Linear Spaces

We have seen (12.1-12.3 of Apostol) that n-tuple space V_n has the following properties:

Addition:

1. (Commutativity) $A + B = B + A$.
2. (Associativity) $A + (B+C) = (A+B) + C$.
3. (Existence of zero) There is an element $\underline{0}$ such that $A + \underline{0} = A$ for all A.
4. (Existence of negatives) Given A, there is a B such that $A + B = \underline{0}$.

Scalar multiplication:

5. (Associativity) $c(dA) = (cd)A$.
6. (Distributivity) $(c+d)A = cA + dA$, $c(A+B) = cA + cB$.
7. (Multiplication by unity) $1A = A$.

Definition. More generally, let V be *any* set of objects (which we call vectors). And suppose there are two operations on V, as follows: The first is an operation (denoted $+$) that assigns to each pair A, B of vectors, a vector denoted $A + B$. The second is an operation that assigns to each real number c and each vector A, a vector denoted cA. Suppose also that the seven preceding properties hold. Then V, with these two operations, is called a linear space (or a *vector space*). The seven properties are called the *axioms for a linear space*.

There are many examples of linear spaces besides n-tuple space V_n. The study of linear spaces and their properties is dealt with in a subject called Linear Algebra. We shall treat only those aspects of linear algebra needed for calculus. Therefore we will be concerned only with n-tuple space V_n and with certain of its subsets called "linear subspaces" :

Definition. Let W be a non-empty subset of V_n ; suppose W is closed under vector addition and scalar multiplication. Then W is called a linear subspace of V_n (or sometimes simply a subspace of V_n.)

To say W is closed under vector addition and scalar multiplication means that for every pair A, B of vectors of W, and every scalar c, the vectors $A + B$ and cA belong to W. Note that it is automatic that the zero vector $\underline{0}$ belongs to W, since for any A in W, we have $\underline{0} = 0A$. Furthermore, for each A in W, the vector $-A$ is also in W. This means (as you can readily check) that W is a linear space in its own right (i.e., it satisfies all the axioms for a linear space).

Subspaces of V_n may be specified in many different ways, as we shall see.

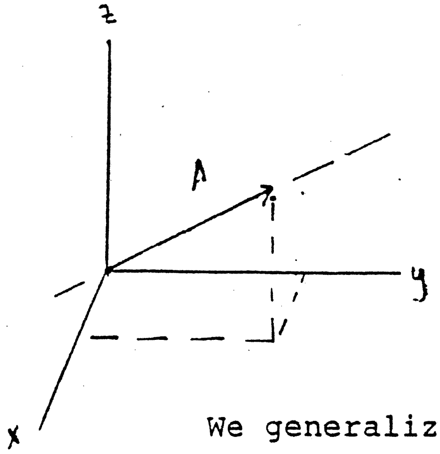
Example 1. The subset of V_n consisting of the 0-tuple alone is a subspace of V_n; it is the "smallest possible" subspace. And of course V_n is by definition a subspace of V_n; it is the "largest possible" subspace.

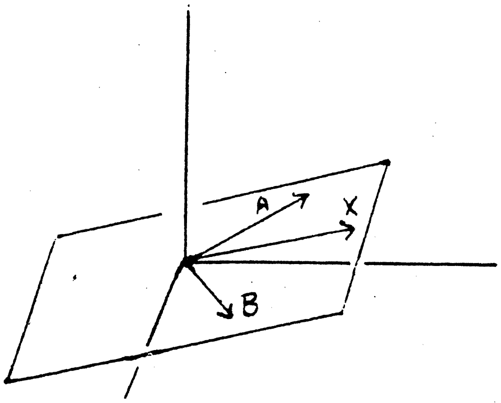
Example 2. Let A be a fixed non-zero vector. The subset of V_n consisting of all vectors X of the form $X = cA$ is a subspace of V_n. It is called the subspace spanned by A. In the case $n = 2$ or 3, it can be pictured as consisting of all vectors lying on a line through the origin.

Example 3. Let A and B be given non-zero vectors that are not parallel. The subset of V_n consisting of all vectors of the form

$$X = cA + dB$$

is a subspace of V_n. It is called the subspace spanned by A and B. In the case $n = 3$, it can be pictured as consisting of all vectors lying in the plane through the origin that contains A and B.





We generalize the construction given in the preceding examples as follows:

Definition. Let $S = \{A_1, \ldots, A_k\}$ be a set of vectors in V_n. A vector X of V_n of the form

$$X = c_1A_1 + \ldots + c_kA_k$$

is called a linear combination of the vectors $A_1,\ldots,A_k$. The set W of all such vectors X is a subspace of V_n, as we will see; it is said to be the subspace spanned by the vectors $A_1,\ldots,A_k$. It is also called the linear span of $A_1,\ldots,A_k$ and denoted by L(S).

Let us show that W is a subspace of V_n. If X and Y belong to W, then

$$X = c_1A_1 + \cdots + c_kA_k \quad \text{and} \quad Y = d_1A_1 + \cdots + d_kA_k,$$

for some scalars c_i and d_i. We compute

$$X + Y = (c_1+d_1)A_1 + \cdots + (c_k+d_k)A_k,$$
$$aX = (ac_1)A_1 + \cdots + (ac_k)A_k,$$

so both $X + Y$ and aX belong to W by definition. Thus W is a subspace of V_n.

Giving a spanning set for W is one standard way of specifying W. Different spanning sets can of course give the same subspace. For example, it is intuitively clear that, for the plane through the origin in Example 3, any two non-zero vectors C and D that are not parallel and lie in this plane will span it. We shall give a proof of this fact shortly.

Example 4. The n-tuple space V_n has a natural spanning set, namely the vectors

$$E_1 = (1,0,0,\ldots,0),$$
$$E_2 = (0,1,0,\ldots,0),$$
$$\cdots$$
$$E_n = (0,0,0,\ldots,1).$$

These are often called the unit coordinate vectors in V_n. It is easy to see that they span V_n, for if $X = (x_1,\ldots,x_n)$ is an element of V_n, then

$$X = x_1E_1 + \cdots + x_nE_n.$$

In the case where $n = 2$, we often denote the unit coordinate vectors E_1 and E_2 in V_2 by $\vec{i}$ and $\vec{j}$, respectively. In the case where $n = 3$, we often denote E_1, E_2, and E_3 by $\vec{i}$, $\vec{j}$, and $\vec{k}$ respectively. They are pictured as in the accompanying figure.

$\vec{j}$
$\vec{i}$
$\vec{k}$
$\vec{j}$
$\vec{i}$

Example 5. The subset W of V_3 consisting of all vectors of the form $(a,b,0)$ is a subspace of V_3. For if X and y are 3-tuples whose third component is 0, so are $X + Y$ and cX. It is easy to see that W is the linear span of $(1,0,0)$ and $(0,1,0)$.

Example 6. The subset of V_3 consisting of all vectors of the form $X = (3a+2b,a-b,a+7b)$ is a subspace of V_3. It consists of all vectors of the form

$$X = a(3,1,1) + b(2,-1,7),$$

so it is the linear span of $(3,1,1)$ and $(2,-1,7)$.

Example 7. The set W of all tuples (x_1,x_2,x_3,x_4) such that

$$3x_1 - x_2 + 5x_3 + x_4 = 0$$

is a subspace of V_4, as you can check. Solving this equation for x_4, we see that a 4-tuple belongs to W if and only if it has the form

$$X = (x_1, x_2, x_3, -3x_1 + x_2 - 5x_3),$$

where x_1 and x_2 and x_3 are arbitrary. This element can be written in the form

$$X = x_1(1,0,0,-3) + x_2(0,1,0,1) + x_3(0,0,1,-5).$$

It follows that $(1,0,0,-3)$ and $(0,1,0,1)$ and $(0,0,1,-5)$ span W.

Exercises

1. Show that the subset of V_3 specified in Example 5 is a subspace of V_3. Do the same for the subset of V_4 specified in Example 7. What can you say about the set of all $(x_1,\ldots,x_n)$ such that $a_1x_1 + \ldots + a_nx_n = 0$ in general? (Here we assume $A = (a_1,\ldots,a_n)$ is not the zero vector.) Can you give a geometric interpretation?

2. In each of the following, let W denote the set of all vectors (x,y,z) in V_3 satisfying the condition given. (Here we use (x,y,z) instead of (x_1,x_2,x_3) for the general element of V_3.) Determine whether W is a subspace of V_3. If it is, draw a picture of it or describe it geometrically, and find a spanning set for W.

(a) $x = 0$.

(b) $x + y = 0$.

(c) $x + y = 1$.

(d) $x = y$ and $2x = z$.

(e) $x = y$ or $2x = z$.

(f) $x^2 - y^2 = 0$.

(g) $x^2 + y^2 = 0$.

3. Consider the set F of all real-valued functions defined on the interval $[a,b]$.

(a) Show that F is a linear space if $f + g$ denotes the usual sum of functions and cf denotes the usual product of a function by a real number. What is the zero vector?

(b) Which of the following are subspaces of F?

(i) All continuous functions.

(ii) All integrable functions.

(iii) All piecewise-monotonic functions.

(iv) All differentiable functions.

(v) All functions f such that $f(a) = 0$.

(vi) All polynomial functions.

Linear independence

Definition. We say that the set $S = \{A_1,\ldots,A_k\}$ of vectors of V_n spans the vector X if X belongs to $L(S)$, that is, if

$$X = c_1A_1 + \ldots + c_kA_k$$

for some scalars c_i. If S spans the vector X, we say that S spans X uniquely if the equations

$$X = \sum_{i=1}^{k} c_iA_i \qquad \text{and} \qquad X = \sum_{i=1}^{k} d_iA_i$$

imply that $c_i = d_i$ for all i.

It is easy to check the following:

Theorem 1. Let $S = \{A_1,\ldots,A_k\}$ be a set of vectors of V_n; let X be a vector in $L(S)$. Then S spans X uniquely if and only if S spans the zero vector $\underline{0}$ uniquely.

Proof. Note that $\underline{0} = \sum 0A_i$. This means that S spans the zero vector uniquely if and only if the equation

$$\underline{0} = \sum_{i=1}^{k} c_i A_i$$

implies that $c_i = 0$ for all i.

Suppose S spans $\underline{0}$ uniquely. To show S spans X uniquely, suppose

$$X = \sum_{i=1}^{k} c_i A_i \quad \text{and} \quad X = \sum_{i=1}^{k} d_i A_i \ .$$

Subtracting, we see that

$$\underline{0} = \sum_{i=1}^{k} (c_i - d_i) A_i \ ,$$

whence $c_i - d_i = 0$, or $c_i = d_i$, for all i.

Conversely, suppose S spans X uniquely. Then

$$X = \sum_{i=1}^{k} x_i A_i$$

for some (unique) scalars x_i. Now if

$$\underline{0} = \sum_{i=1}^{k} c_i A_i \ ,$$

it follows that

$$X = X + \underline{0} = \sum_{i=1}^{k} (x_i + c_i) A_i \ .$$

Since S spans X uniquely, we must have $x_i = x_i + c_i$, or $c_i = 0$, for all i. □

This theorem implies that if S spans one vector of L(S) uniquely, then it spans the zero vector uniquely, whence it spans every vector of L(S) uniquely. This condition is important enough to be given a special name:

Definition. The set $S = \{A_1, \ldots, A_k\}$ of vectors of V_n is said to be linearly independent (or simply, independent) if it spans the zero vector uniquely. The vectors themselves are also said to be independent in this situation.

If a set is not independent, it is said to be dependent.

Example 8. If a subset T of a set S is dependent, then S itself is dependent. For if T spans $\underline{0}$ non-trivially, so does S. (Just add on the additional vectors with zero coefficients.)

This statement is equivalent to the statement that if S is independent, then so is any subset of S.

Example 9. Any set containing the zero vector $\underline{0}$ is dependent. For example, if $S = \{A_1, \ldots, A_k\}$ and $A_1 = \underline{0}$, then

$$\underline{0} = 1A_1 + 0A_2 + \ldots + 0A_k .$$

Example 10. The unit coordinate vectors $E_1, \ldots, E_n$ in V_n span $\underline{0}$ uniquely, so they are independent.

Example 11. Let $S = \{A_1, \ldots, A_k\}$. If the vectors A_i are non-zero and mutually orthogonal, then S is independent. For suppose

$$\underline{0} = \sum_{i=1}^{k} c_i A_i .$$

Taking the dot product of both sides of this equation with A_1 gives the equation

$$0 = c_1 \, A_1 \cdot A_1$$

(since $A_i \cdot A_1 = 0$ for $i \neq 1$). Now $A_1 \neq \underline{0}$ by hypothesis, whence $A_1 \cdot A_1 \neq 0$, whence $c_1 = 0$. Similarly, taking the dot product with A_j for the fixed index j shows that $c_j = 0$.

Sometimes it is convenient to replace the vectors A_i by the vectors $B_i = A_i / \|A_i\|$. Then the vectors $B_1, \ldots, B_k$ are of unit length and are mutually orthogonal. Such a set of vectors is called an orthonormal set. The coordinate vectors $E_1, \ldots, E_n$ form such a set.

Example 12. A set consisting of a single vector A is independent if $A \neq \underline{0}$. A set consisting of two non-zero vectors A,B is independent if and only if the vectors are not parallel. More generally, one has the following result:

Theorem 2. The set $S = \{A_1,\dots,A_k\}$ is independent if and only if none of the vectors A_j can be written as a linear combination of the others.

Proof. Suppose first that one of the vectors equals a linear combination of the others. For instance, suppose that

$$A_1 = c_2A_2 + \cdots + c_kA_k;$$

then the following non-trivial linear combination equals zero:

$$A_1 - c_2A_2 - \cdots - c_kA_k = \underline{0}.$$

Conversely, if

$$c_1A_1 + c_2A_2 + \cdots + c_kA_k = \underline{0},$$

where not all the c_i are equal to zero, we can choose m so that $c_m \neq 0$, and obtain the equation

$$A_m = -(c_1/c_m)A_1 - \cdots - (c_k/c_m)A_k,$$

where the sum on the right extends over all indices different from m. □

Given a subspace W of V_n, there is a very important relation that holds between spanning sets for W and independent sets in W :

Theorem 3. Let W be a subspace of V_n that is spanned by the k vectors $A_1,\dots,A_k$. Then any independent set of vectors in W contains at most k vectors.

Proof. Let $\{B_1,\ldots,B_m\}$ be a set of vectors of W; let $m > k$. We wish to show that these vectors are dependent. That is, we wish to find scalars $x_1,\ldots,x_m$, not all zero, such that

$$\sum_{j=1}^{m} x_j B_j = \underline{0} .$$

Since each vector B_j belongs to W, we can write it as a linear combination of the vectors A_i. We do so, using a "double-indexing" notation for the coefficents, as follows:

$$B_j = a_{1j} A_1 + a_{2j} A_2 + \cdots + a_{kj} A_k .$$

Multiplying the equation by x_j and summing over j, and collecting terms, we have the equation

$$\sum_{j=1}^{m} x_j B_j = \Big(\sum_{j=1}^{m} x_j a_{1j}\Big) A_1 + \Big(\sum_{j=1}^{m} x_j a_{2j}\Big) A_2 + \cdots + \Big(\sum_{j=1}^{m} x_j a_{kj}\Big) A_k .$$

In order for $\sum x_j B_j$ to equal $\underline{0}$, it will suffice if we can choose the x_j so that coefficient of each vector A_i in this equation equals 0. Now the numbers a_{ij} are given, so that finding the x_j is just a matter of solving a (homogeneous) system consisting of k equations in m unknowns. Since $m > k$, there are more unknowns than equations. In this case the system always has a non-trivial solution X (i.e., one different from the zero vector). This is a standard fact about linear equations, which we now prove. □

First, we need a definition.

Definition. Given a homogeneous system of linear equations, as in (*) following, a solution of the system is a vector $(x_1,\ldots,x_n)$ that satisfies each equation of the system. The set of all solutions is a linear subspace of V_n (as you can check). It is called the solution space of the system.

It is easy to see that the solution set is a subspace. If we let

$$A_j = (a_{j_1}, a_{j_2}, \ldots, a_{j_n})$$

be the n-tuple whose components are the coefficents appearing in the j^{th} equation of the system, then the solution set consists of those X such that $A_j \cdot X = 0$ for all j. If X and Y are two solutions, then

$$A_j \cdot (X + Y) = A_j \cdot X + A_j \cdot Y = \underline{0}$$

and

$$A_j \cdot (cX) = c(A_j \cdot X) = 0$$

Thus $X + Y$ and cX are also solutions, as claimed.

Theorem 4. Given a homogeneous system of k linear equations in n unknowns. If k is less than n, then the solution space contains some vector other than **0**.

Proof. We are concerned here only with proving the *existence* of some solution other than **0**, not with actually finding such a solution in practice, nor with finding all possible solutions. (We will study the practical problem in much greater detail in a later section.)

We start with a system of k equations in n unknowns:

$$(*)\qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0, \\ &\vdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n &= 0. \end{aligned}$$

Our procedure will be to reduce the size of this system step-by-step by eliminating first x_1, then x_2, and so on. After $k - 1$ steps, we will be reduced to solving just one equation and this will be easy. But a certain amount of care is needed in the description—for instance, if $a_{11} = \cdots = a_{k1} = 0$, it is nonsense to speak of "eliminating" x_1, since all its coefficients are zero. We have to allow for this possibility.

To begin then, if all the coefficients of x_1 are zero, you may verify that the vector $(1, 0, \ldots, 0)$ is a solution of the system which is different from **0**, and you are done. Otherwise, at least one of the coefficients of x_1 is nonzero, and we may suppose for convenience that the equations have been arranged so that this happens in the first equation, with the result that $a_{11} \neq 0$. We multiply the first equation by the scalar a_{21}/a_{11} and then subtract it from the second, eliminating the x_1-term from the second equation. Similarly, we eliminate the x_1-term in each of the remaining equations. The result is a *new* system of linear equations of the form

$$(**)\qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0, \\ \boxed{\begin{aligned} b_{22}x_2 + \cdots + b_{2n}x_n &= 0, \\ &\vdots \\ b_{k2}x_2 + \cdots + b_{kn}x_n &= 0. \end{aligned}} \end{aligned}$$

Now any solution of this new system of equations is also a solution of the old system (*), because we can recover the old system from the new one: we merely multiply the first equation of the system (**) by the same scalars we used before, and then we *add* it to the corresponding later equations of this system.

The crucial thing about what we have done is contained in the following statement: If the smaller system enclosed in the box above has a solution other than the zero vector, then the larger system (**) also has a solution other than the zero vector [so that the original system (*) we started with has a solution other than the zero vector]. We prove this as follows: Suppose $(d_2, \ldots, d_n)$ is a solution of the smaller system, different from $(0, \ldots, 0)$. We substitute into the first equation and solve for x_1, thereby obtaining the following vector,

$$\left((-1/a_{11})(a_{12}d_2 + \cdots + a_{1n}d_n), d_2, \ldots, d_n\right),$$

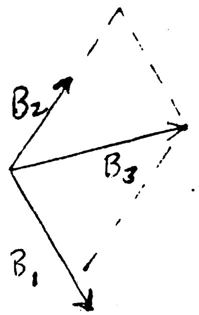
which you may verify is a solution of the larger system (**).

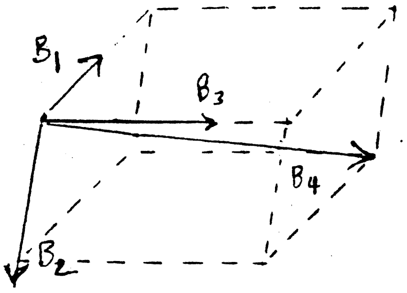
In this way we have reduced the size of our problem; we now need only to prove our theorem for a system of $k - 1$ equations in $n - 1$ unknowns. If we apply this reduction a second time, we reduce the problem to proving the theorem for a system of $k - 2$ equations in $n - 2$ unknowns. Continuing in this way, after $k - 1$ elimination steps in all, we will be down to a system consisting of only one equation, in $n - k + 1$ unknowns. Now $n - k + 1 \geq 2$, because we assumed as our hypothesis that $n > k$; thus our problem reduces to proving the following statement: a "system" consisting of *one* linear homogeneous equation in *two or more* unknowns always has a solution other than **0**.

We leave it to you to show that this statement holds.(Be sure you consider the case where one or more or all of the coefficents are zero.) □

Example 13. We have already noted that the vectors $E_1,\ldots,E_n$ span all of V_n. It follows, for example, that any three vectors in V_2 are dependent, that is, one of them equals a linear combination of the others. The same holds for any four vectors in V_3. The accompanying picture makes these facts plausible.

Similarly, since the vectors $E_1,\ldots,E_n$ are independent, any spanning set of V_n must contain at least n vectors. Thus no two vectors can span V_3, and no set of three vectors can span V_4.





Theorem 5. Let W be a subspace of V_n that does not consist of $\underline{0}$ alone. Then:

(a) The space W has a linearly independent spanning set.

(b) Any two linearly independent spanning sets for W have the same number k of elements; furthermore, $k < n$ unless W is all of V_n.

Proof. (a) Choose $A_1 \neq \underline{0}$ in W. Then the set $\{A_1\}$ is independent. In general, suppose $\{A_1,\ldots,A_i\}$ is an independent set of vectors of W. If this set spans W, we are finished. Otherwise, we can choose a vector A_{i+1} of W that is not in $L(A_1,\ldots,A_i)$. Then the set $\{A_1,\ldots,A_i,A_{i+1}\}$ is independent: For suppose that

$$c_1A_1 + \ldots + c_iA_i + c_{i+1}A_{i+1} = \underline{0}$$

for some scalars c_i not all zero. If $c_{i+1} = 0$, this equation contradicts independence of $\{A_1,\ldots,A_i\}$, while if $c_{i+1} \neq 0$, we can solve this equation for A_{i+1}, contradicting the fact that A_{i+1} does not belong to $L(A_1,\ldots,A_i)$.

Continuing the process just described, we can find larger and larger independent sets of vectors in W. The process stops only when the set we obtain spans W. Does it ever stop? Yes, for W is contained in V_n, and V_n contains no more than n independent vectors. So the process cannot be repeated indefinitely!

(b) Suppose $S = \{A_1,\ldots,A_k\}$ and $T = \{B_1,\ldots,B_j\}$ are two linearly independent spanning sets for W. Because S is independent and T spans W, we must have $k \leq j$, by the preceding theorem. Because S spans W and T is independent, we must have $k \geq j$. Thus $k = j$.

Now V_n contains no more than n independent vectors; therefore we must have $k \leq n$. Suppose that W is not all of V_n. Then we can choose a vector A_{k+1} of V_n that is not in W. By the argument just given, the set $\{A_1,\ldots,A_k,A_{k+1}\}$ is independent. It follows that $k+1 \leq n$, so that $k < n$. □

Definition. Given a subspace W of V_n that does not consist of $\underline{0}$ alone, it has a linearly independent spanning set. Any such set is called a basis for W, and the number of elements in this set is called the dimension of W. We make the convention that if W consists of $\underline{0}$ alone, then the dimension of W is zero.

Example 14. The space V_n has a "natural" basis consisting of the vectors $E_1,\ldots,E_n$. It follows that V_n has dimension n. (Surprise!) There are many other bases for V_n. For instance, the vectors

$$A_1 = (1,0,0,\ldots,0)$$
$$A_2 = (1,1,0,\ldots,0)$$
$$A_3 = (1,1,1,\ldots,0)$$
$$\ldots$$
$$A_n = (1,1,1,\ldots,1)$$

form a basis for V_n, as you can check.

Exercises

1. Consider the subspaces of V_3 listed in Exercise 2, p. A6. Find bases for each of these subspaces, and find spanning sets for them that are not bases.

2. Check the details of Example 14.

3. Suppose W has dimension k. (a) Show that any independent set in W consisting of k vectors spans W. (b) Show that any spanning set for W consisting of k vectors is independent.

4. Let $S = \{A_1,\ldots,A_m\}$ be a spanning set for W. Show that S contains a basis for W. [Hint: Use the argument of Theorem 5.]

5. Let $\{A_1,\ldots,A_k\}$ be an independent set in V_n. Show that this set can be extended to a basis for V_n. [Hint: Use the argument of Theorem 5.]

6. If V and W are subspaces of V_n and V_k, respectively, a function $T : V \rightarrow W$ is called a linear transformation if it satisfes the usual linearity properties:

$$T(X + Y) = T(X) + T(Y),$$

$$T(cX) = cT(X).$$

If T is one-to-one and carries V onto W, it is called a linear isomorphism of vector spaces.

Suppose $A_1,\ldots,A_k$ is a basis for V; let $B_1,\ldots,B_k$ be arbitrary vectors of W. (a) Show there exists a linear transformation $T : V \rightarrow W$ such that $T(A_i) = B_i$ for all i. (b) Show this linear transformation is unique.

7. Let W be a subspace of V_n; let $A_1,\ldots,A_k$ be a basis for W. Let X, Y be vectors of W. Then $X = \sum x_i A_i$ and $Y = \sum y_i A_i$ for unique scalars x_i and y_i. These scalars are called the components of X and Y, respectively, relative to the basis $A_i,\ldots,A_k$.

(a) Note that $X + Y = \sum (x_i + y_i)A_i$ and $cX = \sum (cx_i)A_i$. Conclude that the function $T : V_k \rightarrow W$ defined by $T(x_1,\ldots,x_k) = \sum x_i A_i$ is a linear isomorphism.

(b) Suppose that the basis $A_1,\ldots,A_k$ is an orthonormal basis. Show that $X\cdot Y = \sum x_i y_i$. Conclude that the isomorphism T of (a) preserves the dot product, that is, $T(X)\cdot T(Y) = X\cdot Y$.

8. Prove the following:

Theorem. If W is a subspace of positive dimension of V_n, then W has an orthonormal basis.

Proof. Step 1. Let $B_1,\ldots,B_m$ be mutually orthogonal non-zero vectors in V_n ; let A_{m+1} be a vector not in $L(B_1,\ldots,B_m)$. Given scalars $c_1,\ldots,c_m$, let

$$B_{m+1} = A_{m+1} + c_1B_1 + \ldots + c_mB_m .$$

Show that B_{m+1} is different from $\underline{0}$ and that $L(B_1,\ldots,B_m,B_{m+1}) = L(B_1,\ldots,B_m,A_{m+1})$. Then show that the c_i may be so chosen that B_{m+1} is orthogonal to each of $B_1,\ldots,B_m$.

Step 2. Show that if W is a subspace of V_n of positive dimension, then W has a basis consisting of vectors that are mutually orthogonal. [Hint: Proceed by induction on the dimension of W.]

Step 3. Prove the theorem.

Gauss-Jordan elimination

If W is a subspace of V_n, specified by giving a spanning set for W, we have at present no constructive process for determining the dimension of W nor of finding a basis for W, although we know these exist. There is a simple procedure for carrying out this process; we describe it now.

Definition. The rectangular array of numbers

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & \cdots & \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{bmatrix}$$

is called a matrix of size k by n. The number a_{ij} is called the entry of A in the i^{th} row and j^{th} column. Suppose we let A_i be the vector

$$A_i = (a_{i1}, a_{i2}, \ldots, a_{in})$$

for $i = 1,\ldots k$. Then A_i is just the i^{th} row of the matrix A. The subspace of V_n spanned by the vectors $A_1,\ldots,A_k$ is called the row space of the matrix A.

We now describe a procedure for determining the dimension of this space. It involves applying operations to the matrix A, of the following types:

(1) Interchange two rows of A.

(2) Replace row i of A by itself plus a scalar multiple of another row, say row m.

(3) Multiply row i of A by a non-zero scalar.

These operations are called the elementary row operations. Their usefulness comes from the following fact:

Theorem 6. Suppose B is the matrix obtained by applying a sequence of elementary row operations to A,successively. Then the row spaces of A and B are the same.

Proof. It suffices to consider the case where B is obtained by applying a single row operation to A. Let $A_1,\ldots,A_k$ be the rows of A, and let $B_1,\ldots,B_k$ be the rows of B.

If the operation is of type (1), these two sets of vectors are the same (only their order is changed), so the spaces they span are the same. If the operation is of type (3), then

$$B_i = cA_i \quad \text{and} \quad B_j = A_j \quad \text{for } j \neq i.$$

Clearly, any linear combination of $B_1,\ldots,B_k$ can be written as a linear combination of $A_1,\ldots,A_k$. Because $c \neq 0$, the converse is also true. Finally, suppose the operation is of type (2). Then

$$B_i = A_i + dA_m \quad \text{and} \quad B_j = A_j \quad \text{for } j \neq i.$$

Again, any linear combination of $B_1,\ldots,B_k$ can be written as a linear combination of $A_1,\ldots,A_k$. Because

$$A_i = B_i - dA_m = B_i - dB_m ,$$

and

$$A_j = B_j \quad \text{for } j \neq i ,$$

the converse is also true. □

The Gauss-Jordan procedure consists of applying elementary row operations to the matrix A until it is brought into a form where the dimension of its row space is obvious. It is the following:

<u>Gauss-Jordan</u> <u>elimination.</u> Examine the first column of your matrix.

(I) If this column consists entirely of zeros, nothing needs to be done. Restrict your attention now to the matrix obtained by deleting the first column, and begin again.

(II) If this column has a non-zero entry, exchange rows if necessary to bring it to the top row. Then add multiplesof the top row to the lower rows so as to make all remaining entries in the first column into zeros. Restrict your attention now to the matrix obtained by deleting the first column and first row, and begin again.

The procedure stops when the matrix remaining has only one row.

Let us illustrate the procedure with an example.

Problem. Find the dimension of the row space of the matrix

$$A = \begin{bmatrix} 0 & 1 & 4 & 1 & 2 \\ -1 & -2 & 0 & 9 & -1 \\ 1 & 2 & 0 & -6 & 1 \\ 2 & 5 & 4 & -10 & 4 \end{bmatrix}.$$

Solution. First step. Alternative (II) applies. Exchange rows (1) and (2), obtaining

$$\begin{bmatrix} -1 & -2 & 0 & 9 & -1 \\ 0 & 1 & 4 & 1 & 2 \\ 1 & 2 & 0 & -6 & 1 \\ 2 & 5 & 4 & -10 & 4 \end{bmatrix}$$

Replace row (3) by row (3) + row (1); then replace (4) by (4) + 2 times (1).

$$\left[\begin{array}{c|cccc|} -1 & -2 & 0 & 9 & -1 \\ \cline{2-5} 0 & 1 & 4 & 1 & 2 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 1 & 4 & 8 & 2 \\ \cline{2-5} \end{array}\right]$$

Second step. Restrict attention to the matrix in the box. (II) applies. Replace row (4) by row (4) - row (2), obtaining

$$\left[\begin{array}{cc|ccc|} -1 & -2 & 0 & 9 & -1 \\ 0 & 1 & 4 & 1 & 2 \\ \cline{3-5} 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 7 & 0 \\ \cline{3-5} \end{array}\right]$$

Third step. Restrict attention to the matrix in the box. (I) applies, so nothing needs be done. One obtains the matrix

$$\left[\begin{array}{ccccc}
-1 & -2 & 0 & 9 & -1 \\
0 & 1 & 4 & 1 & 2 \\ \cline{4-5}
0 & 0 & 0 & \multicolumn{1}{|c}{3} & \multicolumn{1}{c|}{0} \\
0 & 0 & 0 & \multicolumn{1}{|c}{7} & \multicolumn{1}{c|}{0} \\ \cline{4-5}
\end{array}\right]$$

Fourth step. Restrict attention to the matrix in the box. (II) applies. Replace row (4) by row (4) $- \frac{7}{3}$ row (3) , obtaining

$$B = \left[\begin{array}{ccccc}
\textcircled{-1} & -2 & 0 & 9 & -1 \\ \cline{1-1}
0 & \multicolumn{1}{|c}{\textcircled{1}} & 4 & 1 & 2 \\ \cline{2-3}
0 & 0 & 0 & \multicolumn{1}{|c}{\textcircled{3}} & 0 \\ \cline{4-5}
0 & 0 & 0 & 0 & 0
\end{array}\right]$$

The procedure is now finished. The matrix B we end up with is in what is called echelon or "stair-step"form. The entries beneath the steps are zero. And the entries -1, 1, and 3 that appear at the "inside corners" of the stairsteps are non-zero. These entries that appear at the "inside corners" of the stairsteps are often called the pivots in the echelon form.

You can check readily that the non-zero rows of the matrix B are independent. (We shall prove this fact later.) It follows that the non-zero rows of the matrix B form a basis for the row space of B, and hence a basis for the row space of the original matrix A. Thus this row space has dimension 3.

The same result holds in general. If by elementary operations you reduce the matrix A to the echelon form B, then the non-zero rows are B are independent, so they form a basis for the row space of B, and hence a basis for the row space of A.

Now we discuss how one can continue to apply elementary operations to reduce the matrix B to an even nicer form. The procedure is this:

Begin by considering the last non-zero row. By adding multiples of this row to each row above it, one can bring the matrix to the form where each entry lying above the pivot in this row is zero. Then continue the process, working now with the next-to-last non-zero row. Because all the entries above the last pivot are already zero, they remain zero as you add multiples of the next-to-last non-zero row to the rows above it. Similarly one continues. Eventually the matrix reaches the form where all the entries that are directly above the pivots are zero. (Note that the stairsteps do not change during this process, nor do the pivots themselves.)

Applying this procedure in the example considered earlier, one brings the matrix B into the form

$$C = \begin{bmatrix} -1 & 0 & 8 & 0 & 3 \\ 0 & 1 & 4 & 0 & 2 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that up to this point in the reduction process , we have used only elementary row operations of types (1) and (2). It has not been necessary to multiply a row by a non-zero scalar. This fact will be important later on.

We are not yet finished. The final step is to multiply each non-zero row by an appropriate non-zero scalar, chosen so as to make the pivot entry into 1. This we can do, because the pivots are non-zero. At the end of this process, the matrix is in what is called reduced echelon form.

The reduced echelon form of the matrix C above is the matrix

$$D = \begin{bmatrix} 1 & 0 & -8 & 0 & -3 \\ 0 & 1 & 4 & 0 & 2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

As we have indicated, the importance of this process comes from the following theorem:

Theorem 7. Let A be a matrix; let W be its row space. Suppose we transform A by elementary row operations into the echelon matrix B, or into the reduced echelon matrix D. Then the non-zero rows of B are a basis for W, and so are the non-zero rows of D.

Proof. The rows of B span W, as we noted before; and so do the rows of D. It is easy to see that no non-trivial linear combination of the non-zero rows of D equals the zero vector, because each of these rows has an entry of 1 in a position where the others all have entries of 0. Thus the dimension of W equals the number r of non-zero rows of D. This is the same as the number of non-zero rows of B. If the rows of B were not independent, then one would equal a linear combination of the others. This would imply that the row space of B could be spanned by fewer than r rows, which would imply that its dimension is less than r.

Exercises

1. Find bases for the row spaces of the following matrices:

$$A = \begin{bmatrix} 1 & 1 & 3 \\ 2 & -1 & 4 \\ 0 & -1 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} 3 & 2 & 1 \\ 5 & 3 & 3 \\ 7 & 4 & 5 \\ 1 & 1 & -1 \end{bmatrix}$$

$$B = \begin{bmatrix} 3 & 2 & 1 \\ 5 & 3 & 3 \\ 1 & 1 & -1 \end{bmatrix} \qquad E = \begin{bmatrix} 1 & -2 & 1 & 2 \\ 2 & 3 & -1 & -5 \\ 4 & -1 & 1 & -1 \\ 5 & -3 & 2 & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} 3 & 2 & 1 \\ 5 & 3 & 3 \\ 7 & 4 & 5 \end{bmatrix}$$

2. Reduce the matrices in Exercise 1 to reduced echelon form.
Save your answers for later use!

*3. Prove the following:

Theorem. The reduced echelon form of a matrix is unique.

Proof. Let D and D' be two reduced echelon matrices, whose rows span the same subspace W of V_n. We show that D = D'.

Let $R_1,\ldots,R_k$ be the non-zero rows of D ; and suppose that the pivots (first non-zero entries) in these rows occur in columns $j_1,\ldots,j_k$, respectively.

(a) Show that the pivots of D' occur in the columns $j_1,\ldots,j_k$. [Hint: Let R be a row of D'; suppose its pivot occurs in column p. We have $R = c_1R_1 + \ldots + c_kR_k$ for some scalars c_i . (Why?) Show that $c_i = 0$ if $j_i < p$. Derive a contradiction if p is not equal to any of $j_1,\ldots,j_k$.]

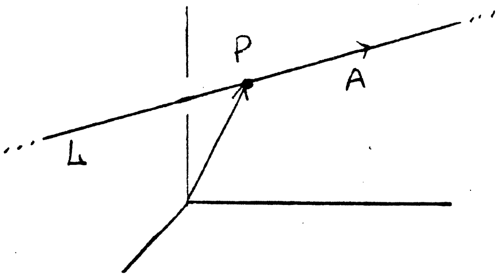
(b) If R is a row of D' whose pivot occurs in column j_m , show that $R = R_m$. [Hint: We have $R = c_1R_1 + \ldots + c_kR_k$ for some scalars c_i . Show that $c_i = 0$ for $i \neq m$, and $c_m = 1$.]

Parametric equations of lines and planes in V_n

Given n-tuples P and A, with $A \neq \underline{0}$, the line through P determined by A is defined to be the set of all points X such that

$$(*) \qquad X = P + tA$$

for some scalar t. It is denoted by $L(P;A)$. The vector A is called a direction vector for the line. Note that if $P = \underline{0}$, then L is simply the 1-dimensional subspace of V_n spanned by A.



The equation (*) is often called a parametric equation for the line, and t is called the parameter in this equation. As t ranges over all real numbers, the corresponding point X ranges over all points of the line L. When $t = 0$, then $X = P$; when $t = 1$, then $X = P + A$; when $t = \frac{1}{2}$, then $X = P + \frac{1}{2}A$; and so on. All these are points of L.

Occasionally, one writes the vector equation out in scalar form as follows:

$$\begin{aligned} x_1 &= p_1 + ta_1 \\ x_2 &= p_2 + ta_2 \\ &\dots \\ x_n &= p_n + ta_n \end{aligned}$$

where $P = (p_1,\ldots,p_n)$ and $A = (a_1,\ldots,a_n)$. These are called the *scalar parametric equations* for the line.

Of course, there is no uniqueness here; a given line can be represented by many different parametric equations. The following theorem makes this result precise:

Theorem 8. The lines $L(P;A)$ *and* $L(Q;B)$ *are equal if and only if they have a point in common and* A *is parallel to* B.

Proof. If $L(P;A) = L(Q;B)$, then the lines obviously have a point in common. Since P and $P + A$ lie on the first line they also lie on the second line, so that

$$P = Q + t_1B \qquad \text{and} \qquad P + A = Q + t_2B$$

for distinct scalars t_1 and t_2. Subtracting, we have $A = (t_2-t_1)B$, so A is parallel to B.

Conversely, suppose the lines intersect in a point R, and suppose A and B are parallel. We are given that

$$P + t_1A = R = Q + t_2B$$

for some scalars t_1 and t_2, and that $A = cB$ for some $c \neq 0$. We can solve these equations for P in terms of Q and B:

$$P = Q + t_2B - t_1A = Q + (t_2-t_1c)B.$$

Now, given any point $X = P + tA$ of the line $L(P;A)$, we can write

$$X = P + tA = Q + (t_2-t_1c)B + tcB.$$

Thus X belongs to the line $L(Q;B)$.

Thus every point of $L(P;A)$ belongs to $L(Q;B)$. The symmetry of the argument shows that the reverse holds as well. □

Definition. It follows from the preceding theorem that given a line, its direction vector is uniquely determined up to a non-zero scalar multiple. We define two lines to be *parallel* if their direction vectors are parallel.

Corollary 9. Distinct parallel lines cannot intersect.

Corollary 10. Given a line L and a point Q, there is exactly one line containing Q that is parallel to L.

Proof. Suppose L is the line $L(P;A)$. Then the line $L(Q;A)$ contains Q and is parallel to L. By Theorem 8, any other line containing Q and parallel to L is equal to this one. □

Theorem 11. Given two distinct points P and Q, there is exactly one line containing them.

