Determinant and Trace

Area and mappings from the plane to itself: Recall that in the last set of notes we found a linear mapping to take the unit square $S = \{0 \leq x \leq 1, 0 \leq y \leq 1\}$ to any parallelogram P with one corner at the origin. We can write the parallelogram P as

$$P = \{xv + yw : 0 \leq x \leq 1, 0 \leq y \leq 1\},$$

where v and w are the two vectors which form the edges of P starting at the origin $(0,0)$. Then we can write the linear transformation as

$$T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = xv + yw, \qquad [T] = \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \end{bmatrix},$$

where $v = (v_1, v_2)$ and $w = (w_1, w_2)$ in components. Notice that the mapping T is invertible precisely when it does not collapse S down to a line segment (or a point), which happens precisely when the area of the parallelogram P is non-zero. Also, recall that we said T is invertible precisely when its determinant $\det([T]) = v_1 w_2 - v_2 w_2$ is nonzero. We have

$$\det([T]) \neq 0 \Leftrightarrow T \text{ invertible } \Leftrightarrow \text{Area}(P) \neq 0.$$

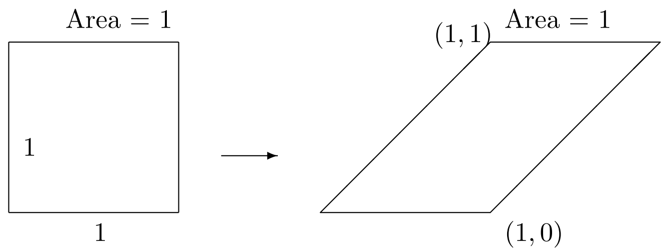
We'll see next that $\det([T])$ *is* the area of P, up to a sign. This is easiest to see with the shear map we examined in the last set of notes. Start with the sheer map T whose matrix representation is

$$[T] = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}.$$

(In the earlier set of notes we wrote the entry in the upper right corner of $[T]$ as a, but it will turn out to be convenient to call it b for our later discussion.) In this case, T maps the unit square $S = \{0 \leq x \leq 1, 0 \leq y \leq 1\}$ to the parallelogram P spanned by the two vectors $v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $w = \begin{bmatrix} b \\ 1 \end{bmatrix}$; in other words,

$$T(S) = P = \{xv + yw : 0 \leq x \leq 1, 0 \leq y \leq 1\} = \left\{ x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y \begin{bmatrix} b \\ 1 \end{bmatrix} : 0 \leq x \leq 1, 0 \leq y \leq 1 \right\}.$$

We reproduce a picture here:



We already know that the unit square S has area 1, but let's see that P also has area 1. The area of a parallelogram is equal to its base times its height, and the height and base of P are both 1, so the area of P is $1 \cdot 1 = 1$. On the other hand,

$$\det([T]) = \det\left(\begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix}\right) = 1 \cdot 1 - 0 \cdot b = 1 = \text{Area}(P).$$

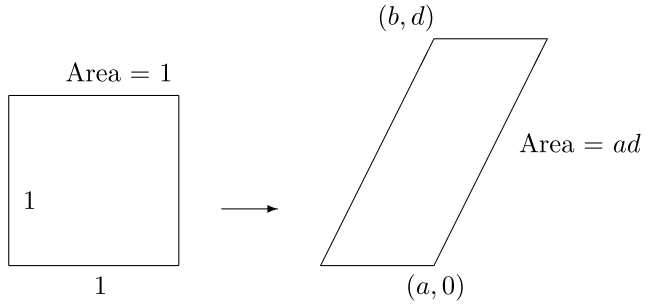
Now we can rescale the sheer T by a in the horizontal direction and by d and vertical direction, to have something more general. This time we have

$$[T] = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix},$$

and

$$T(S) = P = \{xv + yw : 0 \leq x \leq 1, 0 \leq y \leq 1\} = \left\{ x \begin{bmatrix} a \\ 0 \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix} : 0 \leq x \leq 1, 0 \leq y \leq 1 \right\},$$

and the picture looks like



(In this particular picture $a = 1/2$ and $d = 2$, but this choice of scaling factors is not important.) This time the height of the parallelogram P is d while its base is a, so $\text{Area}(P) = \text{base}\cdot\text{height} = ad$. Again, we have

$$|\det([T])| = \left|\det\left(\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}\right)\right| = |a \cdot d - b \cdot 0| = |ad| = \text{Area}(P).$$

Notice that the absolute value here is necessary, because a and d could have opposite signs.

