alpha + \beta + K = \overline{\alpha + \beta}$, and $a\bar{\alpha} = a(\alpha + K) = a\alpha + aK \subset a\alpha + K = \overline{a\alpha}$. Notice that $a\bar{\alpha} = \overline{a\alpha}$ when $\bar{\alpha}$ and $\overline{a\alpha}$ are considered to be elements of $\bar{U}$ and scalar mutliplication is the induced operation, but that $a\bar{\alpha}$ and $\overline{a\alpha}$ may not be the same when they are viewed as subsets of U (for example, let $a = 0$). However, since $a\bar{\alpha} \subset \overline{a\alpha}$ the set $a\bar{\alpha}$ determines the desired coset in $\bar{U}$ for the induced operations. Thus we can compute effectively in $\bar{U}$ by doing the corresponding operations with representatives. This is precisely what is done when we compute in residue classes of integers modulo an integer m.

Definition. $\bar{U}$ with the induced operations is called a *factor space* or *quotient space*. In order to designate the role of the subspace K which defines the equivalence relations, $\bar{U}$ is usually denoted by U/K.

In our discussion of solutions of linear problems we actually encountered quotient spaces, but the discussion was worded in such a way as to avoid introducing this more sophisticated concept. Given the linear transformation σ of U into V, let K be the kernel of σ and let $\bar{U} = U/K$ be the corresponding quotient space. If α_1 and α_2 are solutions of the linear problem, $\alpha(\xi) = \beta$, then $\sigma(\alpha_1 - \alpha_2) = 0$ so that α_1 and α_2 are in the same coset of K. Thus for each $\beta \in \mathrm{Im}(\sigma)$ there corresponds precisely one coset of K. In fact the correspondence between U/K and $\mathrm{Im}(\sigma)$ is an isomorphism, a fact which is made more precise in the following theorem.

Theorem 10.2. *(First homomorphism theorem). Let σ be a linear transformation of U into V. Let K be the kernel of σ. Then σ can be written as the product of a canonical mapping η of U onto $\bar{U} = U/K$ and a monomorphism σ_1 of $\bar{U}$ into V.*

PROOF. The canonical mapping η has already been defined. To define σ_1, for each $\bar{\alpha} \in \bar{U}$ let $\sigma_1(\bar{\alpha}) = \sigma(\alpha)$ where α is any representative of $\bar{\alpha}$. Since $\sigma(\alpha) = \sigma(\alpha')$ for $\alpha \sim \alpha'$, σ_1 is well defined. It is easily seen that σ_1 is a monomorphism since σ must have different values in different cosets. □

The homomorphism theorem is usually stated by saying, "The homomorphic image is isomorphic to the quotient space of U modulo the kernel."

Theorem 10.3. *(Mapping decomposition theorem). Let σ be a linear transformation of U into V. Let K be the kernel of σ and I the image of σ. Then σ can be written as the product $\sigma = \iota\sigma_1\eta$, where η is the canonical mapping of*

U onto $\bar{U} = U/K$, σ_1 is an isomorphism of $\bar{U}$ onto I, and ι is the injection of I into V.

PROOF. Let σ' be the linear transformation of U onto I induced by restricting the codomain of σ to the image of σ. By Theorem 10.2, σ' can be written in the form $\sigma' = \sigma_1\eta$. □

Theorem 10.4. *(Mapping factor theorem). Let S be a subspace of U and let $\bar{U} = U/S$ be the resulting quotient space. Let σ be a linear transformation of U into V, and let K be the kernel of σ. If $S \subset K$, then there exists a linear transformation σ_1 of $\bar{U}$ into V such that $\sigma = \sigma_1\eta$ where η is the canonical mapping of U onto $\bar{U}$.*

PROOF. For each $\bar{\alpha} \in \bar{U}$, let $\sigma_1(\bar{\alpha}) = \sigma(\alpha)$ where $\alpha \in \bar{\alpha}$. If α' is another representative of $\bar{\alpha}$, then $\alpha - \alpha' \in S \subset K$. Thus $\sigma(\alpha) = \sigma(\alpha')$ and σ_1 is well defined. It is easy to check that σ_1 is linear. Clearly, $\sigma(\alpha) = \sigma_1(\bar{\alpha}) = \sigma_1(\eta(\alpha))$ for all $\alpha \in U$, and $\sigma = \sigma_1\eta$. □

We say that σ *factors* through $\bar{U}$.