Proof. Let $A = Q - P$; then $A \neq \underline{0}$. The line $L(P;A)$ contains both P (since $P = P + 0A$) and Q (since $Q = P + 1A$).

Now suppose $L(R;B)$ is some other line containing P and Q. Then

$$P = R + t_1 B,$$
$$Q = R + t_2 B,$$

for distinct scalars t_1 and t_2. It follows that

$$Q - P = (t_2 - t_1)B,$$

so that the vector $A = Q - P$ is parallel to B. It follows from Theorem 8 that

$$L(R;B) = L(P;A). \quad \square$$

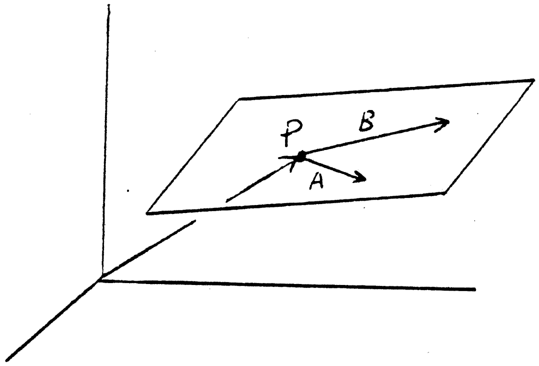
Now we study planes in V_n.

Definition. If P is a point of V_n and if A and B are independent vectors of V_n, we define the plane through P determined by A and B to be the set of all points X of the form

$$(*) \qquad X = P + sA + tB,$$

where s and t run through all real numbers. We denote this plane by $M(P;A,B)$.

The equation (*) is called a parametric equation for the plane, and s and t are called the parameters in this equation. It may be written out as n scalar equations, if desired. When $s = t = 0$, then $X = P$; when $s = 1$ and $t = 0$, then $X = P + A$; when $s = 0$ and $t = 1$, then $X = P + B$; and so on.



Note that if $P = \underline{0}$, then this plane is just the 2-dimensional subspace of V_n spanned by A and B.

Just as for lines, a plane has many different parametric representations. More precisely, one has the following theorem:

Theorem 12. The planes $M(P;A,B)$ and $M(Q;C,D)$ are equal if and only if they have a point in common and the linear span of A and B equals the linear span of C and D.

Proof. If the planes are equal, they obviously have a point in common. Furthermore, since P and $P + A$ and $P + B$ all lie on the first plane, they lie on the second plane as well. Then

$$P = Q + s_1C + t_1D,$$
$$P + A = Q + s_2C + t_2D,$$
$$P + B = Q + s_3C + t_3D,$$

are some scalars s_i and t_i. Subtracting, we see that

$$A = (s_2-s_1)C + (t_2-t_1)D,$$
$$B = (s_3-s_1)C + (t_3-t_1)D.$$

Thus A and B lie in the linear span of C and D. Symmetry shows that C and D lie in the linear span of A and B as well. Thus these linear spans are the same.

Conversely, suppose that the planes intersect in a point R and that $L(A,B) = L(C,D)$. Then

$$P + s_1A + t_1B = R = Q + s_2C + t_2D$$

for some scalars s_i and t_i. We can solve this equation for P as follows:

$$P = Q + (\text{linear combination of } A,B,C,D).$$

Then if X is any point of the first plane $M(P;A,B)$, we have

$$\begin{aligned} X &= P + sA + tB \qquad \text{for some scalars } s \text{ and } t, \\ &= Q + (\text{linear combination of } A,B,C,D) + sA + tB \\ &= Q + (\text{linear combination of } C,D), \end{aligned}$$

since A and B belong to $L(C,D)$.

Thus X belongs to $M(Q;C,D)$.

Symmetry of the argument shows that every point of $M(Q;C,D)$ belongs to $M(P;A,B)$ as well. □

Definition. Given a plane $M = M(P;A,B)$, the vectors A and B are not uniquely determined by M, but their linear span is. We say the planes $M(P;A,B)$ and $M(Q;C,D)$ are parallel if $L(A,B) = L(C,D)$.

Corollary 13. Two distinct parallel planes cannot intersect.

Corollary 14. Given a plane M and a point Q, there is exactly one plane containing Q that is parallel to M.

Proof. Suppose $M = M(P;A,B)$. Then $M(Q;A,B)$ is a plane that contains Q and is parallel to M. By Theorem 12 any other plane containing Q parallel to M is equal to this one. □

Definition. We say three points P,Q,R are collinear if they lie on a line.

Lemma 15. The points P,Q,R are collinear if and only if the vectors $Q-P$ and $R-P$ are dependent (i.e., parallel).

Proof. The line $L(P; Q-P)$ is the one containing P and Q, and the line $L(P;R-P)$ is the one containing P and R. If $Q-P$ and $R-P$ are parallel, these lines are the same, by Theorem 8, so P, Q, and R are collinear. Conversely, if P, Q, and R are collinear, these lines must be the same, so that $Q-P$ and $R-P$ must be parallel. □

Theorem 16. Given three non-collinear points P, Q, R, there is exactly one plane containing them.

Proof. Let $A = Q - P$ and $B = R - P$; then A and B are independent. The plane $M(P; A,B)$ contains P and $P + A = Q$ and $P + B = R$.

Now suppose $M(S;C,D)$ is another plane containing P, Q, and R. Then

$$P = S + s_1C + t_1D$$
$$Q = S + s_2C + t_2D$$
$$R = S + s_3C + t_3D$$

for some scalars s_i and t_i. Subtracting, we see that the vectors $Q - P = A$ and $R - P = B$ belong to the linear span of C and D. By symmetry, C and D belong to the linear span of A and B. Then Theorem 12 implies that these two planes are equal.

Exercises

1. We say the line L is parallel to the plane $M = M(P;A,B)$ if the direction vector of L belongs to $L(A,B)$. Show that if L is parallel to M and intersects M, then L is contained in M.

2. Show that two vectors A_1 and A_2 in V_n are linearly dependent if and only if they lie on a line through the origin.

3. Show that three vectors A_1, A_2, A_3 in V_n are linearly dependent if and only if they lie on some plane through the origin.

4. Let $P = (1,0,-1)$, $Q = (0,0,0)$, $R = (-2,5,0)$. Let $A = (1,-1,0)$, $B = (2,0,1)$.

(a) Find parametric equations for the line through P and Q, and for the line through R with direction vector A. Do these lines intersect?

(b) Find parametric equations for the plane through P, Q, and R, and for the plane through P determined by A and B.

5. Let L be the line in V_3 through the points $P = (1,0,2)$ and $Q = (-1,1,3)$. Let L' be the line through $\underline{0}$ parallel to the vector $A = (3,1,-1)$. Find parametric equations for the line that intersects both L and L' and is orthogonal to both of them.

Parametric equations for k-planes in V_n.

Following the pattern for lines and planes, one can define, more generally, a k-plane in V_n as follows:

Definition. Given a point P of V_n and a set $A_1,\ldots,A_k$ of k independent vectors in V_n, we define the k-plane through P determined by $A_1,\ldots,A_k$ to be the set of all vectors X of the form

$$X = P + t_1A_1 + \cdots + t_kA_k,$$

for some scalars t_i. We denote this set of points by $M(P;A_1,\ldots,A_k)$.

Said differently, X is in the k-plane $M(P;A_1,\ldots,A_k)$ if and only if $X - P$ is in the linear span of $A_1,\ldots,A_k$.

Note that if $P = \underline{0}$, then this k-plane is just the k- dimensional linear subspace of V_n spanned by $A_1,\ldots,A_k$.

Just as with the case of lines (1-planes) and planes (2-planes), one has the following results:

Theorem 17. Let $M_1 = M(P;A_1,\ldots,A_k)$ and $M_2 = M(Q;B_1,\ldots,B_k)$ be two k-planes in V_n. Then $M_1 = M_2$ if and only if they have a point in common and the linear span of $A_1,\ldots,A_k$ equals the linear span of $B_1,\ldots,B_k$.

Definition. We say that the k-planes M_1 and M_2 of this theorem are parallel if the linear span of $A_1,\ldots,A_k$ equals the linear span of $B_1,\ldots,B_k$.

Theorem 18. Given a k-plane M in V_n and a point Q, there is exactly one k-plane in V_n containing Q and parallel to M.

Lemma 19. Given points $P_0,\ldots,P_k$ in V_n, they are contained in a plane of dimension less than k if and only if the vectors

$P_1 - P_0, \ldots, P_k - P_0$ are dependent.

<u>Theorem</u> <u>20.</u> Given $k+1$ distinct points $P_0, \ldots, P_k$ in V_n. If these points do not lie in any plane of dimension less than k, then there is exactly one k-plane containing them; it is the k-plane

$$M(P_0; P_1 - P_0, \ldots, P_k - P_0).$$

More generally, we make the following definition:

<u>Definition.</u> If $M_1 = M(P; A_1, \ldots, A_k)$ is a k-plane, and $M_2 = M(Q; B_1, \ldots, B_m)$ is an m-plane, in V_n, and if $k \leq m$, we say M_1 is <u>parallel</u> to M_2 if the linear span of $A_1, \ldots, A_k$ is contained in the linear span of $B_1, \ldots, B_m$.

<u>Exercises</u>

1. Prove Theorems 17 and 18.

2. Prove Theorems 19 and 20.

3. Given the line $L = L(\underline{O}; A)$ in V_3, where $A = (1,-1,2)$. Find parametric equations for a 2-plane containing the point $P = (1,1,1)$ that is parallel to L. Is it unique? Can you find such a plane containing both the point P and the point $Q = (-1,0,2)$?

4. Given the 2-plane M_1 in V_4 containing the points $P = (1,-1,2,-1)$ and $Q = (0,1,1,0)$ and $R = (1,1,0,3)$. Find parametric equations for a 3-plane in V_4 that contains the point $S = (1,1,1,1)$ and is parallel to M_1. Is it unique? Can you find such a 3-plane that contains both S and the point $T = (0,1,0,2)$?

18.024 Multivariable Calculus with Theory
Spring 2011

Matrices

We have already defined what we mean by a matrix. In this section, we introduce algebraic operations into the set of matrices.

Definition. If A and B are two matrices of the same size, say k by n, we define $A + B$ to be the k by n matrix obtained by adding the corresponding entries of A and B, and we define cA to be the matrix obtained from A by multiplying each entry of A by c. That is, if a_{ij} and b_{ij} are the entries of A and B, respectively, in row i and column j, then the entries of $A + B$ and of cA in row i and column j are

$$a_{ij} + b_{ij} \qquad \text{and} \qquad ca_{ij} ,$$

respectively.

Note that for fixed k and n, the set of all k by n matrices satisfies all the properties of a linear space. This fact is hardly surprising, for a k by n matrix is very much like a $k\cdot n$ tuple; that only difference is that the components are written in a rectangular array instead of a linear array.

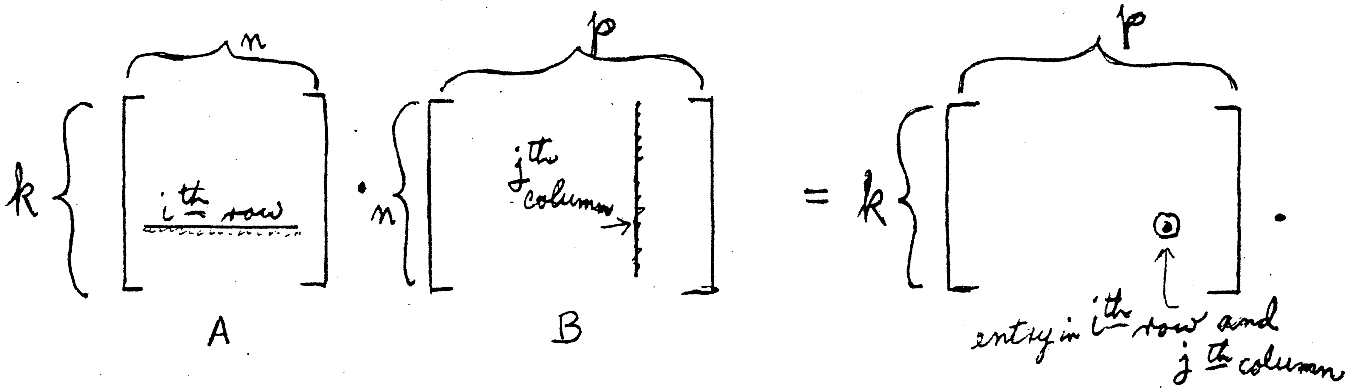
Unlike tuples, however, matrices have a further operation, a product operation. It is defined as follows:

Definition. If A is a k by n matrix, and B is an n by p matrix, we define the product $D = A\cdot B$ of A and B to be the matrix of size k by p whose entry d_{ij} in row i and column j is given by the formula

$$d_{ij} = \sum_{s=1}^{n} a_{is}b_{sj}.$$

Here $i = 1,\ldots,k$ and $j = 1,\ldots,p$.

The entry d_{ij} is computed, roughly speaking, by taking the "dot product" of the $i^{\underline{th}}$ row of A with the $j^{\underline{th}}$ column of B. Schematically,



This definition seems rather strange, but it is in fact extremely useful. Motivation will come later! One important justification for this definition is the fact that this product operation satisfies some of the familar "laws of algebra" :

Theorem 1. Matrix multiplication has the following properties: Let A, B, C , D *be matrices.*

(1) (*Distributivity*) *If* A·(B + C) *is defined, then*

$$A\cdot(B + C) = A\cdot B + A\cdot C .$$

Similarly, if (B + C)·D *is defined, then*

$$(B + C)\cdot D = B\cdot D + C\cdot D.$$

(2) (*Homogeneity*) *If* A·B *is defined, then*

$$(cA)\cdot B = c(A\cdot B) = A\cdot(cB) .$$

(3) (*Associativity*) *If* A·B *and* B·C *are defined, then*

$$A\cdot(B\cdot C) = (A\cdot B)\cdot C .$$

(4) (*Existence of identities*) *For each* m, *there is an* m *by* m *matrix* I_m *such that for matrices* A *and* B, *we have*

$$I_m \cdot A = A \quad \text{and} \quad B \cdot I_m = B$$

whenever these products are defined.

Proof. We verify the first distributivity formula. In order for $B + C$ to be defined, B and C must have the same size, say n by p. Then in order for $A \cdot (B + C)$ to be defined, A must have n columns. Suppose A has size k by n. Then $A \cdot B$ and $A \cdot C$ are defined and have size k by p; thus their sum is also defined. The distributivity formula now follows from the equation

$$\sum_{s=1}^{n} a_{is}(b_{sj}+c_{sj}) = \sum_{s=1}^{n} a_{is}b_{sj} + \sum_{s=1}^{n} a_{is}c_{sj}.$$

The other distributivity formula and the homogeneity formula are proved similarly. We leave them as exercises.

Now let us verify associativity.

If A is k by n and B is n by p, then $A \cdot B$ is k by p. The product $(A \cdot B) \cdot C$ is thus defined provided C has size p by q. The product $A \cdot (B \cdot C)$ is defined in precisely the same circumstances. Proof of equality is an exercise in summation symbols: The entry in row i and column j of $(A \cdot B) \cdot C$ is

$$\sum_{t=1}^{p} \left(\sum_{s=1}^{n} a_{is}b_{st}\right)c_{tj} ;$$

and the corresponding entry of $A \cdot (B \cdot C)$ is

$$\sum_{s=1}^{n} a_{is}\left(\sum_{t=1}^{p} b_{st}c_{tj}\right).$$

These two expressions are equal.

Finally, we define matrices I_m that act as identity elements. Given m, let I_m be the m by m matrix whose general entry is δ_{ij}, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. The matrix I_m is a square matrix that has 1's down the "main diagonal" and 0's elsewhere. For instance, I_4 is the matrix

$$I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Now the product $I_m \cdot A$ is defined in the case where A has m rows. In this case, the general entry of the product $C = I_m \cdot A$ is given by the equation

$$c_{ij} = \sum_{s=1}^{m} \delta_{is} a_{sj} .$$

Let i and j be fixed. Then as s ranges from 1 to m, all but one of the terms of this summation vanish. The only one that does not vanish is the one for which $s = i$, and in that case $\delta_{is} = 1$. We conclude that

$$c_{ij} = 0 + \cdots + 0 + \delta_{ii} a_{ij} + 0 + \cdots + 0 = a_{ij}.$$

An entirely similar proof shows that $B \cdot I_m = B$ if B has m columns. □

<u>Remark</u>. If $A \cdot B$ is defined, then $B \cdot A$ need not be defined. And even if it is defined, the two products need not be equal. For example,

$$\begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ -1 & -3 \end{bmatrix}, \quad \text{and}$$

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

Remark. A natural question to ask at this point concerns the existence of multiplicative inverses in the set of matrices. We shall study the answer to this question in a later section.

Exercises

1. Verify the other half of distributivity.

2. Verify homogeneity of matrix multiplication.

3. Show the identity element is unique. [Hint: If I'_m and I''_m are two possible choices for the identity element of size m by m, compute $I'_m \cdot I''_m$.]

4. Find a non-zero 2 by 2 matrix A such that $A \cdot A$ is the zero matrix. Conclude that there is no matrix B such that $B \cdot A = I_2$.

5. Consider the set of m by m matrices; it is closed under addition and multiplication. Which of the field axioms (the algebraic axioms that the real numbers satisfy) hold for this set? (Such an algebraic object is called in modern algebra a "ring with identity.")

Systems of linear equations

Given numbers a_{ij} for $i = 1,\ldots,k$ and $j = 1,\ldots,n$, and given numbers $c_1,\ldots,c_k$, we wish to study the following, which is called a system of k linear equations in n unknowns:

$$(*)\quad \begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = c_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = c_2 \\ \qquad\qquad\ldots \\ a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kn}x_n = c_k. \end{cases}$$

A solution of this system is a vector $X = (x_1,\ldots,x_n)$ that satisfies each equation. The solution set of the system consists of all such vectors; it is a subset of V_n .

We wish to determine whether this system has a solution, and if so, what the nature of the general solution is. Note that we are not assuming anything about the relative size of k and n; they may be equal, or one may be larger than the other.

Matrix notation is convenient for dealing with this system of equations. Let A denote the k by n matrix whose entry in row i and column j is a_{ij}. Let X and C denote the matrices

$$X = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \qquad \text{and} \qquad C = \begin{bmatrix} c_1 \\ \cdot \\ \cdot \\ \cdot \\ c_k \end{bmatrix} .$$

These are matrices with only one column; accordingly, they are called column matrices . The system of equations (*) can now be written in matrix form as

$$A\cdot X = C.$$

A solution of this matrix equation is now, strictly speaking, a column matrix rather than an n-tuple. However, one has a natural correspondence

$$(x_1,\ldots,x_n) \longrightarrow \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

between n-tuples and column matrices of size n by 1. It is a one-to-one correspondence, and even the vector space operations correspond. What this means is that we can identify V_n with the space of all n by 1 matrices if we wish; all this amounts to is a change of notation.

Representing elements of V_n as column matrices is so convenient that we will adopt it as a convention throughout this section, whenever we wish.

Example 1. Consider the system

$$\begin{aligned} 2x + y + z &= 1 \\ x - y \phantom{{}+z} &= 2 \\ 3x \phantom{{}+y} + z &= 0 \ . \end{aligned}$$

[Here we use x, y, z for the unknowns instead of x_1, x_2, x_3, for convenience.] This system has no solution, since the sum of the first two equations contradicts the third equation.

Example 2. Consider the system

$$\begin{aligned} 2x + y + z &= 1 \\ x - y \quad &= 2 \\ 3x \quad + z &= 3 \end{aligned}$$

This system has a solution; in fact, it has more than one solution. In solving this sytem, we can ignore the third equation, since it is the sum of the first two. Then we can assign a value to y arbitrarily, say $y = t$, and solve the first two equations for x and z. We obtain the result

$$\begin{aligned} x &= 2 + y = 2 + t \\ y &= t \\ z &= 1 - 2x - y = 1 = 2(2+t) - t = -3 - 3t. \end{aligned}$$

The solution set consists of all matrices of the form

$$X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2+t \\ t \\ -3-3t \end{bmatrix} .$$

Shifting back to tuple notation, we can say that the solution set consists of all vectors X such that

$$X = (x,y,z) = (2+t, t, -3-3t)$$

or

$$X = (2,0,-3) + t(1,1,-3) .$$

This expression shows that the solution set is a line in V_3, and in "solving" the system, we have written the equation of this line in parametric form.

Now we tackle the general problem. We shall prove the following result:

Suppose one is given a system of k linear equations in n unknowns. Then the solution set is either (1) empty, or (2) it consists of a single point, or (3) it consists of the points of an m-plane in V_n, for some $m > 0$. In case (1), we say the system is inconsistent, meaning that it has no solution.

In case (2), the solution is unique. And in case (3), the system has infinitely many solutions.

We shall apply Gauss-Jordan elimination to prove these facts. The crucial result we shall need is stated in the following theorem:

Theorem 2. Consider the system of equations $A\cdot X = C$, where A is a k by n matrix and C is a k by 1 matrix. Let B be the matrix obtained by applying an elementary row operation to A, and let C' be the matrix obtained by applying the same elementary row operation to C. Then the solution set of the system $B\cdot X = C'$ is the same as the solution set of the system $A\cdot X = C$.

Proof. Exchanging rows i and j of both matrices has the effect of simply exchanging equations i and j of the system. Replacing row i by itself plus c times row j has the effect of replacing the i^{th} equation by itself plus c times the j^{th} equation. And multiplying row i by a non-zero scalar d has the effect of multiplying both sides of the i^{th} equation by d. Thus each solution of the first system is also a solution of the second system.

Now we recall that the elementary operations are invertible. Thus the system $A\cdot X = C$ can be obtained by applying an elementary operation to both sides of the equation $B\cdot X = C'$. It follows that every solution of the second system is a solution of the first system.

Thus the two solution sets are identical. □

We consider first the case of a homogeneous system of equations, that is, a system whose matrix equation has the form

$$A\cdot X = \underline{0} .$$

In this case, the system obviously has at least one solution, namely the trivial solution $X = \underline{0}$. Furthermore, we know that the set of solutions is a linear subspace of V_n, that is, an m-plane through the origin for some m. We wish to determine the dimension of this solution space, and to find a basis for it.

Definition. Let A be a matrix of size k by n. Let W be the row space of A; let r be the dimension of W. Then r equals the number of non-zero rows in the echelon form of A. It follows at once that $r \leq k$. It is also true that $r \leq n$, because W is a subspace of V_n. The number r is called the rank of A (or sometimes the row rank of A).

Theorem 3. Let A be a matrix of size k by n. Let r be the rank of A. Then the solution space of the system of equations $A \cdot X = \underline{0}$ is a subspace of V_n of dimension $n - r$.

Proof. The preceding theorem tells us that we can apply elementary operations to both the matrices A and $\underline{0}$ without changing the solution set. Applying elementary operations to $\underline{0}$ leaves it unchanged, of course.

So let us apply elementary operations to A so as to bring A into reduced echelon form D, and consider the system $D \cdot X = \underline{0}$. The number of non-zero rows of D equals the dimension of the row space of A, which is r. Now for a zero row of D, the corresponding equation is automatically satisfied, no matter what X we choose. Only the first r equations are relevant.

Suppose that the pivots of D appear in columns $j_1, \ldots, j_r$. Let J denote the set of indices $\{j_1, \ldots, j_r\}$ and let K consist of the remaining indices from the set $\{1, \ldots, n\}$. Each unknown x_j for which j is in J appears with a non-zero coefficient in only one of the equations of the system $D \cdot X = \underline{0}$. Therefore, we can "solve" for each of these unknowns in terms of the remaining unknowns x_k, for k in K. Substituting these expressions for $x_{j_1}, \ldots, x_{j_r}$ into the n-tuple $X = (x_1, \ldots, x_n)$, we see that the general solution of the system can be written as a vector of which each component is a linear combination of the x_k, for k in K. (Of course, if k is in K, then the linear combination that appears in the k^{th} component consists merely of the single term x_k!)

Let us pause to consider an example.

Example 3. Let A be the 4 by 5 matrix given on p.A20. The equation $A\cdot X = \underline{0}$ represents a system of 4 equations in 5 unknowns. Now A reduces by row operations to the reduced echelon matrix

$$D = \begin{bmatrix} 1 & 0 & -8 & 0 & -3 \\ 0 & 1 & 4 & 0 & 2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} .$$

Here the pivots appear in columns 1,2 and 4; thus J is the set $\{1,2,4\}$ and K is the set $\{3,5\}$. The unknowns x_1, x_2, and x_4 each appear in only one equation of the system. We solve for theese unknowns in terms of the others as follows:

$$\begin{aligned} x_1 &= 8x_3 + 3x_5 \\ x_2 &= -4x_3 - 2x_5 \\ x_4 &= 0. \end{aligned}$$

The general solution can thus be written (using tuple notation for convenience)

$$X = (8x_3 + 3x_5, -4x_3 - 2x_5, x_3, 0, x_5), \quad \text{or}$$

$$X = (8x_3, -4x_3, x_3, 0, 0) + (3x_5, -2x_5, 0, 0, x_5), \quad \text{or}$$

$$X = x_3(8,-4,1,0,0) + x_5(3,-2,0,0,1) .$$

The solution space is thus spanned by two vectors $(8,-4,1,0,0)$ and $(3,-2,0,0,1)$.

The same procedure we followed in this example can be followed in general. Once we write X as a vector of which each component is a linear combination of the x_k, then we can write it as a sum of vectors each of which involves only one of the unknowns x_k, and then finally as a linear combination, with coefficients x_k, of vectors in V_n. There are of course $n - r$ of the unknowns x_k, and hence $n - r$ of these vectors.

It follows that the solution space of the system has a spanning set consisting of $n - r$ vectors. We now show that these vectors are independent; then the theorem is proved. To verify independence, it suffices to show that if we take the vector X, which equals a linear combination with coefficents x_k of these vectors, then $X = \underline{0}$ if and only if each x_k (for k in K) equals 0. This is easy. Consider the first expression for X that we wrote down, where each component of X is a linear combination of the unknowns x_k. The k^{th} component of X is simply x_k. It follows that the equation $X = \underline{0}$ implies in particular that for each k in K, we have $x_k = 0$.

For example, in the example we just considered, we see that the equation $X = \underline{0}$ implies that $x_3 = 0$ and $x_5 = 0$, because x_3 is the third component of X and x_5 is the fifth component of X. □

This proof is especially interesting because it not only gives us the dimension of the solution space of the system, but it also gives us a method for finding a basis for this solution space, in practice. All that is involved is Gauss-Jordan elimination.

<u>Corollary 4.</u> Let A be a k by n matrix. If the rows of A are independent, then the solution space of the system $A \cdot X = \underline{0}$ has dimension $n - k$. □

Now we consider the case of a general system of linear equations, of the form $A \cdot X = C$. For the moment, we assume that the system has at least one solution, and we determine what the general solution looks like in this case.

<u>Theorem 5.</u> Let A be a k by n matrix. Let r equal the rank of A. If the system $A \cdot X = C$ has a solution, then the solution set is a plane in V_n of dimension $m = n - r$.

Proof. Let $X = P$ be a solution of the system. Then $A \cdot P = C$. If X is a column matrix such that $A \cdot X = C$, then $A \cdot (X - P) = \underline{0}$, and conversely. The solution space of the system $A \cdot X = \underline{0}$ is a subspace of V_n of dimension $m = n - r$; let $A_1, \ldots, A_m$ be a basis for it. Then X is a solution of the system $A \cdot X = C$ if and only if $X - P$ is a linear combination of the vectors A_i, that is, if and only if

$$X = P + t_1 A_1 + \ldots + t_m A_m$$

for some scalars t_i. Thus the solution set is an m-plane in V_n. □

Now let us try to determine when the system $A \cdot X = C$ has a solution. One has the following general result:

Theorem 6. Let A be a k by n matrix. Let r equal the rank of A.

(a) If $r < k$, then there exist vectors C in V_k such that the system $A \cdot X = C$ has no solution.

(b) If $r = k$, then the system $A \cdot X = C$ always has a solution.

Proof. We consider the system $A \cdot X = C$ and apply elementary row operations to both A and C until we have brought A into echelon form B. (For the moment, we need not go all the way to reduced echelon form.) Let C' be the column matrix obtained by applying these same row operations to C. Consider the system $B \cdot X = C'$.

Consider first the case $r < k$. In this case, the last row at least of B is zero. The equation corresponding to this row has the form

$$0x_1 + \ldots + 0x_n = c_k',$$

where c_k' is the entry of C' in row k. If c_k' is not zero, there are no values of $x_1, \ldots, x_n$ satisfying this equation, so the system has no solution.

Let us choose C' to be a k by 1 matrix whose last entry is non-zero. Then apply the same elementary operations as before, in reverse order, to both B and C'. These operations transform B back to A; when we apply them to C', the result is a matrix C such that the system $A\cdot X = C$ has no solution.

Now in the case $r = k$, the echelon matrix B has no zero rows, so the difficulty that occurred in the preceding paragraph does not arise. We shall show that in this case the system has a solution.

More generally, we shall consider the following two cases at the same time: Either (1) B has no zero rows, or (2) whenever the i^{th} row of B is zero, then the corresponding entry c_i' of C' is zero. We show that in either of these cases, the system has a solution.

Let us consider the system $B\cdot X = C'$ and apply further operations to both B and C', so as to reduce B to reduced echelon form D. Let C'' be the matrix obtained by applying these same operations to C'. Note that the zero rows of B, and the corresponding entries of C', are not affected by these operations, since reducing B to reduced echelon form requires us to work only with the non-zero rows.

Consider the resulting system of equations $D\cdot X = C''$. We now proceed as in the proof of Theorem 3. Let J be the set of column indices in which the pivots of D appear, and let K be the remaining indices. Since each x_j, for j in J, appears in only one equation of the system, we can solve for each x_j in terms of the numbers c_i'' and the unknowns x_k. We can now assign values arbitrarily to the x_k and thus obtain a particular solution of the system. The theorem follows. □

The procedure just described actually does much more than was necessary to prove the theorem. It tells us how to determine, in a particular case, whether or not there is a solution; and it tells us, when there is one, how to express the solution set in parametric form as an m-plane in V_n .

Consider the following example:

Example 4. Consider once again the reduced echelon matrix of Example 3:

$$D = \begin{bmatrix} 1 & 0 & -8 & 0 & -3 \\ 0 & 1 & 4 & 0 & 2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The system

$$D\cdot X = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

has no solution because the last equation of the system is

$$0x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 = 1 .$$

On the other hand, the system

$$D\cdot X = \begin{bmatrix} -1 \\ 3 \\ 7 \\ 0 \end{bmatrix}$$

does have a solution. Following the procedure described in the preceding proof, we solve for the unknowns x_1, x_2, and x_4 as follows:

$$x_1 = -1 + 8x_3 + 3x_5$$
$$x_2 = 3 - 4x_3 - 2x_5$$
$$x_4 = 7.$$

The general solution is thus the 2-plane in V_5 specified by the parametric equation

$$X = (-1,3,0,7,0) + x_3(8,-4,1,0,0) + x_5(3,-2,0,0,1).$$

Remark. Solving the system $A \cdot X = C$ in practice involves applying elementary operations to A, and applying these same operations to C. A convenient way to perform these calculations is to form a new matrix from A by adjoining C as an additional column. This matrix is often called the augmented matrix of the system. Then one applies the elementary operations to this matrix, thus dealing with both A and C at the same time. This procedure is described in 16.18 of vol. I of Apostol.

Exercises

1. Let A be a k by n matrix. (a) If $k < n$, show that the system $A \cdot X = \underline{0}$ has a solution different from $\underline{0}$.(Is this result familiar?) What can you say about the dimension of the solution space? (b) If $k > n$, show that there are values of C such that the system $A \cdot X = C$ has no solution.

2. Consider the matrix A of p. A23. (a) Find the general solution of the system $A \cdot X = \underline{0}$. (b) Does the system $A \cdot X = C$ have a solution for arbitrary C?

3. Repeat Exercise 2 for the matrices C, D, and E of p. A23.

4. Let B be the matrix of p. A23. (a) Find the general solution of the system

$$B \cdot X = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

(b) Find conditions on a, b, and c that are necessary and sufficient for the system $B \cdot X = C$ to have a solution, where $C = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. [Hint: What happens to C when you reduce B to echelon form?]

5. Let A be the matrix of p. A20. Find conditions on a, b, c, and d that are necessary and sufficient for the system $A \cdot X = C$ to have a solution, where

$$C = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} .$$

6. Let A be a k by n matrix; let r be the rank of A. Let R be the set of all those vectors C of V_k for which the system

$$A \cdot X = C$$

has a solution. (That is, R is the set of all vectors of the form $A \cdot X$, as X ranges over V_n.)

(a) Show that R is a subspace of V_k.

(b) Show that R has dimension r. [Hint: Let W be the solution space of the system $A \cdot X = \underline{0}$. Then W has dimension $m = n - r$. C oose a basis $A_1, \ldots, A_m$ for W. By adjoining vectors one at a time, extend this to a basis $A_1, \ldots, A_m, B_1, \ldots, B_r$ for all of V_n. Show the vectors $A \cdot B_1, \ldots, A \cdot B_r$ span R; this follows from the fact that $A \cdot A_i = \underline{0}$ for all i. Show these vectors are independent.]

(c) Conclude that if $r < k$, there are vectors C in V_k such that the system $A \cdot X = C$ has no solution; while if $r = k$, this system has a solution for all C. (This provides an alternate proof of Theorem 6.)

7. Let A be a k by n matrix. The columns of A, when looked at as elements of V_k, span a subspace of V_k that is called the column space of A. The row space and column space of A are very different, but it is a totally unexpected fact that they have the same dimension ! Prove this fact as follows: Let R be the subspace of V_k defined in Exercise 6. Show that R is spanned by the vectors $A \cdot E_1, \ldots, A \cdot E_n$; conclude that R equals the column space of A.

Cartesian equations of k-planes in V_n.

There are two standard ways of specifying a k-plane M in V_n. One is by an equation in parametric form:

$$X = P + t_1A_1 + \dots + t_kA_k ,$$

where $A_1,\dots,A_k$ are independent vectors in V_n. (If these vectors were not independent, this equation would still specify an m-plane for some m, but some work would be required to determine m. We normally require the vectors to be independent in the parametric form of the equation of a k-plane.)

Another way to specify a plane in V_n is as the solution set of a system of linear equations

$$A\cdot X = C,$$

where the rows of A are independent. If A has size k by n, then the plane in question has dimension $n - k$. The equation is called a caretesian form for the equation of a plane. (If the rows of A were not independent, then the solution set would be either empty , or an m-plane for some m, but some work would be required to determine m.)

The process of "solving" the system of equations $A\cdot X = C$ that we described in the preceding section is an algorithm for passing from a cartesian equation for M to a parametric equation for M. One can ask whether there is a process for the reverse, for passing from a parametric equation for M to a cartesian equation. The answer is "yes," as we shall see shortly. The other question one might ask is, "Why should one care?" The answer is that sometimes one form is convenient, and other times the other form is more useful. Particularly is this true in the case of 3-dimensional space V_3 , as we shall see.

Definition. Let A be a matrix of size k by n. Let $A_1,\ldots,A_k$ be the rows of A; let W be the subspace of V_n they span. Now the vector X is a solution of the system $A\cdot X = \underline{0}$ if and only if X is orthogonal to each of the vectors A_i. This statement is equivalent to the statement that X is orthogonal to every vector belonging to W. The solution space of this system is for this reason sometimes called the orthogonal complement of W. It is often denoted $W^\perp$ (read "W perp".)

We have the following result:

Theorem 7. If W is a subspace of V_n of dimension k, then its orthogonal complement has dimension $n - k$. Furthermore, W is the orthogonal complement of $W^\perp$; that is, $(W^\perp)^\perp = W$.

Proof. That $W^\perp$ has dimension $n - k$ is an immediate consequence of Theorem 3; for W is the row space of a k by n matrix A with independent rows A_i, whence $W^\perp$ is the solution space of the system $A\cdot X = \underline{0}$.

The space $(W^\perp)^\perp$ has dimension $n - (n - k)$, by what we just proved. And it contains each vector A_i (since $A_i\cdot X = \underline{0}$ for each X in $W^\perp$.) Therefore it equals the space spanned by $A_1,\ldots,A_k$. □

Theorem 8. Suppose a k-plane M in V_n is specified by the parametric equation

$$X = P + t_1A_1 + \ldots + t_kA_k ,$$

where the vectors A_i are independent. Let W be the space they span; and let $B_1,\ldots,B_m$ be a basis for $W^\perp$. If B is the matrix with rows $B_1,\ldots,B_m$, then the equation $B\cdot(X-P)=0$, or

$$B\cdot X = B\cdot P,$$

is a cartesian equation for M.

Proof. The vector X lies in M if and only if $X - P$ belongs to W. This occurs if and only if $X - P$ is orthogonal to each of the vectors B_i, and this occurs if and only if $B\cdot(X - P) = \underline{0}$. $\square$

The preceding proof actually tells us how to find a cartesian equation for M. One takes the matrix A whose rows are the vectors A_i; one finds a basis $B_1,\ldots,B_m$ for the solution space of the system $A\cdot X = \underline{0}$, using the Ga$^{u}_{\wedge}$ss-Jordan algorithm; and then one writes down the equation $B\cdot X = B\cdot P$.

We now turn to the special case of V_3, whose model is the familiar 3-dimensional space in which we live. In this space, we have only lines (1-planes) and planes (2-planes) to deal with. And we can use either the parametric or cartesian form for lines and planes, as we prefer. However, in this situation we tend to prefer:

parametric form for a line, and

cartesian form for a plane.

Let us explain why.

If L is a line given in parametric form $X = P + tA$, then A is uniquely determined up to a scalar factor. (The point P is of course not determined.) The equation itself then exhibits some geometric information about the line; one can for instance tell by inspection whether or not two lines are parallel.

On the other hand, if M is a plane given in parametric form by the equation $X = P + sA + tB$, one does not have as much geometric information immediately at hand. However, let us seek to find a cartesian equation for this plane. We note that the orthogonal complement of $L(A,B)$ is one-dimensional, and is thus spanned by a single non-zero vector

$N = (a_1, a_2, a_3)$. We call N a <u>normal vector</u> to the plane M ; it is uniquely determined up to a scalar factor. (In practice, one finds N by solving the system of equations

$$A \cdot N = \underline{0},$$
$$B \cdot N = \underline{0} \,.)$$

Then a cartesian equation for M is the equation

$$N \cdot (X - P) = 0 \,.$$

If P is the point (p_1, p_2, p_3) of the plane M, this equation has the form

$$(*) \qquad a_1(x_1 - p_1) + a_2(x_2 - p_2) + a_3(x_3 - p_3) = 0.$$

We call this the <u>equation of the plane through</u> $P = (p_1, p_2, p_3)$ <u>with normal vector</u> $N = (a_1, a_2, a_3)$.

We have thus proved the first half of the following theorem:

<u>Theorem 9.</u> If M is a 2-plane in V_3, then M has a cartesian equation of the form

$$a_1x_1 + a_2x_2 + a_3x_3 = b \quad ,$$

where $N = (a_1, a_2, a_3)$ is non-zero. Conversely, any such equation is the cartesian equation of a plane in V_3; the vector N is a normal vector to the plane.

<u>Proof.</u> To prove the converse, we note that this equation is a system consisting of 1 equation in 3 unknowns, and the matrix $A = [a_1 \; a_2 \; a_3]$ has rank 1. Therefore the solution space of the system $A \cdot X = [b]$ is a plane of dimension $3 - 1 = 2$. □

Now we see why the cartesian equation of a plane is useful; it contins some geometric information about the plane. For instance, one can tell by inspection whether two planes given by cartesian equations are parallel.

For they are parallel if and only if their normal vectors are parallel, and that can be determined by inspection of the two equations.

Similarly, one can tell readily whether the line $X = P + tA$ is parallel to a plane M; one just checks whether or not A is orthogonal to the normal vector of M.

Many theorems of 3-dimensional geometry are now easy to prove. us consider some examples.

Theorem 10. Three planes in V_3 intersect in a single point if and only if their normal vectors are independent.

