Math 522 - Spring 2021

Instructor: Mark Williams

Supplementary notes.

Contents

1 Introduction

Dear Math 522 students,

In this course we will make much use of both linear algebra and metric spaces. I realize that at UNC a linear algebra course is not a prerequisite for this course, and also that some of you did not study metric spaces in Math 521. One of the main reasons for these notes is to fill these gaps. The good news is that you already know a lot about both linear algebra and metric spaces, even if you don't realize it.

The two main concepts of linear algebra are *vector space* and *linear transformation.* Other important concepts are *linear combination*, *linear independence*, and *basis.* For example, $\mathbb{R}^n$ is a classic example of a vector space, and an $n \times n$ matrix A defines a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^n$. You used linear algebra in Math 383 whenever you solved systems of algebraic equations to find eigenvectors. When you wrote the general solution to a homogeneous, second order, linear ODE as a linear combination of two linearly independent solutions, you were using the fact that those two independent solutions form a *basis* for the set of all solutions, a set that is itself another example of a vector space.

A *metric space* is a set equipped with a nonnegative function that gives the "distance" between any two points of the space. You also know a lot about metric spaces, since again $\mathbb{R}^n$ is one of the most important examples of a metric space. The standard *metric* on $\mathbb{R}^n$ is just the usual Euclidean distance between two points in $\mathbb{R}^n$. In general metric spaces we can define *open and closed sets, convergence of sequences, Cauchy sequences, completeness, compactness, and connectedness* in ways that are often (not always) close analogues of definitions you have already encountered in the setting of Euclidean space.

The (possibly) bad news is that you may not yet be familiar with linear algebra and metric spaces at the level of generality that we will need in this course. These notes and material I give in the lectures will provide that extra generality. But as you read these notes, always keep in mind the familiar examples involving Euclidean space.

In these notes I will also provide extra material to clarify topics I have talked about in class, introduce some new topics, finish up things that were not quite finished in class, correct mistakes, and so on. New installments will often be associated with the topic of a recent class.

An inserted "why?" asks about a point that is supposed to be obvious (or perhaps, obvious after some thought). In each case, this is a question you should be able to answer if you have understood the surrounding material.

Best,
Mark Williams

2 Vector spaces.

The two main concepts of linear algebra are "vector spaces" and "linear transformations". Both will play an important role in this course. We begin by discussing the concept of vector spaces.[1]

A vector space V is a set on which two operations satisfying certain axioms are defined. The operations are usually referred to as "addition" and "scalar multiplication". The axioms that must be satisfied in order for the set V with its two operations to qualify as a vector space are listed below. In this course the scalars will always be taken to be either real numbers or complex numbers.[2] In the former case we refer to the "real vector space" $(V, \mathbb{R})$; in the latter case we refer to the "complex vector space" $(V, \mathbb{C})$.

For the moment we consider the case of real scalars. A classic example of a real vector space is

[1]Our text discusses vector spaces and linear transformations in Chapter 1.

[2]In general the scalars can be elements of any "field". Wikipedia provides the definition of a "field".

$(\mathbb{R}^n, \mathbb{R})$, that is, the set [3]

$$\mathbb{R}^n := \{(x_1, \ldots, x_n) : x_j \in \mathbb{R}, j = 1, \ldots, n\}$$

together with the operations of component-wise addition and scalar multiplication that you are familiar with from multivariable calculus:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) := (x_1 + y_1, \ldots, x_n + y_n); \ \alpha(x_1, \ldots, x_n) := (\alpha x_1, \ldots, \alpha x_n) \tag{2.1}$$

where $\alpha \in \mathbb{R}$, the set of scalars. In this case the elements or "points" of the vector space are n-tuples of real numbers. We will give several other examples shortly.

Keep in mind that the notion of vector space is very general, and that there are examples of vector spaces in which the operations of "addition" and "scalar multiplication" are quite different from what you might normally think of as addition and scalar multiplication.

We proceed now to list the axioms that must be satisfied in order for $(V, \mathbb{R})$ to be a real vector space. The axioms for $(V, \mathbb{C})$ are exactly analogous. You should quickly convince yourself that each axiom does hold in the special case of $(\mathbb{R}^n, \mathbb{R})$ with the above two operations. For now we'll use Greek letters α, β, γ, etc., to denote (arbitrary) scalars and lower case Roman letters x, y, z, a, b, c to denote (arbitrary) elements of the space V. Scalars *do not* belong to V except in special cases like $(V, \mathbb{R}) = (\mathbb{R}, \mathbb{R})$.

Vector space axioms.

1. The set V is *closed* under addition and scalar multiplication in the following sense: given any elements $x, y \in V$ and scalar $\alpha \in \mathbb{R}$, one has $x + y \in V$ and $\alpha x \in V$.
2. Associativity properties: $x + (y + z) = (x + y) + z$ and $\alpha(\beta x) = (\alpha\beta)x$.
3. Commutativity of addition: $x + y = y + x$.
4. Distributivity properties: $\alpha(x + y) = \alpha x + \alpha y$ and $(\alpha + \beta)x = \alpha x + \beta x$.
5. Identity elements: There exists an element denoted "0" $\in V$ and called the *zero vector* such that $x + 0 = x$ for all $x \in V$.[4] Thus, the element $0 \in V$ is an "additive identity element".

   Also, $1x = x$, where $1 \in \mathbb{R}$. That is, the scalar $1 \in \mathbb{R}$ is an "identity element for scalar multiplication".[5]
6. Additive inverses: Given any $x \in V$ there exists an element denoted "$-x$" $\in V$ (and called the "additive inverse" of x) such that $x + (-x) = 0$, where 0 is the additive identity element.

Remark 2.1. *After reading the axioms you might naturally wonder if the additive identity element "0" in a vector space is* unique *or if, given $x \in V$, the additive inverse $-x$ is unique. Similarly, the question arises whether or not, given $x \in V$, it is true that $-x = (-1)x$, or whether $0x = 0$. That the answer in all these cases is "yes" follows easily from the next Proposition. We give the proof of $(-1)x = -x$ after the proof of the proposition.*

Proposition 2.2 (Uniqueness of solutions)**.** *Any equation of the form $x + a = b$, where $a, b \in V$, has a unique solution $x \in V$. Here a, b are given elements of V and x is the unknown.*

[3] The symbol " := " indicates a *definition*.

[4] Some authors denote the zero vector by $\mathbf{0}$ or $\underline{0}$ or $\overline{0}$ to distinguish it from the scalar $0 \in \mathbb{R}$. Sometimes we'll do this, but usually we just rely on the context to make this distinction clear. For example, for $x \in V$ the equation $x + 0 = x$ makes no sense in general if the "0" that appears there is in $\mathbb{R}$. The equation only makes sense if $0 \in V$. In the case of $(\mathbb{R}^n, \mathbb{R})$ the zero vector is the $n-$tuple $(0, \ldots, 0)$.

[5] The property $1x = x$ is a separate *axiom*. It is *not* something that is obvious, or that we could deduce in some way from the other axioms.

Proof. First we *suppose* that a solution exists and prove uniqueness. Thus, suppose $x + a = b$ for some $x \in V$. Then[6]

(2.2)
$$(x+a)+(-a) = b+(-a) \Rightarrow x+(a+(-a)) = b+(-a) \Rightarrow x+0 = b+(-a) \Rightarrow x = b+(-a).$$

This implies uniqueness, since it shows that *if* a solution x exists, it must be $b + (-a)$.

To show a solution exists, we simply plug in $b + (-a)$ for x and show that it works. (Do that using the axioms.)

□

To see that $(-1)x = -x$, observe that each side of this equation is a solution of the equation $x + y = 0$ (where y is the unknown). That is clear for $-x$, and you should check this for $(-1)x$. Thus, Proposition 2.2 implies $(-1)x = -x$.

More examples of vector spaces.

Try to quickly convince yourself that $(V, \mathbb{R})$ (or $(V, \mathbb{C})$) in each of the following examples is a vector space.

1. Let $(V, \mathbb{R})$ be the set of all convergent sequences of real numbers:

(2.3) $$V = \{(a_n)_{n=1}^{\infty} : a_n \in \mathbb{R} \text{ for all } n \text{ and } \lim_{n\to\infty} a_n = L \text{ for some } L \in \mathbb{R}\}.$$

Take the operations of addition and scalar multiplication to be the usual (term-by-term) operations by which we add sequences and multiply sequences by real numbers. What mathematical theorem is needed to verify that V is closed under the two operations? Given some $x \in V$ (so x here is a sequence), what is $-x$? What is the element $0 \in V$?

2. Let $(V, \mathbb{R})$ be the set of all real-valued *functions* on $[a, b]$:[7]

(2.4) $$V = \{f : [a, b] \to \mathbb{R}\}$$

Take the operations of addition and scalar multiplication to be the usual operations by which we add functions and multiply functions by scalars. That is, given $f, g \in V$ define $f + g \in V$ by

(2.5) $$(f + g)(x) := f(x) + g(x) \text{ for all } x \in [a, b].$$

Given $\alpha \in \mathbb{R}$ define $\alpha f \in V$ by

(2.6) $$(\alpha f)(x) := \alpha f(x) \text{ for all } x \in [a, b].$$

Note that the elements (or "points") of V are *functions*, and that in our notation we are careful to distinguish the function f from the value of that function at x, namely $f(x)$. Thus, f and $f(x)$ are two completely *different* mathematical objects. The first is a function, the second is a particular real number. The right side of (2.6) is the product of two real numbers. The left side of (2.6) is the real number obtained by evaluating the function αf at x.

Given $f \in V$, what is $-f$, the additive inverse? What is the element $0 \in V$ in this case?

[6] You should identify which axiom is being used for each implication.

[7] We read "$\{f : [a, b] \to \mathbb{R}\}$" as "the set of all functions f mapping $[a, b]$ into $\mathbb{R}$."

3. Let $(V, \mathbb{R})$ be the set of all real-valued *continuous* functions on $[a, b]$:

$$V = \{f : [a, b] \to \mathbb{R} : f \text{ is continuous}\} \tag{2.7}$$

with the same two operations as in example 2. How do we know V is closed under the two operations; that is, what theorem are we using when we assert that V is closed under the two operations?[8]

4. Let $(V, \mathbb{R})$ be the set of all real-valued *differentiable* functions on $[a, b]$:

$$V = \{f : [a, b] \to \mathbb{R} : f \text{ is differentiable}\} \tag{2.8}$$

with the same two operations as in example 2. What theorem are we using now when we assert that V is closed under the two operations?

5. Let $V = C^k([a, b]) := \{f : [a, b] \to \mathbb{R} : f, f', f'', \dots, f^{(k)}$ are all continuous on $[a, b]\}$, where $k \in \mathbb{N}_0$.[9] Again, why is V closed under the two operations?

6. Let $(V, \mathbb{C}) = (\mathbb{C}^n, \mathbb{C})$, where $\mathbb{C}^n := \{(z_1, z_2, \dots, z_n) : z_j \in \mathbb{C}, j = 1, \dots, n\}$. Thus, V is the set of all $n-$tuples of complex numbers. The operations of addition and scalar multiplication are defined just as for $\mathbb{R}^n$, except now the scalars are complex. This is our first example of a complex vector space.

7. Another example of a complex vector space is the set of all complex-valued functions defined on $[a, b]$, $V = \{f : [a, b] \to \mathbb{C}\}$, with the obvious two operations.

8. For $N \in \mathbb{N}$ let $(V, \mathbb{R}) = (\mathcal{M}_N, \mathbb{R})$ be the set of all $N \times \text{N}$ matrices with real entries, with the usual operations of addition and scalar multiplication. The complex vector space $(\mathcal{M}_N, \mathbb{C})$ is defined analogously.

Remark 2.3. *1. Examples 2-5 are of fundamental importance for the study of differential equations. They are examples of "function spaces", that is, vector spaces whose elements are functions.*

2. In contrast to example #1, consider the set V of all sequences of real numbers that fail *to converge, with the same two operations as in example #1. Is V a vector space? Why or why not?*

3 Subspaces, linear span.

In the previous section we gave eight examples of vector spaces (not counting $(\mathbb{R}^n, \mathbb{R})$). Let's denote by V_k the vector space defined in example k, where $k = 1, \dots, 8$. So V_1 is the space of convergent sequences and so on.

Note that the vector space V_3 is a *subset* of the vector space V_2.[10]Also, the vector space V_4 is a subset of the vector space V_3 (why?).

Definition 3.1 (Subspace)**.** *If V is a given vector space and if W is a subset of V that is a vector space in its own right with respect to the operations of addition and scalar multiplication of V, then we call W a* subspace *of V.*

[8]Answer: We use the theorem that the sum of two continuous functions is continuous, and that the product of a continuous function by a scalar is a continuous function.

[9]We read this as "the set of all functions f mapping $[a, b]$ into $\mathbb{R}$ such that f and all its derivatives up to order k are continuous on $[a, b]$." Here $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. Usually we write $C([a, b])$ instead of $C^0([a, b])$.

[10]In these notes we say "A is a subset of B" and write $A \subset B \Leftrightarrow (a \in A \Rightarrow a \in B)$. So if $A = B$ it is true that $A \subset B$.

Thus, for example, V_4 is a subspace of V_3. Similarly, $C^k([a,b])$ is a subspace of $C^m([a,b])$ if $m \leq k$. (why?). Given any vector space V, the trivial subspace $\{0\}$ and V itself are both subspaces of V. The empty set ϕ is a subset of V, but not a subspace of V (why?).

Remark 3.2. *It follows directly from Definition 3.1 that if the vector space $V \neq \{0\}$, any subspace of W such that $W \neq \{0\}$ (including V itself) must contain* infinitely *many distinct elements. For let $x \in W$ be such that $x \neq 0$. Then, for example, the elements $x, 2x, 3x, \ldots$ are distinct elements of W.(why distinct?)*

Now suppose we are given a vector space V and a nonempty *subset* $W \subset V$, and suppose that we want to determine if W is actually a *subspace* of V. The following proposition can save us a lot of work.

Proposition 3.3 (Subspaces)**.** *Let V be a vector space and suppose W is a* nonempty *subset of V. Then, in order to conclude that W is a subspace of V, it is enough to show just that W is* closed *under the two operations of V . (The converse, namely that if W is a subspace, then it is closed under the two operations, is immediate.)*

Proof. We have to show that W (with the same operations as V) satisfies all the axioms of a vector space. By assumption W is closed under the two operations on V. Axioms 2,3,4 are automatically satisfied for W, since every element of W is an element of V, which is a vector space. For the same reason the axiom $1x = x$ holds for $x \in W$.

Now W is nonempty, so let $x \in W$. Then $0x = 0 \in W$, since W is closed under scalar multiplication.[11] Also if $w \in W$, then $-w = (-1)w \in W$, since W is closed under scalar multiplication. Thus, W satisfies all the axioms of a vector space.

□

Examples 3.4. *1. The subset of $\mathbb{R}^3$ given by any (full) line containing the origin is a subspace of $\mathbb{R}^3$. Similarly, any plane containing the origin is a subspace of $\mathbb{R}^3$ (check using Proposition 3.3). A line or plane in $\mathbb{R}^3$ which does not contain the origin is* not *a subspace of $\mathbb{R}^3$. It does not contain the additive identity.*

2. The set of all real-valued polynomials of degree at most $n \in \mathbb{N}_0$, namely

$$\mathcal{P}_n := \{a_0 + a_1 x + \cdots + a_n x^n : a_j \in \mathbb{R},\ j = 0, 1, \ldots, n\},$$

is a subspace of $C(\mathbb{R})$. By Proposition 3.3 we just need to check closure of $\mathcal{P}$ under the two operations. Closure under addition is equivalent to the (obviously true) statement that the sum of two polynomials of degree at most n is a polynomial of degree at most n.

The set $\mathcal{P}_n$ is also a subspace of $C^\infty(\mathbb{R}) := \cap_{k=0}^{\infty} C^k(\mathbb{R})$, the space of functions whose derivatives of all *orders are continuous on $\mathbb{R}$.*

3. The set of all polynomials

$$\mathcal{P} := \cup_{n=0}^{\infty} \mathcal{P}_n$$

is a subspace of $C(\mathbb{R})$.

4. The set of functions $V = \{f \in C^1([a,b]) : f(a) = f'(a) = 0\}$ is a subspace of $C^1([a,b])$. Again, by Proposition 3.3 we just need to check closure of V under the two operations. To check closure under addition, suppose f_1 and f_2 are in V. This means *that each f_i has 3 properties: it is C^1,*

[11] Here we used the fact that $0x = 0$; this follows from Proposition 2.2 by considering the equation $0x+y = 0x$.(how?). By the way, the 0 appearing on the left of $0x = 0$ has to be a real number, while the 0 that appears on the right has to be an element of V, simply because no other interpretation would make any sense in this context.

$f_i(a) = 0$, *and* $f_i'(a) = 0$. *Closure under addition holds because the sum* $f_1 + f_2$ *has those* same *three properties. Here we use the facts from Calculus that the sum of two continuous functions is continuous, and the sum of two differentiable functions is differentiable.*

5. Is the set of functions $W = \{f \in C^1([a,b]) : f(a) = 0, f'(a) = 1\}$ *a subspace of* $C^1([a,b])$*? Why or why not?*

We can use linear combinations of elements of V to construct subspaces of a vector space V.

Definition 3.5 (Linear combination, linear span)**.** *(a) Let* $x_1, \ldots, x_n$ *be elements of a vector space* $(V, \mathbb{R})$. *A* linear combination *of* $x_1, \ldots, x_n$ *is any element of* V *of the form*

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n,$$

where $a_j \in \mathbb{R}$, $j = 1, \ldots, n$. *Linear combinations of elements of complex vector spaces* $(V, \mathbb{C})$ *are defined analogously.*

(b) The linear span *of the set* $T = \{x_1, \ldots, x_n\} \subset V$ *is the set of* all possible *linear combinations of* $x_1, \ldots, x_n$. *We write*

$$\text{span } T := \{a_1x_1 + \cdots + a_nx_n : a_j \in \mathbb{R}, j = 1, \ldots, n\} \subset V. \tag{3.1}$$

Observe that span T *is a subspace of* V *(why?)*

(c) If T *is an infinite subset of* V, *we define the linear span of* T, span T, *to be the set of all possible* finite *linear combinations of elements of* V.[12] *Again,* span T *is a subspace of* V.

4 Linear independence, basis, and dimension

A fundamental notion in linear algebra is the notion of linear independence.

Definition 4.1 (Linear independence)**.** *Let* V *be a vector space. The elements* $x_1, \ldots, x_n$ *in* V *are said to be* linearly independent *if and only if the only possible linear combination of* $x_1, \ldots, x_n$ *that is equal to the zero vector* $0 \in V$ *is the linear combination where all scalar coefficients are zero. Briefly, the elements* $x_1, \ldots, x_n$ *are linearly independent if and only if*

$$a_1x_1 + \cdots + a_nx_n = 0 \Rightarrow a_j = 0 \textit{ for all } j = 1, \ldots, n. \tag{4.1}$$

If $x_1, \ldots, x_n$ *fail to be linearly independent, we say they are linearly dependent. Thus, the elements* $x_1, \ldots, x_n$ *are linearly dependent if and only if there exists* some *linear combination* $a_1x_1 + \cdots + a_nx_n$ *with coefficients* not *all zero, such that*

$$a_1x_1 + \cdots + a_nx_n = 0. \tag{4.2}$$

If the elements $x_1, \ldots, x_n$ *are linearly independent, we will often say that the* set $T = \{x_1, \ldots, x_n\}$ *is linearly independent.*

Remark 4.2. *Suppose the vectors* $x_1, \ldots, x_n$ *are linearly dependent and that* (4.2) *is satisfied with, say,* $a_2 \neq 0$. *Then we can express* x_2 *as a linear combination of* $x_1, x_3, \ldots, x_n$ *(how?). So in that sense* x_2 *"depends on" the other elements. If the elements* $x_1, \ldots, x_n$ *are linearly independent, it is not possible to write any element in this way as a linear combination of the others.(why?) This justifies the terminology of "linear dependence" and "linear independence".*

[12] In a general vector space, infinite linear combinations make no sense, since we have no notion of convergence of partial sums (or convergence of sequences) in a general vector space.

Examples 4.3. *1.) The vectors* $x_1 = (1, 1, 1)$, $x_2 = (1, 0, 1)$, $x_3 = (2, 1, 2)$ *in* $\mathbb{R}^3$ *are linearly dependent because the following linear combination, with coefficients not all zero, gives the zero vector:* $x_1 + x_2 - x_3 = 0$.

2.) The vectors $x_1 = (1, 0, 0)$, $x_2 = (0, 1, 0)$, $x_3 = (0, 0, 1)$ *in* $\mathbb{R}^3$ *are linearly independent because (check!)*

$$a_1x_1 + a_2x_2 + a_3x_3 = 0 \textit{ implies } a_1 = a_2 = a_3 = 0. \tag{4.3}$$

3) The functions f, g, h *defined by* $f(x) = 2x + x^2$, $g(x) = x$, $h(x) = x^2$ *in* $C^\infty(\mathbb{R})$ *are linearly dependent because the following linear combination, with coefficients not all zero, gives the zero function:* $f - 2g - h = 0$. *(The right side of this equation is the* function *that is identically zero for all* $x \in \mathbb{R}$.*)*

4) The functions f, g, h, i *in* $\mathcal{P}_3$ *(recall the Examples 3.4) defined for all* $x \in \mathbb{R}$ *by* $f(x) = 1$, $g(x) = x$, $h(x) = x^2$, $i(x) = x^3$ *are linearly independent. Here is a proof:*

Proof. Suppose $a_0f + a_1g + a_2h + a_3i = 0$. This means that for all $x \in \mathbb{R}$ we have

$$a_01 + a_1x + a_2x^2 + a_3x^3 = 0. \tag{4.4}$$

We must show that all a_j must be zero. Evaluate (4.4) at $x = 0$ to obtain $a_0 = 0$. Then differentiate (4.4) and evaluate the resulting equation at $x = 0$ to obtain $a_1 = 0$. Continue in the same way to obtain $a_2 = 0$ and $a_3 = 0$. □

5) Similarly, the functions $f_0, f_1, \ldots, f_n$ *in* $\mathcal{P}_n$ *given by* $1, x, x^2, \ldots, x^n$ *are linearly independent. This can be shown, using mathematical induction on* n*, using the idea of part (4) of this remark.*

6) Consider the three points x_1, x_2, x_3 *in* $\mathbb{R}^3$*. They are linearly dependent if and only if there is a single plane passing through the origin which contains all three.*

7) Let x_1, x_2 *be elements in any vector space* V*. They are linearly dependent if and only if one of them can be expressed as a scalar multiple of the other.*

8) The exponential functions f, g, h *in* $C^\infty(\mathbb{R})$ *given by* $f(x) = e^{\alpha_1 x}$, $f_2(x) = e^{\alpha_2 x}$, $f_3 = e^{\alpha_3 x}$, *where the* α_i *are distinct real numbers, are linearly independent.*

9). Consider a collection of elements of the form $0, x_1, x_2, \ldots, x_n$ *in any vector space* V*. These elements are linearly dependent (why?).*

4.1 Basis and dimension

Consider again the vector space $\mathbb{R}^n$ and the subset $B = \{e_1, \ldots, e_n\}$, where the element $e_j \in \mathbb{R}^n$ has j-component equal to 1 and all other components zero. It is clear that the set B is linearly independent (recall Example 4.3, part 2) and also that span $B = \mathbb{R}^n$ (because any $(a_1, \ldots, a_n)$ in $\mathbb{R}^n$ can be written $a_1e_1 + \cdots + a_ne_n$). A subset of $\mathbb{R}^n$ with both of these properties is called a *basis* of $\mathbb{R}^n$. More generally we have the following definition:

Definition 4.4 (Basis)**.** *Let* V *be a vector space. Suppose* $B = \{v_1, \ldots, v_n\} \subset V$ *is linearly independent and* span $B = V$*. Then* B *is called a* basis *of* V*.*

The basis $B = \{e_1, \ldots, e_n\}$ given above is often referred to as the *standard basis* of $\mathbb{R}^n$. The next proposition is a simple fact about bases that you should know.

Proposition 4.5. *Let* $B = \{v_1, \ldots, v_n\}$ *be a basis for a vector space* V*. Then any element* $v \in V$ *can be expressed as a linear combination of the elements of* B in only one way*.*

Proof. Let $v \in V$ and suppose $v = c_1v_1 + \cdots + c_nv_n$ and $v = d_1v_1 + \cdots + d_nv_n$. Then

$$0 = (c_1 - d_1)v_1 + \cdots + (c_n - v_n)v_n.$$

Since B is linearly independent, we must have $c_1 - d_1 = 0, \ldots, c_n - d_n = 0$. □

Remark 4.6. *In view of Proposition 4.5, it makes sense to define $c_1, \ldots, c_n$ as the* coordinates *of v* with respect to the basis *B.*

By definition a basis is *finite*. Therefore, not every vector space has a basis. For example, the vector space $\mathcal{P}$ of all real-valued polynomials (Example 3.4, part 3) has no basis (why?).

If V is a vector space that has a basis, it is clear that the basis need not be *unique*. For example, the set $B_2 = \{(0,0,1),(0,1,1),(1,1,1)\}$ is another basis of $\mathbb{R}^3$ in addition to the standard basis (check). However, it turns out that if V has a basis B_1, then any other basis B_2 must have the same number of elements. This fact is the content of the next important theorem, proved in the next section.

Theorem 4.7. *Let V be a vector space and suppose $B_1 = \{v_1, \ldots, v_m\}$ and $B_2 = \{w_1, \ldots, w_n\}$ are bases of V. Then $m = n$. In other words, all bases of V have the same number of elements.*

As a consequence of Theorem 4.7, the following definition of the *dimension* of a vector space V makes sense.

Definition 4.8 (Dimension)**.** *Let V be a vector space. If V has a basis, then the* dimension *of V is the number of elements in any basis. If V does not have a basis, then we say V is* infinite dimensional.

Here is a closely related, useful fact which is also proved in the next section.

Proposition 4.9. *Let V be a vector space of dimension n and suppose $S = \{v_1, \ldots, v_n\} \subset V$ is* any *linearly independent subset containing n elements. Then S is a basis of V.*

5 Proof of Theorem 4.7.

The main ingredient we need to prove that any two bases of a vector space V have the same number of elements is the following proposition.

Proposition 5.1. *Suppose V is a vector space with a basis $B = \{v_1, \ldots, v_m\}$. Let $Y = \{w_1, \ldots, w_n\} \subset V$, where $n > m$. Then Y is linearly dependent (LD for short).*

The proof uses two lemmas.

Lemma 5.2. *Let $S = \{b_1, \ldots, b_k, \overline{0}\} \subset V$. Then S is linearly dependent.*

Proof. Consider the linear combination $c_1b_1 + \cdots + c_kb_k + 1 \cdot \overline{0}$, where all the $c_j = 0$. Why does this prove the result? □

Lemma 5.3. *Suppose* span $\{v_1, v_2, \ldots, v_m\} = V$*. Let $w \in V$ and suppose*

$$v_1 = c_1w + c_2v_2 + \cdots + c_mv_m \text{ for some } c_j \in \mathbb{R}. \tag{5.1}$$

Then span $\{w, v_2 \ldots, v_m\} = V$*.*

Proof. Clearly, span $\{w, v_2 \ldots, v_m\} \subset V$ (why?), so it remains to show $V \subset \text{span}\,\{w, v_2 \ldots, v_m\}$. If $v \in V$, then for some $\alpha_j \in \mathbb{R}$ we have $v = \alpha_1 v_1 + \cdots + \alpha_m v_m = \alpha_1(c_1 w + c_2 v_2 + \cdots + c_m v_m) + \alpha_2 v_2 + \cdots + \alpha_m v_m \in \text{span}\,\{w, v_2 \ldots, v_m\}$. (why?) □

We can now prove Proposition 5.1:

Proof of Proposition 5.1. **1.** If any element of Y is $\overline{0}$ (the zero vector), then by Lemma 5.1 Y is LD and we are done. So we now suppose all $w_j \in Y$ are nonzero.

2. We can write $w_1 = c_1 v_1 + c_2 v_2 + \cdots + c_m v_m$, where $c_j \in \mathbb{R}$ and not all c_j are zero. (why?). Suppose without loss of generality (wlog) that $c_1 \neq 0$. Then

$$v_1 = \frac{1}{c_1}(w_1 - c_2 v_2 - \cdots - c_m v_m).$$

Thus, Lemma 5.3 implies $S_1 := \{w_1, v_2, \ldots, v_m\}$ spans V (that is, the span of S_1 is V).

3. Next consider $w_2 \in Y$. We can write (why?)

$$w_2 = d_1 w_1 + d_2 v_2 + d_3 v_3 + \cdots + d_m v_m. \tag{5.2}$$

If all the coefficients $d_2, \ldots, d_m$ are zero, then Y is LD and we are done (why?). Call this "early termination". Suppose then that early termination does not occur and, wlog, that $d_2 \neq 0$. Then

$$v_2 = \frac{1}{d_2}(w_2 - d_1 w_1 - d_3 v_3 - \cdots - d_m v_m).$$

Applying Lemma 5.3 again (with S_1 now playing the role of the given spanning set), we conclude (how?) that $S_2 := \{w_1, w_2, v_3, \ldots, v_m\}$ spans V.

4. Continuing in the same way, we see that if early termination does not occur, the set $S_m := \{w_1, w_2, \ldots, w_m\}$ spans V. But then for some $\beta_j \in \mathbb{R}$, we have

$$w_{m+1} = \beta_1 w_1 + \cdots + \beta_m w_m.$$

This implies Y is LD (why?), and we are done.

□

A simple consequence of Proposition 5.1 is Theorem 4.7, restated here.

Theorem 5.4. *Suppose $B_1 = \{v_1, \ldots, v_m\}$ and $B_2 = \{w_1, \ldots, w_n\}$ are bases of V. Then $m = n$. In other words, all bases of V have the same number of elements.*

Proof. If $n > m$, then (since B_1 is a basis of V) B_2 is LD by Proposition 5.1. That cannot be (why?), so $n \leq m$. Similarly, if $m > n$ then B_1 is LD, and this cannot be. So $m \leq n$. We conclude $m = n$.

□

Next we prove Proposition 4.9, restated here.

Proposition 5.5. *Let V be a vector space of dimension n and suppose $S = \{v_1, \ldots, v_n\} \subset V$ is* any *linearly independent subset containing n elements. Then S is a basis of V.*

Proof. We only need to show span $S = V$ (why). Clearly, span $S \subset V$ (why?), so it remains to show $V \subset \text{span}\, S$. If $w \in V$ is not in the span of S, we claim $S' := \{w, v_1, \ldots, v_n\}$ is a linearly independent set, which is a contradiction (why?).

Thus it only remains to establish the above claim. So suppose

$$c_0 w + c_1 v_1 + \cdots + c_n v_n = \overline{0}$$

for some $c_j \in \mathbb{R}$. We must have $c_0 = 0$, since otherwise $w \in \operatorname{span} S$ (why?). But then

$$c_1 v_1 + \cdots + c_n v_n = \overline{0},$$

which implies $c_1 = c_2 = \cdots = c_n = 0$ (why?). This establishes the claim. □

6 Linear transformations and linear differential operators

Linear transformations are in a sense the simplest non-constant functions that map one vector space into another. They are interesting in their own right, and are often useful as approximations to more general nonlinear functions.

Definition 6.1. *A linear transformation $T : V \to W$, where V and W are real vector spaces, is a* function *which has the following additional special properties:*

$$\text{(6.1)} \qquad \begin{aligned} &(a)\ T(v_1 + v_2) = T(v_1) + T(v_2) \textit{ for all } v_1, v_2 \in V \\ &(b)\ T(\alpha v_1) = \alpha T(v_1) \textit{ for all } \alpha \in \mathbb{R}. \end{aligned}$$

Linear transformations on complex vector spaces are defined in the obvious analogous way.

Linear transformations are also referred to as linear *operators*, so the words "transformation" and "operator" are interchangeable here. It should be emphasized that the properties (6.1) are very special. In a sense that could be made precise, "most" functions $f : V \to W$ are not linear.

Remark 6.2. *a) Observe that Definition 6.1 would make no sense if V and W were not both sets on which operations of addition and multiplication are defined, or if either V or W failed to be closed under those operations. Thus, linear transformations naturally have vector spaces as their domains and target spaces.*

b) The meaning *of addition (resp. scalar multiplication) generally changes from the left side of* (6.1) *(a) (resp.* (6.1) *(b)) to the right side. In the case $V = W$ these meanings do not change.*

Examples:

1. If $a \in \mathbb{R}^3$, the map $T : \mathbb{R}^3 \to \mathbb{R}$ defined by $T(x) = a \cdot x$ (dot product) is a linear transformation (check).

2. Any $m \times n$ matrix A (m rows, n columns) with real entries defines a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$ by its "usual" action on elements of $\mathbb{R}^n$. Recall, for example, that if A is the 2×3 matrix

$$A = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \text{ and } x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

then

$$Ax := \begin{pmatrix} a_1x_1 + a_2x_2 + a_3x_3 \\ b_1x_1 + b_2x_2 + b_3x_3 \end{pmatrix} \in \mathbb{R}^2.$$

Note that the first and second rows of Ax can be written as dot products, so the linearity of A follows from example 1.

3. An important example for the study of differential equations is the linear operator $D : C^1(\mathbb{R}) \to C(\mathbb{R})$ given by $D(f) := f'$ for $f \in C^1(\mathbb{R})$. Observe that we also have $D : C^k(\mathbb{R}) \to C^{k-1}(\mathbb{R})$ for any $k \in \mathbb{N}$. The operator D on these domains satisfies property (6.1)(a), for example, simply because the derivative of the sum of two functions is the sum of the derivatives. Note also that the derivative of a function in $C^1(\mathbb{R})$ (respectively, $C^k(\mathbb{R})$) is in $C(\mathbb{R})$ (respectively, $C^{k-1}(\mathbb{R})$), simply as a consequence of the *definition* of $C^1(\mathbb{R})$ (respectively, $C^k(\mathbb{R})$) (why?).

4. An example of a linear transformation mapping $C(\mathbb{R})$ to $C(\mathbb{R})$ is the map defined by $T(f) := af$, where $a \in C(\mathbb{R})$. This is just the map given by "multiplication by the function a". Sometimes we will write $af = (aI)(f)$, where $I(f) = f$, the identity operator. Here the identity operator is being *composed* with the operator of multiplication by a.

5. The following functions mapping $\mathbb{R} \to \mathbb{R}$ are *not* linear (why?): $f_1(x) = x^2$, $f_2(x) = 2x + 1$, $f_3(x) = x^n$, $n \in \{2, 3, 4, \dots\}$, $f_4(x) = \sin x$, $f_5(x) = |x|$, $f_6(x) = 7$.

Proposition 6.3. *(a) If S, T are linear transformations mapping V into W, then $S + T : V \to W$ is a linear transformation, and $\alpha S : V \to W$ is a linear transformation (for $\alpha \in \mathbb{R}$).*

(b) If $S : V \to W$ and $T : W \to Z$ are linear transformations, then the composition $T \circ S : V \to Z$ is a linear transformation. Observe that $S \circ T$ is not defined, unless the vector spaces Z and V are the same.

(c) If $T : V \to W$ is linear, then $T(0_V) = 0_W$, where 0_V denotes the additive identity in V.[13]

The exact analogues of (a), (b), (c) hold for complex vector spaces $(V, \mathbb{C})$.