Note that the homomorphism theorem is a special case of the factor theorem in which $K = S$.

Theorem 10.5. *(Induced mapping theorem). Let U and V be vector spaces over F, and let τ be a linear transformation of U into V. Let U_0 be a subspace of U and let V_0 be a subspace of V. If $\tau(U_0) \subset V_0$, it is possible to define in a natural way a mapping $\bar{\tau}$ of U/U_0 into V/V_0 such that $\sigma_2\tau = \bar{\tau}\sigma_1$ where σ_1 is the canonical mapping U onto $\bar{U}$ and σ_2 is the canonical mapping of V onto $\bar{V}$.*

PROOF. Consider $\sigma = \sigma_2\tau$, which maps U into $\bar{V}$. The kernel of σ is $\tau^{-1}(V_0)$. By assumption, $U_0 \subset \tau^{-1}(V_0)$. Hence, by the mapping factor theorem, there is a linear transformation $\bar{\tau}$ such that $\bar{\tau}\sigma_1 = \sigma_2\tau$. □

We say that $\bar{\tau}$ is *induced* by τ.

Numerical calculations with quotient spaces can usually be avoided in problems involving finite dimensional vector spaces. If U is a vector space over F and K is a subspace of U, we know from Theorem 4.9 of Chapter I that K is a direct summand. Let $U = K \oplus W$. Then the canonical mapping η maps W isomorphically onto U/K. Thus any calculation involving U/K can be carried out in W.

Although there are many possible choices for the complementary subspace W, the Hermite normal form provides a simple and effective way to select a W and a basis for it. This typically arises in connection with a linear problem. To see this, reexamine the proof of Theorem 5.1. There we let $k_1, k_2, \ldots, k_\rho$ be those indices for which $\sigma(\alpha_{k_i}) \notin \sigma(U_{k_i-1})$. We showed there that $\{\beta'_1, \ldots, \beta'_k\}$ where $\beta'_i = \sigma(\alpha_{k_i})$ formed a basis of $\sigma(U)$. $\{\alpha_{k_1}, \alpha_{k_2}, \ldots, \alpha_{k_\rho}\}$ is a basis for a suitable W which is complementary to $K(\sigma)$.

Example. Consider the linear transformation σ of R^5 into R^3 represented by the matrix

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

It is easy to determine that the kernel K of σ is 2-dimensional with basis $\{(1, -1, -1, 1, 0), (0, 0, -1, 0, 1)\}$. This means that σ has rank 3 and the image of σ is all of R^3. Thus $\overline{R^5} = R^5/K$ is isomorphic to R^3.

Consider the problem of solving the equation $\sigma(\xi) = \beta$, where β is represented by (b_1, b_2, b_3). To solve this problem we reduce the augmented matrix

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & b_1 \\ 0 & 1 & 0 & 1 & 0 & b_2 \\ 1 & 0 & 0 & -1 & 0 & b_3 \end{bmatrix}$$

to the Hermite normal form

$$\begin{bmatrix} 1 & 0 & 0 & -1 & 0 & b_3 \\ 0 & 1 & 0 & 1 & 0 & b_2 \\ 0 & 0 & 1 & 1 & 1 & b_1 - b_3 \end{bmatrix}.$$