Proof. Take a cartesian equation for each plane; collectively, they form a system $A \cdot X = C$ of three equations in three unknowns. The rows of A are the normal vectors. The solution space of the system (which consists of the points common to all three planes) consists of a a single point if and only if the rows of A are independent. □

Theorem 11. Two non-parallel planes in V_3 intersect in a straight line.

Proof. Let $N_1 \cdot X = b_1$ and $N_2 \cdot X = b_2$ be cartesian equations for the two planes. Their intersection consists of those points X that satisfy both equations. Since N_1 and N_2 are not zero and are not parallel, the matrix having rows N_1 and N_2 has rank 2. Hence the solution of this system of equations is a 1-plane in V_3. □

Theorem 12. Let L be a line, and M a plane, in V_3. If L is parallel to M, then their intersection is either empty or all of L. If L is not parallel to M, then their intersection is a single point.

Proof. Let L have parametric equation $X = P + tA$; let M have cartesian equation $N \cdot X = b$. We wish to determine for what values of t the point $X = P + tA$ lies on the plane M; that is, to determine the solutions of the equation

$$N \cdot (P + tA) = b .$$

Now if L is parallel to M, then the vector A is perpendicular to the normal vector N to M; that is, $N\cdot A = 0$. In this case, the equation

$$N\cdot(P + tA) = b$$

holds for all t if it happens that $N\cdot P = b$, and it holds for no t if $N\cdot P \neq b$. Thus the intersection of L and M is either all of L, or it is empty.

On the other hand, if L is not parallel to M, then $N\cdot A \neq 0$. In this case the equation can be solved uniquely for t. Thus the intersection of L and M consists of a single point. □

Example 5. Consider the plane $M = M(P;A,B)$ in V_3, where $P = (1, -7, 0)$ and $A = (1, 1, 1)$ and $B = (-1, 2, 0)$. To find a normal vector $N = (a_1, a_2, a_3)$ to M, we solve the system

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

One can use the Gauss-Jordan algorithm, or in this simple case , proceed almost by inspection. One can for instance set $a_2 = 1$. Then the second equation implies that $a_1 = 2$; and then the first equation tells us that $a_3 = -a_1 - a_2 = -3$. The plane thus has cartesian equation

$$2(x_1 - 1) + (x_2 + 7) - 3(x_3 - 0) = 0,$$

or

$$2x_1 + x_2 - 3x_3 = -5.$$

Exercises

1. The solution set of the equation

$$3x_1 + 2x_2 - x_3 = 15$$

is a plane in V_3; write the equation of this plane in parametric form.

2. Write parametric equations for the line through (1,0,0) that is perpendicular to the plane $x_1 - x_3 = 5$.

3. Write a parametric equation for the line through (0,5,-2) that is parallel to the planes $2x_2 = x_3$ and $5x_1 + x_2 - 7x_3 = 4$.

4. Show that if P and Q are two points of the plane M, then the line through P and Q is contained in M.

5. Write a parametric equation for the line of intersection of the planes of Exercise 3.

6. Write a cartesian equation for the plane through P = (-1,0,2) and Q = (3,1,5) that is parallel to the line through R = (1,1,1) with direction vector A = (1,3,4).

7. Write cartesian equations for the plane M(P;A,B) in V_4, where P = (1, -1, 0, 2) and A = (1, 0, 1, 0) and B = (2, 1, 0, 1).

8. Show that every n - 1 plane in V_n is the solution set of an equation of the form $a_1x_1 + \dots + a_nx_n = b$, where $(a_1,\dots,a_n) \neq \underline{0}$; and conversely.

9. Let M_1 and M_2 be 2-planes in V_4; assume they are not parallel. What can you say about the intersection of M_1 and M_2? Give examples to illustrate the possibilities.

18.024 Multivariable Calculus with Theory
Spring 2011

The inverse of a matrix

We now consider the problem of the existence of multiplicatiave inverses for matrices. At this point, we must take the non-commutativity of matrix multiplication into account. For it is perfectly possible, given a matrix A, that there exists a matrix B such that $A\cdot B$ equals an identity matrix, without it following that $B\cdot A$ equals an identity matrix. Consider the following example:

Example 6. Let A and B be the matrices

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 0 \\ 3 & -2 \\ -1 & 1 \end{bmatrix}$$

Then $A\cdot B = I_2$, but $B\cdot A \neq I_3$, as you can check.

Definition. Let A be a k by n matrix. A matrix B of size n by k is called an inverse for A if both of the following equations hold:

$$A\cdot B = I_k \qquad \text{and} \qquad B\cdot A = I_n .$$

We shall prove that if $k \neq n$, then it is impossible for both these equations to hold. Thus only square matrices can have inverses. We also show that if the matrices are square and one of these equations holds, then the other equation holds as well!

Theorem 13. Let A be a matrix of size k by n. Then A has an inverse if and only if $k = n = \operatorname{rank} A$. If A has an inverse, that inverse is unique.

Proof. Step 1. If B is an n by k matrix, we say B is a right inverse for A if $A \cdot B = I_k$. We say B is a left inverse for A if $B \cdot A = I_n$. Let A be a k by n matrix.

Let r be the rank of A. We show that if A has a right inverse, then $r = k$; and if A has a left inverse, then $r = n$. The "only if" part of the theorem follows.

First, suppose B is a right inverse for A. Then $A \cdot B = I_k$. It follows that the system of equations $A \cdot X = C$ has a solution for arbitrary C, for the vector $X = B \cdot C$ is one such solution, as you can check. Theorem 6 then implies that r must equal k.

Second, suppose B is a left inverse for A. Then $B \cdot A = I_n$. It follows that the system of equations $A \cdot X = \underline{0}$ has only the trivial solution, for the equation $A \cdot X = \underline{0}$ implies that $B \cdot (A \cdot X) = \underline{0}$, whence $X = \underline{0}$. Now the dimension of the solution space of the system $A \cdot X = \underline{0}$ is $n - r$; it follows that $n - r = 0$.

Step 2. Now let A be an n by n matrix of rank n. We show there is a matrix B such that $A \cdot B = I_n$.

Because the rows of A are independent, the system of equations $A \cdot X = C$ has a solution for arbitrary C. In particular, it has a solution when C is one of the unit coordinate vectors E_i in V_n. Let us choose B_i so that

$$A \cdot B_i = E_i ,$$

for $i = 1,\ldots,n$. Then if B is the n by n matrix whose successive columns are $B_1,\ldots,B_n$, the product $A \cdot B$ equals the matrix whose successive columns are $E_1,\ldots,E_n$; that is, $A \cdot B = I_n$.

Step 3. We show that if A and B are n by n matrices and $A \cdot B = I_n$, then $B \cdot A = I_n$. The "if" part of the theorem follows.

Let us note that if we apply Step 1 to the case of a square matrix of size n by n, it says that if such a matrix has either a right inverse <u>or</u> a left inverse, then its rank must be n.

Now the equation $A \cdot B = I_n$ says that A has a right inverse and that B has a left inverse. Hence both A and B must have rank n. Applying Step 2 to the matrix B, we see that there is a matrix C such that $B \cdot C = I_n$. Now we compute

$$\begin{aligned} A\cdot(B\cdot C) &= (A\cdot B)\cdot C \,, \\ A\cdot I_n &= I_n\cdot C \,, \\ A &= C \,. \end{aligned}$$

The equation $B \cdot C = I_n$ now becomes $B \cdot A = I_n$, as desired.

<u>Step</u> 4. The computation we just made shows that if a matrix has an inverse, that inverse is unique. Indeed, we just showed that if B has an left inverse A and a right inverse C, then $A = C$. □

Let us state the result proved in Step 3 as a separate theorem:

<u>Theorem</u> 14. If A and B are n by n matrices such that $A \cdot B = I_n$, then $B \cdot A = I_n$. □

We now have a theoretical criterion for the existence of A^{-1}. But how can one find A^{-1} in practice? For instance, how does one compute $B = A^{-1}$ if A is a given nonsingular 3 by 3 matrix? By Theorem 14, it will suffice to find a matrix

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

such that $A \cdot B = I_3$. But this problem is just the problem of solving three systems of linear equations

$$A \cdot \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad A \cdot \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad A \cdot \begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} .$$

Thus the Gauss-Jordan algorithm applies. An efficient way to apply this algorithm to the computation of A^{-1} is outlined on p. 612 of Apostol, which you should read now.

There is also a formula for A^{-1} that involves determinants. It is given in the next section.

Remark . It remains to consider the question whether the existence of the inverse of a matrix has any practical significance, or whether it is of theoretical interest only. In fact, the problem of finding the inverse of a matrix in an efficient and accurate way is of great importance in engineering. One way to explain this is to note that often in a real-life situation, one has a fixed matrix A, and one wishes to solve the system $A \cdot X = C$ repeatedly, for many different values of C. Rather than solving each one of these systems separately, it is much more efficient to find the inverse of A, for then the solution $X = A^{-1} \cdot C$ can be computed by simple matrix multiplication.

Exercises

1. Give conditions on a,b,c,d,e,f such that the matrix

$$B = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}$$

is a right inverse to the matrix A of Example 6. Find two right inverses for A.

2. Let A be a k by n matrix with $k < n$. Show that A has no left inverse. Show that if A has a right inverse, then that right inverse is not unique.

3. Let B be an n by k matrix with $k < n$. Show that B has no right inverse. Show that if B has a left inverse, then that left inverse is not unique.

Determinants

The determinant is a function that assigns, to each square matrix A, a real number. It has certain properties that are expressed in the following theorem:

Theorem 15. There exists a function that assigns, to each n by n matrix A, a real number that we denote by $\det A$. It has the following properties:

(1) If B is the matrix obtained from A by exchanging rows i and j of A, then $\det B = -\det A$.

(2) If B is the matrix obtained form A by replacing row i of A by itself plus a scalar multiple of row j (where $i \neq j$), then $\det B = \det A$.

(3) If B is the matrix obtained from A by multiplying row i of A by the scalar c, then $\det B = c \cdot \det A$.

(4) If I_n is the identity matrix, then $\det I_n = 1$.

We are going to assume this theorem for the time being, and explore some of its consequences. We will show, among other things, that these four properties characterize the determinant function completely. Later we shall construct a function satisfying these properties.

First we shall explore some consequences of the first three of these properties. We shall call properties (1)-(3) listed in Theorem 15 the elementary row properties of the determinant function.

Theorem 16. Let f be a function that assigns, to each n by n matrix A, a real number. Suppose f satisfies the elementary row properties of the determinant function. Then for every n by n matrix A,

$$(*) \qquad f(A) = f(I_n) \cdot \det A .$$

This theorem says that any function f that satisfies properties (1), (2), and (3) of Theorem 15 is a scalar multiple of the determinant function. It also says that if f satisfies property (4) as well, then f must equal the determinant function. Said differently, there is at most one function that satisfies all four conditions.

Proof. Step 1. First we show that if the rows of A are dependent, then $f(A) = 0$ and $\det A = 0$. Equation (*) then holds trivially in this case.

Let us apply elementary row operations to A to bring it to echelon form B. We need only the first two elementary row operations to do this, and they change the values of f and of the determinant function by at most a sign. Therefore it suffices to prove that $f(B) = 0$ and $\det B = 0$. The last row of B is the zero row, since A has rank less than n. If we multiply this row by the scalar c, we leave the matrix unchanged, and hence we leave the values of f and $\det$ unchanged. On the other hand, this operation multiplies these values by c. Since c is arbitrary, we conclude that $f(B) = 0$ and $\det B = 0$.

Step 2. Now let us consider the case where the rows of A are independent. Again, we apply elementary row operations to A. However, we will do it very carefully, so that the values of f and $\det$ do not change.

As usual, we begin with the first column. If all entries are zero, nothing remains to be done with this column.

We move on to consider columns $2,\ldots,n$ and begin the process again.

Otherwise, we find a non-zero entry in the first column. If necessary, we exchange rows to bring this entry up to the upper left-hand corner; this changes the sign of both the functions f and $\det$, so we then multiply this row by -1 to

change the signs back. Then we add multiples of the first row to each of the remaining rows so as to make all the remaining entries in the first column into zeros. By the preceding theorem and its corollary, this does not change the values of either f or det.

Then we repeat the process, working with the second column and with rows $2,\ldots,n$. The operations we apply will not affect the zeros we already have in column 1.

Since the rows of the original matrix were independent, then we do not have a zero row at the bottom when we finish, and the "stairsteps" of the echelon form go over just one step at a time.

In this case, we have brought the matrix to a form where all of the entries below the main diagonal are zero. (This is what is called upper triangular form.) Furthermore, all the diagonal entries are non-zero. Since the values of f and det remain the same if we replace A by this new matrix B, it now suffices to prove our formula for a matrix of the form

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & b_{22} & \cdots & b_{2n} \\ & \cdot\ \cdot\ \cdot & & \\ 0 & 0 & \cdots & b_{nn} \end{bmatrix} .$$

where the diagonal entries are non-zero.

Step 3. We show that our formula holds for the matrix B. To do this we continue the Gauss-Jordan elimination process. By adding a multiple of the last row to the rows above it, then adding multiples of the next-to-last row to the rows lying above it, and so on, we can bring the matrix to the form where all the non-diagonal entries vanish. This form is called diagonal form. The values of both f and det remain the same if we replace B by this new matrix C. So now it suffices to prove our formula for a matrix of the form

$$C = \begin{bmatrix} b_{11} & 0 & 0 & \dots & 0 \\ 0 & b_{22} & 0 & \dots & 0 \\ & & \cdot\;\cdot\;\cdot & & \\ 0 & 0 & 0 & \dots & b_{nn} \end{bmatrix}.$$

(Note that the diagonal entries of B remain unchanged when we apply the Gauss-Jordan process to eliminate all the non-zero entries above the diagonal. Thus the diagonal entries of C are the same as those of B.)

We multiply the first row of C by $1/b_{11}$. This action multiplies the values of both f and det by a factor of $1/b_{11}$. Then we multiply the second row by $1/b_{22}$, the third by $1/b_{33}$, and so on. By this process, we transform the matrix C into the identity matrix I_n. We conclude that

$$f(I_n) = (1/b_{11})\dots(1/b_{nn})\ f(C)\ , \quad \text{and}$$

$$\det I_n = (1/b_{11})\dots(1/b_{nn}) \det C.$$

Since $\det I_n = 1$ by hypothesis, it follows from the second equation that

$$\det C = b_{11}\, b_{22} \dots b_{nn}\ .$$

Then it follows from the first equation that

$$f(C) = f(I_n)\cdot \det C,$$

as desired. □

Besides proving the determinant function unique, this theorem also tells us one way to compute determinants. One applies this version of the Gauss-Jordan algorithm to reduce the matrix to echelon form. If the matrix that results has a zero row, then the determinant is zero. Otherwise, the matrix that results is in upper triangular form with non-zero diagonal entries, and the determinant is the product of the diagonal entries. [illegible] diago[illegible]

[illegible] but some of the diagonal entries are zero.)

The proof of this theorem tells us something else: If the rows of A are not independent, then $\det A = 0$, while if they are independent, then $\det A \neq 0$. We state this result as a theorem:

Theorem 16. Let A be an n by n matrix. Then A has rank n if and only if $\det A \neq 0$. □

An n by n matrix A for which $\det A \neq 0$ is said to be non-singular . This theorem tells us that A has rank n if and only if A is non-singular.

Now we prove a totally unexpected result:

Theorem 17. Let A and B be n by n matrices. Then

$$\det(A\cdot B) = (\det A)\cdot(\det B) .$$

Proof. This theorem is almost impossible to prove by direct computation. Try the case $n = 2$ if you doubt me ! Instead, we proceed in another direction:

Let B be a fixed n by n matrix. Let us define a function f of n by n matrices by the formula

$$f(A) = \det(A\cdot B).$$

We shall prove that f satisfies the elementary row properties of the determinant function. From this it follows that

$$f(A) = f(I_n)\cdot \det A ,$$

which means that

$$\begin{aligned} \det(A\cdot B) &= \det(I_n\cdot B)\cdot \det A \\ &= \det B \cdot \det A , \end{aligned}$$

and the theorem is proved.

First, let us note that if $A_1,\ldots,A_n$ are the rows of A, considered as row matrices, then the rows of $A\cdot B$ are (by the definition of matrix multiplication) the row matrices $A_1\cdot B,\ldots,A_n\cdot B$. Now exchanging rows i and j of A, namely A_i and A_j, has the effect of exchanging rows i and j of $A\cdot B$. Thus this operation changes the value of f by a factor of -1. Similarly, replacing the i^{th} row A_i of A by $A_i + cA_j$ has the effect on $A\cdot B$ of replacing its i^{th} row $A_i\cdot B$ by

$$\begin{aligned} (A_i + cA_j)\cdot B &= A_i\cdot B + c\, A_j\cdot B \\ &= (\text{row } i \text{ of } A\cdot B) + c(\text{row } j \text{ of } A\cdot B). \end{aligned}$$

Hence it leaves the value of f unchanged. Finally, replacing the i^{th} row A_i of A by cA_i has the effect on $A\cdot B$ of replacing the i^{th} row $A_i\cdot B$ by

$$(cA_i)\cdot B = c\,(A_i\cdot B) = c\,(\text{row } i \text{ of } A\cdot B).$$

Hence it multiplies the value of f by c. □

The determinant function has many further properties, which we shall not explore here. (One reference book on determinants runs to four volumes!) We shall derive just one additional result, concerning the inverse matrix.

Exercises

1. Suppose that f satisfies the elementary row properties of the determinant function. Suppose also that x, y, z are numbers such that

$$f\begin{bmatrix} x & y & z \\ 3 & 0 & 2 \\ 1 & 1 & 1 \end{bmatrix} = 1,$$

Compute the value of f for each of the following matrices:

(a) $\begin{bmatrix} 2x & 2y & 2z \\ 3/2 & 0 & 1 \\ 3 & 3 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} x & y & z \\ 3x+3 & 3y & 3z+2 \\ x+2 & y+2 & z+2 \end{bmatrix}$ (c) $\begin{bmatrix} x-1 & y-1 & z-1 \\ 1 & 1 & 1 \\ 4 & 1 & 3 \end{bmatrix}$

2. Let f be the function of Exercise 1. Calculate $f(I_n)$. Express f in terms of the determinant function.

3. Compute the determinant of the following matrix, using Gauss-Jordan elimination.

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

4. Determine whether the following sets of vectors are linearly independent, using determinants.

(a) $A_1 = (1,-1,0)$, $A_2 = (0,1,-1)$, $A_3 = (2,3,-1)$.

(b) $A_1 = (1,-1,2,1)$, $A_2 = (-1,2,-1,0)$, $A_3 = (3,-1,1,0)$, $A_4 = (1,0,0,1)$.

(c) $A_1 = (1,0,0,0,1)$, $A_2 = (1,1,0,0,0)$, $A_3 = (1,0,1,0,1)$, $A_4 = (1,1,0,1,1)$, $A_5 = (1,0,0,0,0)$.

(d) $A_1 = (1,-1)$, $A_2 = (0,1)$, $A_3 = (1,1)$.

A formula for A^{-1}

We know that an n by n matrix A has an inverse if and only if it has rank n, and we know that A has rank n if and only if $\det A \neq 0$. Now we derive a formula for the inverse that involves determinants directly.

We begin with a lemma about the evaluation of determinants.

Lemma 18. Given the row matrix $[a_1 \ \ldots \ a_n]$, let us define a function f of $(n-1)$ by $(n-1)$ matrices B by the formula

$$f(B) = \det \left[\begin{array}{c|c|c} a_1 \ \cdots & a_j & \cdots \ \ a_n \\ \hline & 0 & \\ B_1 & \vdots & B_2 \\ & 0 & \end{array}\right]$$

where B_1 consists of the first $j-1$ columns of B, and B_2 consists of the remainder of B. Then

$$f(B) = (-1)^{j+1} a_j \cdot \det B.$$

Proof. You can readily check that f satisfies properties (1)-(3) of the determinant function. Hence $f(B) = f(I_{n-1}) \cdot \det B$. We compute

$$f(I_n) = \det \left[\begin{array}{c|c|c} a_1 \ \cdots & a_j & \cdots \ a_n \\ \hline I_{j-1} & 0 & \mathbf{O} \\ & \vdots & \\ \cline{1-1}\cline{3-3} \mathbf{O} & 0 & I_{n-j} \end{array}\right]$$

where the large zeros stand for zero matrices of the appropriate size. A sequence of $j-1$ interchanges of adjacent rows gives us the equation

$$f(I_n) = (-1)^{j-1}\det\left[\begin{array}{c|c|c} I_{j-1} & \begin{matrix}0\\ \vdots\\ 0\end{matrix} & \bigcirc \\ \hline a_1 \;\cdots & a_j & \cdots\; a_n \\ \hline \bigcirc & \begin{matrix}0\\ \vdots\\ 0\end{matrix} & I_{n-j} \end{array}\right].$$

One can apply elementary operations to this matrix, without changing the value of the determinant, to replace all of the entries $a_1,\ldots,a_{j-1},a_{j+1},\ldots,a_n$ by zeros. Then the resulting matrix is in diagonal form. We conclude that

$$f(I_n) \;=\; (-1)^{j-1}\, a_j \;=\; (-1^{j+1}\, a_j \;.\; \square$$

<u>Corollary</u> 19. Consider an n by n matrix of the form

$$A \;=\; \begin{array}{cc} \swarrow\text{column } j & \\ \left[\begin{array}{c|c|c} B_1 & \begin{matrix}0\\ \vdots\\ 0\end{matrix} & B_2 \\ \hline a_{i1} \;\cdots & a_{ij} & \cdots\; a_{in} \\ \hline B_3 & \begin{matrix}0\\ \vdots\\ 0\end{matrix} & B_4 \end{array}\right] & \leftarrow\text{row } i \end{array}$$

where $B_1,\ldots,B_4$ are matrices of appropriate size. Then

$$\det A = (-1)^{i+j} a_{ij}\cdot\det\begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}$$

<u>Proof</u>. A sequence of $i-1$ interchanges of adjacent rows will bring the matrix A to the form given in the preceding lemma. □

<u>Definition</u>. In general, if A is an n by n matrix, then the matrix of size $(n-1)$ by $(n-1)$ obtained by deleting the i^{th} row and the j^{th} column of A is called the <u>(i,j)-minor</u> of A, and is denoted A_{ij}.

The preceding corollary can then be restated as follows:

Corollary 20. If all the entries in the j^{th} column of A are zero except for the entry a_{ij} in row i, then $\det A = (-1)^{i+j} a_{ij} \cdot \det A_{ij}$.

The number $(-1)^{i+j} \det A_{ij}$ that appears in this corollary is also given a special name. It is called the (i,j)-cofactor of A. Note that the signs $(-1)^{i+j}$ follows the pattern

$$\begin{bmatrix} + & - & + & - & + & \cdots \\ - & + & - & + & - & \cdots \\ + & - & + & - & + & \cdots \\ & & \cdots & & & \end{bmatrix}$$

Now we derive our formula for A^{-1}.

Theorem 21. Let A be an n by n matrix with $\det A \neq 0$. If $A \cdot B = I_n$, then

$$b_{ij} = (-1)^{j+i} \det A_{ji} / \det A.$$

(That is, the entry of B in row i and column j equals the (j,i)-cofactor of A, divided by det A. This theorem says that you can compute B by computing det A and the determinants of n^2 different $(n-1)$ by $(n-1)$ matrices. This is certainly not a practical procedure except in low dimensions!)

Proof. Let X denote the j^{th} column of B. Then $x_i = b_{ij}$. Because $A \cdot B = I_n$, the column matrix X satisfies the equation

$$A \cdot X = (j^{th} \text{ column of } I_n) = E_j .$$

(Here E_j is the column matrix consisting of zeros except for an entry of 1 in row j.) Furthermore, if A_i denote the i^{th} column of A, then because $A \cdot I_n = A$, we have the equation

$$A \cdot (i^{th} \text{ column of } I_n) = A \cdot E_i = A_i .$$

Now we introduce a couple of weird matrices for reasons that will become clear. Using the two preceding equations, we put them together to get the following matrix equation:

$$(*) \quad A \cdot [E_1 \ldots E_{i-1} \;\; X \;\; E_{i+1} \cdots \;\; E_n] \;=\; [A_1 \cdots A_{i-1} \;\; E_j \;\; A_{i+1} \cdots A_n].$$

It turns out that when we take determinants of both sides of this equation, we get exactly the equation of our theorem! First, we show that

$$\det [E_1 \cdots E_{i-1} \;\; X \;\; E_{i+1} \cdots E_n] = x_i .$$

Written out in full, this equation states that

$$\det \left[\begin{array}{c|c|c} I_{i-1} & \begin{matrix} x_1 \\ \vdots \end{matrix} & O \\ \hline 0 \;\ldots\; 0 & x_i & 0 \;\ldots\; 0 \\ \hline O & \begin{matrix} \vdots \\ x_n \end{matrix} & I_{n-i} \end{array}\right] = x_i .$$

If $x_i = 0$, this equation holds because the matrix has a zero row. If $x_i \neq 0$, we can by elementary operations replace all the entries above and beneath x_i in its column by zeros. The resulting matrix will be in diagonal form, and its determinant will be x_i.

Thus the determinant of the left side of equation (*) equals $(\det A) \cdot x_i$, which equals $(\det A) \cdot b_{ij}$. We now compute the determinant of the right side of equation (*). Corollary 20 applies, because the i^{th} column of this matrix consists of zeros except for an entry of 1 in row j. Thus the right side of (*) equals $(-1)^{j+i}$ times the determinant of the matrix obtained by deleting row j and column i. This is exactly the same matrix as we would obtain by deleting row j and column i of A. Hence the right side of (*) equals $(-1)^{j+i} \det A_{ji}$, and our theorem is proved. □

Remark 1. If A is a matrix with general entry a_{ij} in row i and column j, then the transpose of A (denoted A^{tr}) is the matrix whose entry in row i and column j is a_{ji}.

Thus if A has size k by n, then A^{tr} has size n by k; it can be pictured as the matrix obtained by flipping A around the line $y = -x$. For example,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^{tr} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}.$$

Of course, if A is square, then the transpose of A has the same dimensions as A.

Using this terminology, the theorem just proved says that the inverse of A can be computed by the following four-step process:

(1) Form the matrix whose entry in row i and column j is the number $\det A_{ij}$. (This is called the matrix of minor determinants.)

(2) Prefix the sign $(-1)^{i+j}$ to the entry in row i and column j, for each entry of the matrix. (This is called the matrix of cofactors.)

(3) Transpose the resulting matrix.

(4) Divide each entry of the matrix by $\det A$.

In short, this theorem says that

$$A^{-1} = \frac{1}{\det A}(\text{cof } A)^{tr}.$$

This formula for A^{-1} is used for computational purposes only for 2 by 2 or 3 by 3 matrices; the work simply gets too great otherwise. But it is important for theoretical purposes. For instance, if the entries of A

are continuous functions of a parameter t, this theorem tells us that the entries of A^{-1} are also continuous functions of t, provided det A is never zero.

Remark 2. This formula does have one practical consequence of great importance. It tells us that if det A is small as compared with the entries of A, then a small change in the entries of A is likely to result in a large change in the computed entries of A^{-1}. This means, in an engineering problem , that a small error in calculating A (even round-off error) may result in a gross error in the calculated value of A^{-1}. A matrix for which det A is relatively small is said to be *ill-conditioned.* If such a matrix arises in practice, one usually tries to reformulate the problem to avoid dealing with such a matrix.

Exercises

1. Use the formula for A^{-1} to find the inverses of the following matrices , assuming the usual definition of the determinant in low dimensions.

(a) $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, assuming $ad - bc \neq 0$.

(b) $\begin{bmatrix} a & b & 0 \\ 0 & c & d \\ 0 & 0 & e \end{bmatrix}$, assuming $ace \neq 0$.

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

2. Let A be a square matrix all of whose entries are integers. Show that if $\det A = \pm 1$, then all the entries of A^{-1} are integers.

3. Consider the matrices A,B,C,D,E of p. A.23. Which of these matrices have inverses?

4. Consider the following matrix function:

$$A(t) = \begin{bmatrix} t & t^2 & t^3 \\ 0 & 1 & t \\ 2 & 0 & t \end{bmatrix}$$

For what values of t does A^{-1} exist? Give a formula for A^{-1} in terms of t.

5. Show that the conclusion of Theorem 20 holds if A has an entry of a_{ij} in row i and column j, and all the other entries in row i equal 0.

*6. <u>Theorem</u> Let A, B, C be matrices of size k by k, and m by k, and m by m, respectively. Then

$$\det \begin{bmatrix} A & 0 \\ B & C \end{bmatrix} = (\det A)(\det C).$$

(Here 0 is the zero matrix of appropriate size.)

<u>Proof.</u> Let B and C be fixed. For each k by k matrix A, define

$$f(A) = \det \begin{bmatrix} A & 0 \\ B & C \end{bmatrix}.$$

(a) Show f satisfies the elementary row properties of the determinant function.

(b) Use Exercise 5 to show that $f(I_k) = \det C$.

(c) Complete the proof.

Construction of the determinant when $n \leq 3$.

The actual definition of the determinant function is the least interesting part of this entire discussion. The situation is similar to the situation with respect to the functions $\sin x$, $\cos x$, and e^x. You will recall that their actual definitions (as limits of power series) were not nearly as interesting as the properties we derived from simple basic assumptions about them.

We first consider the case where $n \leq 3$, which is doubtless familiar to you. This case is in fact all we shall need for our applications to calculus.

We begin with a lemma:

Lemma 21. Let $f(A)$ be a real-valued function of n by n matrices. Suppose that:

(i) Exchanging any two rows of A changes the value of f by a factor of -1.

(ii) For each i, f is linear as a function of the i^{th} row.

Then f satisfies the elementary row properties of the determinant function.

Proof. By hypothesis, f satisfies the first elementary row property. We check the other two.

Let $A_1, \ldots, A_n$ be the rows of A. To say that f is linear as a function of row i alone is to say that (when f is written as a function of the rows of A):

$$(*) \quad f(A_1, \ldots, cX + dY, \ldots, A_n) = cf(A_1, \ldots, X, \ldots, A_n) + df(A_1, \ldots, Y, \ldots, A_n),$$

where $cX + dY$ and X and Y appear in the i^{th} component.

The special case $d = 0$ tells us that multiplying the i^{th} row of A by c has the effect of multiplying the value of f by c.

We now consider the third type of elementary operation. Suppose that B is the matrix obtained by replacing row i of A by itself plus c times row j. We then compute (assuming $j > i$ for convenience in notation),

$$f(B) = f(A_1,\ldots,\underset{\uparrow_{i}\text{th}}{A_i + cA_j},\ldots,\underset{\uparrow_{j}\text{th}}{A_j},\ldots,A_n\ )$$

$$= f(A_1,\ldots,\underset{\uparrow_{i}\text{th}}{A_i},\ldots,\underset{\uparrow_{j}\text{th}}{A_j},\ldots,A_n) \ + \ c\ f(A_1,\ldots,\underset{\uparrow_{i}\text{th}}{A_j},\ldots,\underset{\uparrow_{j}\text{th}}{A_j},\ldots,A_n\ ).$$

The second term vanishes, since two rows are the same. (Exchanging them does not change the matrix, but by Step 1 it changes the value of f by a factor of -1.) □

<u>Definition.</u> We define

$$\det[a] = a.$$

$$\det\begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} = a_1b_2 - a_2b_1.$$

$$\det\begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix} = a_1\cdot\det\begin{bmatrix} b_2 & b_3 \\ c_2 & c_3 \end{bmatrix} - a_2\cdot\det\begin{bmatrix} b_1 & b_3 \\ c_1 & c_3 \end{bmatrix} + a_3\cdot\det\begin{bmatrix} b_1 & b_2 \\ c_1 & c_2 \end{bmatrix}$$

<u>Theorem 22.</u> The preceding definitions satisfy the four conditions of the determinant function.

<u>Proof.</u> The fact that the determinant of the identity matrix is 1 follows by direct computation. It then suffices to check that (i) and (ii) of the preceding theorem hold .

In the 2 by 2 case, exchanging rows leads to the determinant $b_1a_2 - b_2a_1$, which is the negative of what is given.

In the 3 by 3 case, the fact that exchanging the last two rows changes the sign of the determinant follows from the 2 by 2 case. The fact that exchanging the first two rows also changes the sign follows similarly if we rewrite the formula defining the determinant in the form

$$\det\begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}\cdot c_3 \quad - \quad \det\begin{bmatrix} a_1 & a_3 \\ b_1 & b_3 \end{bmatrix}\cdot c_2 \quad + \quad \det\begin{bmatrix} a_2 & a_3 \\ b_2 & b_3 \end{bmatrix}\cdot c_1 \quad .$$

Finally, exchanging rows 1 and 3 can be accomplished by three exchanges of adjacent rows [namely, $(A,B,C) \to (A,C,B) \to (C,A,B) \to (C,B,A)$], so it changes the sign of the determinant.

To check (ii) is easy. Consider the 3 by 3 case, for example. We know that any function of the form

$$f(X) = [\,a \;\; b \;\; c\,]\cdot X = ax_1 + bx_2 + cx_3$$

is linear, where X is a vector in V_3 . The function

$$f(X) = \det\begin{bmatrix} x_1 & x_2 & x_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}$$

has this form, where the coefficients a, b, and c involve the constants b_i and c_j . Hence f is linear as a function of the first row. The "row-exchange property" then implies that f is linear as a function of each of the other rows. □

Exercise

*1. Let us define

$$\det\begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix} = a_1\cdot\det\begin{bmatrix} b_2 & b_3 & b_4 \\ c_2 & c_3 & c_4 \\ d_2 & d_3 & d_4 \end{bmatrix} - a_2\cdot\det\begin{bmatrix} b_1 & b_3 & b_4 \\ c_1 & c_3 & c_4 \\ d_1 & d_3 & d_4 \end{bmatrix}$$

$$+\ a_3\cdot\det\begin{bmatrix} b_1 & b_2 & b_4 \\ c_1 & c_2 & c_4 \\ d_1 & d_2 & d_4 \end{bmatrix} - a_4\cdot\det\begin{bmatrix} b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \end{bmatrix}.$$

(a) Show that det $I_4 = 1$.

(b) Show that exchanging any two of the last three rows changes the sign of the determinant.

(c) Show that exchanging the first two rows changes the sign. [Hint: Write the expression as a sum of terms involving $\det\begin{bmatrix} a_i & a_j \\ b_i & b_j \end{bmatrix}$.]

(d) Show that exchanging any two rows changes the sign.

(e) Show that det is linear as a function of the first row.

(f) Conclude that det is linear as a function of the i^{th} row.

(g) Conclude that this formula satisfies all the properties of the determinant function.

Construction of the Determinant Function in general. Suppose we take the positive integers 1, 2, . . . , k and write them down in some arbitrary order, say $j_1, j_2, \ldots, j_k$. This new ordering is called a *permutation* of these integers. For each integer j_i in this ordering, let us count how many integers *follow* it in this ordering, but *precede* it in the natural ordering 1, 2, . . . , k. This number is called the *number of inversions caused by* the integer j_i. If we determine this number for each integer j_i in the ordering and add the results together, the number we get is called the total number of inversions which occur in this ordering. If the number is odd, we say the permutation is an *odd permutation*; if the number is even, we say it is an *even permutation.*

For example, consider the following reordering of the integers between 1 and 6:

$$2, 5, 1, 3, 6, 4.$$

If we count up the inversions, we see that the integer 2 causes one inversion, 5 causes three inversions, 1 and 3 cause no inversions, 6 causes one inversion, and 4 causes none. The sum is five, so the permutation is odd.

If a permutation is odd, we say the *sign* of that permutation is $-$; if it is even, we say its sign is $+$. A useful fact about the sign of a permutation is the following:

> **Theorem 22.** If we interchange two adjacent elements of a permutation, we change the sign of the permutation.

Proof. Let us suppose the elements j_i and j_{i+1} of the permutation $j_1, \ldots, j_i, j_{i+1}, \ldots, j_k$ are the two we interchange, obtaining the permutation

$$j_1, \ldots, j_{i+1}, j_i, \ldots, j_k.$$

The number of inversions caused by the integers $j_1, \ldots, j_{i-1}$ clearly is the same in the new permutation as in the old one, and so is the number of inversions caused by $j_{i+2}, \ldots, j_k$. It remains to compare the number of inversions caused by j_{i+1} and by j_i in the two permutations.

Case I: j_i precedes j_{i+1} in the natural ordering $1, \ldots, k$. In this case, the number of inversions caused by j_i is the same in both permutations, but the number of inversions caused by j_{i+1} is one larger in the second permutation than in the first, for j_i follows j_{i+1} in the second permutation, but not in the first. Hence the total number of inversions is increased by one.

Case II: j_i follows j_{i+1} in the natural ordering $1, \ldots, k$. In this case, the number of inversion caused by j_{i+1} is the same in both permutations, but the number of inversions caused by j_i is one less in the second permutation than in the first.

In either case the total number of inversions changes by one, so that the sign of the permutation changes. □

Example. If we interchange the second and third elements of the permutation considered in the previous example, we obtain 2, 1, 5, 3, 6, 4, in which the total number of inversions is four, so the permutation is even.

Definition. Consider a k by k matrix

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}.$$

Pick out one entry from each row of A; do this in such a way that these entries all lie in different columns of A. Take the product of these entries,

$$a_{1j_1}a_{2j_2}a_{3j_3} \cdots a_{kj_k},$$

and prefix a $\pm$ sign according as the permutation $j_1, \ldots, j_k$ is even or odd. (Note that we arrange the entries in the order of the rows they come from, and then we compute the sign of the resulting permutation of the column indices.)

If we write down *all possible* such expressions and add them together, the number we get is defined to be the *determinant* of A.

Remark. We apply this definition to the general 2 by 2 matrix, and obtain the formula

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

If we apply it to a 3 by 3 matrix, we find that

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{array}{l} + a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} \\ - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} \\ + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}. \end{array}$$

The formula for the determinant of a 4 by 4 matrix involves 24 terms, and for a 5 by 5 matrix it involves 120 terms; we will not write down these formulas. The reader will readily believe that the definition we have given is not very useful for computational purposes!

The definition is, however, very convenient for theoretical purposes.

Theorem 24. The determinant of the identity matrix is 1.

Proof. Every term in the expansion of det I_n has a factor of zero in it except for the term $a_{11}a_{22}\dots a_{kk}$, and this term equals 1. □

Theorem 25. If A' is obtained from A by interchanging rows i and i+1, then det A' = - det A.

Proof. Note that each term in the expansion of det A' also appears in the expansion of det A, because we make *all possible* choices of one entry from each row and column when we write down this expansion. The only thing we have to do is to compare what signs this term has when it appears in the two expansions.

Let $a_{1j_1} \cdots a_{ij_i}a_{i+1,j_{i+1}} \cdots a_{kj_k}$ be a term in the expansion of det A. If we look at the corresponding term in the expansion of det A', we see that we have the same factors, *but they are arranged differently.* For to compute the sign of this term, we agreed to arrange the entries in the order of the rows they came from, and then to take the sign of the corresponding permutation of the column indices. Thus in the expansion of det A', this term will appear as

$$a_{1j_1} \cdots a_{i+1,j_{i+1}}a_{i,j_i} \cdots a_{kj_k}.$$

The permutation of the column indices here is the same as above except that elements j_i and j_{i+1} have been interchanged. By Theorem 8.4, this means that this term appears in the expansion of det A' with the sign opposite to its sign in the expansion of det A.

Since this result holds for each term in the expansion of det A', we have det $A' = -$ det A. □

Theorem 26. The function det is linear as a function of the i^{th} row.

Proof. Suppose we take the constant matrix A, and replace its i^{th} row by the row vector $[x_1 \dots x_k]$. When we take the determinant of this new matrix, each term in the expression equals a constant times x_j, for some j. (This happens because in forming this term, we picked out exactly one entry from each row of A.) Thus this function is a linear combination of the components x_i; that is, it has the form

$[c_1 \dots\dots c_k] \cdot X$, for some constants c_i . □

Exercises

1. Use Theorem 25 to show that exchanging any two rows of A changes the sign of the determinant.

2. Consider the term $a_{1j_1} \cdot a_{2j_2} \cdots a_{kj_k}$ in the definition of the determinant. (The integers $j_1, j_2, \ldots, j_k$ are distinct.) Suppose we arrange the factors in this term in the order of their column indices, obtaining an expression of the form

$$a_{i_1 1} \cdot a_{i_2 2} \cdots a_{i_k k} .$$

Show that the sign of the permutation $i_1, i_2, \ldots, i_k$ equals the sign of the permutation $j_1, j_2, \ldots, j_k$.

