

**An Introduction to Linear Algebra in  
Parallel Distributed Processing**

---

M. I. JORDAN

Many of the properties of the models described in this book are captured by the mathematics of linear algebra. This chapter serves as an introduction to linear algebra and is a good starting place for the reader who wishes to delve further into the models presented in other parts of the book. I will focus on the aspects of linear algebra most essential for the analysis of parallel distributed processing models, particularly the notions of a vector space, the inner product, and linearity. I will also discuss some simple PDP models, and show how their workings correspond to operations on vectors.

---

**VECTORS**

---

A vector is a useful way to describe a pattern of numbers. Consider for example the pattern of numbers that describe the age, height, and weight of an average person. Suppose that Joe is 37 years old, 72 inches tall, and weighs 175 pounds. This information can be summarized in a vector or ordered list of numbers. For each person, there is a corresponding vector, as in Figure 1A. Each vector has three components: age, height, and weight. There is no reason to limit ourselves

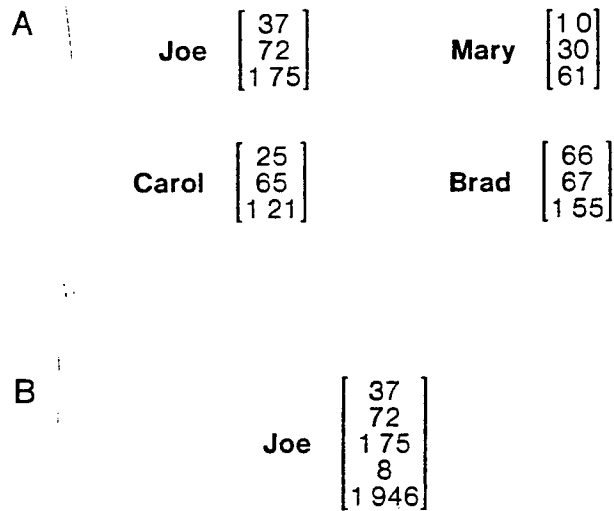


FIGURE 1.

to only three components, however. If, for example, we also wanted to keep track of Joe's shoe size and year of birth, then we would simply make a vector with five components, as in Figure 1B.

One important reason for the great utility of linear algebra lies in the simplicity of its notation. We will use bold, lower-case letters such as  $\mathbf{v}$  to stand for vectors. With this notation, an arbitrarily long list of information can be designated by a single symbol.

When a vector has no more than three components, it can be represented graphically by a point or an arrow in three-dimensional space. An example with three components is given in Figure 2 for the vector corresponding to Mary. Each axis in the figure corresponds to one of the three components of the vector.

It will prove helpful to try and visualize vectors as points or arrows in two- and three-dimensional space in proceeding through this chapter in order to develop geometric intuition for the operations on vectors. Notice, however, that there is no fundamental distinction between such vectors and vectors with more than three components. All of the operations upon vectors described in later sections apply equally well to vectors with any finite number of components.

In a parallel distributed processing model, many quantities are best represented by vectors. The pattern of numbers representing the activations of many processing units is one example. Other examples are the set of weights on the input lines to a particular processing unit, or the set of inputs to a system.

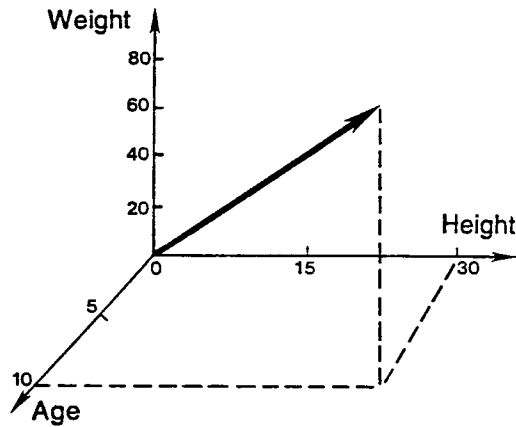


FIGURE 2.

## BASIC OPERATIONS

### Multiplication by Scalars

In linear algebra, a single real number is referred to as a *scalar*. A vector can be multiplied by a scalar by multiplying every component of the vector by the scalar.

*Examples:*

$$2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \qquad 5 \begin{bmatrix} -3 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} -15 \\ 20 \\ 5 \end{bmatrix}$$

Geometrically, scalar multiplication corresponds to lengthening or shortening the vector, while leaving it pointing in the same or opposite direction. As can be seen in Figure 3, multiplying a vector by 2 leaves it pointing in the same direction but twice as long. In general, multiplying a vector by a positive scalar produces a new vector that is longer or shorter by an amount corresponding to the magnitude of the scalar. Multiplication by a negative scalar produces a vector pointing in the opposite direction. It, too, is longer or shorter depending on the magnitude of the scalar. Two vectors that are scalar multiples of one another are said to be *collinear*.

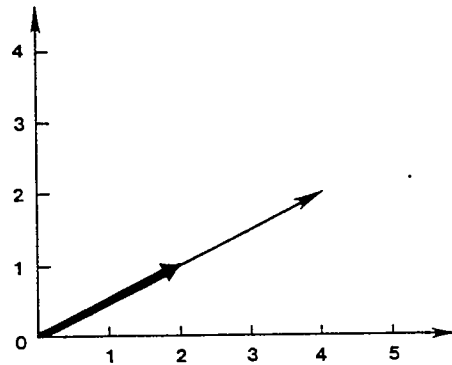


FIGURE 3.

### Addition of Vectors

Two or more vectors can be added by adding their components. The vectors must have the same number of components to be added; otherwise the operation is undefined.

*Examples:*

$$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \end{bmatrix}$$

Vector addition is associative (the vectors can be grouped in any manner) and commutative (the order of addition is unimportant) just like addition in ordinary algebra. This is true because if we consider one component at a time, vector addition *is* just addition in ordinary algebra.

How can vector addition be represented graphically? Consider Figure 4, where the vectors  $v_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$  and  $v_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$  are being added. It can be seen that the sum  $v_1 + v_2$  is a vector  $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$  which lies between  $v_1$  and  $v_2$ . Forming the parallelogram with sides  $v_1$  and  $v_2$ , we see that the sum of

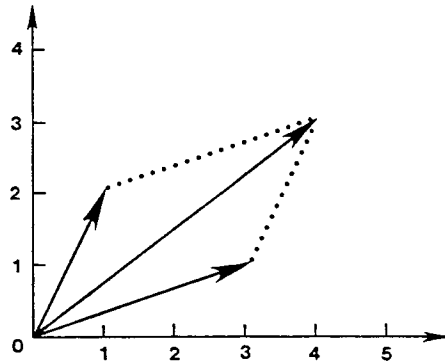


FIGURE 4.

the two vectors is the diagonal of this parallelogram. In two and three dimensions this is easy to visualize, but not when the vectors have more than three components. Nevertheless, it will be useful to imagine vector addition as forming the diagonal of a parallelogram. One implication of this view, which we will find useful, is that the sum of two vectors is a vector that lies in the same plane as the vectors being added.

