Outline on Determinants

Usually, in a mathematics course, as new material is presented, the significant results are demonstrated or derived, rather than simply stated. The goal is to present the material in such a way that the student can see for him or herself that each statement is valid. One topic that has traditionally been included in linear algebra courses is that of determinants. Among the teachers of linear algebra, there is a growing consensus that this topic is of less importance than others, and should not be given the emphasis that it used to receive. On the other hand, there are still applications where an understanding of determinants is required. As a compromise, I will simply present a brief overview of the topic, giving you an indication of what is true, and of some of the concepts, without attempting to give a complete mathematical development of the material. This will be done in one lecture and assignment set. Our text gives a thorough treatment on determinants, filling up all of chapter 4. You should hang onto that as a reference.

Geometric Definition of Determinant

The determinant operation or function is defined for square matrices. We know that an $n \times n$ matrix A can be thought of as the representation of a linear transformation from $\mathbf{R^n}$ to $\mathbf{R^n}$, namely, the mapping under which $\mathbf{x}$ is sent to $A\mathbf{x}$.

We can apply this mapping to each point in a subset of $\mathbf{R^n}$ and the result is another subset of $\mathbf{R^n}$. For example, in $\mathbf{R^2}$ we can apply the mapping to each point of the square with vertices (0,0), (0,1), (1,1) and (1,0). The result will be a parallelogram. Applying A to the vertices of the original square gives us the vertices of that parallelogram. Comparing the areas of the square and the parallelogram is a key part of the definition of determinant.

As a matter of fact, if A is applied to any figure in $\mathbf{R^2}$, the area of the resulting figure is a fixed multiple of the area of the original figure. So, in the preceding paragraph, if the area of the parallelogram is 2 times the area of the square, then applying A to *any* figure produces a new result that has any area 2 times as great. In this case we can think of 2 as an expansion factor for areas connected with A. The general idea is that every matrix has some expansion (or contraction) factor, and that applying the matrix to any figure increases (or decreases) the area by this factor.

These ideas can be extended to 1 and 3 dimensions. In the 3 dimensional case, the figures of interest are solids, and we focus on *volume*, rather than area. For example, in $\mathbf{R^3}$ we can think about applying A to each point in a cube with vertices (0,0,0), (0,1,0), (1,1,0), (1,0,0), (0,0,1), (0,1,1), (1,1,1), and (1,0,1). This time the result will be what is called a parallelopiped. It is the volume swept out by a parallelogram that is dragged through space for a fixed distance along a straight line. Now the original cube has a volume of 1, because its length, width, and depth are all 1. Suppose that the parallelopiped has a volume of 1.5. In analogy with the situation for two dimensions, we can say that A has expanded the volume of the original cube by a factor of 1.5, and this same factor will apply to any other figure. That is, if F is a solid figure, and $A(F)$ is the result of applying A to each point of F, then the volume of $A(F)$ will be 1.5 times the volume of F. This is a general principle for any 3×3 matrix: when we apply the matrix to every point of a figure, the volume of the result is a fixed multiple of the volume of the original figure. The multiplier is an expansion (or contraction) factor for volumes in, similar to the expansion factors for areas in the two dimensional case. Of course, not every matrix has an expansion factor of 1.5, but every matrix has *some* expansion factor.

The one dimensional case is so simple that it is only of interest because of the way it completes the pattern that holds for 2 and 3 dimensions. In any case, in one dimension, we deal with 1×1 matrices, which really behave just as numbers. In this setting, length takes the place of area, and as for the other dimensions, applying A to each point of an interval has the effect of multiplying the length by a fixed factor, given by the absolute value of the single entry of A.