Conclude that $\det A^{tr} = \det A$ in general.

3. Let A be an n by n matrix, with general entry a_{ij} in row i and column j. Let m be a fixed index. Show that

$$\det A = \sum_{j=1}^{n} a_{mj} \cdot (-1)^{m+j} \det A_{mj} .$$

Here A_{mj} denotes , as usual, the (m,j)-minor of A. This formula is called the"formula for expanding $\det A$ according to the cofactors of the m^{th} row." [Hint: Write the m^{th} row as the sum of n vectors, each of which has a single non-zero component. Then use the fact that the determinant function is linear as a function of the m^{th} row.]

The cross-product in V_3

If $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$ are vectors in V_3, we define their cross product to be the vector

$$A \times B = (\det\begin{bmatrix} a_2 & a_3 \\ b_2 & b_3 \end{bmatrix}, -\det\begin{bmatrix} a_1 & a_3 \\ b_1 & b_3 \end{bmatrix} \det\begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}) .$$

We shall describe the geometric significance of this product shortly. But first, we prove some properties of the cross product:

Theorem 27. For all vectors A, B in V_3, we have

(a) $B \times A = -A \times B.$

(b) $A \times (B + C) = A \times B + A \times C,$

$(B + C) \times A = B \times A + C \times A .$

(c) $(cA) \times B = c(A \times B) = A \times (cB) .$

(d) $A \times B$ is orthogonal to both A and B.

(e) $\|A \times B\|^2 = \|A\|^2 \cdot \|B\|^2 - (A \cdot B)^2 .$

Proof. (a) follows because exchanging two rows of a determinant changes the sign; and (b) and (c) follows because the determinant is linear as a function of each row separately. To prove (d), we note that if $C = (c_1, c_2, c_3)$, then

$$C \cdot (A \times B) = \det\begin{bmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix} ,$$

by definition of the determinant. It follows that $A \cdot (A \times B) = B \cdot (A \times B) = 0$ because the determinant vanishes if two rows are equal. The only proof that requires some work is (e). For this, we recall that $(a + b)^2 = a^2 + b^2 + 2ab$, and $(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc$. Equation (e) can be written in the form

$$(a_2b_3 - a_3b_2)^2 + (a_1b_3 - a_3b_1)^2 + (a_1b_2 - a_2b_1)^2 + (a_1b_1 + a_2b_2 + a_3b_3)^2 = (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) .$$

We first take the squared terms on the left side and show they equal the right side. Then we take the "mixed" terms on the left side and show they equal zero. The squared terms on the left side are

$$(a_2b_3)^2 + (a_3b_2)^2 + (a_1b_3)^2 + (a_3b_1)^2 + (a_1b_2)^2 + (a_2b_1)^2 + (a_1b_1)^2 + (a_2b_2)^2 + (a_3b_3)^2,$$

which equals the right side,

$$\sum_{i,j=1}^{3} (a_ib_j)^2 .$$

The mixed terms on the left side are

$$-2a_2b_3a_3b_2 - 2a_1b_3a_3b_1 - 2a_1b_2a_2b_1 + 2a_1b_1a_2b_2 + 2a_1b_1a_3b_3 + 2a_2b_2a_3b_3 = 0. \square$$

In the process of proving the previous theorem, we proved also the following:

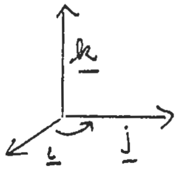
<u>Theorem</u> 28. Given A, B, C , we have $A\cdot(B\times C) = (A\times B)\cdot C$.

<u>Proof.</u> This follows from the fact that

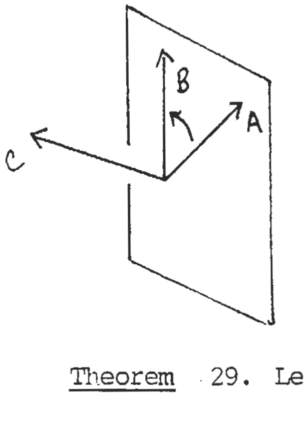
$$\det\begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix} = \det\begin{bmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix}. \square$$

<u>Definition</u>. The ordered 3-tuple of independent vectors (A,B,C) of vectors of V_3 is called a <u>positive</u> triple if $A\cdot(B\times C) > 0$. Otherwise, it is called a <u>negative</u> triple. A positive triple is sometimes said to be a <u>right-handed</u> triple, and a negative one is said to be <u>left-handed</u>.

The reason for this terminology is the following: (1) the triple $(\underline{i}, \underline{j}, \underline{k})$ is a positive triple, since $\underline{i}\cdot(\underline{j}\times\underline{k}) = \det I_3 = 1$, and (2) if we draw the vectors $\underline{i}$, $\underline{j}$, and $\underline{k}$ in V_3 in the usual way, and if one curls the fingers of one's right hand in the direction from the first to the second, then one's thumb points in the direction of the third.



Furthermore, if one now moves the vectors around in V_3, perhaps changing their lengths and the angles between them, but never letting them become dependent, and if one moves one's right hand around correspondingly, then the fingers still correspond to the new triple (A,B,C) in the same way, and this new triple is still a positive triple, since the determinant cannot have changed sign while the vectors moved around. (Since they did not become dependent, the determinant did not vanish.)



Theorem 29. Let A and B be vectors in V_3. If A and B are dependent, then $A\times B = \underline{0}$. Otherwise, $A\times B$ is the unique vector orthogonal to both A and B having length $\|A\|\,\|B\|\sin\theta$ (where θ is the angle between A and B), such that the triple $(A,B,A\times B)$ forms a positive (i.e., right-handed) triple.

Proof. We know that $A \times B$ is orthogonal to both A and B. We also have

$$\begin{aligned} \|A \times B\|^2 &= \|A\|^2 \cdot \|B\|^2 - (A \cdot B)^2 \\ &= \|A\|^2 \cdot \|B\|^2 (1 - \cos^2\theta) = \|A\|^2 \|B\|^2 \sin^2\theta . \end{aligned}$$

Finally, if $C = A \times B$, then (A,B,C) is a positive triple, since

$$A \cdot (B \times C) = (A \times B) \cdot C = (A \times B) \cdot (A \times B) = \|A \times B\|^2 > 0 . \square$$

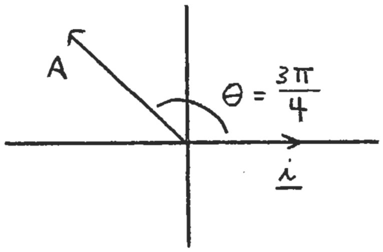
Polar coordinates

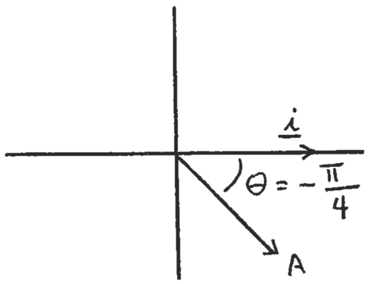
Let $A = (a,b)$ be a point of V_2 different from $\underline{0}$. We wish to define what we mean by a "polar angle" for A. The idea is that it should be the angle between the vector A and the unit vector $\underline{i} = (1,0)$. But we also wish to choose it so its value reflects whether A lies in the upper or lower half–plane. So we make the following definition:

Definition. Given $A = (a,b) \neq \underline{0}$. We define the number

(*) $$\theta = \pm \arccos (A \cdot \underline{i}/\|A\|)$$

to be a *polar angle* for A, where the sign in this equation is specified to be + if $b > 0$, and to be – if $b < 0$. Any number of the form $2m\pi + \theta$ is also defined to be a polar angle for A.





If $b = 0$, the sign in this equation is not determined, but that does not matter. For if $A = (a,0)$ where $a > 0$, then $\arccos (A \cdot i/\|A\|) = \arccos 1 = 0$, so the sign does not matter. And if $A = (-a,0)$ where $a > 0$, then $\arccos (A \cdot \underline{i}/\|A\|) = \arccos (-1) = \pi$. Since the two numbers $+\pi$ and $-\pi$ differ by a multiple of 2π, the sign does not matter, for since one is a polar angle for A, so is the other.

Note: The polar angle θ for A is *uniquely* determined if we require $-\pi < \theta \leq \pi$. But that is a rather artificial restriction.

Theorem. *Let* $A = (a,b) \neq \underline{0}$ *be a point of* V_2. *Let* $r = (a^2+b^2)^{1/2} = \|A\|$; *let* θ *be a polar angle for* A. *Then*

$$A = (r \cos \theta, r \sin \theta).$$

Proof. If $A = (a,0)$ with $a > 0$, then $r = a$ and $\theta = 0 + 2m\pi$; hence

$$r \cos \theta = a \quad \text{and} \quad r \sin \theta = 0.$$

If $A = (-a,0)$ with $a > 0$, then $r = a$ and $\theta = \pi + 2m\pi$, so that

$$r \cos \theta = -a \quad \text{and} \quad r \sin \theta = 0.$$

Finally, suppose $A = (a,b)$ with $b \neq 0$. Then $A \cdot \underline{i}/\|A\| = a/r$, so that

$$\theta = 2m\pi \pm \arccos(a/r).$$

Then

$$a/r = \cos(\pm(\theta - 2m\pi)) = \cos \theta, \quad \text{or} \quad a = r \cos \theta.$$

Furthermore,

$$b^2 = r^2 - a^2 = r^2(1 - \cos^2 \theta) = r^2 \sin^2 \theta,$$

so

$$b = \pm r \sin \theta.$$

We show that in fact $b = r \sin \theta$. For if $b > 0$, then $\theta = 2m\pi + \arccos(a/r)$, so that

$$2m\pi < \theta < 2m\pi + \pi$$

and $\sin \theta$ is positive. Because b, r, and $\sin \theta$ are all positive, we must have $b = r \sin \theta$ rather than $b = -r \sin \theta$.

On the other hand, if $b < 0$, then $0 = 2m\pi - \arccos(a/r)$, so that

$$2m\pi - \pi < \theta < 2m\pi$$

and $\sin \theta$ is negative. Since r is positive, and b and $\sin \theta$ are negative, we must have $b = r \sin \theta$ rather than $b = -r \sin \theta$. □

Planetary Motion

In the text, Apostol shows how Kepler's three (empirical) laws of planetary motion can be deduced from the following two laws:

(1) Newton's second law of motion: $\underline{F} = m\underline{a}$.

(2) Newton's law of universal gravitation:

$$\|\underline{F}\| = G \frac{mM}{r^2}.$$

Here m, M are the masses of the two objects, r is the distance between them, and G is a universal constant.

Here we show (essentially) the reverse—how Newton's laws can be deduced from Kepler's.

More precisely, suppose a planet P of mass m moves in the xy plane with the sun whose mass is M at the origin. Newton's laws tell us that the acceleration of P is given by the equation

$$\underline{a} = \frac{1}{m}\underline{F} = \frac{1}{m}\left[-G\frac{mM}{r^2}\right]\underline{\mu}_r = -\frac{GM}{r^2}\underline{\mu}_r.$$

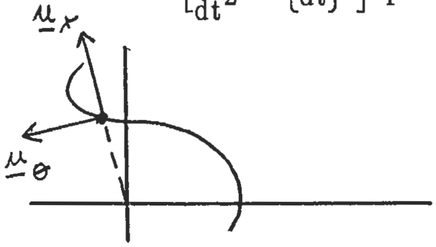
That is, Newton's laws tell us that there is a number λ such that

$$\underline{a} = -\frac{\lambda}{r^2}\underline{\mu}_r,$$

and that λ is the <u>same</u> for all planets in the solar system. (One needs to consider other systems to see that λ involves the mass of the sun.)

This is what we shall prove. We use the formula for acceleration in polar coordinates (Apostol, p. 542):

$$\underline{a} = \left[\frac{d^2r}{dt^2} - r\left[\frac{d\theta}{dt}\right]^2\right]\underline{\mu}_r + \left[2\frac{dr}{dt}\frac{d\theta}{dt} + r\frac{d^2\theta}{dt^2}\right]\underline{\mu}_\theta$$



We also use some facts about area that we shall not actually prove until Units VI and VII of this course.

<u>Theorem. Suppose a planet P moves in the xy plane with the sun at the origin.</u>

(a) <u>Kepler's second law implies that the acceleration is radial.</u>

(b) <u>Kepler's first and second laws imply that</u>

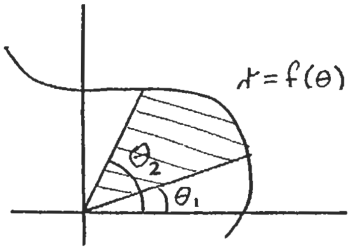
$$\underline{a} = -\frac{\lambda_P}{r^2}\underline{\mu}_r\,,$$

where λ_P is a number that may depend on the particular planet P.

(c) Kepler's three laws imply that λ_P is the same for all planets.

Proof. (a) We use the following formula for the area swept out by the radial vector as the planet moves from polar angle θ_1 to polar angle θ_2:

$$A = \int_{\theta_1}^{\theta_2} \frac{1}{2} r^2 \, d\theta.$$



Here it is assumed the curve is specified by giving r as a function of θ.

Now in our present case both θ and r are functions of time t. Hence the area swept out as time goes from t_0 to t is (by the substitution rule) given by

$$A(t) = \int_{t_0}^{t} \left[\frac{1}{2} r^2 \frac{d\theta}{dt}\right] dt.$$

Differentiating, we have $\frac{dA}{dt} = \frac{1}{2} r^2 \frac{d\theta}{dt}$, which is constant by Kepler's second law. That is,

$$(*) \qquad \boxed{2\frac{dA}{dt} = r^2 \frac{d\theta}{dt} = K}$$

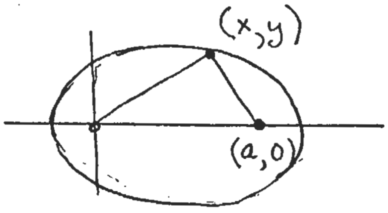
for some K.

Differentiating, we have

$$2r \frac{dr}{dt}\frac{d\theta}{dt} + r^2 \frac{d^2\theta}{dt^2} = 0.$$

The left side of this equation is just the transverse component (the μ_θ component) of a! Hence a is radial.

(b) To apply Kepler's first law, we need the equation of an ellipse with focus at the origin.



We put the other focus at (a,0), and use the fact that an ellipse is the locus of all points (x,y) the sum of whose distances from (0,0) and (a,0) is a constant b > a.

The algebra is routine:

$$\sqrt{x^2 + y^2} + \sqrt{(x-a)^2 + y^2} = b,$$

or
$$r + \sqrt{r^2 - 2a(r \cos \theta) + a^2} = b,$$

$$r^2 - 2a(r \cos \theta) + a^2 = (b-r)^2 = b^2 - 2br + r^2,$$

$$2br - 2ar \cos \theta = b^2 - a^2,$$

$$r = \frac{(b^2-a^2)/2b}{1 - \frac{a}{b}\cos \theta},$$

(**)
$$\boxed{r = \frac{c}{1 - e \cos \theta},\ \text{where } c = \frac{b^2 - a^2}{2b} \text{ and } e = a/b.}$$

(The number e is called the eccentricity of the ellipse, by the way.) Now we compute the radial component of acceleration, which is

$$\left[\frac{d^2r}{dt^2} - r\left[\frac{d\theta}{dt}\right]^2\right].$$

Differentiating (**), we compute

$$\frac{dr}{dt} = c\left[\frac{-1}{(1-e \cos \theta)^2}(e \sin \theta)\frac{d\theta}{dt}\right].$$

Simplifying,

$$\frac{dr}{dt} = \frac{1}{c}(-1)r^2(e \sin \theta)\frac{d\theta}{dt}.$$

Then using (*) from p. B60, we have

$$\frac{dr}{dt} = \frac{1}{c}(e \sin \theta)K.$$

Differentiating again, we have

$$\frac{d^2r}{dt^2} = -\frac{1}{c}(e \cos \theta)\frac{d\theta}{dt}K,$$

or

$$\frac{d^2r}{dt^2} = -\frac{1}{c}(e \cos \theta)\left[\frac{K}{r^2}\right]K, \text{ using (*) to get rid of } d\theta/dt.$$

Similarly,

$$-\,r\left[\frac{d\theta}{dt}\right]^2 = -\,r\left[\frac{K}{r^2}\right]^2 \text{ using (*) again to get rid of } d\theta/dt.$$

Hence the radial component of acceleration is (adding these equations)

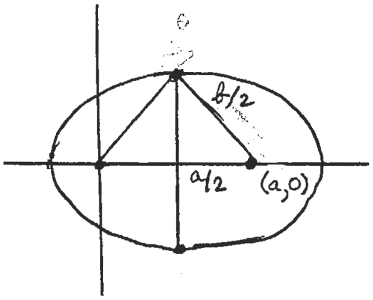
$$\begin{aligned} -\tfrac{1}{c}(e\cos\theta)\frac{K^2}{r^2} - \frac{K^2}{r^3} &= -\frac{K^2}{r^2}\left[\frac{e\ \cos\ \theta}{c} + \frac{1}{r}\right] \\ &= -\frac{K^2}{r^2}\left[\frac{e\ \cos\ \theta}{c} + \frac{1-e\ \cos\ \theta}{c}\right] \\ &= -\left[\frac{K^2}{c}\right]\frac{1}{r^2}\,. \end{aligned}$$

Thus, as desired,

(***) $$\boxed{\underline{a} = -\frac{\lambda_P}{r^2}\underline{\mu}_r, \quad \text{where} \quad \lambda_P = \frac{K^2}{c}\,.}$$

(c) To apply Kepler's third law, we need a formula for the area of an ellipse, which will be proved later, in Unit VII. It is

$$\text{Area} = \pi\frac{(\text{major axis})}{2}\frac{(\text{minor axis})}{2}.$$



The minor axis is easily determined to be given by:

$$\text{minor axis} = 2\sqrt{b^2/4-a^2/4} = \sqrt{b^2-a^2}.$$

It is also easy to see that

$$\text{major axis} = b.$$

Now we can apply Kepler's third law. Since area is being swept out at the constant rate $\frac{1}{2}K$, we know that (since the period is the time it takes to sweep out the entire area),

$$\text{Area} = (\tfrac{1}{2}K)(\text{Period}).$$

Kepler's third law states that the following number is the same for all planets:

$$\frac{(\text{Period})^2}{(\text{major axis})^3} = \frac{4(\text{Area})^2/K^2}{(\text{major axis})^3}$$

$$= \frac{4}{16}\,\frac{\pi^2(\text{major axis})^2(\text{minor axis})^2/K^2}{(\text{major axis})^3}$$

$$= \frac{\pi^2}{4}\,\frac{(\text{minor axis})^2}{(\text{major axis})}\,\frac{1}{K^2}$$

$$= \frac{\pi^2}{4}\,\frac{(b^2-a^2)}{b}\,\frac{1}{K^2}$$

$$= \frac{\pi^2}{2}\left[\frac{c}{K^2}\right] \quad \text{by } (**)$$

$$= \frac{\pi^2}{2}\,\frac{1}{\lambda_P} \quad \text{by } (***).$$

Thus the constant λ_P is the same for all planets. □

SUPPLEMENTARY EXERCISES FOR UNIT III

(1) Let L be a line in V_n with direction vector A; let P be a point not on L. Show that the point X on the line L closest to P satisfies the condition that $X-P$ is perpendicular to A.

(2) Find parametric equations for the curve C consisting of all points of V_2 equidistant from the point P = (0,1) and the line y = -1. If X is any point of C, show that the tangent vector to C at X makes equal angles with the vector $X-P$ and the vector $\vec{j}$. (This is the reflection property of the parabola.)

(3) Consider the curve $f(t) = (t, t\cos(\pi/t))$ for $0 < t \le 1$,

$$= (0,0) \qquad \text{for } t = 0.$$

Then f is continuous. Let P be the partition

$$P = \{0, 1/n, 1/(n-1), \ldots, 1/3, 1/2, 1\}.$$

Draw a picture of the inscribed polygon $\pi(P)$ in the case n = 5. Show that in general, $\pi(P)$ has length

$$|\pi(P)| \ge 1 + 2(1/2 + 1/3 + \ldots + 1/n).$$

Conclude that f is not rectifiable.

(4) Let $\underline{u}$ be a fixed unit vector. A particle moves in V_n in such a way that its position vector $\underline{r}(t)$ satisfies the equation $\underline{r}(t)\cdot\underline{u} = 5t^3$ for all t, and its velocity vector makes a constant angle θ with $\underline{u}$, where $0 < \theta < \pi/2$.

(a) Show that $\|\underline{v}\| = 15t^2/\cos\theta$.

(b) Compute the dot product $\underline{a}(t)\cdot\underline{v}(t)$ in terms of t and θ.

(5) A particle moves in 3—space so as to trace out a curve of constant curvature $\kappa = 3$. Its speed at time t is e^{2t}. Find $\|\underline{a}(t)\|$, and find the angle between $\underline{v}$ and $\underline{a}$ at time t.

(6) Consider the curve given in polar coordinates by the equation $r = e^{-\theta}$ for $0 \leq \theta \leq 2\pi M$, where M is a positive integer. Find the length of this curve. What happens as M becomes arbitrarily large?

(7) (a) Derive the following formula, which can be used to compute the curvature of a curve in R^n:

$$(\underline{v}\cdot\underline{v})^2 \kappa \underline{N} = (\underline{v}\cdot\underline{v})\underline{a} - (\underline{a}\cdot\underline{v})\underline{v}.$$

(b) Find the curvature of the curve $\underline{r}(t) = (1+t,\ 3t,\ 2+t^2,\ 2t^2)$.

18.024 Multivariable Calculus with Theory
Spring 2011

Derivatives of vector functions.

Recall that if $\underline{x}$ is a point of R^n and if $f(\underline{x})$ is a scalar function of $\underline{x}$, then the derivative of f (if it exists) is the vector

$$\vec{\nabla} f = (D_1 f, \ldots, D_n f)$$

$$= (\partial f/\partial x_1, \ldots, \partial f/\partial x_n).$$

For some purposes, it will be convenient to denote the derivative of f by a row matrix rather than by a vector. When we do this, we usually denote the derivative by Df rather than $\vec{\nabla} f$. Thus

$$Df(\underline{a}) = [D_1 f(\underline{a}) \quad D_2 f(\underline{a}) \quad \ldots \quad D_n f(\underline{a})].$$

If we use this notation, the definition of the derivative takes the following form:

$$f(\underline{a}+\underline{h}) - f(\underline{a}) = Df(\underline{a}) \cdot \underline{h} + \varepsilon(\underline{h}) \|\underline{h}\|,$$

where $\varepsilon(\underline{h}) \longrightarrow 0$ as $\underline{h} \longrightarrow \underline{0}$. Here the dot denotes matrix multiplication, so we must write $\underline{h}$ as a column matrix in order for the formula to work;

$$\underline{h} = \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}.$$

This is the formula that will generalize to vector functions $\underline{f}$.

Definition. Let S be a subset of R^n. If $\underline{f} : S \longrightarrow R^k$, then $\underline{f}(\underline{x})$ is called a vector function of a vector variable. In scalar form, we can write $\underline{f}(\underline{x})$ out in the form

$$\underline{f}(\underline{x}) = (f_1(x_1,\ldots,x_n),\ldots,f_k(x_1,\ldots,x_n)).$$

Said differently, $\underline{f}$ consists of "k real-valued functions of n variables." Suppose now that $\underline{f}$ is defined in an open ball about the point $\underline{a}$. We say that $\underline{f}$ is differentiable at $\underline{a}$ if each of the functions $f_1(\underline{x}),\ldots,f_k(\underline{x})$ is differentiable at $\underline{a}$ (in the sense already defined). Furthermore, we define the derivative of $\underline{f}$ at $\underline{a}$ to be the matrix

$$D\underline{f}(\underline{a}) = \begin{bmatrix} D_1f_1(\underline{a}) & D_2f_1(\underline{a}) & \ldots & D_nf_1(\underline{a}) \\ D_1f_2(\underline{a}) & D_2f_2(\underline{a}) & \ldots & D_nf_2(\underline{a}) \\ & & \ldots & \\ D_1f_k(\underline{a}) & D_2f_k(\underline{a}) & \ldots & D_nf_k(\underline{a}) \end{bmatrix}.$$

That is, $D\underline{f}(\underline{a})$ is the matrix whose $i^{\underline{th}}$ row is the derivative $Df_i(\underline{a})$ of the $i^{\underline{th}}$ coordinate function of $\underline{f}$.

Said differently, the derivative $D\underline{f}(\underline{a})$ of $\underline{f}$ at $\underline{a}$ is the k by n matrix whose entry in row i and column j is

$$D_jf_i(\underline{x}) = \partial f_i/\partial x_j;$$

it is often called the *Jacobian matrix* of $\underline{f}(\underline{x})$. Another notation for this matrix is the notation

$$\frac{\partial(f_1,\ldots,f_k)}{\partial(x_1,\ldots,x_n)}\ .$$

With this notation, many of the formulas we proved for a scalar function $f(\underline{x})$ hold without change for a vector function $\underline{f}(\underline{x})$. We consider some of them here:

Theorem 1. *The function* $\underline{f}(\underline{x})$ *is differentiable at* $\underline{a}$ *if and only if*

$$\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a}) = D\underline{f}(\underline{a})\cdot\underline{h} + \underline{E}(\underline{h})\|\underline{h}\|,$$

where $\underline{E}(\underline{h}) \longrightarrow 0$ *as* $\underline{h} \longrightarrow \underline{0}$.
(Here $\underline{f}$, $\underline{h}$, and $\underline{E}$ are written as column matrices.)

Proof: Both sides of this equation represent column matrices. If we consider the $i^{\underline{th}}$ entries of these matrices, we have the following equation:

$$f_i(\underline{a}+\underline{h}) - f_i(\underline{a}) = Df_i(\underline{a})\cdot\underline{h} + E_i(\underline{h})\|\underline{h}\|.$$

Now $\underline{f}$ is differentiable at $\underline{a}$ if and only if each function f_i is. And f_i is differentiable at $\underline{a}$ if and only if $E_i(\underline{h}) \longrightarrow 0$ as $\underline{h} \longrightarrow \underline{0}$. But $E_i(\underline{h}) \longrightarrow 0$ as $\underline{h} \longrightarrow \underline{0}$, for each i, if and only if $\underline{E}(\underline{h}) \longrightarrow \underline{0}$ as $\underline{h} \longrightarrow \underline{0}$. □

Theorem 2. If $\underline{f}(\underline{x})$ is differentiable at $\underline{a}$, then $\underline{f}$ is continuous at $\underline{a}$.

Proof. If $\underline{f}$ is differentiable at $\underline{a}$, then so is each function f_i. Then in particular, f_i is continuous at $\underline{a}$, whence $\underline{f}$ is continuous at $\underline{a}$.

The general chain rule.

Before considering the general chain rule, let us take the chain rule we have already proved and reformulate it in terms of matrices.

Assume that $f(\underline{x}) = f(x_1,\ldots,x_n)$ is a scalar function defined in an open ball about $\underline{a}$, and that $\underline{x}(t) = (x_1(t),\ldots, x_n(t))$ is a parametrized curve passing through $\underline{a}$. Let $\underline{x}(t_0) = \underline{a}$. If $f(\underline{x})$ is differentiable at $\underline{a}$, and if $\underline{x}(t)$ is differentiable at t_0, and we have shown that the composite $f(\underline{x}(t))$ is differentiable at t_0, and its derivative is given by the equation

$$\frac{d}{dt} f(\underline{x}(t)) = \vec{\nabla} f(\underline{x}(t)) \cdot \underline{x}'(t)$$

when $t = t_0$.

We can rewrite this formula in scalar form as follows:

$$\frac{d}{dt} f(\underline{x}(t)) = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt} ;$$

or we can rewrite it in the following matrix form:

$$\frac{d}{dt}\ f(\underline{x}(t)) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \cdot \begin{bmatrix} dx_1/dt \\ \vdots \\ dx_n/dt \end{bmatrix}$$

Recalling the definition of the Jacobian matrix $D\underline{f}$, we see that the latter formula can be written in the form

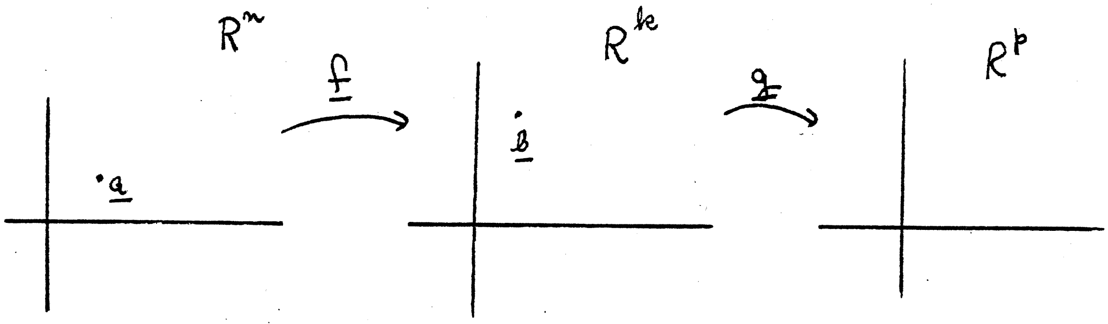
$$\frac{d}{dt}\ f(\underline{x}(t)) = Df(\underline{x}(t)) \cdot D\underline{x}(t).$$

(Note that the matrix Df is a row matrix, while the matrix $D\underline{x}$ is by its definition a column matrix.)

This is the form of the chain rule that we find especially useful, for it is the formula that generalizes to higher dimensions.

Let us now consider a composite of vector functions of vector variables. For the remainder of this section, we assume the following:

Suppose $\underline{f}$ *is defined on an open ball in* R^n *about* $\underline{a}$, *taking values in* R^k, *with* $\underline{f}(\underline{a}) = \underline{b}$. *Suppose* $\underline{g}$ *is defined in an open ball about* $\underline{b}$, *taking values in* R^p. *Let* $\underline{F}(\underline{x}) = \underline{g}(\underline{f}(\underline{x}))$ *denote the composite function.*



We shall write these functions as

$$\underline{f}(\underline{x}) = \underline{f}(x_1,\ldots,x_n) \quad \text{and} \quad \underline{g}(\underline{y}) = \underline{g}(y_1,\ldots,y_k).$$

If $\underline{f}$ and $\underline{g}$ are differentiable at $\underline{a}$ and $\underline{b}$ respectively, it is easy to see that the partial derivatives of $\underline{F}(\underline{x})$ exist at $\underline{a}$, and to calculate them. After all, the $i^{\underline{th}}$ coordinate function of $\underline{F}(\underline{x})$ is given by the equation

$$F_i(\underline{x}) = g_i(\underline{f}(\underline{x})).$$

If we set each of the variables x_ℓ, except for the single variable x_j, equal to the constant a_ℓ, then both sides are functions of x_j alone. The chain rule already proved then gives us the formula

$$(*) \qquad \frac{\partial F_i}{\partial x_j} = \frac{\partial g_i}{\partial y_1}\frac{\partial f_1}{\partial x_j} + \frac{\partial g_i}{\partial y_2}\frac{\partial f_2}{\partial y_2} + \cdots + \frac{\partial g_i}{\partial y_k}\frac{\partial f_k}{\partial x_j}.$$

Thus

$$D_jF_i = [D_1g_i \quad D_2g_i \quad \cdots \quad D_kg_i] \cdot \begin{bmatrix} D_jf_1 \\ D_jf_2 \\ \vdots \\ D_jf_k \end{bmatrix}.$$

$$= [i^{\underline{th}} \text{ row of } D\underline{g}] \cdot \begin{bmatrix} j^{\underline{th}} \text{ column} \\ \text{of } D\underline{f} \end{bmatrix}.$$

domains, then the composite $\underline{F}(\underline{x}) = \underline{g}(\underline{f}(\underline{x}))$ *is continuously differentiable on its domain, and*

$$D\underline{F}(\underline{x}) = D\underline{g}(\underline{f}(\underline{x})) \cdot Df(\underline{x}).$$

This theorem is adequate for all the chain-rule applications we shall make.

Note: The matrix form of the chain rule is nice and neat, and it is useful for theoretical purposes. In practical situations, one usually uses the scalar formula (*) when one calculates partial derivatives of a composite function, however.

The following proof is included solely for completeness ; we shall not need to use it:

Theorem 4. Let $\underline{f}$ *and* $\underline{g}$ *be as above. If* f *is differentiable at* $\underline{a}$ *and* $\underline{g}$ *is differentiable at* $\underline{b}$, *then*

$$\underline{F}(\underline{x}) = \underline{g}(\underline{f}(\underline{x}))$$

is differentiable at $\underline{a}$, *and*

$$D\underline{F}(\underline{a}) = D\underline{g}(\underline{b}) \cdot D\underline{f}(\underline{a}).$$

Proof. We know that

$$\underline{g}(\underline{b}+\underline{k}) - \underline{g}(\underline{k}) = D\underline{g}(\underline{b}) \cdot \underline{k} + \underline{E}_1(\underline{k})\|\underline{k}\|,$$

where $\underline{E}_1(k) \longrightarrow 0$ as $\underline{k} \longrightarrow \underline{0}$. Let us set $\underline{k} = \underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})$ in this formula. Then

Thus the Jacobian matrix of $\underline{F}$ satisfies the matrix equation

$$D\underline{F}(\underline{a}) = D\underline{g}(\underline{b}) \cdot D\underline{f}(\underline{a}).$$

This is our generalized version of the chain rule.

There is, however, a problem here. We have just shown that if $\underline{f}$ and $\underline{g}$ are differentiable, then the partial derivatives of the composite function $\underline{F}$ exist. But we know that the mere *existence* of the partial derivatives of the function F_i is not enough to guarantee that F_i is differentiable. One needs to give a separate proof that if both $\underline{f}$ and $\underline{g}$ are differentiable, then so is the composite $\underline{F}(\underline{x}) = \underline{f}(\underline{g}(\underline{x}))$. (See Theorem 4 following.)

One can avoid giving a separate proof that $\underline{F}$ is differentiable by assuming a stronger hypothesis, namely that both $\underline{f}$ and $\underline{g}$ are *continuously* differentiable. In this case, the partials of $\underline{f}$ and $\underline{g}$ are continuous on their respective domains; then the formula

$$D_j F_i(\underline{x}) = \sum_{\ell=1}^{k} D_\ell g_i(\underline{f}(\underline{x})) \cdot D_j f_\ell(\underline{x}),$$

which we have proved, shows that $D_j F_i$ is also a continuous function of $\underline{x}$. Then by our basic theorem, F_i is differentiable for each i, so that $\underline{F}$ is differentiable, by definition.

We summarize these facts as follows:

Theorem 3. Let $\underline{f}$ *be defined on an open ball in* R^n *about* a, *taking values in* R^k; *let* $f(\underline{a}) = \underline{b}$. *Let* $\underline{g}$ *be defined in an open ball about* $\underline{b}$, *taking values in* R^p. *If* $\underline{f}$ *and* $\underline{g}$ *are continuously differentiable on their respective*

$$(**)\qquad \underline{g}(\underline{f}(\underline{a}+\underline{h})) - \underline{g}(\underline{f}(\underline{a})) = D\underline{g}(\underline{b}) \cdot (\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a}))$$

$$+ \underline{E}_1(\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a}))\|\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})\| .$$

Now we know that

$$\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a}) = D\underline{f}(\underline{a}) \cdot \underline{h} + \underline{E}_2(\underline{h})\|\underline{h}\| ,$$

where $\underline{E}_2(\underline{h}) \longrightarrow \underline{0}$ as $\underline{h} \longrightarrow \underline{0}$. Plugging this into (**), we get the equation

$$\underline{g}(\underline{f}(\underline{a}+\underline{h}) - \underline{g}(\underline{f}(\underline{a})) = D\underline{g}(\underline{b}) \cdot D\underline{f}(\underline{a}) \cdot \underline{h} + D\underline{g}(\underline{b}) \cdot \underline{E}_2(\underline{h})\|\underline{h}\|$$

$$+ E_1(\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a}))\|\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})\| .$$

Thus

$$\underline{F}(\underline{a}+\underline{h}) - \underline{F}(\underline{a}) = D\underline{g}(\underline{b}) \cdot D\underline{f}(\underline{a}) \cdot \underline{h} + \underline{E}_3(\underline{h})\|\underline{h}\| ,$$

where

$$\underline{E}_3(\underline{h}) = D\underline{g}(\underline{b}) \cdot \underline{E}_2(\underline{h}) + \underline{E}_1(\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a}))\frac{\|\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})\|}{\|\underline{h}\|} .$$

We must show that $\underline{E}_3 \longrightarrow \underline{0}$ as $\underline{h} \longrightarrow \underline{0}$. The first term is easy, since $Dg(\underline{b})$ is constant and $\underline{E}_2(\underline{h}) \longrightarrow \underline{0}$ as $h \longrightarrow 0$. Furthermore, as $\underline{h} \longrightarrow \underline{0}$, the expression $\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})$ approaches $\underline{0}$ (since $\underline{f}$ is continuous), so that

$\underline{E}_1(\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})) \longrightarrow \underline{0}$. We need finally to show that the expression

$$\|\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})\| / \|\underline{h}\|$$

is bounded as $\underline{h} \longrightarrow \underline{0}$. Then we will be finished. Now

$$\frac{\|\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})\|}{\|\underline{h}\|} = \|D\underline{f}(\underline{a}) \cdot \frac{\underline{h}}{\|\underline{h}\|} + \underline{E}_2(\underline{h})\|$$

$$\leqslant \|D\underline{f}(\underline{a}) \cdot \underline{u}\| + \|\underline{E}_2(\underline{h})\|,$$

where $\underline{u}$ is a unit vector. Now $E_2(\underline{h}) \longrightarrow \underline{0}$ as $\underline{h} \longrightarrow \underline{0}$, and it is easy to see that $\|Df(\underline{a}) \cdot \underline{u}\| \leqslant nk \max|D_i f_j(\underline{a})|$. (Exercise!) Hence the expression $\|\underline{f}(\underline{a}+\underline{h}) - \underline{f}(\underline{a})\| / \|\underline{h}\|$ is bounded, and we are finished. □

<u>Differentiating</u> <u>inverse</u> <u>functions</u>.

Recall that if $f(x)$ is a differentiable real-valued function of a single real variable x, and if $f'(x) > 0$ for $a \leqslant x \leqslant b$, then f is strictly increasing, so it has an inverse g. Furthermore, g is differentiable and its derivative satisfies the formula

$$g'(f(x)) = \frac{1}{f'(x)} .$$

Part, but not all, of this theorem generalizes to vector functions. We shall show that <u>if</u> $\underline{f}$ has an inverse $\underline{g}$, and *if* $\underline{g}$ is differentiable, then there is a formula for $D\underline{g}$ analogous to this one. Specifically, we prove the following:

Theorem 5. *Let* S *be a subset of* R^n. *Suppose that* $\underline{f} : A \longrightarrow R^n$ *and that* $\underline{f}(\underline{a}) = \underline{b}$. *Suppose also that* $\underline{f}$ *has an inverse* $\underline{g}$.

If $\underline{f}$ *is differentiable at* $\underline{a}$, *and if* $\underline{g}$ *is differentiable at* $\underline{b}$, *then*

$$D\underline{g}(\underline{b}) = [D\underline{f}(\underline{a})]^{-1}.$$

Proof. Because $\underline{g}$ is inverse to $\underline{f}$, the equation $\underline{g}(\underline{f}(\underline{x})) = \underline{x}$ holds for all $\underline{x}$ in S. Now both $\underline{f}$ and $\underline{g}$ are differentiable and so is the composite function $\underline{g}(\underline{f}(\underline{x}))$. Thus we can use the chain rule to compute

$$D\underline{g}(\underline{b}) \cdot D\underline{f}(\underline{a}) = D(\text{identity}) = I_n.$$

Since the matrices involved are n by n, this equation implies that

$$Dg(\underline{b}) = [Df(\underline{a})]^{-1}. \qquad \square$$

Remark 1. This theorem shows that in order for the differentiable function $\underline{f}$ to have a differentiable inverse, it is *necessary* that the Jacobian matrix $D\underline{f}(\underline{a})$ have rank n. Roughly speaking, this condition is also *sufficient* for $\underline{f}$ to have an inverse.