Proof. **a.** We have $(S + T)(v_1 + v_2) = S(v_1 + v_2) + T(v_1 + v_2) = S(v_1) + S(v_2) + T(v_1) + T(v_2) = (S + T)(v_1) + (S + T)(v_2)$. Here the first and last equalities follow from the definition of $S + T$. The second equality uses the assumption of linearity. The rest of the proof of part (a) (involving scalar multiplication) is left to you.

b. We have $(S \circ T)(v_1 + v_2) = S(T(v_1 + v_2)) = S(T(v_1) + T(v_2)) = S(T(v_1)) + S(T(v_2)) = (S \circ T)(v_1) + (S \circ T)(v_2)$. The first and last equalities follow from the definition of $S \circ T$. The second and third equalities use the linearity assumption. The rest of the proof of part (b) is left to you.

c. We have $T(0_V) = T(0_V + 0_V) = T(0_V) + T(0_V)$, so Proposition 2.2 implies $T(0_V) = 0_W$ (why?). □

From Proposition 6.3 we can deduce several things useful for the study of ODEs. First, the operator

$$D^n = D \circ D \circ D \cdots \circ D \text{ (n times)} : C^n(\mathbb{R}) \to C(R) \tag{6.2}$$

is a linear transformation ("linear" for short), since it is a composition of linear transformations. Also, if $a \in C(\mathbb{R})$ the operator

$$aD^n : C^n(\mathbb{R}) \to C(\mathbb{R}) \tag{6.3}$$

is linear, since it is the composition of the linear operator D^n with the linear operator of multiplication by the function a. Explicitly,

$$(aD^n f)(x) := a(x) f^{(n)}(x). \tag{6.4}$$

We can also form a linear operator $L(x, D) : C^n(\mathbb{R}) \to C(\mathbb{R})$ by adding together operators like aD^n:

$$L(x, D) := a_n D^n + a_{n-1} D^{n-1} + \cdots + a_1 D + a_0 I, \text{ where } a_j \in C(\mathbb{R}) \text{ for all } j. \tag{6.5}$$

[13]Consequently, if $T : V \to W$ and $T(0_V) \neq 0_W$, we can conclude immediately that T is not linear.

Definition 6.4. *The operator $L(x, D)$ in* (6.5) *is referred to as* the general linear scalar differential operator of order n *mapping* $C^n(\mathbb{R})$ *to* $C(\mathbb{R})$. *If the coefficients a_j are functions in $C([a, b])$, then we say $L(x, D)$ maps $C^n([a, b])$ to $C([a, b])$ Usually we assume that the "leading coefficient" a_n is never zero on its domain.*

Remark 6.5. *The* general (scalar) linear ordinary differential equation of order n *is given by*

$$L(x, D)y = f, \tag{6.6}$$

where $L(x, D)$ is as in (6.5) *and $f \in C(\mathbb{R})$.*[14] *The use of the word "linear" to describe these equations is justified by the fact that the operator $L(x, D) : C^n(\mathbb{R}) \to C(\mathbb{R})$ is indeed a* linear operator *in the sense of Definition 6.1. We say this operator is "scalar" because the coefficients a_j are scalar-valued functions. Similarly, a system of linear differential equations is given by a linear differential operator whose coefficients A_j are* matrix-valued *functions of x.*

7 More on linear transformations: null space and range

Now we define two important subspaces associated with any linear transformation $T : V \to W$.

Definition 7.1 (Null space and range)**.** *Let V and W be vector spaces and suppose $T : V \to W$ is a linear transformation. The* null space *(or* kernel*) of T, denoted $N(T)$, is the subspace of V given by*

$$N(T) = \{v \in V : Tv = 0\}.$$

The range *of T, denoted $R(T)$ is the subspace of W defined by*

$$R(T) = \{Tv : v \in V\}.$$

Let's check that $N(T)$, which is obviously a subset of V, is a subspace of V. We just need to check it is closed under addition and scalar multiplication. Suppose $v_1, v_2 \in N(T)$. Then $T(v_1 + v_2) = Tv_1 + Tv_2 = 0 + 0 = 0 \in W$. Also, for $\alpha \in \mathbb{R}$, we have $T(\alpha v_1) = \alpha T v_1 = \alpha 0 = 0$. Thus, $v_1 + v_2 \in N(T)$ and $\alpha v_1 \in N(T)$, so we are done.[15]. We leave it to you to check that $R(T)$ is a subspace of W.

Examples 7.2. *1. Consider the linear transformation $A : \mathbb{R}^3 \to \mathbb{R}^3$ defined by the 3×3 matrix $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 6 & 9 \\ 1 & 1 & 1 \end{pmatrix}$. Then $N(T)$ is the set of all solutions $(x, y, z) \in \mathbb{R}^3$ of the following system of* 3 *equations in* 3 *unknowns:*

$$2x + y = 0, \quad 6y + 9z = 0, \quad x + y + z = 0. \tag{7.1}$$

2. Consider the linear differential operator $L(D) : C^2(\mathbb{R}) \to C(\mathbb{R})$ given by $D^2 + 10D + 3$.[16] *Then $N(L(D))$ is the set of all solutions $y \in C^2(\mathbb{R})$ of the differential equation*

$$y'' + 10y' + 3y = 0.$$

It will follow from Corollary 7.8 below that the dimension of $N(L(D))$, $\dim N(L(D))$, is two. Thus, if we can find two independent solutions y_1, y_2 of $L(D)y = 0$, it follows from Proposition 4.9 that

[14] One can also make sense of equations like (6.6) when the right side f is not continuous.

[15] Here we have started to write $T(v) = Tv$ when T is linear, and we have omitted the W on 0_W in the equations of this proof. We will often do these things.

[16] We can also write $L(D)$ as $D^2 + 10D + 3I$, where I is the identity operator, $Iy = y$.

any *solution can be written as* $c_1y_1 + c_2y_2$ *for some constants* c_1, c_2 *(why?). The fact that* $N(L(D))$ *is a subspace of* $C^2(\mathbb{R})$ *is equivalent to the following statement: the sum of two solutions is a solution (the "superposition principle"), and a constant multiple of any solution is a solution.(why?)*

3. (a). Explain why the linear transformation D^k *(taking* k *derivatives) maps* $C^n(\mathbb{R})$ *to* $C^{n-k}(\mathbb{R})$ *for any* $n > k$*. (here* $n, k \in \{0, 1, 2, \dots\}$*).*

(b) What is the null space of $D^3 : C^3(\mathbb{R}) \to C(\mathbb{R})$*? Prove your answer.*

The next proposition rephrases the properties of injectivity (one-to-one-ness) and surjectivity (onto-ness) for linear transformations in term of null space and range.

Proposition 7.3. *Let* $T : V \to W$ *be a linear transformation. Then* T *is a injective if and only if* $N(T) = \{0\}$*, and* T *is surjective if and only if* $R(T) = W$*.*

Proof. The statement about surjectivity is immediate.

Suppose T is injective and that $v \in N(T)$. Then $Tv = 0$. But $T0 = 0$; hence $v = 0$ (why?).

Conversely, suppose $N(T) = \{0\}$ and that $Tv_1 = Tv_2$. Then $T(v_1 - v_2) = 0$, so $v_1 - v_2 = 0$ (why?). □

The following general result about the equation $Tv = f$, where $T : V \to W$ is a linear transformation, is often used in ODE theory.

Proposition 7.4. *Let* $T : V \to W$ *be a linear transformation and let* $f \in W$*. Then* any *solution to the equation* $Tv = f$ *(*if *it has a solution) has the form* $v_p + v_n$*, where* v_p *is a particular solution and* $v_n \in N(T)$*.*

Proof. Suppose v_p is solution. Then clearly $v_p + v_n$ is a solution if $v_n \in N(T)$ (why?). Conversely, if $v \in V$ is any solution, then $T(v - v_p) = Tv - Tv_p = f - f = 0$, so $v - v_p := v_n \in N(T)$. □

The next proposition illustrates connections between several concepts related to linear transformations. In the proof we assume that you are familiar with the use of "row operations" on a matrix to reduce the matrix to "row echelon" form. That is usually taught in Math 383 as a technique for solving systems of algebraic equations, a technique useful for finding eigenvectors of $n \times n$ matrices. We also rely on your familiarity with determinants from previous courses. In particular, we will use the fact that if B is an $n \times n$ matrix obtained from A by a row operation, then $\det A \neq 0 \Leftrightarrow \det B \neq 0$.[17] This fact implies that $\det A \neq 0$ if and only if the row echelon form of A is the identity matrix I_n.(why?)

Proposition 7.5. *Let* $A : \mathbb{C}^n \to \mathbb{C}^n$ *be an* $n \times n$ *matrix. The following statements are equivalent:*

a) $\det A \neq 0$.

b) The system $Ax = b$ *has a unique solution for all* $b \in \mathbb{C}^n$*.*

c) A^{-1} *exists.*

d) $Ax = 0$ *has the unique solution* $0 \in \mathbb{C}^n$*.*

e) The columns of A *are linearly independent in* $\mathbb{C}^n$*.*

Proof. (a⇒ b): The row echelon system equivalent to $Ax = b$ has the form $I_nx = \tilde{b}$.

(b⇒ c): The columns of the inverse matrix are the solutions to $Ab_i = e_i$, $i = 1, 2, 3$.

(c⇒d): Apply A^{-1} to both sides of $Ax = 0$.

(d⇔e) $Ax = 0 \Leftrightarrow \sum_{i=1}^n x_ia_i = 0$, where the a_i are the columns of A.

[17] This follows easily from the discussion of the determinant in section 10.3.

(d⇒a). If $\det A = 0$, the row echelon system equivalent to $Ax = 0$ would be $A_{RE}x = 0$, where A_{RE} has at least one zero row.

□

A matrix with these properties is called *nonsingular*.

7.1 Applications to scalar ODEs

Applications of linear algebra to the study of *scalar* ODEs are largely concerned with the following four results. First, we state the fundamental existence and uniqueness theorem (FEUT) for $n-$th order scalar linear ODEs. For now try mainly to understand the statements of the results. Later we will give analogues of these results for linear first order *systems*.

Theorem 7.6 (FEUT for scalar linear ODEs)**.** *Consider the initial value problem on* $[a, b]$*:*

$$\begin{aligned} &L(x, D)y = f, \\ &y(a) = c_0, y'(a) = c_1, \dots, y^{n-1}(a) = c_{n-1}. \end{aligned} \tag{7.2}$$

Here $L(x, D) : C^n([a, b] \to C([a, b])$ *is any linear* $n-$*th order differential operator (recall* (6.5)*) with* $a_j \in C([a, b])$ *and such that the leading coefficient* a_n *is never zero on* $[a, b]$*. The constants* c_j *(the "initial data") are given, and the function* $f \in C([a, b])$ *(the "interior forcing" or "interior data" term) is given.*

This problem has a unique solution $y \in C^n([a, b])$*.*

The next Corollary is an immediate consequence of the theorem:

Corollary 7.7. *Let* $L(x, D) : C^n([a, b] \to C([a, b])$ *be as in the FEUT. Then* $R(L(x, D)) = C([a, b])$ *(why?).*

A simple (but less immediate) consequence of the FEUT is the following result. We use it when we claim to write down the "general solution" of a linear ODE.

Corollary 7.8. *Let* $L(x, D) : C^n([a, b] \to C([a, b])$ *be an* $n-$*th order operator as in the FEUT. Then* $\dim N(L(x, D)) = n$*.*

Remark 7.9. *Corollary 7.8 (with Proposition 4.9) implies that if we can find* n *linearly independent solutions* $y_1, \dots, y_n$ *of* $L(x, D)y = 0$*, then any solution of* $L(x, D)y = 0$ *can be written as a linear combination of* $y_1, \dots, y_n$ *(why?). In other words* $\{y_1, \dots, y_n\}$ *is a basis of* $N(L(x, D))$*.*

An immediate consequence of Proposition 7.4 and Remark 7.9 is the following proposition.

Proposition 7.10. *Let* $L(x, D)$ *be an* $n-$*th order operator as in the FEUT, and let* $y_1, \dots, y_n$ *be linearly independent solutions of the homogeneous equation* $L(x, D)y = 0$*. If* $f \in C([a, b])$*, the general solution of* $L(x, D)y = f$ *is given by* $y_p + c_1y_1 + \dots + c_ny_n$*, where* y_p *is a particular solution.*

8 Metric spaces

Why study metric spaces? Many texts (such as the one by Abbott which is often used here) do all of Math 521 "on the real line". That is, all (or most of) the main theorems of the course are proved for real-valued functions defined on $\mathbb{R}$. But this is much too restrictive; as you saw in Math 233 we often need to work with functions which are either defined on more general spaces (for example, $\mathbb{R}^n$), or which take values in more general spaces (for example, $\mathbb{R}^m$). For many important applications of analysis (for example, to the study of differential equations), we need to work with functions, such as differential or integral operators, defined on spaces *other than* $\mathbb{R}^n$, namely spaces like $C([a,b])$, the set of continuous functions on the interval $[a,b]$, or $C^k([a,b])$, the set of functions each of whose derivatives up to order k is continuous on $[a,b]$.

In this section you will see that central concepts like distance, sequences, open and closed sets, boundedness, convergence, uniform convergence, completeness, compactness, connectedness, continuity, uniform continuity, and others make perfectly good sense in the general context of metric spaces. Working with metric spaces is partly motivated by considerations of "economy of effort". When we prove a theorem for general metric spaces, in one stroke we have a theorem that applies not just to the real line, but also to $\mathbb{R}^n$, $\mathbb{C}^n$, $C([a,b])$, and many other frequently used spaces. We will also find that there are some important results that we can prove only if we use the tools associated with metric spaces, for example, the fundamental existence and uniqueness theorem for systems of ODEs that we will study later in the course. Finally, the perspective provided by working in this more general context will give you a better understanding of analysis on the real line, because it helps clarify what properties of the real line are essential for the main results. It is much more difficult to see what is essential if the real line is the *only* setting in which results are proved.

Our text discusses metric spaces in Appendix A.1. Another source for this material is the classic text by Walter Rudin, "Principles of Mathematical Analysis".

Definitions. A metric space is a set X with the property that we can speak of the "distance" between any two of its elements. This "distance" must satisfy certain properties. This is made precise in the next definition:

Definition 8.1. *A metric space (X,d) is a set X together with a distance function $d : X \times X \to \{x \in \mathbb{R} : x \geq 0\}$ such that for all $x, y, z \in X$ we have:*[18]

1) $d(x,y) \geq 0$ and $d(x,y) = 0 \Leftrightarrow x = y$.
2) $d(x,y) = d(y,x)$ (d is symmetric)
3) $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality)

If d satisfies these properties we call it a "metric" on X.

These three properties are natural properties for a distance function to satisfy. Here are some examples. In each case you should quickly convince yourself that $d(x,y)$ satisfies the three properties.

1. $X = \mathbb{R}$, $d(x,y) = |x-y|$.
2. $X = \mathbb{R}^n$, $d(x,y) = |x-y| := \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ (the usual Euclidean distance)
3. $X = \mathbb{R}^n$, $d(x,y) = \sum_{i=1}^n |x_i - y_i|$.
4. $X = \mathbb{R}^n$, $d(x,y) = \sup_{i=1,\ldots,n} |x_i - y_i|$.
5. $X = C([0,1];\mathbb{R}) := \{f : [0,1] \to \mathbb{R} : f \text{ is continuous}\}$,[19] $d(f,g) = \sup\{|f(x) - g(x)| : x \in [0,1]\}$.
6. If (X,d) is any metric space and $F \subset X$, then (F,d) is also a metric space. We say that F "inherits" the metric from X.

Remark 8.2. *a) Example #5 is a fundamental new type of example that will be very important in this course, a metric space whose "elements" (or "points") are* functions. *For example, the function*

[18] $X \times X := \{(a,b) : a, b \in X\}$

[19] This is the set of all continuous real-valued functions defined on the closed interval $[0,1]$.

f_1 defined by $f_1(x) = \sin x$ is an element of this metric space. So is the function f_2 defined by $f_2(x) = e^x$.

b) What facts from Calculus 1 guarantee that the supremum that defines $d(f,g)$ in #5 is a finite *nonnegative real number?*

c) Suppose $X = C((0,1);\mathbb{R}) := \{f : (0,1) \to \mathbb{R} : f \text{ is continuous}\}$, and $d(f,g) := \sup\{|f(x) - g(x)| : x \in (0,1)\}$. Is (X,d) a metric space? The answer is No. Why?

d) Note that in part (a) of this remark, we distinguish *between the function f and the value of that function at a particular $x \in [0,1]$, namely $f(x)$. Sometimes (e.g., now) it is important to do that. The elements of X in example #5 are* functions f, *not values $f(x)$ taken by functions. The function f and the real number $f(x)$ are totally different mathematical objects. It is common for authors to be sloppy about this distinction, and later we will be sloppier probably.*

e) Probably I will often use (E,d) as well as (X,d) to denote a metric space.

More examples:

6. Let X be any set. Define $d(x,y) = 1$ if $x \neq y$, $d(x,y) = 0$ if $x = y$. This defines a metric. (Check!) This shows that absolutely any set can be made into a metric space.

7. Note that if we take $X = \mathbb{R}^n$ and $d(x,y) = \inf_{i=1,\dots,n} |x_i - y_i|$, then (X,d) is NOT a metric space (why?). If we take $d(x,y) = |x_2 - y_2|$, again (X,d) is not a metric space.(why not?)

Definition 8.3 (Sequences)**.** *a)Let (X,d) be a metric space. A* sequence *in X is a function $p : \mathbb{N} \to X$. We usually denote $p(n)$ by p_n. We often represent the sequence by $(p_n)_{n=1}^{\infty}$ or simply (p_n) or even more simply (and with some abuse) p_n.*

b) Let (p_n) be a sequence in X and suppose $p \in X$. We say the sequence converges to p if and only if given $\epsilon > 0$ there exists $N = N(\epsilon)$ such that if $n \geq N$ we have $d(p_n, p) < \epsilon$.

c) A subsequence *of (p_n) is defined to be a sequence $(q_m)_{m=1}^{\infty} := (p_{n_m})_{m=1}^{\infty}$, where the n_m are natural numbers satisfying $n_1 < n_2 < n_3 < \dots$. Observe that $q : \mathbb{N} \to X$.*

d) A sequence (p_n) in X is called a Cauchy sequence *if and only if given $\epsilon > 0$ there exists $N = N(\epsilon)$ such that if $m, n \geq N$ we have $d(p_n, p_m) < \epsilon$.*

Note that in part (b) of this definition if we replace X by $\mathbb{R}$ and $d(x,y)$ by $|x-y|$ we get the familiar definition of convergence of a sequence of reals.

Definition 8.4 (bounded sets)**.** *Let (X,d) be a metric space and suppose $S \subset X$. If S is empty, S is* bounded. *Otherwise, we say S is* bounded *if and only if there exists $q \in X$ and $r > 0$ such that $S \subset \{p \in X : d(p,q) < r\}$. Equivalently, we say that any $S \subset X$ is bounded when the range of $d : S \times S \to \mathbb{R}$ is bounded.*

We call the set $\{p \in X : d(p,q) < r\}$ the ball *centered at q of radius r in X, and we denote it by $B(q,r)$. Later, after we define the word "open", we'll call this set the* open ball *of radius r in X.*

In Math 521 even if you did not work with metric spaces, you studied proofs of a number of facts about sequences of real numbers, including the following:
a) if the limit of a sequence exists, it is unique.
b) a convergent sequence is bounded.
c) every subsequence of a convergent sequence converges to the same limit.

These same facts hold by essentially the same proofs for sequences in general metric spaces. (Just replace $|x-y|$ by $d(x,y)$ in the proofs!)

For sequences (a_n) with $a_n \in \mathbb{R}$, we also have the well-known facts that the sum of two convergent sequences converges to the sum of the limits, the product of two convergent sequences converges to the product of the limits, and similarly for quotients. The last-mentioned facts do not even make sense in many metric spaces (why?).

8.1 Open and closed sets in metric spaces.

First recall the definition of a ball.

Definition 8.5 (a ball)**.** *Let (X, d) be a metric space and let $p \in X$, $r > 0$. We define* the ball centered at p of radius r *to be* $\{q \in X : d(q, p) < r\}$.

Note that we use $<$ here, not $\leq$.

If X is $\mathbb{R}$ with the usual distance, then, for example, $B(1, 2)$ is the interval $(-1, 3)$, right? If $X = \mathbb{R}^2$ with the usual Euclidean distance and $p = (1, 2)$, then $B(p, 5)$ is the set consisting of all points strictly inside the circle of radius 5 centered at the point $(1, 2)$, right? In $\mathbb{R}^3$ with the usual Euclidean distance and $p = (0, 7, 8)$, the ball $B(p, 10)$ is what?

Example 8.6. *Suppose $X = C([0, 1], \mathbb{R})$ with its usual metric. Let p denote the function whose value at $x \in [0, 1]$ is $p(x) = \sin x$. Then for $r > 0$ we have*

$$B(p, r) = \{q \in C([0, 1], \mathbb{R}) : d(p, q) < r\} = \{q \in C([0, 1], \mathbb{R}) : \sup_{x \in [0,1]} |p(x) - q(x)| < r\}. \tag{8.1}$$

What could you say if you were asked to describe this ball pictorially*? Note that the "points" q in $B(p, r)$ are* functions*.*

Example 8.7. *If (X, d) is a metric space and $F \subset X$, recall that (F, d) is also a metric space. If $p \in F$ observe that the ball in F of radius r centered at p, namely $B_F(p, r) := \{q \in F : d(q, p) < r\}$ is* not *the same (in general) as the ball in X of radius r centered at p (why?).*

For example, consider $[0, 1] = F \subset \mathbb{R}$. We have $B_F(0, \frac{1}{2}) = [0, \frac{1}{2})$, while the corresponding ball in $(\mathbb{R}, |x - y|)$ is $B(0, \frac{1}{2}) = (-\frac{1}{2}, \frac{1}{2})$.

Definition 8.8 (open and closed sets)**.** *(a) Let (X, d) be a metric space and let $S \subset X$. We say S is an* open set *(or open) if, given* any *point $p \in S$ there exists $r > 0$ such that $B(p, r) \subset S$.*

(b) A set $S \subset X$ is closed *if its complement in X, $S^c := \{x \in X : x \notin S\} = X \setminus S$, is open.*

Thus, a set $S \subset X$ is open if and only if every point of S is the center of some ball that is completely contained in S.

In a given metric space some sets may be neither open nor closed. For example, in $\mathbb{R}$ with the usual metric the interval $(1, 3)$ is open (check), the interval $[1, 3]$ is closed, but the interval $(1, 3]$ is neither open nor closed. You should draw examples of sets in $\mathbb{R}^2$ with the usual metric that are open, examples that are closed, and examples that are neither open nor closed.

Proposition 8.9 (Basic properties of open sets in (X, d))**.**

1. The set X is (clearly) open. The empty set ϕ is (by definition, or by default) open.

2. Finite intersections of open sets are open. That is, if $\{U_i : i \in I\}$ is a collection of open sets where I is finite, then $\cap_{i \in I} U_i$ is open.

3. Arbitrary unions of open sets are open. That is, if $\{U_i : i \in I\}$ is a collection of open sets where I is finite or infinite, then $\cup_{i \in I} U_i$ is open.

Proof. To prove (2) let $p \in \cap_{i \in I} U_i$. We must find $r > 0$ such that $B(p,r) \subset \cap_{i \in I} U_i$. For each i there exists $r_i > 0$ such that $B(p, r_i) \subset U_i$. Since the index set I is finite, $r = \min_{i \in I} r_i > 0$. Moreover, $B(p,r) \subset \cap_{i \in I} U_i$ (check). The proof of (3) is similar but even easier (do it). □

Be sure to note how the finiteness is used in the proof of (2).

Example 8.10. *Here's an example of a countably infinite collection of open intervals whose intersection is not open:* $\{(-1/n, 1/n) : n \in \mathbb{N}\}$. *It is easy to give similar examples in* $\mathbb{R}^n$.

Thus, the finiteness assumption in part (2) of proposition 8.9 is really *needed. Although the proof of (2) clearly* used *finiteness, we* must *have examples like this to be sure that finiteness is necessary in (2). Without such examples, we have to recognize the possibility that our proof is suboptimal in the sense that it used something we didn't really have to use. With the examples that possibility is ruled out.*

Note that part 1 of proposition 8.9 implies that the sets X and the empty set ϕ are each *both* open and closed. For example, $X^c = \phi$, which is open, so X is closed.

Corollary 8.11. *1. In (X,d) arbitrary intersections of closed sets are closed.*
2. Finite unions of closed sets are closed.

Proof. To prove (1) we use the fact that $(\cap_{i \in I} S_i)^c = \cup_{i \in I} S_i^c$, for any index set I and any sets $S_i \subset X$. If the S_i are closed this (together with Proposition 8.9, part (3)) shows that the complement of their intersection is open, and thus their intersection is closed. The proof of (2) is very similar and uses $(\cup_{i \in I} S_i)^c = (\cap_{i \in I} S_i^c)$. □

Example 8.12. *Here is an example of a countable union of closed sets that is not closed:* $\cup_{n=1}^{\infty} [3 + \frac{1}{n}, 6 - \frac{1}{n}]$. *Thus, the finiteness assumption in part (2) of the corollary is really needed.*

Now we show that balls are open.

Proposition 8.13. *Let $p \in (X,d)$ and $r > 0$. Then $B(p,r)$ is open.*

Proof. Let $q \in B(p,r)$. We must find $r' > 0$ such that $B(q, r') \subset B(p,r)$. I claim $r' = r - d(p,q)$ works.[20] To see this we must show that if $t \in B(q, r')$, then $t \in B(p,r)$. Well,

$$d(t,p) \leq d(t,q) + d(q,p) < r' + d(q,p) = (r - d(p,q)) + d(p,q) = r,$$

so $t \in B(p,r)$. □

So from now on we refer to $B(p,r)$ as the *open* ball of radius r in X.

The next proposition shows how open sets in (F,d) are related to open sets in (X,d) when $F \subset X$.

Proposition 8.14. *Let (X,d) be a metric space and suppose $F \subset X$. Then a set U is open in (F,d) if and only if U can be written $U = F \cap \mathcal{O}$ for some open set $\mathcal{O} \subset X$.*

The proof follows easily from the observations that for any $p \in F$ and $r > 0$, we have $B_F(p,r) = F \cap B(p,r)$, and that in any metric space an open set can always be written as the union of a family of open balls.

We can use open balls to characterize convergence of sequences in any metric space.

Proposition 8.15. *Let (p_n) be a sequence in (X,d) and let $p \in X$. Then $p_n \to p$ if and only if given any $\epsilon > 0$ there exists an $N = N(\epsilon)$ such that $n \geq N$ implies $p_n \in B(p, \epsilon)$.*

Proof. This is immediate from Definition 8.3(b) and the fact that $q \in B(p,\epsilon) \Leftrightarrow d(q,p) < \epsilon$. □

[20] Draw a picture of these balls in the case of $\mathbb{R}^2$.

8.2 Closedness in terms of sequences.

Next we give an alternative characterization of closed sets.

Proposition 8.16. *Let (E, d) be a metric space and suppose $S \subset E$. Then S is closed if and only if whenever $p_n \in S$ and $p_n \to p \in E$, then $p \in S$.*

Proof. ($\Rightarrow$) Suppose S is closed, $p_n \in S$, and $p_n \to p \in E$. If $p \notin S$ there exists $\epsilon > 0$ such that $B(p, \epsilon) \subset S^c$. Since $p_n \to p$ there exists N such that $n \geq N$ implies $p_n \in B(p, \epsilon) \subset S^c$, a contradiction.

($\Leftarrow$). We prove the contrapositive. Suppose S is not closed. We will exhibit a sequence p_n in S such that $p_n \to p$, but $p \notin S$. If S is not closed, then S^c is not open. Thus, there exists a point $p \in S^c$ such that for all $\epsilon > 0$, $B(p, \epsilon)$ contains points in S. So for each $n \in \mathbb{N}$ there exists a point $p_n \in S$ such that $p_n \in B(p, \frac{1}{n})$. Thus, $p_n \to p$ and $p_n \in S$ for all n, but $p \notin S$.

□

Remark 8.17. *This proposition shows that the convergent sequences in a metric space completely determine the open sets (since the convergent sequences completely determine the closed sets). More precisely, the direction ($\Leftarrow$) shows that the convergent sequences determine the closed sets. We showed in Proposition 8.15 that convergence of sequences can be defined using open sets, so we already knew that the open sets determine the convergent sequences in a metric space.*

8.3 The closure, interior, and boundary of a set $S \subset (X, d)$.

Suppose $S \subset (X, d)$. The *closure* of S, denoted $\overline{S}$, is the smallest closed set in X that contains S. Equivalently,

$$\overline{S} = \{p \in X : \text{ there exists a sequence of points } (x_n) \text{ in } S \text{ such that } x_n \to p\}. \tag{8.2}$$

The *interior* of S, denoted $\mathring{S}$, is the largest open set of X contained in S. Equivalently,

$$\mathring{S} = \{p \in S : \text{ for some } \delta > 0, B(p, \delta) \subset S\}. \tag{8.3}$$

The *boundary* of S, denoted bS, is $\overline{S} \setminus \mathring{S}$. Equivalently,[21]

$$bS = \{p \in X : \text{ for any } \delta > 0, B(p, \delta) \text{ meets both } S \text{ and } S^c\}. \tag{8.4}$$

Draw a picture in $\mathbb{R}^2$ of a set S which is the union of finitely many open and closed disks, line segments, and isolated points. Determine $\overline{S}$, $\mathring{S}$, and bS for that set.

8.4 Completeness.

The definition of completeness for metric spaces is the exact analogue of the definition of completeness of $\mathbb{R}$ given in Math 521.

Definition 8.18. *The metric space (X, d) is* complete *provided every Cauchy sequence (p_n) converges to an element $p \in X$.*

In Math 521 you saw that in the case of $\mathbb{R}$, this definition is equivalent to the *least upper bound property*. Of course, the least upper bound property makes no sense in a general metric space (why?).

[21] We say that a set A *meets* B when $A \cap B \neq \emptyset$.

Examples 8.19. *1. The metric space* $\mathbb{Q}$ *(the rationals) with the metric inherited from* $\mathbb{R}$ *is not complete. This was proved in Math 521. (Consider a sequence of rationals converging to* $\sqrt{2}$*, for example.)*

2. The metric space $\mathbb{R}^n$ *with its usual metric is complete. To see why this must be true, first think about the case* $\mathbb{R}^2$*. Suppose* $((x_k, y_k))$ *is a Cauchy sequence. Then each of the sequences* (x_k)*,* (y_k) *must be a Cauchy sequence in* $\mathbb{R}$ *(why?), so we immediately get a candidate for a limit of* $((x_k, y_k))$ *(how?). Try to finish this proof, which immediately generalizes to* $\mathbb{R}^n$*.*

3. The space $C([0,1],\mathbb{R})$ *(often denoted just* $C([0,1])$*, which was discussed in Remark 8.2, is a complete metric space. Proof: Suppose* (f_n) *is a Cauchy sequence of functions in this space. Using the definition of the metric for this space, we see that this is equivalent to saying that the sequence* (f_n) *is* uniformly Cauchy on $[0,1]$ *in the sense you defined in Math 521. So we can apply the Math 521 result that such a sequence converges uniformly to a continuous function* f *on* $[0,1]$*. But this just* means *that* $f_n \to f$ *in the metric space* $C([0,1],\mathbb{R})$ *(since* $f_n \to f$ *uniformly on* $[0,1]$ *if and only if given any* $\epsilon > 0$ *there exists* N *such that* $n \geq N$ *implies* $d(f_n, f) < \epsilon$*, where* $d(f,g) := \sup_{x\in[0,1]} |f(x) - g(x)|$*).*

4. The space $C^1([a,b]) = \{f : [a,b] \to \mathbb{R} : f, f' \text{ are continuous}\}$*, which was defined in the section on vector spaces, is a complete metric space* if *we use the metric*

$$d_1(f,g) := \sup_{x\in[a,b],k=0,1} |f^{(k)}(x) - g^{(k)}(x)|. \tag{8.5}$$

This completeness assertion is again just a rephrasing of a corresponding Math 521 result (which one?).

5. If F *is a closed subset of a complete metric space* (X,d)*, then* (F,d) *is a complete metric space (why?).*

6. See Corollary 8.27 below for another family of examples.

8.5 Compactness

We begin by defining an *open cover* of a set in a metric space (X,d). It's pretty much just what it sounds like.

Definition 8.20. *(a) Let* $S \subset X$*. A family of open sets,* $\mathcal{F} = \{\mathcal{O}_\alpha,\ \alpha \in I\}$*, is an* open cover of S *when* $S \subset \cup_{\alpha\in I}\mathcal{O}_\alpha$*.*

(b) We say that an open cover $\mathcal{F}$ *of* S *can be* reduced to a finite subcover *when there exists a finite subcollection of* $\mathcal{F}$*, say* $\{\mathcal{O}_{\alpha_1}, \ldots, \mathcal{O}_{\alpha_N}\}$ *such that* $S \subset \cup_{i=1}^N \mathcal{O}_{\alpha_i}$*.*

In this definition the original open cover $\mathcal{F}$ can be finite, countably infinite, or uncountably infinite. Next we define compactness in terms of open covers.

Definition 8.21 (Compact set)**.** *A subset* $K \subset X$ *is* compact *when any open cover of* K *can be reduced to a finite subcover. We say that "the metric space* (X,d) *is compact" when any open cover of* X *can be reduced to a finite subcover.*

Pay close attention to the quantifiers in these definitions, "there exists" in the first, "any" in the second.

Examples 8.22 (and counterexamples)**.** *1. In any metric space a finite set* $S = \{x_1, \ldots, x_M\}$ *is compact. Proof: Let* $\mathcal{F}$ *be an open cover, possibly infinite. Each point* x_i *belongs to some element of* $\mathcal{F}$*, say* $x_i \in \mathcal{O}_i \in \mathcal{F}$*. Then* $\{\mathcal{O}_i,\ i = 1, \ldots, M\} \subset \mathcal{F}$ *is a finite subcover.*

2. The open interval $(0,1) \subset \mathbb{R}$ *is not compact. Proof: The open cover* $\{(\frac{1}{n}, 1-\frac{1}{n}), n = 5, \ldots, \infty\}$ *cannot be reduced to a finite subcover. (Why not?)*

3. The metric space $\mathbb{R}$ itself with the usual metric $|x-y|$ is not compact. (You should exhibit an open cover of $\mathbb{R}$ that cannot be reduced to a finite subcover.)

4. In $(\mathbb{R}^n, |x-y|)$ any closed and bounded set is compact. This is a nontrivial theorem, proved in Proposition A.1.13 of our text. It is important to note that this theorem does not *hold in general metric spaces. See Remark 8.23.*

5. The subset $\mathbb{N} \subset (\mathbb{R}, |x-y|)$ is not compact. (You should exhibit an open cover of $\mathbb{N}$ that cannot be reduced to a finite subcover.)

Remark 8.23. *It is not hard to show that in any metric space a compact set must be closed and bounded, but the converse is not necessarily true in a general metric space. The converse* is *true in $(\mathbb{R}^n, |x-y|)$ though. Thus, we have that in $(\mathbb{R}^n, |x-y|)$ a set is compact if and only if it is closed and bounded.*

The following proposition is closely related to Proposition 8.14.