*Example: Calculating averages.* We can demonstrate the use of the two operations thus far defined in calculating the average vector. Suppose we want to find the average age, height, and weight of the four individuals in Figure 1A. Clearly this involves summing the components separately and then dividing each sum by 4. Using vectors, this corresponds to adding the four vectors and then multiplying the resulting sum by the scalar  $1/4$ . Using  $\mathbf{u}$  to denote the average vector,

$$\mathbf{u} = \frac{1}{4} \left\{ \begin{bmatrix} 37 \\ 72 \\ 175 \end{bmatrix} + \begin{bmatrix} 10 \\ 30 \\ 61 \end{bmatrix} + \begin{bmatrix} 25 \\ 65 \\ 121 \end{bmatrix} + \begin{bmatrix} 66 \\ 67 \\ 155 \end{bmatrix} \right\} = \begin{bmatrix} 34.5 \\ 58.5 \\ 128 \end{bmatrix}.$$

Using vector notation, if we denote the four vectors by  $\mathbf{v}_1$ ,  $\mathbf{v}_2$ ,  $\mathbf{v}_3$ , and  $\mathbf{v}_4$ , then we can write the averaging operation as

$$\mathbf{u} = \frac{1}{4} (\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 + \mathbf{v}_4).$$

The vector  $\mathbf{u}$ , then, is a vector whose components are the averages of the components of the four individual vectors. Notice that the same result is obtained if each vector is first multiplied by  $1/4$ , and the resulting vectors are added. This shows that multiplication by scalars and vector addition obey a distributive law, as in ordinary algebra.

## LINEAR COMBINATIONS AND LINEAR INDEPENDENCE

### Linear Combinations of Vectors

The average vector calculated in the last section is an example of a *linear combination* of vectors. In this section, we pursue this idea further.

Consider the vectors  $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ ,  $\mathbf{v}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ , and  $\mathbf{u} = \begin{bmatrix} 9 \\ 10 \end{bmatrix}$ . Can  $\mathbf{u}$  be written as the sum of scalar multiples of  $\mathbf{v}_1$  and  $\mathbf{v}_2$ ? That is, can scalars  $c_1$  and  $c_2$  be found such that  $\mathbf{u}$  can be written in the form

$$\mathbf{u} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2?$$

If so, then  $\mathbf{u}$  is said to be a linear combination of the vectors  $\mathbf{v}_1$  and  $\mathbf{v}_2$ . The reader can verify that  $c_1 = 3$  and  $c_2 = 2$  will work, and thus  $\mathbf{u}$  is a linear combination of  $\mathbf{v}_1$  and  $\mathbf{v}_2$ .

This can also be seen directly in Figure 5, where these vectors are plotted. Remembering that multiplication by a scalar shortens or

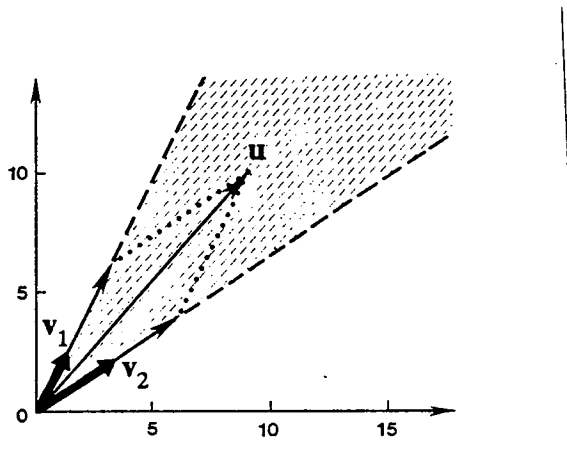


FIGURE 5.

lengthens a vector and that vector addition corresponds to forming the diagonal of a parallelogram, it seems clear that we can find scalars to adjust  $\mathbf{v}_1$  and  $\mathbf{v}_2$  to form a parallelogram that yields  $\mathbf{u}$ . This is indicated in the figure. It also seems clear that, using positive scalars, any vector in the shaded area of the figure can be generated this way. By using both negative and positive scalars, any vector in the plane can be written as a linear combination of  $\mathbf{v}_1$  and  $\mathbf{v}_2$ . This is true because multiplication by a negative scalar reverses the direction of a vector as well as shortening or lengthening it. The vectors  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are said to *span* the plane, because any vector in the plane can be generated from these two vectors.

In general, given a set  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  of vectors, a vector  $\mathbf{v}$  is said to be a linear combination of the  $\mathbf{v}_i$  if scalars  $c_1, c_2, \dots, c_n$  can be found such that

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_n\mathbf{v}_n. \quad (1)$$

The set of *all* linear combinations of the  $\mathbf{v}_i$  is called the set *spanned* by the  $\mathbf{v}_i$ .

*Example.* The three vectors  $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ ,  $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$  and  $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$  span all of three-dimensional space since any vector  $\mathbf{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$  can be written as a linear combination  $\mathbf{v} = a \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + c \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ . The vectors are referred to as the standard basis for three-dimensional space (more on the idea of a basis in the next section).

### Linear Independence

To say that a set of vectors span a space is to say that all vectors in the space can be generated from the original set by linear combination. We have shown examples in which two vectors span two-dimensional space and three vectors span three-dimensional space. We might be led to expect that, in general,  $n$  vectors suffice to span  $n$ -dimensional space. In fact, we have been using the term "dimension" without defining what it means; it would seem that a good definition of  $n$ -dimensional space is the set of vectors spanned by  $n$  vectors.

To make this definition work, we would require that the same size space be generated by any set of  $n$  vectors. However, this is not the case, as can be easily shown. Consider any pair of collinear vectors, for example. Such vectors lie along a single line, thus any linear combination of the vectors will lie along the same line. The space spanned by these two vectors is therefore only a one-dimensional set. The collinear vectors  $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$  and  $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$  are a good example. Any linear combination of these vectors will have equal components, thus they do not span the plane.

Another example is a set of three vectors that lie on a plane in three-dimensional space. Any parallelograms that we form will be in the same plane, thus all linear combinations will remain in the plane and we can't span all of three-dimensional space.

The general rule arising from these examples is that of a set of  $n$  vectors, if at least one can be written as a linear combination of the others, then the vectors span something less than a full  $n$ -dimensional space. We call such a set of vectors *linearly dependent*. If, on the other hand, none of the vectors can be written as a linear combination of the others, then the set is called *linearly independent*. We now revise the definition of dimensionality as follows:  $n$ -dimensional space is the set of vectors spanned by a set of  $n$  linearly independent vectors. The  $n$  vectors are referred to as a *basis* for the space.

*Examples:*

1.  $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$  and  $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$  are linearly dependent. They span only a one-dimensional space.
2.  $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$  and  $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$  are linearly independent. Thus they span the plane, a two-dimensional space.
3.  $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ ,  $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ , and  $\begin{bmatrix} -1 \\ 3 \end{bmatrix}$  are linearly dependent since 7 times the first vector minus 4 times the second vector is equal to the third vector.
4.  $\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$ ,  $\begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix}$ , and  $\begin{bmatrix} 9 \\ 10 \\ 0 \end{bmatrix}$  are linearly dependent. Clearly they cannot span all of three-dimensional space, because no vector with a nonzero third component can be generated from this set.



Notice the relationship between examples (2) and (3). The vectors in example (2) are linearly independent, therefore they span the plane. Thus any other vector with two components is a linear combination of these two vectors. In example (3), then, we know that the set will be linearly dependent before being told what the third vector is. This suggests the following rule: There can be no more than  $n$  linearly independent vectors in  $n$ -dimensional space.

A linearly independent set of vectors has the important property that a vector can be written as a linear combination of the set in only one way. In other words, the coefficients  $c_i$  in Equation 1 are unique if the vectors  $\mathbf{v}_i$  are linearly independent. This fact can be easily seen, for example, in the case of the standard basis, for there is only one vector in the basis which has a nonzero entry for any given component.