This means the solution ξ is represented by

$$(b_3, b_2, b_1 - b_3, 0, 0) + x_4(1, -1, -1, 1, 0) + x_5(0, 0, -1, 0, 1).$$

$$(b_3, b_2, b_1 - b_3, 0, 0) = b_1(0, 0, 1, 0, 0) + b_2(0, 1, 0, 0, 0) + b_3(1, 0, -1, 0, 0)$$

is a particular solution and a convenient basis for a subspace W complementary to K is $\{(0, 0, 1, 0, 0), (0, 1, 0, 0, 0), (1, 0, -1, 0, 0)\}$. σ maps $b_1(0, 0, 1, 0, 0) + b_2(0, 1, 0, 0, 0) + b_3(1, 0, -1, 0, 0)$ onto (b_1, b_2, b_3). Hence, W is mapped isomorphically onto R^3.

This example also provides an opportunity to illustrate the working of the first homomorphism theorem. For any $(x_1, x_2, x_3, x_4, x_5) \in R^5$.

$$\begin{aligned}(x_1, x_2, x_3, x_4, x_5) = {} & (x_1 + x_3 + x_5)(0, 0, 1, 0, 0) \\ & + (x_2 + x_4)(0, 1, 0, 0, 0) \\ & + (x_1 - x_4)(1, 0, -1, 0, 0) \\ & + x_4(1, -1, -1, 1, 0) + x_5(0, 0, -1, 0, 1).\end{aligned}$$

Thus $(x_1, x_2, x_3, x_4, x_5)$ is mapped onto the coset $(x_1 + x_3 + x_5)(0, 0, 1, 0, 0) + (x_2 + x_4)(0, 1, 0, 0, 0) + (x_1 - x_4)(1, 0, -1, 0, 0) + K$ under the natural homomorphism onto R^5/K. This coset is then mapped isomorphically onto $(x_1 + x_3 + x_5, x_2 + x_4, x_1 - x_y) \in R^3$. However, it is somewhat contrived to

work out an example of this type. The main importance of the first homomorphism theorem is theoretical and not computational.

*11 | Hom(U, V)

Let U and V be vector spaces over F. We have already observed in Section 1 that the set of all linear transformations of U into V can be made into a vector space over F by defining addition and scalar multiplication appropriately. In this section we will explore some of the elementary consequences of this observation. We shall call this vector space Hom(U, V), "The space of all homomorphisms of U into V."

Theorem 11.1. *If* $\dim U = n$ *and* $\dim V = m$, *then* $\dim \mathrm{Hom}(U, V) = mn$.

PROOF. Let $\{\alpha_1, \ldots, \alpha_n\}$ be a basis of U and let $\{\beta_1, \ldots, \beta_m\}$ be a basis of V. Define the linear transformation of σ_{ij} by the rule

$$\begin{aligned}\sigma_{ij}(\alpha_k) &= \delta_{jk}\beta_i \\ &= \sum_{r=1}^{m} \delta_{ri}\delta_{jk}\beta_r \end{aligned} \tag{11.1}$$

Thus σ_{ij} is represented by the matrix $[\delta_{ri}\delta_{jk}] = A_{ij}$. A_{ij} has a zero in every position except for a 1 in row i column j.

The set $\{\sigma_{ij}\}$ is linearly independent. For if a linear relation existed among the σ_{ij} it would be of the form

$$\sum_{i,j} a_{ij}\sigma_{ij} = 0.$$

This means $\sum_{i,j} a_{ij}\sigma_{ij}(\alpha_k) = 0$ for all α_k. But $\sum_{i,j} a_{ij}\sigma_{ij}(\alpha_k) = \sum_{i,j} a_{ij}\delta_{jk}\beta_i = \sum_i a_{ik}\beta_i = 0$. Since $\{\beta_i\}$ is a lineary independent set, $a_{ik} = 0$ for $i = 1, 2, \ldots, m$. Since this is true for each k, all $a_{ij} = 0$ and $\{\sigma_{ij}\}$ is linearly independent.