As the foregoing discussion shows, in one, two, and three dimensions, multiplying figure by a matrix expands its length, area, or volume by a fixed factor. The idea of length, area, and volume can be extended to higher dimensions. Indeed, the term *measure* is used in mathematics to refer to amount space taken up by a figure in any number of dimensions. So in the plane, measure means area; in space it means volume; and on a line, measure refers to length. In higher dimensions, measure has an analogous meaning. The full explanation of what that meaning is for higher dimensions is beyond the scope of this brief treatment of determinants. But it should be understood by analogy with the three dimensions we know. And in any number of dimensions, applying a matrix to any figure expands or contracts the measure by a fixed factor.

From this point on, the term *volume* will be used in place of *measure*, because the concept of volume is more familiar. It will be generally helpful to visualize the ideas presented in the setting of $\mathbf{R^3}$. But do keep in mind that volume really corresponds to the more general concept of measure, and so means length in one dimension and area in two dimensions. With this convention, it is correct to say that for any matrix there is a fixed expansion factor that tells how the volume of a figure changes under the action of the matrix. This expansion factor is the first of two fundamental concepts that goes into the definition of determinants.

The other ingredient involves what is called orientation. This can be understood geometrically in 2 and 3 dimensions. Given two vectors in the plane, in a specified order, we can think about rotating the first vector about the origin in the direction of the other vector, that is, in the direction that requires the shortest rotation to reach the second vector. If the direction is counterclockwise, we say that the order in which the vectors are given has positive orientation, otherwise negative orientation. Observe that the standard vectors $\mathbf{e_1}$, $\mathbf{e_2}$, in that order, have a positive orientation. Similar imagery applies in three dimensions. Given three independent vectors, $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$, consider the plane P containing the origin $\mathbf{u}$ and $\mathbf{v}$. Sit in space at $\mathbf{w}$ which is outside that plane, and watch as the plane is rotated about the origin to carry $\mathbf{u}$ in the direction of $\mathbf{v}$. If the rotation appears counterclockwise, then the three vectors, as ordered, have positive orientation, otherwise negative. As before, the standard vectors $\mathbf{e_1}$, $\mathbf{e_2}$, and $\mathbf{e_3}$, in that order, have positive orientation. In both of these settings, the idea of orientation is related to mirror images. In general, the mirror image of a figure has the opposite orientation of the original. So for instance, figures illustrating a right and left hand would have opposite orientation; two right hands would have the same orientation.

It is a fact about linear transformations that they either preserve or reverse orientation. Given a pair of vectors in $\mathbf{R^2}$, ordered with a positive orientation, we can apply A to each and obtain another pair of vectors. Either A always retains the orientation of pairs of vectors, or it always reverses them. In the first case we say that A preserves orientation. Similarly in three dimensions, either a matrix A always maintains the orientation of sets of three vectors, or it reverses the orientation. In the first case we say that A preserves orientation. In a similar way, these ideas can be extended to n dimensions. The basic idea is that there is a way to distinguish between a positive and negative orientation of sets of n independent vectors; the standard vectors $\mathbf{e_1}$ through $\mathbf{e_n}$, in that order have a positive orientation. A matrix A must either preserve or reverse orientation.

Now we can define the determinant of a square matrix: If A is a square matrix, the determinant of A, (written $|A|$, $\det A$, or $\det(A)$) is equal to the volume expansion factor for A, if A preserves orientations, and is equal to the negative of the expansion factor for A if A does not preserve orientations.

Example: A rigid rotation has a determinant of 1.

Example: A reflection through a line in $\mathbf{R^2}$ or through a plane in $\mathbf{R^3}$ has a determinant of -1.

Example: An elementary row operation matrix that simply interchanges two rows has a determinant of -1.

