LINEAR TRANSFORMATIONS

Definition 2–1. A *linear transformation*, or *linear operator*, from a vector space $\mathcal{V}_1$ to a vector space $\mathcal{V}_2$ is a function A which associates with each vector $\mathbf{x}$ in $\mathcal{V}_1$ a unique vector $A(\mathbf{x})$ in $\mathcal{V}_2$ in such a way that

$$A(\mathbf{x}_1 + \mathbf{x}_2) = A(\mathbf{x}_1) + A(\mathbf{x}_2) \tag{2–1}$$

and

$$A(\alpha\mathbf{x}) = \alpha A(\mathbf{x}) \tag{2–2}$$

for all vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}$ in $\mathcal{V}_1$, and all scalars α.

In other words, a linear transformation is a function, or mapping, from one vector space to another which sends sums into sums and scalar products into scalar products (see Fig. 2–1). These requirements are sometimes referred to by saying that a linear transformation is "compatible" with the algebraic operations of addition and scalar multiplication defined on vector spaces, and it is just this compatibility which accounts for the importance of such functions in linear algebra.

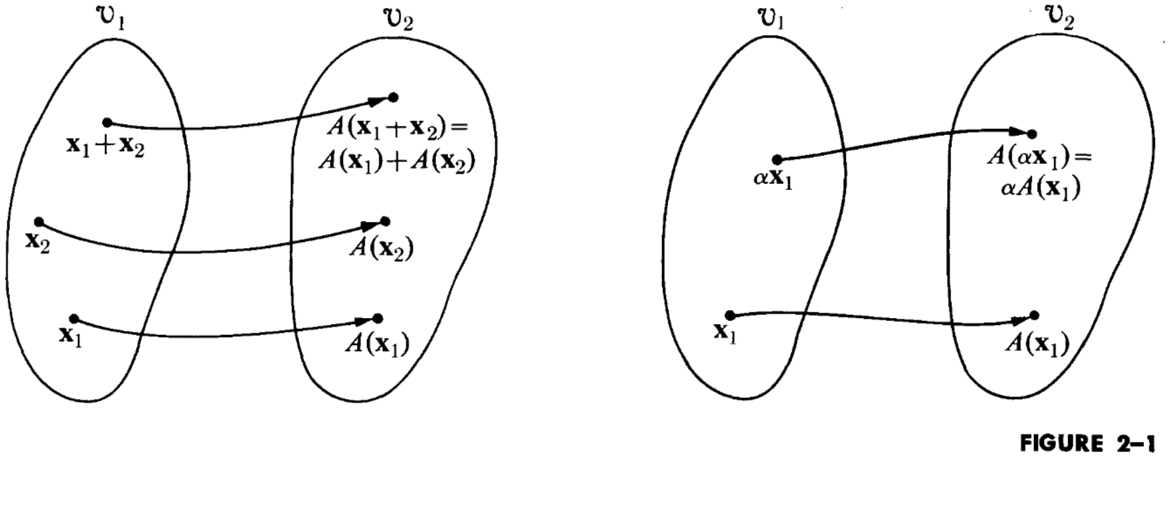


FIGURE 2–1

One consequence of Definition 2–1 is that a linear transformation *always* maps the zero vector of $\mathcal{V}_1$ onto the zero vector of $\mathcal{V}_2$; that is,

$$A(\mathbf{0}) = \mathbf{0}. \tag{2–3}$$

Another is that

$$A(\alpha_1\mathbf{x}_1 + \cdots + \alpha_n\mathbf{x}_n) = \alpha_1 A(\mathbf{x}_1) + \cdots + \alpha_n A(\mathbf{x}_n) \tag{2–4}$$

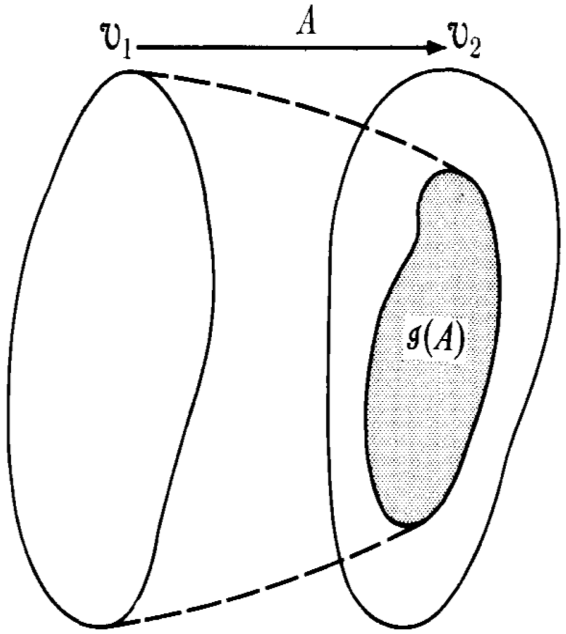
for any finite collection of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathcal{V}_1$ and scalars $\alpha_1, \ldots, \alpha_n$. The first of these assertions can be established by setting $\alpha = 0$ in (2–2), the second by repeated use of (2–1) and (2–2) in the obvious fashion. In particular, when $n = 2$, (2–4) becomes

$$A(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) = \alpha_1 A(\mathbf{x}_1) + \alpha_2 A(\mathbf{x}_2). \tag{2–5}$$

We call attention to this equation in order to remark that, by itself, it can be (and often is) taken as the definition of a linear transformation, since (2–1) and (2–2) are satisfied if and only if (2–5) is. From time to time we shall use this fact when proving that a function is a linear transformation.

If A is a linear transformation from $\mathcal{V}_1$ to $\mathcal{V}_2$ we write $A: \mathcal{V}_1 \to \mathcal{V}_2$ (read "A maps $\mathcal{V}_1$ into $\mathcal{V}_2$"), and refer to $\mathcal{V}_1$ as the *domain* of A. In this case the set of all vectors $\mathbf{y}$ in $\mathcal{V}_2$ such that $\mathbf{y} = A(\mathbf{x})$ for some $\mathbf{x}$ in $\mathcal{V}_1$ is called the *image* or *range* of A, and is denoted by $\mathcal{I}(A)$. Lest it be overlooked, we point out that the image of A need not be all of $\mathcal{V}_2$, a possibility which is made explicit by saying that A maps $\mathcal{V}_1$ *into* $\mathcal{V}_2$ (Fig. 2–2). Of course, it may happen that $\mathcal{I}(A) = \mathcal{V}_2$, in which case the term *onto* is used.