More precisely, one has the following result, which is the famous "Inverse Function Theorem" of Analysis:

Suppose $\underline{f}$ is defined and continuously differentiable in an open ball of R^n about $\underline{a}$, taking values in R^n. If $D\underline{f}(\underline{a})$ has rank n, then there is some (probably smaller) open ball B about $\underline{a}$, such that $\underline{f}$ carries B in a 1-1 fashion onto an open set C in R^n. Furthermore, the inverse function $\underline{g} : C \longrightarrow B$ is continuously differentiable, and $D\underline{g}(\underline{f}(\underline{x})) = [D\underline{f}(\underline{x})]^{-1}$.

Remark 2. For a function of a single variable, $y = f(x)$, the rule for the derivative of the inverse function $x = g(y)$ is often written in the form

$$\frac{dx}{dy} = \frac{1}{dy/dx} \cdot$$

This formula is easy to remember; the Leibnitz notation for derivatives "does the work for you". It is tempting to think that a similar result should hold for a function of several variables. It does not.

For example, suppose

$$\underline{f}(x,y) = (u,v)$$

is a differentiable transformation from the $x - y$ plane to the $u - v$ plane. And suppose it has an inverse; given by

$$(x,y) = \underline{g}(u,v).$$

Our theorem tells us that if $\underline{f}(\underline{a}) = \underline{b}$, then

$$D\underline{g}(\underline{b}) = [D\underline{f}(\underline{a})]^{-1}.$$

If we write out these matrices in Leibnitz notation, we obtain the equation

$$\begin{bmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{bmatrix} = \begin{bmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial v/\partial y \end{bmatrix}^{-1}.$$

Now the formula for the inverse of a matrix gives (in the case of a 2 by 2 matrix) the formula

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Applying this formula, we obtain the equation

$$\begin{bmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{bmatrix} = \frac{1}{\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}}\begin{bmatrix} \partial v/\partial y & -\partial u/\partial y \\ -\partial v/\partial x & \partial u/\partial x \end{bmatrix}.$$

This means, for example, that

$$\frac{\partial x}{\partial v} = \frac{1}{\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}}\left(- \frac{\partial u}{\partial y}\right).$$

Thus the simplistic idea that $\partial x/\partial v$ "should be" the reciprocal of $\partial v/\partial x$ is very far from the truth. The Leibnitz notation simply doesn't "do the work for you" in dimensions greater than 1. Matrix notation does.

Implicit differentiation.

Suppose $\underline{F}$ is a function from R^{n+k} to R^n; let us write it in the form

$$\underline{F}(\underline{x},\underline{y}) = \underline{F}(x_1,\ldots,x_n,y_1,\ldots,y_k).$$

Let $\underline{c}$ be a point of R^n, and consider the equation

$$\underline{F}(\underline{x},\underline{y}) = \underline{c}.$$

This equation represents a system of n equations in $n + k$ unknowns. In general, we would expect to be able to solve this system for $\underline{n}$ of the unknowns in terms of the others. For instance, in the present case we would expect to be able to solve this system for $\underline{x}$ in terms of $\underline{y}$. We would also expect the resulting function $\underline{x} = \underline{g}(\underline{y})$ to be differentiable.

Assuming this expectation to be correct, one can then calculate the derivative of the resulting function $\underline{g}$ by using the chain rule. One understands best how this is done by working through a number of examples. Apostol works several in sections 9.6 and 9.7. At this point, you should read 9.6 and Examples 1,2,3, and 6 of 9.7.

A natural question to ask now is the following: to what extent our assumptions are correct, that the given equation determines $\underline{x}$ as a function of $\underline{y}$. We discuss that question now.

First let us consider the problem discussed on p. 294 of the text. It involves an equation of the form

$$F(x,y,z) = 0 ,$$

where F is continuously differentiable. Assuming that one can in theory solve this equation for z as a function of x and y, say $z = f(x,y)$, Apostol derives equations for the partials of this unknown function:

$$\frac{\partial f}{\partial x}(x,y) = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial z}} \quad \text{and} \quad \frac{\partial f}{\partial y} = -\frac{\frac{\partial F}{\partial y}}{\frac{\partial F}{\partial z}} .$$

Here the functions on the right side of these equations are evaluated at the point $(x,y,f(x,y))$.

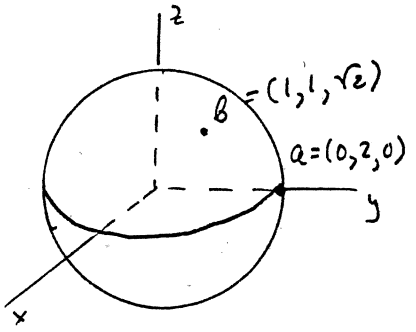
Note that it was necessary to assume that $\partial F/\partial z \neq 0$, in order to carry out these calculations. It is a remarkable fact that the condition $\partial F/\partial z \neq 0$ is also sufficient to justify the assumptions we made in carrying them out. This is a consequence of a famous theorem of Analysis called the Implicit Function Theorem. One consequence of this theorem is the following: If one has a point (x_0,y_0,z_0) that satisfies the equation $F(x,y,z) = 0$, and if $\partial F/\partial z \neq 0$ at this point, then there exists a unique differentiable function $f(x,y)$, defined in an open set B about (x_0,y_0), such that $f(x_0,y_0) = z_0$ and such that

$$F(x,y,f(x,y)) = 0$$

for all (x,y) in B. Of course, once one knows that f exists and is differentiable, one can find its partials by implicit differentiation, as explained in the text.

Example 1. Let $F(x,y,z) = x^2 + y^2 + z^2 + 1$. The equation $F(x,y,z) = 0$ cannot be solved for z in terms of x and y; for in fact there is no point that satisfies the equation.

Example 2. Let $F(x,y,z) = x^2 + y^2 + z^2 - 4$. The equation $F(x,y,z) = 0$ is satisfied by the point $a = (0,2,0)$. But $\partial F/\partial z = 0$ at the point a, so the implicit function theorem does not apply. This fact is hardly surprising, since it is clear from the picture that z is not determined as a function of (x,y) in an open set about the point $(x_0,y_0) = (0,2)$.



However, the point $b = (1,1,\sqrt{2})$ satisfies the equation also, and $\partial F/\partial z \neq 0$ at this point. The implicit function theorem implies that there is a function $f(x,y)$ defined in a neighborhood of $(x_0,y_0) = (1,1)$ such that $f(1,1) = \sqrt{2}$ and f satisfies the equation $F(x,y,z) = 0$ identically. Note that f is not uniquely determined unless we specify its value at (x_0,y_0). There are two functions f defined in a neighborhood of $(1,1)$ that satisfy the equation $f(x,y,z) = 0$, namely,

$$z = [4 - x^2 - y^2]^{\frac{1}{2}} \quad \text{and} \quad z = - [4 - x^2 - y^2]^{\frac{1}{2}} .$$

However, only one of them satisfies the condition $f(1,1) = \sqrt{2}$.

Note that at the point $a = (0,2,0)$ we do have the condition $\partial F/\partial y \neq 0$. Then the implicit function theorem implies that y is determined as a function of (x,z) near this point. The picture makes this fact clear.

Now let us consider the more general situation discussed on p. 296 of the text. We have two equations

$$(*)\qquad \begin{aligned} F(x,y,z,w) &= 0 \\ G(x,y,z,w) &= 0 \end{aligned}$$

where F and G are continuously differentiable. (We have inserted an extra variable to make things more interesting.) Assuming there are functions $x = X(z,w)$ and $y = Y(z,w)$ that satisfy these equations for all points in an open set in the (z,w) plane, we have the identities

$$F(X,Y,z,w) = 0 \quad \text{and} \quad G(X,Y,z,w) = 0 \ ,$$

whence (differentiating with respect to z),

$$\frac{\partial F}{\partial x}\frac{\partial X}{\partial z} + \frac{\partial F}{\partial y}\frac{\partial Y}{\partial z} + \frac{\partial F}{\partial z} = 0 \ ,$$

$$\frac{\partial G}{\partial x}\frac{\partial X}{\partial z} + \frac{\partial G}{\partial y}\frac{\partial Y}{\partial z} + \frac{\partial G}{\partial z} = 0 \ .$$

These are linear equations for $\partial X/\partial z$ and $\partial Y/\partial z$; we can solve them if the coefficient matrix

$$\begin{bmatrix} \partial F/\partial x & \partial F/\partial y \\ \partial G/\partial x & \partial G/\partial y \end{bmatrix} = \frac{\partial F,G}{\partial x,y}$$

is non-singular. One can use Cramer's rule, as in the text, or one can write the solution in the form

$$\begin{bmatrix} \partial X/\partial z \\ \partial Y/\partial z \end{bmatrix} = -\left(\frac{\partial F,G}{\partial x,y}\right)^{-1} \begin{bmatrix} \partial F/\partial z \\ \partial G/\partial z \end{bmatrix} \ .$$

The functions on the right side of this equation are evaluated at the point $(X(z,w),Y(z,w),z,w)$, so that both sides of the equation are functions of z and w alone.

You can check that one obtains an equation for the other partials of X and Y if one replaces z by w throughout.

All this is discussed in the text. But now let us note that in order to carry out these calculations, it was necessary to assume that the matrix $\partial F,G/\partial x,y$ was non-singular. Again , it is a remarkable fact that this condition is also sufficient to justify the assumptions we have made. Specifically, the Implicit Function Theorem tells us that if (x_0,y_0,z_0,w_0) is a point satisfying the equations (*), and if the matrix $\partial F,G/\partial x,y$ is non-singular at this point, then there do exist unique differentiable functions $X(z,w)$ and $Y(z,w)$ defined in an open set about (z_0,w_0) , such that

$$X(z_0,w_0) = x_0 \quad \text{and} \quad Y(z_0,w_0) = y_0 ,$$

and such that F and G vanish identically when X and Y are substituted for x and y. Thus under this assumption all our calculations are justified.

Example 3. Consider the equations

$$F(x,y,z,w) = 3x^2z + 6wy^2 - 2z + 1 = 0 ,$$
$$G(x,y,z,w) = xz - 4y/z - 3w - 7 = 0 .$$

The points $(x_0,y_0,z_0,w_0) = (1,2,-1,0)$ and $(x_1,y_1,z_1,w_1) = (1,\frac{1}{2},2,-2)$ satisfy these equations, as you can check. We calculate

$$\partial F,G/\partial x,y = \begin{bmatrix} 6xz & 12wy \\ z & -4/z \end{bmatrix} .$$

At the point (x_0,y_0,z_0,w_0), this matrix equals $\begin{bmatrix} -6 & 0 \\ -1 & 4 \end{bmatrix}$, which is non-singular. Therefore, there exist unique functions $x = X(z,w)$ and $y = Y(z,w)$, defined in a neighborhood of $(z_0,w_0) = (-1,0)$ that satisfy these equations identically, such that $X(-1,0) = 1$ and $Y(-1,0) = 2$.

Since we know the values of X and Y at the point $(-1,0)$, we can find the values of their partial derivatives at this point also. Indeed,

$$\begin{bmatrix} \partial X/\partial z \\ \partial Y/\partial z \end{bmatrix} = -\begin{bmatrix} 6Xz & 12wY \\ z & -4/z \end{bmatrix}^{-1} \cdot \begin{bmatrix} 3X^2 - 2 \\ X + 4Y/z^2 \end{bmatrix}$$

$$= -\begin{bmatrix} 6 & 0 \\ -1 & 4 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 \\ 9 \end{bmatrix} = -\frac{1}{24}\begin{bmatrix} 4 \\ 55 \end{bmatrix} .$$

On the other hand, at the point $(x_1,y_1,z_1,w_1) = (1,\frac{1}{2},2,-2)$ the matrix $\partial F,G/\partial x,y$ equals

$$\begin{bmatrix} 12 & -12 \\ 2 & -2 \end{bmatrix},$$

which is singular. Therefore we do not expect to be able to solve for x and y in terms of z and w near this point. However, at this point, we have

$$\partial F,G/\partial x,w = \begin{bmatrix} 6xz & 6y^2 \\ z & -3 \end{bmatrix} = \begin{bmatrix} 12 & 3/2 \\ 2 & -3 \end{bmatrix} .$$

Therefore, the implicit function theorem implies that we can solve for x and w in terms of y and z near this point.

Exercises

1. Given the continuously differentiable scalar field $f(x,y)$, let $\phi(t) = f(t^2,t^3 + 1)$. Find $\phi'(1)$, given that $\vec{\nabla} f(1,2) = 5\vec{i} - \vec{j}$.

2. Find the point on the surface $z = xy$ nearest the point $(2,2,0)$.

3. A rectangular box, open at the top, is to hold 256 cubic inches. Find the dimensions that minimize surface area.

4. Find parametric equations for the tangent line to the curve of intersection of the surfaces

$$x^2 + y^2 + 2z^2 = 13,$$

$$z = x^2 - xy^3 ,$$

at the point $(2,1,2)$.

5. Let f be a scalar function of 3 variables. Define

$$F(t) = f(3t^2, 2t+1, 3-t^3).$$

Express $F'(1)$ in terms of the first order partials of f at the point $(3,3,2)$.

Express $F''(1)$ in terms of the first and second order partials of f at the point $(3,3,2)$.

6. Let $\underline{f} : R^2 \longrightarrow R^2$ and let $\underline{g} : R^2 \longrightarrow R^3$. Suppose that

$\underline{f}(0,0) = (1,2)$ $\qquad$ $\underline{f}(1,2) = (0,0)$.

$\underline{g}(0,0) = (1,3,2)$ $\qquad$ $\underline{g}(1,2) = (-1,0,1)$.

Suppose that

$$D\underline{f}(0,0) = \begin{bmatrix} -1 & 2 \\ 6 & 3 \end{bmatrix} \qquad D\underline{f}(1,2) = \begin{bmatrix} -1 & 3 \\ -2 & 4 \end{bmatrix}$$

$$D\underline{g}(0,0) = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \qquad D\underline{g}(1,2) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 2 & 1 \end{bmatrix}.$$

a) If $\underline{h}(\underline{x}) = \underline{g}(\underline{f}(\underline{x}))$, find $D\underline{h}(0,0)$.

b) If $\underline{f}$ has an inverse $\underline{k} : R^2 \longrightarrow R^2$, find $D\underline{k}(0,0)$.

7. Consider the functions of Example 3. Find the partials $\partial X/\partial w$ and $\partial Y/\partial w$ at the point $(z_0,w_0) = (-1, 0)$.

8. For the functions F and G of Example 3, compute $\partial(F,G)/\partial(x,y)$ at the point $(1,\frac{1}{2},2,-2,)$. Given the equations $F = 0$, $G = 0$, for which pairs of variables is it possible to solve in terms of the other two near this point?

The second-derivative test for extrema of a function of two variables.

Theorem. *Suppose that* $f(x_1,x_2)$ *has continuous second-order partial derivatives in a ball* B *about* a. *Suppose that* D_1f *and* D_2f *vanish at* a. *Let*

$$A = D_{1,1}f(\underline{a}), \qquad B = D_{1,2}f(\underline{a}), \qquad C = D_{2,2}f(\underline{a}).$$

(a) *If* $B^2 - AC > 0$, *then* f *has a saddle point at* a.

(b) *If* $B^2 - AC < 0$ *and* $A > 0$, *then* f *has a relative minimum at* a.

(c) *If* $B^2 - AC < 0$ *and* $A < 0$, *then* f *has a relative maximum at* a.

(d) *If* $B^2 - AC = 0$, *the test is inconclusive*.

Proof. *Step 1*. We first prove a version of Taylor's theorem with remainder for functions of two variables:

Suppose $f(x_1,x_2)$ has continuous second-order partials in a ball B centered at $\underline{a}$. Let $\underline{v}$ be a fixed vector; say $\underline{v} = (h,k)$. Then

$$(*) \quad \begin{aligned} f(\underline{a}+t\underline{v}) &= f(\underline{a}) + [D_1f(\underline{a})\cdot h + D_2f(\underline{a})\cdot k]t \\ &+ \{D_{1,1}f(\underline{a}^*)h^2 + 2D_{1,2}f(\underline{a}^*)hk + D_{2,2}f(\underline{a}^*)k^2\}\frac{t^2}{2}, \end{aligned}$$

where $\underline{a}^*$ is some point on the line segment from $\underline{a}$ to $\underline{a} + t\underline{v}$.

We derive this formula from the single-variable form of Taylor's theorem. Let $g(t) = f(\underline{a}+t\underline{v})$, i.e., Let $F(t\underline{v})$ denote the left side of this equation. We will be concerned about the sign of $F(t\underline{v})$ when t is small, because that sign will depend on whether f has a local maximum or a local minimum at $\underline{a}$, or neither.

Step 3. From now on, let $\underline{v} = (h,k)$ be a unit vector. Consider the quadratic function

$$Q(\underline{v}) = Q(h,k) = Ah^2 + 2Bhk + Ck^2.$$

We shall determine what values Q takes as $\underline{v}$ varies over the unit circle.

Case 1. If $B^2 - AC < 0$, then we show that $Q(\underline{v})$ has the same sign as A, for all unit vectors $\underline{v}$.

Proof. When $\underline{v} = (1,0)$, then $Q(\underline{v}) = A$; thus $Q(\underline{v})$ has the same sign as A in this case. Consider the continuous function $Q(\cos t, \sin t)$. As t ranges over the interval $[0,2\pi]$, the vector $(\cos t, \sin t)$ ranges over all unit vectors in V_2. If this function takes on a value whose sign is different from that of A, then by the intermediate-value theorem, there must be a t_o such that $Q(\cos t_o, \sin t_o) = 0$. That is,

$$Q(h_o,k_o) = 0$$

for some unit vector (h_o,k_o). Now if $h_o \neq 0$, this means that the number k_o/h_o is a real root of the equation

$$A + 2Bx + Cx^2 = 0.$$

$$g(t) = f(a_1 + th,\ a_2 + tk).$$

We know that $g(t) = g(0) + g'(0)\cdot t + g''(c)\cdot t^2/2!$ where c is between 0 and t. Calculating the derivatives of g gives

$$g'(t) = D_1 f(a_1+th, a_2+tk)\cdot h + D_2 f(a_1+th, a_2+tk)\cdot k,$$

$$g''(t) = (D_{1,1}f)h^2 + (D_{1,2}f)hk + (D_{2,1}f)kh + (D_{2,2}f)k^2,$$

from which formula (*) follows. Here $\underline{a}^* = \underline{a} + c\underline{v}$, where c is between 0 and t.

Step 2. In the present case, the first partials of f vanish at $\underline{a}$, so that

$$f(\underline{a}+t\underline{v}) - f(\underline{a}) \sim \{Ah^2 + 2Bhk + Ck^2\}t^2/2.$$

The only reason this is an approximation rather than an equality is that the second partials are evaluated at the unknown point $\underline{a}^*$ instead of at $\underline{a}$. This matter will be disposed of by using elementary epsilonics. Formally, we have the equation

$$(**)\quad \begin{aligned} \frac{2}{t^2}[f(\underline{a}+t\underline{v}) - f(\underline{a})] &= \{Ah^2 + 2Bhk + Ck^2\} \\ &+ [D_{1,1}f(\underline{a}^*)-A]h^2 + 2[D_{1,2}f(\underline{a}^*)-B]hk + [D_{2,2}f(\underline{a}^*)-C]k^2. \end{aligned}$$

Note that the last three terms are small if $\underline{a}^*$ is close to $\underline{a}$, because the second partials are continuous.

But this equation has a real root only if $B^2 - AC > 0$. Similarly, if $k_o \neq 0$, the number h_o/k_o is a real root of the equation

$$Ax^2 + 2Bx + C = 0;$$

again we conclude that $B^2 - AC > 0$. Thus in either case we are led to a contradiction.

Case 2. If $B^2 - AC > 0$, then we show that $Q(\underline{v})$ takes on both positive and negative values.

Proof. If $A \neq 0$, the equation $Ax^2 + Bx + C = 0$ has two distinct real roots. Thus the equation $y = Ax^2 + 2Bx + C$ represents a parabola that crosses the x-axis at two distinct points. On the other hand, if $A = 0$, then $B \neq 0$ (since $B^2 - AC > 0$); in this case the equation $y = Ax^2 + 2Bx + C$ represents a line with non-zero slope. It follows that in either case, there is a number x_o for which

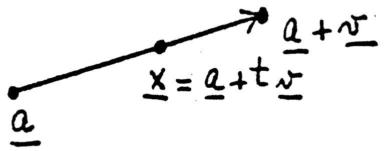
$$Ax_o^2 + 2Bx_o + C < 0,$$

and a number x_1 for which

$$Ax_1^2 + 2Bx_1 + C > 0.$$

Let (h_o,k_o) be a unit vector with $h_o/k_o = x_o$ and let (h_1,k_1) be a unit vector with $h_1/k_1 = x_1$. Then $Q(h_o,k_o) < 0$ and $Q(h_1,k_1) > 0$.

Step 4. We prove part (a) of the theorem. Assume $B^2 - AC > 0$. Let $\underline{v}_o$ be a unit vector for which $Q(\underline{v}_o) > 0$. Examining formula (**), we see that the expression


$\underline{a}+\underline{v}$
$\underline{x}=\underline{a}+t\underline{v}$
$\underline{a}$

$2[f(\underline{a}+t\underline{v}) - f(\underline{a})]/t^2$ approaches $Q(v_o)$ as $t \longrightarrow 0$. Let $\underline{x} = \underline{a} + t\underline{v}$ and let $t \longrightarrow 0$. Then $\underline{x}$ approaches $\underline{a}$ along the straight line from $\underline{a}$ to $\underline{a} + \underline{v}_o$, and the expression $f(\underline{x}) - f(\underline{a})$ approaches zero through positive values. On the other hand, if $\underline{v}_1$ is a point at which $Q(\underline{v}_1) < 0$, then the same argument shows that as $\underline{x}$ approaches $\underline{a}$ along the straight line from $\underline{a}$ to $\underline{a} + \underline{v}_1$, the expression $f(\underline{x}) - f(\underline{a})$ approaches $\underline{0}$ through negative values.

We conclude that f has a saddle point at $\underline{a}$.

Step 5. We prove parts (b) and (c) of the theorem. Examining equation (**) once again. We know that $|Q(\underline{v})| > 0$ for all unit vectors $\underline{v}$. Then $|Q(\underline{v})|$ has a positive minimum m, as $\underline{v}$ ranges over all unit vectors. (Apply the extreme-value theorem to the continuous function $|Q(\cos t, \sin t)|$, for $0 \le t \le 2\pi$.) Now choose δ small enough that each of the three square-bracketed expressions on the right side of (**) is less than $m/3$ whenever $\underline{a}^*$ is within δ of $\underline{a}$. Here we use continuity of the second-order partials. If $0 < t < \delta$, then $\underline{a}^*$ is on the line from $\underline{a}$ to $\underline{a} + \delta\underline{v}$; since $\underline{v}$ is a unit vector, then the right side of (*) has the same sign as A whenever $0 < t < \delta$. If $A > 0$, this means that $f(\underline{x}) - f(\underline{a}) > 0$ whenever

$0 < |\underline{x}-\underline{a}| < \delta$, so f has a relative minimum at $\underline{a}$. If $A < 0$, then $f(\underline{x}) - f(\underline{a}) < 0$ whenever $0 < |\underline{x}-\underline{a}| < \delta$, so f has a relative maximum at $\underline{a}$.

For examples illustrating (d), see the exercises. □

Exercises

1. Show that the function $x^4 + y^4$ has a relative minimum at the origin, while the function $x^4 - y^4$ has a saddle point there. Conclude that the second-derivative test is inconclusive if $B^2 - AC = 0$.

2. Use Taylor's theorem to derive the second derivative test for maxima and minima of a function $f(x)$ of a single variable. If $f'(a) = f''(a) = 0$ and $f'''(a) \neq 0$, what can you say about the existence of a relative maximum or minimum at f at a?

3. Suppose $f(x)$ has continuous derivatives of orders $1,\ldots,n+1$ near $x = a$. Suppose

$$f'(a) = f''(a) = \cdots = f^{(n)}(a) = 0$$

and $f^{(n+1)}(a) \neq 0$. What can you say about the existence of a relative maximum or minimum of f at a? Prove your answer correct.

4. (a) Suppose $f(x_1,x_2)$ has continuous third-order partials near $\underline{a}$. Derive a third-order version of formula (*) of the preceding theorem.

(b) Derive the general version of Taylor's theorem for functions of two variables.

[The following "operator notation" is convenient

$$(hD_1+kD_2)f\big|_{\underline{x}=\underline{a}} = hD_1f(\underline{a}) + kD_2f(\underline{a}),$$

$$(hD_1+kD_2)^2f\big|_{\underline{x}=\underline{a}} = h^2D_1D_1f(\underline{a}) + 2hkD_1D_2f(\underline{a}) + h^2D_2D_2f(\underline{a}),$$

and similarly for $(hD_1+kD_2)^n$.]

The extreme-value theorem and the small-span theorem.

The proofs of the extreme-value theorem and small-span theorem for rectangles given in Apostol are sufficiently condensed to cause some students difficulty. Here are the details. We shall prove the theorems only for R^2, but the proofs go through without difficulty in R^n.

A rectangle Q in R^2 is the Cartesian product of two closed intervals $[a,b]$ and $[c,d]$;

$$Q = [a,b] \times [c,d] = \{(x,y) \mid a \leq x \leq b \text{ and } c \leq y \leq d\}.$$

The intervals $[a,b]$ and $[c,d]$ are called the component intervals of Q. If

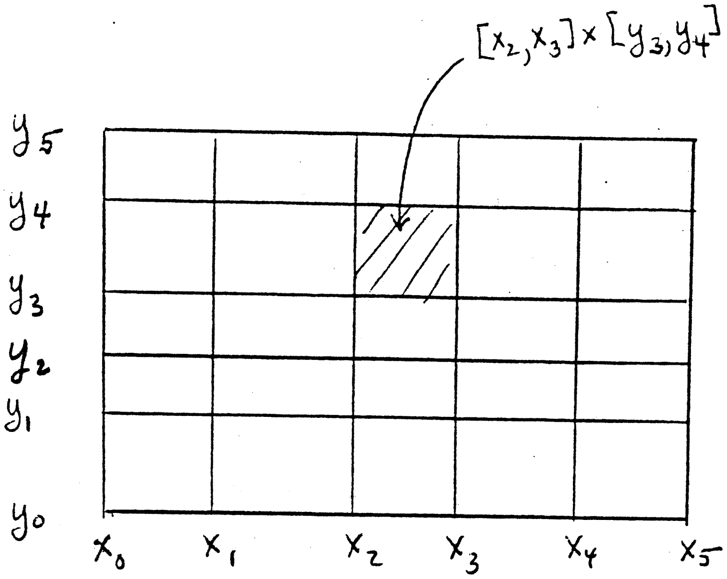
$$P_1 = \{x_0, x_1, \ldots, x_n\}$$

is a partition of $[a,b]$, and if

$$P_2 = \{y_0, y_1, \ldots, y_m\}$$

is a partition of $[c,d]$, then the cartesian product $P_1 \times P_2$ is said to be a partition of Q. Since P_1 partitions $[a,b]$ into n subintervals and P_2 partitions $[c,d]$ into m subintervals, the partition $P = P_1 \times P_2$ partitions Q into mn subrectangles, namely the rectangles

$$[x_{i-1}, x_i] \times [y_{j-1}, y_j].$$



Theorem (small-span theorem). Let f be a scalar function that is continuous on the rectangle

$$Q = [a,b] \times [c,d]$$

in R^2. Then, given $\epsilon > 0$, there is a partition of Q such that f is bounded on every subrectangle of the partition and such that the span of f in every subrectangle of the partition is less than ϵ.

Proof. For purposes of this proof, let us use the following terminology: If Q_0 is any rectangle contained in Q, let us say that a partition of Q_0 is "ϵ-nice" if f is bounded on every subrectangle R of the partition and if the span of f in every subrectangle of the partition is less than ϵ. We recall that the span of f in the set S is defined by the equation

$$\text{span}_S\, f = \sup\{f(x) \mid x \in S\} - \inf\{f(x) \mid x \in S\}.$$

Recall also that if S_1 is a subset of S, then

$$\text{span}_{S_1} f \leq \text{span}_S f.$$

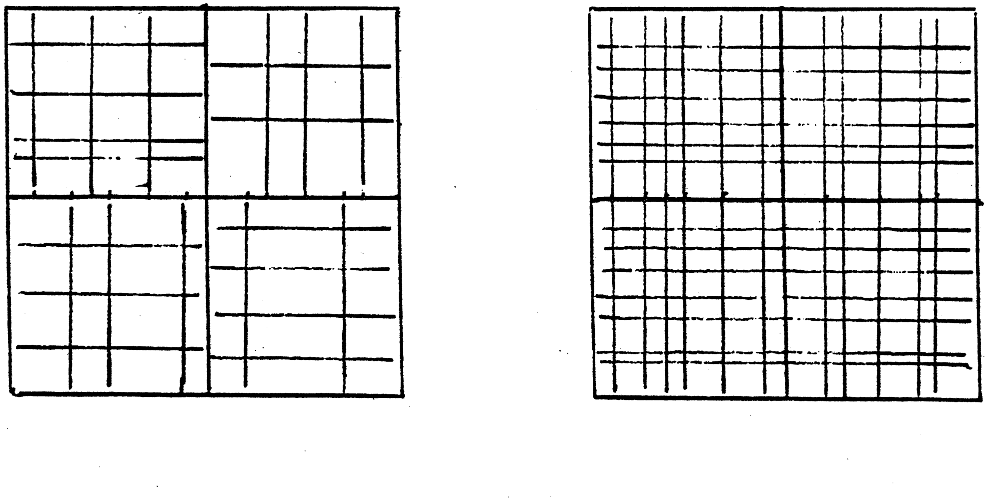
To begin, we note the following elementary fact: Suppose

$$Q_0 = [a_0,b_0] \times [c_0,d_0]$$

is any rectangle contained in Q. Let us bisect the first component interval $[a_0,b_0]$ of Q_0 into two subintervals $I_1 = [a_0,p]$ and $I_2 = [p,b_0]$, where p is the midpoint of $[a_0,b_0]$. Similarly, let us bisect $[c_0,d_0]$ into two subintervals J_1 and J_2. Then Q_0 is written as the union of the four rectangles

$$I_1 \times J_1 \quad \text{and} \quad I_2 \times J_1 \quad \text{and} \quad I_1 \times J_2 \quad \text{and} \quad I_2 \times J_2.$$

Now if each of these rectangles has a partition that is ϵ_0-nice, then we can put these partitions together to get a partition of Q_0 that is ϵ_0-nice. The figure indicates the proof; each of the subrectangles of the new partition is contained in a subrectangle of one of the old partitions.



Now we prove the theorem. We suppose the theorem is false and derive a contradiction. That is, we assume that for some $\epsilon_0 > 0$, the rectangle Q has no partition that is ϵ_0-nice.

Let us bisect each of the component intervals of Q, writing Q as the union of four rectangles. Not all of these smaller rectangles have partitions that are ϵ_0-nice, for if they did, then Q would have such a partition. Let Q_1 be one of these smaller rectangles, chosen so that Q_1 does not have a partition that is ϵ_0-nice.

Now we repeat the process. Bisect each component interval of Q_1 into four smaller rectangles. At least one of these smaller rectangles has no partition that is ϵ_0-nice; let Q_2 denote one such.

Continuing similarly, we obtain a sequence of rectangles

$$Q,\ Q_1,\ Q_2, \ldots$$

none of which have partitions that are ϵ_0-nice. Consider the left-hand end points of the first component interval of each of these rectangles. Let s be their least upper bound. Similarly, consider the left-hand end points of the second component interval of each of these rectangles, and let t be their least upper bound. Then the point (s,t) belongs to all of the rectangles $Q, Q_1, Q_2, \ldots$.

Now we use the fact that f is continuous at the point (s,t). We choose a ball of radius r centered at (s,t) such that the span of f in this ball is less than ϵ_0. Because the rectangles Q_m become arbitrarily small as m increases, and because they all contain the point (s,t), we can choose m large enough that Q_m lies within this ball.

Now we have a contradiction. Since Q_m is contained in the ball of radius r centered at (s,t), the span of f in Q_m is less than ϵ_0. But this implies that there *is* a partition of Q_m that is ϵ_0-nice, namely the trivial partition! Thus we have reached a contradiction. □

Corollary. *Let* f *be a scalar function that is continuous on the rectangle* Q. *Then* f *is bounded on* Q.

Proof. Set $\epsilon = 1$, and choose a partition of Q that is ϵ-nice. This partition divides Q into a certain number of subrectangles, say $R_1,\ldots,R_{mn}$. Now f is bounded on each of these subrectangles, by hypothesis; say

$$|f(\underline{x})| \leq M_i \quad \text{for} \quad \underline{x} \in R_i.$$

Then if $M = \max\{M_1,\ldots,M_{mn}\}$, we have

$$|f(\underline{x})| \leq M$$

for all $\underline{x} \in Q$. □

Theorem (extreme-value theorem). *Let* f *be a scalar function that is continuous on the rectangle* Q. *Then there are points* x_0 *and* x_1 *of* Q *such that*

$$f(\underline{x}_0) \leq f(\underline{x}) \leq f(\underline{x}_1)$$

for all $x \in Q$.

Proof. We know f is bounded on Q; let

$$M = \sup\{f(\underline{x}) \mid \underline{x} \in Q\}.$$

We wish to show there is a point $\underline{x}_1$ of Q such that $f(\underline{x}_1) = M$.

Suppose there is no such a point. Then the function $M - f(\underline{x})$ is continuous and positive on Q, so that the function

$$g(\underline{x}) = \frac{1}{M - f(\underline{x})}$$

is also continuous and positive on Q. By the preceding corollary g is bounded on Q; let C be a positive constant such that $g(\underline{x}) \leq C$ for $\underline{x} \in Q$. Then

$$\frac{1}{M - f(\underline{x})} \leq C, \text{ or}$$

$$f(\underline{x}) \leq M - (1/C)$$

for all $\underline{x}$ in Q. Then $M - (1/C)$ is an upper bound for the set of values of $f(\underline{x})$ for $\underline{x}$ in Q, contradicting the fact that M is the least upper bound for this set.

A similar argument proves the existence of a point $\underline{x}_0$ of Q such that

$$f(\underline{x}_0) = \inf\{f(\underline{x}) \mid \underline{x} \in Q\}. \quad \square$$

Exercises on line integrals

1. Find the centroid of a homogeneous wire in shape of the parabolic arc

$$y = x^2 \quad \text{for} \quad -1 \leq x \leq 1.$$

[Use a table of integrals if you wish.]

2. Let

$$\underline{f}(x,y) = \frac{-y\underline{i}+x\underline{j}}{x^2+y^2}\,,$$

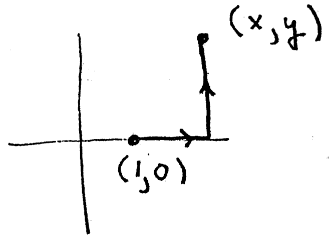
on the set S consisting of all $(x,y) \neq 0$.

(a) Show that $D_2f_1 = D_1f_2$ on S.

(b) Compute the line integral $\int_C \underline{f} \cdot d\underline{\alpha}$ from $(a,0)$ to $(-a,0)$ when C is the upper half of the circle $x^2 + y^2 = a^2$. Compute it when C is the lower half of the same circle.

3. Let $\underline{f}$ be as in problem 2. Let U be the set of all (x,y) with $x > 0$. Find a potential function for $\underline{f}$ that is defined in U. Hint: Set

$$\phi(x,y) = \int_C \underline{f}\cdot d\underline{\alpha} \quad \text{where } C \text{ is the curve}$$



4. Let $\underline{f}$ be a continuous vector field defined in the open, connected subset S of R^n. Suppose that $\underline{f} = \vec{\nabla}\phi_1$ and $\underline{f} = \vec{\nabla}\phi_2$ in S. Show that $\phi_1 - \phi_2$ is a constant function. [Hint: Apply Thoerem 10.3.]

18.024 Multivariable Calculus with Theory
Spring 2011

Notes on double integrals.

(Read 11.1-11.5 of Apostol.)

Just as for the case of a single integral, we have the following condition for the existence of a double integral:

Theorem 1 (Riemann condition). Suppose f is defined on $Q = [a,b] \times [c,d]$. Then f is integrable on Q if and only if given any $\varepsilon > 0$, there are step functions s and t with $s \leqslant f \leqslant t$ on Q, such that

$$\iint_Q t - \iint_Q s < \varepsilon.$$

Let A be a number. If these step functions s and t satisfy the further condition that

$$\iint_Q s \leqslant A \leqslant \iint_Q t,$$

then $A = \iint_Q f.$

The proof is almost identical with the corresponding proof for the single integral.

Using this condition, one can readily prove the three basic properties--linearity, additivity, and comparison--for the integral $\iint_Q f$. We state them as follows:

Theorem 2. (a) Suppose f and g are integrable on Q. Then so is $cf(x) + dg(x)$; furthermore,

$$\iint_Q (cf + dg) = c \iint_Q f + d \iint_Q g.$$

(b) *Let* Q *be subdivided into two rectangles* Q_1 *and* Q_2. *Then* f *is integrable over* Q *if and only if it is integrable over both* Q_1 *and* Q_2; *furthermore,*

$$\iint_Q f = \iint_{Q_1} f + \iint_{Q_2} f.$$

(c) *If* $f \leqslant g$ *on* Q, *and if* f *and* g *are integrable over* Q, *then*

$$\iint_Q f \leqslant \iint_Q g.$$

To prove this theorem, one first verifies these results for step functions (see 11.3), and then uses the Riemann condition to prove them for general integrable functions. The proofs are very similar to those given for the single integral.

We give one of the proofs as an illustration. For example, consider the formula

$$\iint_Q (f + g) = \iint_Q f + \iint_Q g,$$

where f and g are integrable. We choose step functions s_1, s_2, t_1, t_2 such that

$$s_1 \leqslant f \leqslant t_1 \quad \text{and} \quad s_2 \leqslant g \leqslant t_2$$

on Q, and such that

$$\iint_Q (t_1 - s_1) < \varepsilon/2 \quad \text{and} \quad \iint_Q (t_2 - s_2) < \varepsilon/2.$$

We then find a single partition of Q relative to which all of s_1, s_2, t_1, t_2 are step functions; then $s_1 + s_2$ and $t_1 + t_2$ are also step functions relative to this partition. Furthermore, one adds the earlier inequalities to obtain

$$s_1 + s_2 \leqslant f + g \leqslant t_1 + t_2.$$

Finally, we compute

$$\iint_Q (t_1+t_2) - (s_1+s_2) = \iint_Q (t_1-s_1) + \iint_Q (t_2-s_2) < \varepsilon;$$

this computation uses the fact that linearity has already been proved for step functions. Thus $\iint_Q (f + g)$ exists. To calculate this integral, we note that

$$\iint_Q s_1 \leqslant \iint_Q f \leqslant \iint_Q t_1,$$

$$\iint_Q s_2 \leqslant \iint_Q g \leqslant \iint_Q t_2,$$

by definition. Then

$$\iint_Q (s_1+s_2) \leqslant \iint_Q f + \iint_Q g \leqslant \iint_Q (t_1+t_2);$$

here again we use the linearity of the double integral for step functions. It follows from the second half of the Riemann condition that $\iint_Q (f + g)$ must equal the number $A = \iint_Q f + \iint_Q g$.

Up to this point, the development of the double integral has been remarkably similar to the development of the single integral. Now things begin to change. We have the following basic questions to answer:

(1) Under what conditions does $\iint_Q f$ exist?

(2) If $\iint_Q f$ exists, how can one evaluate it?