The matrix A has an inverse if and only if $|A| \neq 0$

We already know that A has an inverse if and only if the columns of A are linearly independent. On the other hand, the only way $|A|$ can equal 0 is if the volume expansion factor of A is zero. In particular, if the expansion factor is 0, when A is applied to the *cube* define by the standard $\mathbf{e_i}$ the resulting figure must have 0 volume. That happens if and only if the image vectors are linearly dependent. This is clear in 2 and 3 dimensions. The images of the standard vectors will define area 0 in the plane only if they fall on a single line, and they will define volume 0 in space only if they all fall into a single plane. These are exactly the cases in 2 and 3 dimensions in which the image vectors are linearly dependent. In higher dimensions similar reasoning holds as well. So, $|A| = 0$ is equivalent to the images of the $\mathbf{e_i}$ vectors being linearly dependent. But those images are precisely the columns of A. Therefore, A has determinant 0 just when the columns of A are dependent, which in turn occurs just when A is not invertible.

The $|AB| = |A| \cdot |B|$

If you ignore the signs involved, this is immediately clear from the interpretation of AB as the matrix that represents the composite operation *first apply B then apply A to the result.* If B expands volumes by a factor b and A by a factor a, then it is clear that first applying B and then A will have a cumulative expansion factor of ab. So all that we need to worry about is the way the signs combine for orientation preserving and reversing choices of A and B. If A and B both preserve orientation, it is clear that the composite operation will preserve orientation. Likewise, if one but not both of A and B reverses orientation, the composite operation must also reverse orientation. It is less self evident, but nevertheless true, that if A and B both reverse orientation, the composite operation preserves orientation. These facts assure that the signs work out correctly to have $|AB| = |A||B|$.

The determinants of elementary operation matrices

Consider a swap of two rows. The elementary matrix for that operation is just the identity matrix with two rows interchanged (which is the same as having two columns interchanged). So the *cube* defined by the $\mathbf{e_i}$ vectors gets mapped to itself. This shows that the volume expansion factor is 1. On the other hand, because two of the $\mathbf{e_i}$ are swapped by the matrix, the orientation of the standard vectors is not preserved. You can check this directly in 2 and 3 dimensions, for higher dimensions it requires a more careful development of the idea of orientation. The result is that the determinant of a row swap operation matrix is -1.

Now consider scaling one row by a constant. If the constant is positive, the effect on the cube defined by the $\mathbf{e_i}$ vectors is simply to scale the length of the cube in the direction parallel to one axis. So the expansion factor will be equal to the scale factor. If the scale factor is negative, the effected dimension is also reflected through the origin, and that has the effect of reversing the orientation of the images of the standard vectors. In either case the determinant is just exactly equal to the scale factor.

Finally, if we add a multiple of one row to another, that is simply a shear in one direction, and it as no effect on either volume or orientation. So the determinant of that kind of row operation is 1.

The determinant of a diagonal matrix

If D is a diagonal matrix, with entries $d_1, d_2, \cdots, d_n$ on the diagonal, then we can express D as a product $D = E_1 E_2 \cdots E_n$ where each E_i looks just like the identity matrix except with d_i at the ith entry on the diagonal. But that means each E_i is a row operation matrix for scaling row i by d_i, and we know that $|E_i| = d_i$. Now take the determinant of both sides of the equation $D = E_1 E_2 \cdots E_n$, using the previous result about determinants of products of matrices. That gives

$$|D| = |E_1||E_2| \cdots |E_n| = d_1 d_2 \cdots d_n$$

That is, the determinant of a diagonal matrix is the product of the diagonal entries.

The determinant of a triangular matrix

Consider first a lower triangular matrix. If there is a zero on the diagonal, the determinant will be 0. To see this, consider the first row having a diagonal entry of 0. It might be in the first row, in which case the first row is all zeros. Then the matrix is not invertible and the determinant is 0. If the first diagonal 0 is not in the first row, then the row it is in is a linear combination of the rows above it (draw a few diagrams and you will see why that works). In that case too the matrix is not invertible and the determinant is 0. So if there is a 0 on the diagonal the determinant is 0.