FIGURE 2–2



EXAMPLE 1. Let $\mathbf{x} = (x_1, x_2)$ be an arbitrary vector in $\mathcal{R}^2$, and set

$$A(\mathbf{x}) = (x_1, -x_2).$$

Geometrically A can be described as the linear transformation mapping $\mathcal{R}^2$ onto itself by reflection across the x_1-axis.

EXAMPLE 2. A second linear transformation for which we reserve a special symbol is the *identity transformation* I mapping a vector space $\mathcal{V}$ onto itself. The defining equation for I is

$$I(\mathbf{x}) = \mathbf{x}$$

for all x in $\mathcal{V}$; its linearity is obvious.

EXAMPLE 3. Consider the space $\mathcal{C}[a, b]$ of all real valued continuous functions on the interval $[a, b]$, and for each $\mathbf{f}$ in $\mathcal{C}[a, b]$ set

$$A(\mathbf{f}) = \int_a^x \mathbf{f}(t)\,dt, \qquad a \leq x \leq b.$$

Then since $A(\mathbf{f})$ is continuous on $[a, b]$, A can be viewed as a mapping of $\mathcal{C}[a, b]$ into itself. As such it is linear since

$$\begin{aligned} A(\alpha_1\mathbf{f}_1 + \alpha_2\mathbf{f}_2) &= \int_a^x [\alpha_1\mathbf{f}_1(t) + \alpha_2\mathbf{f}_2(t)]\,dt \\ &= \int_a^x \alpha_1\mathbf{f}_1(t)\,dt + \int_a^x \alpha_2\mathbf{f}_2(t)\,dt \\ &= \alpha_1\int_a^x \mathbf{f}_1(t)\,dt + \alpha_2\int_a^x \mathbf{f}_2(t)\,dt \\ &= \alpha_1 A(\mathbf{f}_1) + \alpha_2 A(\mathbf{f}_2). \end{aligned}$$

EXAMPLE 4. For the same reasons as those just given, the mapping $A\colon \mathcal{C}[a,b] \to \mathcal{R}^1$ defined by

$$A(\mathbf{f}) = \int_a^b \mathbf{f}(x)\,dx$$

is also linear.

EXAMPLE 5. Let $\mathcal{C}^1[a, b]$ denote the space of all continuously differentiable functions on $[a, b]$, and let D denote the operation of differentiation on this space; that is, $D(\mathbf{f}) = \mathbf{f}'$. Then the familiar identities

$$D(\mathbf{f}_1 + \mathbf{f}_2) = D(\mathbf{f}_1) + D(\mathbf{f}_2), \qquad D(\alpha\mathbf{f}) = \alpha D(\mathbf{f})$$

imply that D is a linear transformation from $\mathcal{C}^1[a, b]$ to $\mathcal{C}[a, b]$. More generally, the operation of taking nth derivatives is a linear transformation mapping the space of n-times continuously differentiable functions on an interval $[a, b]$ into the space $\mathcal{C}[a, b]$.

Definition 2–2. Let A and B be linear transformations from $\mathcal{V}_1$ to $\mathcal{V}_2$. Then their *sum*, $A + B$, is the transformation from $\mathcal{V}_1$ to $\mathcal{V}_2$ defined by

$$(A + B)(\mathbf{x}) = A(\mathbf{x}) + B(\mathbf{x}) \tag{2–6}$$

for all $\mathbf{x}$ in $\mathcal{V}_1$.

This, of course, is just the familiar addition of functions here applied to linear transformations, and it is an easy matter to show that $A + B$ is again a linear transformation from $\mathcal{V}_1$ to $\mathcal{V}_2$. Indeed, if $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to $\mathcal{V}_1$, and α_1 and α_2 are scalars, then

$$\begin{aligned}
(A + B)(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) &= A(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) + B(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2)\\
&= \alpha_1 A(\mathbf{x}_1) + \alpha_2 A(\mathbf{x}_2) + \alpha_1 B(\mathbf{x}_1) + \alpha_2 B(\mathbf{x}_2)\\
&= \alpha_1[A(\mathbf{x}_1) + B(\mathbf{x}_1)] + \alpha_2[A(\mathbf{x}_2) + B(\mathbf{x}_2)]\\
&= \alpha_1(A + B)(\mathbf{x}_1) + \alpha_2(A + B)(\mathbf{x}_2).
\end{aligned}$$

Thus $A + B$ satisfies Eq. (2–5), and is therefore linear, as asserted.

EXAMPLE 1. Let D and D^2 denote, respectively, the operations of taking first and second derivatives in $\mathcal{C}^2[a, b]$. Then the sum $D^2 + D$ is the linear transformation from $\mathcal{C}^2[a, b]$ to $\mathcal{C}[a, b]$ which sends each function $\mathbf{y}$ in $\mathcal{C}^2[a, b]$ onto the continuous function $\mathbf{y}'' + \mathbf{y}'$; that is,

$$(D^2 + D)(\mathbf{y}) = D^2\mathbf{y} + D\mathbf{y}.$$

EXAMPLE 2. Let $\mathbf{K}(t)$ be a fixed function in $\mathcal{C}[a, b]$, and let A be the linear transformation mapping $\mathcal{C}[a, b]$ into itself given by

$$A(\mathbf{f}) = \int_a^x \mathbf{K}(t)\mathbf{f}(t)\,dt, \qquad a \leq x \leq b.$$