For linearly dependent vectors, however, the situation is different. If a vector can be written as a linear combination of a linearly dependent set of vectors, then there are an infinite number of sets of coefficients that will work. Let us attempt to demonstrate this fact with the aid of geometric intuition. Suppose that we wish to write vector  $\mathbf{v}$  as a linear combination of three vectors  $\mathbf{v}_1$ ,  $\mathbf{v}_2$ , and  $\mathbf{v}_3$  in the plane. Let us choose any arbitrary coefficient  $c_1$  for the vector  $\mathbf{v}_1$ . As shown in Figure 6, there must be a vector  $\mathbf{w}$  such that  $\mathbf{v} = c_1\mathbf{v}_1 + \mathbf{w}$ . Thus, if we can write  $\mathbf{w}$  as a linear combination of  $\mathbf{v}_2$  and  $\mathbf{v}_3$ , i.e.,  $\mathbf{w} = c_2\mathbf{v}_2 + c_3\mathbf{v}_3$ , then we have succeeded in writing  $\mathbf{v}$  as a linear combination of  $\mathbf{v}_1$ ,  $\mathbf{v}_2$ , and  $\mathbf{v}_3$ . But clearly we can do this, because  $\mathbf{w}$  is a vector in the plane, and  $\mathbf{v}_2$  and  $\mathbf{v}_3$  together span the plane.

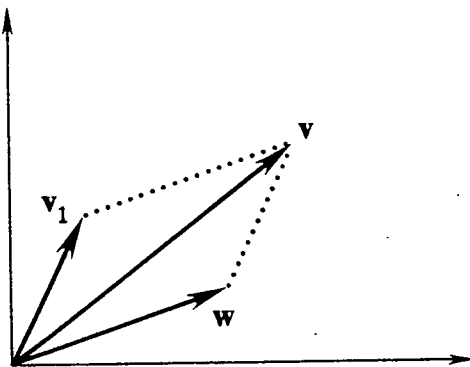


FIGURE 6.

## VECTOR SPACES

Let us pause to reflect for a moment upon what a vector is. I have implied that a vector is a list of numbers, and I have also used the term to refer to a point or an arrow in space. Are both of these objects vectors, or is one just a heuristic representation for the other? Are there other objects that should be called vectors? Just what is a vector?

As is often the case in mathematics, these kinds of questions are solved by being avoided. Consider the following definition of an abstract vector space, and try to decide what a vector is.

A *vector space* is a set  $V$  of elements, called vectors, with the following properties:

- To every pair,  $\mathbf{u}$  and  $\mathbf{v}$ , of vectors in  $V$ , there corresponds a vector  $\mathbf{u} + \mathbf{v}$  also in  $V$ , called the sum of  $\mathbf{u}$  and  $\mathbf{v}$ , in such a way that addition is commutative and associative.
- For any scalar  $c$  and any vector  $\mathbf{v}$  in  $V$ , there is a vector  $c\mathbf{v}$  in  $V$ , called the product of  $c$  and  $\mathbf{v}$ , in such a way that multiplication by scalars is associative and distributive with respect to vector addition.<sup>1</sup>

The answer to the question is that a vector is an undefined object in linear algebra, much like a line in geometry. The definition of a vector space simply lists the properties that vectors must have, without specifying what a vector must be. Thus, any set of objects that obey these properties can be called a vector space. Lists of numbers are vectors when addition is defined as adding components separately and scalar multiplication is defined as multiplying all the components by the scalar, because these operations fill all the requirements of a vector space. Arrows or points in space are also vectors when addition is defined geometrically as taking the diagonal of a parallelogram and scalar multiplication is defined as lengthening or shortening the arrow, because again, these operations fill the requirements of a vector space. A seemingly unrelated example of a vector space is the set of polynomials of order  $n$ , with addition and scalar multiplication defined in the obvious way.

This sort of abstraction is common in mathematics. It is useful because any theorem that is true about a general vector space must be

---

<sup>1</sup> I have left out certain technicalities usually included as axioms for a vector space. These include the axiom that there must be a zero vector, and for every vector, there is an additive inverse.

true about any instantiation of a vector space. We can therefore discuss general properties of vector spaces without being committed to choosing a particular representation such as a list of numbers. Much of the discussion about linear combinations and linear independence was of this nature.

When we do choose numbers to represent vectors, we use the following scheme. First we choose a basis for the space. Since every vector in the space can be written as a linear combination of the basis vectors, each vector has a set of coefficients  $c_1, c_2, \dots, c_n$  which are the coefficients in the linear combination. These coefficients are the numbers used as the components of the vector. As was shown in the previous section, the coefficients of a given vector are unique because basis vectors are linearly independent.

There is a certain arbitrariness in assigning the numbers, since there are infinitely many sets of basis vectors, and each vector in the space has a different description depending on which basis is used. That is, the coefficients, which are referred to as *coordinates*, are different for different choices of basis. The implications of this fact are discussed further in a later section where I also discuss how to relate the coordinates of a vector in one basis to the coordinates of the vector in another basis. Chapter 22 contains a lengthy discussion of several issues relating to the choice of basis.

## INNER PRODUCTS

As of yet, we have no way to speak of the length of a vector or of the similarity between two vectors. This will be rectified with the notion of an inner product.

The inner product of two vectors is the sum of the products of the vector components. The notation for the inner product of vectors  $\mathbf{v}$  and  $\mathbf{w}$  is  $\mathbf{v} \cdot \mathbf{w}$ . As with vector addition, the inner product is defined only if the vectors have the same number of components.

*Example:*

$$\mathbf{v} = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

$$\mathbf{v} \cdot \mathbf{w} = (3 \cdot 1) + (-1 \cdot 2) + (2 \cdot 1) = 3.$$

The inner product is a kind of multiplication between vectors, although somewhat of a strange sort of multiplication, since it produces a single number from a pair of vectors. What does this single number "measure"?

### Length

As a special case, consider taking the inner product of a vector with itself. An example is the vector  $\mathbf{v} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$  in Figure 7. The inner product of  $\mathbf{v}$  with itself is

$$\mathbf{v} \cdot \mathbf{v} = 3^2 + 4^2 = 25.$$

Consider the right triangle in Figure 7 with sides corresponding to the components of  $\mathbf{v}$ , and hypotenuse  $\mathbf{v}$  itself. The Pythagorean theorem tells us that the square of the length of  $\mathbf{v}$  is equal to the sum of the squares of the sides. Since this is exactly what is calculated by the inner product  $\mathbf{v} \cdot \mathbf{v}$ , it appears that a reasonable definition of the *length* of a vector is the square root of the inner product of the vector with itself. Thus we define the length of a vector  $\mathbf{v}$ , denoted by  $\|\mathbf{v}\|$ , as

$$\|\mathbf{v}\| = (\mathbf{v} \cdot \mathbf{v})^{1/2}.$$

Although the definition was motivated by an example in two dimensions, it can be applied to any vector. Notice that many of the

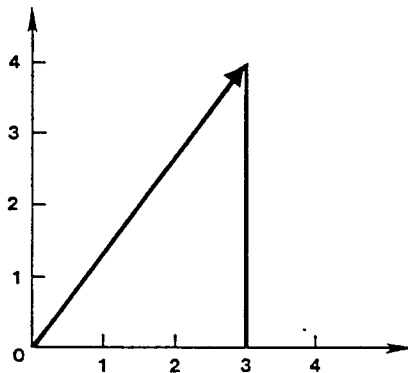


FIGURE 7.

properties we intuitively associate with length are included in this definition. For example, if a vector has larger components than another vector, it will be longer, because the squared components will contribute to a larger inner product. Multiplying a vector by a scalar produces a new vector whose length is the absolute value of the scalar times the length of the old vector:

$$\|c\mathbf{v}\| = |c| \|\mathbf{v}\|.$$

This is a property that can be easily proved. Somewhat harder to prove is the so-called triangle inequality, which states that the length of the sum of two vectors is less than or equal to the sum of the lengths of the two vectors:

$$\|\mathbf{v}_1 + \mathbf{v}_2\| \leq \|\mathbf{v}_1\| + \|\mathbf{v}_2\|.$$

Geometrically, the triangle inequality corresponds to the statement that one side of a triangle is no longer than the sum of the lengths of the other two sides.