If there is no zero on the diagonal, each diagonal can be used as a pivot entry to clear the remaining entries in its column, using only the row operations that add a multiple of one row to another. That means we have an equation of the form

$$E_m \cdots E_2 E_1 A = D$$

where the E's represent row operations and D is a diagonal matrix with the same values on its diagonal as in A. Take the determinant of both sides of this equation using the fact that the determinant of a product can be applied factor by factor. Also, each of the E matrices has a determinant of 1. So $|A| = |D|$. Using the previous result, this gives $|A|$ as the product of the diagonal entries. So for a lower triangular matrix A, the determinant is the product of the diagonal entries. Very similar arguments show that the same conclusion applies to upper triangular matrices.

Determinant and triangular factorization

We have seen that row reducing A to the upper triangular form U can be expressed in terms of a matrix factorization. If the reduction requires no row exchanges, we can write $A = LU$ where L and U are each triangular and L has only 1's on the diagonal. That means that $|A| = |L||U|$. But the previous result shows that the determinant of each triangular matrix is the product of the diagonal entries. For L that gives a determinant of 1. Consequently, $|A|$ is just the product of the diagonal entries of U. If row exchanges are required in the row reduction operation, they only modify this situation slightly. Each time we apply a row exchange to A, that is multiplying by a row operation matrix with determinant -1. So the net effect on the determinant of the row reduced matrix is a sign change. This gives the following result: Reduce A to an upper triangular form U without using any row operations that scale rows of the matrix. In the process count the number of row exchanges performed. If that count is even, $|A|$ is just the product of the diagonal entries of U; otherwise $|A|$ is the negative of that product. This is a very efficient way to compute the determinant of a matrix.

Exercise: Show using row operations that

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Your proof should be valid for all choices of a, b, c, and d. One approach is to consider three cases: (1) $a \neq 0$; (2) $a = 0$ but $c \neq 0$; and (3) $a = c = 0$.

$|A^T| = |A|$

If A is not invertible, then neither is A^T and vice versa. So in that case each has a determinant of 0, and they are equal. Otherwise, A can be written as a product of elementary row operation matrices, so A^T is the product of the transposes of those row operation matrices. We can compute the determinant of A^T by multiplying the determinants of all the transposed elementary row operation matrices. For this reason, it suffices to show that transposing an elementary row operation matrix has no effect on its determinant. This is straightforward, and works in a very similar fashion to the discussion of determinants of row operation matrices above.

Cofactor Expansions

There is another way to compute the determinant of a matrix using determinants of some of its submatrices. This is primarily of theoretical interest, since it is impractical for actual computation for matrices much larger than 10 by 10, or so. It is difficult to prove that the procedure is correct, especially in the way that the ideas have been developed here. For completeness here is a brief description of how it works. A more expanded version is given on pages 181 and 182 of the text.

Here is what to do.

- Choose any row or column of your matrix. It doesn't matter which you choose, all will lead to the same result. In Figure 1 below, the third column is selected.

$$\begin{bmatrix} 11 & 12 & \boxed{13} & 14 \\ 21 & 22 & \boxed{23} & 24 \\ 31 & 32 & \boxed{33} & 34 \\ 41 & 42 & \boxed{43} & 44 \end{bmatrix}$$

Figure 1.

$$\begin{bmatrix} \not{11} & \not{12} & \not{13} & \not{14} \\ 21 & 22 & \not{23} & 24 \\ 31 & 32 & \not{33} & 34 \\ 41 & 42 & \not{43} & 44 \end{bmatrix}$$

Figure 2.

- Write down the entries of your row or column, leaving plenty of space between the entries. By referring to the figure, you can see that for the example it illustrates, the entries would be written down like this:

$$13 \qquad 23 \qquad 33 \qquad 43$$

- Attach a + sign to the first entry if you picked an odd numbered row or column, attach a - sign otherwise. For the example in the figure we picked an odd column (the third column) so we attach a + sign to the 13.