Then the sum $A + I$, I the identity transformation on $\mathcal{C}[a, b]$, is the linear transformation mapping $\mathcal{C}[a, b]$ into itself whose defining equation is

$$(A + I)(\mathbf{f}) = \int_a^x \mathbf{K}(t)\mathbf{f}(t)\,dt + \mathbf{f}.$$

The addition of linear transformations defined above has a number of familiar and suggestive properties. In the first place, it is clear that

$$A + (B + C) = (A + B) + C \tag{2-7}$$

and

$$A + B = B + A \tag{2-8}$$

whenever A, B, and C are linear transformations from $\mathcal{V}_1$ to $\mathcal{V}_2$. Secondly, the zero mapping from $\mathcal{V}_1$ to $\mathcal{V}_2$ acts as a "zero" for this addition since

$$A + O = O + A = A \tag{2-9}$$

for all $A: \mathcal{V}_1 \to \mathcal{V}_2$. And finally, if A is any linear transformation from $\mathcal{V}_1$ to $\mathcal{V}_2$, and if we define $-A$ by the equation

$$(-A)(\mathbf{x}) = -A(\mathbf{x}) \tag{2-10}$$

for all $\mathbf{x}$ in $\mathcal{V}_1$, we obtain a linear transformation from $\mathcal{V}_1$ to $\mathcal{V}_2$ with the property that

$$A + (-A) = -A + A = O. \tag{2-11}$$

In short, the addition of linear transformations from $\mathcal{V}_1$ to $\mathcal{V}_2$ satisfies all of the axioms postulated for addition in a vector space.

To complete what should by now be an obvious sequence of ideas we introduce a scalar multiplication on the set of linear transformations from $\mathcal{V}_1$ to $\mathcal{V}_2$. The relevant definition is as follows:

Definition 2–3. The *product* of a real number α and a linear transformation $A: \mathcal{V}_1 \to \mathcal{V}_2$ is the mapping αA from $\mathcal{V}_1$ to $\mathcal{V}_2$ given by

$$(\alpha A)(\mathbf{x}) = \alpha A(\mathbf{x}) \tag{2-12}$$

for all $\mathbf{x}$ in $\mathcal{V}_1$. In other words, αA is the function whose value at $\mathbf{x}$ is computed by forming the scalar product of α and the vector $A(\mathbf{x})$.

We omit the proof that αA is linear, as well as the easy sequence of arguments required to show that the remaining axioms in the definition of a real vector space are now satisfied. Granting the truth of these facts, we have

Theorem 2–1. *The set of linear transformations from $\mathcal{V}_1$ to $\mathcal{V}_2$ is itself a real vector space under the definitions of addition and scalar multiplication given above.*

Under suitable hypotheses, it is also possible to define a multiplication of transformations. And, as we shall see, this single fact makes their study much richer in content and quite different in spirit from that of vector spaces alone.

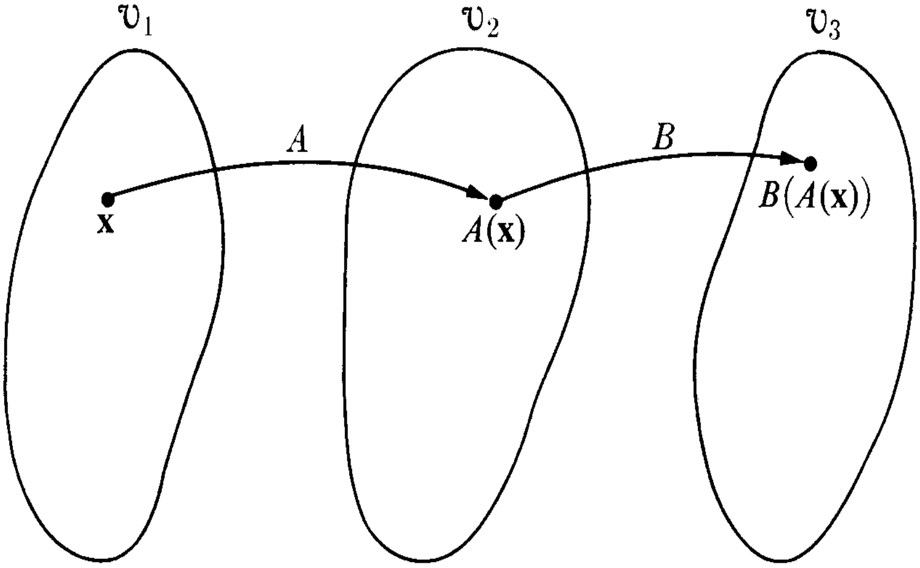


FIGURE 2–5

To introduce this multiplication, let $\mathcal{V}_1$, $\mathcal{V}_2$, $\mathcal{V}_3$ be vector spaces, and consider a pair of linear transformations

$$A: \mathcal{V}_1 \to \mathcal{V}_2, \qquad B: \mathcal{V}_2 \to \mathcal{V}_3.$$

Then, for each $\mathbf{x}$ in $\mathcal{V}_1$, $A(\mathbf{x})$ is a vector in $\mathcal{V}_2$, and it therefore makes sense to speak of applying B to $A(\mathbf{x})$ to obtain the vector $B(A(\mathbf{x}))$ in $\mathcal{V}_3$ (see Fig. 2–5). Thus A and B can be combined, or multiplied, to produce a function from $\mathcal{V}_1$ to $\mathcal{V}_3$ which will be denoted by BA, and called the *product* of A and B in that order, viz., *first A, then B.* This is the content of

Definition 2–4. If $A: \mathcal{V}_1 \to \mathcal{V}_2$ and $B: \mathcal{V}_2 \to \mathcal{V}_3$ are linear transformations, then their *product*, BA, is the mapping from $\mathcal{V}_1$ to $\mathcal{V}_3$ defined by the equation

$$BA(\mathbf{x}) = B(A(\mathbf{x})) \tag{2–13}$$

for all $\mathbf{x}$ in $\mathcal{V}_1$.