Thus, in the special case where the operands are the same vector, the inner product is closely related to the idea of length. What if the operands are different vectors?

### Angle

The angle between two vectors  $\mathbf{v}$  and  $\mathbf{w}$  is defined in terms of the inner product by the following definition:

$$\cos \theta = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|} \quad (2)$$

where  $\theta$  is the angle between  $\mathbf{v}$  and  $\mathbf{w}$ . Note that all of the quantities on the right hand side of the equation are easily calculated for  $n$ -dimensional vectors. At the end of this section, I will show geometrically why this formula is correct in two-dimensional space, using the ordinary geometrical definition of angle.

*Example.* Find the angle  $\theta$  between the vectors  $\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$  and  $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ . First, we calculate the necessary inner product and lengths:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = 1 \quad \|\mathbf{v}_1\| = 1 \quad \|\mathbf{v}_2\| = \sqrt{2},$$

and then substitute these values in Equation 2:

$$\cos \theta = \frac{1}{1 \cdot \sqrt{2}} = 0.707.$$

Thus,

$$\theta = \cos^{-1} (0.707) = 45^\circ.$$

This result could also have been found using basic trigonometry, but clearly the inner product method is superior in general (consider finding the angle between vectors with forty components!).

The inner product is often said to measure the "match" or "similarity" between two vectors. In a vague sense, this seems to be the case from the definition of the inner product as the sum of products. Equation 2, however, shows this in a clearer way: Writing out the equation in terms of the components of the vectors gives

$$\cos \theta = \frac{\sum_{i=1}^n v_i w_i}{\left(\sum_{i=1}^n v_i^2\right)^{1/2} \left(\sum_{i=1}^n w_i^2\right)^{1/2}}.$$

This is the formula for the correlation between two sets of numbers with zero means.

We can use our geometrical intuitions about angles and our understanding of correlation to turn Equation 2 around and gain a better understanding of the inner product. This understanding is important for the analysis of PDP models, because as will be seen, PDP models often compute inner products. Let us imagine moving two vectors around in space like the hands on a clock. If we hold the lengths of the vectors constant, then Equation 2 says that the inner product is proportional to the cosine of the angle:  $\mathbf{v} \cdot \mathbf{w} = \|\mathbf{v}\| \|\mathbf{w}\| \cos \theta$ . For example, if the angle between the vectors is zero, where the cosine is at a maximum, the inner product must therefore be at a maximum. As the two vectors move farther apart, the cosine decreases, thus the inner product decreases. It reaches zero when the angle is  $90^\circ$ , and its most negative value when the angle between the vectors is  $180^\circ$ , that is, when the vectors point in opposite directions. Thus, the closer the two vectors are, the larger the inner product. The more the vectors point in opposite directions, the more negative the inner product.

We must be careful, however, in claiming that two vectors are closer together than two others because they have a larger inner product. We

must remember to divide the inner product by the lengths of the vectors involved to make such comparative statements.

An important special case occurs when the inner product is zero. In this case, the two vectors are said to be *orthogonal*. Plugging zero into the right side of Equation 2 gives

$$\cos \theta = 0.$$

which implies that the angle between the vectors is  $90^\circ$ . Thus, orthogonal vectors are vectors which lie at right angles to one another.

We will often speak of a set of orthogonal vectors. This means that every vector in the set is orthogonal to every other vector in the set. That is, every vector lies at a right angle to every other vector. A good example in three-dimensional space is the standard basis referred to earlier. Although we will skip the proof, it is probably clear that any orthogonal set is linearly independent. Indeed, orthogonality is stronger than linear independence: whereas every orthogonal set is linearly independent, there are very many linearly independent sets of vectors that are not orthogonal. An example in two-dimensional space is the pair  $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$  and  $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ . When we choose a basis for a space, we typically choose an orthogonal basis. In fact, in much of classical physics and mathematics, there is not the slightest hint that a basis should be anything but orthogonal.

## Projections

A further application of the inner product, closely related to the ideas of length and angle, is the notion of a projection of one vector onto another. An example is given in Figure 8. The distance  $x$  is the projection of  $\mathbf{v}$  on  $\mathbf{w}$ . In two dimensions, we readily know how to calculate the projection. It is

$$x = \|\mathbf{v}\| \cos \theta \quad (3)$$

where  $\theta$  is the angle between  $\mathbf{v}$  and  $\mathbf{w}$ . This formula generalizes, and for any vectors  $\mathbf{v}$  and  $\mathbf{w}$ , the projection of  $\mathbf{v}$  on  $\mathbf{w}$  is given by Equation 3. It is a scalar which can be thought of as indicating how much  $\mathbf{v}$  is pointing in the direction of  $\mathbf{w}$ .

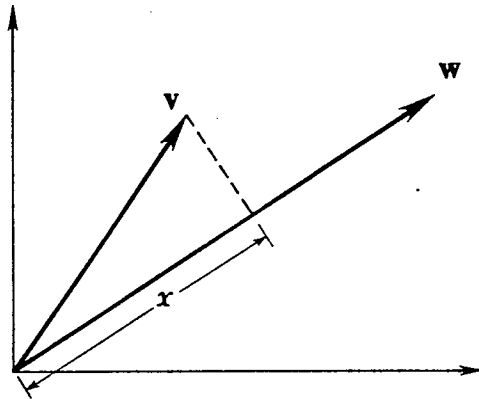


FIGURE 8.

There is a close relationship between the inner product and the projection. Using Equation 2, we can rewrite the formula for the projection:

$$\begin{aligned} x &= \|v\| \cos \theta \\ &= \|v\| \frac{v \cdot w}{\|v\| \|w\|} \\ &= \frac{v \cdot w}{\|w\|}. \end{aligned}$$

Thus, the projection is the inner product divided by the length of  $w$ . In particular, if  $w$  has length one, then  $\|w\| = 1$ , and the projection of  $v$  on  $w$  and the inner product of  $v$  and  $w$  are the same thing. This way of thinking about the inner product is consistent with our earlier comments. That is, if we hold the lengths of  $v$  and  $w$  constant, then we know that the inner product gets larger as  $v$  moves toward  $w$ . From the picture, we see that the projection gets larger as well. When the two vectors are orthogonal, the projection as well as the inner product are zero.

### Inner Products in Two Dimensions

Equation 2 can be shown to be correct in two-dimensional space with the help of some simple geometry. Let  $v$  and  $w$  be two vectors in the plane, and  $\theta$  be the angle between them, as shown in Figure 9. Denote the  $x$  and  $y$  coordinates of  $v$  and  $w$  by  $v_x$ ,  $v_y$  and  $w_x$ ,  $w_y$ , respectively.



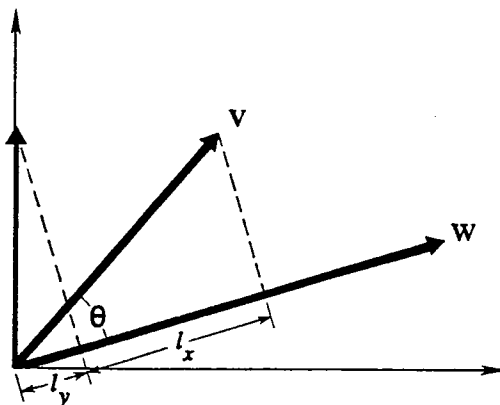


FIGURE 9.