Now that we know $|\det([T])|$ gives the area of the image of the unit square if $[T] = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$, it's not too hard to see this is true for any linear map. We'll first need a technical fact.

Exercise: Prove $\det(AB) = \det(A)\det(B)$ for 2×2 matrices A and B. (Really, just multiply it out.) Notice that this means $\det(AB) = \det(BA)$ for any pair of 2×2 matrices.

Exercise: Show that for any angle θ we have

$$\det([R_\theta]) = \det\left(\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\right) = 1.$$

Now let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a linear mapping of the plane to itself, and suppose

$$[T] = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

This means $T(e_1) = \begin{bmatrix} a \\ c \end{bmatrix}$ and $T(e_2) = \begin{bmatrix} b \\ d \end{bmatrix}$, where $e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ as before. Now, the vector $T(e_1) = \begin{bmatrix} a \\ c \end{bmatrix}$ makes some angle θ with the positive x axis, so we apply the rotation $R_{-\theta}$ to T to get a new mapping

$$\tilde{T} = R_{-\theta} \circ T, \qquad [\tilde{T}] = [R_{-\theta}][T] = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \tilde{a} & \tilde{b} \\ 0 & \tilde{d} \end{bmatrix},$$

and

$$\det([\tilde{T}]) = \det([R_{-\theta}][T]) = \det([R_{-\theta}])\det([T]) = \det([T]).$$

By the computation we did above, $\text{Area}(\tilde{P}) = |\det([\tilde{T}])|$. We also have that T sends the unit square S to a parallelogram P, and $\tilde{T}$ sends S to a parallelogram $\tilde{P}$. These two parallelograms P and $\tilde{P}$ differ by a rotation, so they have the same area. Thus we see

$$\text{Area}(P) = \text{Area}(\tilde{P}) = |\det([\tilde{T}])| = |\det([T])|.$$

In particular, we have just proven that $\det([T]) \neq 0$ precisely when T is invertible, because this is precisely when the image parallelogram P has nonzero area.

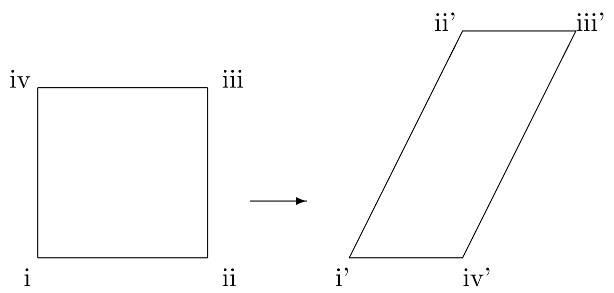
Orientation and the sign of the determinant: As we saw in the previous notes, there are actually two linear transformations which map the unit square S onto this parallelogram P, we can also have

$$T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} b \\ d \end{bmatrix}, \qquad T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} a \\ 0 \end{bmatrix}, \qquad [T] = \begin{bmatrix} b & a \\ d & 0 \end{bmatrix}.$$

In this case we see that

$$\det([T]) = -ad = -\text{Area}(P).$$

Why do we have the minus sign? To understand what's going on, it will help to label the corners of the unit square S and the parallelogram P as in the picture below.



What does this labeling mean? The mapping T sends the vector e_1, which goes from i to ii in the square on the left to the vector $\begin{bmatrix} b \\ d \end{bmatrix}$. which also goes from i to ii in the parallelogram on the right. Similarly, the mapping T sends the vector e_2, which goes from i to iv in the square on the left to the vector $\begin{bmatrix} a \\ 0 \end{bmatrix}$. which also goes from i to iv in the parallelogram on the right. Now, if we follow the labeling of the corners of the square in order, as in i to ii to iii to iv, then we traverse along the boundary of the square counter-clockwise. However, if we follow the labeling of th ecorners of the parallelogram in order, as in i' to ii' to iii' to iv', we traverse along the boundary of the parallelogram clockwise. This means the mapping T reversed the direction we traversed along the boundary of the shape. In other words, T reversed the orientation. We have discovered the following general principle:

$$\det([T]) < 0 \Leftrightarrow T \text{ reverses orientation.}$$

This principle is exactly why we wrote $|\det([T])| = \text{Area}(P)$ before. In general, if $T : \mathbb{R}^2 \to \mathbb{R}^2$ preserves orientation then $\det([T]) = \text{Area}(P)$, but if T reverses orientation then $\det([T]) = -\text{Area}(P)$.