The essential fact about such products is that they are always linear. Indeed, if $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to $\mathcal{V}_1$, and α_1 and α_2 are arbitrary real numbers, then

$$\begin{aligned} BA(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) &= B[A(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2)] \\ &= B[\alpha_1 A(\mathbf{x}_1) + \alpha_2 A(\mathbf{x}_2)] \\ &= \alpha_1 B\big(A(\mathbf{x}_1)\big) + \alpha_2 B\big(A(\mathbf{x}_2)\big) \\ &= \alpha_1 BA(\mathbf{x}_1) + \alpha_2 BA(\mathbf{x}_2). \end{aligned}$$

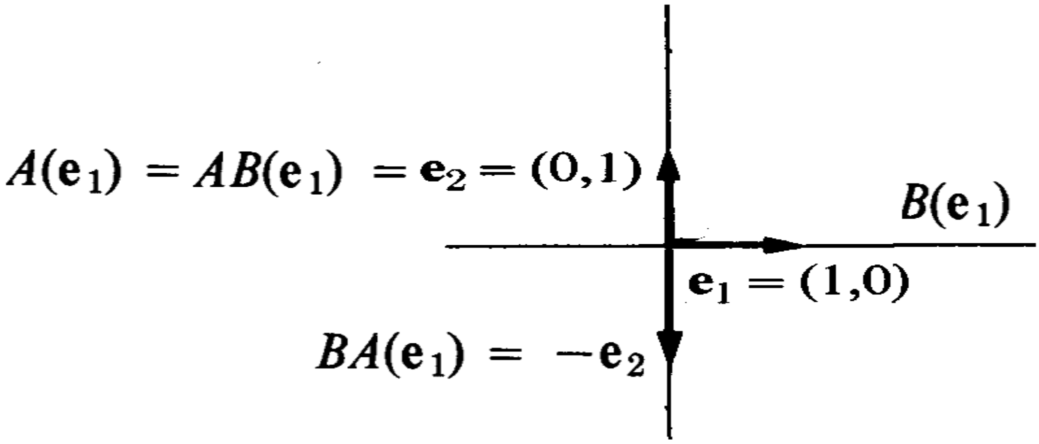
Hence BA satisfies (2–5), and its linearity has been established.

Having established the convention that the symbol BA stands for the product of A and B, in that order, we observe that this product is defined *only* when the image of A is contained in the domain of B. Thus one of the products AB or BA may exist and the other not, a phenomenon which will reappear later when we introduce the subject of matrices. But even when both A and B map a given vector space into itself, in which case AB and BA are linear transformations on the same space, it is by no means true that they must be equal. A simple example of this disturbing fact can be given in $\mathcal{R}^2$ by letting A be a counterclockwise rotation of 90° about the origin, and B a reflection across any line through the origin, say the x-axis. Then, with $\mathbf{e}_1$ and $\mathbf{e}_2$ the standard basis vectors,

$$AB(\mathbf{e}_1) = \mathbf{e}_2 \text{ while } BA(\mathbf{e}_1) = -\mathbf{e}_2, \text{ and } AB \neq BA.$$

In short, *the multiplication of linear transformations is noncommutative.*

$$AB(\mathbf{e}_1) = \mathbf{e}_2 \qquad\qquad BA(\mathbf{e}_1) = -\mathbf{e}_2$$



The foregoing example illustrates one of the ways in which this multiplication differs from "ordinary" multiplication. Why then call it multiplication at all? The answer is provided by the following identities which show that *most* of the properties usually associated with the term multiplication are still valid when phrased in terms of linear transformations. Specifically, assuming that all of the indicated products are defined, we have

$$A(BC) = (AB)C, \tag{2–14}$$

$$(A_1 + A_2)B = A_1B + A_2B, \qquad A(B_1 + B_2) = AB_1 + AB_2, \tag{2–15}$$

$$(\alpha A)B = A(\alpha B) = \alpha(AB), \quad \alpha \text{ a scalar}, \tag{2–16}$$

$$\begin{matrix} AI = A, \\ IA = A, \end{matrix} \quad I \text{ the identity map.} \tag{2–17}$$

The first of these identities asserts that the multiplication of linear transformations is *associative*, the next two that it is *distributive* over addition, and the fourth that it commutes with the operation of scalar multiplication. Finally, (2–17) implies that the identity transformation plays the same role in operator multiplication that the number one plays in arithmetic. Two different identity maps are usually involved here, and, strictly speaking, if $A\colon \mathcal{V}_1 \to \mathcal{V}_2$, then (2–17) ought to be written

$$AI_{\mathcal{V}_1} = A,$$

$$I_{\mathcal{V}_2}A = A,$$

where $I_{\mathcal{V}_1}$ denotes the identity map on $\mathcal{V}_1$, $I_{\mathcal{V}_2}$ the identity map on $\mathcal{V}_2$. But this notation is rarely used since the meaning of the unidentified symbol I is always clear from the context.

The proof of each of the above identities is an easy exercise in the definitions of the operations involved. Thus to establish (2–14) suppose that $C\colon \mathcal{V}_1 \to \mathcal{V}_2$, $B\colon \mathcal{V}_2 \to \mathcal{V}_3$, and $A\colon \mathcal{V}_3 \to \mathcal{V}_4$. Then each of the products $A(BC)$ and $(AB)C$ is a linear transformation from $\mathcal{V}_1$ to $\mathcal{V}_4$, and to prove their equality we simply apply Definition 2–4, twice for each product. This gives

$$[A(BC)](\mathbf{x}) = A\big(BC(\mathbf{x})\big) = A\Big(B\big(C(\mathbf{x})\big)\Big),$$

and

$$[(AB)C](\mathbf{x}) = AB\big(C(\mathbf{x})\big) = A\Big(B\big(C(\mathbf{x})\big)\Big),$$

and (2–14) now follows from the equality of the right-hand sides of these expressions. The remaining proofs have been left to the exercises.

EXAMPLE 1. *Powers of a linear transformation.* If A is a linear transformation on a fixed vector space $\mathcal{V}$ (i.e., $A\colon \mathcal{V} \to \mathcal{V}$) we can form the product of A with itself any finite number of times, thereby obtaining a sequence of linear transformations on $\mathcal{V}$ known as the *powers* of A. The associativity of operator multiplication implies that each of these powers is independent of the grouping of its factors and hence can be denoted without ambiguity by A^n, n a positive integer. Thus

$$A^1 = A,\ A^2 = AA,\ A^3 = AA^2, \ldots .$$

In addition, it is customary to let A^0 denote the identity transformation on $\mathcal{V}$, i.e., $A^0 = I$, so that all of the familiar rules for manipulating (nonnegative) exponents become valid. In particular, we have

$$A^m A^n = A^n A^m = A^{m+n},$$
$$(A^m)^n = A^{mn}$$

for all nonnegative integers m and n.