Let  $l$  denote the projection of  $v$  on  $w$ . We have  $l = \|v\| \cos\theta$  from geometry. We can break  $l$  into two pieces  $l_x$  and  $l_y$  as shown in the figure.  $l_y$  can be computed from the diagram by noticing that triangles  $OAD$  and  $COB$ , in Figure 10, are similar triangles. Thus, the ratio of corresponding sides is constant:

$$\frac{l_y}{v_y} = \frac{w_y}{\|w\|},$$

giving

$$l_y = \frac{v_y w_y}{\|w\|}.$$

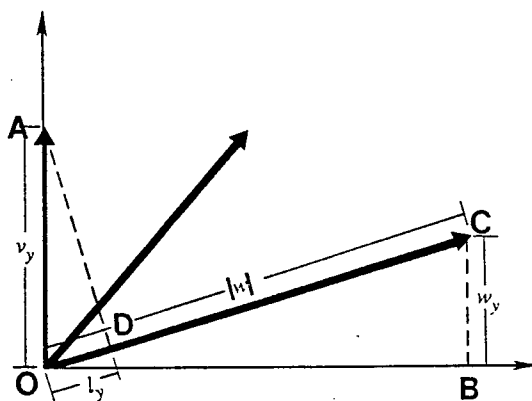


FIGURE 10.

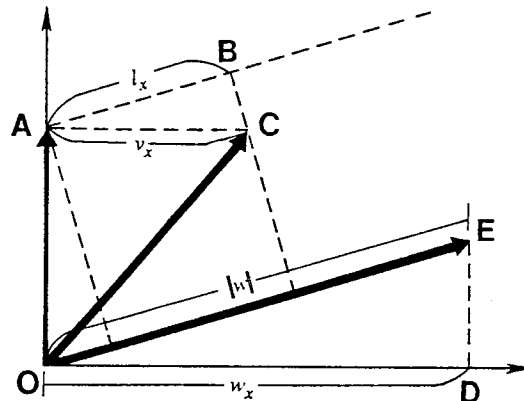


FIGURE 11.

In Figure 11, we see how to compute  $l_x$ , by observing that triangles EOD and CAB are similar. Thus,

$$\frac{l_x}{v_x} = \frac{w_x}{\|w\|},$$

giving

$$l_x = \frac{v_x w_x}{\|w\|}.$$

We can now write  $l = l_x + l_y$ , which yields

$$l = \|v\| \cos \theta = l_x + l_y = \frac{v_x w_x}{\|w\|} + \frac{v_y w_y}{\|w\|} = \frac{v \cdot w}{\|w\|}.$$

Thus,

$$\cos \theta = \frac{v \cdot w}{\|v\| \|w\|}.$$

### Algebraic Properties of the Inner Product

In this section, we collect together some useful algebraic theorems concerning inner products. Most of these theorems can be easily proved using the definition of the inner product and properties of real

numbers. In what follows,  $c$  and  $c_i$  will be any scalars, and the  $\mathbf{v}$  and  $\mathbf{w}$  will be  $n$ -dimensional vectors.

$$\mathbf{v} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{v} \quad (4)$$

$$c(\mathbf{v} \cdot \mathbf{w}) = (c\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (c\mathbf{w}) \quad (5)$$

$$\mathbf{w} \cdot (\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{w} \cdot \mathbf{v}_1 + \mathbf{w} \cdot \mathbf{v}_2 \quad (6)$$

The first theorem says simply that order is unimportant; the inner product is commutative. The second and third theorems show that the inner product is a *linear* function, as we will discuss at length in a later section. We can combine these two equations to get  $\mathbf{w} \cdot (c_1\mathbf{v}_1 + c_2\mathbf{v}_2) = c_1(\mathbf{w} \cdot \mathbf{v}_1) + c_2(\mathbf{w} \cdot \mathbf{v}_2)$ . It is also well worth our while to use mathematical induction to generalize this formula, giving us

$$\begin{aligned} \mathbf{w} \cdot (c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n) = \\ c_1(\mathbf{w} \cdot \mathbf{v}_1) + c_2(\mathbf{w} \cdot \mathbf{v}_2) + \cdots + c_n(\mathbf{w} \cdot \mathbf{v}_n). \end{aligned} \quad (7)$$

This important result tells us how to calculate the inner product of  $\mathbf{w}$  and a linear combination of vectors.

Another useful theorem is

$$|\mathbf{v} \cdot \mathbf{w}| \leq \|\mathbf{v}\| \|\mathbf{w}\| \quad (8)$$

This is known as the Cauchy-Schwartz inequality. It gives an upper bound on the inner product.

## ONE UNIT IN A PARALLEL DISTRIBUTED PROCESSING SYSTEM

In this section, we show how some of the concepts we have introduced can be used in analyzing a very simple PDP model. Consider the processing unit in Figure 12 which receives inputs from the  $n$  units below. Associated with each of the  $n + 1$  units there is a scalar *activation value*. We shall use the scalar  $u$  to denote the activation of the output unit and the vector  $\mathbf{v}$  to denote the activations of the  $n$  input units. That is, the  $i$ th component of  $\mathbf{v}$  is the activation of the  $i$ th input unit. Since there are  $n$  input units,  $\mathbf{v}$  is an  $n$ -dimensional vector.

Associated with each link between the input units and the output unit, there is a scalar weight value, and we can think of the set of  $n$

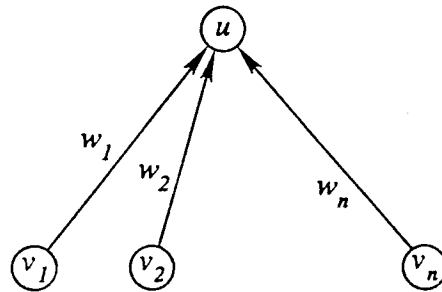


FIGURE 12.

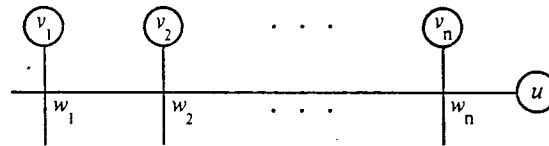


FIGURE 13.

weights as an  $n$ -dimensional vector  $w$ . This is the *weight vector* corresponding to the output unit. Later we will discuss a model with many output units, each of which will have its own weight vector.

Another way to draw the same model is shown in Figure 13. Here we have drawn the  $n$  input units at the top with the output unit on the right. The components of the weight vector are stored at the junctions where the vertical input lines meet the horizontal output line. Which diagram is to be preferred (Figure 12 or Figure 13) is mostly a matter of taste, although we will see that the diagram in Figure 13 generalizes better to the case of many output units.

Now to the operation of the model: Let us assume that the activation of each input unit is multiplied by the weight on its link, and that these products are added up to give the activation of the output unit. Using the definition of the inner product, we translate that statement into mathematics as follows:

$$u = w \cdot v.$$

The activation of the output unit is the inner product of its weight vector with the vector of input activations.

The geometric properties of the inner product give us the following picture to help in understanding what the model is computing. We imagine that the set of possible inputs to the model is a vector space. It is an  $n$ -dimensional space, where  $n$  is the number of input lines. The weight vector also has  $n$  components, thus we can plot the weight vector in the input space. The advantage of doing this is that we can now state how the system will respond to the various inputs. As we have seen, the inner product gives an indication of how close two vectors are. Thus, in this simple PDP model, the output activation gives an indication or measurement of how close the input vector is to the stored weight vector. The inputs lying close to the weight vector will yield a large positive response, those lying near  $90^\circ$  will yield a zero response, and those pointing in the opposite direction will yield a large negative response. If we present a succession of input vectors of constant length, the output unit will respond most strongly to that input vector which is closest to its weight vector, and will drop off in response as the input vectors move away from the weight vector.