- Attach signs to the rest of the entries alternating + and -, starting with the sign you put on the first entry. For the example, that produces

$$+13 \qquad -23 \qquad +33 \qquad -43$$

- Directly following each entry write down a determinant, which I will call the subdeterminant, as follows. You are looking at one entry in the line that is being created: find that same entry in the original matrix. Mentally draw a line through the entire row and column of that entry, and then copy all of the other entries into the subdeterminant that is being written. If the original matrix was $n \times n$, then the subdeterminants will each be $(n-1) \times (n-1)$. This is illustrated for the first subdeterminant in our example in Figure 2. The complete formula for that example would look like this:

$$+13\begin{vmatrix} 21 & 22 & 24 \\ 31 & 32 & 34 \\ 41 & 42 & 44 \end{vmatrix} -23\begin{vmatrix} 11 & 12 & 14 \\ 31 & 32 & 34 \\ 41 & 42 & 44 \end{vmatrix} +33\begin{vmatrix} 11 & 12 & 14 \\ 21 & 22 & 24 \\ 41 & 42 & 44 \end{vmatrix} -43\begin{vmatrix} 11 & 12 & 14 \\ 21 & 22 & 24 \\ 31 & 32 & 34 \end{vmatrix}$$

- Now multiply each entry in your line by its adjacent subdeterminant, and add them together according to the + and - signs.

Note that this procedure is recursive. If you start with a 4 by 4 matrix, you will obtain using this method an expression including 4 determinants of 3 by 3 matrices. Then each of the 3 by 3 determinants will have to be expanded to a series of 2 by 2 determinants and so on. No matter what size matrix you begin with, eventually, you reach 2 by 2 determinants, and these can be directly computed using the formula

$$\left| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right| = ad - bc$$

One idea for using this approach is to pick a row or column that has many 0 entries. Also, if a subdeterminant is triangular, you can calculate the determinant directly by multiplying the diagonal entries as described above.

$\det[\ \mathbf{a_1} \quad \mathbf{a_2} \quad \cdots \quad \mathbf{a_n}\ ]$

If we think of the matrix A in terms of its columns, then $\det(A)$ can be thought of as a function of those columns. That is, given n column vectors $\mathbf{a_1}, \mathbf{a_2}, \cdots, \mathbf{a_n} \in \mathbf{R^n}$ define $\det(\mathbf{a_1}, \mathbf{a_2}, \cdots, \mathbf{a_n})$ to be the determinant of the matrix with columns given by the $\mathbf{a}$'s. This function is linear in each variable, meaning that

$$\det[\mathbf{a_1}\mathbf{a_2}\cdots(\mathbf{a_j}+\mathbf{u})\cdots\mathbf{a_n}] = \det[\mathbf{a_1}\mathbf{a_2}\cdots\mathbf{a_j}\cdots\mathbf{a_n}] + \det[\mathbf{a_1}\mathbf{a_2}\cdots\mathbf{u}\cdots\mathbf{a_n}]$$

and

$$\det[\mathbf{a_1}\mathbf{a_2}\cdots r\mathbf{a_j}\cdots\mathbf{a_n}] = r\det[\mathbf{a_1}\mathbf{a_2}\cdots\mathbf{a_j}\cdots\mathbf{a_n}]$$

The second of these equations is easy to prove using what we know about elementary matrices and determinants of products. First, note that

$$[\mathbf{a_1}\mathbf{a_2}\cdots r\mathbf{a_j}\cdots\mathbf{a_n}] = [\mathbf{a_1}\mathbf{a_2}\cdots\mathbf{a_j}\cdots\mathbf{a_n}]\begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & r & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

where the matrix at the far right is a diagonal matrix with an r in the j^{th} diagonal position and 1's in the others. The determinant of this diagonal matrix is found by multiplying all the diagonal entries, resulting in r. Therefore, taking the determinant of each matrix leads to the equation we were trying to prove:

$$\det[\mathbf{a_1}\mathbf{a_2}\cdots r\mathbf{a_j}\cdots\mathbf{a_n}] = r\det[\mathbf{a_1}\mathbf{a_2}\cdots\mathbf{a_j}\cdots\mathbf{a_n}]$$

The first linearity equation is trickier to justify on the basis of the preceding results. One can make careful use of row operations, but there are several cases that have to be considered and it gets pretty involved. For these reasons, I won't try to prove the addition part of the linearity of the determinant function. But it is an important property of the determinant, and you should keep it in mind.