EXAMPLE 2. Let D be the differentiation operator on the space of polynomials $\mathcal{P}_n$. Then D is a linear transformation mapping $\mathcal{P}_n$ into itself, and its powers are simply the derivatives of orders two, three, etc. Since differentiation lowers the degree of every nonzero polynomial by one, the nth power of D maps *every* polynomial in $\mathcal{P}_n$ onto zero, and D^n is the zero transformation on $\mathcal{P}_n$. However, if $n > 1$, then D^{n-1}, and hence D itself, is certainly different from zero, and we have therefore shown that *a power of a nonzero linear transformation may be zero.*

In general, a nonzero linear transformation $A: \mathcal{V} \to \mathcal{V}$ with the property that $A^n = 0$ for some $n > 1$ is said to be *nilpotent* on $\mathcal{V}$, and the smallest integer n such that $A^n = 0$ is called the *degree of nilpotence* of A. We call attention to the fact that the property of being nilpotent actually depends upon the vector space under consideration as well as the linear transformation involved.

EXAMPLE 3. *Polynomials in A.* If A is a linear transformation on a vector space $\mathcal{V}$, we can use the powers of A together with the operations of addition and scalar multiplication to form polynomials in A. Thus if

$$p(x) = a_0 + a_1x + \cdots + a_nx^n$$

is a polynomial in x with real coefficients, we define $p(A)$ to be the linear transformation on $\mathcal{V}$ obtained by substituting A for x in $p(x)$. In other words,

$$p(A) = a_0I + a_1A + \cdots + a_nA^n,$$

or

$$p(A) = a_0 + a_1A + \cdots + a_nA^n,$$

the factor I being understood in the first term of this expression just as $x^0 = 1$ is understood in $p(x)$. Hence if $\mathbf{x}$ is any vector in $\mathcal{V}$,

$$p(A)(\mathbf{x}) = a_0\mathbf{x} + a_1A(\mathbf{x}) + \cdots + a_nA^n(\mathbf{x}).$$

Multiplicatively, these polynomials obey all of the familiar rules of polynomial algebra with the single exception that products can sometimes vanish without any of their factors vanishing, as was shown in the example above. In particular, *the multiplication of polynomials in a linear transformation is commutative* since the identity $p(x)q(x) = q(x)p(x)$ for "ordinary" polynomials p and q implies that $p(A)q(A) = q(A)p(A)$. This, in turn, implies that such polynomials can be factored, for, as the reader will remember, factorization of polynomials depends only on the commutativity of multiplication and its distributivity over addition.

Example 4. Let $\mathcal{C}^\infty[a, b]$ denote the space of all infinitely differentiable functions defined on the interval $[a, b]$, and again let D be differentiation. Then D maps $\mathcal{C}^\infty[a, b]$ into itself, and we can therefore form polynomials in D, which in this setting are expressions of the type

$$a_n D^n + a_{n-1} D^{n-1} + \cdots + a_1 D + a_0,$$

$a_0, \ldots, a_n$ real numbers. (Such expressions are known as *constant coefficient linear differential operators*, and it should be observed that they can also be interpreted as linear transformations from $\mathcal{C}^n[a, b]$ to $\mathcal{C}[a, b]$.) The polynomial $D^2 + D - 2$ is a typical example, and if y is any function in $\mathcal{C}^\infty[a, b]$ (or $\mathcal{C}^2[a, b]$), then

$$(D^2 + D - 2)y = \frac{d^2y}{dx^2} + \frac{dy}{dx} - 2y.$$

By virtue of the remarks made in the preceding example, we know that $D^2 + D - 2$ may be rewritten in either of the equivalent forms $(D + 2)(D - 1)$ or $(D - 1)(D + 2)$, and in this case it is easy to verify directly that these factorizations are correct. Indeed,

$$\begin{aligned}(D + 2)(D - 1)y &= (D + 2)[(D - 1)y] \\ &= (D + 2)\left(\frac{dy}{dx} - y\right) \\ &= \frac{d}{dx}\left(\frac{dy}{dx} - y\right) + 2\left(\frac{dy}{dx} - y\right) \\ &= \frac{d^2y}{dx^2} + \frac{dy}{dx} - 2y \\ &= (D^2 + D - 2)y,\end{aligned}$$

while a similar calculation yields the equality $(D - 1)(D + 2) = D^2 + D - 2$.

EXAMPLE 5. Let A and B denote the linear transformations mapping $\mathcal{C}^\infty[a, b]$ into itself defined by

$$A = xD + 1, \qquad B = D - x;$$

that is,

$$A(y) = x\frac{dy}{dx} + y, \qquad B(y) = \frac{dy}{dx} - xy$$

for each y in $\mathcal{C}^\infty[a, b]$. Then

$$\begin{aligned} AB(y) &= (xD + 1)[(D - x)y] \\ &= (xD + 1)\left(\frac{dy}{dx} - xy\right) \\ &= x\frac{d}{dx}\left(\frac{dy}{dx} - xy\right) + \frac{dy}{dx} - xy \\ &= x\frac{d^2y}{dx^2} + (1 - x^2)\frac{dy}{dx} - 2xy \\ &= [xD^2 + (1 - x^2)D - 2x]y, \end{aligned}$$

and hence

$$(xD + 1)(D - x) = xD^2 + (1 - x^2)D - 2x. \tag{2–18}$$

On the other hand,

$$
\begin{aligned}
BA(y) &= (D - x)\left(x\frac{dy}{dx} + y\right) \\
&= \frac{d}{dx}\left(x\frac{dy}{dx} + y\right) - x\left(x\frac{dy}{dx} + y\right) \\
&= x\frac{d^2y}{dx^2} + (2 - x^2)\frac{dy}{dx} - xy \\
&= [xD^2 + (2 - x^2)D - x]y,
\end{aligned}
$$

and hence

$$
(D - x)(xD + 1) = xD^2 + (2 - x^2)D - x. \tag{2–19}
$$

Comparing these results, we see that

$$
(xD + 1)(D - x) \neq (D - x)(xD + 1),
$$

and we have another illustration of the noncommutativity of operator multiplication. The reader should note that in this case neither the product AB nor BA can be evaluated by using the rules of elementary algebra. This is another of the unpleasant consequences of a noncommutative multiplication, and, as we shall see, has a decisive effect upon the study of linear differential equations.