One way to describe the functioning of the processing unit is to say that it splits the input space into two parts, the part where the response is negative and the part where the response is positive. We can easily imagine augmenting the unit in the following way: if the inner product is positive, output a 1; if the inner product is negative, output a 0. This unit, referred to as a *linear threshold unit*, explicitly computes which part of the space the input lies in.

In some models, the weight vector is assumed to be normalized, that is,  $\|\mathbf{w}\| = 1$ . As we have seen, in this case, the activation of the output unit is simply the projection of the input vector on the weight vector.

---

## MATRICES AND LINEAR SYSTEMS

---

The first section introduced the concepts of a vector space and the inner product. We have seen that vectors may be added together and multiplied by scalars. Vectors also have a length, and there is an angle between any pair of vectors. Thus, we have good ways of describing the structure of a set of vectors.

The usefulness of vectors can be broadened considerably by introducing the concept of a matrix. From an abstract point of view, matrices are a kind of "operator" that provide a mapping from one vector space

to another vector space. They are at the base of most of the models in this book which take vectors as inputs and yield vectors as outputs.

First, we will define matrices and show that they have an algebra of their own which is analogous to that of vectors. In particular, matrices can be added together and multiplied by scalars.

## MATRICES

A matrix is simply an array of real numbers. If the array has  $m$  rows and  $n$  columns, then we will refer to the matrix as an  $m \times n$  matrix. Capital letters will be used to denote matrices.

*Examples:*

$$\mathbf{M} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix} \quad \mathbf{N} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{P} = \begin{bmatrix} 10 & -1 \\ -1 & 27 \end{bmatrix}$$

$\mathbf{M}$  is a  $2 \times 3$  matrix,  $\mathbf{N}$  is a  $3 \times 3$  matrix, and  $\mathbf{P}$  is a  $2 \times 2$  matrix.

*Some special matrices.* There are several classes of matrices that are useful to identify. A *square* matrix is a matrix with the same number of rows and columns. The matrices  $\mathbf{N}$  and  $\mathbf{P}$  are examples of square matrices. A *diagonal* matrix is a square matrix that is zero everywhere except on its main diagonal. An example is matrix  $\mathbf{N}$ . A *symmetric* matrix is a square matrix whose  $i, j$ th element is equal to its  $j, i$ th element. Any diagonal matrix is symmetric. Matrix  $\mathbf{P}$  is an example of a symmetric matrix that is not diagonal. Finally, the diagonal matrix that has all ones on its main diagonal is referred to as the identity matrix, and is denoted  $\mathbf{I}$ .

### Multiplication by Scalars

A matrix can be multiplied by a scalar by multiplying every element in the matrix by that scalar.

*Example:*

$$3\mathbf{M} = 3 \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 12 & 15 \\ 3 & 0 & 3 \end{bmatrix}$$

### Addition of Matrices

Matrices are added together by adding corresponding elements. Only matrices that have the same number of rows and columns can be added together.

*Example:*

$$\mathbf{M} + \mathbf{N} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 2 \\ 4 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 7 \\ 5 & 1 & 0 \end{bmatrix}$$

Notice that there is a close relationship between these definitions and the corresponding definitions for vectors. In fact, for fixed integers  $m$  and  $n$ , the set of all  $m \times n$  matrices is another example of a vector space. However, we will not exploit this fact, rather, we will think about matrices in another way, in terms of functions from one vector space to another. This is the subject of the next section.

### Multiplication of a Vector by a Matrix

We now link up vectors and matrices by showing how a vector can be multiplied by a matrix to produce a new vector. Consider the matrix

$$\mathbf{W} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix} \text{ and the vector } \mathbf{v} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}. \text{ We wish to define a vector } \mathbf{u}$$

which is the product of  $\mathbf{W}$  and  $\mathbf{v}$ , and denoted

$$\mathbf{u} = \mathbf{W}\mathbf{v} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}.$$

To define this operation, first imagine breaking the matrix into its rows. Each row of the matrix is a list of three numbers. We can think of the row as a three-dimensional vector and speak of the *row vectors* of the matrix. There are two such row vectors. Now consider forming the inner products of each of these row vectors with the vector  $\mathbf{v}$ . This will yield two numbers. These two numbers can be thought of as a two-dimensional vector  $\mathbf{u}$ , which is defined to be the product  $\mathbf{W}\mathbf{v}$ .

*Example:*

$$\mathbf{u} = \mathbf{W}\mathbf{v} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \cdot 1 + 4 \cdot 0 + 5 \cdot 2 \\ 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 2 \end{bmatrix} = \begin{bmatrix} 13 \\ 3 \end{bmatrix}$$

The components of  $\mathbf{u}$  are the inner products of  $\mathbf{v}$  with the row vectors of  $\mathbf{W}$ .

For a general  $m \times n$  matrix  $\mathbf{W}$  and an  $n$ -dimensional vector  $\mathbf{v}$ ,<sup>2</sup> the product  $\mathbf{W}\mathbf{v}$  is an  $m$ -dimensional vector  $\mathbf{u}$ , whose elements are the inner products of  $\mathbf{v}$  with the row vectors of  $\mathbf{W}$ . As suggested by Figure 14, the  $i$ th component of  $\mathbf{u}$  is the inner product of  $\mathbf{v}$  with the  $i$ th row vector of  $\mathbf{W}$ . Thus, the multiplication of a vector by a matrix can be thought of as simply a shorthand way to write down a series of inner products of a vector with a set of other vectors. The vector  $\mathbf{u}$  tabulates the results. This way of thinking about the multiplication operation is a good way to conceptualize what is happening in a PDP model with many output units, as we will see in the next section.

There is another way of writing the multiplication operation that gives a different perspective on what is occurring. If we imagine breaking the matrix up into its columns, then we can equally well speak of the *column vectors* of the matrix. It can then be easily shown that the multiplication operation  $\mathbf{W}\mathbf{v}$  produces a vector  $\mathbf{u}$  that is a linear combination of the column vectors of  $\mathbf{W}$ . Furthermore, the coefficients of the linear combination are the components of  $\mathbf{v}$ . For example, letting  $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$  be the column vectors of  $\mathbf{W}$ , we have

$$\mathbf{u} = v_1\mathbf{w}_1 + v_2\mathbf{w}_2 + v_3\mathbf{w}_3 = 1 \begin{bmatrix} 3 \\ 1 \end{bmatrix} + 0 \begin{bmatrix} 4 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 13 \\ 3 \end{bmatrix}$$

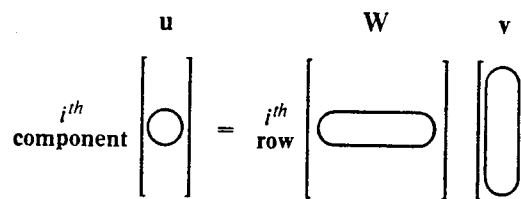


FIGURE 14.

<sup>2</sup> The dimensionality of  $\mathbf{v}$  must be equal to the number of columns of  $\mathbf{W}$  so that the inner products can be defined.



where the  $v_i$  are the components of  $\mathbf{v}$ . This way of viewing the multiplication operation is suggested in Figure 15 for a matrix with  $n$  columns.

If we let the term *column space* refer to the space spanned by the column vectors of a matrix, then we have the following interesting result: The vector  $\mathbf{u}$  is in the column space of  $\mathbf{W}$ .