Proposition 8.24. *Let $K \subset (X,d)$. Then K is a compact subset of X if and only if the metric space (K,d) is compact.*

To understand this proposition notice that the statement that "K is a compact subset of X" is a statement about open covers of K using open sets in X, while the statement that "(K,d) is a compact metric space" is a statement about open covers of K using open sets in (K,d). The proof uses Proposition 8.14 and is straightforward.

In metric spaces compactness can be characterized in terms of sequences.

Definition 8.25 (sequential compactness)**.** *A set $K \subset (X,d)$ is said to be* sequentially compact *if any sequence in K has a subsequence that converges to a point of K.*

Theorem 8.26. *A set $K \subset (X,d)$ is compact if and only if it is sequentially compact.*

This nontrivial result is proved in our text (Theorem A.1.10).[22] Some of you probably took sections of Math 521 in which compactness was *defined* for subsets of $\mathbb{R}$ (or $\mathbb{R}^n$) using sequences as in Definition 8.25, without reference to open covers. Theorem 8.26 shows that this is permissible for metric spaces. However, in some topological spaces which are not metric spaces, compactness (defined in terms of open covers) is not equivalent to sequential compactness.[23]

Here is a direct corollary of Theorem 8.26 (prove it).

Corollary 8.27. *A compact metric space is complete.*

8.6 Connectedness

Definition 8.28. *(a) We say that (X,d) is* connected *if X cannot be written as the union of two disjoint, nonempty, open sets.*

(b)If $S \subset X$ we say that S is connected if the metric space (S,d) is connected.

The following Proposition provides an equivalent characterization that is often useful when one is trying to show that some nonempty subset of a metric space is in fact the whole space.

Proposition 8.29. *The metric space (X,d) is connected if and only if the only subsets of X that are both open and closed are X and the empty set $\emptyset$.*

---

[22] Our text refers to this theorem as the Heine-Borel theorem, but I think most authors use that name to refer to the result in Example 8.22 (4).

[23] See Wikipedia for the definition of "topological space". It is a mathematical structure more general than a metric space. In this course we have no need to work with topological spaces that are not metric spaces.

Proof. (⇒) Suppose $A \neq X$, $A \neq \emptyset$ is both open and closed. Then $X = A \cup A^c$ exhibits X as the union of two disjoint, nonempty, open sets.

(⇐) You do this.

□

When a metric space (X, d) can be written $X = A \cup B$, where A, B are disjoint, nonempty, open sets, I will refer to this as a "disconnection" of X.[24]

Examples 8.30. *In most cases connected sets are exactly what you expect them to be.*

1. A subset of $\mathbb{R}$ is connected if and only if it is an interval. A proof is given on p. 391 of our text.

2. In $\mathbb{R}^n$ open and closed balls are connected.

3. If (X, d) has the property that for any $p, q \in X$, there exists a continuous map $\gamma : [0, 1] \to X$ (see Definition 9.3) with $\gamma(0) = p$, $\gamma(1) = q$, then we say X is path connected. A consequence of (1) is that any path-connected metric space is connected. This will become clear after we discuss continuous functions on metric spaces in the next section; in particular, see Proposition 9.7.

9 Functions on metric spaces

Let (X, d_X) and (Y, d_Y) be metric spaces.[25]. Parallel to the definition given in class for real-valued functions defined on $\mathbb{R}^n$, if $f : X \to Y$ we say that

$$\lim_{x \to a} f(x) = L \in Y$$

if and only if given $\epsilon > 0$ there exists $\delta = \delta(\epsilon) > 0$ such that $0 < d(x, a) < \delta$ implies $d(f(x), L) < \epsilon$.

Definition 9.1. *The function $f : X \to Y$ is continuous at $a \in X$ if $\lim_{x \to a} f(x) = f(a)$.*

It is useful to have equivalent formulations of continuity.

Proposition 9.2 (Continuity)**.** *The following properties of $f : X \to Y$ are equivalent:*

a) f is continuous at a.

b) Given any $\epsilon > 0$ there exists a $\delta > 0$ such that $d(x, a) < \delta$ implies $d(f(x), f(a)) < \epsilon$. (Here $\delta = \delta(a, \epsilon)$ may depend on both a and ϵ.)

c) Whenever $x_n \to a$, then $f(x_n) \to f(a)$. (sequential continuity)

d) If $\mathcal{O}$ is any open set containing $f(a)$, then the preimage $f^{-1}(\mathcal{O})$ contains $B(a, \delta)$ for some $\delta > 0$.

Proof. The equivalence of (a), (b), (c) is proved just like in Math 521 (replace $| \cdot - \cdot |$ by $d(\cdot, \cdot)$), and the equivalence of (b) and (d) follows from the definitions of balls and open sets.

□

So continuity is a property that is initially defined *pointwise*. If $f : X \to Y$ we say that f is continuous on X (or just continuous) if f is continuous at every point of X. The following proposition is an immediate consequence of Proposition 9.2(d).

[24] I think this usage is probably nonstandard.

[25] That was probably the last time I'll write d_X, d_Y. Instead I'll use just d for both and rely on the *context* to make clear which metric I'm using. But when the metric space is $(\mathbb{R}, |x - y|)$ or $(\mathbb{R}^n, |x - y|)$, we *should* use $|x - y|$ instead of $d(x, y)$.

Proposition 9.3. *Let $f : X \to Y$. Then f is continuous if and only if for any open set $\mathcal{O} \subset Y$ the preimage $f^{-1}(\mathcal{O})$ is open in X.*

Definition 9.4 (Uniform continuity)**.** *Let $f : X \to Y$. We say f is* uniformly continuous *on X if, given $\epsilon > 0$ there exists $\delta = \delta(\epsilon) > 0$ such that $d(x_1, x_2) < \delta$ implies $d(f(x_1), f(x_2)) < \epsilon$. (Here $\delta = \delta(\epsilon)$ may depend only on ϵ.)*

The next three propositions generalize results for real-valued functions defined on $\mathbb{R}$ (which ones?) and can be proved by essentially the same proofs. The use of sequential compactness gives a "two line" proof of Prop. 9.6 (do it), and the use of Prop. 9.3 yields a short proof of Prop.9.7.

Proposition 9.5. *Let (X, d) be compact and suppose $f : X \to Y$ is continuous. Then f is uniformly continuous on X.*[26]

Proposition 9.6. *Let $f : X \to Y$ be continuous and suppose (X, d) is compact. Then $f(X)$ is compact.*

Proposition 9.7. *Let $f : X \to Y$ be continuous and suppose (X, d) is connected. Then $f(X)$ is connected.*

Remark 9.8. *a) A direct corollary of Proposition 9.6 is the "extreme value theorem" from Calc I: If $f : [a, b] \to \mathbb{R}$ is continuous, then f attains its max and min on $[a, b]$.*

b) A direct corollary of Proposition 9.7 is the "intermediate value theorem" from Calc I: If $f : [a, b] \to \mathbb{R}$ is continuous, then f assumes every value between $f(a)$ and $f(b)$.

c) The result of (a) holds if $[a, b]$ is replaced by any compact metric space, and the result of (b) holds if $[a, b]$ is replaced by any connected metric space.[27]

9.1 Sequences of functions.

Let $f : X \to Y$ and for $n \in \mathbb{N}$ let $f_n : X \to Y$.

Definition 9.9. *(a) The sequence (f_n) converges* pointwise *to f if for each $x \in X$ we have $f_n(x) \to f(x)$. In other words, given any $x \in X$ and $\epsilon > 0$, there exists $N = N(x, \epsilon) \in \mathbb{N}$ such that $n \geq N$ implies $d(f_n(x), f(x)) < \epsilon$.*

(b) The sequence (f_n) converges uniformly *to f if: given any $\epsilon > 0$, there exists $N = N(\epsilon) \in \mathbb{N}$ such that $n \geq N$ implies $d(f_n(x), f(x)) < \epsilon$ for* all *$x \in X$.*

The following result demonstrates that uniform convergence is relevant to questions involving the interchange of two limits.

Proposition 9.10. *Suppose $f_n : X \to Y$ are continuous and $f_n \to f$ uniformly on X. Then $f : X \to Y$ is continuous.*

The proof is by the same "$\epsilon/3$ argument" using the triangle inequality that was used for the parallel result in Math 521 (do it). Recall that the *pointwise* limit of a sequence of continuous functions is not necessarily continuous as the example of $f_n(x) = x^n$ on $[0, 1]$ illustrates. A good exercise at this point is to use the sequence (x^n) on $[0, 1]$ to show that $C([0, 1], \mathbb{R})$ is a complete metric

[26] A slightly less general version of this is proved in Prop. A.1.16 of our text.

[27] We use this extension of the extreme value theorem in our application of Taylor's theorem to local minima. There we take (X, d) to be $(S^{n-1}, |x - y|)$.

space in which some closed and bounded sets are not compact; indeed, in this space the closed unit ball is not compact.[28]

Suppose now that X is a compact metric space and Y is a complete metric space. We let

$$C(X,Y) = \{f : X \to Y : f \text{ is continuous}\}. \tag{9.1}$$

The standard way to make $C(X,Y)$ a metric space is to define its metric D by

$$D(f,g) = \sup_{x\in X} d(f(x), g(x)). \tag{9.2}$$

You should check D is a metric, and note that if X is not compact, D is not in general a metric (why?).

Remark 9.11. *Note that $C(X,Y)$ is not in general a vector space, since there may be no way to add elements of Y or multiply them by scalars. But if Y is a vector space, then $C(X,Y)$ is a vector space.*

The next proposition generalizes Example 8.19 (3).

Proposition 9.12. *If X is compact and Y is complete, then $(C(X,Y), D)$ is a complete metric space.*

A good way to see how to prove this is to revisit the proof that $C([0,1],\mathbb{R})$ is complete, and review how the compactness of $[0,1]$ and the completeness of $\mathbb{R}$ were used in that proof. The proof of Proposition 9.12 is given in Prop. A.1.17 of our text. To make sure you understand this proposition and the concept of uniform convergence, you should reformulate Prop. 9.12 in the language of uniform convergence of sequences of functions.

As another application of uniform convergence we first recall the following Math 521 result.

Proposition 9.13. *Let $I \subset \mathbb{R}$ be an open interval and suppose $f_n : I \to \mathbb{R}$, $n \in \mathbb{N}$ are C^1 functions. Let $f, g : I \to \mathbb{R}$ and suppose that given any compact subset $K \subset I$ the sequences f_n and f_n' converge uniformly on K to f and g respectively. Then f is C^1 and $g = f'$.*

Here is an extension of this result to functions defined on $\mathbb{R}^n$.

Proposition 9.14. *Let $\mathcal{O} \subset \mathbb{R}^n$ be open and suppose $f_n : \mathcal{O} \to \mathbb{R}$, $n \in \mathbb{N}$ are C^1 functions. For $i = 1, \dots, n$ let $f, g_i : \mathcal{O} \to \mathbb{R}$ and suppose that given any compact subset $K \subset \mathcal{O}$ the sequences f_n and $\partial_i f_n$ converge uniformly on K to f and g_i respectively for all i. Then f is C^1 and $g_i = \partial_{x_i} f$ for all i.*

Finally, let's recall a Math 521 result relating uniform convergence to integrals.

Proposition 9.15. *Let $f_n : [a,b] \to \mathbb{R}$, $n \in \mathbb{N}$, be continuous. Suppose $f_n \to f$ uniformly on $[a,b]$. Then f is continuous and*

$$\lim_{n\to\infty} \int_a^b f_n(t)dt = \int_a^b f(t)dt.$$

Remark 9.16. *Each of Propositions 9.10, 9.14, and 9.15 states conditions under which a particular pair of "limit processes" can be interchanged (which ones?). In each case there are simple examples showing that the conclusion is generally false if the assumption of uniform convergence is replaced by pointwise convergence.*

[28] It is much easier to give an example of a metric space which is *not* complete in which some closed, bounded sets are not compact.

10 Dual spaces and multilinear forms

In this section we present some background that will be important for our later study of differential forms.

Definition 10.1. *Let V be a vector space. The* dual space *of V, denoted V', is the set of all linear maps $f : V \to \mathbb{R}$. The elements of V' are often referred to as* linear functionals *on V.*

It is clear that V' is a vector space (why?). Moreover, if V is n-dimensional, so is V'.

Proposition 10.2. *If $\dim V = n$, then $\dim V' = n$.*

Proof. Let $B = \{v_1, \ldots, v_n\}$ be a basis of V. For $i = 1, \ldots, n$ let $\omega_i \in V'$ be the linear functional satisfying

$$\omega_i(v_j) = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{if } j \neq i \end{cases}, \qquad \text{or equivalently, } \omega_i(v_j) = \delta_{i,j}. \tag{10.1}$$

We show now that $B' := \{\omega_1, \ldots, \omega_n\}$ is a basis of V'.

Suppose first that

$$\sum_i c_i \omega_i = 0. \tag{10.2}$$

Evaluating both sides at v_k we get $c_k = 0$ for all k. Thus, B' is linearly independent. To show $\operatorname{span} B' = V'$ let $f \in V'$ and set $c_j = f(v_j)$ for all j. Then $f = \sum_k c_k \omega_k$ (why?).

□

Definition 10.3 (Dual basis)**.** *We refer to the basis $B' = \{\omega_j, j = 1, \ldots, n\}$ defined by* (10.1) *as the* dual basis *of $B = \{v_1, \ldots, v_n\}$.*

Next we define bilinear forms.

Definition 10.4. *Let V be any vector space. A function $\alpha : V \times V \to \mathbb{R}$ is called* a bilinear form *if it is linear in each argument when the other is held fixed. That is, for $u, v, w \in V$ and $c \in \mathbb{R}$ we have*

$$\alpha(cu + v, w) = c\alpha(u, w) + \alpha(v, w) \text{ and } \alpha(w, cu + v) = c\alpha(w, u) + \alpha(w, v). \tag{10.3}$$

Notations 10.5. *We let $\mathcal{T}^2(V)$ denote the set of all bilinear forms $\alpha : V \times V \to \mathbb{R}$.*

Example 10.6. *If $\phi, \omega \in V'$ then the map $\phi \otimes \omega : V \times V \to \mathbb{R}$ defined by*

$$\phi \otimes \omega(v, w) := \phi(v)\omega(w) \text{ for } v, w \in V \tag{10.4}$$

is a bilinear form (why?).

Proposition 10.7. *Suppose $\dim V = n$. Then $\dim \mathcal{T}^2(V) = n^2$.*

Proof. Let $B = \{v_1, \ldots, v_n\}$ be a basis of V and $B' = \{\omega_1, \ldots, \omega_n\}$ its dual basis. Let $\omega_j \otimes \omega_k$ be the element of $\mathcal{T}^2(V)$ defined by[29]

$$\omega_j \otimes \omega_k(v_p, v_q) = \omega_j(v_p)\omega_k(v_q) \text{ for all } j, k, p, q. \tag{10.5}$$

[29]Check that this condition determines a well-defined element of $\mathcal{T}^2(V)$.

We will show $B'' = \{\omega_j \otimes \omega_k : j,k = 1,\dots,n\}$ is a basis of $\mathcal{T}^2(V)$.

Suppose first that $\sum_{j,k} c_{j,k}\omega_j \otimes \omega_k = 0$. Evaluate at (v_l, v_m) to get $c_{l,m} = 0$. Thus B'' is linearly independent. Next suppose that $\alpha \in \mathcal{T}^2(V)$ and set $c_{m,n} := \alpha(v_m, v_n)$. Then $\alpha = \sum_{i,j} c_{i,j}\omega_i \otimes \omega_j$ (why?).[30] So span $B'' = \mathcal{T}^2(V)$.

□

Examples 10.8. *1)The dot product $D : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by $D(a,b) = a \cdot b$ is bilinear.*

2) Let A be a real $n \times n$ matrix. Then the map $\alpha : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $\alpha(x,y) := Ax \cdot y$ is a bilinear (check). Observe that $A = (a_{i,j})$, where $a_{i,j} = Ae_j \cdot e_i = \alpha(e_j, e_i)$.

3) If $a, b \in \mathbb{R}^2$ and $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(a,b) := \det(a\ b),$$

where $(a\ b)$ is the matrix whose columns are a and b, then f is bilinear.[31]

Remark 10.9. *Observe that a bilinear form on $V \times V$ is generally* not *a linear map of $V \times V$ into $\mathbb{R}$.*

10.1 Multilinear $k-$forms

Suppose $k \in \mathbb{N}$ and let V^k be the cartesian product of k copies of V. A multilinear $k-$form on V is simply a function $\alpha : V^k \to \mathbb{R}$ that is linear in each argument when the others are held fixed. Thus, a bilinear form is a multilinear $2-$form. We write $\mathcal{T}^k(V)$ for the vector space of all multilinear $k-$forms on V. We have $\mathcal{T}^1(V) = V'$. By convention we set $\mathcal{T}^0(V) := \mathbb{R}$, the scalars, which we can think of as constant functions on V. [32].

Generalizing Example 10.8, we see that the determinant of a real $k \times k$ matrix defines an element of $\mathcal{T}^k(\mathbb{R}^k)$.

The proof of the following proposition is easily obtained by mimicking the proof of Proposition 10.7.

Proposition 10.10. *If* $\dim V = n$, *then* $\dim \mathcal{T}^k(V) = n^k$. *With the usual notation a basis of $\mathcal{T}^k(V)$ is given by*

$$\{\omega_{i_1} \otimes \cdots \otimes \omega_{i_k} : 1 \le i_1, \dots, i_k \le n\}. \tag{10.6}$$

Definition 10.11. *If $\alpha \in \mathcal{T}^p(V)$ and $\beta \in \mathcal{T}^q(V)$ we can define the (tensor) product $\alpha \otimes \beta \in \mathcal{T}^{p+q}(V)$ in the obvious way:*

$$\alpha \otimes \beta(v,w) := \alpha(v)\beta(w) \text{ for } v = (v_1,\dots,v_p) \in V^p,\ w = (w_1,\dots,w_q) \in V^q. \tag{10.7}$$

This generalizes Example 10.6.

The product $\otimes$ is clearly associative and distributive over addition, but not commutative. We next define the *pullback* operation on multilinear $k-$forms.

[30] Answer: Each side of the equation is a bilinear form, so each side is determined by its action on arguments of the form (v_m, v_n). The two sides clearly agree on such arguments.

[31] We have not given our "official" definition of the determinant yet; see Definition 10.22 for that. Until then I rely on your knowledge of determinants from previous courses.

[32] The $\mathcal{T}$ in $\mathcal{T}^k(V)$ stands for "tensor". Multilinear $k-$forms are also referred to as (covariant) $k-$tensors, and the "product" $\otimes$ appearing in (10.6) and below is called a "tensor product".

Definition 10.12 (Pullbacks)**.** *Let $A : V \to W$ be a linear transformation and suppose $\beta \in \mathcal{T}^k(W)$. Then we define $A^*\beta \in \mathcal{T}^k(V)$ by*

$$A^*\beta(v_1, \ldots, v_k) := \beta(Av_1, \ldots, Av_k). \tag{10.8}$$

Remark 10.13. *Clearly $A^* : \mathcal{T}^k(W) \to \mathcal{T}^k(V)$ is linear. Also, it follows immediately from the definitions that*

$$A^*(\alpha \otimes \beta) = A^*\alpha \otimes A^*\beta. \tag{10.9}$$

This fact is of fundamental importance for the theory of differential forms.

10.2 Alternating multilinear $k-$forms.

A multilinear $k-$form α is *alternating* if the sign of α is reversed whenever two arguments are transposed:

$$\alpha(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_k) = -\alpha(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_k). \tag{10.10}$$

It will be useful to rephrase this definition. Let S_k denote the group of permutations of the set $\{1, 2, \ldots, k\}$; note that S_k has $k!$ elements. A permutation $\sigma \in S_k$ is any bijective map

$$\sigma : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}. \tag{10.11}$$

If $\sigma, \tau \in S_k$ the composition $\sigma \circ \tau \in S_k$ and the inverse $\sigma^{-1} \in S_k$ are defined in the usual way for functions. We say that a permutation is *even* or *odd* if it can be expressed as the product (i.e., composition) of an even or odd number of transpositions.[33] For example, the permutation σ defined by $(1, 2, 3, 4) \to (2, 4, 1, 3)$ is odd.[34]. Let us define the *sign of the permutation* σ by:[35]

$$(-1)^\sigma = \begin{cases} 1, & \text{if } \sigma \text{ is even} \\ -1, & \text{if } \sigma \text{ is odd} \end{cases} . \tag{10.12}$$

Observe that (why?)[36]

$$(-1)^{\sigma \circ \tau} = (-1)^\sigma (-1)^\tau, \text{ so } \operatorname{sgn} \sigma = \operatorname{sgn} \sigma^{-1}. \tag{10.13}$$

If α is a multilinear $k-$form on V^k and $\sigma \in S_k$, define another multilinear $k-$form by

$$\alpha^\sigma(v_1, \ldots, v_k) := \alpha(v_{\sigma(1)}, \ldots, v_{\sigma(k)}). \tag{10.14}$$

Then it's not hard to check that the alternating multilinear $k-$forms are those satisfying

$$\alpha^\sigma = (-1)^\sigma \alpha \text{ for all } \sigma \in S_k. \tag{10.15}$$

Notations 10.14. *We denote by $\Lambda^k(V) \subset \mathcal{T}^k(V)$ the subspace of alternating multilinear $k-$forms.*

[33] It is not immediately obvious (to me at least) that the notion of an even or odd permutation is well-defined. How do we know that a given permutation might not be expressible in one way as the product of an even number of transpositions and in another way as the product of an odd number of transpositions? This point is discussed in our text just below (1.4.21).

[34] Here $\sigma(1) = 2$, $\sigma(2) = 4$, etc.. An example of a transposition is the element of S_4 given by $(1, 2, 3, 4) \to (4, 2, 3, 1)$.

[35] Some authors write $\operatorname{sgn} \sigma := (-1)^\sigma$.

[36] Answer: Count transpositions.

Example 10.15. *We have seen that the determinant of a real $k \times k$ matrix defines an element of $\mathcal{T}^k(\mathbb{R}^k)$. This element is alternating; interchanging two columns of a $k \times k$ matrix causes the determinant to change sign.*

Given any element of $\mathcal{T}^k(V)$ there is a simple way to "convert it" into an *alternating* multilinear $k-$form. If $\alpha \in \mathcal{T}^k(V)$, we define a linear transformation $\text{Alt} : \mathcal{T}^k(V) \to \Lambda^k(V)$ by

$$\text{Alt } \alpha := \sum_{\sigma \in S_k} (-1)^\sigma \alpha^\sigma. \tag{10.16}$$

Let's check that Alt α is alternating in a special case. Suppose $\alpha = \omega_i \otimes \omega_j \in \mathcal{T}^2(V)$, where $\omega_i, \omega_j \in \Lambda^1(V)$ are as in Definition 10.3. Then

$$\text{Alt } (\omega_i \otimes \omega_j) = \omega_i \otimes \omega_j - \omega_j \otimes \omega_i. \tag{10.17}$$

The first term on the right corresponds to the identity permutation (even) and the second to the permutation $(1, 2) \to (2, 1)$ (odd). The right side of (10.17) is clearly alternating (why?) At this point it would be a good idea for you to write Alt $(\omega_p \otimes \omega_q \otimes \omega_r)$ explicitly in a form similar to (10.17), and check that it is alternating. Look at what happens if two of the indices are equal.

If α is alternating to begin with, we see that

$$\text{Alt } \alpha = k!\alpha, \tag{10.18}$$

since every term on the right in (10.16) equals α. Some authors include a factor of $\frac{1}{k!}$ on the right in (10.16) so that one obtains Alt $\alpha = \alpha$ instead. We will not do that, since we want $\omega_i \wedge \omega_j(v_i, v_j) = 1$; see Notations 10.16.

Notations 10.16. *The right side of* (10.17) *is often written as $\omega_i \otimes \omega_j - \omega_j \otimes \omega_i := \omega_i \wedge \omega_j$. This is our first example of a* wedge product.

10.3 Wedge products

We would like to have a product that takes $\alpha \in \Lambda^p(V)$ and $\beta \in \Lambda^q(V)$ and produces an element of $\Lambda^{p+q}(V)$. Unfortunately, $\alpha \otimes \beta$ does not work. It's in $\mathcal{T}^{p+q}(V)$, but it is not alternating in general. Instead we have:

Definition 10.17. *For $\alpha \in \Lambda^p(V)$ and $\beta \in \Lambda^q(V)$ we define the* wedge product $\alpha \wedge \beta$ *by*

$$\alpha \wedge \beta = \text{Alt } (\alpha \otimes \beta). \tag{10.19}$$

The linearity of Alt implies that $\wedge$ distributes over addition. It is more difficult to check that this product is associative:

Proposition 10.18. *Let α, β, γ be alternating multilinear forms of degrees p, q, and r respectively. Then*

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma) = \text{Alt}(\alpha \otimes \beta \otimes \gamma) \in \Lambda^{p+q+r}(V). \tag{10.20}$$

For the proof we refer to [GP], p. 156, (or many other sources). This proposition allows us to write $\alpha \wedge \beta \wedge \gamma$ without ambiguity. Moreover, the formula $\alpha \wedge \beta \wedge \gamma = \text{Alt } (\alpha \otimes \beta \otimes \gamma)$ immediately extends to any number of factors.

Remark 10.19. *a) The wedge product is* not *commutative. For example, it is anticommutative on elements of* $\Lambda^1(V)$*. If* $\alpha, \beta \in \Lambda^1(V)$*, the computation* (10.17) *shows that*

$$\alpha \wedge \beta = -\beta \wedge \alpha \text{ and } \alpha \wedge \alpha = 0. \tag{10.21}$$

b) The following analogue of (10.9) *holds. Let* $\alpha \in \Lambda^p(V)$*,* $\beta \in \Lambda^q(V)$ *and suppose* $A : V \to V$ *is a linear transformation. Then*

$$A^*(\alpha \wedge \beta) = A^*\alpha \wedge A^*\beta. \tag{10.22}$$

In the special case where $\alpha, \beta \in \Lambda^1(V)$*, this is immediate from* (10.9) *and* (10.17) *(why?). The general case has essentially the same proof. Note that* (10.22) *extends immediately to any number of factors, since* $\wedge$ *is associative.*

We now have the tools to find a basis for $\Lambda^k(V)$. First some notation:

Notations 10.20. *a) Fix* $n, k \in \mathbb{N}$*. Let* $\mathcal{I}_k$ *denote the set of all* $k-$*tuples* $I = (i_1, \ldots, i_k)$*, where each* $i_p \in \{1, \ldots, n\}$*.*

b) Suppose $\dim V = n$ *and let* $B' = \{\omega_i, i = 1, \ldots, n\}$ *be a basis of* $V' = \Lambda^1(V)$ *as in Definition 10.3. Then for* $I = (i_1, \ldots, i_k) \in \mathcal{I}_k$ *we set*

$$\omega_{I,\otimes} := \omega_{i_1} \otimes \omega_{i_2} \otimes \cdots \otimes \omega_{i_k} \in \mathcal{T}^k(V). \tag{10.23}$$

c) Let $\omega_I = \mathrm{Alt}\, \omega_{I,\otimes} = \omega_{i_1} \wedge \omega_{i_2} \wedge \cdots \wedge \omega_{i_k} \in \Lambda^k(V)$*.*

d) b) If $k \leq n$*, let* $\mathcal{I}_{k,\nearrow} \subset \mathcal{I}_k$ *denote the subset of* $k-$*tuples* I *satisfying* $i_1 < i_2 \cdots < i_k$*.*

Observe that the cardinality of $\mathcal{I}_k$, denoted $|\mathcal{I}_k|$, equals n^k, while $|\mathcal{I}_{k\nearrow}| = \binom{n}{k} := \frac{n!}{k!(n-k)!}$. (why?)

Proposition 10.21. *Suppose* $\dim V = n$*. Fix* $k \in \mathbb{N}$ *with* $k \leq n$*. Then* $\dim \Lambda^k(V) = \binom{n}{k}$ *and a basis of* $\Lambda^k(V)$ *is given by* $\{\omega_I : I \in \mathcal{I}_{k\nearrow}\}$*. If* $k > n$*, we have* $\Lambda^k(V) = \{0\}$*.*

Proof. **1.** Let $k \in \mathbb{N}$. First, we claim that span $\{\omega_I : I \in \mathcal{I}_k\} = \Lambda^k(V)$. To see this let $\alpha \in \Lambda^k(V) \subset \mathcal{T}^k(V)$. Then by Proposition 10.10 we can write

$$\alpha = \sum_{I \in \mathcal{I}_k} c_I \omega_{I,\otimes} \text{ for some } c_I \in \mathbb{R}. \tag{10.24}$$

Now apply Alt to both sides, and use (10.18) and Alt $\omega_{I,\otimes} = \omega_I$ to get

$$k!\alpha = \sum_{I \in \mathcal{I}_k} c_I \omega_I. \tag{10.25}$$

This establishes the claim.

2. Now suppose $k \leq n$ and that $I, J \in \mathcal{I}_k$ are the same except for their orderings. Then repeated application of (10.21) implies that $\omega_I = \pm\omega_J$ (one or the other), and that if any two indices of I are the same then $\omega_I = 0$. Combining "redundant" terms and eliminating zero terms from (10.25) we have

$$k!\alpha = \sum_{I \in \mathcal{I}_k} c_I \omega_I = \sum_{I \in \mathcal{I}_{k\nearrow}} d_I \omega_I \text{ for some new } d_I. \tag{10.26}$$

This shows $\mathcal{B} := \{\omega_I : I \in \mathcal{I}_{k\nearrow}\}$ spans $\Lambda^k(V)$.

3. It remains to show $\mathcal{B}$ is linearly independent. Suppose $\sum_{I \in \mathcal{I}_{k\nearrow}} c_I \omega_I = 0$. For $J \in \mathcal{I}_{k\nearrow}$ and v_i as in Definition 10.3, evaluate the left side at $v_J = (v_{j_1}, \dots, v_{j_k})$ to get $c_J = 0$ (why?).[37]

4. If $k > n$ and $I = (i_1, \dots, i_k) \in \mathcal{I}_k$, then at least one index i_p must be repeated, so $\omega_I = 0$. Thus, step **1** implies $\Lambda^k(V) = 0$ if $k > n = \dim V$.

∎

From Proposition 10.21 we see that $\dim \Lambda^n(\mathbb{R}^n) = 1$, since $\binom{n}{n} = 1$. Thus, any two alternating multilinear $n-$forms on $\mathbb{R}^n$ must be scalar multiples of each other. In particular, there is only one $\alpha \in \Lambda^n(\mathbb{R}^n)$ with the property that $\alpha(e_1, e_2, \dots, e_n) = 1$, where $\{e_i, i = 1, \dots, n\}$ is the standard basis of $\mathbb{R}^n$. This leads us to our "official" definition of the determinant.

Definition 10.22 (Determinant). *Let $\{e_i, i = 1, \dots, n\}$ be the standard basis of $\mathbb{R}^n$. We denote by* $\det$ *(the determinant) the unique element of $\Lambda^n(\mathbb{R}^n)$ such that* $\det(e_1, e_2, \dots, e_n) = 1$.

Let $B' = \{\omega_i, i = 1, \dots, n\}$ be the dual basis of the standard basis of $\mathbb{R}^n$.

Proposition 10.23. $\det = \omega_1 \wedge \omega_2 \wedge \cdots \wedge \omega_n$.

Proof. The right side belongs to $\Lambda^n(\mathbb{R}^n)$ and gives 1 when applied to $(e_1, \dots, e_n)$. ∎

We can now derive a classical formula for $\det(a_1, \dots, a_n)$, where $a_j \in \mathbb{R}^n$. Let us write $a_j = (a_{1j}, \dots, a_{nj})$, as is natural if we think of a_j as the $j-$th column of an $n \times n$ matrix.

Proposition 10.24. $\det(a_1, \dots, a_n) = \sum_{\sigma \in S_n} (-1)^\sigma a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)}$.

Proof. We have

$$\begin{aligned} &\det(a_1, \dots, a_n) = \omega_1 \wedge \cdots \wedge \omega_n(a_1, \dots, a_n) = \mathrm{Alt}(\omega_1 \otimes \cdots \otimes \omega_n)(a_1, \dots, a_n) = \\ &\sum_{\sigma \in S_n} (-1)^\sigma (\omega_1 \otimes \cdots \otimes \omega_n)^\sigma (a_1, \dots, a_n) = \sum_{\sigma \in S_n} (-1)^\sigma (\omega_1 \otimes \cdots \otimes \omega_n)(a_{\sigma(1)}, \dots, a_{\sigma(n)}) = \\ &\sum_{\sigma \in S_n} (-1)^\sigma a_{1\sigma(1)} a_{2\sigma(2)} \dots a_{n\sigma(n)}. \end{aligned} \tag{10.27}$$

∎

We conclude this section with a proposition that, together with the change of variable formula for Riemann integrals, turns out to provide the basis for a natural definition of $\int_M \alpha$, when α is a differential $n-$form (to be defined) on an $n-$dimensional manifold M (to be defined).

Proposition 10.25. *Suppose* $\dim V = n$, $B = \{v_1, \dots, v_n\}$ *is a basis of V, and $B' = \{\omega_1, \dots, \omega_n\}$ the dual basis of $\Lambda^1(V) = V'$. Suppose $T : V \to V$ is a linear transformation. Then*[38]

$$T^*(\omega_1 \wedge \cdots \wedge \omega_n) = (\det T)\omega_1 \wedge \cdots \wedge \omega_n. \tag{10.28}$$

Proof. **1.** For each j write

$$T v_j = \sum_i a_{ij} v_i, \tag{10.29}$$

[37] Answer: Every term in the sum defining $\omega_J(v_J) = (\mathrm{Alt}\ \omega_{J,\otimes})(v_J)$ is zero except the term corresponding to the identity permutation, which is one. If $J' \in \mathcal{I}_{k\nearrow}$ is not equal to J, then $\omega_{J'}(v_J) = 0$ by (10.1). If this does not seem clear, I suggest you look at the case $k = 2$ first and use (10.17).

[38] Here we use the definition of $\det T$ given in HW2.

and observe that[39]

$$T^*\omega_j = \sum_i a_{ji}\omega_i. \tag{10.30}$$

2. Using (10.22), we have[40]

$$\begin{aligned} &T^*(\omega_1\wedge\cdots\wedge\omega_n) = T^*\omega_1\wedge\cdots\wedge\cdots T^*\omega_n = \\ &\left(\sum_i a_{1i}\omega_i\right)\wedge\cdots\wedge\left(\sum_i a_{ni}\omega_i\right) = \\ &\sum_{i_1,\dots,i_n} a_{1i_1}\cdots a_{ni_n}\ \omega_{i_1}\wedge\cdots\wedge\omega_{i_n} = \\ &\left(\sum_{\sigma\in S_n}(-1)^\sigma a_{1\sigma(1)}\cdots a_{n\sigma(n)}\right)\ \omega_1\wedge\cdots\wedge\omega_n = (\det T)\omega_1\wedge\cdots\wedge\omega_n. \end{aligned} \tag{10.31}$$

□

10.4 Orientation of a vector space.

From Math 233 (and possibly also from physics courses) you may remember defining "orientations" for "coordinate systems" or for "surfaces" using "right-hand rules".[41] In higher dimensions our right hands are not adequate to the task, so we need other ways to define orientations. An elegant way is to use the fact that $\dim\Lambda^k(V) = \binom{k}{k} = 1$ when $\dim V = k$.