One thing that linearity shows right away is that as a general rule, $|A+B|$ will not be equal to $|A|+|B|$. Consider the 2×2 case. If $A = [A_1 \;\; A_2]$ and $B = [B_1 \;\; B_2]$, then applying linearity three times shows that

$$\begin{aligned} |A+B| &= |A_1+B_1 \quad A_2+B_2| \\ &= |A_1 \quad A_2+B_2| + |B_1 \quad A_2+B_2| \\ &= |A_1 \quad A_2| + |A_1 \quad B_2| + |B_1 \quad A_2| + |B_1 \quad B_2| \\ &= |A| + |A_1 \quad B_2| + |B_1 \quad A_2| + |B| \end{aligned}$$

There are special cases where $|A+B|$ happens to work out to the same result as $|A|+|B|$, but it is rare, and you definitely cannot count on it.

Another interesting consequence of linearity is a result called Cramer's rule, a formula for the solution of a linear system using determinants. More specifically, if the system $A\mathbf{x} = \mathbf{b}$ has an invertible matrix A, then each entry of the solution vector $\mathbf{x}$ can be expressed in terms of determinants. Here is a derivation.

Use the columns $\mathbf{a_1}, \mathbf{a_2}, \cdots, \mathbf{a_n}$ of A to define a function f from $\mathbf{R^n}$ to $\mathbf{R^n}$ as follows. Given any vector $\mathbf{u}$, $f(\mathbf{u}) = \det(\mathbf{u}, \mathbf{a_2}, \mathbf{a_3}, \cdots, \mathbf{a_n})$. Now apply f to both sides of the vector equation form of $A\mathbf{x} = \mathbf{b}$:

$$x_1\mathbf{a_1} + x_2\mathbf{a_2} + \cdots + x_n\mathbf{a_n} = \mathbf{b}$$

Because f is linear, we obtain

$$x_1 f(\mathbf{a_1}) + x_2 f(\mathbf{a_2}) + \cdots + x_n f(\mathbf{a_n}) = f(\mathbf{b})$$

Now $f(\mathbf{a_1}) = \det A \neq 0$, because A is invertible. All the other terms on the left are 0. This is because $f(\mathbf{a_j}) = \det(\mathbf{a_j}, \mathbf{a_2}, \mathbf{a_3}, \cdots, \mathbf{a_n})$ and that is a determinant with two columns equal to $\mathbf{a_j}$ for $j > 1$. If two columns are equal, then the columns are certainly not linearly independent, and that means the matrix with those two equal columns is not invertible, and that means the determinant is 0. Therefore, we have

$$x_1 \det A = \det[\mathbf{b}, \mathbf{a_2}, \cdots \mathbf{a_n}]$$

and hence

$$x_1 = \frac{\det[\mathbf{b}, \mathbf{a_2}, \cdots \mathbf{a_n}]}{\det A}$$

We can describe this in words. The first component of $\mathbf{x}$ can be found as follows: replace the first column of A with $\mathbf{b}$ and compute the determinant. Divide that by the determinant of A. Similar reasoning shows that the same conclusion is valid if you replace the word *first* with *jth*. That gives

> The j^{th} component of x can be found as follows: replace the j^{th} column of A with $\mathbf{b}$ and compute the determinant. Divide that by the determinant of A.

That is Cramer's rule.

Exercises

1. In 4.1: 1,9,25,27,29
2. In 4.2: 5, 7, 29, 30, 31, 37
3. In 4.3: 1, 5, 27