If $\sigma \in \mathrm{Hom}(U, V)$ and $\sigma(\alpha_k) = \sum_{i=1}^{m} a_{ik}\beta_i$,
then

$$\begin{aligned}\sigma(\alpha_k) &= \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{ij}\delta_{jk}\right)\beta_i \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\sigma_{ij}(\alpha_k) \\ &= \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\sigma_{ij}\right)(\alpha_k).\end{aligned}$$

Thus $\{\sigma_{ij}\}$ spans Hom(U, V), which is therefore of dimension mn. □

If V_1 is a subspace of V, every linear transformation of U into V_1 also defines a mapping of U into V. This mapping of U into V is a linear transformation of

U into V. Thus, with each element of $\mathrm{Hom}(U, V_1)$ there is associated in a natural way an element of $\mathrm{Hom}(U, V)$. We can identify $\mathrm{Hom}(U, V_1)$ with a subset of $\mathrm{Hom}(U, V)$. With this identification $\mathrm{Hom}(U, V_1)$ is a subspace of $\mathrm{Hom}(U, V)$.

Now let U_1 be a subspace of U. In this case we cannot consider $\mathrm{Hom}(U_1, V)$ to be a subset of $\mathrm{Hom}(U, V)$ since a linear transformation in $\mathrm{Hom}(U_1, V)$ is not necessarily defined on all of U. But any linear transformation in $\mathrm{Hom}(U, V)$ is certainly defined on U_1. If $\sigma \in \mathrm{Hom}(U, V)$ we shall consider the mapping obtained by applying σ only to elements in U_1 to be a new function and denote it by $R(\sigma)$. $R(\sigma)$ is called the *restriction* of σ to U_1. We can consider $R(\sigma)$ to be an element of $\mathrm{Hom}(U_1, V)$.

It may happen that different linear transformations defined on U produce the same restriction on U_1. We say that σ_1 and σ_2 are equivalent on U_1 if and only if $R(\sigma_1) = R(\sigma_2)$. It is clear that $R(\sigma + \tau) = R(\sigma) + R(\tau)$ and $R(a\sigma) = aR(\sigma)$ so that the mapping of $\mathrm{Hom}(U, V)$ into $\mathrm{Hom}(U_1, V)$ is linear. We call this mapping R, the *restriction mapping*.

The kernel of R is clearly the set of all linear transformations in $\mathrm{Hom}(U, V)$ that vanish on U_1. Let us denote this kernel by U_1^*.

If $\underline{\sigma}$ is any linear transformation belonging to $\mathrm{Hom}(U_1, V)$, it can be extended to a linear transformation belonging to $\mathrm{Hom}(U, V)$ in many ways. If $\{\alpha_1, \ldots, \alpha_n\}$ is a basis of U such that $\{\alpha_1, \ldots, \alpha_r\}$ is a basis of U_1, then let $\sigma(\alpha_j) = \underline{\sigma}(\alpha_j)$ for $j = 1, \ldots, r$, and let $\sigma(\alpha_j)$ be defined arbitrarily for $j = r + 1, \ldots, n$. Since σ is then the restriction of σ, we see that R is an epimorphism of $\mathrm{Hom}(U, V)$ onto $\mathrm{Hom}(U_1, V)$. Since $\mathrm{Hom}(U, V)$ is of dimension mn and $\mathrm{Hom}(U_1, V)$ is of dimension mr, U_1^* is of dimension $m(n - r)$.

Theorem 11.2. $\mathrm{Hom}(U_1, V)$ *is canonically isomorphic to* $\mathrm{Hom}(U, V)/U_1^*$. □

Note: It helps the intuitive understanding of this theorem to examine the method by which we obtained an extension of $\underline{\sigma}$ on U_1, to σ on U. U_1^* is the set of all extensions of σ when $\underline{\sigma}$ is the zero mapping, and one can see directly that the dimension of U_1^* is $(n - r)m$.