When $\dim V = k$, we can define an equivalence relation on $\Lambda^k(V)\setminus\{0\}$ by declaring $\alpha\sim\beta$ when α is a *positive* scalar multiple of β. If $\gamma\in\Lambda^k(V)\setminus\{0\}$ is a given fixed element, we can then write

$$\Lambda^k(V)\setminus\{0\} = \Lambda^k_+(V)\cup\Lambda^k_-(V), \tag{10.32}$$

where we declare that $\Lambda^k_+(V)$ consists of all β such that $\beta\sim\gamma$, and $\Lambda^k_-(V)$ consists of all β such that $\beta\sim-\gamma$. If we replace γ by some other $\tilde\gamma\in\Lambda^k(V)$ in this definition, $\Lambda^k_\pm(V)$ will be *unchanged* as long as $\tilde\gamma\sim\gamma$. Otherwise the new $\Lambda^k_+(V)$ equals the old $\Lambda^k_-(V)$.

Definition 10.26 (Orientation of V)**.** *Each of the equivalence classes $\Lambda^k_+(V)$, $\Lambda^k_-(V)$ defined above is said to be* an orientation of V. *Any element $\alpha\in\Lambda^k_+(V)$ (resp. $\Lambda^k_-(V)$) is said to determine the* positive orientation *(resp.* negative orientation*).*

Remark 10.27. *We see that a vector space V has two possible orientations. It is arbitrary which one we designate as "positive", but* after *we make that choice it* is *sometimes important to be able to determine whether any given $\alpha\in\Lambda^k(V)\setminus\{0\}$ determines the positive or negative orientation; see Proposition 10.28.*

To make the above definition more concrete, let's take a basis $B = \{v_1,\dots,v_n\}$ of V and $B' = \{\omega_1,\dots,\omega_n\}$ it's dual basis. Next we declare that $\omega_1\wedge\cdots\wedge\omega_n$ determines the positive orientation. This completes the process of orienting V, because the equivalence classes $\Lambda^k_\pm(V)$ are clearly determined.

[39] To see this, apply both sides of (10.30) to v_i, and use the definition of T^* and (10.29).
[40] You should justify each equality in (10.31).
[41] Apologies for all the quotes-it's because I have not defined these terms yet.

Here is an equivalent way that is often used to specify *the same* orientation of V. First observe that

$$\omega_1 \wedge \cdots \wedge \omega_n(v_1, v_2, \ldots, v_n) = 1, \tag{10.33}$$

and that

$$\omega_1 \wedge \cdots \wedge \omega_n(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}) = \begin{cases} 1 \text{ when } (-1)^\sigma = 1 \\ -1 \text{ when } (-1)^\sigma = -1 \end{cases}. \tag{10.34}$$

In view of (10.34) we say that the *ordered basis* $\{v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}\}$ determines the positive orientation of V when σ is positive and the negative orientation when σ is negative. Here we think of the given ordered basis of V as determining either $\pm\omega_1 \wedge \cdots \wedge \omega_n$, where we choose the sign according to which choice gives the answer 1 when the form is applied to that ordered basis.

Here is a corollary of Proposition (10.25).

Proposition 10.28. *Suppose $B = \{v_1, \ldots, v_n\}$ is an ordered basis of V and $T : V \to V$ a nonsingular linear transformation. Then the ordered basis $\{Tv_1, \ldots, Tv_n\}$ determines the same orientation as B if and only if* $\det T > 0$.

Proof. Observe that if $\{\omega_1, \ldots, \omega_n\}$ is the dual basis of B, then $\{(T^{-1})^*\omega_1, \ldots, (T^{-1})^*\omega_n\}$ is the dual basis of $\{Tv_1, \ldots, Tv_n\}$.(why?) The ordered basis $\{Tv_1, \ldots, Tv_n\}$ determines the orientation given by $(T^{-1})^*\omega_1 \wedge \cdots \wedge (T^{-1})^*\omega_n$. By (10.28) with T^{-1} in place of T, that is the same as the orientation given by $\omega_1 \wedge \cdots \wedge \omega_n$. □

Remark 10.29 (Right hand rule). *Let $V = \mathbb{R}^3$. It is often said that the "standard" or positive orientation of $\mathbb{R}^3$ is given by the* ordered *basis $\{e_1, e_2, e_3\}$ according to "the right hand rule." This usage reflects the biological accident that when we point the fingers of our right hand in the direction e_1, and then curl our fingers in the direction e_2, our thumb happens to point in the direction e_3. In the language of this section, this rule is equivalent to saying that the ordered standard basis $\{e_1, e_2, e_3\}$ gives the positive orientation, which means that $\omega_1 \wedge \omega_2 \wedge \omega_3 =$* det *gives the positive orientation, where $\{\omega_1, \omega_2, \omega_3\}$ is the dual basis. Observe that $\{e_2, e_1, e_3\}$ gives the negative orientation, while $\{e_2, e_3, e_1\}$ gives the positive orientation.*[42]

11 What is an $m-$dimensional surface in $\mathbb{R}^n$?

In this section we define the notion of a "C^1, $m-$dimensional surface in $\mathbb{R}^n$", where $m \leq n$.[43] This makes precise and generalizes the notion of 2−dimensional surface in $\mathbb{R}^3$ that we worked with earlier in this course, and that you studied in multivariable calculus. Our text gives a nice treatment surfaces in section 3.2. My goal here is to give what may be a more easily readable distillation of some of the main points, and to provide extra examples and motivation.

Definition 11.1. *(a) Let $m \leq n$. We say that a set $M \subset \mathbb{R}^n$ is a* a C^1, $m-$dimensional surface in $\mathbb{R}^n$ *if, given any $p \in M$, there is an open set $U \ni p$ in M and a C^1 map $\phi : \mathcal{O} \to U$, from an open set $\mathcal{O} \subset \mathbb{R}^m$ bijectively to U, with ϕ' injective at each point, and continuous inverse $\phi^{-1} : U \to \mathcal{O}$.*[44] *Such a map is called a* coordinate chart on M *and we call $U \subset M$ a* coordinate patch.

[42] Check that the last ordered basis is "right-handed".

[43] The term "surface of class C^1" is also used.

[44] Since $M \subset \mathbb{R}^n$, we know that $(M, |x - y|)$ is a metric space, so it makes sense to speak of an "open set in M" containing p. When we say that the map $\phi : \mathcal{O} \to U$ is C^1, we mean that the map $\phi : \mathcal{O} \to \mathbb{R}^n$ is C^1. The map ϕ has n scalar components.

(b) A C^k, $m-$dimensional surface in $\mathbb{R}^n$ is defined in exactly the same way for any $k \geq 1$ or for $k = \infty$, except that we require ϕ to be C^k.

This is perhaps the first "global" definition of a surface you have ever seen. Before (in HW 3, for example) we have worked mainly with surfaces defined in neighborhood of a single point.

A C^1, $m-$dimensional surface in $\mathbb{R}^n$ is a special case of a more general mathematical object known as a C^1, $m-$dimensional *manifold*. Like our surfaces, manifolds are locally homeomorphic to open subsets in euclidean space, but they are not required to be subsets of $\mathbb{R}^n$ for any n. We will discuss manifolds later in the course if time permits.[45]

Examples 11.2. *(a) The subsets $M_1 \subset \mathbb{R}^n$ given by $\{x \in \mathbb{R}^n : |x| = 1\}$ and $M_2 = \{x = (x', x_n) \in \mathbb{R}^n : |(x', x_n) - (0, x_n)| = 1\}$ are C^k, $(n-1)-$dimensional surfaces in $\mathbb{R}^n$ for all k.[46] The first is the unit sphere, and the second is an infinite cylinder of radius 1 centered on the x_n axis.*

(b)The graph of a C^1 function $g : \mathcal{O} \subset \mathbb{R}^{n-1} \to \mathbb{R}$, namely $M_3 := \{(x, g(x)) : x \in \mathcal{O}\}$ is a C^1, $(n-1)-$dimensional surface in $\mathbb{R}^n$.

(c) Let $I \subset \mathbb{R}$ be an open interval and $r : I \to \mathbb{R}^n$ a C^1 map such that $r'(t) \neq 0$ for $t \in I$. Suppose that for each $p \in r(I)$ there is an open set $U \ni p$ in $r(I)$ and an open subinterval $I_U \subset I$ such that $r : I_U \to U$ is a homeomorphism. Then $M_4 = r(I)$ is a C^1 one-dimensional surface (or "curve") in $\mathbb{R}^n$.

(d) Suppose $\psi : \mathbb{R}^n \to \mathbb{R}$ is C^1 and $S = \{x \in \mathbb{R}^n : \psi(x) = 0\}$. Suppose $\psi(a) = 0$ and $\nabla\psi(a) \neq 0$. The implicit function theorem implies that there is an open set V in $\mathbb{R}^n$, $V \ni a$, such that $M_5 = S \cap V$ is a C^1 $(n-1)-$dimensional surface in $\mathbb{R}^n$ (check).

(e) Suppose $\psi_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2$, are C^1 maps and $S = \{x \in \mathbb{R}^n : \psi_1(x) = \psi_2(x) = 0\}$. Suppose $a \in S$ and that $\nabla\psi_1(a)$ and $\nabla\psi_2(a)$ are linearly independent. The implicit function theorem implies that there is an open set V in $\mathbb{R}^n$, $V \ni a$, such that $M_6 = S \cap V$ is a C^1 $(n-2)-$dimensional surface in $\mathbb{R}^n$ (check). One can produce $(n-\ell)-$dimensional surfaces in $\mathbb{R}^n$ in the same way.

(f) As in Math 233 let $\mathcal{O} \subset \mathbb{R}^2$ be open, suppose $r : \mathcal{O} \to \mathbb{R}^3$ is C^1, write points in $\mathcal{O}$ as (s, t), and suppose that r_s and r_t are linearly independent at $(0, 0)$. Then there is an open set $\mathcal{O}' \subset \mathcal{O}$ containing $(0, 0)$ such that $M_7 = r(\mathcal{O}')$ is the graph of a C^1 function defined on an open subset of $\mathbb{R}^2$ (why?), and is thus a C^1 two-dimensional surface in $\mathbb{R}^3$ (by example (b)). We say r "parametrizes" the surface. For example, $r(s, t) = (\sin s \cos t, \sin s \sin t, \cos s)$ defined on $\mathcal{O} := (0, \pi)_s \times (0, 2\pi)_t$ parametrizes (much of) the unit sphere.

(g) An open subset of $\mathbb{R}^n$ is for any k a C^k $n-$dimensional surface in $\mathbb{R}^n$. (why?)

11.1 Tangent spaces to surfaces: are they well-defined?

As in Definition 11.1 let $M \subset \mathbb{R}^n$ be a C^1, $m-$dimensional surface, and let $\phi : \mathcal{O} \to U$ be a coordinate chart defined on some open set $\mathcal{O} \subset \mathbb{R}^m$. If $\phi(x_0) = p \in U$, recall that $\phi'(x_0) : \mathbb{R}^m \to \mathbb{R}^n$ and define the *tangent space to M at p* by

$$T_pM := \text{Range } \phi'(x_0) \subset \mathbb{R}^n. \tag{11.1}$$

Since $\phi'(x_0)$ is injective, this is an $m-$dimensional subspace of $\mathbb{R}^n$ (check!). We refer to each $v \in T_pM$ as *a tangent vector to M at p.*

Remark 11.3. *(a) If $\phi : \mathcal{O} \to U$ is a C^1 coordinate chart as above, we have*

$$\phi(x) = \phi(x_0) + \phi'(x_0)(x - x_0) + r(x, x_0), \text{ where } \lim_{x \to x_0} \frac{r(x, x_0)}{|x - x_0|} = 0. \tag{11.2}$$

[45] They are defined starting just above (3.2.100) in our text.

[46] Thus, we refer to them as C^∞ surfaces.

The m−dimensional "plane" in $\mathbb{R}^n$ which is the range of $L(x) = \phi(x_0) + \phi'(x_0)(x - x_0)$ is just the translate of T_pM by $\phi(x_0) = p$. The equation (11.2) expresses (and makes precise) the fact that $p + T_pM$ is the "plane" which best approximates M near p.

(b) If $\mathcal{O}$ is an open subset of $\mathbb{R}^n$, and thus a C^∞, n−dimensional surface in $\mathbb{R}^n$, then (why?)

$$T_p\mathcal{O} = \mathbb{R}^n \text{ for any } p \in \mathcal{O}. \tag{11.3}$$

With Ω an open subset of $\mathbb{R}^m$, suppose now that $\psi : \Omega \to V \subset \mathbb{R}^n$ is another C^1 coordinate chart on M such that $\psi(y_0) = p = \phi(x_0) \in M$. For convenience assume $\psi(\Omega) = \phi(\mathcal{O}) = U$ (as can be arranged by shrinking Ω and $\mathcal{O}$ if necessary). Let us confirm that T_pM, "the tangent space to M at p," is well-defined by checking that

$$\text{Range } \psi'(y_0) = \text{Range } \phi'(x_0). \tag{11.4}$$

Let $F : \mathcal{O} \to \Omega$ be given by $F = \psi^{-1} \circ \phi$.[47] It is clear that F is continuous with a continuous inverse. In fact, F can be shown to be a C^1 diffeomorphism; this is the content of Lemma 3.2.1. Clearly, we have $\phi = \psi \circ F$, so by the chain rule

$$\phi'(x_0) = \psi'(y_0) \circ F'(x_0). \quad \text{(Note: } F(x_0) = y_0. \text{ (why?))} \tag{11.5}$$

Thus, (11.4) holds. When coordinate charts overlap, they define the same tangent spaces on the intersection.

Remark 11.4. *The proof of Lemma 3.2.1 is simple. First "augment" the function $\phi : \mathcal{O} \to U$ to a C^1 function $\Phi : \mathcal{O} \times \mathbb{R}^{n-m} \to \mathbb{R}^n$, by defining $\Phi(x, z) = \phi(x) + Az$ for a matrix A chosen so that $\Phi'(x_0, 0)$ is invertible. The inverse function theorem now yields an open set $\tilde{\mathcal{O}}$ in $\mathcal{O} \times \mathbb{R}^{n-m}$ containing $(x_0, 0)$ and an open set $\tilde{U}$ in $\mathbb{R}^n$ containing $\phi(x_0) = p$ such that $\Phi : \tilde{\mathcal{O}} \to \tilde{U}$ is a C^1 diffeomorphism. Define a local C^1 diffeomorphism $\Psi(y, z) = \psi(y) + Bz$ augmenting $\psi(y)$ similarly, and conclude that F and F^{-1} are C^1 by observing for x near x_0 and y near y_0 that*

$$\Psi^{-1} \circ \Phi(x, 0) = (F(x), 0) \text{ and } \Phi^{-1} \circ \Psi(y, 0) = (F^{-1}(y), 0). \tag{11.6}$$

11.2 Functions mapping between surfaces.

Let M and N be C^1 surfaces of (respective) dimensions m and l in $\mathbb{R}^n$. There are two natural ways to define the notion of a C^1 map $f : M \to N$. Each definition makes use of the definition we already have of C^1 maps between open subsets of euclidean spaces.

Definition 11.5 (I)**.** *A function $f : M \to N$ is C^1 if, given any $p \in M$ there exists an open set $\mathcal{U} \subset \mathbb{R}^n$ containing p such that $f|_{M \cap \mathcal{U}}$ extends to a C^1 function $\tilde{f} : \mathcal{U} \to \mathbb{R}^n$.*[48]

Definition 11.6 (II)**.** *A function $f : M \to N$ is C^1 if for each $p \in M$ there are open sets $U \ni p$ in M and $V \ni f(p)$ in N, and charts $\phi : \mathcal{O} \to U$ and $\psi : \Omega \to V$ such that $\psi^{-1} \circ f \circ \phi : \mathcal{O} \to \Omega$ is a C^1 map.*

These definitions are actually equivalent *when* both M and N are C^1 surfaces, as we have assumed.[49] This can be shown using C^1 maps Φ and Ψ as constructed in Remark 11.4, which "augment" ϕ and ψ. For example, if f extends to $\tilde{f}$ near p, we have

$$\Psi^{-1} \circ \tilde{f} \circ \Phi(x, 0) = (\psi^{-1} \circ f \circ \phi(x), 0) \text{ for } x \text{ near } x_0, \tag{11.7}$$

and the map on the left is clearly C^1 in x. The proof of the other direction is part of Homework 7.

[47] This is a good time to look at Figure 3.2.1 in our text.

[48] The original map $f : M \to N \subset \mathbb{R}^n$ has n components.

[49] Definition 11.5 makes sense for maps f between *arbitrary* subsets M, N of $\mathbb{R}^n$!

Definition 11.7. *(a) Suppose M and N are both C^1, $m-$dimensional surfaces in $\mathbb{R}^n$. We say that a C^1 map $f : M \to N$ is a* local C^1 diffeomorphism *at $p \in M$ if there exist open sets $U \ni p$ and $V \ni f(p)$ in M and N respectively, such that $f : U \to V$ is bijective with C^1 inverse.*

(b)We say that M and N are C^1-diffeomorphic *if there exists a bijective map $f : M \to N$ which is a local C^1 diffeomorphism at every $p \in M$.*

Proposition 11.8. *Let $\phi : \mathcal{O} \to U$, where $\phi(x_0) = p$, be a coordinate chart on a C^1, $m-$dimensional surface $M \subset \mathbb{R}^n$. Then ϕ is a C^1 diffeomorphism of $\mathcal{O}$ onto U.*

Proof. We know ϕ is a homeomorphism of $\mathcal{O}$ onto U, and ϕ is C^1. Thus, it remains just to show that $\phi^{-1} : U \to \mathcal{O}$ is C^1. We'll use Definition 11.5. Consider a C^1-diffeomorphism $\Phi : \tilde{\mathcal{O}} \to \tilde{U}$ exactly as in Remark 11.4, where $\tilde{\mathcal{O}} \ni (x_0, 0)$ and $\tilde{U} \ni \phi(x_0) = p \in U$. If $\tilde{u} \in \tilde{U}$, then $\Phi^{-1}(\tilde{u}) = (x, z) \in \tilde{\mathcal{O}}$ for some (x, z). Moreover,

$$\Phi^{-1}(\phi(x)) = (x, 0) \text{ for } x \text{ near } x_0, \tag{11.8}$$

so $\pi_x \Phi^{-1}$ is a C^1 map between open sets in euclidean spaces that extends ϕ^{-1} near p.[50] □

Remark 11.9. *(a) The action of the function $\phi^{-1} : U \to \mathcal{O} \subset \mathbb{R}^n$ can be written $\phi^{-1}(p) = (x_1(p), \ldots, x_m(p))$. Proposition 11.8 implies that the functions $x_i : U \to \mathbb{R}$ are C^1.*

(b) Recall that $\phi'(x_0) : \mathbb{R}^m \to T_pM$. Using (11.8) *we can see that $\phi'(x_0)^{-1}$, which is defined on T_pM, is given by $(\pi_x \Phi^{-1})'|_{T_pM}$, the restriction to T_pM of an $m \times n$ matrix.*

(c) The above three definitions and Proposition 11.8 have obvious analogues where all occurrences of C^1 are replaced by C^k for some $k > 1$ or by $k = \infty$. Note that we only speak of C^k maps $f : M \to N$ when both M and N are C^k surfaces.

Given a C^1 map $f : M \to N$ between C^1 surfaces, our next task is to define the linear transformation $f'(p) : T_pM \to T_{f(p)}N$. We want this definition to satisfy two properties:

(i) For maps between open sets in euclidean spaces the new definition should agree with the old one.

(ii) The chain rule should hold for $g \circ f$, when $f : M \to N$ and $g : N \to P$:

$$(g \circ f)'(p) = g'(f(p)) \circ f'(p).$$

It is not too hard to see that there is only one *possible* definition with these properties. Let $\phi : \mathcal{O} \to U \subset M$ and $\psi : \Omega \to V \subset N$ be C^1 charts as in Definition 11.6 with $\phi(x_0) = p$ and $\psi(y_0) = f(p)$, and let $h : \mathcal{O} \to \Omega$ be $h = \psi^{-1} \circ f \circ \phi$.[51] Then the only acceptable definition is[52]

$$f'(p) = \psi'(y_0) \circ h'(x_0) \circ \phi'(x_0)^{-1}. \tag{11.9}$$

Another *candidate* for a definition is to let $\tilde{f}$ be a C^1 extension of f to an open set in $\mathbb{R}^n$ containing p and to define

$$f'(p) = (\tilde{f})'(p)|_{T_pM}. \tag{11.10}$$

[50]Here we define $\pi_x(x, z) = x$.

[51]Draw a picture of a square with M, N at the top two corners, $\mathcal{O}, \Omega$ at the bottom two corners, and the maps f, ϕ, ψ, h indicated by arrows connecting the corners. Let's call this a "standard corner diagram".

[52]Look at your picture to see this.

It remains to see that these definitions are well-defined, that they are equivalent, and that they satisfy properties (i),(ii) above. To see that $f'(p)$ in (11.9) is well-defined, one must check that the right side is independent of the charts used. To see that $f'(p)$ in (11.10) is well-defined, one must check that the range of $f'(p)$ is actually a subset of $T_{f(p)}N$ and that the definition is independent of the extension $\tilde{f}$.

Proposition 11.10. *The definitions of $f'(p) : T_pM \to T_{f(p)}N$ given by* (11.9) *and* (11.10) *are well-defined, equivalent, and satisfy properties (i),(ii).*

Proof. **1.** Take augmentations Φ, Ψ of ϕ, ψ as in Remark 11.4, take an extension $\tilde{f}$ of f, and define[53]

$$H = \Psi^{-1} \circ \tilde{f} \circ \Phi. \tag{11.11}$$

It is clear that $\tilde{f} = \Psi \circ H \circ \Phi^{-1}$, and since all maps in (11.11) are maps between euclidean spaces, we have by the original chain rule (suppressing evaluations):

$$(\tilde{f})' = \Psi' \circ H' \circ (\Phi^{-1})'. \tag{11.12}$$

2. We claim

$$\Psi' \circ H' \circ (\Phi^{-1})'|_{T_pM} = \psi'(y_0)h'(x_0)\phi'(x_0)^{-1}. \tag{11.13}$$

Let us assume (11.13) for a moment. With (11.12), the equality (11.13) implies:

a) The map $(\tilde{f})'|_{T_pM}$ has range in $T_{f(p)}N$ and is independent of the choice of extension of f.
b) The definition 11.9 is independent of the charts used, since $\tilde{f}$ is independent of those charts.
c) The definitions of $f'(p)$ given by (11.9) and (11.10) are equivalent.

3. Since definition (11.10) clearly satisfies properties (i),(ii), it only remains to check the claim (11.13). Using (11.8) and Remark 11.9 we have

$$(\Phi^{-1})'|_{T_pM} = \begin{pmatrix} (\pi_x \Phi^{-1})'|_{T_pM} \\ 0 \end{pmatrix} = \begin{pmatrix} \phi'(x_0)^{-1} \\ 0 \end{pmatrix}. \tag{11.14}$$

The map H satisfies $H(x,0) = (h(x),0)$, so the $n \times n$ matrix $H'(x_0,0)$ has the form

$$H'(x_0,0) = \begin{pmatrix} H_x(x_0,0) & H_z(x_0,0) \end{pmatrix} = \begin{pmatrix} h'(x_0) & * \\ 0 & * \end{pmatrix}, \tag{11.15}$$

where $\begin{pmatrix} * \\ * \end{pmatrix} = H_z(x_0,0)$. Since $\Psi'(y_0,0) = \begin{pmatrix} \psi'(y_0) & * \end{pmatrix}$, we obtain (11.13).

□

We end this section with an extension of the inverse function theorem to maps between C^1 surfaces M and N.

Theorem 11.11 (Inverse function theorem)**.** *Suppose $f : M \to N$ is a C^1 map whose derivative $f'(p) : T_pM \to T_{f(p)}N$ is an isomorphism.*[54] *Then f is a local C^1- diffeomorphism at p.*

The proof is a direct application, using charts, of the inverse function theorem for maps between open subsets of euclidean space.

[53]Draw a standard corner diagram for these maps!
[54]This implies M and N have the same dimension.

12 Differential forms

We begin with a section to motivate our study of differential forms.

12.1 Why study differential forms?

Even back in Calculus I, it was apparent that *orientation* has to play a role in parts of integration theory.[55] For example, when we use the fundamental theorem of calculus to write

$$\int_a^b f'(x)dx = f(b) - f(a), \tag{12.1}$$

it is important that the orientation of $[a, b]$, the domain of integration on the left, and the orientation of $\{a, b\}$, which we can view as the domain of "integration" on the right, are in a sense *consistent* with each other. If we integrate *from* a to b on the left, then on the right we must have $f(b) - f(a)$, not $f(a) - f(b)$. If we integrate from b to a on the left, then we must have $f(a) - f(b)$ on the right to get a true statement. We can express this consistency condition on orientations by saying that the given orientation $a \to b$ on $[a, b]$ *induces* a corresponding orientation on the boundary of $[a, b]$, namely $\{a, b\}$, and that this induced orientation must be used on the right in 12.1.

In Math 233 you saw that orientation plays a similar role in Green's theorem, which says that if $F = (M, N)$ is a C^1 vector field in the plane, and if $D \subset \mathbb{R}^2$ is an open set enclosed by a C^1 simple closed curve $C = \partial D$, then

$$\int_D (N_x - M_y)dxdy = \int_C Mdx + Ndy, \tag{12.2}$$

where the curve C must be oriented *counterclockwise*. Here the appearance of $N_x - M_y$ rather than $M_y - N_x$ in the integrand on the left indicates that the region D is "oriented", using the method of Math 233, by a unit normal vector pointing in the k direction. This orientation on D "induces" the counterclockwise orientation on the curve C by the right hand rule.[56] If the orientations on D and ∂D are not related in this way, the two sides of (12.2) will differ by a factor of -1.

The results expressed by (12.1) and (12.2) are both special cases of the Generalized Stokes Theorem, which says that if ω is any smooth $(k-1)-$form on a smooth, compact oriented $k-$dimensional surface M with smooth boundary ∂M, then[57]

$$\int_M d\omega = \int_{\partial M} \omega, \tag{12.3}$$

provided ∂M carries the orientation induced by the orientation of M.[58] Here $d\omega$ is a $k-$form, the "exterior derivative" of ω. Understanding and proving this theorem is our main goal for the near future.

When doing calculus on surfaces, two of the main challenges are to understand precisely how various constructions depend on particular choices of coordinate charts, and to formulate the main

[55] In this preliminary discussion I will sometimes use the word "orientation" in an informal, intuitive sense that is compatible with the strict sense of section 10.4.

[56] If you point the thumb of your right hand in the k direction, the fingers naturally curl in the counterclockwise direction.

[57] Every unfamiliar word in this sentence and the next will be defined shortly.

[58] Additional special cases of this theorem are the "divergence theorem" and the version of Stokes theorem for surfaces in $\mathbb{R}^3$ that you may also have seen in multivariable calculus or physics courses. All these results, which have obvious features in common, are derived from the fundamental theorem of calculus.

results in a way that is *independent* of particular choices of coordinate charts.[59] We have already encountered these issues in our (class) discussion of surface integrals, where we first defined the metric tensor $G(x)$ using a particular chart $\phi : \mathcal{O}_x \to U \subset M$, and then checked that the tensor $H(y)$ computed using another chart $\psi : \Omega_y \to U$ satisfies:

$$G(x) = F'(x)^t H(y) F'(x), \text{ where } y = F(x) \text{ and } F = \psi^{-1} \circ \phi : \mathcal{O} \to \Omega. \tag{12.4}$$

Using (12.4) and the change of variable formula, we showed that for any continuous function f supported in the coordinate patch U:

$$\int_{\mathcal{O}} f \circ \phi(x) \sqrt{g(x)} dx = \int_{\Omega} f \circ \psi(y) \sqrt{h(y)} dy, \text{ where } g(x) := \det G(x). \tag{12.5}$$

This follows from the change of variables formula since $\sqrt{g(x)} = |\det F'(x)| \sqrt{h(y)}$. The equality (12.5) then *allowed* us to define the surface integral $\int_U f dS$ by

$$\int_U f dS := \int_{\mathcal{O}} f \circ \phi(x) \sqrt{g(x)} dx. \tag{12.6}$$

Without (12.5) the definition (12.6) would not be well-defined. Moreover, *because* (12.6) is well-defined, it makes sense to define $\int_M f dS$ for functions f that are *not* supported in a single coordinate patch $U \subset M$, using (12.6) and partitions of unity as we have done.

People often summarize this situation by saying that $\sqrt{g(x)}dx$ and $\sqrt{h(y)}dy$ are two different representations in local coordinates of the same object dS which "lives on M".[60] In section 12.2 we will define a different kind of object that lives on M, a differential form α, for which integration on M makes sense. As with dS, the integral of α will first be defined on a coordinate patch using a local coordinate representation of α. The fact that the value of the integral is independent of the coordinate representation used will be a consequence of the change of variable formula and the properties of α.

We explained above that orientation plays a role in the fundamental theorem of calculus and its generalizations. We will see that differential forms are perfectly designed to serve as integrands in situations where orientation is important.[61]

Sections 10 and 11, together with the change of variables formula, provide the groundwork for the theory of differential forms on surfaces. We proceed to define them.

12.2 Definition of differential forms

When V is an $m-$dimensional real vector space, we defined in section 10.2 the space of alternating multilinear $k-$forms on V, $\Lambda^k(V)$. Suppose now that M is a C^∞ $m-$dimensional surface in $\mathbb{R}^n$. Then for any $p \in M$, we know that T_pM is an $m-$dimensional real vector space; in fact, it is a subspace of $\mathbb{R}^n$. It is now a small step to define the space of differential $k-$forms on M.

Definition 12.1 (Differential $k-$form)**.** *A $k-$form on M is a function ω that assigns to each point $p \in M$ an* alternating *multilinear $k-$form $\omega(p)$ on T_pM. That is, for each $p \in M$ we have $\omega(p) \in \Lambda^k(T_pM)$. If ω is "smooth" (that is, C^∞, in a sense to be defined below), then we say ω is a* differential $k-$form*, and with some abuse of notation we write $\omega \in \Lambda^k(M)$. We define the $0-$forms on M, $\Lambda^0(M)$, to be the set of functions $C^\infty(M, \mathbb{R})$.*

[59] These issues rarely arise when doing calculus on open subsets of $\mathbb{R}^n$, since we can fix a single global chart defined using the standard basis once and for all.

[60] That object is called a "density'.

[61] Orientation did *not* play a role in our discussion of surface integrals.

If ω_1 and ω_2 are $k-$forms and $c \in \mathbb{R}$, it is clear that we can define $k-$forms $\omega_1 + \omega_2$ and $c\omega_1$ by

$$(\omega_1 + \omega_2)(p) = \omega_1(p) + \omega_2(p) \text{ and } (c\omega_1)(p) = c\omega_1(p). \tag{12.7}$$

Similarly, if ω is a $q-$form and θ is an r form, we can define a $q + r$ form $\omega \wedge \theta$ by

$$(\omega \wedge \theta)(p) = \omega(p) \wedge \theta(p). \tag{12.8}$$

When $q = 0$ we *define* the right side of (12.8) to be the product of $\omega(p) \in \mathbb{R}$ with $\theta(p)$.

The pull-back operation (recall Definition 10.12) can also be defined pointwise. If $f : M \to N$ is C^1 and ω is a $k-$form on N, we define $f^*\omega$, a $k-$form on M by

$$(f^*\omega)(p) = f'(p)^*(\omega(f(p)), \text{ when } k \geq 1. \tag{12.9}$$

If $k = 0$ we *define* $f^*(\omega) = \omega \circ f$, a smooth function on M.

The properties of pullbacks discussed in section 10 immediately imply (check!)

$$\begin{aligned} &(a) f^*(\omega_1 + \omega_2) = f^*\omega_1 + f^*\omega_2 \\ &(b) f^*(\omega \wedge \theta) = f^*\omega \wedge f^*\theta \\ &(c) (f \circ h)^*\omega = h^* f^*\omega. \end{aligned} \tag{12.10}$$

12.3 Forms on open subsets of $\mathbb{R}^n$

Take $M = \mathcal{O}$, an open subset of $\mathbb{R}^n$, and let $f : \mathcal{O} \to \mathbb{R}$ be C^∞. For $x_0 \in \mathcal{O}$ we have $f'(x_0) : T_{x_0}\mathcal{O} \to T_{f(x_0)}\mathbb{R}$ or, equivalently, $f'(x_0) : T_{x_0}\mathcal{O} = \mathbb{R}^n \to \mathbb{R}$; recall (11.3). The map $x_0 \to f'(x_0) \in \Lambda^1(T_{x_0}\mathcal{O})$ is a $1-$form that is often denoted $df \in \Lambda^1(\mathcal{O})$ and is referred to as the *differential of* f. Observe that for any $v = (v_1, \ldots, v_n) \in T_{x_0}\mathcal{O}$, we have

$$df(x_0)(v_1, \ldots, v_n) = \nabla f(x_0) \cdot v. \tag{12.11}$$

In particular when $f = x_i$, one of the coordinate functions on $\mathcal{O}$, we obtain $dx_i(x_0)(v_1, \ldots, v_n) = v_i$.[62] If we take $v = e_j$, an element of the standard basis of $\mathbb{R}^n = T_{x_0}\mathcal{O}$, we obtain $dx_i(x_0)(e_j) = \delta_{i,j}$; recall (10.1). This shows that at each $x_0 \in \mathcal{O}$ the set $B' = \{dx_i(x_0), i = 1, \ldots, n\}$ is the dual basis to the standard basis $\{e_i, i = 1, \ldots, n\}$ of $T_{x_0}\mathcal{O}$. We have

$$\Lambda^1(T_{x_0}\mathcal{O}) = \text{span } \{dx_i(x_0), i = 1, \ldots, n\}. \tag{12.12}$$

From this we obtain immediately that any $1-$form ω, smooth or not, can be expressed uniquely in the form

$$\omega = \sum_{i=1}^n a_i dx_i, \tag{12.13}$$

for *some* functions $a_i : \mathcal{O} \to \mathbb{R}$. When $a_i \in C^\infty(\mathcal{O}; \mathbb{R})$, we say that ω is C^∞ (or smooth) and write $\omega \in \Lambda^1(\mathcal{O})$. In particular, if $f \in C^\infty(\mathcal{O}; \mathbb{R})$ we have (why?)[63]

$$df = \sum_{i=1}^n \partial_{x_i} f \ dx_i. \tag{12.14}$$

[62] The "coordinate function" $x_i : \mathcal{O} \to \mathbb{R}$ is the function given by $x_i(c_1, \ldots, c_n) = c_i$.

[63] Be sure you understand the meaning of each term in (12.14).

Now we can use the basis theorem for $\Lambda^k(V)$, Proposition 10.21, to express any $k-$form on $\mathcal{O}$ in terms of the dx_i. For each strictly increasing sequence $I = (i_1, \dots, i_k) \in I \in \mathcal{I}_{k,\nearrow}$ with $i_r \in \{1, \dots, n\}$, let[64]

$$dx_I := dx_{i_1} \wedge \cdots \wedge dx_{i_k} \in \Lambda^k(\mathcal{O}). \tag{12.15}$$

With Definition 12.1, Proposition 10.21 implies that any $k-$form ω on $\mathcal{O}$ can be written

$$\omega = \sum_{I \in \mathcal{I}_{k,\nearrow}} a_I dx_I, \tag{12.16}$$

for a unique choice of functions $a_I : \mathcal{O} \to \mathbb{R}$.(why?) When the $a_I \in C^\infty(\mathcal{O};\mathbb{R})$, we say that ω is a differential form and write $\omega \in \Lambda^1(\mathcal{O})$.