Finally, it is important to understand what is happening on an abstract level. Notice that for each vector  $\mathbf{v}$ , the operation  $\mathbf{W}\mathbf{v}$  produces another vector  $\mathbf{u}$ . The operation can thus be thought of as a mapping or function from one set of vectors to another set of vectors. That is, if we consider an  $n$ -dimensional vector space  $\mathbf{V}$  (the domain) and an  $m$ -dimensional vector space  $\mathbf{U}$  (the range), then the operation of multiplication by a fixed matrix  $\mathbf{W}$  is a function from  $\mathbf{V}$  to  $\mathbf{U}$ , as shown in Figure 16. It is a function whose domain and range are both vector spaces.

### Algebraic Properties of Matrix Mapping

Several properties of matrix-vector multiplication follow directly from the properties of the inner product. In all cases, the number of

$$\begin{array}{ccc}
 \mathbf{W} & \mathbf{v} & \mathbf{u} \\
 \left[ \begin{array}{c|c|c} | & & | \\ \mathbf{w}_1 & \cdots & \mathbf{w}_n \\ | & & | \end{array} \right] & \left[ \begin{array}{c} v_1 \\ \cdot \\ \cdot \\ v_n \end{array} \right] & = \left[ \begin{array}{c} | \\ v_1\mathbf{w}_1 + \cdots + v_n\mathbf{w}_n \\ | \end{array} \right]
 \end{array}$$

FIGURE 15.

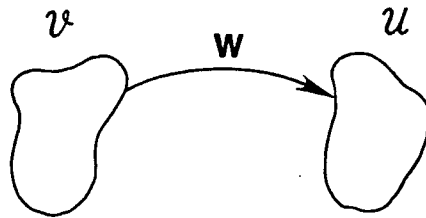


FIGURE 16.

components of the vector must be the same as the number of columns of the matrix.

$$W(av) = aWv \quad (9)$$

$$W(u + v) = Wu + Wv \quad (10)$$

These equations are the counterparts to Equations 5 and 6. As in that section, they can be combined and generalized to general linear combinations:

$$W(c_1v_1 + c_2v_2 + \dots + c_nv_n) = c_1(Wv_1) + c_2(Wv_2) + \dots + c_n(Wv_n) \quad (11)$$

In the next theorem, the matrices  $M$  and  $N$  must have the same number of rows and columns.

$$Mv + Nv = (M + N)v \quad (12)$$

### ONE LAYER OF A PARALLEL DISTRIBUTED PROCESSING SYSTEM

I now generalize the simple model presented earlier to show how matrices can be used in analyzing PDP models. Consider Figure 17, which is the generalization of Figure 12 to the case of many output units. Suppose that there are  $m$  output units, each one connected to all of the  $n$  input units. Denote the activation of the output units by  $u_1, u_2, \dots, u_m$ . Each output unit has its own weight vector  $w_i$ , separate from the other output units. As before, the activation rule

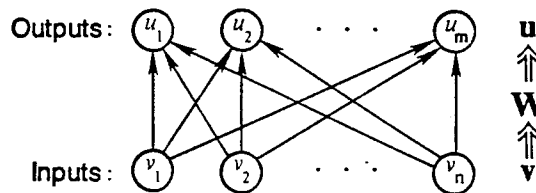


FIGURE 17.

says that the activation of an output unit is given by the inner product of its weight vector with the input vector, thus,

$$u_i = \mathbf{w}_i \cdot \mathbf{v}.$$

If we form a matrix  $\mathbf{W}$  whose row vectors are the  $\mathbf{w}_i$ , then we can use the rule for matrix-vector multiplication to write all of the computations at once. Let  $\mathbf{u}$  be the vector whose components are the  $u_i$ . Then

$$\mathbf{u} = \mathbf{W}\mathbf{v}.$$

This is a very succinct expression of the computation performed by the network. It says that for each input vector  $\mathbf{v}$ , the network produces an output vector  $\mathbf{u}$  whose components are the activations of the output units.

Another way to draw the network is shown in Figure 18, which is the generalization of Figure 13 to the case of many output units. At each junction in the diagram there is a weight connecting an input unit with an output unit.<sup>3</sup> The weight vectors associated with each output unit appear on the horizontal lines. When drawn this way, it is clear why a matrix appears in the equation linking the output vector to the input vector: The array of junctions in the diagram is exactly the weight matrix  $\mathbf{W}$ .

Now let us attempt to understand geometrically what is being computed by the model. Each output unit is computing the inner product

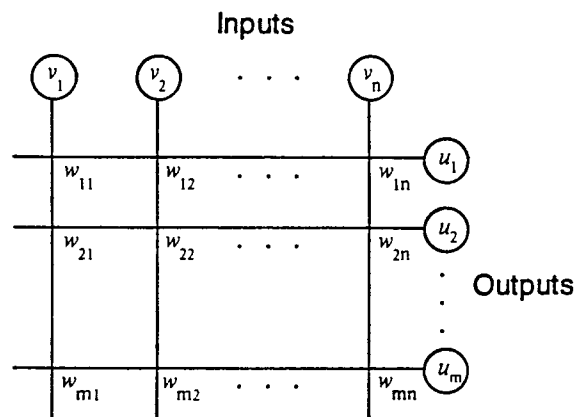


FIGURE 18.

<sup>3</sup> Note that the weight in the  $i$ th row and  $j$ th column connects the  $j$ th input unit to the  $i$ th output unit.

of its weight vector and the input vector (which is common to all output units). Thus, each unit can be thought of as computing how close its weight vector is to the input vector. A larger activation is attained the closer the two vectors are. If all of the weight vectors have the same length, then that output unit with the largest activation will be the unit whose weight vector is closest to the input vector.

In the model with only one output unit, we imagined plotting the weight vector in the input vector space. This enabled us to see directly which input vectors led to a large response and which input vectors led to a small response. In the model with several output units, we can generalize by plotting each weight vector in the input space. Now we can see for each unit which inputs it responds to. If the weight vectors are spread around in the space, then every input will lead to some response. Also, the different units will respond to different inputs. If the weight vectors are assumed to have unit length, then the activation of the  $i$ th output unit is just the projection of  $v$  on the  $i$ th weight vector. For a given input, we can draw the projections of the input on the weight vectors. This gives us a graphic representation of the output of the network. It should be emphasized, however, that this representation is useful mostly as a conceptual tool. The graphic approach cannot be used in most systems, which can have hundreds or thousands of input lines.

Another perspective on the operation of the model can be obtained by focusing on the columns of the weight matrix rather than on its rows. Whereas the rows of the matrix are the weights on the lines coming *in* to the processing units, the columns correspond to the weights on the lines going *out* from the processing units. Each unit on the lower row in Figure 17 is associated with such a vector: The components of the vector are the weights linking that unit with the output units above. These vectors are referred to as the *outgoing weight vectors*, as contrasted with the *incoming weight vectors* which are the rows of the weight matrix.<sup>4</sup> In the previous section, it was seen that when a matrix multiplies a vector, the resulting vector is a linear combination of the columns of the matrix. This view applies to the PDP model as follows: The output vector  $u$  is a linear combination of the outgoing weight vectors from the input units. The coefficients in the linear combination are the activations of the input units. Thus, in this perspective, each input unit multiplies its outgoing weight vector by its activation, and the resulting vectors are added to yield the output vector of the system.

In general, as will be discussed further in a later section, a unit can

---

<sup>4</sup> This is not standard terminology, and I will continue to use the term *weight vector* to refer to the incoming weight vectors.

appear in a multilayer system and thus have both an incoming weight vector and an outgoing weight vector, as shown in Figure 19. In this case, both views of matrix-vector multiplication can be useful: The unit can be thought of as matching its incoming weight vector to the current input using the inner product, and sending the result of this match multiplied by the outgoing weight vector to the next level.