THE NULL SPACE AND IMAGE; INVERSES

Let A be a linear transformation from $\mathcal{V}_1$ to $\mathcal{V}_2$, and let $\mathfrak{N}(A)$ denote the set of all $\mathbf{x}$ in $\mathcal{V}_1$ such that $A(\mathbf{x}) = \mathbf{0}$. Then, as we have already observed, $\mathfrak{N}(A)$ always contains the zero vector of $\mathcal{V}_1$. Actually we can say much more than this, for if $A(\mathbf{x}_1) = A(\mathbf{x}_2) = \mathbf{0}$, then

$$A(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) = \alpha_1 A(\mathbf{x}_1) + \alpha_2 A(\mathbf{x}_2) = \mathbf{0}$$

for all scalars α_1, α_2, and it follows that $\mathfrak{N}(A)$ *is a subspace of* $\mathcal{V}_1$. This subspace is called the *null space* or *kernel* of A, and is of fundamental importance in studying the behavior of A on $\mathcal{V}_1$.

Of equal importance with the null space of A is its image, $\mathfrak{I}(A)$, which, we recall, is the set of all $\mathbf{y}$ in $\mathcal{V}_2$ such that $\mathbf{y} = A(\mathbf{x})$ for some $\mathbf{x}$ in $\mathcal{V}_1$. It too is a subspace—this time in $\mathcal{V}_2$—since if $\mathbf{y}_1$ and $\mathbf{y}_2$ belong to $\mathfrak{I}(A)$ with $\mathbf{y}_1 = A(\mathbf{x}_1)$, $\mathbf{y}_2 = A(\mathbf{x}_2)$, then

$$A(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) = \alpha_1 A(\mathbf{x}_1) + \alpha_2 A(\mathbf{x}_2) = \alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2,$$

and $\alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2$ is also in the image of A, as required.

EXAMPLE 1. Let $I: \mathcal{V} \to \mathcal{V}$ be the identity transformation. Then $\mathfrak{N}(I) = \mathcal{O}$, the trivial subspace of $\mathcal{V}$, while $\mathfrak{I}(I) = \mathcal{V}$.

EXAMPLE 2. If $O: \mathcal{V}_1 \to \mathcal{V}_2$ is the zero transformation, then, by its very definition, $\mathfrak{N}(O) = \mathcal{V}_1$, $\mathfrak{I}(O) = \mathcal{O}$.

Example 3. Let D be the differentiation operator on the space of polynomials $\mathcal{P}_n$. Then the null space of D consists of all polynomials of degree zero together with the zero polynomial, while its image consists of the zero polynomial and all polynomials of degree $< n - 1$.

Example 4. Let $\mathcal{C}^2(-\infty, \infty)$ denote the space of all twice continuously differentiable functions on $(-\infty, \infty)$, and let $A: \mathcal{C}^2(-\infty, \infty) \to \mathcal{C}(-\infty, \infty)$ be the linear transformation $D^2 - I$. Then

$$A(y) = \frac{d^2y}{dx^2} - y,$$

and the null space of A is the set of all functions y in $\mathcal{C}^2(-\infty, \infty)$ for which

$$\frac{d^2y}{dx^2} - y = 0.$$

Thus, $\mathfrak{N}(A)$ is the set of solutions of a certain differential equation, and the problem of finding all solutions of this equation is identical with that of finding the null space of $D^2 - I$.

Example 5. Let $\mathcal{R}^\infty$ be the space of all infinite sequences $\{x_1, x_2, x_3, \ldots\}$ of real numbers, with addition and scalar multiplication defined termwise, and let A and B be the linear transformations on $\mathcal{R}^\infty$ defined by

$$A\{x_1, x_2, x_3, \ldots\} = \{x_2, x_3, x_4, \ldots\},$$
$$B\{x_1, x_2, x_3, \ldots\} = \{0, x_1, x_2, \ldots\}.$$

Then $\mathfrak{N}(A)$ is the subspace of $\mathcal{R}^\infty$ consisting of all sequences of the form $\{x_1, 0, 0, \ldots\}$, with x_1 arbitrary, while $\mathfrak{N}(B) = \mathcal{O}$. On the other hand, $\mathfrak{g}(A) = \mathcal{R}^\infty$, while, by definition, $\mathfrak{g}(B)$ consists of all sequences whose first entry is zero.

Now that we have introduced the null space and image of a linear transformation we propose to take a closer look at those transformations $A: \mathcal{V}_1 \to \mathcal{V}_2$ for which either

$$\text{(i)}\quad \mathfrak{N}(A) = \mathcal{O}, \qquad \text{or} \qquad \text{(ii)}\quad \mathfrak{g}(A) = \mathcal{V}_2,$$

or both. The second of these equations asserts that A maps $\mathcal{V}_1$ *onto* $\mathcal{V}_2$, and implies that for each $\mathbf{y}$ in $\mathcal{V}_2$ there exists at least one $\mathbf{x}$ in $\mathcal{V}_1$ such that $\mathbf{y} = A(\mathbf{x})$. The first, which says that the null space of A contains only the zero vector, turns out to be equivalent to the assertion that A is *one-to-one* in the sense of the following definition.

Definition 2–5. A linear transformation $A: \mathcal{V}_1 \to \mathcal{V}_2$ is said to be *one-to-one* if and only if $A(\mathbf{x}_1) = A(\mathbf{x}_2)$ implies that $\mathbf{x}_1 = \mathbf{x}_2$.