Next we show explicitly how these forms behave under pullbacks.

Proposition 12.2. *Let $\mathcal{O} \subset \mathbb{R}^m$ and $\Omega \subset \mathbb{R}^l$ be open sets, and suppose $f = (f_1, \dots, f_l) : \mathcal{O} \to \Omega$ is C^∞. Denote the coordinate functions on $\mathcal{O}$ and Ω by $x_1, \dots, x_m$ and $y_1, \dots, y_l$ respectively. Then*

$$f^* dy_i = df_i = \sum_{j=1}^{m} \partial_{x_j} f_i \, dx_j. \tag{12.17}$$

Proof. The second equality follows from (12.14). If $f(x_0) = y_0$ and $v \in T_{x_0}\mathcal{O} = \mathbb{R}^m$, we simply unravel the definitions to find[65]

$$(f^* dy_i)(x_0)(v) = [(f'(x_0))^* dy_i(f(x_0))](v) = dy_i(y_0)(f'(x_0)v) = \nabla f_i(x_0) \cdot v = df_i(x_0)(v). \tag{12.18}$$

□

An easy way to remember this proposition is to view it as saying that f^* and d commute, when f^*d is applied to y_i. This commutation property will be generalized later to allow the $0-$form y_i to be replaced by any $k-$form.

The action of f^* on $0-$forms and the $1-$forms dy_i determines it completely. The next proposition is an immediate consequence of the properties (12.10) and Proposition 12.2.

Proposition 12.3. *With $\mathcal{O}$, Ω, and f as in Proposition 12.2, let $\omega = \sum_{I \in \mathcal{I}_{k,\nearrow}} a_I dy_I$ be an arbitrary element of $\Lambda^k(\Omega)$. Then*

$$f^*\omega = \sum_{I \in \mathcal{I}_{k,\nearrow}} (f^* a_I) f^* dy_I = \sum_{I \in \mathcal{I}_{k,\nearrow}} (a_I \circ f) df_{i_1} \wedge \cdots \wedge df_{i_k}. \tag{12.19}$$

The following corollary will be used to show that integration of forms is well-defined.

Corollary 12.4. *In Proposition 12.3 take $\mathcal{O}$ and Ω to be open sets in $\mathbb{R}^m$, and suppose $a dy_1 \wedge \cdots \wedge dy_m$ is any element of $\Lambda^m(\Omega)$. Then*

$$f^*(a dy_1 \wedge \cdots \wedge dy_m) = (a \circ f) \det(f') \, dx_1 \wedge \cdots \wedge dx_m, \text{ where } \det(f')(x) := \det(f'(x)). \tag{12.20}$$

In particular when $a = 1$ we have

$$f^*(dy_1 \wedge \cdots \wedge dy_m) = \det(f') \, dx_1 \wedge \cdots \wedge dx_m, \text{ where } \det(f')(x) := \det(f'(x)). \tag{12.21}$$

[64] Recall Notations 10.20.

[65] This proposition is fundamental, so be sure you understand each equality here. We have used Definition 10.12, of course.

Proof. We will write the same proof in two slightly different ways.[66]

1. Observe that $T_x\mathcal{O} = T_{f(x)}\Omega = \mathbb{R}^m$, and that the standard basis of linear functionals on $\mathbb{R}^m$ can be written as $\{dx_i(x), i = 1, \dots, m\}$ for $T_x\mathcal{O}$ and as $\{dy_i(f(x)), i = 1, \dots, m\}$ for $T_{f(x)}\Omega$. In the case when $a = 1$ the equality (12.20) is a direct consequence of Proposition 10.25, and the general case follows immediately .

2. Taking $a = 1$, we have

$$f^*(dy_1 \wedge \cdots \wedge dy_m) = df_1 \wedge \cdots \wedge df_m = (\sum_{j=1}^m \partial_{x_j} f_1\, dx_j) \wedge \cdots \wedge (\sum_{j=1}^m \partial_{x_j} f_m\, dx_j), \tag{12.22}$$

and we conclude by repeating the calculation that begins on the second line of (10.31).

□

12.4 Forms on surfaces.

We first explain what it means for a form ω on a C^∞ surface M to be C^∞ (that is, smooth). In section 12.3 we defined smoothness of forms defined on open subsets of euclidean space.

Definition 12.5. *Let ω be a form on M. Then ω is smooth if $\phi^*\omega$ is smooth for every smooth chart of the form $\phi : \mathcal{O} \to U \subset M$.*[67]

As with forms on open subsets of $\mathbb{R}^n$, the building blocks of differential forms on smooth surfaces $M \subset \mathbb{R}$ are smooth functions $f : M \to \mathbb{R}^n$ and their "differentials".

Definition 12.6. *Let M be a smooth $m-$dimensional surface in $\mathbb{R}^n$ and $f \in C^\infty(M, \mathbb{R}) = \Lambda^0(M)$. The* differential *of f is the map that sends $p \in M$ to $f'(p) \in \Lambda^1(T_pM)$. We denote this map by df and write $df(p) = f'(p)$. The smoothness of df follows from Proposition 12.7 and* (12.14)*, so $df \in \Lambda^1(M)$.*

The next proposition extends the commutation property we observed in Proposition 12.2.

Proposition 12.7. *Let M and N be smooth surfaces in $\mathbb{R}^n$. Suppose $f : M \to N$ is smooth (write $f \in C^\infty(M; N)$) and $\phi \in C^\infty(N; \mathbb{R})$. Then*

$$f^*(d\phi) = d(f^*\phi). \tag{12.23}$$

Proof. For $p \in M$ and $v \in T_pM$ we unravel definitions to check that

$$f^*(d\phi)(p)(v) = d(f^*\phi)(p)(v). \tag{12.24}$$

The right side of (12.24) is $d(\phi \circ f)(p)(v) = \phi'(f(p)) \circ f'(p)v$. The left side is

$$(f^*d\phi)(p)(v) = [df(p)^*d\phi(f(p))]v = d\phi(f(p)) \circ df(p)v. \tag{12.25}$$

□

[66] I prefer the second.

[67] Using the third property of pullbacks in (12.10), one sees that it suffices to check smoothness of $\phi^*(\omega)$ just for *one* collection of charts that covers M. This definition is readily checked to be equivalent to the definition in section 12.3 of smooth forms on open subsets of $\mathbb{R}^n$.

We proceed to describe all possible differential $1-$forms on M, a smooth $m-$dimensional surface in $\mathbb{R}^n$. Let $\phi : \mathcal{O} \to U \subset M$ be a smooth chart. The version of Proposition 11.8 for C^∞ maps shows that $\phi^{-1} : U \to \mathcal{O}$ is smooth, so if we write $\phi^{-1}(p) = (x_1(p), \ldots, x_m(p)$ for $p \in U$, we see that these "coordinate functions on U" satisfy $x_i \in C^\infty(U, \mathbb{R})$. Thus, their differentials $dx_i \in \Lambda^1(U)$. The next Proposition implies immediately that any $\omega \in \Lambda^1(U)$ can be written in the form

$$\omega = \sum_{i=1}^{m} a_i dx_i \tag{12.26}$$

for a unique choice of $a_i \in C^\infty(U, \mathbb{R})$, $i = 1, \ldots, m$.

Proposition 12.8. *For any $p \in U$ the set $B' = \{dx_i(p), i = 1, \ldots, m\}$, defined using the chart $\phi : \mathcal{O} \to U \subset M$, is a basis of $\Lambda^1(T_pM)$.*

Proof. We'll show that B' is the dual basis to the basis of T_pM given by $B = \{\phi'(x_0)e_j, j = 1, \ldots, m\}$, where $\phi(x_0) = p$. We have

$$dx_i(p)[\phi'(x_0)e_j] = d(x_i \circ \phi)(x_0)(e_j) = dx_i(x_0)(e_j) = \delta_{i,j}. \tag{12.27}$$

The occurrence of x_i in the final term of (12.27) is the coordinate function on $\mathcal{O}$ given by $x_i(c_1, \ldots, c_m) = c_i$. Here we have used that $x_i \circ \phi = \phi^* x_i$ is the coordinate function x_i on $\mathcal{O}$. □

Remark: Dual role of the symbol x_i. (a) One tipoff to the reader that the meaning of x_i changes from the second term to the third in (12.27) is the *evaluation point*. In the first two terms that point is $p \in U$. In the last term it is $x_0 \in \mathcal{O}$.

(b) Proposition 12.7 implies

$$dx_i(x_0) = d(\phi^* x_i)(x_0) = \phi^* dx_i(x_0). \tag{12.28}$$

Here the first occurrence of x_i is the coordinate function on $\mathcal{O}$, while the last is the coordinate function on U. Again one tipoff is the evaluation point x_0. Another is $\phi^* x_i$, which only makes sense if the x_i to which ϕ^* is applied is a function on U. Similarly, we have (why?)[68]

$$dx_i(p) = d((\phi^{-1})^* x_i)(p) = (\phi^{-1})^* dx_i(p), \tag{12.29}$$

where the first occurrence of x_i is the coordinate function on U, while the last is the coordinate function on $\mathcal{O}$.

(c) Proposition 12.8 (and its proof) may be summarized by saying that since $\{dx_i(x_0), i = 1, \ldots, m\}$ is a basis of $\Lambda^1(T_{x_0}\mathcal{O})$ dual to $\{e_i, i = 1, \ldots, m\}$, then $\{((\phi^{-1})^* dx_i)(p), i = 1, \ldots, m\}$ is a basis of $\Lambda^1(T_pU)$ dual to $\{\phi'(x_0)e_i, i = 1, \ldots, m\}$.

The *dual role* of the symbol x_i occurs frequently in the literature.[69]

Exercise.[70] (a) Suppose $\phi : \mathcal{O}_x \to U \subset M$ is a smooth chart on an $m-$dimensional surface M, and let $f \in C^\infty(U; \mathbb{R})$. Determine explicitly the coefficients $a_i \in C^\infty(U, \mathbb{R})$ in the expansion[71]

$$df = \sum_{i=1}^{m} a_i dx_i, \text{ where } dx_i \in \Lambda^1(U). \tag{12.30}$$

[68] Observe that the second and third terms (12.29) make sense *only* if interpreted as $[d((\phi^{-1})^* x_i)](p)$ and $[(\phi^{-1})^* dx_i](p)$, respectively.

[69] One could avoid the dual role of the symbol x_i by defining $u_i = \phi^* x_i$, where x_i is a coordinate function on U.

[70] This is part of HW 8.

[71] Compare your result for this part to (12.14).

(b) Suppose $\psi : \Omega_y \to U$ is another such chart, and let $df = \sum_{k=1}^m b_k dy_k$. Give an explicit formula expressing b_k in terms of the a_i. Specify where every function that appears is evaluated.

(Hint: In part (a) first expand $\phi^* df$.)

The following proposition gives a result parallel to (12.16).

Proposition 12.9. *Suppose M is a smooth $m-$dimensional surface, and let $\phi : \mathcal{O} \to U \subset M$ be a smooth chart with $\phi(x_0) = p$. Then any differential $k-$form $\omega \in \Lambda^k(U)$ can be written*

$$\omega = \sum_{I \in \mathcal{I}_{k,\nearrow}} a_I dx_I, \tag{12.31}$$

for a unique choice of functions $a_I \in C^\infty(U;\mathbb{R})$. Here, of course, the x_{i_j} appearing in dx_I lie in $C^\infty(U;\mathbb{R})$.

Proof. By Definition 12.5 we have $\phi^*\omega \in \Lambda^k(\mathcal{O})$ and thus (12.16) implies

$$\phi^*\omega = \sum_{I \in \mathcal{I}_{k,\nearrow}} b_I dx_I, \tag{12.32}$$

for a unique choice of functions $b_I \in C^\infty(\mathcal{O};\mathbb{R})$. Now (why?)[72]

$$\omega = (\phi^{-1})^*\phi^*\omega = \sum_{I \in \mathcal{I}_{k,\nearrow}} (b_I \circ \phi^{-1})(\phi^{-1})^* dx_{i_1} \wedge \cdots \wedge (\phi^{-1})^* dx_{i_k} = \sum_{I \in \mathcal{I}_{k,\nearrow}} (b_I \circ \phi^{-1}) dx_I. \tag{12.33}$$

□

We conclude this section with a few computations involving an $m-$form ω on an $m-$dimensional surface M that should be compared with Corollary 12.4. Be sure you can explain all the expressions and equalities that follow. Whenever a dx_i or a dy_j appears, is it a form on M or on an open subset of euclidean space? Where is every function that appears evaluated? In what space Λ^k space does each form that appears belong? Begin by drawing a standard corner diagram with three corners.

Let $\phi : \mathcal{O}_x \to U \subset M$ and $\psi : \Omega_y \to U$ be smooth charts, and define $F : \mathcal{O} \to \Omega$ by $F = \psi^{-1} \circ \phi$. Let $\omega = a\, dy_1 \wedge \cdots \wedge dy_m \in \Lambda^m(U)$. Then we have

(12.34)
- *(a)* $\psi^*\omega = (a \circ \psi)\, dy_1 \wedge \cdots \wedge dy_m$
- *(b)* $\phi^*\omega = (a \circ \phi) \det F'\, dx_1 \wedge \cdots \wedge dx_m$
- *(c)* $F^*\psi^*\omega = \phi^*\omega$
- *(d)* $\omega = (\phi^{-1})^*\phi^*\omega = a \det(F' \circ \phi^{-1})\, dx_1 \wedge \cdots \wedge dx_m$ and $\omega = (\psi^{-1})^*\psi^*\omega = a\, dy_1 \wedge \cdots \wedge dy_m$.

Remark 12.10. *We speak of $\psi^*\omega$ and $\phi^*\omega$ as being two different "local coordinate representations" of $\omega \in \Lambda^m(U)$. Since $\phi : \mathcal{O} \to U$ is a smooth diffeomorphism, in some sense "no information is lost" in passing from ω to $\phi^*\omega$.*[73] *Indeed,* (12.34)*(d) shows that we can recover ω from $\phi^*\omega$ (or from $\psi^*\omega$). As we'll see in section 12.6 the factor* $\det F'$ *that appears in* (12.34)*(b) is essential for making sense of $\int_M \omega$. Look back at Proposition 10.25 and Corollary 12.4 to understand* why *that factor is there. In particular, by examining* (12.22) *try to convince yourself that "it all comes down to" $dx_i \wedge dx_j = -dx_j \wedge dx_i$, the anticommutativity of $1-$forms.*

[72] Answer: We use (12.10)(a),(b),(c) and (12.29) here.

[73] If ϕ happened to be the zero function, we'd have $\phi^*\omega = 0$ and *all* information about ω would be lost.

12.5 Orientation of a surface.

We recall from section 10.4 that if V is an $m-$dimensional real vector space, an orientation of V is determined by specifying *either* a choice of an element $\alpha \in \Lambda^m(V) \setminus \{0\}$ or a choice of ordered basis $\{v_1, \ldots, v_m\}$. Those choices determine the *same* orientation if and only if $\alpha(v_1, \ldots, v_m) > 0$. Moreover, if $\beta \in \Lambda^m(V) \setminus \{0\}$ then β determines the same orientation as α if and only if $\beta = c\alpha$ for some $c > 0$. It is now a small step to define orientation for smooth surfaces.

Definition 12.11. *Let M be a smooth $m-$dimensional surface. We say that M is* orientable *if there exists a nowhere vanishing element ω of $\Lambda^m(M)$.*[74] *Such an ω is said to* determine *an orientation of M. If γ is an another such form, then γ determines the same orientation if $\gamma = c\omega$ for some* positive *function $c \in C^\infty(M;\mathbb{R})$. In that case we write $\omega \sim \gamma$. When M is orientable, an equivalence class of nowhere vanishing $m-$forms on M determined by this equivalence relation is said to be* an orientation of M.[75] *If an orientation of M has been chosen, we say M is* oriented.

In the following remark we unpack this simple-looking definition.

Remark A

(a) If M is orientable with orientation determined by $\omega \in \Lambda^m(M)$ and p is any point of M, then $\omega(p) \in \Lambda^m(T_pM) \setminus \{0\}$. Because ω is smooth it therefore determines a *smoothly varying* choice of orientations for all the tangent spaces T_pM (recall Definition 10.26) as p varies throughout M.

(b) For a given $p \in M$ an ordered basis $\{v_1, \ldots, v_m\}$ of T_pM is said to determine the same orientation of T_pM as $\omega(p)$ provided $\omega(p)(v_1, \ldots, v_m) > 0$. Thus, you will see some authors describe an orientation of M as a "smoothly varying choice of ordered bases" of the tangent spaces T_pM as p varies throughout M.

(c) Let's relate our definition of orientation to the definition you may have seen in Math 233 or physics courses for $2-$dimensional surfaces in $\mathbb{R}^3$. If M is such a surface, then an ordered basis $\{v_1, v_2\}$ for a given T_pM determines a unit normal vector $n(p)$ to the surface at p, namely,

$$n(p) = \frac{v_1 \times v_2}{|v_1 \times v_2|}.$$

Thus, an orientation of M may be described as a smoothly varying choice of unit normal vectors to the tangent spaces T_pM as p varies throughout M. For example, the unit sphere $S^2 \subset \mathbb{R}^n$ can be oriented simply by choosing the outward unit normal at every point or the inward unit normal at every point. This description of orientation in terms of a normal vector to M can be generalized to the case where M is an $(n-1)-$dimensional surface in $\mathbb{R}^n$.

(d) An $m-$dimensional surface M can *always* be "oriented locally". Let $\phi : \mathcal{O} \to U \subset M$ be a smooth coordinate chart, and let the ordered standard basis $\{e_1, \ldots, e_m\}$ determine the orientation of $T_{x_0}\mathcal{O} = \mathbb{R}^m$ for every $x_0 \in \mathcal{O}$. Then as x_0 varies throughout $\mathcal{O}$, the points $\phi(x_0) = p$ vary throughout U, and the sets $\{\phi'(x_0)e_i, i = 1, \ldots, m\}$ give a smoothly varying choice of ordered bases of T_pM as p varies throughout U. We will refer to this orientation of U as the *orientation determined by the chart $\phi : \mathcal{O} \to U$*.[76]

(e) Here is a neater way to describe the orientation of U determined by the chart $\phi : \mathcal{O} \to U$. Let $x_i \in C^\infty(\mathcal{O};\mathbb{R})$ be the standard coordinate functions on $\mathcal{O}$.[77] Then $dx_1 \wedge \cdots \wedge dx_m \in \Lambda^m(O)$ determines

[74] This means $\omega(p) \neq 0$ for all $p \in M$.

[75] When M is orientable, authors often makes an arbitrary choice of one equivalence class and designate it *the positive* orientation.

[76] For any $x_0 \in \mathcal{O}$ the ordered basis $\{\phi'(x_0)e_i, i = 1, \ldots, m\}$ gives the "orientation of T_pU determined by $\{e_1, \ldots, e_m\}$ and $\phi'(x_0)$", in the terminology of problem 5 of HW 8.

[77] Recall $x_i(c_1, \ldots, c_m) = c_i$.

the same orientation on $\mathcal{O}$ as the standard ordered basis $\{e_1, \ldots, e_m\}$, and $dx_1 \wedge \cdots \wedge dx_m \in \Lambda^m(U)$ gives the orientation of U determined by the chart $\phi : \mathcal{O} \to U$. To see this note that for $x_0 \in \mathcal{O}$ and $\phi(x_0) = p$, we have (why?)[78]

$$dx_1(p) \wedge \cdots \wedge dx_m(p)(\phi'(x_0)e_1, \ldots, \phi'(x_0)e_m) = 1. \tag{12.35}$$

(f) Some surfaces are not orientable! The classic and simplest example is the 2−dimensional Möbius strip, formed by taking a thin rectangular strip of paper, giving one end a half-twist, and then joining ends to form a loop (do it). Pick an arbitrary point p on the central curve C that bisects the strip and place a toothpick at p pointing perpendicular to the strip. Think of the toothpick as a unit normal vector to the strip at p. Now transport this normal toothpick completely around C one time, and notice that when the toothpick returns to p, its normal direction is *opposite* to its starting normal direction. Thus, one cannot specify a smoothly varying choice of unit normal on this surface.[79]

(g) If M is a *connected* orientable surface of dimension m, it has exactly two possible orientations. To prove this, fix a nowhere vanishing $\omega \in \Lambda^m(M)$, and let β be some other nowhere vanishing $m-$form. Define a function $f : M \to \{1, 2\}$ by

$$f(p) = \begin{cases} 1, & \text{if } \beta(p) = c(p)\alpha(p) \text{ with } c(p) > 0 \\ 2, & \text{if } \beta(p) = c(p)\alpha(p) \text{ with } c(p) < 0 \end{cases} .$$

The function f is continuous (why?), so it must be constant since M is connected.

Definition 12.12. *Let M and N be oriented surfaces of the same dimension m, and suppose $f \in C^\infty(M; N)$ is a diffeomorphism. Let $\beta \in \Lambda^m(N)$ determine the given orientation of N. We say that f is* orientation preserving *if $f^*\beta$ determines the given orientation of M. In the special case when M (or N) is an open subset of $\mathbb{R}^m$, we will* always *use the "default" orientation determined by $dx_1 \wedge \cdots \wedge dx_m$, where the $x_i : M \to \mathbb{R}^n$ are the standard coordinate functions on M, $x_i(c_1, \ldots, c_m) = c_i$, unless otherwise stated.*

Remark B

(a) This definition of an orientation-preserving map $f \in C^\infty(M; N)$ is consistent with the definition of an orientation-preserving linear transformation $T : V \to W$ between two oriented vector spaces given in problem 4 of HW 8.

(b) A special case of Definition 12.12 that we'll often use is the case where "f" is a smooth coordinate chart $\phi : \mathcal{O} \to U \subset M$ and U has some given orientation. Then the given orientation of U is the same as the orientation determined by the chart (as defined in (d) or (e) of the previous remark) if and only if ϕ is orientation-preserving.(check!)

Proposition 12.13. *Let M be a smooth, oriented $m-$dimensional surface with orientation given by $\omega \in \Lambda^m(M)$. Suppose $\phi : \mathcal{O}_x \to U$ and $\psi : \Omega_y \to U$ are both charts on M that determine the given orientation on U. Then the map $F : \mathcal{O} \to \Omega$ given by $F = \psi^{-1} \circ \phi$ is orientation-preserving, or equivalently,* $\det F' > 0$.

Proof. Let $\omega = a dy_1 \wedge \cdots \wedge dy_m \in \Lambda^m(U)$ determine the given orientation of U. Since ψ is orientation-preserving, the function a must be positive.(why?) Since ϕ is also orientation-preserving, the equality (12.34)(b) implies that the function $(a \circ \phi) \det F'$ is positive, and therefore $\det F'$ is positive. The fact that F is orientation-preserving if and only if $\det F' > 0$ follows directly from (12.21). □

[78] Recall (12.27).

[79] This is not quite a rigorous proof, of course, but this argument can be made into one. Equivalent approaches are to look at one of M.C. Escher's ants before and after a roundtrip along C, or to transport a penny starting heads up once around C on a transparent Möbius strip.

12.6 Integration of differential $m-$forms on m-dimensional surfaces.

We now have all the machinery we need to define integration of $m-$forms on $m-$dimensional surfaces.[80] First we take $M = \mathcal{O}$, an open subset of $\mathbb{R}^m$, and we suppose that $\omega = adx_1 \wedge \cdots \wedge dx_m \in \Lambda^m(\mathcal{O})$ has compact support in $\mathcal{O}$, where we define $\operatorname{supp} \omega := \operatorname{supp} a$.

Definition 12.14. *Let $\mathcal{O}$ be an open subset of $\mathbb{R}^m$ with standard coordinate functions $x_i \in C^\infty(\mathcal{O};\mathbb{R})$, and suppose $\omega = adx_1 \wedge \cdots \wedge dx_m \in \Lambda^m(\mathcal{O})$ has compact support in $\mathcal{O}$. We define*[81]

$$\int_{\mathcal{O}} \omega := \int_{\mathcal{O}} a(x)dx. \tag{12.36}$$

Next we define the integral of an $m-$form ω with compact support in a coordinate patch on an $m-$dimensional oriented surface.

Definition 12.15. *Let $M \subset \mathbb{R}^n$ be an $m-$dimensional oriented surface, and let $\psi : \Omega_y \to U$ be a smooth chart on M that determines the given orientation of U. Suppose $\omega \in \Lambda^m(U)$ has compact support in U. Then $\omega = ady_1 \wedge \cdots \wedge dy_m$ for some function $a \in C_c^\infty(U,\mathbb{R})$, and we define*

$$\int_M \omega := \int_\Omega \psi^*\omega = \int_\Omega (a\circ\psi)dy_1 \wedge \cdots \wedge dy_m = \int_\Omega a(\psi(y))dy. \tag{12.37}$$

The first equality is the new definition; the second and third equalities follow from previous definitions (which ones?).[82]

We now verify that this definition is well-defined. All the hard work has been done, so the verification is short. Suppose $\phi : \mathcal{O}_x \to U$ is another smooth chart that determines the given orientation of U. We must show $\int_\Omega \psi^*\omega = \int_{\mathcal{O}} \phi^*\omega$. By (12.34)(b) we have

$$\phi^*\omega = (a\circ\phi)\det F'\, dx_1 \wedge \cdots \wedge dx_m, \text{ where } F = \psi^{-1}\circ\phi : \mathcal{O} \to \Omega.$$

Thus, we must check

$$\int_\Omega a(\psi(y))dy = \int_{\mathcal{O}} a(\phi(x))\det F'(x)\, dx. \tag{12.38}$$

But Proposition 12.13 implies $\det F' = |\det F'|$, so (12.38) is now immediate from the change of variable theorem.

This preparation allows us to integrate a compactly supported $m-$form ω on an oriented $m-$dimensional surface.[83] Here we define $\operatorname{supp}\omega$ to be the closure of the set of points where $\omega(p) \neq 0$.

Definition 12.16. *Let M be an oriented, smooth, $m-$dimensional surface and let $\omega \in \Lambda^m(M)$ have compact support. Choose a finite collection of orientation preserving charts $\phi_i : \mathcal{O}_i \to U_i$, $i = 1, \ldots, K$, such that $\operatorname{supp}\omega \subset \cup_{i=1}^K U_i$ and a smooth partition of unity $\{\rho_i\}$ subordinate to the cover such that $\sum_i \rho_i = 1$ on $\operatorname{supp}\omega$. Now define*

$$\int_M \omega = \sum_i \int_{U_i} \rho_i\omega. \tag{12.39}$$

[80] We will never speak of the integration of a $k-$form on an $m-$dimensional surface if $k \neq m$.
[81] The integral on the right in (12.36) is the usual Riemann integral.
[82] In this definition the variable y_i plays the "dual role" described in the remark just below (12.27).
[83] The compact support assumption is a simple way to insure that we will get a well-defined finite answer for $\int_M \omega$. One can of course relax this condition.

Each individual term on the right in (12.39) is well-defined as we just saw. The argument that $\int_M \omega$ is independent of the choice of partition of unity is essentially the same as the one we gave for $\int_M f dS$; recall problem 3 of HW 7.

Think about the proof of the next proposition, which will be part of HW 10. It's a good exercise in understanding the definitions.

Proposition A.

Let $f : M \to N$ be an orientation-preserving diffeomorphism, where M, N are both $m-$dimensional oriented smooth surfaces. Suppose $\omega \in \Lambda^m(N)$ is compactly supported. Show

$$\int_N \omega = \int_M f^*\omega.$$

12.6.1 Example

Consider $S^2 \subset \mathbb{R}^3$ with the orientation given by the outward unit normal. Let $i : S^2 \to \mathbb{R}^3$ be the inclusion map and let $\omega \in \Lambda^2(\mathbb{R}^3)$ be given by $\omega = x_2 dx_1 \wedge dx_3$. Set $\alpha = i^*\omega \in \Lambda^2(S^2)$ and compute $\int_{S^2_u} \alpha$ where S^2_u is the part of S^2 in $x_3 > 0$.

Solution: We'll use the chart $\phi : \mathcal{O}_x \to S^2_u$ given by $\phi(x_1, x_2) = (x_1, x_2, \sqrt{1 - |(x_1, x_2)|^2})$, where $\mathcal{O} = \{(x_1, x_2) : |(x_1, x_2)| < 1\}$. The ordered basis $\{\phi'(0)e_1, \phi'(0)e_2\} = \{i, j\}$ gives the correct orientation of $T_{(0,0,1)}S^2_u$ (why?), so this chart determines the given orientation of S^2_u. Now $\phi^*\omega = \phi^* i^* \omega = \phi^*\alpha$, so we compute[84]

$$\phi^*\omega = (x_2 \circ \phi) d(x_1 \circ \phi) \wedge d(x_3 \circ \phi) = x_2 dx_1 \wedge \left(\frac{-x_2}{\sqrt{1 - |(x_1, x_2)|^2}} \right) dx_2; \text{ and thus} \tag{12.40}$$

$$\int_{S^2_u} \alpha := \int_{\mathcal{O}} \phi^*\alpha = \int_{\mathcal{O}} \phi^*\omega = \int_{\mathcal{O}} \frac{-x_2^2}{\sqrt{1 - |(x_1, x_2)|^2}} dx_1 dx_2 = \int_0^{2\pi} \int_0^1 \frac{-r^2 \sin^2 \theta}{\sqrt{1 - r^2}} r dr d\theta. \tag{12.41}$$

12.7 The exterior derivative

Differential forms on a surface M can be differentiated as well as integrated, but as with integration, it takes some care to differentiate a form in a way that is independent of choices of coordinate charts.[85] The *exterior derivative* operator "d" is a linear map

$$d : \Lambda^p(M) \to \Lambda^{p+1}(M), \tag{12.42}$$

which is well-defined in the above sense. This operator has already appeared in the definition of the differential $df \in \Lambda^1(M)$ of a function $f \in \Lambda^0(M) = C^\infty(M; \mathbb{R})$, and in our (so far unexplained) statement of Stokes theorem (12.3). We will see that d generalizes the gradient, curl, and divergence operators from vector calculus.

Definition 12.17 (Exterior derivative). *(a) Let $\mathcal{O} \subset \mathbb{R}^m$ be an open set and $\omega = \sum_I a_I dx_I \in \Lambda^p(\mathcal{O})$. Then we define the* exterior derivative *of ω to be*

$$d\omega = \sum_I da_I \wedge dx_I \in \Lambda^{p+1}(\mathcal{O}). \tag{12.43}$$

[84] In (12.40) note the change in meaning of the symbols x_1, x_2 from the left to the right side of the second equality. The last integral in (12.41) can be done by a trig substitution; let's not do it.

[85] It takes much *less* care, though, in the case of differentiation. Nothing as technically challenging as the change of variable formula is involved.

(b) Let M be a smooth $m-$dimensional surface and suppose $\omega \in \Lambda^p(M)$. We define $d\omega \in \Lambda^{p+1}(M)$ locally as follows. Suppose $\phi : \mathcal{O} \to U \subset M$ is a smooth chart. Define $d\omega$ on U to be

$$d\omega|_U = (\phi^{-1})^* d(\phi^* \omega). \tag{12.44}$$

To see that $d\omega$ is well-defined for $\omega \in \Lambda^p(M)$ we will check that if $\psi : \Omega \to U$ is another smooth chart, then

$$(\phi^{-1})^* d(\phi^* \omega) = (\psi^{-1})^* d(\psi^* \omega). \tag{12.45}$$

This will be a consequence of the following commutation property, which we already know in the case $\omega \in \Lambda^0(\Omega)$ (why?).

Proposition 12.18. *Suppose $F : \mathcal{O} \to \Omega$ is C^∞, where $\mathcal{O}$ and Ω are open subsets of euclidean spaces. Then for every differential form $\omega \in \Lambda^k(\Omega)$ we have*

$$d(F^* \omega) = F^*(d\omega). \tag{12.46}$$

Before proving Proposition 12.18 we list a few properties of d acting on forms defined on an open subset of $\mathbb{R}^m$.

Proposition 12.19. *For differential forms defined on an open set $\mathcal{O} \subset \mathbb{R}^m$, the exterior derivative operator satisfies:*

$$\begin{aligned} &(a)\ d(\omega_1 + \omega_2) = d\omega_1 + d\omega_2, \\ &(b)\ d(\omega \wedge \theta) = (d\omega) \wedge \theta + (-1)^p \omega \wedge d\theta, \qquad \text{if } \omega \in \Lambda^p(\mathcal{O}), \\ &(c)\ d(d\omega) = 0. \end{aligned} \tag{12.47}$$

Sketch of proof. Property (a) is obvious. Property (b) is proved by direct computation using the product rule and the property of $1-$forms: $\alpha \wedge \beta = -\beta \wedge \alpha$. To prove (c) we let $\omega = \sum_I a_I dx_I$ and compute[86]

$$d(d\omega) = \sum_I \sum_i \left(\sum_j \partial_j \partial_i a_I dx_j \right) \wedge dx_i \wedge dx_I. \tag{12.48}$$

Property (c) follows by using $dx_j \wedge dx_i = -dx_i \wedge dx_j$ together with

$$\partial_i \partial_j a_I = \partial_j \partial_i a_I$$

to cancel terms in the sum two by two. □

We now show that Proposition 12.18 is an immediate consequence of Proposition 12.19, (12.10)(b), and the fact that d commutes with F^* when applied to $\omega \in \Lambda^0(M)$.

Proof of Proposition 12.18. Let $F = (f_1, \ldots, f_m)$ and $\omega = \sum a_I dy_{i_1} \wedge \cdots \wedge dy_{i_k}$.. Then (why?)[87]

$$\begin{aligned} &F^*\omega = \sum (a_I \circ F) df_{i_1} \wedge \cdots \wedge df_{i_k} \text{ and } dF^*\omega = \sum d(a_I \circ F) \wedge df_{i_1} \wedge \cdots \wedge df_{i_k}, \text{ while} \\ &d\omega = \sum da_I \wedge dy_{i_1} \wedge \cdots \wedge dy_{i_k} \text{ and } F^* d\omega = \sum d(a_I \circ F) \wedge df_{i_1} \wedge \cdots \wedge df_{i_k}. \end{aligned} \tag{12.49}$$

□

[86] Here we let $\partial_j = \partial_{x_j}$.

[87] Note especially where we have used Proposition 12.19(b),(c).

Now we can prove that the operator d is well-defined on $\omega \in \Lambda^k(M)$.

Proposition 12.20. *Let M be a smooth $m-$dimensional surface. Then $d : \Lambda^p(M) \to \Lambda^{p+1}(M)$ is well-defined by* (12.44).

Proof. We must check (12.45). Let $F = \psi^{-1} \circ \phi$. Then by Proposition 12.18

$$F^* d(\psi^* \omega) = d(F^* \psi^* \omega) = d(\phi^* \omega). \tag{12.50}$$

Now apply $(\phi^{-1})^*$ on the left in (12.50) to finish.