(3) Is there a version of the substitution rule for double integrals?

(4) What are the applications of the double integral?

We shall deal with questions (1), (2), and (4) now, postponing question (3) until the next unit.

Let us tackle question (2) first. How can one evaluate the integral if one knows it exists? The answer is that such integrals can almost always be evaluated by repeated one-dimensional integration. More precisely, one has the following theorem:

Theorem 3 (Fubini theorem). Let f be defined and bounded on a rectangle $Q = [a,b] \times [c,d]$, and assume that f is integrable on Q. For each fixed y in $[c,d]$, assume that the one-dimensional integral

$$A(y) = \int_a^b f(x,y)\,dx$$

exists. Then the integral $\int_c^d A(y)\,dy$ exists, and furthermore,

$$\int_c^d \left[\int_a^b f(x,y)\,dx\right] dy = \iint_Q f(x,y)\,dx\,dy.$$

Proof. We need to show that $\int_c^d A(y)dy$ exists and equals the double integral $\iint_Q f$.

Choose step functions $s(x,y)$ and $t(x,y)$, defined on Q, such that $s(x,y) \leqslant f(x,y) \leqslant t(x,y)$, and

$$\iint_Q t - \iint_Q s < \varepsilon.$$

This we can do because $\iint_Q f$ exists. For convenience, choose s and t so they are constant on the partition lines. (This does not affect their double integrals.) Then the one-dimensional integral

$$\int_a^b s(x,y)dx$$

exists. [For, given fixed y in $[c,d]$, the function $s(x,y)$ is either constant (if y is a partition point) or a step function of x; hence it is integrable.] Now I claim that the function $S(y) = \int_a^b s(x,y)dx$ is a step function on the interval $c \leq y \leq d$. For there are partitions $x_o,\ldots,x_m$ and $y_o,\ldots,y_n$ of $[a,b]$ and $[c,d]$, respectively, such that $s(x,y)$ is constant on each open rectangle $(x_{i-1},x_i) \times (y_{j-1},y_j)$. Let $\overline{y}$ and $\overline{\overline{y}}$ be any two points of the interval (y_{j-1},y_j). Then $s(x,\overline{y}) = s(x,\overline{\overline{y}})$ holds for all x. (This is immediate if x is in (x_{i-1},x_i); if x is a partition point, it follows from the fact that s is constant on the partition lines.) Therefore

$$\int_a^b s(x,\bar{y})\,dx = \int_a^b s(x,\bar{\bar{y}})\,dx.$$

Hence $S(y)$ is constant on (y_{j-1},y_j), so it is a step function.

A similar argument shows that the function

$$T(y) = \int_a^b t(x,y)\,dx$$

is a step function for $c \leq y \leq d$.

Now since $s \leq f \leq t$ for all (x,y), we have

$$\int_a^b s(x,y)\,dx \leq \int_a^b f(x,y)\,dx \leq \int_a^b t(x,y)\,dx,$$

by the comparison theorem. (The middle integral exists by hypothesis.) That is, for all y in $[c,d]$,

$$S(y) \leq A(y) \leq T(y).$$

Thus S and T are step functions lying beneath and above A, respectively. Furthermore

$$\iint_Q s = \int_c^d S(y)\,dy \qquad \text{and} \qquad \iint_Q t = \int_c^d T(y)\,dy,$$

(see p. 356), so that

$$\int_c^d T(y)\,dy - \int_c^d S(y)\,dy < \varepsilon.$$

It follows that $\int_c^d A(y)\,dy$ exists, by the Riemann condition.

Now that we know $A(y)$ is integrable, we can conclude from an earlier inequality that

$$\int_c^d S(y)\,dy \leq \int_c^d A(y)\,dy \leq \int_c^d T(y)\,dy;$$

that is,

$$\iint_Q s \leq \int_c^d A(y)\,dy \leq \iint_Q t.$$

But it is also true that

$$\iint_Q s \leq \iint_Q f \leq \iint_Q t,$$

by definition. Since the integrals of s and t are less than ε apart, we conclude that $\int_c^d A(y)\,dy$ and $\iint_Q f$ are within ε of each other. Because ε is arbitrary, they must be equal. □

With this theorem at hand, one can proceed to calculate some specific double integrals. Several examples are worked out in 11.7 and 11.8 of Apostol.

Now let us turn to the first of our basic questions, the one concerning the existence of the double integral. We readily prove the following:

Theorem 4. The integral $\iint_Q f$ exists if f is continuous on the rectangle Q.

Proof. All one needs is the small-span theorem of p. C.29.

Given ε', choose a partition of Q such that the span of f on each subrectangle of the partition is less than ε'. If Q_{ij} is a subrectangle, let

$$s_{ij} = \min f(x) \quad \text{on} \quad Q_{ij}; \qquad t_{ij} = \max f(x) \quad \text{on} \quad Q_{ij}.$$

Then $t_{ij} - s_{ij} < \varepsilon'$. Use the numbers s_{ij} and t_{ij} to obtain step functions s and t with $s \leqslant f \leqslant t$ on Q. One then has

$$\iint_Q (t - s) < \varepsilon'(d - c)(b - a).$$

This number equals ε if we begin the proof by setting $\varepsilon' = \varepsilon/(d-c)(b-a)$. □

In practice, this existence theorem is not nearly strong enough for our purposes, either theoretical or practical. We shall derive a theorem that is much stronger and more useful.

First, we need some definitions:

Definition. If $Q = [a,b] \times [c,d]$ is a rectangle, we define the **area** of Q by the equation

$$\text{area } Q = \iint_Q 1.$$

Of course, since 1 is a step function, we can calculate this integral directly as the product $(d-c)(b-a)$.

Additivity of $\iint$ implies that if we subdivide Q into two rectangles Q_1 and Q_2, then

$$\text{area } Q = \text{area } Q_1 + \text{area } Q_2.$$

Applying this formula repeatedly, we see that if one has a partition of Q, then

$$\text{area } Q = \textstyle\sum_{i,j} \text{area } Q_{ij},$$

where the summation extends over all subrectangles of the partition.

It now follows that if A and Q are rectangles and $A \subset Q$, then area $A \leq$ area Q.

Definition. Let D be a subset of the plane. Then D is said to have content zero if for every $\varepsilon > 0$, there is a finite set of rectangles whose union contains D and the sum of whose areas does not exceed ε.

Examples.

(1) A finite set has content zero.

(2) A horizontal line segment has content zero.

(3) A vertical line segment has content zero.

(4) A subset of a set of content zero has content zero.

(5) A finite union of sets of content zero has content zero.

(6) The graph of a continuous function

$$y = \phi(x); \quad a \leq x \leq b$$

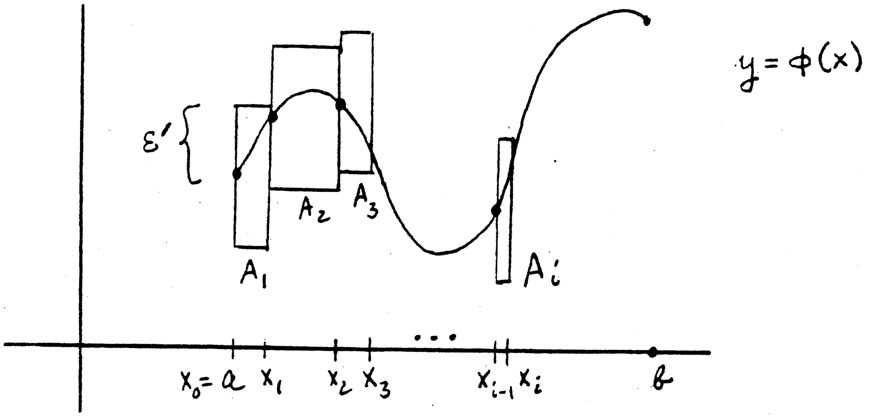
has content zero.

(7) The graph of a continuous function

$$x = \psi(y); \quad c \leq y \leq d$$

has content zero.

Most of these statements are trivial to prove; only the last two require some care. Let us prove (6). Let $\varepsilon' > 0$. Given the continuous function ϕ, let us use the small-span theorem for functions of a single variable to choose a



partition $a = x_0 < x_1 < \ldots < x_n = b$ of $[a,b]$ such that the span of ϕ on each subinterval is less than ε'. Consider the rectangles

$$A_i = [x_{i-1}, x_i] \times [\phi(x_{i-1}) - \varepsilon', \phi(x_{i-1}) + \varepsilon']$$

for $i = 1,\ldots,n$. They cover the graph of ϕ, because $|\phi(x) - \phi(x_{i-1})| < \varepsilon'$ whenever x is in the interval $[x_{i-1}, x_i]$. The total area of the rectangles A_i equals

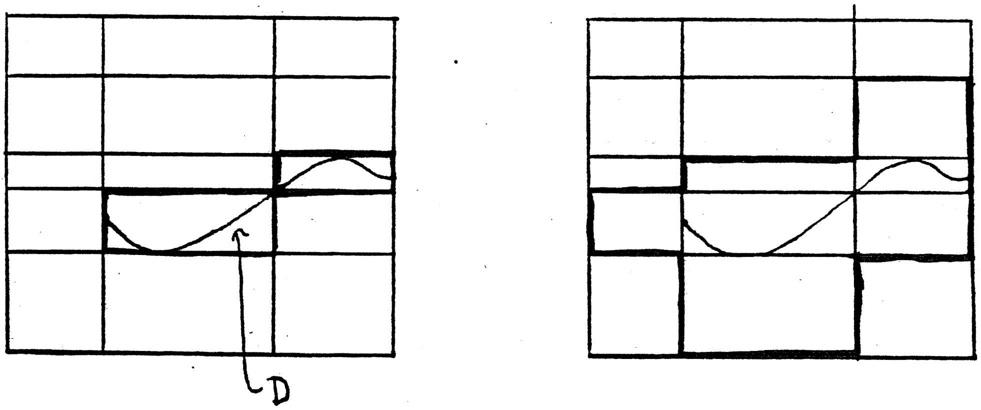
$$\sum_{i=1}^{n} (x_i - x_{i-1}) 2\varepsilon' = 2\varepsilon'(b - a).$$

This number equals ε if we begin the proof by setting $\varepsilon' = \varepsilon/2(b-a)$.

We now prove an elementary fact about sets of content zero:

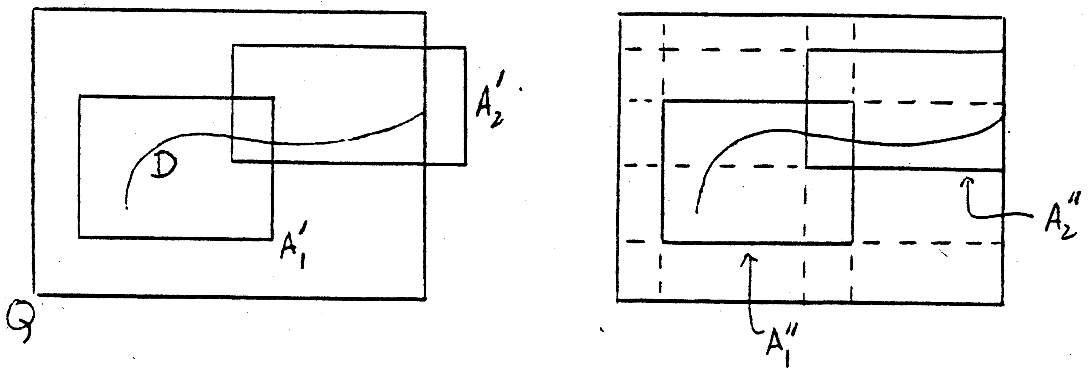
Lemma 5. Let Q *be a rectangle. Let* D *be a subset of* Q *that has content zero. Given* $\varepsilon > 0$, *there is a partition of* Q *such that those subrectangles of the partition that contain points of* D *have total area less than* ε.

Note that this lemma does not state merely that D is *contained in* the union of finitely many subrectangles of the partition having total area less than ε, but that the sum of the areas of *all* the subrectangles that contain points of D is less than ε. The following figure illustrates the distinction; D is contained in the union of two subrectangles, but there are seven subrectangles that contain points of D.



Proof. First, choose finitely many rectangles $A_1,\ldots,A_n$ of total area less than $\varepsilon/2$ whose union contains D. "Expand" each one slightly. That is, for each i, choose a rectangle A_i' whose interior contains A_i, such that the area of A_i' is no more than twice that of A_i. Then the union of the sets $\text{Int } A_i'$ contains D, and the rectangles A_i' have total area less than ε. Of course, the rectangle A_i' may extend outside Q, so let A_i'' denote the rectangle that is the intersection of A_i' and Q. Then the rectangles A_i'' also have total area less than ε.

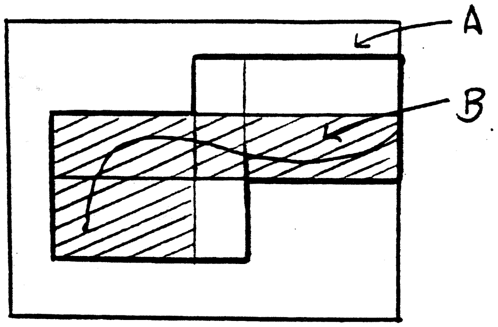
Now use the end points of the component intervals of the rectangles A_i'' to define a partition P of the rectangle Q. See the figure.



We show that this is our desired partition.

Note that by construction, the rectangle A_k'' is partitioned by P, so that it is a union of subrectangles Q_{ij} of P.

Now if a subrectangle Q_{ij} contains a point of D, then it contains a point of $\text{Int } A_k'$ for some k, so that it actually lies in A_k' and hence in A_k''. Suppose we let B denote the union of all the subrectangles Q_{ij} that contain points of D; and let A be the union of the rectangles $A_1'',\ldots,A_n''$. Then $B \subset A$.



It follows that

$$\sum_{Q_{ij}\subset B} \text{area } Q_{ij} \leq \sum_{Q_{ij}\subset A} \text{area } Q_{ij}.$$

Now on the other hand, by additivity of area for rectangles,

$$\sum_{Q_{ij}\subset A_k''} Q_{ij} = \text{area } A_k''.$$

It follows that

$$\sum_{Q_{ij}\subset A} Q_{ij} \leq \sum_{k=1}^{n} \text{area } A_k''.$$

This last inequality is in general strict, because some subrectangles Q_{ij} belong to more than one rectangle A_k'', so their areas are counted more than once in the sum on the right side of the inequality.

It follows that

$$\sum_{Q_{ij}\subset B} \text{area } Q_{ij} < \varepsilon,$$

as desired. □

Now we prove our basic theorem on existence of the double integral $\iint_Q f$.

Theorem 6. *If* f *is bounded on* Q, *and is continuous on* Q *except on a set of content zero, then* $\iint_Q f$ *exists*.

Proof. Step 1. We prove a preliminary result:

Suppose that given $\epsilon > 0$, there exist functions g and h that are integrable over Q, such that

$$g(x) \leq f(x) \leq h(x) \qquad \text{for } x \text{ in } Q$$

and

$$\iint_Q h - \iint_Q g < \epsilon.$$

Then f is integrable over Q.

We prove this result as follows: Because h and g are integrable, we can find step functions s_1, s_2, t_1, t_2 such that

$$s_1 \leq g \leq t_1 \quad \text{and} \quad s_2 \leq h \leq t_2,$$

and such that

$$\iint_Q t_1 - \iint_Q s_1 < \epsilon \quad \text{and} \quad \iint_Q t_2 - \iint_Q s_2 < \epsilon\,.$$

Consider the step functions s_1 and t_2. We know that

$$s_1 \leq g \leq f \leq h \leq t_2$$

so s_1 is beneath f, and t_2 is above f. Furthermore, because the integral of g is between the integrals of s_1 and of t_1, we know that

$$\iint_Q g - \iint_Q s_1 < \epsilon.$$

Similarly,

$$\iint_Q t_2 - \iint_Q h < \epsilon.$$

If we add these inequalities and the inequality

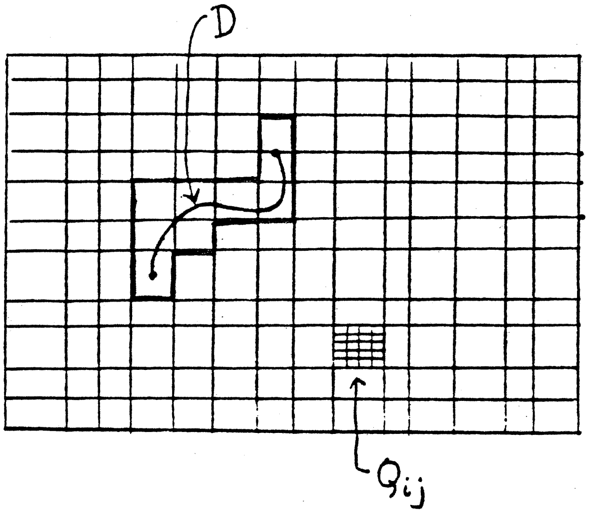
$$\iint_Q h - \iint_Q g < \epsilon,$$

we have

$$\iint_Q t_2 - \iint_Q s_1 < 3\epsilon.$$

Since ϵ is arbitrary, the Riemann condition is satisfied, so f is integrable over Q.

Step 2. Now we prove the theorem. Let D be a set of zero content containing the discontinuities of f. Choose M so that $|f(x)| \leq M$ for x in Q; then given $\epsilon > 0$, set $\epsilon' = \epsilon/2M$. Choose a partition P of Q such that those subrectangles that contain points of D have total area less than ϵ'. (Here we use the preceding lemma.)



Now we define functions g and h such that $g \leq f \leq h$ on Q. If Q_{ij} is one of the subrectangles that does not contain a point of D, set

$$g(x) = f(x) = h(x)$$

for $x \in Q_{ij}$. Do this for each such subrectangle. Then for any other x in Q, set

$$g(x) = -M \quad \text{and} \quad h(x) = M.$$

Then $g \leq f \leq h$ on Q.

Now g is integrable over each subrectangle Q_{ij} that does not contain a point of D, since it equals the continuous function f there. And g is integrable over each subrectangle Q_{ij} that does contain a point of D, because it is a step function on such a subrectangle. (It is constant on the interior of Q_{ij}.) The additivity property of the integral now implies that g is integrable over Q.

Similarly, h is integrable over Q. Using additivity, we compute the integral

$$\iint_Q (h-g) = \sum \iint_{Q_{ij}} (h-g) = 2M \sum (\text{area } Q_{ij} \text{ that contain points of D})$$

$$< 2M\epsilon' = \epsilon.$$

Thus the conditions of Step 1 hold, and f is integrable over Q. □

Theorem 7. Suppose f *is bounded on* Q, *and* f *equals* 0 *except on a set* D *of content zero. Then* $\iint_Q f$ *exists and equals zero*

Proof. We apply Step 2 of the preceding proof to the function f.

Choose M so that $|f(x)| \leq M$ for x in Q; given $\epsilon > 0$, set $\epsilon' = \epsilon/2M$. Choose a partition P such that those subrectangles that contain points of D have total area less than ϵ'.

Define functions g and h as follows: If Q_{ij} is one of the subrectangles that does not contain a point of D, set $g(x) = f(x) = 0$ and $h(x) = f(x) = 0$ on Q_{ij}. Do this for each such subrectangle. For any other x in Q, set

$$g(x) = -M \quad \text{and} \quad h(x) = M.$$

Then $g \leq f \leq h$ on Q.

Now g and h are step functions on Q, because they are constant on the interior of each subrectangle Q_{ij}. We compute

$$\iint_Q h = M\,(\textstyle\sum \text{(area } Q_{ij} \text{ that contain points of D))}$$
$$< 2M\epsilon' = \epsilon/2.$$

Similarly,

$$\iint_Q g > -M\epsilon' = -\epsilon/2.$$

Hence $\iint_Q (h-g) < \epsilon$, so that f is integrable over Q. Furthermore,

$$-\epsilon/2 < \iint_Q g \leq \iint_Q f \leq \iint_Q h < \epsilon/2.$$

Since ϵ is arbitrary, $\iint_Q f = 0$. □

Corollary 8. If $\iint_Q f$ *exists, and if* g *is a bounded function that equals* f *except on a set of content zero, then* $\iint_Q g$ *exists and equals* $\iint_Q f$.

Proof. We write $g = f + (g-f)$. Now f is integrable by hypothesis, and $g - f$ is integrable by the preceding corollary. Then g is integrable and

$$\iint_Q g = \iint_Q f + \iint_Q (g-f) = \iint_Q f. \square$$

Double integrals extended over more general regions.

(Read section 11.12 of Apostol.) In this section, Apostol defines $\iint_S f$ for a function f defined on a bounded set S, but then he quickly restricts himself to the special case where S is a region of Types I or II. We discuss here the general case.

First, we prove the following basic existence theorem:

Theorem 9. *Let* S *be a bounded set in the plane. If* Bd S *has content zero, and if* f *is* bounded on S and *continuous at each point of* Int S, *then* $\iint_S f$ *exists*.

Proof. Let Q be a rectangle containing S. As usual, let $\tilde{f}$ equal f on S, and let $\tilde{f}$ equal 0 outside S. Then $\tilde{f}$ is continuous at each point x_0 of the interior of S (because it equals f in an open ball about x_0, and f is continuous at x_0). The function $\tilde{f}$ is also continuous at each point x_1 of the exterior of S, because it equals zero on an open ball about x_1. The only points where $\tilde{f}$ can fail to be continuous are points of the boundary of S, and this set, by assumption, has content zero. Hence $\iint_Q \tilde{f}$ exists. $\square$

Note: Adjoining or deleting boundary points of S changes the value of f only on a set of content zero, so that value of $\iint_S f$ remains unchanged. Thus $\iint_S f = \iint_{\text{Int } S} f$, for instance.

Let us remark on a more general existence theorem than that stated in Theorem 9. If S is a bounded set, and if $\text{Bd } S$ has content zero, and if f is continuous on $\text{Int } S$ except on a set D of content zero, then $\iint_S f$ exists. For in this case the discontinuities of the extended function $\tilde{f}$ lie in the union of the sets $\text{Bd } S$ and D, and this set has content zero because both $\text{Bd } S$ and D do.

There are more general existence theorems even than this, but we shall not consider them.

Now we note that the basic properties of the double integral hold also for this extended integral:

Theorem 10. Let S be a bounded set in the plane. One has the following properties:

(a) Linearity.

$$\iint_S cf + dg = c \iint_S f + d \iint_S g;$$

the left side exists if the right side does.

(b) Comparison. If $f \leq g$ on the set S, then

$$\iint_S f \leq \iint_S g,$$

provided both integrals exist.

(c) Additivity. Let $S = S_1 \cup S_2$. If $S_1 \cap S_2$ has content zero, then

$$\iint_S f = \iint_{S_1} f + \iint_{S_2} f,$$

provided the right side exists.

Proof. (a) Given f, g defined on S, let $\tilde{f}, \tilde{g}$ equal f, g, respectively, on S and equal 0 otherwise. Then $c\tilde{f} + d\tilde{g}$ equals $cf + dg$ on S and 0 otherwise. Let Q be a rectangle containing S. We know that

$$\iint_Q c\tilde{f} + d\tilde{g} = c \iint_Q \tilde{f} + d \iint_Q \tilde{g};$$

from this linearity follows.

(b) Similarly, if $f \leq g$, then $\tilde{f} \leq \tilde{g}$, from which we conclude that

$$\iint_S f = \iint_Q \tilde{f} \leq \iint_Q \tilde{g} = \iint_S g.$$

(c) Let Q be a rectangle containing S. Let f_1 equal f on S_1, and equal 0 elsewhere. Let f_2 equal f on S_2, and equal 0 elsewhere. Let f_3 equal f on S, and equal 0 elsewhere. Consider the function

$$f_4 = f_1 + f_2 - f_3;$$

it equals f on the set $S_1 \cap S_2$, and equals zero elsewhere. Because $S_1 \cap S_2$ has content zero, $\iint_Q f_4$ exists and equals zero. Now

$$f_3 = f_1 + f_2 - f_4;$$

linearity implies that

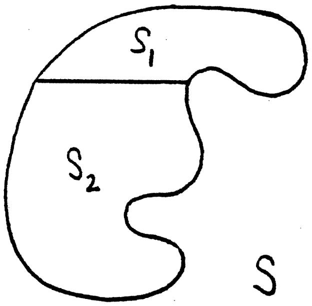
$$\iint_Q f_3 = \iint_Q f_1 + \iint_Q f_2 - \iint_Q f_4,$$

or

$$\iint_S f = \iint_{S_1} f + \iint_{S_2} f. \quad \square$$

How can one evaluate $\iint_S f$ when S is a general region? The computation is easy when S is a region of type I or II and f is continuous on the interior of S; one evaluates $\iint_S f$ by iterated integration. This result is proved on p. 367 of Apostol.

Using additivity, one can also evaluate $\iint_S f$ for many other regions as well. For example, to integrate a continuous function f over the region S pictured, one can



break it up as indicated into two regions S_1 and S_2 that intersect in a set of content zero. Since S_1 is of type I and S_2 is of type II, we can compute the integrals $\iint_{S_1} f$ and $\iint_{S_2} f$ by iterated integration. We add the results to obtain $\iint_S f$.

Area.

We can now construct a rigorous theory of area. We already have defined the area of the rectangle $Q = [a,b] \times [c,d]$ by the equation

$$\text{area } Q = \iint_Q 1.$$

We use this same equation for the general definition.

Definition. Let S be a bounded set in the plane. We say that S is *Jordan-measurable* if $\iint_S 1$ exists; in this case, we define

$$\text{area } S = \iint_S 1.$$

Note that if $\text{Bd } S$ has content zero, then S is Jordan-measurable, by by Theorem 9. The converse also holds; the proof is left as an exercise.

The area function has the following properties:

Theorem 11. *Let* S *and* T *be Jordan-measurable sets in the plane*.

(1) (*Monotonicity*). *If* $S \subset T$, *then* area $S \leq$ area T.

(2) (*Positivity*). Area $S \geq 0$, *and equality holds if and only if* S *has content zero*.

(3) (Additivity) If $S \cap T$ is a set of content zero, then $S \cup T$ is Jordan-measurable and

$$\text{area}(S \cup T) = \text{area } S + \text{area } T.$$

(4) $\text{Area } S = \text{Area}(\text{Int } S) = \text{Area}(S \cup \text{Bd } S)$.

Proof. Let Q be a rectangle containing S and T. Let

$$1_S(x) = 1 \quad \text{for} \quad x \in S$$

$$= 0 \quad \text{for} \quad x \notin S.$$

Define 1_T similarly.

(1) If S is contained in T, then $1_S(x) \leq 1_T(x)$. Then by the comparison theorem,

$$\text{area } S = \iint_S 1 = \iint_Q 1_S \leq \iint_Q 1_T = \iint_T 1 = \text{area } T.$$

(2) Since $0 < 1$, we have by the comparison theorem,

$$0 = \iint_S 0 \leq \iint_S 1 = \text{area } S,$$

for all S. If S has content zero, then $\iint_S 1 = \iint_Q 1_S = 0$, by Corollary 7.

Conversely, suppose $\iint_S 1 = 0$. Then $\iint_Q 1_S = 0$. Given $\varepsilon > 0$, there must be a step function $t \geqslant 1_S$ defined on Q such that $\iint_Q t < \varepsilon$. Let P be a partition relative to which t is a step function. Now if a subrectangle Q_{ij} of this partition contains a point of S in its interior, then the value of t on this subrectangle must be at least 1. Thus these subrectangles have total area less than ε. Now S is contained in the union of these subrectangles (of total area less than ε) and the partition lines. Thus S has content zero.

(3) Because $\iint_S 1$ and $\iint_T 1$ exist and $S \cap T$ has content zero, it follows from additivity that $\iint_{S \cup T} 1$ exists and equals $\iint_S 1 + \iint_T 1$.

(4) Since the part of S not in $\text{Int } S$ lies in $\text{Bd } S$, it has content zero. Then additivity implies that

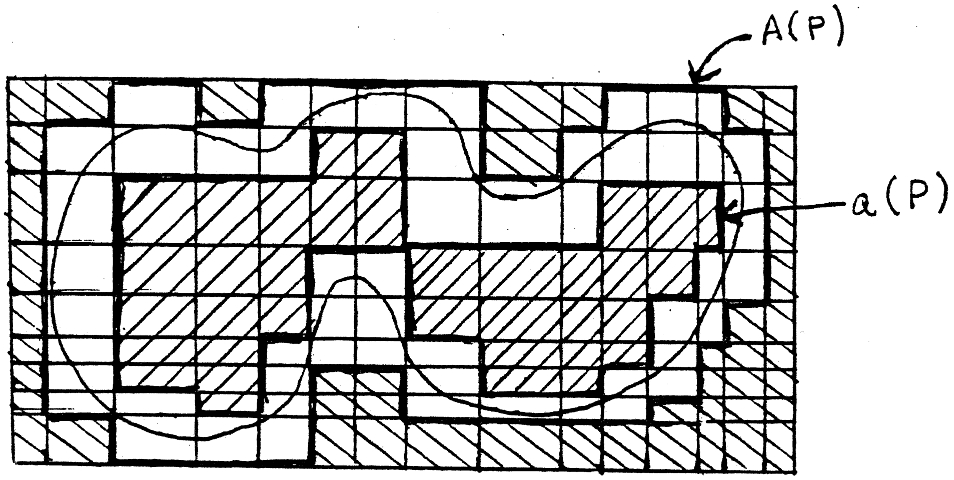
$$\begin{aligned} \text{area } S &= \text{area}(\text{Int } S) + \text{area}(S - \text{Int } S) \\ &= \text{area}(\text{Int } S). \end{aligned}$$

A similar remark shows that

$$\begin{aligned} \text{area}(S \cup \text{Bd } S) &= \text{area}(\text{Int } S) + \text{area}(\text{Bd } S) \\ &= \text{area}(\text{Int } S). \quad \square \end{aligned}$$

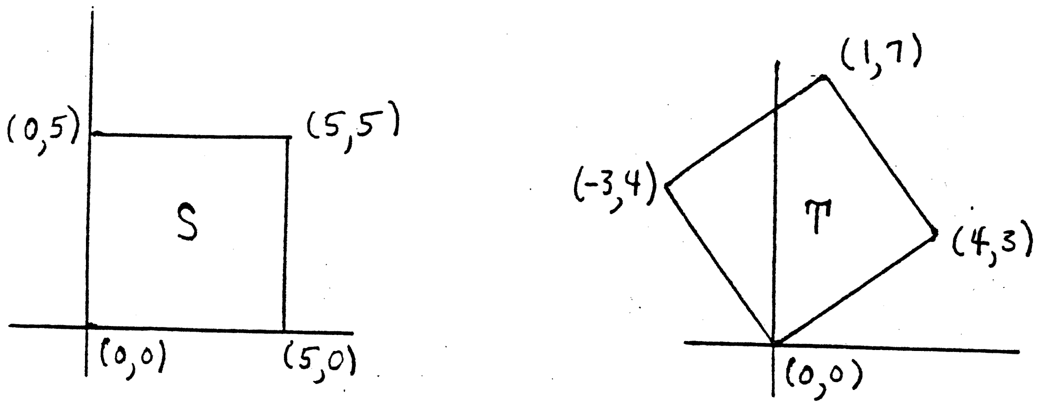
Remark. Let S be a bounded set in the plane. A direct way of defining the area of S, without developing integration theory, is as follows: Let Q be a rectangle containing S.

Given a partition P of Q, let a(P) denote the total area of all subrectangles of P that are contained in S, and let A(P) denote the total area of all subrectangles of P that contain points of S. Define the inner area of S be the supremum of the numbers a(P), as P ranges over all partitions of Q; and define the outer area of S to be the infemum of the numbers A(P). If the inner area and outer area of S are equal, their common value is called the area of S.



We leave it as a (not too difficult) exercise to show that this definition of area is the same as the one we have given.

Remark. There is just one fact that remains to be proved about our notion of area. We would certainly wish it to be true that if two sets S and T in the plane are "congruent" in the sense of elementary geometry, then their areas are the same. This fact is not immediate from the definition of area, for we used rectangles with sides parallel to the coordinate axes to form the partitions on which we based our notion of "integral", and hence of "area". It is not immediate, for instance, that the rectangles S and T pictured below have the same area, for the area of T is defined



by approximating T by rectangles with vertical and horizontal sides. [Of course, we can write equations for the curves bounding T and compute its area by integration, if we wish.]

Proof of the invariance of area under "congruence" will have to wait until we study the problem of change of variables in a double integral.

Exercises

1. Show that if $\iint_S 1$ exists, then Bd S has content zero. [Hint: Choose Q so that $S \subset Q$. Since $\iint_Q 1_S$ exists, there are functions s and t that are step functions relative to a partition P of Q, such that $s \leq 1_S \leq t$ on Q and $\iint_Q (t - s) < \varepsilon$. Show that the subrectangles determined by P that contain points of S have total volume less than ε.]

2. (a) Let S and T be bounded subsets of R^2. Show that Bd $(S \cup T) \subset$ (Bd $S \cup$ Bd T). Give an example where equality does not hold.

(b) Show that if S and T are Jordan-measurable, then so are $S \cup T$ and $S \cap T$, and furthermore

$$\text{area}(S \cup T) = \text{area } S + \text{area } T - \text{area } (S \cap T).$$

3. Express in terms of iterated integrals the double integral $\iint_S x^2y^2$, where S is the bounded portion of the first quadrant lying between the curves $xy = 1$ and $xy = 2$ and the lines $y = x$ and $y = 4x$. (Do not evaluate the integrals.)

4. A solid is bounded above by the surface $z = x^2 - y^2$, below by the xy-plane, and by the plane $x = 2$. Make a sketch; express its volume as an integral; and find the volume.

5. Express in terms of iterated integrals the volume of the region in the first octant of R^3 bounded by: (a) The surfaces $z = xy$ and $z = 0$ and $x + 2y + z = 1$. (b) The surfaces $z = xy$ and $z = 0$ and $x + 2y - z = 1$.

Let Q denote the rectangle $[0,1] \times [0,1]$ in the following exercises.

6. (a) Let $f(x,y) = 1/(y-x)$ if $x \neq y$,

$f(x,y) = 0$ if $x = y$.

Does $\iint_Q f$ exist?

(b) Let $g(x,y) = \sin(1/(y-x))$ if $x \neq y$,

$g(x,y) = 0$ if $x = y$.

Does $\iint_Q g$ exist?

7. Let $f(x,y) = 1$ if $x = 1/2$ and y is rational,

$f(x,y) = 0$ otherwise

Show that $\iint_Q f$ exists but $\int_0^1 f(x,y)dy$ fails to exist when $x = 1/2$.

8. Let $f(x,y) = 1$ if (x,y) has the form $(a/p,b/p)$,

where a and b are integers and p is prime,

$f(x,y) = 0$ otherwise.

Show that $\int_0^1 \int_0^1 f(x,y)dy\,dx$ exists but $\iint_Q f$ does not.

D28

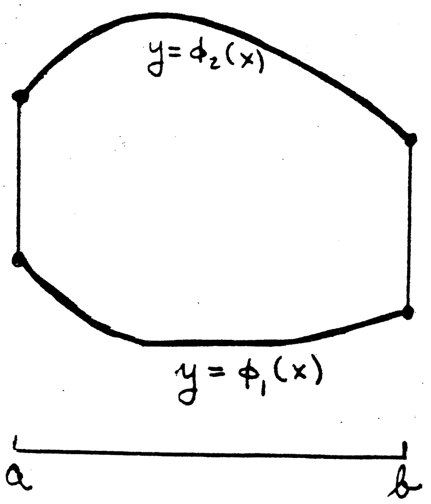
18.024 Multivariable Calculus with Theory
Spring 2011

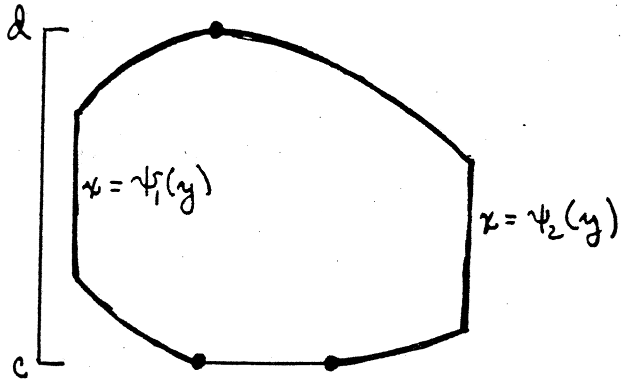
GREEN'S THEOREM AND ITS APPLICATIONS

The discussion in 11.19 - 11.27 of Apostol is not complete nor entirely rigorous, as the author himself points out. We give here a rigorous treatment.

§11.19 Green's Theorem in the Plane

We already know what is meant by saying that a region in the plane is of Type I or of Type II or that it is of both types simultaneously. Apostol proves Green's Theorem for a region that is of both types. Such a region R can be described in two different ways, as follows:





R: $a < x < b$

$\phi_1(x) < y < \phi_2(x)$

R: $c < y < d$

$\psi_1(y) < x < \psi_2(y)$

The author's proof is complete and rigorous except for one gap, which arises from his use of the intuitive notion of "counter-clockwise".

Specifically, what he does is the following: For the first part of the proof he orients the boundary C of R as follows:

(*) By increasing x, on the curve $y = \phi_1(x)$;
By increasing y, on the line segment $x = b$;
By decreasing x, on the curve $y = \phi_2(x)$; and
By decreasing y, on the line segment $x = a$.

Then in the second part of the proof, he orients C as follows:

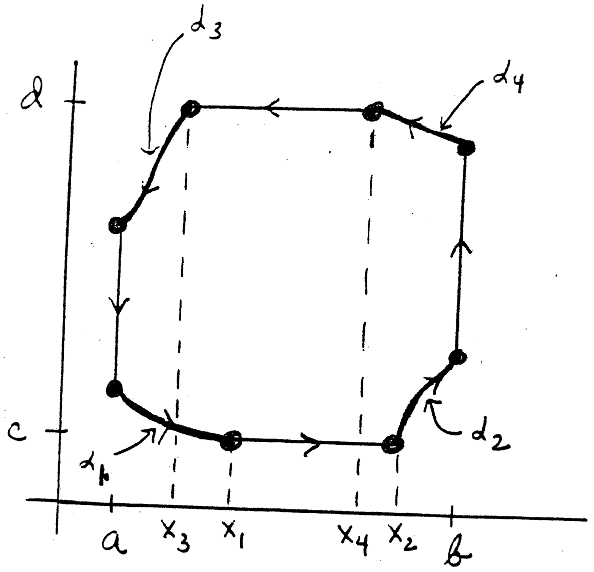
(**) By decreasing y, on the curve $x = \psi_1(y)$;
By increasing x, on the line segment $y = c$;
By increasing y, on the curve $x = \psi_2(y)$; and
By decreasing x, on the line segment $y = d$.

(The latter line segment collapses to a single point in the preceding figure.)

The crucial question is: <u>How does one know these two orientations of</u> C <u>are the same</u>?

One can in fact see that these two orientations are the same, by simply analyzing a bit more carefully what one means by a region of Types I and II.