Higher dimensions: So far we've seen that the determinant of a 2×2 matrix is the area (up to a sign) of the parallelogram which is the image of the unit square. In fact, a similar thing is true in higher dimensions. Let $[T]$ be an $n \times n$ matrix, which we've seen corresponds to a linear map $T : \mathbb{R}^n \to \mathbb{R}^n$. Then T sends the unit cube $S = \{0 \leq x_i \leq 1 : i = 1, 2, \ldots, n\}$ to a parallelpiped P, which is spanned by the columns of $[T]$. Then $|\det([T])| = \mathrm{Vol}(P)$, where Vol gives the n-dimensional volume.

We begin with a quick illustrative example. Consider

$$[T] = \begin{bmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & e \end{bmatrix}, \qquad e > 0.$$

Then the image of the unit cube S under T is

$$P = \{(x, y, z) : (x, y) \in \bar{P}, 0 \leq z \leq e\},$$

where

$$[\bar{T}] = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \bar{S} = \{0 \leq x \leq 1, 0 \leq y \leq 1\}, \quad \bar{P} = T(\bar{S}).$$

By slicing P with horizontal slices, we see

$$\mathrm{Vol}(P) = e\,\mathrm{Area}(\bar{P}) = e|\det([\bar{T}])|.$$

So, by any reasonable definition of the determinant for 3×3 matrices which fits with our definition for 2×2 matrices, we must have

$$\det([T]) = e \det([\bar{T}]) = e(ad - bc).$$

Exercise: Let $[T]$ be a 3×3 matrix. Show that you can always perform a rotation to make the last colum of $[T]$ into $\begin{bmatrix} 0 \\ 0 \\ e \end{bmatrix}$. (Hint: what are the columns of $[T]$?)

At this point, we can write down a reasonable formula for the determinant of a 3×3 matrix. Let

$$[T] = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix},$$

then

$$\det[T] = c \det\left(\begin{bmatrix} d & e \\ g & h \end{bmatrix}\right) - f \det\left(\begin{bmatrix} a & b \\ g & h \end{bmatrix}\right) + i \det\left(\begin{bmatrix} a & b \\ d & e \end{bmatrix}\right).$$

Here we've singled out the last column, but we can do the same thing by picking out any row or column. To do this properly, we need some notation. Let $[T] = [A_{ij}]$, so that the entry of $[T]$ in the ith row, jth column is A_{ij}. Also, let $[\bar{T}_{ij}]$ be the 2×2 matrix you get from $[T]$ by crossing out the ith row and jth column. Then for any choice of $j = 1, 2, 3$ we can write

$$\det([T]) = (-1)^{1+j} A_{1j} \det([\bar{T}_{1j}]]) + (-1)^{2+j} A_{2j} \det([\bar{T}_{2j}]) + (-1)^{3+j} A_{3j} \det([\bar{T}_{3j}]),$$

which computes $\det([T])$ by exanding along the jth column. Alternatively, for any choice of $i = 1, 2, 3$ we can write

$$\det([T]) = (-1)^{i+1} A_{i1} \det([\bar{T}_{i1}]) + (-1)^{i+2} A_{i2} \det([\bar{T}_{i2}]) + (-1)^{i+3} A_{i3} \det([\bar{T}_{i3}]),$$

which computes $\det([T])$ by expanding along the ith row.