□

The following proposition is now immediate:

Proposition 12.21. *The properties of the exterior derivative stated in Propositions 12.18 and 12.19 continue to hold when $\mathcal{O}$ and Ω are replaced by smooth surfaces M and N. Moreover, when $f \in \Lambda^0(M)$, df agrees with the earlier definition.*

The next exercise relates d to the classical gradient, curl, and divergence operators of vector calculus on $\mathbb{R}^3$.
Exercise: part of HW 9

(a) Let $f \in \Lambda^0(\mathbb{R}^3)$. What is the relation between df and the gradient of f?
(b) Let $\omega = f_1 dx_1 + f_2 dx_2 + f_3 dx_3 \in \Lambda^1(\mathbb{R}^3)$. What is the relation between $d\omega$ and curl$F$, where F is the vector field (f_1, f_2, f_3)?
(c) Let $\omega = f_1 dx_2 \wedge dx_3 + f_2 dx_3 \wedge dx_1 + f_3 dx_1 \wedge dx_2 \in \Lambda^2(\mathbb{R}^3)$. What is the relation between $d\omega$ and div$F$ where again $F = (f_1, f_2, f_3)$?[88]
(d) If $\omega \in \Lambda^3(\mathbb{R}^3)$, compute $d\omega$ with justification.
In each of parts a-c, you are being asked to *find* the relation, and to describe what it is.

12.8 Surfaces with boundary

The generalized Stokes formula (12.3) involves $\int_{\partial M} \omega$, the integral of a differential $(k-1)-$form over the "boundary" of a $k-$dimensional surface M in $\mathbb{R}^n$. In this section we define and discuss "smooth surfaces with boundary". First we extend the notion of a (smooth) diffeomorphism to arbitrary subsets of $\mathbb{R}^n$.

Definition 12.22. *(a) Let X be an* arbitrary *subset of* $\mathbb{R}^n$*. A map $f : X \to \mathbb{R}^m$ is called* smooth *(or C^∞) if it may be locally extended to a smooth map on open sets; that is, if for each $p \in X$ there exists an open set $U \ni p$ in $\mathbb{R}^n$ and a smooth map $\tilde{f} : U \to \mathbb{R}^m$ such that $\tilde{f}$ equals f on $U \cap X$.*

(b) Let X and Y be arbitrary subsets of two euclidean spaces. A smooth map $f : X \to Y$ is a diffeomorphism *is it is bijective and if the inverse map $f^{-1} : Y \to X$ is smooth.*

Remark 12.23. *Observe that $M \subset \mathbb{R}^n$ is a smooth $m-$dimensional surface if and only if it is* locally diffeomorphic *to $\mathbb{R}^m$. This means that for each $p \in M$ there is an open set $U \ni p$ in M which is diffeomorphic to an open set $\mathcal{O} \subset \mathbb{R}^m$.*[89]

Our basic example of a smooth $m-$dimensional "surface with boundary" is the "left-halfspace" in $\mathbb{R}^m$:

$$H^m = \{x = (x_1, x') \in \mathbb{R}^m : x_1 \leq 0\}. \tag{12.51}$$

[88]Note the cyclic occurrence of 1,2,3 in the three terms of ω.
[89]The diffeomorphism $\phi : \mathcal{O} \to U$ is what we usually call a coordinate chart.

Definition 12.24 ($m-$dimensional surface with boundary). *(a) A subset $M \subset \mathbb{R}^n$ is a* smooth $m-$dimensional surface with boundary *if every point $p \in M$ is contained in an open set $U \subset M$ that is diffeomorphic to an open set $\mathcal{O}$ in H^m.*[90] *As before we call such a diffeomorphism $\phi : \mathcal{O} \to U$ a smooth coordinate chart on M.*

(b) The boundary *of M, denoted ∂M, consists of those points that belong to the image of the boundary of H^m under some coordinate chart.*

(c) The interior *of M,* Int $(M) := M \setminus \partial M$.[91]

Examples

1. The closed unit ball in $\mathbb{R}^3$ is a $3-$dimensional surface with boundary equal to S^2.

2. Let S^2_+ be the closed upper unit half-sphere in $\mathbb{R}^3$. This is a $2-$dimensional surface with boundary $\partial S^2_+ = \{(x, y, 0) \in \mathbb{R}^3 : |(x, y)| = 1\}$. Then $\text{Int}(S^2_+) = S^2_+ \cap \{z > 0\}$ is a smooth $2-$dimensional surface.

3. Let $\mathcal{C}$ be the cylindrical surface in $\mathbb{R}^3$ given $\{(x, y, z) \in \mathbb{R}^3 : |(x, y)| = 1, 0 \leq z \leq 10\}$. This is a $2-$dimensional surface with boundary

$$\partial\mathcal{C} = \{(x, y, 0) \in \mathbb{R}^3 : |(x, y)| = 1\} \cup \{(x, y, 10) \in \mathbb{R}^3 : |(x, y)| = 1\}.$$

4. Let $M = \{(x, y, 0) \in \mathbb{R}^3 : |(x, y)| \leq 1\}$, a $2-$dimensional surface with boundary $\partial M = \{(x, y, 0) : |(x, y)| = 1\}$ in $\mathbb{R}^3$. Considered as a subset of the metric space $(\mathbb{R}^3, |x - y|)$, we have $\mathring{M} = \emptyset$, $bM = M$. What about $N = \{(x, y) \in \mathbb{R}^2 : |(x, y)| \leq 1\}$?

5. Any smooth $m-$dimensional surface M in $\mathbb{R}^n$ is a smooth $m-$dimensional surface M with boundary, where $\partial M = \emptyset$.

Remark 12.25. *When M is a smooth $m-$dimensional surface with boundary, tangent spaces and derivatives are easily defined at boundary points $p \in \partial M$. For example, if $\phi : \mathcal{O} \to U$ is a smooth chart with $U \ni p = \phi(x_0)$, we define $\phi'(x_0)$ and T_pM as usual after first taking a smooth extension of ϕ to an open set containing x_0 in $\mathbb{R}^m$.*[92] *By continuity these definitions are independent of the choice of extension. Thus, if $f : M \to N$ is a smooth map between two surfaces with boundary, we may define $f'(p) : T_pM \to T_{f(p)}N$ as before when $p \in \partial M$, and the chain rule remains valid. The definition and properties of $\Lambda^k(M)$ extend immediately to the case when M is a smooth surface with boundary.*

It is easy to see that if M is a smooth $m-$dimensional surface with boundary, then $\text{Int}(M)$ is a smooth $m-$dimensional surface, but it is not so obvious that ∂M is a smooth $(m - 1)-$dimensional surface.

Proposition 12.26. *If M is a smooth $m-$dimensional surface with boundary, then ∂M is a smooth $(m - 1)-$dimensional surface.*

The main point in the proof is to show that if $p \in \partial M$ with respect to one coordinate chart, then it is in the boundary with respect to any chart. If $p \in \partial M$ there is a coordinate chart $\phi : \mathcal{O} \to U$, where $\mathcal{O}$ is an open subset of H^m and $U \ni p$ is an open subset of M. One just needs to check that

[90] An open set in H^m may well contain points where $x_1 = 0$.

[91] These notions of ∂M and $\text{Int}(M)$ must not be confused with bM and $\mathring{M}$, the formerly defined notions of the boundary and interior of M considered as subset of the metric space $(\mathbb{R}^n, |x - y|)$. See Examples (2),(3),(4) below, where the distinction is important.

[92] So if $p \in \partial M$, the space T_pM has dimension m just as for $p \in \text{Int}(M)$. Observe that T_pM and $T_p(\partial M)$ are *different* objects when $p \in \partial M$.

$\phi(\partial\mathcal{O}) = \partial U$, for then ϕ restricts to a diffeomorphism of $\partial\mathcal{O} = \mathcal{O} \cap \partial H^m$, an open set in $\mathbb{R}^{m-1}$, with $\partial U = \partial M \cap U$, an open set in ∂M containing p. The containment $\phi(\partial\mathcal{O}) \subset \partial U$ holds by definition. The reverse containment can be proved by contradiction using the inverse function theorem.[93]

If M is an oriented smooth surface with boundary, there is a natural notion of induced orientation on ∂M.

Definition 12.27 (Induced orientation on ∂M)**.** *(a) Suppose M is a smooth $m-$dimensional surface with boundary. We say that M is orientable if there is a nowhere vanishing element of $\Lambda^m(M)$; we then orient M as before.*

(b) Let M be an oriented smooth $m-$dimensional surface with boundary. Suppose $\phi : \mathcal{O}_x \to U \subset M$ is a coordinate chart with $p \in \partial U$ that determines the given orientation of U. This means that the $m-$form $dx_1 \wedge \cdots \wedge dx_m \in \Lambda^m(U)$ determines the given orientation of U.[94] *Then the* induced orientation on ∂U *is the orientation determined by the $(m-1)-$form $dx_2 \wedge \cdots \wedge dx_m$.*[95] *Using a collection of charts we define in this way an orientation of ∂M.*

Exercise. Check that this procedure determines an orientation on ∂M, that is, a nowhere vanishing element of $\Lambda^{m-1}(\partial M)$.

Remark 12.28 (Relation to the right hand rule for surfaces in $\mathbb{R}^3$)**.** *Suppose a $2-$dimensional surface M in $\mathbb{R}^3$ has an orientation given by a choice of unit normal $n(p)$ at each point. At $p \in M$ this orientation can be specified by a chart $\phi : \mathcal{O}_x \to U \subset M$ using $dx_1(p) \wedge dx_2(p)$, where the $x_j \in C^\infty(U, \mathbb{R})$,* or *by a choice of ordered basis $\{T_1, T_2\}$ for T_pM. In particular we say that each of $dx_1(p) \wedge dx_2(p)$ and $\{T_1, T_2\}$ determines the same orientation as $n(p)$ precisely when*

$$dx_1(p) \wedge dx_2(p)(T_1, T_2) > 0 \Leftrightarrow \text{ the ordered basis } \{T_1, T_2\} \text{ satisfies } \frac{T_1 \times T_2}{|T_1 \times T_2|} = n(p). \tag{12.52}$$

Now suppose $p \in \partial M$ and let us choose T_1 tangent to M but perpendicular to the boundary at p, and T_2 tangent to the boundary at p. (Recall $x_1 = 0$ defines ∂U and $x_1(q) < 0$ for $q \in$ Int(M)*). If $dx_2(p)(T_2) > 0$, which means that T_2 gives the induced orientation on $T_p(\partial M)$, and if also $dx_1(p) \wedge dx_2(p)(T_1, T_2) > 0$, so that $T_1 \times T_2$ gives the correct normal direction, then we* must *have $dx_1(p)(T_1) > 0$. That is, T_1* must point in the direction of increasing x_1.*(why?)*[96] *So, to apply the right hand rule to determine boundary orientation,* first *place your right hand on the surface near the boundary with your thumb in the given normal direction and your straightened fingers pointing in the direction* outward *from M (like T_1).* Then *curl your fingers to find the correct direction on the boundary (T_2).*[97] *Since $dx_2(p)(T_2) > 0$, that is the direction of increasing x_2.*

12.9 The generalized Stokes theorem

With all the machinery we have developed, we can now prove the generalized Stokes theorem as a direct consequence of the fundamental theorem of calculus. At this point every object appearing in the Stokes formula (12.3) has been defined.

[93] We refer to [GP], Chapter 2, for additional detail.

[94] Recall that the x_i here are the component functions of $\phi^{-1} : U \to \mathcal{O}$ *and* that $x_1(q) < 0$ for $q \in \text{Int}(U)$. Moreover, the functions $x_2, \ldots, x_m$ are coordinate functions on ∂U.

[95] We use Remark 12.25 here.

[96] Answer: $T_1 = \phi'(x_0)v$ for some v. Thus, T_1 is a tangent vector to the curve $r(s) = \phi(x_0 + sv)$ at $p = \phi(x_0)$: indeed, $r'(0) = \phi'(x_0)v$. Also, $\frac{d}{ds}|_{s=0}\, x_1(\phi(x_0 + sv)) = dx_1(p)\phi'(x_0)v = dx_1(p)T_1$.

[97] It might be better to say: bend your fingers to form a right angle with the palm of your hand to find the correct direction on the boundary.

Theorem 12.29 (Generalized Stokes theorem)**.** *Let ω be a compactly supported differential $(m-1)-$form on an oriented $m-$dimensional smooth surface M with boundary. Assume that ∂M is given the induced orientation. Then*

$$\int_M d\omega = \int_{\partial M} i^*\omega, \tag{12.53}$$

where $i : \partial M \to M$ is the inclusion map.

Proof. **1.** Using a partition of unity, we see that it suffices to prove this in the case when $\omega \in \Lambda^{m-1}(M)$ has compact support in a coordinate patch $U \subset M$. Let $\phi : \mathcal{O}_x \to U$ be a smooth chart with $\mathcal{O} \subset H^m = \{(x_1, x') \in \mathbb{R}^m : x_1 \le 0\}$. We may reduce to considering ω of the form (why?)

$$\omega = b_j dx_1 \wedge \cdots \wedge \widehat{dx_j} \wedge \cdots \wedge dx_m,$$

where $\widehat{dx_j}$ indicates an omitted factor.[98]. We have

$$d\omega = (-1)^{j-1}[\partial_{x_j}(b_j \circ \phi) \circ \phi^{-1}]dx_1 \wedge \cdots \wedge dx_m.$$

2. The case $j > 1$. In this case $i^*\omega = 0$ since $d(x_1 \circ i) = 0$. With $x'' = (x_1, \ldots, \hat{x}_j, \ldots, x_m)$ we have[99]

$$\int_M d\omega = \int_{\mathcal{O}} \phi^* d\omega = (-1)^{j-1} \int \left[\int_{-\infty}^{\infty} \partial_{x_j}(b_j \circ \phi)(x) dx_j\right] dx'' = 0. \tag{12.54}$$

3. The case $j = 1$. Set $\mathcal{O}_0 = \{x' \in \mathbb{R}^{m-1} : (0, x') \in \mathcal{O}\}$ and define the chart $\phi_0 : \mathcal{O}_0 \to \partial M$ by $\phi_0(x') = \phi(0, x')$. In this case (12.54) is replaced by

$$\begin{aligned} \int_M d\omega = \int_{\mathcal{O}} \phi^* d\omega = \int \left[\int_{-\infty}^{0} \partial_{x_1}(b_1 \circ \phi)(x) dx_1\right] dx' = \int_{\mathcal{O}_0} (b_1 \circ \phi)(0, x') dx' = \\ \int_{\mathcal{O}_0} (\phi_0)^* i^* b_1 \, dx_2 \wedge \cdots \wedge dx_m = \int_{\partial M} i^*\omega. \end{aligned} \tag{12.55}$$

□

Remark 12.30 ("Hard" versus "soft")**.** *Apart from the fundamental theorem of calculus, the proof of this theorem was largely a matter of "unraveling definitions and notation". Some of those definitions, though, depend on technically challenging - some would say "hard" as opposed to "soft" - results. For example, the definition of $\int_M \omega$ depends on the change of variable formula, a "hard" result. Also, much of our discussion of surfaces relied on the inverse function theorem, another "hard" result. For example, our proof that a smooth coordinate chart $\phi : \mathcal{O} \to U$ is actually a smooth diffeomorphism (Proposition 11.8) used "augmentations", and these were defined using the inverse function theorem. Augmentations were also used in Proposition 11.10 about the definition and properties of the derivative map $f'(p) : T_pM \to T_{f(p)}N$. The discussion of the exterior derivative used pullbacks (and thus $f'(p)$), but otherwise depended on rather "soft" results like the chain rule, product rule, and Clairaut's theorem.*

"Soft" is not a synonym for easy. "Hard" results are often, but not always, hard. Mathematicians who think the change of variable formula is a "hard" result don't all think it is hard. Some would say

[98] Here $b_j, x_j \in C^\infty(U, \mathbb{R})$.

[99] Here supp $\partial_{x_j}(b_j \circ \phi) \subset \mathcal{O} \subset H^m$ and, as usual, we extend $\partial_{x_j}(b_j \circ \phi)$ smoothly to be 0 in $H^m \setminus \mathcal{O}$. Also, observe the changes in meaning of the symbol x_i in the lines below.

that it is "straightforward" in the sense that its proof involves extending the relatively easy special case of invertible linear *changes of variable to the general case by carefully estimating the error term in the linear approximation. The fundamental theorem of calculus is a "hard" result from the 17th century. My opinion on the generalized Stokes theorem: it's a "soft" result that builds on at least 3 "hard" ones.*

The multilinear algebra results of section 10 were also essential for Theorem 12.29. Would you consider them to be "hard" or "soft"? Each of the main fields of pure mathematics -analysis, algebra, geometry and topology, logic and foundations - contains both "hard" and "soft" results. In analysis, the proof of a "hard" result usually requires careful estimates. [100]

12.10 The classical Green's, Stokes, and Divergence theorems of vector calculus

You may have studied these classical results in Math 233 or a physics course.[101] In HW 10 you will be asked to extract each of these theorems from the generalized Stokes theorem. Part of the work is done in HW 9. We will state them (mostly) in the language and notation of a vector calculus course.

Theorem 12.31 (Green's theorem)**.** *Let Ω be a bounded, connected open subset of $\mathbb{R}^2$ with a smooth boundary $\partial\Omega$ oriented counterclockwise.*[102] *Let $f, g \in C^\infty(\mathbb{R}^2, \mathbb{R})$. Then*

$$\int_\Omega (g_x - f_y)dxdy = \int_{\partial\Omega} fdx + gdy. \tag{12.56}$$

Theorem 12.32 (Stokes theorem)**.** *Let S be a smooth compact oriented $2-$dimensional surface with boundary in $\mathbb{R}^3$. Let $F = (f_1, f_2, f_3) \in C^\infty(\mathbb{R}^3, \mathbb{R}^3)$, give ∂S the induced orientation, and let n be the unit normal vector to S determined by the given orientation of S. Then*

$$\int_S (\text{curl}F \cdot n)dS = \int_{\partial S} f_1dx_1 + f_2dx_2 + f_3dx_3. \tag{12.57}$$

Note that Green's theorem is a special case of Theorem 12.32.(why?)

Theorem 12.33 (Divergence theorem)**.** *Let W be a bounded connected open set in $\mathbb{R}^3$ with smooth boundary ∂W and suppose $F = (f_1, f_2, f_3) \in C^\infty(\mathbb{R}^3, \mathbb{R}^3)$. Then*

$$\int_W \text{div}F\ dxdydz = \int_{\partial W} F \cdot n\ dS, \tag{12.58}$$

where n is the outward unit normal to ∂W.

13 Integration and mappings: degree theory

The generalized Stokes theorem is a remarkable result partly because it links two different fields of mathematics; it describes a relationship that holds in all dimensions among the purely *analytic*

[100] The words "deep", "beautiful", "surprising", and "useful" are also used to describe mathematical results. "Deep" results are usually "hard" and hard. "Beautiful" results are not necessarily "hard" or hard. The generalized Stokes theorem is "beautiful"; so is the Pythagorean theorem, and Euclid's proof that there are infinitely many prime numbers, and the ancient proof that $\sqrt{2}$ is irrational. I would say the inverse function theorem is "hard", deep, beautiful, surprising, and useful. The change of variable theorem is "hard" and useful. The contraction mapping theorem is beautiful and extremely useful.

[101] Green's theorem, at least, is usually covered in Math 233. These results used to be taught routinely in "Calc 4".

[102] The connectedness assumption is not essential. The regularity assumption here and in the next two results can also be relaxed.

operations $\int$ and d on forms, and the purely geometric (or topological) operation ∂ on smooth surfaces. The theorem is not an end in itself; the link it provides turns out to be useful. In particular one can use Stokes theorem to relate the analytic operation of integration to the topological behavior of mappings. We will explore a few of these applications.

Here is a question to motivate our study of degree theory. We know that if $f : M \to N$ is a smooth, *orientation-preserving, diffeomorphism* of two compact oriented smooth surfaces of dimension m and if $\omega \in \Lambda^m(N)$, then

$$\int_M f^*\omega = \int_N \omega. \tag{13.1}$$

This fact, a direct consequence of the definitions, is proved in HW 10. If f is orientation-reversing then (13.1) holds with a minus sign on the right. We emphasize that this holds for *any* $\omega \in \Lambda^m(N)$. Now let $f : M \to N$ be an *arbitrary* smooth map and again let ω be any element of $\Lambda^m(N)$.

$$\text{What can we say in general about the relationship between } \int_M f^*\omega \text{ and } \int_N \omega? \tag{13.2}$$

If I didn't know better, I would guess that there is absolutely nothing we could say in general. I would expect the relation between the two integrals to depend in a complicated way on particular choices of f and ω. But that is wrong. There is an amazing relationship that holds for *all* such f and all such ω. It is given by the *degree formula*,

$$\int_M f^*\omega = \deg(f) \int_N \omega, \tag{13.3}$$

where $\deg(f)$ is a constant that depends just on topological properties of f.[103] The existence of *any* such constant satisfying (13.3) for all $\omega \in \Lambda^m(N)$ would be remarkable enough, but it actually turns out that $\deg(f)$ is an *integer* with a simple geometric interpretation. To repeat: the same integer works for arbitrary $\omega \in \Lambda^m(N)$. The formula (13.3) has many applications. For example, we will use it to give a short proof of:

Theorem 13.1 (Fundamental Theorem of Algebra)**.** *Any polynomial $p(z) = a_0 + a_1 z + \cdots + a_m z^m$ of degree $m \geq 1$ has a root in $\mathbb{C}$.*[104]

Along the way we will prove some other beautiful results such as:

Theorem 13.2 (Brouwer fixed point theorem)**.** *Any smooth map $f : B \to B$, where B is the closed unit ball of $\mathbb{R}^n$, must have a fixed point; that is, $f(x) = x$ for some $x \in B$.*

Theorem 13.3 (Vector fields on spheres)**.** *There is no smooth nonvanishing vector field on S^n if n is even.*[105]

The generalized Stokes theorem is a key tool in proving the degree formula and theorem 13.2. In the next section we define the concept of homotopy of maps. This makes precise the sense in which a map $f : M \to N$ might be "continuously deformed".

[103] The number doesn't change if f is "deformed in a continuous way". We'll make that precise shortly. In general, the word "topological" is applied to properties of sets or mappings that are invariant under continuous deformations.

[104] Once we know there is at least one root, it is a simple matter to show there are exactly m roots in $\mathbb{C}$.(why?)

[105] You can't comb the hair on a coconut!

13.1 Prelude: Brouwer results

The next proposition is fundamental for degree theory and has other nice applications as well. It is a simple consequence of Stokes theorem.

Proposition 13.4. *Let M and N be compact oriented smooth surfaces of dimension m, and suppose $M = \partial W$, where W is a compact oriented smooth surface of dimension $m+1$. Suppose $f : M \to N$ is a smooth map which extends smoothly to all of W. Then for every $\omega \in \Lambda^m(N)$ we have $\int_M f^*\omega = 0$.*

Proof. Let $F : W \to N$ be a smooth extension of f and $i : \partial W \to W$ the inclusion map. Since $F = f$ on $M = \partial W$,

$$\int_M f^*\omega = \int_{\partial W} (F \circ i)^*\omega = \int_{\partial W} i^*F^*\omega = \int_W dF^*\omega = \int_W F^*d\omega = 0, \tag{13.4}$$

since $d\omega \in \Lambda^{m+1}(N)$ and $\dim N = m$. □

Our first application of Proposition 13.4 will be to prove some classical topological results of Brouwer. The first is the *no retraction* theorem.

Definition 13.5. *Let W be a smooth surface with boundary ∂W. A* retraction *of W onto its boundary is a map $\phi : W \to \partial W$ (not necessarily smooth) such that $\phi(p) = p$ for all $p \in \partial W$.*

Theorem 13.6 (No retraction theorem)**.** *Let W be a smooth oriented $(m+1)-$dimensional surface with* nonempty *boundary ∂W. There is no* smooth *retraction.*

Before proving this theorem we define the notion of the "volume form", ω_M, on an oriented smooth surface M. It is closely related to the volume (or "area"') element dS we defined earlier.

Definition 13.7 (Volume form)**.** *Let M be an oriented smooth $m-$dimensional surface and suppose $\phi : \mathcal{O}_x \to U \subset M$ is any orientation-preserving chart on M; that is, the chart determines the given orientation on M. We define ω_M on M by setting $\omega_M|_U = \sqrt{g} \cdot dx_1 \wedge \cdots \wedge dx_m$ for any such chart. Here, as before,*

$$g = \det G, \ \text{where } G(x) = \phi'(x)^t\phi'(x) \tag{13.5}$$

is the metric tensor on M.

Let us check that ω_M is well-defined on M. If $\psi : \Omega_y \to U$ is another orientation-preserving chart and $p \in U$, we must check that

$$\sqrt{g(x(p))} \cdot dx_1(p) \wedge \cdots \wedge dx_m(p) = \sqrt{h(y(p))} \cdot dy_1(p) \wedge \cdots \wedge dy_m(p), \tag{13.6}$$

where $h(y) = \det H(y)$ and $H(y) = \psi'(y)^t\psi(y)$. We showed earlier that if $F = \psi^{-1} \circ \phi$, then

$$dy_1(p) \wedge \cdots \wedge dy_m(p) = \det F'(x(p))\, dx_1(p) \wedge \cdots \wedge dx_m(p). \tag{13.7}$$

Now $\det F' > 0$ since both ϕ and ψ are orientation-preserving, so (13.6) holds.[106]

Remark 13.8. *Since $\int_M \omega_M = \int_M dS = \mathrm{vol}(M)$, we call ω_M the volume form.*

[106] Here we use the fact proved earlier that $\sqrt{h(y(p))}|\det F'(x(p))| = \sqrt{g(x(p))}$.

Proof of Theorem 13.6. If a smooth retraction exists, then the identity map $i : \partial W \to W$ extends smoothly to W. Let ω_M be the volume form on ∂W. Proposition 13.4 implies $\int_{\partial W} \omega_M = \int_{\partial W} i^*\omega_M = 0$, a contradiction.

□

Brouwer's fixed point theorem is a corollary of Theorem 13.6.

Corollary 13.9 (Brouwer fixed point theorem)**.** *If $F : B \to B$ is a smooth map on the closed unit ball in $\mathbb{R}^n$, then F has a fixed point; that is, there exists an x in B such that $F(x) = x$.*

Proof. If there is no fixed point, then $F(x) \neq x$ for all $x \in B$. Define $\phi(x)$ to be the point at which the ray from $F(x)$ to x hits ∂B (see Figure 5.3.1 of the text). An explicit formula is

$$\begin{aligned} &\phi(x) = x + t(x - F(x)), \quad t = \frac{\sqrt{b^2 + 4ac} - b}{2a}, \\ &a = |x - F(x)|^2, b = 2x \cdot (x - F(x)), c = 1 - |x|^2, \end{aligned} \tag{13.8}$$

where t is chosen so that $|x + t(x - F(x))|^2 = 1$. The quantity $b^2 + 4ac$ is positive on B, so t is a smooth function of x. Thus, ϕ is a smooth retraction. Contradiction. □

13.2 Homotopy and isotopy

Let M and N be smooth surfaces and consider two smooth maps $f_i : M \to N$, $i = 0, 1$. We think of f_1 as a "smooth deformation" of f_0 if f_0 and f_1 can be "joined by" a smoothly evolving family of maps $f_t : M \to N$. The concept of homotopy, which makes this precise, is one of the fundamental concepts of topology.

Definition 13.10 (Homotopy)**.** *We say that f_0 and f_1 are* homotopic *and write $f_0 \sim f_1$ if there exists a smooth map $F : I \times M \to N$ such that $F(0, x) = f_0(x)$ and $F(1, x) = f_1(x)$; here $I = [0, 1]$. We call F a* homotopy *between f_0 and f_1.*

Note that $\sim$ is an equivalence relation when the maps involved are only required to be continuous. Suppose $f_0 \sim f_1$ with homotopy F and $f_1 \sim f_2$ with homotopy G. Then $H : I \times M \to N$ defined by

$$H(t, x) := \begin{cases} F(2t, x), 0 \leq t \leq \frac{1}{2} \\ G(2t - 1, x), \frac{1}{2} \leq t \leq 1 \end{cases} \tag{13.9}$$

is a continuous homotopy showing $f_0 \sim f_2$. You can easily check reflexivity and symmetry of the relation $\sim$. Observe that the map H in (13.9) is not necessarily smooth, even if F and G are (why not?) Nevertheless, it it is still true that $\sim$ is an equivalence relation when F is required to be smooth.[107]

The next proposition discusses the orientation of the homotopy space $I \times M$ and its boundary.

Proposition 13.11. *Let M be a smooth oriented $m-$dimensional surface oriented by $\omega \in \Lambda^m(M)$. Then the "homotopy space" $I_t \times M$ is a smooth oriented $(m+1)-$dimensional surface with boundary oriented by $dt \wedge \omega$ and (with natural notation):*

$$\partial(I \times M) = M_1 - M_0. \tag{13.10}$$

[107] Reflexivity and symmetry are proved just as in the continuous case. To prove transitivity note first that if f_0 and f_1 are homotopic, there exists a homotopy $\tilde{F}(t, x)$ such that $\tilde{F}(t, x) = f_0(x)$ for $t \in [0, \frac{1}{4}]$ and $\tilde{F}(t, x) = f_1(x)$ for $t \in [\frac{3}{4}, 1]$. For example, if $F(t, x)$ is any homotopy and $\rho \in C_c^\infty(\mathbb{R}, \mathbb{R})$ is 0 for $t \leq \frac{1}{4}$ and 1 for $t \geq \frac{3}{4}$, one can take $\tilde{F}(t, x) = F(\rho(t), x)$.

That is, the surface $\partial(I \times M)$ is a union of two disjoint copies of M; with their induced orientations they are M_1 and $-M_0$.

Proof. Let $p \in M$ and suppose $\psi_i : \mathcal{O}_x \to U \subset M$ are smooth charts such that when $i = 1$, $dx(p) := dx_1 \wedge \ldots dx_m(p) \sim \omega(p)$ and when $i = 0$, $dx(p) \sim -\omega(p)$.[108] Next define charts on $I \times M$ near $(1, p)$ and $(0, p)$ as follows. Set

$$\begin{aligned} &\phi_1 : (-\epsilon, 0] \times \mathcal{O} \to V_1, \text{ where } \phi_1(y_1, x) = (y_1 + 1, \psi_1(x)) = (t, q) \in V_1 \\ &\phi_0 : (-\epsilon, 0] \times \mathcal{O} \to V_0, \text{ where } \phi_0(y_1, x) = (-y_1, \psi_0(x)) = (t, q) \in V_0. \end{aligned} \tag{13.11}$$

Letting y_1 and x_j denote components of ϕ_i^{-1} we see that

$$\begin{aligned} &(dy_1 \wedge dx)(1, p) = (dt \wedge \omega)(1, p) \text{ when } i = 1 \text{ but} \\ &(dy_1 \wedge dx)(0, p) = (-dt \wedge -\omega)(0, p) \text{ when } i = 0. \end{aligned} \tag{13.12}$$

Thus, the induced orientation on M_1 is given by ω, while the induced orientation on M_0 is given by $-\omega$.

□

With this we prove the following corollary of Proposition 13.4.

Corollary 13.12. *Suppose $f_1, f_0 : M \to N$ be homotopic maps of smooth, compact, oriented $m-$dimensional surfaces. Then for every $\omega \in \Lambda^m(N)$ we have*

$$\int_M f_0^* \omega = \int_M f_1^* \omega. \tag{13.13}$$

Proof. Let $F : I \times M \to N$ be a homotopy. Recall $\partial(I \times M) = M_1 - M_0$ and set $\partial F := f_i$ on M_i. We have

$$0 = \int_{\partial(I \times M)} (\partial F)^* \omega = \int_{M_1} (\partial F)^* \omega - \int_{M_0} (\partial F)^* \omega = \int_M f_1^* \omega - \int_M f_0^* \omega, \tag{13.14}$$

where the first equality follows from Proposition 13.4. □

The next proposition generalizes Corollary 13.12; its proof is essentially the same.[109]

Proposition 13.13. *Let M be a compact oriented $m-$dimensional smooth surface and N a smooth surface, not necessarily of the same dimension. Suppose $\omega \in \Lambda^m(N)$ is closed (i.e. $d\omega = 0$.) If $f_0, f_1 : M \to N$ are homotopic, then $\int_M f_0^* \omega = \int_M f_1^* \omega$.*

We will need a special type of homotopy in part of our treatment of degree theory.[110]

Definition 13.14. *An* isotopy *is a homotopy $F : I \times M \to N$ in which each map $F_t : M \to N$ is a diffeomorphism. Two diffeomorphisms are* isotopic *if they can be joined by an isotopy. An isotopy is* compactly supported *if the maps F_t are all equal to the identity outside some fixed compact subset of M.*

Lemma 13.15 (Isotopy lemma)**.** *Let M be a smooth* connected *$m-$dimensional surface. Given points $p, q \in M$ there exists a diffeomorphism $k : M \to M$ such that $k(p) = q$ and k is isotopic to the identity map. The isotopy may be taken to be compactly supported.*

[108] When $i = 1$, the x_j here are the components of ψ_1^{-1}. When $i = 0$, they are the components of ψ_0^{-1}.

[109] The solution to problem 6a of HW 10 gives the details in a special case.

[110] Much of our treatment in section 13 follows [GP].

Proof. **1.** If the statement is true for points p and q, let us call them "isotopic" and write $p \sim q$. The relation $\sim$ is an equivalence relation on M, so M is a disjoint union of equivalence classes. If we show each equivalence class is open, there can be only one since M is connected. To finish, given $p \in M$, we will show it is isotopic to every near enough point q.

2. Fix $r > 0$ and $0 < \delta < r$. First we construct an isotopy $g : I \times \mathbb{R}^m \to \mathbb{R}^m$ such that

$$g_0 = Id,\ g_1(0) = z \in \mathbb{R}^m,\ g_t = Id \text{ outside } B(0,\delta), \tag{13.15}$$

where z is any given point close enough to 0. First assume that $z = (z_1, 0)$. Choose $\chi \in C_c^\infty(\mathbb{R}^m, \mathbb{R})$ such that

$$\operatorname{supp} \chi \subset B(0,\delta) \text{ and } \chi(0) = 1. \tag{13.16}$$

With $x = (x_1, x')$ let $g_t(x) = (x_1 + t\chi(x)z_1, x')$; then the conditions (13.15) hold.

We must show each $g_t : \mathbb{R}^m \to \mathbb{R}^m$ is a diffeomorphism when $|z|$ is small. Bijectivity holds because it holds on each line $x' = c$.[111] To prove smoothness of g_t and g_t^{-1}, a *local* property, we use the inverse function theorem. We have

$$g_t'(x) = \begin{pmatrix} 1 + t\chi_{x_1}(x)z_1 & * \\ 0 & I_{m-1} \end{pmatrix}, \tag{13.17}$$

so for $|z|$ small enough, $g_t'(x)$ is invertible for all x. Thus, g is smooth with smooth inverse by the inverse function theorem.

To treat general z first apply a rotation R such that $Rz = (|z|, 0)$. Then take $g_t = R^{-1}\tilde{g}_t R$, where $\tilde{g}_t$ is constructed as above.

3. Given $p_0 \in M$, choose a chart $\phi : B(0,r) \to U \subset M$ such that $\phi(0) = p_0$. For $0 < \delta < r$ and $q = \phi(z)$ close enough to p_0 construct $g : I \times \mathbb{R}^m \to \mathbb{R}^m$ as in step **2**. To finish set

$$h(t,p) = \begin{cases} \phi(g(t, \phi^{-1}(p))),\ p \in U \\ Id,\ p \notin U \end{cases}.$$

The function $h : I \times M \to M$ satisfies $h_0 = Id$, $h_1(p_0) = q$, $h_t = Id$ outside U, and each $h_t : M \to M$ is a diffeomorphism. ☐

13.3 Regular and critical values of mappings

This section presents some general results on the properties of smooth mappings $f : M \to N$ between smooth surfaces. These results are of interest in themselves and will also serve as preparation for our study of the degree formula.