In other words, A is one-to-one if and only if A maps *distinct* vectors in $\mathcal{V}_1$ onto *distinct* vectors in $\mathcal{V}_2$; whence the name. (See Fig. 2–6.) This said, we now prove

Theorem 2–2. *A linear transformation $A: \mathcal{V}_1 \to \mathcal{V}_2$ is one-to-one if and only if $\mathfrak{N}(A) = \mathcal{O}$.*

Proof. Let A be one-to-one, and suppose that $A(\mathbf{x}) = \mathbf{0}$. Then $A(\mathbf{x}) = A(\mathbf{0})$, and Definition 2–5 implies that $\mathbf{x} = \mathbf{0}$. Thus $\mathfrak{N}(A) = \mathcal{O}$. Conversely, if $\mathfrak{N}(A) = \mathcal{O}$ and $A(\mathbf{x}_1) = A(\mathbf{x}_2)$, then $A(\mathbf{x}_1) - A(\mathbf{x}_2) = \mathbf{0}$, or $A(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0}$. Thus $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{0}$, and $\mathbf{x}_1 = \mathbf{x}_2$, as asserted. ▌

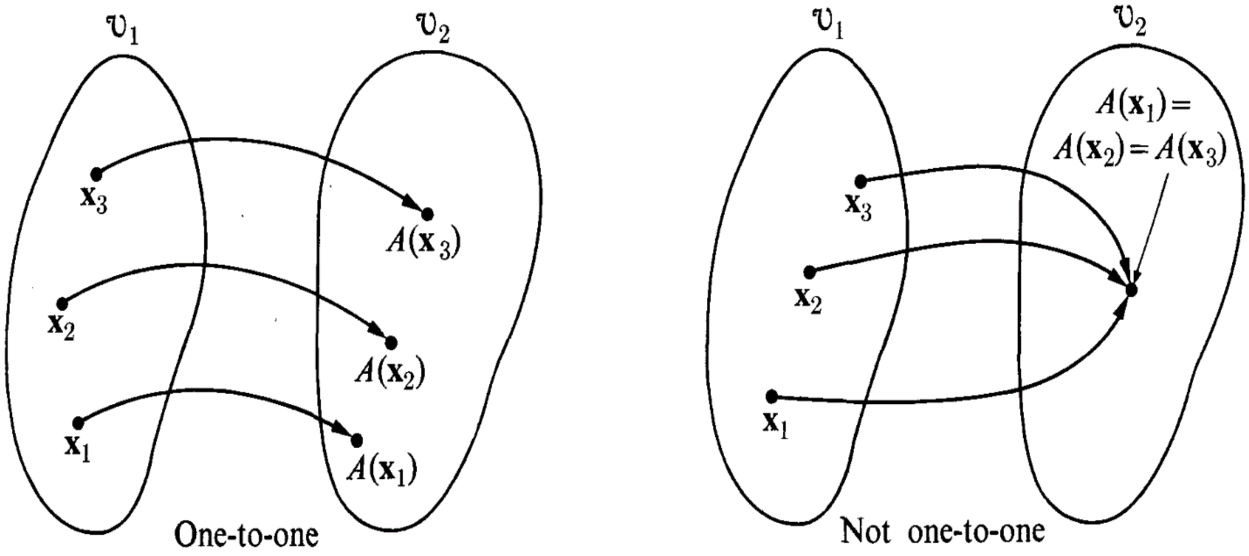


FIGURE 2–6

Linear transformations which are both one-to-one and onto are called *isomorphisms*, and are said to be *invertible*. They are of particular importance since, just as with ordinary one-to-one onto functions, they have inverses, and all of the standard facts concerning inverse functions can then be established. Indeed, if $A: \mathcal{V}_1 \to \mathcal{V}_2$ is one-to-one and onto, *each* vector $\mathbf{y}$ in $\mathcal{V}_2$ is paired with a *unique* vector $\mathbf{x}$ in $\mathcal{V}_1$, and A can therefore be used to define a function from $\mathcal{V}_2$ to $\mathcal{V}_1$. This function is called the *inverse* of A, and is denoted by A^{-1} (read "A inverse"). It can be described explicitly as *the* function from $\mathcal{V}_2$ to $\mathcal{V}_1$ such that

$$A^{-1}(\mathbf{y}) = \mathbf{x}, \quad \text{where} \quad A(\mathbf{x}) = \mathbf{y} \tag{2-20}$$

for each $\mathbf{y}$ in $\mathcal{V}_2$. Loosely speaking, A^{-1} is obtained from A by reading the definition of A from right to left, as suggested in Fig. 2–7, and is clearly a one-to-one map of $\mathcal{V}_2$ onto $\mathcal{V}_1$. Moreover, it is linear, since if $\mathbf{y}_1$ and $\mathbf{y}_2$ belong to $\mathcal{V}_2$ with $\mathbf{y}_1 = A(\mathbf{x}_1)$, $\mathbf{y}_2 = A(\mathbf{x}_2)$, then

$$A(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) = \alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2,$$

and (2–20) implies that

$$A^{-1}(\alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2) = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 = \alpha_1 A^{-1}(\mathbf{y}_1) + \alpha_2 A^{-1}(\mathbf{y}_2).$$

Having observed that A^{-1} is one-to-one, onto, and linear, it follows that it too is invertible, and if we simply parrot the construction given above, this time starting with A^{-1}, we find that $(A^{-1})^{-1} = A$. Finally, if we form the products $A^{-1}A$ and AA^{-1}, each of which is certainly defined, an easy argument reveals that they both reduce to the identity; that is,

$$A^{-1}A(\mathbf{x}) = \mathbf{x} \quad \text{and} \quad AA^{-1}(\mathbf{y}) = \mathbf{y}$$

for all $\mathbf{x}$ in $\mathcal{V}_1$ and all $\mathbf{y}$ in $\mathcal{V}_2$. And with this we have proved the following theorem.