The same idea will compute the determinant of any square matrix inductively. That is, you write the determinant of an $n \times n$ matrix as a sum of determinants of $(n-1) \times (n-1)$ matrices. We write the general formula as follows. Again, we let A_{ij} be the entry of $[T]$ in the ith row, jth column, and we let $[\bar{T}_{ij}]$ be the $(n-1)\times(n-1)$ matrix you get from $[T]$ by crossing out the ith row and the jth column. The for any choice of $j = 1, 2, \dots, n$ we compute $\det([T])$ by expanding along the jth column using the formula

$$\det([T]) = (-1)^{1+j} A_{1j} \det([\bar{T}_{1j}]) + (-1)^{2+j} A_{2j} \det([\bar{T}_{2j}]) + \cdots + (-1)^{n+j} A_{nj} \det([\bar{T}_{nj}]).$$

Alternatively, for any choice of $i = 1, 2, \dots, n$ we compute $\det([T])$ by expanding along the ith row using the formula

$$\det([T]) = (-1)^{i+1} A_{i1} \det([\bar{T}_{i1}]) + (-1)^{i+2} A_{i2} \det([\bar{T}_{i2}]) + \cdots + (-1)^{i+n} A_{in} \det([\bar{T}_{in}]).$$

We summarize some important properties of the determinant here.

1. The determinant is linear in each row and column. That is, if A is an $n \times n$ matrix and $\tilde{A}$ is the same as A except that you multiply the ith row by c, then $\det(\tilde{A}) = c \det(A)$. Also, A_1 and A_2 are the same except at the ith row and A is what you get by adding together the ith row of A_1 and A_2 then $\det(A) = \det(A_1) + \det(A_2)$. The same goes for columns.
2. Consequently, if A is an $n \times n$ matrix and c is a number then $\det(cA) = c^n \det(A)$.
3. An $n \times n$ matrix A is invertible if and only if $\det(A) \neq 0$.
4. In fact, $|\det(A)|$ is the n-dimensional volume of the parallelpiped P which is the image of the unit cube $S = \{0 \leq x_1 \leq 1, \dots, 0 \leq x_n \leq 1\}$ under the linear transformation associated to A. (You can prove this by induction, in a very similar way we got the geometric interpretation for three dimensions from the two-dimensional version.)
5. Let A and B be $n \times n$ matrices, then $\det(AB) = \det(A)\det(B)$.
6. Let A be an $n \times n$ matrix and let $\tilde{A}$ be the matrix you get by swapping adjacent two rows of A (or by swapping two adjacent columns). Then $\det(\tilde{A}) = -\det(A)$

Trace: Another important number associated to an $n \times n$ matrix is its trace. We let $[T]$ be an $n \times n$ matrix with A_{ij} being the entry in the ith row, jth column. Then the trace of $[T]$ is given by

$$\mathrm{tr}([T]) = A_{11} + A_{22} + \cdots + A_{nn},$$

the sum of the entries of $[T]$ on the diagonal running from the top left of $[T]$ to its bottom right. Later on, we'll see that under some conditions (for instance, if $A_{ij} = A_{ji}$) that the trace measures an average distortion of T as a mapping. That is, if the trace is close to n then T doesn't distort distances too much, but if the traces is very different from n then it distorts distances a lot. Remember that this guidline only holds if T is symmetric, that is if $A_{ij} = A_{ji}$. The trace satisfies the following properties:

1. If I is the $n \times n$ identity matrix then $\mathrm{tr}(I) = n$.
2. If A and B are square matrices and c is a number then

$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B), \qquad \mathrm{tr}(cA) = c\,\mathrm{tr}(A), \qquad \mathrm{tr}(AB) = \mathrm{tr}(BA).$$

Some special matrices: We close with some special types of square matrices. Let $[T]$ be a square matrix, with entries A_{ij} in the ith row, jth colum as above. We say $[T]$ is *diagonal* if $A_{ij} = 0$ for $i \neq j$. We say $[T]$ is *upper triangular* if $A_{ij} = 0$ for $i > j$ and that $[T]$ is *lower triangular* if $A_{ij} = 0$ for $i < j$.

Exercise: Why is a diagonal matrix called diagonal? How about upper (or lower) triangular matrices?

Exercise: Let $[T]$ be diagonal, and show that

$$\det([T]) = A_{11}A_{22}\cdots A_{nn}, \qquad \mathrm{tr}([T]) = A_{11} + A_{22} + \cdots + A_{nn}$$

Exercise: Show that the same formula holds for a upper triangular and lower triangular matrices.