## LINEARITY

A distinction is often made between a *linear* system and a *nonlinear* system. In general, linear systems are relatively easy to analyze and understand, whereas nonlinear systems can be difficult. In this section, I will characterize linear systems. Nonlinear systems are defined simply as everything else. In a later section, I will give some specific examples of nonlinear systems.

Suppose that there is a function  $f$  which represents a system in that for each input  $x$  to the system, the output  $y$  is given by

$$y = f(x).$$

The  $x$  and  $y$  might be scalars or they might be vectors, depending on the particular system. The function  $f$  is said to be *linear* if for any inputs  $x_1$  and  $x_2$ , and any real number  $c$ , the following two equations hold:

$$f(cx) = cf(x). \quad (13)$$

$$f(x_1 + x_2) = f(x_1) + f(x_2). \quad (14)$$

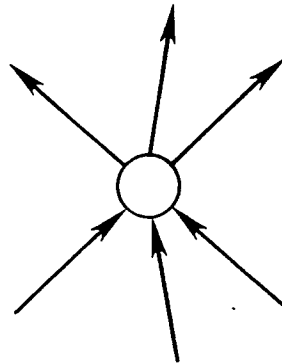


FIGURE 19.

The first of these two equations implies that if we multiply the input by some constant, then the output is multiplied by the same constant. The second equation is more important. Consider presenting the inputs  $x_1$  and  $x_2$  separately to the system and measuring the outputs. In a linear system, knowing how the system responds separately to the inputs is all we need to predict the output of the system when the sum  $x_1 + x_2$  is presented. We simply add the outputs found separately to obtain the response to the sum. In a nonlinear system, on the other hand, we might find that the response to the sum is much larger or smaller than would be expected based on the inputs taken separately. The response to the sum might be zero even when strong responses are obtained separately.

If we restrict ourselves to scalar functions of a scalar variable, then the only linear functions are those in which the output is proportional to the input, i.e., for some real number  $c$ :

$$y = cx.$$

However, many systems are scalar or vector functions of a vector input. For example, for a fixed vector  $\mathbf{w}$ , the function

$$u = \mathbf{w} \cdot \mathbf{v}$$

is a scalar function of a vector input  $\mathbf{v}$ . This function is a linear function because

$$\mathbf{w} \cdot (c\mathbf{v}) = c(\mathbf{w} \cdot \mathbf{v})$$

and

$$\mathbf{w} \cdot (\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{w} \cdot \mathbf{v}_1 + \mathbf{w} \cdot \mathbf{v}_2.$$

The PDP model with one output unit is an example of such a linear system.

A system in which the output is obtained from the input by matrix multiplication is also a linear system, according to Equations 9 and 10. It turns out that these are the only linear vector functions. That is, if a function  $f$  which maps from one vector space to another vector space is linear, then it can be represented by matrix multiplication.<sup>5</sup>

The PDP model discussed in the previous section is an example of a linear system because it is represented by matrix multiplication. In such a system, because of linearity, we know what the output will be when the sum of two vectors is presented if we know the outputs when

<sup>5</sup> Let  $\mathbf{v}_i$  be the  $i$ th standard basis vector and let  $\mathbf{w}_i = f(\mathbf{v}_i)$ . Then if  $\mathbf{W}$  is a matrix whose columns are the  $\mathbf{w}_i$ ,  $f(\mathbf{v}_i) = \mathbf{W}\mathbf{v}$  for all  $\mathbf{v}$ .

the vectors are presented separately. We also know what the output will be to scalar multiples of a vector. These properties imply that if we know the output to all of the vectors in some set  $\{v_i\}$ , then we can calculate the output to any linear combination of the  $v_i$ . That is, if  $v = c_1v_1 + c_2v_2 + \cdots + c_nv_n$ , then the output when  $v$  is presented to the system is

$$\begin{aligned} Wv &= W(c_1v_1 + c_2v_2 + \cdots + c_nv_n) = \\ &c_1(Wv_1) + c_2(Wv_2) + \cdots + c_n(Wv_n) \end{aligned} \quad (15)$$

The terms in the parentheses on the right are known vectors: They are the outputs to the vectors  $v_i$ . Thus, we simply multiply these vectors by the  $c_i$  to calculate the output when  $v$  is presented. If the  $v_i$  are a basis for some vector space, then every vector in the space is a linear combination of the  $v_i$ . Therefore, knowing the outputs of the system to the basis vectors allows us to calculate immediately the output to any other vector in the vector space without reference to the system matrix  $W$ . The preceding statement should be studied carefully, because it expresses an extremely important defining property of linear systems. Another way to say the same thing is as follows: Imagine that we are studying some physical system by measuring its responses to various inputs. The system might be electronic or physiological, for example. If it is a linear system, then we should first measure the responses to a set of inputs that constitute a basis for the input space. We then have no need to make any further measurements. The responses of the system to any other input vector can be immediately calculated based on the measurements that we have already made.

## MATRIX MULTIPLICATION AND MULTILAYER SYSTEMS

The systems considered until now have been *one-layer* systems. That is, the input arrives at a set of input units, is passed through a set of weighted connections described by a matrix, and appears on a set of output units. Let us now arrange two such systems in *cascade*, so that the output of the first system becomes the input to the next system, as shown in Figure 20. The composite system is a *two-layer* system and is described by two matrix-vector multiplications. An input vector  $v$  is first multiplied by the matrix  $N$  to produce a vector  $z$  on the set of intermediate units:

$$z = Nv,$$

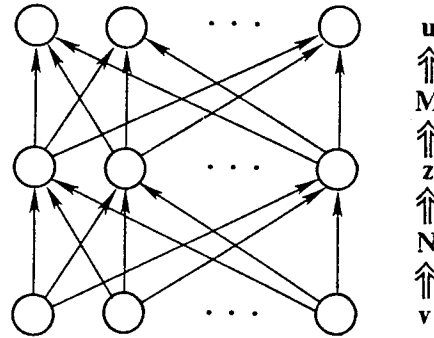


FIGURE 20.

and then  $z$  is multiplied by  $M$  to produce a vector  $u$  on the uppermost set of units:

$$u = Mz.$$

Substituting  $Nv$  for  $z$  yields the response for the composite system:

$$u = M(Nv). \quad (16)$$

This equation relates the input vectors  $v$  to the output vectors  $u$ .

We will now define an operation on matrices, called *matrix multiplication*, which will simplify the analysis of cascaded systems, allowing us to replace the two matrices  $M$  and  $N$  in Equation 16 by a single matrix  $P$ . Matrices  $M$  and  $N$  can be multiplied to produce a matrix  $P = MN$  as follows: The  $i, j$ th element of  $P$  is the inner product of the  $i$ th row of  $M$  with the  $j$ th column of  $N$ . Note that the order of multiplication is important—the product  $MN$  is generally not equal to the product  $NM$ . This is to be expected from the asymmetric treatment of  $M$  and  $N$  in the definition.

*Example:*

$$\begin{bmatrix} 3 & 4 & 5 \\ 1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} (3+8-5) & (6+0+5) \\ (1+0-1) & (2+0+1) \\ (0+2-2) & (0+0+2) \end{bmatrix} = \begin{bmatrix} 6 & 11 \\ 0 & 3 \\ 0 & 2 \end{bmatrix}$$

Another way to think about matrix multiplication follows from the definition of matrix-vector multiplication. Each column vector of  $P$  is the product of the matrix  $M$  with the corresponding column vector of  $N$ . For example, the first column of  $P$  is computed by multiplying the