Definition 13.16 (Regular and critical values)**.** *(a) Let M and N be smooth surfaces and suppose $f : M \to N$ is smooth. A point $q \in N$ is called a* regular value *for f if $f'(p) : T_pM \to T_{f(p)}N$ is surjective at every point p such that $f(p) = q$. Otherwise, q is said to be a* critical value.

(b) A point $p \in M$ is a critical point *of f if $f'(p) : T_pM \to T_{f(p)}N$ is not surjective, and denote by $\mathcal{C} \subset M$ the set of all critical points. Thus, $f(\mathcal{C}) = \{$critical values of $f\}$.*

If q is not in the range of f, it is automatically a regular value (why?). The set of critical points $\mathcal{C}$ can be large. For example, if f is a constant map, we have $\mathcal{C} = M$.

[111] To see that use the fact that if $|z|$ is small enough, $\partial_{x_1} g_t(x) > \frac{1}{2}$ for all x and all $t \in [0,1]$.

We will be mainly interested in the case where M and N have the same dimension. In that case if q is a regular value, then f is a local diffeomorphism at each preimage point (why?). The following theorem tells us that "most" points in N are regular values.[112]

Theorem 13.17 (Sard's theorem)**.** *Let M and N be smooth surfaces of the same dimension and suppose $f : M \to N$ is smooth. For any compact set $K \subset M$, the set $f(\mathcal{C} \cap K)$ is a nil subset of N.*

The theorem is a consequence of the following Proposition about maps on $\mathbb{R}^m$.[113]

Proposition 13.18. *Let $\mathcal{O}$ be an open subset of $\mathbb{R}^m$. If $F : \mathcal{O} \to \mathbb{R}^m$ is a C^2 map, $\mathcal{C} \subset \mathcal{O}$ its set of critical points, and $K \subset \mathcal{O}$ compact, then $F(\mathcal{C} \cap K)$ is a nil subset of $\mathbb{R}^m$.*

Proof. **1.** Fix $\delta > 0$. We can assume K is a cubical cell. Let $\mathcal{P}$ be a partition of K into cubical cells R_α with diam $R_\alpha = \delta$ for all α. Let $\mathcal{P}'$ consist of the cells in $\mathcal{P}$ that intersect $\mathcal{C}$. For $R_\alpha \in \mathcal{P}'$ pick $x_\alpha \in \mathcal{C} \cap R_\alpha$.

2. For any $x_\alpha + y \in R_\alpha$ we have by Taylor's formula.

(13.18)

$$F(x_\alpha + y) = F(x_\alpha) + F'(x_\alpha)y + r_\alpha(y), \text{ where } r_\alpha(y) = \sum_{|\beta|=2} \frac{2}{\beta!} \left(\int_0^1 (1-s)\partial^\beta F(x_\alpha + sy)ds \right) y^\beta.$$

Choose M such that $|\partial^\beta F(x)| \leq M$ for all $x \in K$, $1 \leq |\beta| \leq 2$. Then we have for all α:

$$|r_\alpha(y)| \leq C(M)\delta^2 \text{ for all } x_\alpha + y \in R_\alpha. \tag{13.19}$$

Thus, $F(R_\alpha)$ is contained in a $C(M)\delta^2$ neighborhood of the set $H_\alpha = F(x_\alpha) + F'(x_\alpha)(R_\alpha - x_\alpha)$. Because $F'(x_\alpha)$ is singular, H_α is a parallelepiped of dimension $\leq m-1$ and diameter $\leq M\delta$. Thus[114]

$$\text{cont}^+ F(R_\alpha) \leq C(m)(M\delta)^{m-1}C(M)\delta^2 = C(m,M)\delta^m\delta = \tilde{C}(m,M)V(R_\alpha)\delta, \tag{13.20}$$

which implies

$$\text{cont}^+ F(\mathcal{C} \cap K) \leq \sum_{R_\alpha \in \mathcal{P}'} \text{cont}^+ F(R_\alpha) \leq \tilde{C}(m,M)V(K)\delta. \tag{13.21}$$

□

Theorem 13.19 (Stack of records theorem)**.** *Suppose $f : M \to N$ is a smooth map of $m-$dimensional surfaces, where M is compact, and let $q \in N$ be a regular value of f. Then $f^{-1}(q)$ is a finite set. If $f^{-1}(q) = \{p_1, \dots, p_N\}$ is nonempty, there exists a open set $V \ni q$ in N such that $f^{-1}(V) = U_1 \cup \cdots \cup U_N$, where the U_i are disjoint open sets of M, $U_i \ni p_i$, and for each i, $f : U_i \to V$ is a diffeomorphism. We may take V to be connected.*

Proof. **1.** Suppose $S = f^{-1}(q)$ is infinite. Then since M is compact there exists a sequence p_n of distinct points in S such that $p_n \to p \in M$. By continuity of f we must have $p \in S$. Since $f'(p)$ is invertible, there exists an open set $U \ni p$ such that $f : U \to f(U)$ is a diffeomorphism. But f is not $1-1$ on U.(why?) Contradiction.

[112] We state here a modified special case of the classical Sard's theorem, which says that if $f : M \to N$ is a smooth map (where M and N may have different dimensions), then the set of critical values has Lebesgue measure zero in N.

[113] Use charts and a partition of unity to see this.

[114] Here we use $V(R_\alpha) = C(m)\delta^m$.

2. Suppose S is nonempty. We can choose disjoint open sets $W_i \ni p_i$ such that each $f : W_i \to f(W_i)$ is a diffeomorphism. Then $M \setminus \cup_i W_i$ is compact, and thus $f(M \setminus \cup_i W_i) := K$ is compact and does not contain q. Now define

$$\begin{aligned} V &= f(W_1) \cap \cdots \cap f(W_N) \cap K^c \\ U_i &= f^{-1}(V) \cap W_i. \end{aligned} \tag{13.22}$$

We claim these (open) sets have the stated properties. Clearly, $f : U_i \to f(U_i)$ is a diffeomorphism. (why?)

3. First we check $f(U_i) = V$. The containment $\subset$ is clear. For the other direction, let $q' \in V$. Then $q' \in f(W_i)$ for all i. So $q' = f(p_i')$ for some $p_i' \in W_i$. Also $p_i' \in f^{-1}(V)$ since $q' \in V$. So $p_i' \in U_i$ and thus $q' \in f(U_i)$.

4. Next we check $f^{-1}(V) = U_1 \cup \cdots \cup U_N$; note the right side equals $f^{-1}(V) \cap (\cup_i W_i)$. The containment $\supset$ is clear. For the other direction let $p' \in f^{-1}(V)$. Then $f(p') \in V$, which implies $p' \in \cup_i W_i$, since otherwise $f(p') \in K$.

5. To obtain a connected set V with the desired properties, first choose V as above, and then take the component of V that contains q.

□

13.4 The degree formula

First we prove a local version of the degree formula (13.3).

Proposition 13.20. *[GP] Let $f : M \to N$ be a smooth map of two smooth compact oriented $m-$dimensional surfaces and let $q \in N$ be a regular value of f. There exists an open set $V \ni q$ and an integer D (specified below) such that the formula*

$$\int_M f^*\omega = D \cdot \int_N \omega \tag{13.23}$$

holds for every $\omega \in \Lambda^m(N)$ with support in V.

Proof. **1.** If $q \notin f(M)$, choose $V \ni q$ disjoint from $f(M)$ (how?). Then if ω has support in V, we have $f^*\omega = 0$ (why?), so the formula (13.23) holds with $D = 0$.

2. Now suppose $q \in f(M)$ and is regular. Then $f^{-1}(q) = \{p_1, \ldots, p_N\}$ and we now choose a connected open set V and sets U_i exactly as in Theorem 13.19. If ω has support in V, then $f^*\omega$ has support in $f^{-1}(V) = \cup_i U_i$. Thus,

$$\int_M f^*\omega = \sum_{i=1}^N \int_{U_i} f^*\omega. \tag{13.24}$$

Since $f : U_i \to V$ is a diffeomorphism, we have

$$\int_{U_i} f^*\omega = \sigma_i \int_V \omega, \tag{13.25}$$

where σ_i is ± 1 depending on whether $f : U_i \to V$ preserves or reverses orientation. This proves (13.23) with

$$D = \sum_{i=1}^N \sigma_i. \tag{13.26}$$

□

We would like to define $\deg(f) = D$ for D as in Proposition 13.20, but it is not clear yet that the value of D is independent of the choice of regular value q. The next theorem shows that if q is *any* regular value and if we define D as in Proposition 13.20, then the formula (13.23) actually holds for *any* $\omega \in \Lambda^m(N)$. This implies the value of D is independent of the choice of regular value q.(why?)

Theorem 13.21 (Degree formula)**.** *Let $f : M \to N$ be a smooth map of two smooth compact oriented $m-$dimensional surfaces. Assume N is connected and let $q \in N$ be any regular value of f. Define the integer D for that regular value as in Proposition 13.20. Then for every $\omega \in \Lambda^m(N)$ we have*

$$\int_M f^*\omega = D \cdot \int_N \omega. \tag{13.27}$$

This theorem shows that the following definition is well-defined.

Definition 13.22 (Degree of f)**.** *Let $f : M \to N$ be a smooth map of two smooth compact oriented $m-$dimensional surfaces. Assume N is connected. The* degree of f *is the integer D such that* (13.27) *holds for all $\omega \in \Lambda^m(N)$. We can compute D by selecting any regular value q of f and setting $D = \sum_i \sigma_i$ as in Proposition 13.20, where we take the sum to be zero if $q \notin f(M)$.*

Proof of Theorem 13.21. We'll prove the theorem using Proposition 13.20 and the Isotopy Lemma. Let $\omega \in \Lambda^m(N)$.

1. Let $q \in N$ be any regular value of f. Choose an open set $V \ni q$ and define an integer D as in Proposition 13.20. The objects q, V, D are fixed throughout the proof.

2. By the Isotopy Lemma for every point $z \in N$ we can find a diffeomorphism $h_z : N \to N$ that is isotopic to the identity map and that carries q to z. Thus the collection $\{h_z(V) : z \in N\}$ is an open cover of N. By compactness of N there is a finite subcover $\{h_1(V), \ldots, h_n(V)\}$. With a partition of unity we can reduce to proving (13.27) for ω supported in some $h_j(V)$.

3. Let $\omega \in \Lambda^m(N)$ have support in some $h_j(V)$. Since $h_j \sim Id$, then $h_j \circ f \sim f$ (why?), and so Corollary 13.12 implies with $h_j = h$:

$$\int_M f^*\omega = \int_M (h \circ f)^*\omega = \int_M f^*h^*\omega. \tag{13.28}$$

Now $h^*\omega$ is supported in V, so the local formula of Proposition 13.20 implies

$$\int_M f^*(h^*\omega) = D \cdot \int_N h^*\omega. \tag{13.29}$$

Since $h \sim Id$ another application of Corollary 13.12 gives

$$\int_N h^*\omega = \int_N \omega. \tag{13.30}$$

With (13.28) and (13.29) this implies

$$\int_M f^*\omega = D \cdot \int_N \omega. \tag{13.31}$$

□

Corollary 13.23 (Homotopy invariance of degree)**.** *Homotopic maps have the same degree.*

Proof. Let f_0, f_1 be homotopic maps of M to N for M,N as in Theorem 13.21, and let ω_N be the volume form on N. Then using Corollary 13.12 we obtain

$$\deg\, f_0 = \frac{\int_M f_0^*\omega_N}{\int_N \omega_N} = \frac{\int_M f_1^*\omega_N}{\int_N \omega_N} = \deg\, f_1. \tag{13.32}$$

□

Remark 13.24. This proof of the degree formula is slightly different from the proof in [GP] which inspired it. The authors of [GP] define the degree of f and show it is well defined using the apparatus of *intersection theory* well before they consider the degree formula. Thus, they know the degree is well-defined before they prove their analogue of our Theorem 13.21. Their proof also uses the Isotopy Lemma. We do not know degree is well-defined until after proving Theorem 13.21.

The proof of the degree formula in [MT] relies on a local result like our Proposition 13.20 and Proposition 5.3.6 of [MT]. The latter proposition is of much interest in its own right, but its proof is more difficult than that of the Isotopy Lemma. Proposition 5.3.6 is not accessible to us yet because it depends on several tools we have not had time to study: for example, Haar measure on $\mathrm{SO}(n+1)$, properties of the map $\mathrm{Exp} : \mathrm{Skew}(n+1) \to \mathrm{SO}(n+1)$, the Lie derivative, and some properties of vector fields on surfaces.

Our main point here is to note that degree can be shown to be well-defined and the degree formula can be proved just using the relatively simple tools of Proposition 13.20 and the Isotopy Lemma, Lemma 13.15.

13.5 Applications

Our first application of the degree formula is to prove the Fundamental Theorem of Algebra.

Theorem 13.25 (Fundamental Theorem of Algebra)**.** *Let $m \in \{1, 2, 3, \dots\}$ and $p(z) = z^m + a_{m-1}z^{m-1} + \cdots + a_0$. Then p has at least one zero in $\mathbb{C}$.*

Proof. **1.** For $R > 0$ let B_R be the closed ball in $\mathbb{C}$ of radius R centered at the origin. Define a homotopy between $p(z)$ and z^m by

$$p_t(z) = tp(z) + (1-t)z^m = z^m + t(a_{m-1}z^{m-1} + \cdots + a_0), \quad t \in [0, 1]. \tag{13.33}$$

If R is large enough, p_t has no zeros on ∂B_R since

$$\frac{p_t(z)}{z^m} = 1 + t\left(\frac{a_{m-1}}{z} + \cdots + \frac{a_0}{z^m}\right). \tag{13.34}$$

Thus, for R large enough the map $F_t : \partial B_R \to S^1$ given by $F_t(z) = \frac{p_t(z)}{|p_t(z)|}$ is a homotopy between the maps $\frac{p(z)}{|p(z)|}$ and $\frac{z^m}{R^m}$. Thus, these maps have the same degree by Corollary 13.23.

2. We claim the map $z \to z^m$ has degree m. To see this is, note that the map is orientation preserving and that each $p \in S^1$ has m preimages under the map.[115] Thus, $\sum_i \sigma_i = m$ (recall (13.26)).

3. Since $\frac{p(z)}{|p(z)|}$ has degree $m > 0$, by Proposition 13.4 it must be impossible to extend $p/|p|$ smoothly to all of B_R. That can be so only if p has at least one zero in the interior of B_R.

□

[115] You can check that the map is orientation preserving by using charts $t \to (\cos t, \sin t)$, or by observing that as z traverses S^1 one time in the counterclockwise direction, the image z^m traverses S^1 m times in the counterclockwise direction.

Next we give a few exercises that will lead you to prove a classical topological result about spheres.

Definition 13.26. *(a) A* vector field *on a smooth surface $M \subset \mathbb{R}^n$ is a smooth map $v : M \to \mathbb{R}^n$ such that $v(p) \in T_pM$ for all $p \in M$.*

(b) A point $p \in M$ is a zero *of the vector field v if $v(p) = 0$.*

Exercises[116]

1. Show that if k is odd, there exists a vector field v on S^k (the unit sphere in $\mathbb{R}^{k+1}$) having no zeros. Hint: For $k = 1$ use $(x_1, x_2) \to (-x_2, x_1)$.

It is a remarkable and subtle topological fact that nonvanishing vector fields do not exist on the even-dimensional spheres. We now have the tools to give a short proof of that.

2. Prove that if S^k has a nonvanishing vector field, then its antipodal map, $A : S^k \to S^k$ given by $A(p) = -p$, is homotopic to the identity map. (Hint: Show that you may take $|v(p)| = 1$ for all p. Now rotate p to $-p$ in the direction indicated by $v(p)$.)

3. (a) Compute the degree of the antipodal map $A : S^k \to S^k$ for all $k \in \mathbb{N}$.

(b) Show that if k is even, there is no nonvanishing vector field on S^k. This is Theorem 13.3.

14 Existence and uniqueness of solutions to ordinary differential equations.

In this section we prove classical results on the existence and uniqueness of solutions to ordinary differential equations (ODEs). We begin with a short-time existence result for general nonlinear systems of ODEs that follows readily from the contraction mapping theorem (CMT). Then we prove a uniqueness theorem for C^1 solutions. Next we consider the question of when solutions exist for all time - "global existence". A corollary of this study will be the "fundamental theorem of Math 383", namely, the existence and uniqueness of solutions for all time of linear systems of ODEs.[117]

14.1 Basic existence and uniqueness theory

Consider the following initial value problem for an $n \times n$ system of ODEs

$$\frac{dy}{dt} = F(t, y), \quad y(t_0) = y_0. \quad (IVP) \tag{14.1}$$

Here $y_0 \in \mathbb{R}^n$ and the unknown y is a function of t taking values in $\mathbb{R}^n$.

Theorem 14.1 (Local existence)**.** *Let $y_0 \in \Omega$, an open subset of $\mathbb{R}^n$, and let $I \subset \mathbb{R}$ be an open interval containing t_0. Suppose $F : I \times \Omega \to \mathbb{R}^n$ is continuous and satisfies the following "Lipschitz" estimate in y. There exists $L > 0$ such that*

$$|F(t, y_1) - F(t, y_2| \leq L|y_1 - y_2| \text{ for } t \in I, y_j \in \Omega. \tag{14.2}$$

Then the IVP (14.1) *has a C^1 solution on some open $t-$interval containing t_0.*

[116] These will be part of HW 11.

[117] In this section we follow section 2.3 of our text [MT], but provide a lot of additional detail.

Remark 14.2. *1. We call this a "local existence theorem" because it asserts existence of the solution only on some (possibly very small) open time interval containing* t_0*. A theorem that asserted existence of a solution on the full interval* I *for which* F *is defined would be called a "global existence theorem".*

2. In practice the function F *often satisfies conditions that make the Lipschitz condition easy to verify. For example, suppose* $F \in C^1(\mathbb{R} \times \mathbb{R}^n; \mathbb{R}^n)$*. Let* I *be* any *bounded open interval and let* $\Omega \subset \mathbb{R}^n$ *be any* convex *bounded open set. Then since* $\overline{I \times \Omega}$ *is compact we have for* $t \in I$*,* $y_j \in \Omega$*:*

$$F(t, y_1) - F(t, y_2) = \left(\int_0^1 F_y(t, y_2 + s(y_1 - y_2))ds \right)(y_1 - y_2), \tag{14.3}$$

so the Lipschitz condition is satisfied provided L *is chosen so that* $|F_y(t, y)| \leq L$ *on* $\overline{I \times \Omega}$*.(Why is convexity needed here?)*

Proof of Theorem 14.1. **1.** First note that the IVP (14.1) is equivalent to the following *integral equation* for the unknown function y:

$$y(t) = \int_{t_0}^t F(s, y(s))ds + y_0. \tag{14.4}$$

If y satisfies (14.4), differentiate both sides to see that y satisfies the IVP. If y satisfies the IVP, integrate both sides $\int_{t_0}^t$ to see that y satisfies (14.4). Henceforth we focus on solving (14.4), which has the advantage of being readily solvable by the contraction mapping theorem (CMT).

2. A classical approach to solving (14.4) is to use *Picard iteration.* One defines a sequence of "iterates" y_k, which we hope give better and better approximations to a solution as $k \to \infty$. Take

$$y_0(t) = y_0, \;\; y_1(t) = \int_{t_0}^t F(s, y_0(s))ds + y_0, \;\; \ldots, \;\; y_k(t) = \int_{y_0}^t F(s, y_{k-1}(s))ds + y_0. \tag{14.5}$$

This should remind you of the proof of the contraction mapping theorem. Instead of reproving the CMT in this context by showing that the y_k converge to a solution, we will use the CMT directly, without mention of "iterates".

3. In order to apply the CMT we have to exhibit two things: a complete metric space X on which to work, and an appropriate contraction $\Phi : X \to X$. We want to choose X and Φ so that a fixed point of Φ is a solution to (14.4). The choice of Φ is pretty obvious:

$$(\Phi y)(t) = \int_{t_0}^t F(s, y(s))ds + y_0 \text{ where } y \in X. \tag{14.6}$$

The choice of X requires some care. It must be a complete *function space.* A first guess might be $C(J; \mathbb{R}^n)$, where $J = [t_0 - T, t_0 + T]$ for some small enough $T > 0$ to be chosen. This space with the usual sup norm is complete (why?), but it obviously can't work because its definition ignores Ω; many functions $y \in X$ take values outside Ω and for such values $F(s, y(s))$ is not defined, so (14.6) makes no sense. Instead, we consider functions in $C(J; \mathbb{R}^n)$ with a restricted range and take $X = X_R$ where

$$X_R = \{u \in C(J, \mathbb{R}^n) : u(t_0) = y_0, \;\; \sup_{t \in J} |u(t) - y_0| \leq R\}, \tag{14.7}$$

the constant R is picked so that $\overline{B(y_0, R)} = \{y \in \mathbb{R}^n : |y - y_0| \leq R\} \subset \Omega$, and we suppose $J \subset I$. With the metric

$$d(f, g) := \sup_{s \in J} |f(s) - g(s)| \tag{14.8}$$

the metric space X_R is complete (check!).[118]

Our task now is to choose $T > 0$ so that two conditions are satisfied: $\Phi : X_R \to X_R$ and Φ is a contraction.[119] After doing that, we apply CMT and are done.

4. First we choose T to guarantee $\Phi : X_R \to X_R$. There exists $M > 0$ such that

$$\sup_{s\in J, |y-y_0|\le R} |F(s,y)| \le M. \tag{14.9}$$

If $y \in X_R$, clearly $(\Phi y)(t_0) = y_0$ and

$$|(\Phi y)(t) - y_0| \le |\int_{t_0}^{t} F(s, y(s))ds| \le TM \le R \text{ for all } t \in J, \tag{14.10}$$

provided we take $T \le \frac{R}{M}$. That is, $\Phi : X_R \to X_R$ for $T \le \frac{R}{M}$. This is our first restriction on T.

5. Existence. Similarly, for $y, z \in X_R$ and $t \in J$, we have using the Lipschitz condition:

$$|(\Phi y)(t) - (\Phi z)(t)| \le \left| \int_{t_0}^{t} L|y(s) - z(s)|ds \right| \le TL \sup_{s\in J} |y(s) - z(s)|. \tag{14.11}$$

Thus, Φ is a contraction on X_R provided $T > 0$ also satisfies $T < \frac{1}{L}$. Thus, if we take

$$T = \min\left\{ \frac{R}{M}, \frac{1}{2L} \right\}, \tag{14.12}$$

the CMT gives a unique solution y of (14.4) in X_R. Furthermore, the continuity of F and the equation (14.4) imply that $y \in C^1(J, \mathbb{R}^n)$.

6. Uniqueness. Although the CMT gives a unique solution to the IVP (14.1) in X_R, this does not directly allow to assert that C^1 solutions to the IVP on J are unique. (why?) Suppose y and z are two solutions of the IVP on J. Since both solutions satisfy (14.4) we obtain for $t \in J$:

$$|y(t) - z(t)| \le \left| \int_{t_0}^{t} L|y(s) - z(s)|ds \right| \le TL \sup_{s\in J} |y(s) - z(s)| \le \frac{1}{2} \sup_{s\in J} |y(s) - z(s)|, \tag{14.13}$$

and thus $y = z$ on J.

□

The next proposition gives a better uniqueness result that applies to intervals of any length. It provides yet another illustration of the power of "connectedness arguments".

Proposition 14.3. *[Uniqueness] Consider the IVP* (14.1)*, where $t_0 \in I$, $y_0 \in \Omega$, and $F : I \times \Omega$ is as in Theorem 14.1. Let $I' \subset I$ be an open subinterval containing t_0 on which two C^1 solutions y and z of* (14.1) *are given. Then $y = z$ on I'.*

Proof. **0.** We give a proof that also works under the weaker hypothesis of Theorem 14.5 below, that F satisfies boundedness and Lipschitz estimates on sets $I \times K$ where $K \subset \Omega$ is any compact set.

1. Let $L = \{t \in I' : y(t) = z(t)\}$. Then $t_0 \in L$. It suffices to show that L is both open and closed in I' (why?). Clearly, L is closed, since y and z are continuous.

[118] By the way, observe that X_R is *not* a normed vector space (why?). We need the extra generality of metric spaces over normed vector spaces here.

[119] If T is too large, not even the first condition will be satisfied.(why not?)

2. Suppose $t_1 \in L$. Let $y_1 := y(t_1) = z(t_1) \in \Omega$. Choose $R > 0$ such that $\overline{B(y_1, R)} \subset \Omega$. Since y and z are C^1, there is a $\delta > 0$ such that both y and z lie in

(14.14)
$$X(R,\delta) := \{u \in C(J_\delta, \mathbb{R}^n) : u(t_1) = y_1, \ \sup_{t \in J_\delta} |u(t) - y_1| \le R\}, \text{ where } J_\delta = [t_1 - \delta, t_1 + \delta] \subset I'.$$

Since y and z are solutions of
$$\frac{dy}{dt} = F(t, y(t)), \ \ y(t_1) = y_1,$$
they are both fixed points of
$$(\Phi y)(t) = \int_{t_1}^{t} F(s, y(s))ds + y_1.$$
The proof of Theorem 14.1 shows that if we take $0 < \delta' \le \delta$ small enough, then $\Phi : X(R, \delta') \to X(R, \delta')$ and is a contraction, and thus has a unique fixed point. Thus, $y = z$ on $[t_1 - \delta', t_1 + \delta']$.

□

Our next goal is to understand how solutions can fail to exist globally in time, and to formulate verifiable conditions under which solutions exist globally in time. Some clues are provided by the following counterexample.

Example 14.4. Consider the IVP

$$\frac{dy}{dt} = y^2, y(0) = 1. \tag{14.15}$$

A solution valid on $J = (-\infty, 1)$ is found by separation of variables to be $y(t) = \frac{1}{1-t}$. We say that this solution "blows up in finite time" because its value starts at 1 and becomes arbitrarily large on the interval $[0, 1)$. The solution with initial value $y(0) = 1$ does not exist as a C^1 solution beyond $t = 1$. The function $F(t, y) = y^2$ in this case is defined on $I \times \Omega = \mathbb{R} \times \mathbb{R}$, so global existence fails. Actually, global existence fails if I is taken to be any open interval containing 0 and 1.

The word "blow-up" is used to describe what happens to the solution of (14.4) because the size of $|y(t)|$ grows without bound on a finite time interval. There are other ways that a solution might conceivably break down, resulting in a failure of global existence. For example, a C^1 solution might after a time suddenly develop a jump discontinuity, even though it remains bounded; this is called *shock formation*. Or a solution might start out C^1 and then develop a discontinuity in its first derivative, even though it remains continuous and bounded; this could be called *cusp formation*. Another possibility is that a solution might oscillate rapidly as $t \to t_0$, while remaining C^1 and bounded away from t_0; think of $\sin \frac{1}{t}$ on $(0, 1)$. Proposition 14.7 will show that these latter forms of breakdown *never occur* in solutions to ODEs.[120]

Observe that in the proof of Theorem 14.1, the bound M (14.9) and the Lipschitz hypothesis on F were needed *only* on $\overline{B(y_0, R)}$. Thus, one should expect to be able to extend Theorem 14.1 to a result where such estimates are assumed to hold just on compact subsets of Ω. One might also expect (or at least guess) that if one considers a family of initial value problems of the form (14.1), where F is fixed but y_0 varies in a *compact* subset K of Ω, then there should be a *uniform* time of existence, independent of $y_0 \in K$, for all those problems.

[120]They do occur, however, in solutions to nonlinear PDEs. The study of such breakdown is an active area of current research in PDE.

Theorem 14.5 (Uniform local existence)**.** *I. Consider again the IVP* (14.1) *where we assume* $F \in C(I \times \Omega; \mathbb{R}^n)$ *and:*

a) for each compact $K \subset \Omega$*, there exists* $M_K < \infty$ *such that* $|F(t,x)| \leq M_K$ *for all* $x \in K$*,* $t \in I$*;*

b) for each K *as above, there exists* $L_K < \infty$ *such that* $|F(t,x) - F(t,y)| \leq L_K|x-y|$ *for all* $x, y \in K$*,* $t \in I$*.*

Then the IVP (14.1) *has a unique* C^1 *solution on some open* $t-$*interval containing* t_0*.*

II. Let $K \subset \Omega$ *be compact. Then then there is an open* $t-$*interval* J_K *containing* t_0 *such that for each* $y_0 \in K$*, a unique solution to* (14.1) *exists on* $J_K \cap I$*. In other words there is a* uniform *time of existence for all initial values* $y_0 \in K$*.*

Moreover, the same J_K *works for all* $t_0 \in I$*,* $y_0 \in K$*.*

Proof. We prove part II, which implies part I.

Let $K \subset \Omega$ be any fixed compact set. Then there exists $R_K > 0$ such that

$$\tilde{K} := \{x \in \mathbb{R}^n : \operatorname{dist}(x, K) \leq R_K\} \subset \Omega, \tag{14.16}$$

and $\tilde{K}$ is compact.(why?) So for each $y_0 \in K$ we have $\overline{B(y_0, R_K)} \subset \tilde{K}$. Choose $M_{\tilde{K}}$ and $L_{\tilde{K}}$ as in (a), (b) of Theorem 14.5. For each $y_0 \in K$ define Φ as before and let $X = X_{R_K}$, which is defined as in (14.7) where now J is replaced by[121]

$$J_K = \left\{ t \in I : |t - t_0| \leq \min\left\{ \frac{R_K}{M_{\tilde{K}}}, \frac{1}{2L_{\tilde{K}}} \right\} \right\}. \tag{14.17}$$

The proof of Theorem 14.1 shows that for all $y_0 \in K$ a solution to (14.1) exists on J_K. Uniqueness follows from Proposition 14.3.[122]

□

Remark 14.6. *If* $F \in C^1(\mathbb{R}; \mathbb{R}^n)$*, then* F *satisfies the hypotheses of Theorem 14.5 when* I *is any bounded open interval and* $\Omega = \mathbb{R}^n$*.*

The next result, a key step for understanding global existence, tells us that for F as in Theorem 14.5, if (a,b) is a bounded *open* interval on which the solution $y(t)$ to (14.1) remains in some compact set $K \subset \Omega$, then the solution extends to a strictly larger interval (a_1, b_1), where $a_1 < a$ and $b_1 > b$. Thus, only blow-up on (a,b) can prevent a solution from being extendable to a larger interval (a_1, b_1).

Observe that for the solution in Example 14.4, every interval $(a,b) = (a,1)$ with $a < 1$ *fails* to satisfy the above condition.

Proposition 14.7 (Extendability condition)**.** *Let* F *be as in Theorem 14.5. Assume* $[a,b]$ *is contained in the open interval* I *and assume* y *is a* C^1 *solution of the IVP* (14.1) *on* (a,b)*. Assume there exists a compact set* $K \subset \Omega$ *such that* $y(t) \in K$ *for all* $t \in (a,b)$*. Then there exist* $a_1 < a$ *and* $b_1 > b$ *such that* y *is a* C^1 *solution of* (14.1) *on* (a_1, b_1)*. Moreover, the extension of* y *to* (a_1, b_1) *is unique.*

Proof. Theorem 14.5 implies there exists $\delta > 0$ such that for each $t_1 \in (a,b)$ and $y_1 \in K$ the solution to

$$\frac{dz}{dt} = F(t, z), \quad z(t_1) = y_1 \tag{14.18}$$

[121] Compare with (14.12).

[122] Recall step **0** of the proof of Proposition 14.3.

exists on $(t_1-\delta, t_1+\delta)$. Take $t_1 \in (b-\frac{\delta}{2}, b)$ and $y_1 = y(t_1)$, where y solves (14.1) on (a,b). We claim that solving (14.18) continues y past $t=b$ as a C^1 solution. Similarly, one can continue y to the left past $t=a$.

To establish the claim observe that the C^1 solutions y and z are both defined on $(t_1-\delta, b)$ and they agree at t_1. By Proposition 14.3 they must agree on $(t_1-\delta, b)$. Uniqueness of the extension also follows from Proposition 14.3. □

We can now show that under certain conditions, global existence holds.

Proposition 14.8 (Conditions implying global existence)**.** *Consider the IVP* (14.1) *where $t_0 \in I$, $y_0 \in \Omega$, and $F : I \times \Omega$ is as in Theorem 14.5. Suppose that if $J \subset I$ is* any *bounded open subinterval containing t_0 on which a C^1 solution y exists, there exists a compact set $K \subset \Omega$ such that $y(t) \in K$ for all $t \in J$. Then y extends uniquely to a solution on I.*

Proof. If $J = I$ there is nothing to prove, so suppose $J \neq I$ and y is a solution of (14.1) on J. Letting $J = (a,b)$ and $I = (c,d)$, we consider the case $c \leq a < b < d$, where d may be $+\infty$.[123] Let

$$S = \{b' \in \mathbb{R} : b < b' \leq d \text{ and } y \text{ extends to be a } C^1 \text{ solution on } (a,b')\}. \tag{14.19}$$

By Proposition 14.7 we know $S \neq \emptyset$. We wish to show $\sup S = d$. If $\sup S = d' < d$, then y extends to be a C^1 solution on (a,d'). By Proposition 14.7 we can extend y past d', a contradiction. If $c < a$ we may similarly extend y to the left. Uniqueness follows from Proposition 14.3.

□

This proposition might strike you as rather useless, because you might think that the main hypothesis is too hard to verify. As the next example shows, one can sometimes verify the main hypothesis rather easily. It remains true though, that global existence is the exception rather than the rule for nonlinear ODEs.[124]

Example 14.9. *Consider the 2×2 nonlinear system for $y = (x,v)$:*

$$\frac{dx}{dt} = v, \quad \frac{dv}{dt} = -x^3. \tag{14.20}$$

We take $I = \mathbb{R}$, $\Omega = \mathbb{R}^2$ and $F(t,y) = F(t,x,v) = (v, -x^3)$. Suppose (x,v) is a C^1 solution of (14.20) *for t in a subinterval $J \subset I$. Then*

$$\frac{d}{dt}\left(\frac{v^2}{2} + \frac{x^4}{4}\right) = v\frac{dv}{dt} + x^3\frac{dx}{dt} = 0 \text{ on } J. \tag{14.21}$$

So each solution $y(t) = (x(t), v(t))$ to (14.20) *lies on a level curve $\frac{v^2}{2} + \frac{x^4}{4} = C$, and is therefore confined to a compact subset of $\mathbb{R}^2$. From Proposition 14.8 we conclude that for any initial condition, unique solutions to* (14.20) *exist for all $t \in \mathbb{R}$.*

14.2 Linear systems of ODEs.

In this section we consider the important case of *linear* first-order systems:

$$\frac{dy}{dt} = A(t)y, \quad y(t_0) = y_0, \tag{14.22}$$

[123] The remaining case is treated the same way.

[124] Keep in mind Example 14.4.

where A is a continuous function on an open interval $I \ni t_0$ (which could be $\mathbb{R}$) taking values in the space of real $n \times n$ matrices, $M(n, \mathbb{R})$. This is a system of the form (14.1) where $F(t, y) = A(t)y$, so everything we have proved about (14.1) applies here. But things become much simpler in the linear case; in particular, global existence *always* holds. Let us see why by briefly revisiting the proofs of Theorems 14.1 and 14.5.