Specifically, such a region can be described by four monotonic functions:



$y = \alpha_1(x);\ a \leq x \leq x_1,$

$y = \alpha_2(x);\ x_2 \leq x \leq b,$

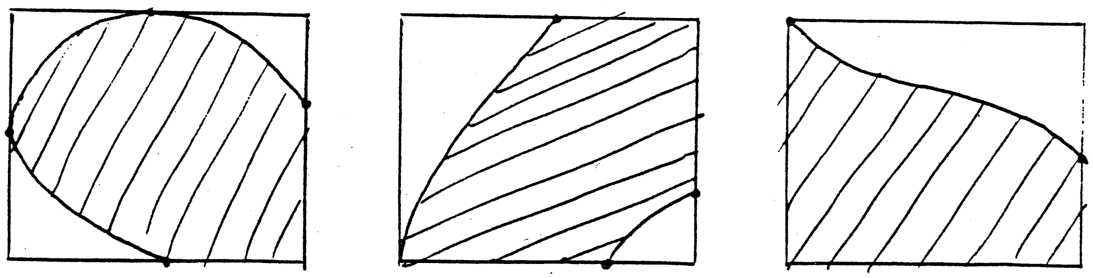
$y = \alpha_3(x);\ a \leq x \leq x_3,$

$y = \alpha_4(x);\ x_4 \leq x \leq b,$

where α_1 and α_4 are strictly decreasing and α_2 and α_3 are strictly increasing. We require that $a \leq x_1 \leq x_2 \leq b$ and $a \leq x_3 \leq x_4 \leq b$; and that $\alpha_1(a) \leq \alpha_3(a)$ and $\alpha_1(x_1) = \alpha_2(x_2)$ and $\alpha_3(x_3) = \alpha_4(x_4)$ and $\alpha_2(b) \leq \alpha_4(b)$, as in the picture.

where $a \leq x_1 \leq x_2 \leq b$ and $a \leq x_3 \leq x_4 \leq b$

[Some or all of the α_i can be missing, of course. Here are pictures of typical such regions:]



The curves α_1 and α_2, along with the line segment $y = c$, are used to define the curve $y = \phi_1(x)$ that bounds the region on the bottom. Similarly, α_3 and α_4 and $y = d$ define the curve $y = \phi_2(x)$ that bounds the region on the top.

Similarly, the *inverse* functions to α_1 and α_3, along with $x = a$, combine to define the curve $x = \psi_1(y)$ that bounds the region on the left; and the inverse functions to α_2 and α_4, along with $x = b$, define the curve $x = \psi_2(y)$.

Now one can choose a direction on the bounding curve C by simply directing each of these eight curves as indicated in the figure, and check that this is the same as the directions specified in (*) and (**). [Formally, one directs these curves as follows:

increasing x = decreasing y on $y = \alpha_1(x)$
increasing x on $y = c$
increasing x = increasing y on $y = \alpha_2(x)$
increasing y on $x = b$
decreasing x = increasing y on $y = \alpha_4(x)$
decreasing x on $y = d$
decreasing x = decreasing y on $y = \alpha_3(x)$
decreasing y on $x = a$.]

We make the following definition:

Definition. Let R be an open set in the plane bounded by a simple closed piecewise-differentiable curve C. We say that R is a *Green's region* if it is possible to choose a direction on C so that the equation

$$\oint_C Pdx + Qdy = \iint_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy$$

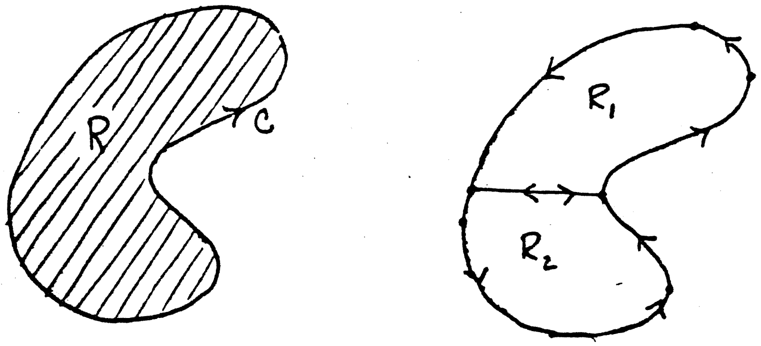
holds for every continuously differentiable vector field $P(x,y)\vec{i} + Q(x,y)\vec{j}$ that is defined in an open set containing R and C.

The direction on C that makes this equation correct is called the *counterclockwise direction*, or the *counterclockwise orientation*, of C.

In these terms, Theorem 11.10 of Apostol can be restated as follows:

Theorem 1. *Let* R *be bounded by a simple closed piecewise-differentiable curve.* *If* R *is of Types* I *and* II, *then* R *is a Green's region.*

As the following figure illustrates, almost any region R you are likely to draw can be shown to be a Green's region by repeated application of this theorem. In such a case, the counterclockwise direction on C is by definition the one for which Green's theorem holds. For example, the region R is a Green's region, and the counterclockwise orientation of its boundary C is as indicated. The figure on the right indicates the proof that it is a Green's region; each of R_1 and R_2 is of Types I and II.

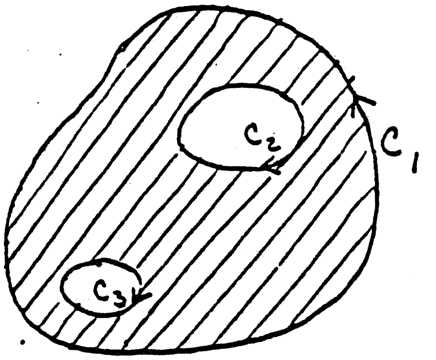


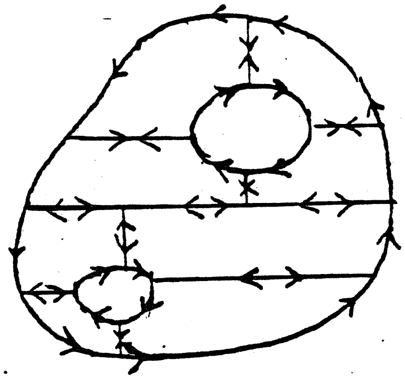
Definition. Let R be a bounded region in the plane whose boundary is the union of the disjoint piecewise-differentiable simple closed curves $C_1, \dots, C_n$. We call R a *generalized Green's region* if it is possible to direct the curves $C_1, \dots, C_n$ so that the equation

$$\int_{C_1+C_2+\ldots+C_n} Pdx + Qdy = \iint_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy$$

holds for every continuously differentiable vector field $P\vec{i} + Q\vec{j}$ defined in an open set about R and C.

Once again, every such region you are likely to draw can be shown to be a generalized Green's region by several applications of Theorem 1. For example, the region R pictured is a generalized Green's region if its boundary is directed as indicated. The proof is indicated in the figure on the right. One applies Theorem 1 to each of the 8 regions pictured and adds the results together.



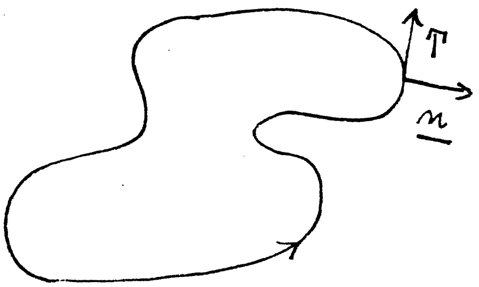


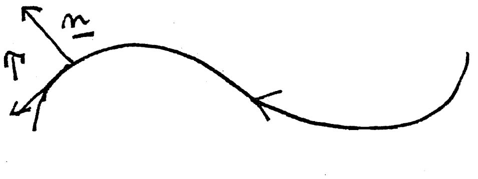
Exercises.

Definition. Let C be a piecewise—differentiable curve in the plane parametrized by the function $\underline{\alpha}(t) = (x(t),y(t))$. The vector $T = (x'(t),y'(t))/\|\underline{\alpha}'(t)\|$ is the unit tangent vector to C. The vector

$$\underline{n} = (y'(t),-x'(t))/\|\underline{\alpha}'(t)\|$$

is called the unit negative normal to C.





If C is a simple closed curve oriented counterclockwise, then $\underline{n}$ is the "outward normal" to C.

(1) If $\underline{f} = P\vec{i} + Q\vec{j}$ is a continuously differentiable vector field defined in an open set containing C, then the integral $\int_C (\underline{f}\cdot\underline{n})dS$ is well–defined; show that it equals the line integral

$$\int_C -Q\,dx + P\,dy.$$

(2) Show that if C bounds a region R that is a Green's region, then $\oint_C (\underline{f}\cdot\underline{n})dS = \iint_R \left[\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right] dx\,dy.$

[<u>Remark</u>. If $\underline{f}$ is the velocity of a fluid, then $\int_C (\underline{f}\cdot\underline{n})dS$ is the area of fluid flowing outward through C in unit time. Thus $\partial P/\partial x + \partial Q/\partial y$ measures the rate of expansion of the fluid, per unit area. It is called the <u>divergence</u> of $\underline{f}$.]

<u>Definition</u>. Let ϕ be a scalar field (continuously differentiable) defined on C. If $\underline{x}$ is a point of C, then $\phi'(\underline{x};\underline{n})$ is the directional derivative of ϕ in the direction of $\underline{n}$. It is equal to $\vec{\nabla}\phi(\underline{x})\cdot\underline{n}$, of course. Physicists and engineers use the (lousy) notation $\frac{\partial\phi}{\partial n}$ to denote this directional derivative.

(3) Let R be a Green's region bounded by C. Let f and g be scalar fields (with continuous first and second partials) in an open set about R and C.

(a) Show $\oint_C \frac{\partial g}{\partial n}\,ds = \iint_R \nabla^2 g\,dx\,dy$

where $\nabla^2 g = \partial^2 g/\partial x^2 + \partial^2 g/\partial y^2$.

(b) Show

$$\oint_C f\frac{\partial g}{\partial n}\,ds = \iint_R (f\nabla^2 g + \vec{\nabla} f\cdot\vec{\nabla} g)dx\,dy.$$

(c) If $\nabla^2 f = 0 = \nabla^2 g$, show

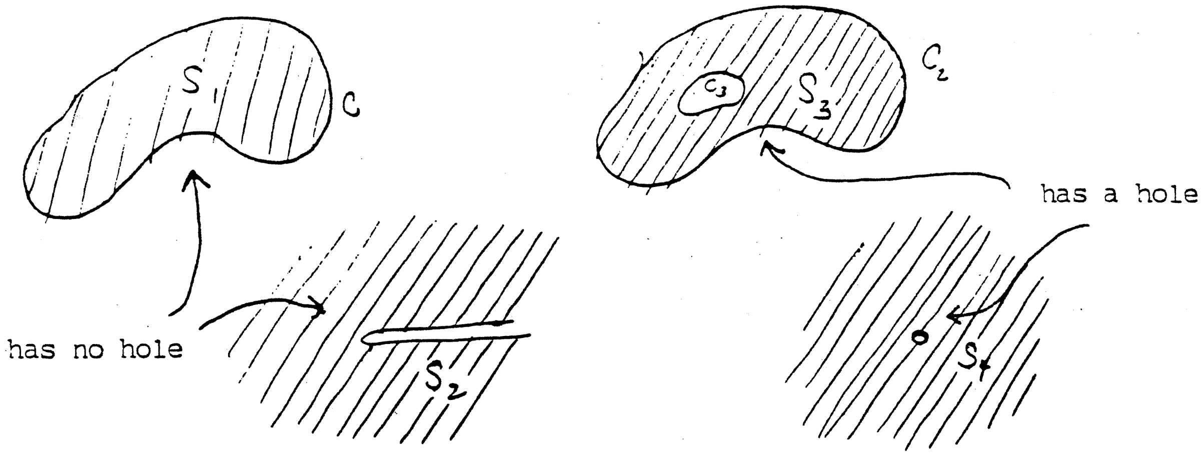
$$\oint_C f\frac{\partial g}{\partial n}\,ds = \oint_C g\frac{\partial f}{\partial n}.$$

These equations are important in applied math and classical physics. A function f with $\nabla^2 f = 0$ is said to be <u>harmonic</u>. Such functions arise in physics: In a region free of charge, electrostatic potential is harmonic; for a body in temperature equilibrium, the temperature function is harmonic.

Conditions Under Which $P\vec{i} + Q\vec{j}$ is a Gradient.

Let $\underline{f} = P\vec{i} + Q\vec{j}$ be a continuously differentiable vector field defined on an open set S in the plane, such that $\partial P/\partial y = \partial Q/\partial x$ on S. In general, we know that $\underline{f}$ need not be a gradient on S. We do know that $\underline{f}$ will be a gradient if S is convex (or even if S is star-convex). We seek to extend this result to a more general class of plane sets.

This more general class may be informally described as consisting of those regions in the plane that have no "holes". For example, the region S_1 inside a simple closed curve C_1 has no holes, nor does the region S_2 obtained from the plane by deleting the non-negative x-axis. On the other hand, the region S_3 consisting of the points inside C_2 and outside C_3 has a hole, and so does the region S_4 obtained from the plane by deleting the origin.

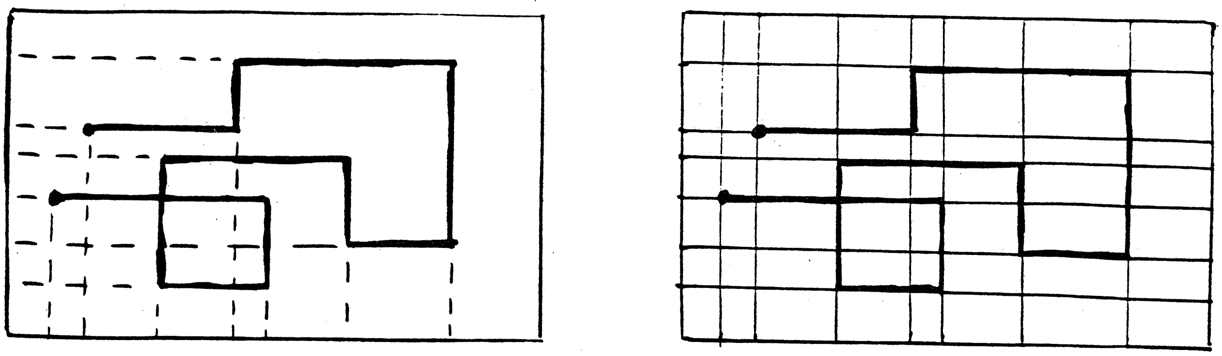


Needless to say, we must make this condition more precise if we are to prove a theorem about it. This task turns out to be surprisingly difficult.

We begin by proving some facts about the geometry of the plane.

Definition. A stairstep curve C in the plane is a curve that is made up of finitely many horizontal and vertical line segments.

For such a curve C, we can choose a rectangle Q whose interior contains C. Then by using the coordinates of the end points of the line segments of the curve C as partition points, we can construct a partition of Q such that C is made up entirely of edges of subrectangles of this partition. This process is illustrated in the following figure:

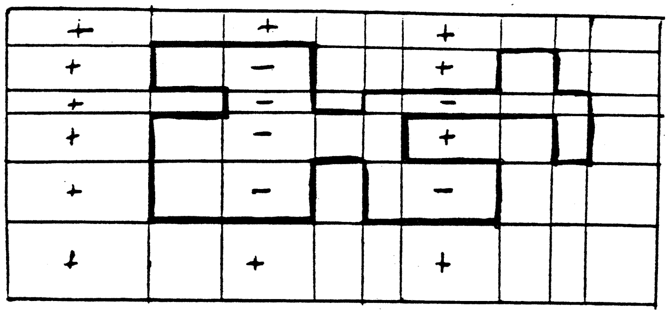


Theorem 2. (*The Jordan curve theorem for stairstep curves*). *Let* C *be a simple closed stairstep curve in the plane. Then the complement of* C *can be written as the union of two disjoint open sets. One of these sets is bounded and the other is unbounded. Each of them has* C *as its boundary.*

Proof. Choose a rectangle Q whose interior contains C, and a partition of Q, say $x_0 < x_1 < \ldots < x_n$ and $y_0 < y_1 < \ldots < y_m$, such that C is made up of edges of subrectangles of this partition.

Step 1. We begin by marking each of the rectangles in the partition + or - by the following rule:

Consider the rectangles in the $i^{\underline{th}}$ "column" beginning with the bottom one. Mark the bottom one with +. In general, if a given rectangle is marked with + or -, mark the one just above it with the *same* sign if their common edge does *not* lie in C, and with the *opposite* sign if this edge *does* lie in C. Repeat this process for each column of rectangles. In the following figure, we have marked the rectangles in columns 1,3, and 6, to illustrate the process.

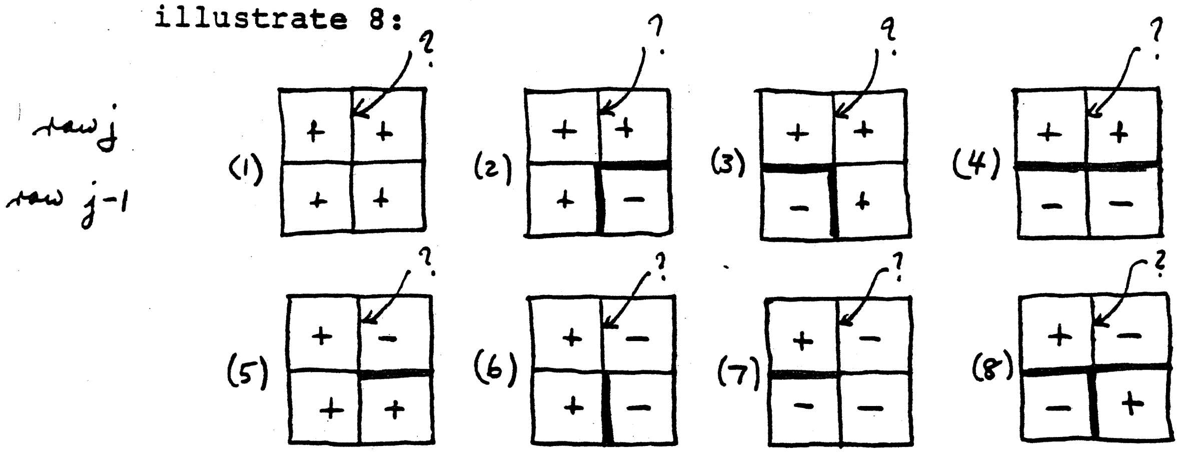


Note that the rectangles in the bottom row are always marked +, and so are those in the first and last columns, (since C does not touch the boundary of Q).

Step 2. We prove the following: If two subrectangles of the partition have an edge in common, then they have opposite signs if that edge is in C, and they have the same sign if that edge is not in C.

This result holds by construction for the horizontal edges. We prove it holds for the vertical edges, by induction.

It is true for each of the lowest vertical edges, those of the form $x_i \times [y_0, y_1]$. (For no such edge is in C, and the bottom rectangles are all marked +.) Supposing now it is true for the rectangles in row j - 1, we prove it true for rectangles in row j. There are 16 cases to consider (!), of which we illustrate 8:



(The other eight are obtained from these by changing all the signs.) We know in each case, by construction, whether the two horizontal edges are in C, and we know from the induction hypothesis whether the lower vectical edge is in C. Those edges that we know are in C are marked heavily in the figure. We seek to determine whether the upper vertical edge (marked "?") is in C or not. We use the fact that C is a simple closed curve, which implies in particular that each vertex in C lies on exactly two edges in C. In case (1), this means that the upper vertical edge is not in C, for otherwise the middle vertex would be on only one edge of C. Similarly, in cases (2), (3), and (4), the upper vertical edge is not in C, for otherwise the middle vertex would lie on three edges of C.

Similar reasoning shows that in cases (5), (6), and (7) the upper vertical edge must lie in C, and it shows that case (8) cannot occur.

Thus Step 2 is proved in these 8 cases. The other 8 are symmetric to these, so the same proof applies.

Step 3. It follows from Step 2 that the top row of rectangles is marked +, since the upper left and upper right rectangles are marked +, and C does not touch the boundary of Q.

Step 4. We divide all of the complement of C into two sets U and V as follows. Into U we put the interiors of all rectangles marked -, and into V we put the interiors of all rectangles marked +. We also put into V all points of the plane lying outside and on the boundary of Q. We still have to decide where to put the edges and vertices of the partition that do not lie in C.

Consider first an edge lying interior to Q. If it does not lie in the curve C, then both its adjacent rectangles lie in U or both lie in V (by Step 2); put this (open) edge in U or in V accordingly. Finally, consider a vertex v that lies interior to Q. If it is not on the curve C, then case (1) of the preceding eight cases (or the case with opposite signs) holds. Then all four of the adjacent rectangles are in U or all four are in V; put v into U or V accordingly.

It is immediately clear from the construction that U and V are open sets; any point of U or V (whether it is interior to a subrectangle, or on an edge, or is a vertex) lies in an open ball contained entirely in U or V . It is also immediate that U is bounded and V is unbounded. Furthermore, C is the common boundary of U and V, because for each edge lying in C, one of the adjacent rectangles is marked + and the other is marked -, by Step 2. □

<u>Definition</u>. Let C be a simple closed stairstep curve in the plane. The bounded open set U constructed in the preceding proof is called the <u>inner region</u> of C, or the region <u>inside</u> C.

It is true that U and V are connected, but the proof is difficult. We shall not need this fact.

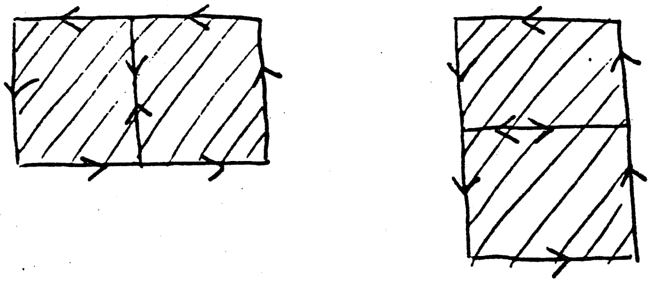
<u>Definition</u>. Let S be an open connected set in the plane. Then S is called <u>simply connected</u>, if, for every simple closed stairstep curve C which lies in S, the inner region of C is also a subset of S.

Theorem 3. *If U is the region inside a simple closed stairstep curve C, then U is a Green's region.*

Proof. Choose a partition of a rectangle Q enclosing U such that C consists entirely of edges of subrectangles of the partition. For each subrectangle Q_{ij} of this partition lying in U, it is true that

$$\int_{C_{ij}} Pdx + Qdy = \iint_{Q_{ij}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy$$

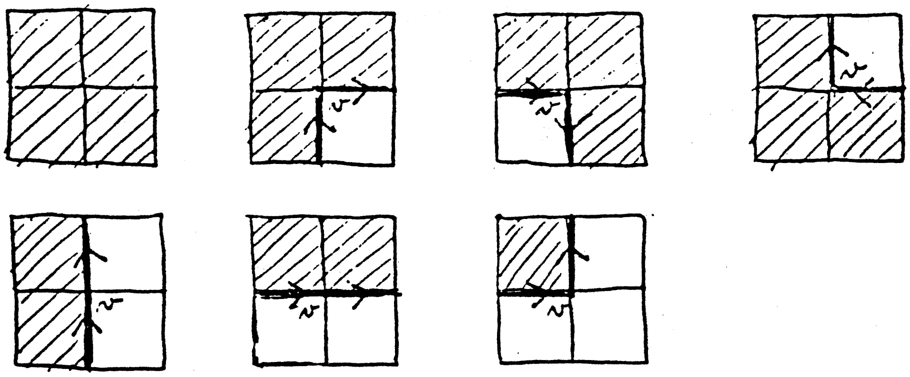
if C_{ij} is the boundary of Q_{ij}, traversed in a counterclockwise direction. (For Q_{ij} is a type I-II region). Now each edge of the partition lying in C appears in only one of these curves C_{ij}, and each edge of the partition not lying in C appears in either none of the C_{ij}, or it appears in two of the C_{ij} with oppositely directed arrows, as indicated:



If we sum over all subrectangles Q_{ij} in U, we thus obtain the equation

$$\int_{(\text{line segments in } C)} Pdx + Qdy = \iint_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy.$$

The only question is whether the directions we have thus given to the line segments lying in C combine to give an orientation of C. That they do is proved by examining the possible cases. Seven of them are as follows; the other seven are opposite to them.



These diagrams show that for each vertex v of the partition such that v is on the curve C, v is the initial point of one of the two line segments of C touching it, and the final point of the other. □

Theorem 4. *Let* S *be an open set in the plane such that every pair of points of* S *can be joined by a stairstep curve in* S. *Let*

$$\underline{f}(x,y) = P(x,y)\vec{i} + Q(x,y)\vec{j}$$

be a vector field that is continuously differentiable in S, *such that*

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$$

on all of S. (a) *If* S *is simply connected*, *then* $\underline{f}$ *is a gradient in* S. (b) *If* S *is not simply connected, then* $\underline{f}$ *may or may not be a gradient in* S.

Proof. The proof of (b) is left as an exercise. We prove (a) here. Assume that S is simply connected.

Step 1. We show that

$$\oint_C Pdx + Qdy = 0$$

for every simple closed stairstep curve C lying in S.

We know that the region U inside C is a Green's region. We also know that the region U lies entirely within S. (For if there were a point p of U that is not in S, then C encircles a point p not in S, so that S has a hole at p. This contradicts the fact that S is simply connected.) Therefore the equation $\partial Q/\partial x = \partial P/\partial y$ holds on all of U; we therefore conclude that

$$\oint_C Pdx + Qdy = \iint_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy = 0,$$

for some orientation of C (and hence for both orientations of C).

Step 2. We show that if

$$\oint_C Pdx + Qdy = 0$$

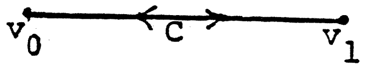
for every simple closed stairstep curve in S, then the same equation holds for every stairstep curve in S.

Assume C consists of the edges of subrectangles in a partition of some rectangle that contains C, as usual.

We proceed by induction on the number of vertices on the curve C. Consider the vertices of C in order:

$$v_0, v_1, \ldots, v_n, v_0.$$

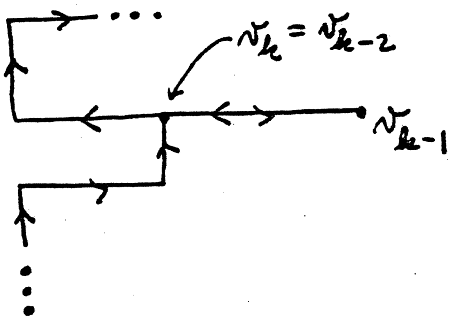
Now C cannot have only one vertex. If it has only two, then C is a path going from v_0 to v_1 and then back to v_0. The line integral vanishes



in this case.

Now suppose the theorem true for curves with fewer than n vertices. Let C have n vertices. If C is a simple curve, we are through. Otherwise, let v_k be the first vertex in this sequence that equals some earlier vertex v_i for $i < k$. We cannot have $v_k = v_{k-1}$, for then $v_{k-1}v_k$ would not be a line segment.

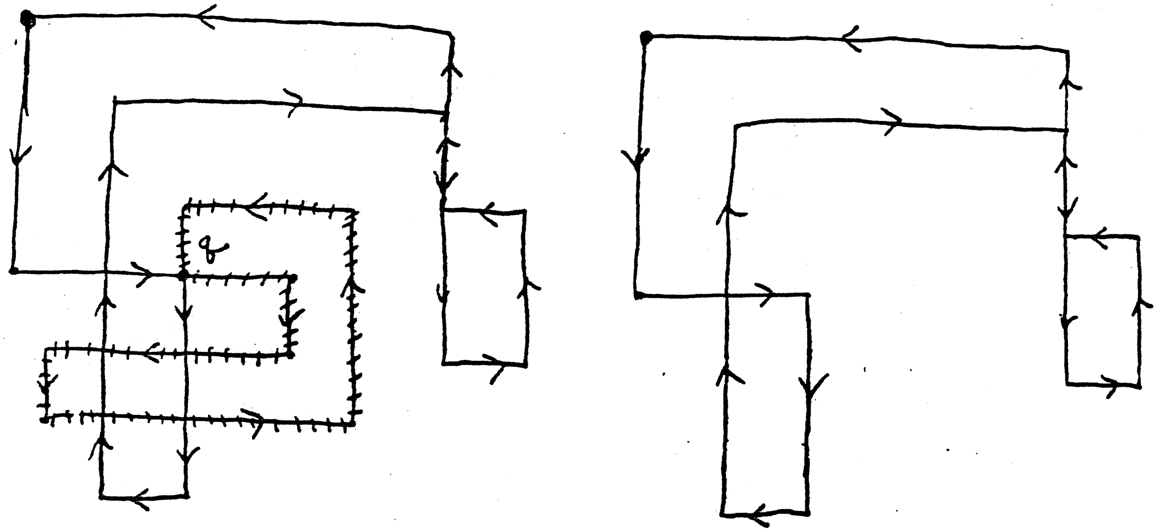
If $v_k = v_{k-2}$, then the curve contains the line segment $v_{k-2}v_{k-1}$, followed by the same line segment in reverse order. Then the integral from v_{k-2} to v_{k-1} and the integral from



v_{k-1} to v_k are negatives of each other. We can delete v_{k-1} from the sequence of vertices without changing the value of the integral. We have a closed curve remaining with fewer line segments than before, and the induction hypothesis applies.

If $i < k-2$, then we can consider the closed curve with vertices $v_i, v_{i+1}, \ldots, v_k$. This is a simple closed curve, since all its vertices are distinct, so the integral around it is zero, by Step 1. Therefore the value of the integral $\int_C Pdx + Qdy$ is not changed if we delete this part of C, i.e., if we delete the vertices $v_i, \ldots, v_{k-1}$ from the sequence. Then the induction hypothesis applies.

Example. In the following case,



the first vertex at which the curve touches a vertex already touched is the point q. One considers the simple closed cross-hatched curve, the integral around which is zero. Deleting this curve, one has a curve remaining consisting of fewer line segments. You can continue the process until you have a simple closed curve remaining.

Step 3. We show that if C_1 and C_2 are any two stairstep curves in S from p to q, then

$$\int_{C_1} Pdx + Qdy = \int_{C_2} Pdx + Qdy.$$

This follows by the usual argument. If $-C_2$ denotes C_2 with the reversed direction, then $C = C_1 + (-C_2)$ in a closed stairstep curve. We have

$$\int_{C_1} - \int_{C_2} = \int_{C_1} + \int_{-C_2} = \oint_C \quad .$$

This last integral vanishes, by Step 2.

Step 4. Now we prove the theorem. Let $\underline{a}$ be a fixed point of S, and define

$$\phi(\underline{x}) = \int_{C(\underline{x})} Pdx + Qdy.$$

where $C(\underline{x})$ is any stairstep curve in S from $\underline{a}$ to $\underline{x}$. There always exists such a stairstep curve (by hypothesis), and the value of the line integral is independent of the choice of the curve (by Step 3). It remains to show that

$$\partial\phi/\partial x = P \text{ and } \partial\phi/\partial y = Q.$$

We proved this once before under the assumption that $C(\underline{x})$ was an arbitrary piecewise smooth curve. But the proof works just as well if we require $C(\underline{x})$ to be a stairstep curve. To compute $\partial\phi/\partial x$, we first computed $[\phi(x+h,y) - \phi(x,y)]/h$. We computed $\phi(x,y)$ by choosing a curve C_1 from $\underline{a}$ to (x,y), and integrated along C_1. We computed $\phi(x+h,y)$ by choosing this same curve C_1 plus the straight line C_2 from (x,y) to $(x+h,y)$. In the present case, we have required C_1 to be a stairstep curve. Then we note that _if C_1 is a stairstep curve, $C_1 + C_2$ is also a stairstep curve_. Therefore the earlier proof goes through without change. □

Remark. It is a fact that if two pair of points of S can be joined by some path in S, then they can be joined by a stairstep path. (We shall not bother to prove this fact.) It follows that the hypothesis of the preceding theorem is merely that S be connected and simply connected.

Exercises

1. Let S be the punctured plane, i.e., the plane with the origin deleted. Show that the vector fields

$$\underline{f} = \frac{x\vec{i} + y\vec{j}}{x^2 + y^2} \qquad \underline{g} = \frac{-y\vec{i} + x\vec{j}}{x^2 + y^2}$$

satisfy the condition $\partial P/\partial y = \partial Q/\partial x$.

(a) Show that $\underline{f}$ is a gradient in S. [_Hint_: First find ϕ so that $\partial\phi/\partial x = x/(x^2 + y^2)$.] (b) Show that $\underline{g}$ is not a gradient in S. [_Hint_: Compute $\int_C \underline{g}\cdot d\underline{\alpha}$ where C is the unit circle centered at the origin.]

2. Prove the following:

Theorem 5. Let C_1 be a simple closed stairstep curve in the plane. Let C_2 be a simple closed stairstep curve that is contained in the inner region of C_1. Show that the region consisting of those points that are in the inner region of C_1 and are not on C_2 nor in the inner region of C_2 is a generalized Green's region, bounded by C_1 and C_2.

[*Hint*: Follow the pattern of the proof of Theorem 3.]

3. Let $\underline{g}$ be the vector field of Exercise 1. Let C be any simple closed stairstep curve whose inner region contains $\underline{0}$. Show that $\int_C \underline{f}\cdot d\underline{r} \neq 0$. [*Hint*: Show this inequality holds if C is the boundary of a rectangle. Then apply Theorem 5.]

*4. Even if the region S is not simply connected, one can usually determine whether a given vector field equals a gradient field in S. Here is one example, where the region S is the punctured plane.

Theorem 6. Suppose that $\underline{f} = P\underline{i} + Q\underline{j}$ *is continuously differentiable and*

$$\partial Q/\partial x = \partial P/\partial y$$

in the punctured plane. Let R *be a fixed rectangle enclosing the origin; orient* Bd R *counterclockwise; let*

$$A = \int_{\text{Bd } R} P\,dx + Q\,dy.$$

(a) *If* C *is any simple closed stairstep curve not touching the origin, then*

$$\int_C P\,dx + Q\,dy$$

either equals $\pm A$ (if the origin is in the inner region of C) *or* 0 (otherwise).

(b) *If* $A = 0$, *then* f *equals a gradient field in the punctured plane*. [Hint: Imitate the proof of Theorem 4.]

(c) *If* $A \neq 0$, *then* $\underline{f}$ *differs from a gradient field by a constant multiple of the vector field*

$$\underline{g}(\underline{x}) = (-y\underline{i} + x\underline{j})/(x^2 + y^2).$$

That is, there is a constant c such that $\underline{f} + c\underline{g}$ equals a gradient field in the punctured plane. (Indeed, $c = -A/2\pi$.)

$$\frac{\partial Q_1}{\partial u} = \left(\frac{\partial Q}{\partial x}\frac{\partial X}{\partial u} + \frac{\partial Q}{\partial y}\frac{\partial Y}{\partial u}\right)\frac{\partial Y}{\partial v} + Q\frac{\partial^2 Y}{\partial u \partial v}$$

$$\frac{\partial P_1}{\partial v} = \left(\frac{\partial Q}{\partial x}\frac{\partial X}{\partial v} + \frac{\partial Q}{\partial y}\frac{\partial Y}{\partial v}\right)\frac{\partial Y}{\partial u} + Q\frac{\partial^2 Y}{\partial v \partial u} .$$

Subtracting, we obtain

$$\frac{\partial Q}{\partial x}\left(\frac{\partial X}{\partial u}\frac{\partial Y}{\partial v} - \frac{\partial X}{\partial v}\frac{\partial Y}{\partial u}\right) = \frac{\partial Q}{\partial x} J(u,v) ,$$

where $\partial Q/\partial x$ is evaluated at $F(u,v)$. Since $\partial Q/\partial x = f$, we have our desired result:

$$\iint_S f(x,y)\,dx\,dy = \pm \iint_T f(F(u,v))\,J(u,v)\,du\,dv. \quad \square$$

One can weaken the hypothesis of this theorem a bit if one wishes. Specifically, it is not necessary that the function $f(x,y)$ which is being integrated be continuous in an entire rectangle containing the region of integration S. It will suffice if $f(x,y)$ is merely continuous on some open set containing S and C. For it is a standard theorem (not too difficult to prove) that in this case one can find a function g that is continuous in the entire plane and equals f on S and C. One then applies the theorem to the function g.

$$\int_C Q\vec{j}\cdot d\underline{\alpha} = \int_a^b Q(\underline{\alpha}(t))\vec{j}\cdot\underline{\alpha}'(t)\ dt$$

$$= \int_a^b Q(\underline{\alpha}(t))\frac{d}{dt}Y(\underline{\beta}(t))\ dt$$

$$= \int_a^b Q(\underline{\alpha}(t))\left(\frac{\partial Y}{\partial u}\beta_1'(t) + \frac{\partial Y}{\partial v}\beta_2'(t)\right)\ dt$$

$$= \int_a^b Q(F(\underline{\beta}(t)))\ \left[\frac{\partial Y}{\partial u}\vec{i} + \frac{\partial Y}{\partial v}\vec{j}\,\right]\cdot\underline{\beta}'(t)\ dt\ ,$$

where the partials are evaluated at $\underline{\beta}(t)$. We can write this last integral as a line integral over the curve D. Indeed, if we define

$$P_1(u,v) = Q(F(u,v))\frac{\partial Y}{\partial u}(u,v)\ ,$$

$$Q_1(u,v) = Q(F(u,v))\frac{\partial Y}{\partial v}(u,v)\ ,$$

then this last integral can be written as

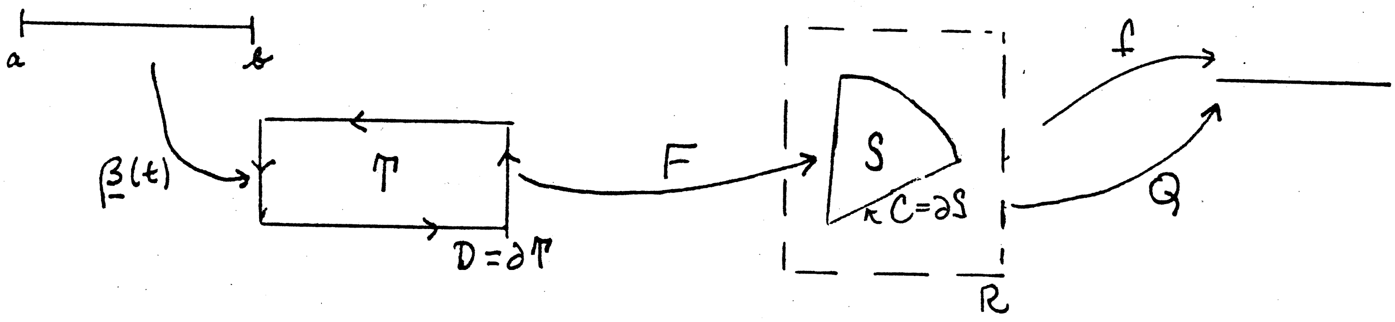
$$\int_D (P_1\vec{i}+Q_1\vec{j})\cdot d\underline{\beta}.$$

Now we apply Green's theorem to express this line integral as a double integral. Since T is by hypothesis a Green's region, this line integral equals

$$\iint_T \left(\frac{\partial Q_1}{\partial u} - \frac{\partial P_1}{\partial v}\right)du\ dv.$$

It remains to compute these partials, using the chain rule. We have

Proof. Let $R = [c,d] \times [c',d']$. Define $Q(x,y) = \int_c^x f(t,y)dt$ for (x,y) in R. Then $\partial Q/\partial x = f(x,y)$ on all of R, because f is continuous. We prove our theorem by applying Green's theorem. Let $(u,v) = \underline{\beta}(t)$ be a parametrization of the curve D, for $a \leq t \leq b$; choose the counterclockwise direction, so Green's theorem holds for T. Then $\underline{\alpha}(t) = F(\underline{\beta}(t))$ is a parametrization of the curve C. It may be constant on some subintervals of the t-axis, but that doesn't matter when we compute line integrals. Also, it may be counterclockwise or clockwise.



We apply Green's theorem to S :

$$\iint_S f(x,y)dx\,dy = \iint_S \partial Q/\partial x\,dx\,dy = \pm \int_C (0\vec{i} + Q\vec{j})\cdot d\underline{\alpha}.$$

This sign is + if $\underline{\alpha}(t)$ parametrizes C in the counterclockwise direction, and - otherwise. Now let us compute this line integral.

The change of variables theorem

Theorem 7. (The change of variables theorem)

Let S be an open set in the (x,y) plane and let T be an open set in the (u,v) plane, bounded by the piecewise-differentiable simple closed curves C and D, respectively. Let $F(u,v) = (X(u,v), Y(u,v))$ be a (continuously differentiable) transformation from an open set of the (u,v) plane into the (x,y) plane that carries T into S, and carries $D = \partial T$ onto $C = \partial S$. As a transformation of D onto C, F may be constant on some segments of D, but otherwise is to be one-to-one.