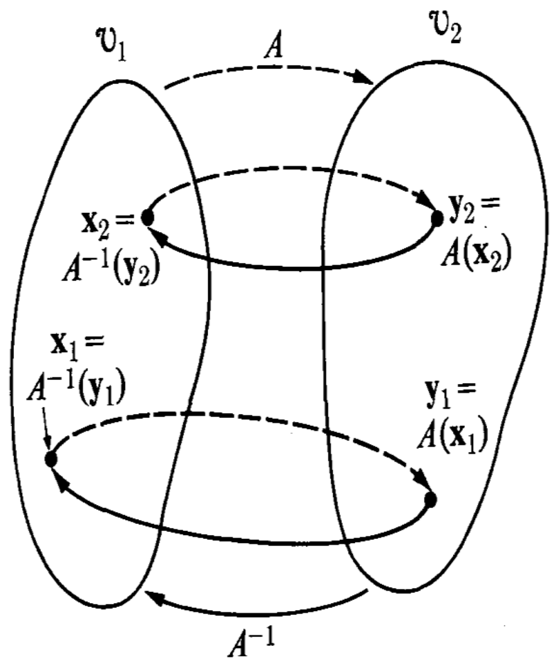


FIGURE 2–7

Theorem 2–3. *Every one-to-one linear transformation A mapping $\mathcal{V}_1$ onto $\mathcal{V}_2$ has a unique inverse from $\mathcal{V}_2$ to $\mathcal{V}_1$ defined by*

$$A^{-1}(\mathbf{y}) = \mathbf{x},$$

where $A(\mathbf{x}) = \mathbf{y}$ for all $\mathbf{y}$ in $\mathcal{V}_2$. A^{-1} is also one-to-one, onto, and linear, with $(A^{-1})^{-1} = A$, and

$$A^{-1}A = I_{\mathcal{V}_1}, \qquad AA^{-1} = I_{\mathcal{V}_2}, \tag{2–21}$$

where $I_{\mathcal{V}_1}$ and $I_{\mathcal{V}_2}$ denote, respectively, the identity maps on $\mathcal{V}_1$ and $\mathcal{V}_2$.

These last equations actually serve to characterize invertible linear transformations—a fact which when stated precisely reads as follows:

Theorem 2–4. *Let $A\colon \mathcal{V}_1 \to \mathcal{V}_2$ and $B\colon \mathcal{V}_2 \to \mathcal{V}_1$ be linear, and suppose that BA and AB are, respectively, the identity maps on $\mathcal{V}_1$ and $\mathcal{V}_2$. Then A is one-to-one and onto, and $B = A^{-1}$.*

Proof. Let $\mathbf{x}$ in $\mathcal{V}_1$ be such that $A(\mathbf{x}) = \mathbf{0}$. Then on the one hand,

$$B(A(\mathbf{x})) = B(\mathbf{0}) = \mathbf{0},$$

and on the other,

$$B(A(\mathbf{x})) = BA(\mathbf{x}) = I(\mathbf{x}) = \mathbf{x}.$$

Thus $\mathbf{x} = \mathbf{0}$, and $\mathfrak{N}(A) = \mathcal{O}$.

Now let $\mathbf{y}$ be an arbitrary vector in $\mathcal{V}_2$. Then

$$\mathbf{y} = I(\mathbf{y}) = AB(\mathbf{y}) = A(B(\mathbf{y})),$$

and it follows that $\mathbf{y}$ is the image under A of the vector $B(\mathbf{y})$ in $\mathcal{V}_1$. Thus $\mathfrak{g}(A) = \mathcal{V}_2$, and we are done. ▌

EXAMPLE 6. If A is any rotation of $\mathcal{R}^2$ about the origin through an angle θ, then A is invertible with A^{-1} the rotation through $-\theta$, since $A^{-1}A = AA^{-1} = I$.

EXAMPLE 7. Let $A: \mathcal{R}^3 \to \mathcal{R}^3$ be defined by

$$A(x_1, x_2, x_3) = (x_1 + x_2, x_2, x_3).$$

Then A is invertible, with A^{-1} given by

$$A^{-1}(x_1, x_2, x_3) = (x_1 - x_2, x_2, x_3),$$

since

$$A^{-1}A(x_1, x_2, x_3) = (x_1, x_2, x_3) = AA^{-1}(x_1, x_2, x_3).$$

Theorem 2-4 suggests a natural and valuable generalization of the notion of the inverse of a linear transformation $A: \mathcal{V}_1 \to \mathcal{V}_2$; to wit, a linear transformation $B: \mathcal{V}_2 \to \mathcal{V}_1$ such that

$$AB = I, \qquad BA \neq I.$$

The fact that such transformations do exist can be seen by looking at Example 5 above where

$$AB\{x_1, x_2, x_3, \ldots\} = \{x_1, x_2, x_3, \ldots\},$$

and

$$BA\{x_1, x_2, x_3, \ldots\} = \{0, x_2, x_3, \ldots\}.$$

Transformations of this sort are encountered fairly often in certain types of problems and are therefore distinguished by name according to the following definition.

Definition 2–6. A linear transformation $B: \mathcal{V}_2 \to \mathcal{V}_1$ is said to be a *right inverse* for $A: \mathcal{V}_1 \to \mathcal{V}_2$ if the product AB is the identity map on $\mathcal{V}_2$. Similarly, B is said to be a *left inverse* for A if BA is the identity map on $\mathcal{V}_1$.

Remark. If B is a right (left) inverse for A, then A is a left (right) inverse for B.

The example given a moment ago shows that a linear transformation may have a right or left inverse without having an inverse. It is easy to show, however, that if A has *both* a right inverse B and a left inverse C, then A is invertible, and $B = C = A^{-1}$. For then

$$AB = I \quad \text{and} \quad CA = I,$$

and it follows that

$$C(AB) = CI = C, \quad (CA)B = IB = B,$$

and hence that $B = C$. Thus $AB = BA = I$, and the assertion that A is invertible with $B = C = A^{-1}$ now follows from Theorem 2–4.

EXAMPLE 8. Let $\mathcal{C}^\infty[a, b]$ be the space of infinitely differentiable functions on $[a, b]$, and let D and L be differentiation and integration, respectively; that is,

$$Dy = \frac{dy}{dx}, \quad Ly = \int_a^x y(t)\,dt.$$

Then

$$LD(y) = \int_a^x y'(t)\,dt = y(x) - y(a),$$

while

$$DL(y) = \frac{d}{dx}\int_a^x y(t)\,dt = y(x),$$

and it follows that $DL = I, LD \neq I$. In other words, the operation of integration on function spaces is only a right inverse, and not an inverse, for differentiation. It is this fact more than any other which motivated us to introduce the notions of right and left inverses in the first place.