Assume S and T are Green's regions. Assume that f(x,y) is continuous in some rectangle R containing S. Then

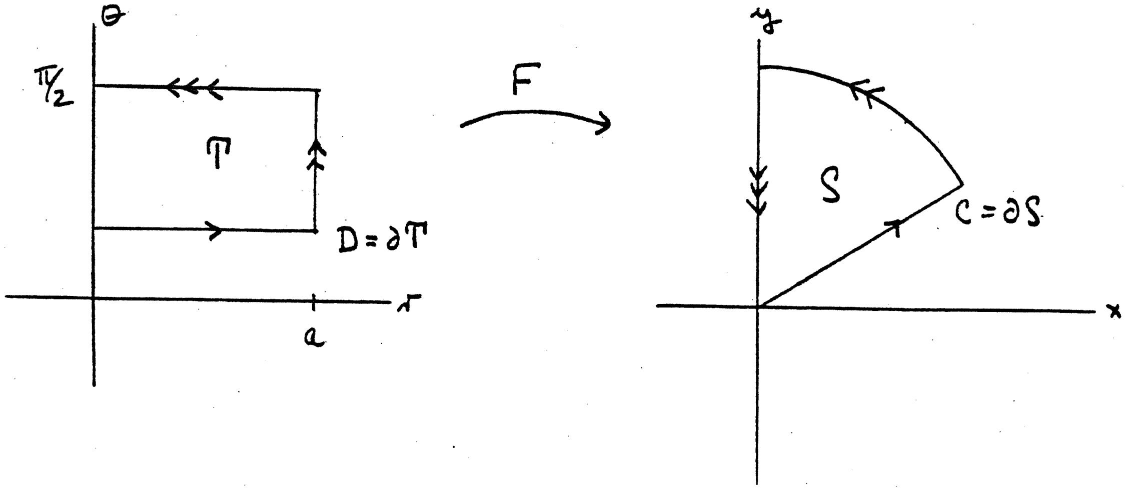
$$\iint_S f(x,y)\,dx\,dy = \pm \iint_T f(F(u,v))\,J(u,v)\,du\,dv\,.$$

Here $J(u,v) = \det \partial X,Y/\partial u,v$. The sign is + if F carries the clockwise orientation of D to the clockwise orientation of C, and is − otherwise.

Example 1. Consider the polar coordinate transformation

$$F(r,\theta) = (r\cos\theta,\ r\sin\theta)\,.$$

It carries the rectangle T in the (r,θ) plane indicated in the figure into the wedge S in the (x,y) plane. It is constant on the left edge of T, but is one-to-one on the rest of T. Note that it carries the counterclockwise orientation of $D = \partial T$ to the counterclockwise orientation of $C = \partial S$.



An alternate version of the change of variables theorem is the following:

Theorem 8. Assume all the hypotheses of the preceding theorem. Assume also that $J(u,v)$ *does not change sign on the region* T.

If $J(u,v) \geqslant 0$ *on all of* T, *the sign in the change of variables formula is* $+$; *while if* $J(u,v) \leqslant 0$ *on all of* T, *the sign is* $-$. *Therefore in either case,*

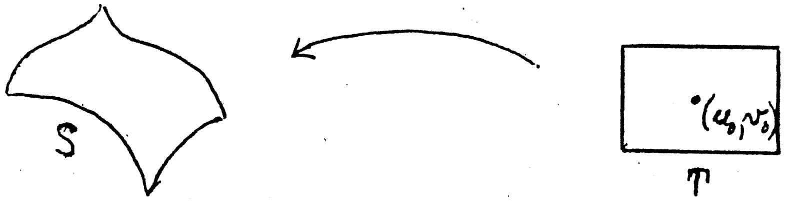
$$\iint\limits_S f(x,y)\,dx\,dy = \iint\limits_T f(F(u,v))\,\left|J(u,v)\right|\,du\,dv.$$

Proof. We apply the preceding theorem to the function $f(x,y) \equiv 1$. We obtain the formula

$$(*) \qquad \iint\limits_S dx\,dy = \pm \iint\limits_T J(u,v)\,du\,dv.$$

The left side of this equation is positive. Therefore if $J(u,v) \geqslant 0$ on all of T, the sign on the right side of the formula must be $+$; while if $J(u,v) \leqslant 0$ on all of T, the sign must be $-$. Now we recall that the sign does not depend on the particular function being integrated, only on the transformation involved. Then the theorem is proved. □

Remark. The formula we have just proved gives a geometric interpretation of the Jacobian determinant of a transformation. If $J(u,v) \neq 0$ at a particular point (u_0,v_0), let us choose a small rectangle T about this point, and consider its image S under the transformation. If T is small enough, $J(u,v)$ will be very close to $J(u_0,v_0)$ on T, and so will not change sign. Assuming S is a Green's region, we have



$$\text{area } S = \iint_S dx\,dy = \iint_T |J(u,v)|\,du\,dv, \quad \text{so}$$

$$\text{area } S \sim |J(u_0,v_0)| \text{ (area } T).$$

Thus, roughly speaking, the magnitude of $J(u,v)$ measures how much the transformation stretches or shrinks areas as it carries a piece of the u, v plane to a piece of the x, y plane. And the sign of $J(u,v)$ tells whether the transformation preserves orientation or not; if the sign is negative, then the transformation "flips over" the region T before shrinking or stretching it to fit onto S.

Isometries

As an application of the change of variables theorem, we shall verify the final property of our notion of area, namely, the fact that congruent regions in the plane have the same area. First, we must make precise what we mean by a "congruence."

Definition. A transformation $\underline{h} : R^2 \longrightarrow R^2$ of the plane to itself is called a congruence or an isometry if it preserves distances between points. That is, $\underline{h}$ is a congruence if

$$\|\underline{h}(\underline{a}) - \underline{h}(\underline{b})\| = \|\underline{a}-\underline{b}\|$$

for every pair $\underline{a}$, $\underline{b}$ of points in the plane.

The following is a purely geometric result:

Lemma 9. *If $\underline{h} : R^2 \longrightarrow R^2$ is a congruence, then h has the form*

$$\underline{h}(x,y) = (ax + by + p, \ cx + dy + q)$$

or, writing vectors as column matrices,

$$\underline{h}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} p \\ q \end{bmatrix},$$

where (a,c) and (b,d) are unit orthogonal vectors. It follows that $ad - bd$, the Jacobian determinant of $\underline{h}$, equals ± 1.

Proof. Let (p,q) denote the point $\underline{h}(0,0)$. Define $\underline{k} : R^2 \longrightarrow R^2$ by the equation

$$\underline{k}(x,y) = \underline{h}(x,y) - (p,q).$$

It is easy to check that $\underline{k}$ is a congruence, since

$$\underline{k}(\underline{a}) - \underline{k}(\underline{b}) = \underline{h}(\underline{a}) - \underline{h}(\underline{b})$$

for every pair of points $\underline{a}$, $\underline{b}$. Let us study the congruence $\underline{k}$, which has the property that $\underline{k}(\underline{0}) = \underline{0}$.

We first show that $\underline{k}$ preserves norms of vectors: By hypothesis,

$$\|\underline{a}-\underline{0}\| = \|\underline{k}(\underline{a}) - \underline{k}(\underline{0})\|, \quad \text{so}$$

$$\|\underline{a}\| = \|\underline{k}(\underline{a}) - \underline{0}\| = \|\underline{k}(\underline{a})\|.$$

Second, we show that $\underline{k}$ preserves dot products: By hypothesis,

$$\|\underline{k}(\underline{a}) - \underline{k}(\underline{b})\|^2 = \|\underline{a}-\underline{b}\|^2, \quad \text{so}$$

$$\|\underline{k}(\underline{a})\|^2 - 2\underline{k}(\underline{a})\cdot\underline{k}(\underline{b}) + \|\underline{k}(\underline{b})\|^2 = \|\underline{a}\|^2 - 2\underline{a}\cdot\underline{b} + \|\underline{b}\|^2.$$

Because $\underline{k}$ preserves norms, we must have

$$\underline{k}(\underline{a}) \cdot \underline{k}(\underline{b}) = \underline{a} \cdot \underline{b}.$$

We now show that $\underline{k}$ is a linear transformation. Let $\underline{e}_1$ and $\underline{e}_2$ be the usual unit basis vectors for R^2; then $(x,y) = x\underline{e}_1 + y\underline{e}_2$. Let

$$\underline{e}_3 = \underline{k}(\underline{e}_1) \quad \text{and} \quad \underline{e}_4 = \underline{k}(\underline{e}_2).$$

Then $\underline{e}_3$ and $\underline{e}_4$ are also unit orthogonal vectors, since $\underline{k}$ preserves dot products and norms. Given $\underline{x} = (x,y)$, consider the vector $\underline{k}(\underline{x})$; because $\underline{e}_3$ and $\underline{e}_4$ form a basis for R^2, we have

$$\underline{k}(\underline{x}) = \alpha(\underline{x})\underline{e}_3 + \beta(\underline{x})\underline{e}_4$$

for some scalars α and β, which are of course functions of $\underline{x}$. Let us compute α and β. We have

$$\begin{aligned}
\alpha(\underline{x}) &= \underline{k}(\underline{x}) \cdot \underline{e}_3 && \text{because } \underline{e}_3 \text{ is orthogonal to } \underline{e}_4, \\
&= \underline{k}(\underline{x}) \cdot \underline{k}(\underline{e}_1) && \text{by definition of } \underline{e}_3, \\
&= \underline{x} \cdot \underline{e}_1 && \text{because } \underline{k} \text{ preserves dot products,} \\
&= x && \text{because } \underline{e}_1 \text{ is orthogonal to } \underline{e}_2.
\end{aligned}$$

Similarly,

$$\beta(\underline{x}) = \underline{k}(\underline{x}) \cdot \underline{e}_4 = \underline{k}(\underline{x}) \cdot \underline{k}(\underline{e}_2) = \underline{x} \cdot \underline{e}_2 = y.$$

We conclude that for all points $\underline{x} = (x,y)$ of R^2,

$$\underline{k}(\underline{x}) = x\underline{e}_3 + y\underline{e}_4.$$

Letting $\underline{e}_3 = (a,c)$ and $\underline{e}_4 = (b,d)$, we can write $\underline{k}$ out in components in the form

$$\underline{k}(\underline{x}) = x(a,c) + y(b,d) = (ax + by,\ cx + dy).$$

Thus $\underline{k}$ is a linear transformation.

Returning now to our original transformation, $\underline{h}$, we recall that

$$\underline{k}(\underline{x}) = \underline{h}(\underline{x}) - (p,q).$$

Therefore we can write out $\underline{h}(\underline{x})$ in components as

$$\underline{h}(\underline{x}) = (ax + by + p,\ cx + dy + q).$$

To compute the Jacobian determinant of $\underline{h}$, we note that because $\underline{e}_3 = (a,c)$ and $\underline{e}_4 = (b,d)$ are unit orthogonal vectors, we have the equation

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a^2+c^2 & ab+cd \\ ab+cd & b^2+d^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Therefore

$$\det \begin{bmatrix} a & c \\ b & d \end{bmatrix} \cdot \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \det \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{or}$$

$$(ad - bc)^2 = 1. \square$$

Theorem 10. *Let* h *be a congruence of the plane to itself, carrying region* S *to region* T. *If both* S *and* T *are Green's regions, then*

$$\text{area } S = \text{area } T.$$

Proof. The transformation carries the boundary of T in a one-to-one fashion onto the boundary of S (since distinct points of R^2 are carried by $\underline{h}$ to distinct points of R^2). Thus the hypotheses of the preceding theorem are satisfied. Furthermore, $|J(u,v)| = 1$. From the equation

$$\iint_S dx\,dy = \iint_T |J(u,v)|\,du\,dv$$

we conclude that

$$\text{area } S = \text{area } T. \;\Box$$

EXERCISES.

1. Let $\underline{h}(\underline{x}) = A \cdot \underline{x}$ be an arbitrary linear transformation of R^2 to itself. If S is a rectangle of area M, what is the area of the image of S under the transformation $\underline{h}$?

2. Given the transformation

$$\underline{h}(x,y) = (ax + by + p,\ cx + dy + q).$$

(a) Show that if (a,c) and (b,d) are unit orthogonal vectors, then $\underline{h}$ is a congruence.

(b) If $ad - bc = \pm 1$, show $\underline{h}$ preserves areas. Is $\underline{h}$ necessarily a congruence?

3. A translation of R^2 is a transformation of the form

$$\underline{g}(\underline{x}) = \underline{x} - \underline{p}$$

where $\underline{p}$ is fixed. A rotation of R^2 is a transformation of the form

$$\underline{h}(\underline{x}) = (x \cos \phi - y \sin \phi, \quad x \sin \phi + y \cos \phi),$$

where ϕ is fixed.

(a) Check that the transformation $\underline{h}$ carries the point with polar coordinates (r,θ) to the point with polar coordinates $(r,\theta+\phi)$.

(b) Show that translations and rotations are congruences. Conversely, show that every congruence with Jacobian $+1$ can be written as the composite of a translation and a rotation.

(c) Show that every congruence with Jacobian -1 can be written as the composite of a translation, a rotation, and the reflection map

$$\underline{k}(x,y) = (-x,y).$$

4. Let A be a square matrix. Show that if the rows of A are orthonormal vectors, then the columns of A are also orthonormal vectors.

5. Let S be the set of all (x,y) with $b^2x^2 + a^2y^2 \leq 1$. Given $f(x,y)$, express the integral $\iint_S f$ as an integral over the unit disc $u^2 + v^2 \leq 1$. Evaluate when $f(x,y) = x^2$.

6. Let C be a circular cylinder of radius a whose central axis is the x-axis. Let D be a circular cylinder of radius $b \leq a$ whose central axis is the z-axis. Express the volume common to the two cylinders as an integral in cylindrical coordinates. [Evaluate when $b = a$ – optional.]

7. Transform the integral in problem 3, p. D.26 by using the substitution $x = u/v$, $y = uv$ with $u, v > 0$. Evaluate the integral.

8. Let S be the parallelogram in the plane with vertices $(0,0)$ and $(1,3)$ and $(2,1)$ and $(3,4)$. Use a suitable linear transformation to transform the integral $\iint_S (x+2y)\,dx\,dy$ into an integral over the unit square $[0,1] \times [0,1]$. Evaluate.

E.36

18.024 Multivariable Calculus with Theory
Spring 2011

Stokes' Theorem

Our text states and proves Stokes' Theorem in 12.11, but it uses the scalar form for writing both the line integral and the surface integral involved. In the applications, it is the vector form of the theorem that is most likely to be quoted, since the notations $dx \wedge dy$ and the like are not in common use (yet) in physics and engineering.

Therefore we state and prove the vector form of the theorem here. The proof is the same as in our text, but not as condensed.

Definition. Let $\vec{F} = P\vec{i} + Q\vec{j} + R\vec{k}$ be a continuously differentiable vector field defined in an open set U of R^3. We define another vector field in U, by the equation

$$\operatorname{curl} \vec{F} = (\partial R/\partial y - \partial Q/\partial z)\,\vec{i} + (\partial P/\partial z - \partial R/\partial x)\,\vec{j} + (\partial Q/\partial x - \partial P/\partial y)\,\vec{k}\,.$$

We discuss later the physical meaning of this vector field.

An easy way to remember this definition is to introduce the symbolic operator "del", defined by the equation

$$\vec{\nabla} = \frac{\partial}{\partial x}\vec{i} + \frac{\partial}{\partial y}\vec{j} + \frac{\partial}{\partial z}\vec{k}\,,$$

and to note that $\operatorname{curl}\vec{F}$ can be evaluated by computing the symbolic determinant

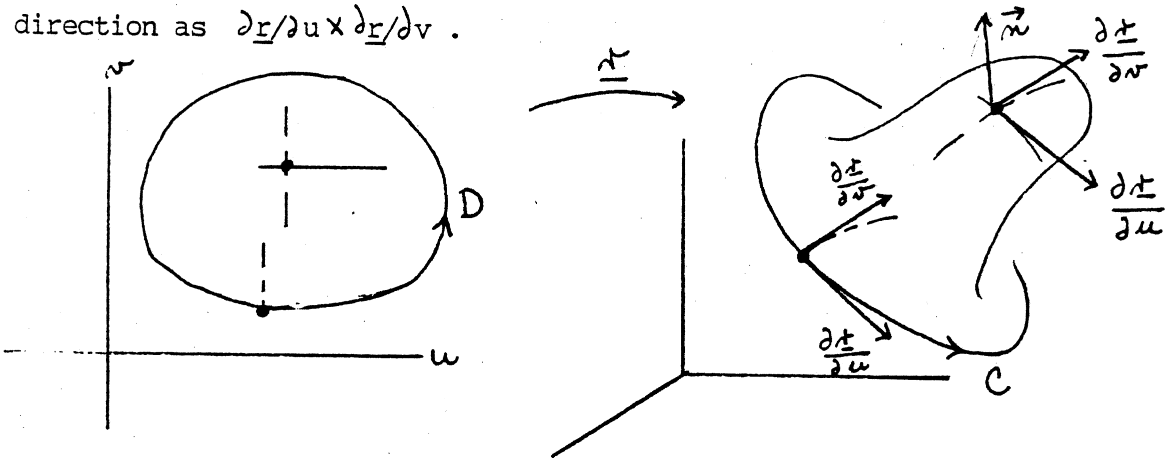
$$\operatorname{curl}\vec{F} = \vec{\nabla}\times\vec{F} = \det\begin{bmatrix} \vec{i} & \vec{j} & \vec{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ P & Q & R \end{bmatrix}.$$

Theorem. (Stokes' theorem). Let S be a simple smooth parametrized surface in R^3, parametrized by a function $\underline{r} : T \rightarrow S$, where T is a region in the (u,v) plane. Assume that T is a Green's region, bounded by a simple closed piecewise-smooth curve D, and that $\underline{r}$ has continuous second-order partial derivatives in an open set containing T and D. Let C be the curve $\underline{r}(D)$.

If F is a continuously differentiable vector field defined in an open set of R^3 containing S and C, then

$$\int_C (\vec{F}\cdot\vec{T})\, ds \quad = \quad \iint_S ((\text{curl}\, \vec{F})\cdot\vec{n})\, dS \ .$$

Here the orientation of C is that derived from the counterclockwise orientation of D; and the normal $\vec{n}$ to the surface S points in the same direction as $\partial\underline{r}/\partial u \times \partial\underline{r}/\partial v$.



v
D
u
r
n
$\frac{\partial r}{\partial v}$
$\frac{\partial r}{\partial u}$
$\frac{\partial r}{\partial v}$
$\frac{\partial r}{\partial u}$
C

Remark 1. The relation between $\vec{T}$ and $\vec{n}$ is often described informally as follows: "If you walk around C in the direction specified by $\vec{T}$, with your head in the direction specified by $\vec{n}$, then the surface S is on your left." The figure indicates the correctness of this informal description.

Remark 2. We note that the equation is consistent with a change of parametrization. Suppose that we reparametrize S by taking a function $\underline{g} : W \rightarrow T$ carrying a region in the (s,t) plane onto T, and use the new parametrization $\underline{R}(s,t) = \underline{r}(\underline{g}(s,t))$. What happens to the integrals in the statement of the theorem? If $\det Dg > 0$, then the left side of the equation is unchanged, for we know that g carries the counterclockwise orientation of ∂W to the counterclockwise orientation of ∂T. Furthermore, because $\partial \underline{R}/\partial s \times \partial \underline{R}/\partial t = (\partial \underline{r}/\partial u \times \partial \underline{r}/\partial v) \det Dg$, the unit normal determined by the parametrization $\underline{R}$ is the same as that determined by $\underline{r}$, so the right side of the equation is also unchanged.

On the other hand, if $\det Dg < 0$, then the counterclockwise orientation of ∂W goes to the opposite direction on C, so that $\vec{T}$ changes sign. But in that case, the unit normal determined by $\underline{R}$ is opposite to that determined by $\underline{r}$. Thus both sides of the equation change sign.

<u>Proof of the theorem.</u> The proof consists of verifying the following three equations:

$$\int_C P\vec{i}\cdot\vec{T}\,ds = \iint_S (\partial P/\partial z\,\vec{j} - \partial P/\partial y\,\vec{k})\cdot\vec{n}\,dS ,$$

$$\int_C Q\vec{j}\cdot\vec{T}\,ds = \iint_S (-\partial Q/\partial z\,\vec{i} + \partial Q/\partial x\,\vec{k})\cdot\vec{n}\,dS ,$$

$$\int_C R\vec{k}\cdot\vec{T}\,ds = \iint_S (\partial R/\partial y\,\vec{i} - \partial R/\partial x\,\vec{j})\cdot\vec{n}\,dS .$$

The theorem follows by adding these equations together.

We shall in fact verify only the first equation. The others are proved similarly. Alternatively, if one makes the substitutions $\vec{i} \to \vec{j}$ and $\vec{j} \to \vec{k}$ and $\vec{k} \to \vec{i}$ and $x \to y$ and $y \to z$ and $z \to x$, then each equation is transformed into the next one. This corresponds to an orientation-preserving change of variables in R^3 , so it leaves the orientations of C and S unchanged.

So let F henceforth denote the vector field $P\vec{i}$; we prove Stokes' theorem in that case.

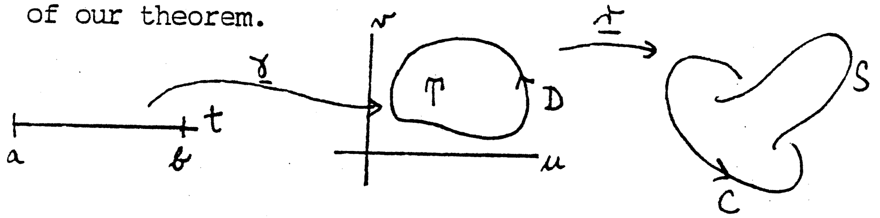
The idea of the proof is to express the line and surface integrals of the theorem as integrals over D and T, respectively, and then to apply Green's theorem to show they are equal.

Let $\underline{r}(u,v) = (X(u,v), Y(u,v), Z(u,v))$, as usual.

Choose a counterclockwise parametrization of D ; call it $\underline{\gamma}(t)$, for $a \leq t \leq b$. Then the function

$$\underline{\alpha}(t) = \underline{r}(\underline{\gamma}(t)) = (X(\underline{\gamma}(t)), Y(\underline{\gamma}(t)), Z(\underline{\gamma}(t)))$$

is the parametrization of C that we need to compute the line integral of our theorem.



We compute as follows:

$$\int_C \vec{F} \cdot d\underline{\alpha} = \int_a^b P(\underline{\alpha}(t))\, \alpha_1'(t) \text{ det}$$

$$= \int_a^b P(\underline{\alpha}(t))[\frac{\partial X}{\partial u}\gamma_1'(t) + \frac{\partial X}{\partial v}\gamma_2'(t)]\, dt ,$$

where $\partial X/\partial u$ and $\partial X/\partial v$ are evaluated at $\underline{\gamma}(t)$, of course. We can write this as a line integral over D. Indeed, if we let p and q be the functions

$$p(u,v) = P(\underline{r}(u,v))\cdot\frac{\partial X}{\partial u}(u,v)$$

$$q(u,v) = P(\underline{r}(u,v))\cdot\frac{\partial X}{\partial v}(u,v) ,$$

then our integral is just the line integral

$$\int_D (p\vec{i} + q\vec{j}) \cdot d\underline{\gamma} .$$

Ncw by Green's theorem, this line integral equals

$$(*) \qquad \iint_T (\partial q/\partial u \; - \; \partial p/\partial v)\, du\, dv .$$

We use the chain rule to compute the integrand. We have

$$\frac{\partial q}{\partial u} = \left(\frac{\partial P}{\partial x}\frac{\partial X}{\partial u} + \frac{\partial P}{\partial y}\frac{\partial Y}{\partial u} + \frac{\partial P}{\partial z}\frac{\partial Z}{\partial u}\right)\frac{\partial X}{\partial v} + P\,\frac{\partial^2 X}{\partial u \partial v}$$

$$\frac{\partial p}{\partial v} = \left(\frac{\partial P}{\partial x}\frac{\partial X}{\partial v} + \frac{\partial P}{\partial y}\frac{\partial Y}{\partial v} + \frac{\partial P}{\partial z}\frac{\partial Z}{\partial v}\right)\frac{\partial X}{\partial u} + P\,\frac{\partial^2 X}{\partial v \partial u} .$$

where P and its partials are evaluated at $\underline{r}(u,v)$, of course. Subtracting, we see that the first and last terms cancel each other. The double integral (*) then takes the form

$$\iint_T \left[- \frac{\partial P}{\partial y}\frac{\partial X,Y}{\partial u,v} + \frac{\partial P}{\partial z}\frac{\partial Z,X}{\partial u,v}\right] du\, dv .$$

Now we compute the surface integral of our theorem. Since $\text{curl}\,\vec{F} = \partial P/\partial z\,\vec{j} - \partial P/\partial y\,\vec{k}$, formula (12.20) on p. 435 of our text tells us we have

$$\iint_S (\text{curl}\,\vec{F})\cdot\vec{n}\,dS = \iint_T \left[\frac{\partial P}{\partial z}\frac{\partial Z,X}{\partial u,v} - \frac{\partial P}{\partial y}\frac{\partial X,Y}{\partial u,v}\right] du\, dv$$

Here $\partial P/\partial z$ and $\partial P/\partial y$ are evaluated at $\underline{r}(u,v)$, of course.

Our theorem is thus proved. □

Exercises on the divergence theorem

1. Let S be the portion of the surface $z = 9 - x^2 - y^2$ lying above the xy plane. Let $\vec{n}$ be the unit upward normal to S. Apply the divergence theorem to the solid bounded by S and the xy-plane to evaluate $\iint_S \vec{F} \cdot \vec{n}\, dS$ if:

(a) $\vec{F} = \sin(y+z)\vec{i} + e^{xz}\vec{j} + (x^2+y^2)\vec{k}$.

(b) $\vec{F} = y^2z\vec{i} + y\vec{j} + z\vec{k}$.

Answers: (a) $81\pi/2$. (b) 81π.

2. Let S_1 denote the surface $z = 1 - x^2 - y^2$; $z \geq 0$. Let S_2 denote the unit disc $x^2 + y^2 \leq 1$, $z = 0$. Let $\vec{F} = x\vec{i} - (2x+y)\vec{j} + zk$; let $\vec{n}_1$ be the unit normal to S_1 and let $\vec{n}_2$ be the unit normal to S_2, both with positive $\vec{k}$ component. Evaluate $\iint_{S_1} \vec{F} \cdot \vec{n}_1\, dS$ and $\iint_{S_2} \vec{F} \cdot \vec{n}_2\, dS$.

Grad, Curl, Div and all that.

We study two questions about these operations:

I. Do they have natural (i.e., coordinate-free) physical or geometric interpretations?

II. What is the relation between them?

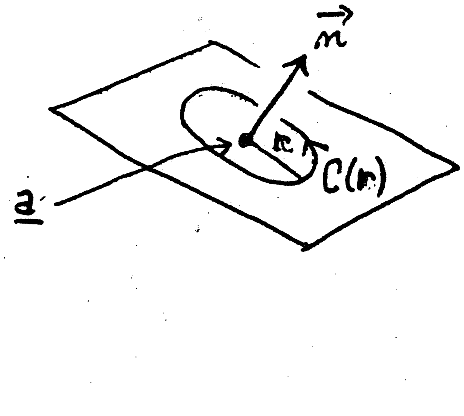
I. We already have a natural interpretation of the gradient.

For divergence, the question is answered in 12.20 of Apostol. The theorem of that section gives a coordinate-free definition of divergence $\vec{F}$, and the subsequent discussion gives a physical interpretation, in the case where $\vec{F}$ is the flux density vector of a moving fluid.

Apostol treats curl rather more briefly. Formula (12.62) on p. 461 gives a coordinate-free expression for $\vec{n} \cdot \operatorname{curl} \vec{F}(\underline{a})$, as follows:

$$(*) \qquad \vec{n} \cdot \operatorname{curl} \vec{F}(\underline{a}) = \lim_{r \to 0} \frac{1}{\pi r^2} \int_{C(r)} \vec{F} \cdot d\vec{\alpha}$$

where $C(r)$ is the circle of radius r centered at $\underline{a}$ lying in the plane perpendicular to $\vec{n}$ and passing through the point $\underline{a}$, and $C(r)$ is directed in a counterclockwise

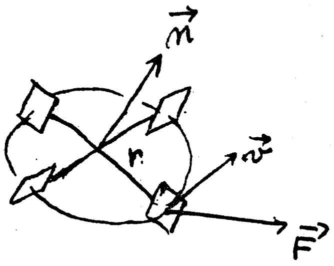


direction when viewed from the tip of $\vec{n}$. This number is called the circulation of $\vec{F}$ at $\underline{a}$ around the vector $\vec{n}$; it is clearly independent of coordinates. Then one has a coordinate-free definition of curl $\vec{F}$ as follows:

> curl $\vec{F}$ at $\underline{a}$ points in the direction of the vector around which the circulation of $\vec{F}$ is a maximum, and its magnitude equals this maximum circulation.

You will note a strong analogy here with the relation between the gradient and the directional derivative.

For a physical interpretation of curl $\vec{F}$, let us imagine $\vec{F}$ to be the velocity vector field of a moving fluid. Let us place a small paddle wheel of radius r in the fluid, with its axis along $\vec{n}$. Eventually, the paddle wheel settles



down to rotating steadily with angular speed ω (considered as positive if counterclockwise as viewed from the tip of $\vec{n}$). The tangential component $\vec{F}\cdot\vec{T}$ of velocity will tend to increase the speed ω if it is positive and to decrease ω if it is negative. On physical grounds, it is reasonable to suppose that

$$\text{average value of } (\vec{F}\cdot\vec{T}) = \text{speed of a point on one of the paddles}$$

$$= r\omega.$$

That is,

$$\frac{1}{2\pi r}\int_C \vec{F}\cdot\vec{T}\,ds = r\omega.$$

It follows that

$$\frac{1}{\pi r^2}\int_C \vec{F}\cdot\vec{T}\,ds = 2\omega,$$

so that by formula (*), we have (if r is very small),

$$\vec{n}\cdot \text{curl } \vec{F}(\underline{a}) = 2\omega.$$

In physical terms then, the vector $[\text{curl } \vec{F}(\underline{a})]$ points in the direction of the axis around which our paddle wheel spins most rapidly (in a counterclockwise direction), and its magnitude equals twice this maximum angular speed.

II. What are the relations between the operations grad, curl, and div? Here is one way of explaining them.

Grad goes from scalar fields to vector fields, Curl goes from vector fields to vector fields, and Div goes from vector fields to scalar fields. This is summarized in the diagram:

$$\begin{tikzcd}
\text{Scalar fields} \arrow[d, bend left, "\text{grad}"] & \phi(\underline{x}) \\
\text{Vector fields} \arrow[d, bend left, "\text{curl}"] & \vec{F}(\underline{x}) \\
\text{Vector fields} \arrow[d, bend left, "\text{div}"] & \vec{G}(\underline{x}) \\
\text{Scalar fields} & \psi(\underline{x})
\end{tikzcd}$$

Let us consider first the top two operations, grad and curl. We restrict ourselves to scalar and vector fields that are continuously differentiable on a region U of R^3.

Here is a theorem we have already proved:

Theorem 1. $\vec{F}$ is a gradient in U if and only if $\oint_C \vec{F} \cdot d\underline{\alpha} = 0$ for every closed piecewise-smooth path in U.

Theorem 2. If $\vec{F} = \text{grad}\ \phi$ for some ϕ, then $\text{curl}\ \vec{F} = \vec{0}$.

Proof. We compute $\text{curl}\ \vec{F}$ by the formula

$$\text{curl}\ \vec{F} = \det \begin{bmatrix} \vec{i} & \vec{j} & \vec{k} \\ D_1 & D_2 & D_3 \\ F_1 & F_2 & F_3 \end{bmatrix}$$

$$= \vec{i}(D_2F_3-D_3F_2) - \vec{j}(D_1F_3-D_3F_1) + \vec{k}(D_1F_2-D_2F_1).$$

We know that if $\vec{F}$ is a gradient, and the partials of F are continuous, then $D_iF_j = D_jF_i$ for all i, j. Hence $\text{curl } \vec{F} = \vec{0}$. □

Theorem 3. *If* $\text{curl } \vec{F} = \vec{0}$ *in a star-convex region* U, *then* $\vec{F} = \text{grad } \phi$ *for some* ϕ *defined in* U.

The function $\psi(x) = \phi(x) + c$ *is the most general function such that* $\vec{F} = \text{grad } \phi$.

Proof. If $\text{curl } \vec{F} = \vec{0}$, then $D_iF_j = D_jF_i$ for all i, j. If U is star-convex, this fact implies that F is a gradient in U, by the Poincaré lemma. □

Theorem 4. *The condition*

$$\text{curl } \vec{F} = \vec{0} \quad \underline{\text{in}} \quad U$$

does not in general imply that $\vec{F}$ *is a gradient in* U.

Proof. Consider the vector field

$$\vec{F}(x,y,z) = (\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}, 0).$$

It is defined in the region U consisting of all of R^3 except for the z-axis. It is easy to check that $\text{curl } \vec{F} = \vec{0}$. To show $\vec{F}$ is not a gradient in U, we let C be the unit circle

$$\underline{\alpha}(t) = (\cos t, \sin t, 0); \quad 0 \leqslant t \leqslant 2\pi$$

in the xy-plane, and compute

$$\oint_C \vec{F} \cdot d\underline{\alpha} = 2\pi \neq 0.$$

It follows from Theorem 1 that $\vec{F}$ cannot be a gradient in U. □

Remark. A region U in R^3 is called "simply connected" if, roughly speaking, every closed curve in U bounds an orientable surface lying in U. The region R^3-(origin) is simply connected, for example, but the region R^3-(z-axis) is not.

It turns out that if U is simply connected and if curl $\vec{F} = \vec{0}$ in U, then $\vec{F}$ is a gradient in U. The proof goes roughly as follows:

Given a closed curve C in U, let S be an orientable surface in U which C bounds. Apply Stokes' theorem to that surface. One obtains the equation

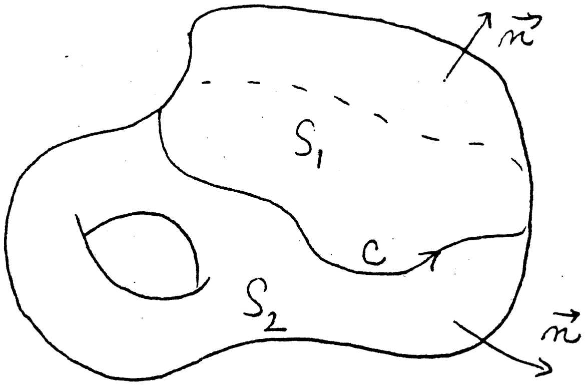
$$\oint_C \vec{F} \cdot d\underline{\alpha} = \iint \text{curl } \vec{F} \cdot \vec{n} \, dS = \iint_S 0 \, dS = 0.$$

Then Theorem 1 shows that $\vec{F}$ is a gradient in U.

Now let us consider the next two operations, curl and div. Again, we consider only fields that are continuously differentiable in a region U of R^3. There are analogues of all the earlier theorems:

Theorem 5. *If* $\vec{G}$ *is a curl in* U, *then* $\iint_S \vec{G}\cdot\vec{n}\; dS = 0$ *for every orientable closed surface* S *in* U.

Proof. Let S be a closed surface that lies in U.



(While we assume that S lies in U, we do not assume that U includes the 3-dimensional region that S bounds.) Break S up into two surfaces S_1 and S_2 that intersect in their common boundary, which is a simple smooth closed curve C. Now by hypothesis, $\vec{G} = \text{curl}\ \vec{F}$ for some $\vec{F}$ defined in U. We compute:

$$\iint_{S_1} \vec{G}\cdot\vec{n}\; dS = \iint_{S_1} \text{curl}\ \vec{F}\cdot\vec{n}\; dS = \int_C \vec{F}\;\; d\underline{\alpha},$$

$$\iint_{S_2} G\cdot n\; dS = \iint_{S_2} \text{curl}\ \vec{F}\cdot\vec{n}\; dS = -\int_C \vec{F}\cdot d\underline{\alpha}.$$

Adding, we see that

$$\iint_S \vec{G}\cdot n\; dS = 0. \square$$

Remark. The converse of Theorem 5 holds also, but we shall not attempt to prove it.

Theorem 6. *If* $\vec{G} = \text{curl}\ \vec{F}$ *for some* $\vec{F}$, *then* $\text{div}\ \vec{G} = 0$.

Proof. By assumption,

$$\vec{G} = \text{curl } \vec{F} = \vec{i}(D_2F_3-D_3F_2)-\vec{j}(D_1F_3-D_3F_1)+\vec{k}(D_1F_2-D_2F_1).$$

Then

$$\text{div } \vec{G} = (D_1D_2F_3-D_1D_3F_2)-(D_2D_1F_3-D_2D_3F_1)+(D_3D_1F_2-D_3D_2F_1)$$

$$= 0. \square$$

Theorem 7. *If* div $\vec{G} = 0$ *in a star-convex region* U, *then* $\vec{G} = \text{curl } \vec{F}$ *for some* $\vec{F}$ *defined in* U.

The function $\vec{H} = \vec{F} + \text{grad } \phi$ *is the most general function such that* $\vec{G} = \text{curl } \vec{H}$.

We shall not prove this theorem in full generality. The proof is by direct computation, as in the Poincaré lemma.

A proof that holds when U is a 3-dimensional box, or when U is all of R^3, is given in section 12.16 of Apostol. This proof also shows how to construct a specific such function $\vec{F}$ in the given cases.

Note that if $\vec{G} = \text{curl } \vec{F}$ and $\vec{G} = \text{curl } \vec{H}$, then $\text{curl}(\vec{H}-\vec{F}) = \vec{0}$. Hence by Theorem 3, $\vec{H} - \vec{F} = \text{grad } \phi$ in U, for some ϕ.

Theorem 8. *The condition*

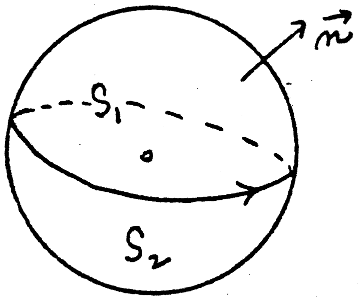
$$\text{div } \vec{G} = 0 \quad \underline{\text{in}} \quad U$$

does not in general imply that $\vec{G}$ *is a curl in* U.

Proof. Let $\vec{G}$ be the vector field

$$\vec{G}(x,y,z) = \frac{x\vec{i} + y\vec{j} + z\vec{k}}{(x^2+y^2+z^2)^{3/2}} ,$$

which is defined in the region U consisting of all of R^3 except for the origin. One readily shows by direct computation that $\operatorname{div} G = 0$.



If S is the unit sphere centered at the origin, then we show that

$$\iint_S \vec{G}\cdot\vec{n}\, dA \neq 0. $$

This will imply (by Theorem 5) that $\vec{G}$ is not a curl.

If (x,y,z) is a point of S, then $\|(x,y,z)\| = 1$, so $\vec{G}(x,y,z) = x\vec{i} + y\vec{j} + z\vec{k} = \vec{n}$. Therefore

$$\iint_S \vec{G}\cdot\vec{n}\, dA = \iint_S 1\, dA = (\text{area of sphere}) \neq 0. \qquad \square$$

Remark. Suppose we say that a region U in R^3 is "two-simply connected" if every closed surface in U bounds a solid region lying in U.* The region $U = R^3-(\text{origin})$ is not "two-simply connected", for example, but the region $U = R^3-(\text{z axis})$ is.

It turns out that if U is "two-simply connected" and if $\text{div}\,\vec{G} = 0$ in U, then $\vec{G}$ is a curl in U. The proof goes roughly as follows:

Given a closed surface S in U, let V be the region it bounds. Since $\vec{G}$ is by hypothesis defined on all of V, we can apply Gauss' theorem to compute

$$\iint_S \vec{G}\cdot\vec{n}\,dS = \iiint_V \text{div}\,\vec{G} = \iiint_V 0 = 0.$$

Then the converse of Theorem 5 implies that $\vec{G}$ is a curl in U.

There is much more one can say about these matters, but one needs to introduce a bit of algebraic topology in order to do so. It is a bit late in the semester for that!

*The proper mathematical term for this is "homologically trivial in dimension two."