In the notation of these theorems we now have $\Omega = \mathbb{R}^n$ and

$$(\Phi y)(t) = \int_{t_0}^{t} A(s)y(s)ds + y_0. \tag{14.23}$$

First let $J \subset I$ be any bounded closed interval containing t_0. If we take $X_T = C(J_T, \mathbb{R}^n)$ where $T > 0$ and $J_T = [t_0 - T, t_0 + T] \subset J$, it is clear that $\Phi : X_T \to X_T$ (why?). Now we don't have to worry about functions y taking values outside Ω because $\Omega = \mathbb{R}^n$.

Clearly, there exists $L > 0$ such that $|A(t)| \leq L$ on J, so we have the Lipschitz estimate[125]

$$|A(t)(y_1 - y_2)| \leq L|y_1 - y_2| \text{ for } t \in J, y_i \in \mathbb{R}^n. \tag{14.24}$$

Unlike the general case, where the choice of Lipschitz constant L depended on a particular compact set $\tilde{K} \subset \Omega$ containing y_0, here the choice of L depends just on $A(t)$ and J, and not at all on y_0 or a choice of $\tilde{K}$. The proof of Theorem 14.1 shows that $\Phi : X_T \to X_T$ is a contraction if $T < \frac{1}{L}$, so we obtain a C^1 solution y on J_T. It is unique by Proposition 14.3.

To see how to extend this solution to all of J, we start by extending it to $[t_0 + T, t_0 + 2T] \subset J$.[126] First solve in the same way:

$$\frac{d\tilde{y}}{dt} = A(t)\tilde{y} \text{ on } [t_0 + T, t_0 + 2T], \;\; \tilde{y}(t_0 + T) = y(t_0 + T), \tag{14.25}$$

noting that we can use the *same* L for this problem (what is the new Φ?). Concatenating y and $\tilde{y}$ gives a C^1 solution on $[t_0 - T, t_0 + 2T]$. Repeating and similarly extending to the left we obtain a unique C^1 solution on J.[127]

Exercise: (a) Show that this solution can be extended to a unique C^1 solution of (14.22) on the original open interval I, which may be unbounded. (Hint: Connectedness!)

(b) Comment briefly on *how* the argument on this page fails if applied to prove global existence for Example 14.4.

The next proposition gives an interesting bound on the growth rate of solutions to (14.22) which implies uniqueness, and with Proposition 14.8, implies global existence. The technique of proof, referred to as an "energy estimate", is classical in both ODE and PDE theory.

Proposition 14.10. *Consider a C^1 solution to the IVP*

$$\frac{dy}{dt} = A(t)y, \;\; y(0) = y_0$$

on an interval $I \ni 0$; here $A \in C(I; M(n, \mathbb{R}))$. If $|A(t)| \leq K$ for $t \in I$, then $y(t)$ satisfies

$$|y(t)| \leq e^{K|t|}|y_0| \text{ on } I. \tag{14.26}$$

[125] As usual, if $B \in M(n, \mathbb{R})$, then $|B| := \sup_{|x| \leq 1} |Bx|$.

[126] Assume here that $t_0 + 2T \in J$. Otherwise extend to the right endpoint of J.

[127] Check that the solutions y and $\tilde{y}$ patch together to give a C^1 solution near $t_0 + T$. Continuity is clear.

Proof. It suffices to prove (14.26) for $t \geq 0$.[128] Then $z(t) := e^{-Kt}y(t)$ satisfies

$$\frac{dz}{dt} = C(t)z, \;\; z(0) = y_0, \;\text{where } C(t) = A(t) - KI. \tag{14.27}$$

We'll show

$$|z(t)| \leq |z(0)| \text{ for } t \geq 0, \tag{14.28}$$

which implies (14.26). We have

$$\frac{d}{dt}|z(t)|^2 = z'(t) \cdot z(t) + z(t) \cdot z'(t) = 2z(t) \cdot (A(t) - K)z(t). \tag{14.29}$$

Since

$$z(t) \cdot A(t)z(t) \leq |z(t)||A(t)z(t)| \leq |A(t)||z(t)|^2, \tag{14.30}$$

the hypothesis $|A(t)| \leq K$ implies $\frac{d}{dt}|z(t)|^2 \leq 0$, which gives (14.28). □

Exercise: (a). Use (14.26) to derive a similar estimate for the IVP (14.22).
(b) Explain briefly why this estimate implies uniqueness for (14.22).
(c) Explain briefly why this estimate implies global existence for (14.22).

15 Compactness in function spaces: the Arzela-Ascoli theorem

We have seen many times the important role that compact sets play in analysis. Most of our experience so far has been with compact subsets of finite-dimensional normed vector spaces like $\mathbb{R}^n$. In $\mathbb{R}^n$ there is a simple criterion for compactness given by the Heine-Borel theorem: a set is compact if and only if it is closed and bounded. You showed in a homework problem that any finite-dimensional normed vector space is homeomorphic to $\mathbb{R}^n$, so the same criterion for compactness works in any finite-dimensional normed vector space.

Most of the normed function spaces used in analysis are infinite-dimensional, like $C^k([a,b], \mathbb{R}^n)$ for example. In such spaces, as in any metric space, compact sets are closed and bounded, but the converse does not hold. In fact, the following result is true: the closed unit ball in *any* infinite-dimensional normed vector space is *not* compact. We have already seen at least one example of this. The sequence of functions $(x^n)_{n=1}^{\infty}$ illustrates that the closed unit ball of $C([0,1]; \mathbb{R})$ is not compact (why?). Here is an example where compactness fails in a different way.

Example 15.1. *Let $V = C_b(\mathbb{R}; \mathbb{R})$ be the set of continuous* bounded *functions on $\mathbb{R}$ with the sup norm. Let $f \in V$ be any element with compact support contained in $(0,1)$ and such that $|f| = 1$. Then the sequence $(f_n)_{n=1}^{\infty}$ defined by $f_n(x) = f(x-n)$ is contained in the closed unit ball of V and illustrates that the closed unit ball is not compact. (why?)*

Next we define a property of sets of functions that will be a key hypothesis in the Arzela-Ascoli theorem.

Definition 15.2 (Equicontinuity)**.** *Let (X,d) be a metric space and $\mathcal{F}$ a family of functions $f : X \to \mathbb{R}$. We say that $\mathcal{F}$ is* equicontinuous *on X if given any $\epsilon > 0$ there exists a $\delta = \delta(\epsilon) > 0$ such that if $d(p,q) < \delta$ then $|f(p) - f(q)| < \epsilon$ for all $f \in \mathcal{F}$.*

[128] To prove (14.26) for $t \leq 0$, consider the problem satisfied by $w(t) = y(-t)$.

We will also need:

Proposition 15.3. *If (X,d) is compact metric space, it has a countable dense subset.*

Proof. For $n \in \mathbb{N}$ we may use the compactness of X to cover X with a collection of finitely many open balls of radius 2^{-n}. Let A_n be the set of centers of these balls. Then $\cup_{n=1}^{\infty} A_n$ is a countable dense subset of X.

□

Recall that if (X,d) is a compact metric space, then $C(X;\mathbb{R})$ is a complete metric space with the sup norm; see Example 8.19(3) and Proposition 9.12. If $K \subset C(X;\mathbb{R})$ we know that the conditions of closedness and boundedness are necessary but not sufficient for K to be compact. The question then arises: what additional conditions on K would guarantee compactness? It is enough just to add equicontinuity.

Theorem 15.4 (Arzela-Ascoli theorem)**.** *Let (X,d) be a compact metric space, and let $K \subset C(X,\mathbb{R})$ be closed, bounded, and equicontinuous. Then K is compact.*

Proof. **1.** Let (f_n) be a sequence of functions in K. It suffices to find a subsequence that is uniformly Cauchy on X (why?). Fix a countable dense subset $A = \{p_1, p_2, p_3, \dots\} \subset X$. The bounded sequence of real numbers $(f_n(p_1))$ has a convergent subsequence $(f_{1,k}(p_1))$, where (f_{1k}) is a subsequence of (f_n). Next consider the sequence $(f_{1,k}(p_2))$. This has a convergent subsequence $(f_{2,k}(p_2))$. Moreover, the subsequence $(f_{2,k})$ of (f_n) converges at both p_1 and p_2. Continue inductively in this way to produce a countable family of subsequences $(f_{m,k})$, $m = 1, \dots, \infty$. Then the *diagonal sequence* $(f_{m,m}) := (g_m)$ converges at all points of A.

2. We claim that (g_m) is uniformly Cauchy on X. Fix $\epsilon > 0$. Using equicontinuity, we choose $\delta > 0$ such that $d(p,q) < \delta$ implies $|g_m(p) - g_m(q)| < \frac{\epsilon}{3}$ for all m. By compactness of X and density of A there exists a finite number of $\delta-$balls covering X centered at points of $S := \{p_1, \dots, p_M\} \subset A$. Let $p \in X$ and choose $p_i \in S$ such that $d(p,p_i) < \delta$. For any k, l we have

$$|g_k(p) - g_l(p)| \le |g_k(p) - g_k(p_i)| + |g_k(p_i) - g_l(p_i)| + |g_l(p_i) - g_l(p)| := I + II + III. \tag{15.1}$$

The terms I and III are $< \frac{\epsilon}{3}$ by the choice of δ. Since $S \subset A$ and is finite, we can choose N large enough and independent of $p_i \in S$ so that $k, l \ge N$ implies $II < \frac{\epsilon}{3}$.

□

Remark 15.5. *1) Theorem 15.4 remains true with the same proof for $C(X;\mathbb{R}^k)$ (with sup norm).*

2). The proof of Theorem 15.4 shows that if $\mathcal{F} \subset C(X;\mathbb{R})$ is bounded and equicontinuous, then it has compact closure in $C(X;\mathbb{R})$.

3). Consider $C(X;Y)$ with the usual metric $d(f,g) = \sup_{x\in X} d_Y(f(x), g(x))$, where X and Y are compact metric spaces.[129] *If $K \subset C(X;Y)$ is closed and equicontinuous, then K is compact. The proof is just like that of Theorem 15.4. Observe that this result implies Theorem 15.4, even though $\mathbb{R}$ is not compact. (why?)*

4) Example 15.1 shows the need for the compactness of X in these results. The sequence (f_n) in that example is closed, bounded, and *equicontinuous, but not relatively compact. How does equicontinuity fail in the example of (x^n) in $C([0,1];\mathbb{R})$?*

5) The "diagonal argument" used in the proof of Theorem 15.4 is a classical one in mathematics. Recall how Cantor used a diagonal argument to prove the reals are uncountable. It's another good argument to keep in your toolbox.

[129]This implies Y is complete.

16 Density and approximation in function spaces

In this section we present results that allow us to approximate a given function in a function space arbitrarily closely using functions in a particular subspace. The first two subsections are preparatory.

16.1 Differentiation under the integral sign

You may have already seen this result in Math 521. It is an example of interchanging two limit processes.

Proposition 16.1. *Let $\Omega \subset \mathbb{R}^2$ be open and suppose $R = \{a \le x \le b,\ c \le t \le d\} \subset \Omega$. Assume $f \in C^1(\Omega, \mathbb{R})$. For $x \in (a,b)$ define $\phi(x) = \int_c^d f(x,t)dt$. Then*[130].

$$\phi'(x) = \int_c^d f_x(x,t)dt. \tag{16.1}$$

and ϕ is C^1 on (a,b).

Proof. Set $g(x) := \int_c^d f_x(x,t)dt$. We have $\frac{\phi(x+h)-\phi(x)}{h} - g(x) =$

$$\int_c^d \frac{f(x+h,t)-f(x,t)}{h}dt - g(x) = \int_c^d [f_x(x+c(h),t) - f_x(x,t)]dt, \tag{16.2}$$

where we used the mean value theorem to get the second equality.

Fix $\epsilon > 0$. Since f_x is uniformly continuous on R, there exists a $\delta > 0$ such that if $|(x_1,t_1) - (x_2,t_2)| < \delta$ and $(x_i,t_i) \in R$, then $|f_x(x_1,t_1) - f_x(x_2,t_2)| < \epsilon$. Thus, if $|h| < \delta$ in (16.2) we obtain

$$\left| \frac{\phi(x+h)-\phi(x)}{h} - g(x) \right| < \epsilon(d-c). \tag{16.3}$$

To obtain the continuity of ϕ' on (a,b), write $\phi'(x_1) - \phi'(x_2)$ as an integral, and again use the uniform continuity of f_x on R.

□

This proposition has obvious generalizations to the case where $(x,t) \in \mathbb{R}^{n+1}$ and t lies in a closed rectangle of $\mathbb{R}^n$. Moreover, if $x \in \mathbb{R}^k$, $t \in \mathbb{R}^l$, one has a similar result for $\nabla_x \phi$ when

$$\phi(x) := \int_R f(x,t)dt,$$

where R is an appropriate closed rectangle. We'll use these generalizations freely.

16.2 Convolution

Suppose we have two functions f and g, where g is smoother than f. Then (under appropriate hypotheses) the convolution of f and g, denoted $f * g$, is a function which has the smoothness of the smoother function g. Moreover, if g is chosen correctly, the function $f * g$ will be close to f in many different norms and metrics.

[130] The proof shows that it is enough to suppose f and f_x are continuous on R.

Definition 16.2 (Convolution)**.** *Let $f \in C(\mathbb{R}^n, \mathbb{R})$ and $g \in C_c(\mathbb{R}^n, \mathbb{R}$.*[131] *Then the convolution of f and g, $f * g$ is given by*

$$(f * g)(x) := \int_{\mathbb{R}^n} f(x-y)g(y)dy = \int_{\mathbb{R}^n} f(y)g(x-y)dy. \tag{16.4}$$

The integrals here are finite because for any given x a continuous function of y is integrated on a compact set. The second equality follows by the change of variable formula.

Proposition 16.3. *Let $f \in C(\mathbb{R}^n, \mathbb{R})$ and $g \in C_c^k(\mathbb{R}^n, \mathbb{R})$, where $k \geq 0$. Then*

a) $\operatorname{supp} f * g \subset \overline{\operatorname{supp} f + \operatorname{supp} g}$;
b) $f * g \in C^k(\mathbb{R}^n, \mathbb{R})$ *and for* $|\alpha| \leq k$ *we have* $\partial^\alpha (f * g) = f * (\partial^\alpha g)$;
c) if $f \in C^k(\mathbb{R}^n, \mathbb{R})$ *then* $\partial^\alpha (f * g) = (\partial^\alpha f) * g = f * (\partial^\alpha g)$

Proof. **(a)** If $x \notin \overline{\operatorname{supp} f + \operatorname{supp} g}$, then $x - y \notin \operatorname{supp} g$ when $y \in \operatorname{supp} f$, so the integrand in the second integral of (16.4) is zero for all y.

(b) and (c) To prove (b) use the second integral in (16.4) and differentiate under the integral sign using Proposition 16.1. The proof of the first equality in (c) is similar. □

16.3 Approximate identities

In this section we discuss a systematic method for approximating C^k functions for any $k \geq 0$, by C^∞ functions. The main tool is an "approximate identity" constructed as follows. Let $g \in C_c^\infty(\mathbb{R}^n, \mathbb{R})$ satisfy

$$g \geq 0, \ \operatorname{supp} g \subset \overline{B(0,1)} = \{x \in \mathbb{R}^n : |x| \leq 1\}, \ \int_{\mathbb{R}^n} g(x)dx = 1. \tag{16.5}$$

Rescale g by defining

$$g_k(x) = k^n g(kx) \text{ for } k = 0, 1, 2, \ldots. \tag{16.6}$$

Observe that (check)

$$g_k \geq 0, \ \operatorname{supp} g_k \subset \overline{B(0, 1/k)}, \text{ and } \int_{\mathbb{R}^n} g_k(x)dx = 1 \text{ for all } k. \tag{16.7}$$

Definition 16.4 (Approximate identity)**.** *The sequence (g_k) given by* (16.6) *is called an* approximate identity*.*

The next proposition justifies this terminology (more or less).

Proposition 16.5. *For $m \geq 0$ let $f \in C^m(\mathbb{R}^n, \mathbb{R})$ and define $f_k \in C^\infty(\mathbb{R}^n, \mathbb{R})$ by*

$$f_k(x) = (f * g_k)(x). \tag{16.8}$$

Then for any compact set $K \subset \mathbb{R}^n$ and for any multi-index α with $|\alpha| \leq m$ we have

$$\partial^\alpha f_k \to \partial^\alpha f \text{ uniformly on } K \text{ as } k \to \infty. \tag{16.9}$$

[131]The subscript "c" indicates compact support.

Proof. **1.** First we do the case $m = 0$. Fix $K \subset \mathbb{R}^n$ compact and fix $\epsilon > 0$. Since f is uniformly continuous on the compact set $\tilde{K} := K + \overline{B(0,1)}$, there exists $\delta > 0$ such that $x, y \in \tilde{K}$ and $|x-y| < \delta$ implies $|f(x) - f(y)| < \epsilon$.

2. For $x \in K$ we have

$$\begin{aligned} |f_k(x) - f(x)| = \left|\int (f(x-y) - f(x))g_k(y)dy\right| = \left|\int (f(x-y) - f(x))k^n g(ky)dy\right| = \\ \left|\int \left(f(x - \frac{v}{k}) - f(x)\right) g(v)dv\right| := A. \end{aligned} \tag{16.10}$$

On supp g we have $|v| \leq 1$ so if we choose k such that $\frac{1}{k} < \delta$, then $A \leq \epsilon \int g(v)dv = \epsilon$.[132]

3. To treat general m observe that for $|\alpha| \leq m$ we have by Proposition 16.3

$$\partial^\alpha f_k = (\partial^\alpha f) * g_k, \tag{16.11}$$

and apply the case $m = 0$ with $\partial^\alpha f$ in the role of f.

□

Remark 16.6. *Proposition 16.5 remains true with essentially the same proof if $\mathbb{R}^n$ is replaced by an open subset $\Omega \subset \mathbb{R}^n$. For this use the fact that if $x \in K \subset \Omega$, we can choose $k = k(K)$ large enough so that $x - y \in \Omega$ for $y \in \text{supp } g_k$.*

16.4 Weierstrass approximation theorem

This theorem allows us to approximate continuous functions on compact intervals by polynomials. Notice how convolution again plays a role in the proof.[133]

Theorem 16.7. *Let $f \in C([a,b], \mathbb{R})$. Then there exists a sequence of polynomials p_n such that $p_n \to f$ uniformly on $[a,b]$.*

Proof. **1.** We can assume $[a,b] = [0,1]$ (why?) and also reduce to the case where $f(0) = f(1) = 0$.[134] Define f to be zero outside $[0,1]$ so that it is uniformly continuous on $\mathbb{R}$.

2. Set $Q_n(x) = c_n(1-x^2)^n$, $n = 1, 2, 3, \ldots$, where c_n is chosen so

$$\int_{-1}^{1} Q_n(x)dx = 1. \tag{16.12}$$

Using $(1-x^2)^n \geq 1 - nx^2$ on $[0,1]$, we can estimate c_n as follows:[135]

$$\int_{-1}^{1} (1-x^2)^n dx = 2\int_0^1 (1-x^2)^n dx \geq 2\int_0^{\frac{1}{\sqrt{n}}} (1-nx^2)dx = \frac{4}{3\sqrt{n}} > \frac{1}{\sqrt{n}}, \tag{16.13}$$

so $c_n < \sqrt{n}$. For any $\delta > 0$ we have $Q_n(x) \leq \sqrt{n}(1-\delta^2)^n$ on $0 < \delta \leq |x| \leq 1$, so

$$Q_n \to 0 \text{ uniformly on } 0 < \delta \leq |x| \leq 1 \text{ as } n \to \infty. \tag{16.14}$$

---

[132] For the first equality of (16.10) we used $\int g_k(x)dx = 1$. To get the third equality we made the change of variables $v = ky$. In the final estimate of A we used $x - \frac{v}{k} \in \tilde{K}$ when $v \in \text{supp } g$.

[133] Here we follow the proof of [WR].

[134] One can subtract a linear polynomial from a given f to make the latter condition hold.

[135] To see that $(1-x^2)^n \geq 1 - nx^2$ consider $h(x) = (1-x^2)^n - 1 + nx^2$. This satisfies $h(0) = 0$, $h'(x) > 0$ on $(0,1)$.

We now redefine Q_n to be 0 outside $[-1, 1]$. Then $Q_n \geq 0$ is continuous on $\mathbb{R}$, a polynomial on $[-1, 1]$, and $Q_n \to 0$ uniformly on $|x| \geq \delta$.

3. For $x \in [0, 1]$ define

$$P_n(x) = (f * Q_n)(x) = \int_{\mathbb{R}} f(x-t)Q_n(t)dt = \int_{-1}^{1} f(x-t)Q_n(t)dt = \int_{0}^{1} f(t)Q_n(x-t)dt, \tag{16.15}$$

and note from the last integral that $P_n(x)$ is a polynomial on $[0, 1]$ (why?).

4. Fix $\epsilon > 0$ and choose $0 < \delta < 1$ so that $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$. With $M = \sup |f(x)|$, using steps **2** and **3** we estimate for $x \in [0, 1]$:

$$\begin{gathered} |P_n(x) - f(x)| = \left| \int_{-1}^{1} [f(x-t) - f(x)]Q_n(t)dt \right| \leq \\ 2M \int_{-1}^{-\delta} Q_n(t)dt + \epsilon \int_{-\delta}^{\delta} Q_n(t)dt + 2M \int_{\delta}^{1} Q_n(t)dt \leq 4M\sqrt{n}(1-\delta^2)^n + \epsilon < 2\epsilon \end{gathered} \tag{16.16}$$

for n large enough.

□

Remark 16.8. *Observe that the sequence of functions Q_n has much in common with the sequence g_k in* (16.7)*. The supports of the Q_n don't shrink with increasing n (they're polynomials!), but we have* (16.14) *as a useful substitute for such shrinkage.*

16.5 Stone-Weierstrass theorem

In this section we state a far-reaching generalization of Theorem 16.7 due to M. Stone. A proof is given Proposition A.5.2 of our text and also in [WR], Chapter 7.

Theorem 16.9 (Stone-Weierstrass)**.** *Let X be a compact metric space and let $\mathcal{A}$ be a subalgebra of $C(X, \mathbb{R})$, the algebra of real-valued continuous functions on X.*[136] *Suppose $1 \in \mathcal{A}$ and that $\mathcal{A}$* separates points *of X, i.e., for distinct $p, q \in X$, there exists $h_{pq} \in \mathcal{A}$ with $h_{pq}(p) \neq h_{pq}(q)$. Then the closure of $\mathcal{A}$ is equal to $C(X, \mathbb{R})$.*

The proof of Theorem 16.9 uses Theorem 16.7. Also, observe that Theorem 16.7 is an immediate corollary of Theorem 16.9.(why?)

Here is the version for algebras of complex-valued functions, which follows readily from Theorem 16.9; see [WR]. We say that an algebra of complex-valued functions is *self-adjoint* if it is closed under complex conjugation.

Theorem 16.10. *Let X be a compact metric space and let $\mathcal{A}$ be a* self-adjoint *subalgebra of $C(X, \mathbb{C})$, the algebra of complex-valued continuous functions on X. Suppose $1 \in \mathcal{A}$ and that $\mathcal{A}$* separates points *of X. Then the closure of $\mathcal{A}$ is equal to $C(X, \mathbb{C})$.*

Corollary 16.11. *Let X be a compact subset of $\mathbb{R}^n$. Every $f \in C(X, \mathbb{C})$ is the uniform limit of a sequence of polynomials on $\mathbb{R}^n$.*

[136] A family of real-valued functions on X is an *algebra* if it is closed under addition, multiplication, and multiplication by real scalars. The definition for a family of complex-valued functions is the same, but the family must be closed under multiplication by $c \in \mathbb{C}$.

Exercise. Let $C_p([0,2\pi],\mathbb{C}) = \{f \in C([0,2\pi],\mathbb{C}) : f(0) = f(2\pi)\}$.[137] Show that the set of all trigonometric polynomials

$$TP := \left\{ \sum_{|k|\le N} a_k e^{ik\theta},\ N = 0,1,2,\dots,\ \text{where } a_k \in \mathbb{C} \right\},$$

is dense in $C_p([0,2\pi],\mathbb{C})$.[138] (Caution: The set TP does not separate points of $[0,2\pi]$.)

17 Introduction to complex analysis

In this section we give another application of Stokes theorem. We use Green's theorem, a corollary of Stokes theorem, to prove a fundamental result about "analytic functions" $f : \mathbb{C} \to \mathbb{C}$, namely, Cauchy's theorem.[139] We also derive a number of remarkable corollaries of Cauchy's theorem. The whole theory of analytic functions is arguably a consequence of this theorem.

First we define the field $\mathbb{C}$ of complex numbers.

Definition 17.1. *The field of complex numbers $\mathbb{C}$ is the set of ordered pairs $(a,b) \in \mathbb{R}^2$ with operations of addition and multiplication given by*

$$(a,b)+(c,d) = (a+c, b+d), \quad (a,b)\cdot(c,d) = (ac-bd, ad+bc). \tag{17.1}$$

We define *$i = (0,1)$ and note that $i^2 = (-1,0)$. If we write an element of $\mathbb{C}$ of the form $(a,0)$ simply as a, then we have*

$$(a,b) = (a,0)+(b,0)(0,1) = a+bi. \tag{17.2}$$

In this way points of $\mathbb{R}^2$ are "identified" with complex numbers. Observe that the definition of multiplication given in (17.1) *can be rewritten as*

$$(a+bi)\cdot(c+di) = (ac-bd) + i(ad+bc).$$

We refer to a and b as the real and imaginary parts of $z = a+bi$, and write $a = \mathrm{Re}\, z$, $b = \mathrm{Im}\, z$. The complex conjugate *of $z = a+bi$ is defined to be $\overline{z} = a-bi$. We define $|z| = \sqrt{z\overline{z}} = \sqrt{a^2+b^2} = |(a,b)|$.*

17.1 Analytic functions and the Cauchy-Riemann equations

Consider now any function $f : \mathbb{C} \to \mathbb{C}$. If we set $z = x+iy \in \mathbb{C}$, we can write $f(z) = u(z)+iv(z)$, or with slight abuse of notation as

$$f(z) = u(x,y) + iv(x,y), \tag{17.3}$$

where the real-valued functions u and v are called the real and imaginary parts of f.[140] Observe that any complex-valued function on $\mathbb{R}^2$ can be viewed as a function $f : \mathbb{C} \to \mathbb{C}$.

[137] The subscript "p" indicates that these functions extend to $2\pi-$periodic functions on $\mathbb{R}$.
[138] This result is important in the theory of Fourier series.
[139] Analytic functions will be defined shortly.
[140] We could also write $f(z) = (u(x,y), v(x,y))$ but that is less often done.

Definition 17.2 (Analytic function)**.** *Let f be a complex-valued C^1 function on an open set $\Omega \subset \mathbb{R}^2$.*[141] *We say that f is* analytic *(or* holomorphic*) on Ω if it is "complex differentiable" in the sense that*

$$\lim_{h\to 0}\frac{f(z+h)-f(z)}{h} := f'(z) \text{ (or } df/dz) \tag{17.4}$$

exists for all $z \in \Omega$.[142] *In this case we write $f \in H(\Omega)$.*

Working directly with difference quotients, one can repeat Math 521 arguments to show:

Proposition 17.3. *(a) If $f, g \in H(\Omega)$ then $fg \in H(\Omega)$ and $(fg)'(z) = f'(z)g(z) + f(z)g'(z)$.*

(b) $\frac{d}{dz}z^k = kz^{k-1}$ for $k \in \mathbb{Z}$, where we restrict to $\mathbb{C} \setminus \{0\}$ for negative k.

(c) Suppose f and g are analytic on open sets G and Ω respectively and $f(G) \subset \Omega$. Then $g \circ f \in H(G)$ and the chain rule holds:

$$(g \circ f)'(z) = g'(f(z))f'(z).$$

Next we give a useful criterion for analyticity in terms of partial derivatives.

Proposition 17.4 (Cauchy-Riemann equations)**.** *Let $f : \Omega \to \mathbb{C}$ be C^1. Then $f = u+iv$ is analytic on Ω if and only if f satisfies the* Cauchy-Riemann *(C-R) equations on Ω:*

$$f_x = \frac{1}{i}f_y, \text{ or equivalently } u_x = v_y,\ u_y = -v_x. \tag{17.5}$$

Proof. **1.** To prove $\Rightarrow$, in (17.4) let $h \to 0$ first along the $x-$axis and then along the $y-$axis (check).

2. Suppose now that (17.5) holds in Ω. For $(x, y) \in \Omega$ and $h = h_1 + ih_2$ small we have

$$f(z+h) - f(z) = u(x+h_1, y+h_2) - u(x, y) + i[v(x+h_1, y+h_2) - v(x, y)]. \tag{17.6}$$

Using the mean value theorem we obtain $u(x + h_1, y + h_2) - u(x, y) =$

$$\begin{aligned}&[u(x+h_1, y+h_2) - u(x, y+h_2)] + [u(x, y+h_2) - u(x, y)] = \\ &\quad u_x(x+c_1, y+h_2)h_1 + u_y(x, y+c_2)h_2 = u_x(x,y)h_1 + u_y(x,y)h_2 + E_1 + E_2,\end{aligned} \tag{17.7}$$

where the E_i are the obvious error terms. Thus, using the C-R equations we find

$$\begin{aligned}&f(z+h) - f(z) = u_x(x,y)h_1 + u_y(x,y)h_2 + i[v_x(x,y)h_1 + v_y(x,y)h_2] + \sum_{j=1}^{4} E_j = \\ &\qquad [u_x(x,y) + iv_x(x,y)](h_1 + ih_2) + \sum_{j=1}^{4} E_j.\end{aligned} \tag{17.8}$$

Since f is C^1 we have $\lim_{h\to 0}\frac{E_j}{h} = 0$ (why?), so we can use (17.8) to take the limit in (17.4) to obtain $f'(z) = u_x(x, y) + iv_x(x, y)$. $\square$

[141] This means that $u = \mathrm{Re}f$ and $v = \mathrm{Im}f$ are C^1.

[142] Here $h \in \mathbb{C}$.

Remark 17.5. *1) Proposition 17.4 tells us that a function $f \in H(\Omega)$ is no more and no less than a map $(u,v) : \Omega \to \mathbb{R}^2$ which is C^1 and satisfies*

$$u_x = v_y,\ u_y = -v_x.$$

The explicit one-to-one correspondence is $f \to (\text{Re}f, \text{Im}f) := (u,v)$ with inverse $(u,v) \to u+iv := f$. For example, the function $f(z) = z^2$ is the function $(x,y) \to (x^2 - y^2, 2xy)$.[143] *Moreover, the proof of Proposition 17.4 shows that if $f \in H(\Omega)$, then*[144]

$$f'(z) = f_x(z) = u_x(x,y) + iv_x(x,y) \text{ and } f'(z) = \frac{1}{i} f_y(z) = \frac{1}{i}(u_y + iv_y) = v_y - iu_y. \tag{17.9}$$

2) The subspace $H(\Omega) \subset C^1(\Omega, \mathbb{C})$ is in some sense quite "small", because the conditions that define it, Definition 17.2 or (17.5) are quite strong. For example, $H(\Omega)$ is not dense in $C^1(\Omega, \mathbb{C})$, when the latter space carries its usual metric space topology.[145]

3) The function $f(z) = \overline{z}$ is not analytic, but it is C^∞ on $\mathbb{C}$. It is the function $(x,y) \to (x, -y)$.

Examples 17.6. *1. The exponential function e^z is defined by*

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}. \tag{17.10}$$

The power series, along with each of its derivatives, converges uniformly on compact sets $K \subset \mathbb{C}$.[146] *Math 521 results together with the C-R equations then imply that $e^z \in H(\mathbb{C})$. Clearly, $\frac{d}{dz}e^z = e^z$ and e^z agrees with e^x for $x \in \mathbb{R}$.*

2. For $z \in \mathbb{C}$ set

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i}, \quad \cos z = \frac{e^{iz} + e^{-iz}}{2}.$$

These functions lie in $H(\mathbb{C})$, agree with $\sin x$ and $\cos x$ for $x \in \mathbb{R}$, and satisfy for all z

$$\frac{d}{dz}\sin z = \cos z, \quad \frac{d}{dz}\cos z = -\sin z \quad \sin^2 z + \cos^2 z = 1.$$

3. Polynomials $p(z)$ lie in $H(\mathbb{C})$ and rational functions, $\frac{p(z)}{q(z)}$ where p, q are polynomials, are analytic away from the zeros of q.

17.2 Cauchy's theorem, the Cauchy integral formula, and some consequences

In preparation for the proof of Cauchy's theorem, we recall Green's theorem and state two slight variants:

Theorem 17.7 (Green's theorem)**.** *(a) Let Ω be a bounded, connected open subset of $\mathbb{R}^2$ with a smooth boundary $\partial\Omega$ oriented "positively".*[147] *Let $f, g \in C^\infty(\overline{\Omega}, \mathbb{R})$. Then*

$$\int_\Omega (g_x - f_y)\,dxdy = \int_{\partial\Omega} f dx + g dy. \tag{17.11}$$

(b) The theorem holds if f and g are complex-valued C^∞ functions on $\overline{\Omega}$ (that is, $f, g \in C^\infty(\overline{\Omega}, \mathbb{C})$).

(c) The theorem holds if f and g are complex-valued C^1 functions on $\overline{\Omega}$ (that is, $f, g \in C^1(\overline{\Omega}, \mathbb{C})$).

---

[143] Clearly the "z" notation has great advantages; imagine z^{1000} in x, y notation.

[144] This remark is helpful, for example, for the exercise that follows Remark 17.17.

[145] In this metric topology, defined in problem 7 of HW 13, a sequence $f_n \to f$ if and only if for $|\alpha| \leq 1$, the sequence $\partial^\alpha f_n \to \partial^\alpha f$ uniformly on compact subsets of Ω.

[146] This is proved just like the analogous statement for the real power series that defines e^x.

[147] This means that the boundary orientation is induced from the standard orientation on $\mathbb{R}^2$. Although Ω is connected, $\partial\Omega$ is not necessarily connected. Think of an annulus. Also, the connectedness assumption is not essential.

The assertion in (b) is proved by applying part (a) to the pairs Ref, Reg and Imf, Img. The assertion in (c) can be checked by observing that the proof of part (a) continues to work under the C^1 assumption; see the statement of Proposition 4.3.1 of our text.[148]

Theorem 17.8 (Cauchy's theorem)**.** *Let Ω be a bounded, connected open subset of $\mathbb{R}^2$ with a smooth boundary $\partial\Omega$ oriented "positively". If $f \in C^1(\overline{\Omega}, \mathbb{C})$ and is analytic in Ω, then*[149]

$$\int_{\partial\Omega} f(z)dz = 0. \tag{17.12}$$

Proof. Apply part (c) of Theorem 17.7 with $g = if$ and the C-R equations (17.5) to get

$$\int_{\partial\Omega} f dz = \int_{\partial\Omega} f(z)(dx + idy) = \int_{\Omega} (if_x - f_y) dxdy = 0. \tag{17.13}$$

□

With Theorem 17.8 we can establish *Cauchy's integral formula*, another foundational result in complex analysis.

Theorem 17.9 (Cauchy integral formula)**.** *Under the same assumptions as Theorem 17.8, if $a \in \Omega$ then*

$$f(a) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{f(z)}{z-a} dz. \tag{17.14}$$

Proof. Choose $r > 0$ so that $B(a, r) \subset \Omega$ and note that $f(z)/(z-a)$ is analytic on $\Omega \setminus \{a\}$. By Cauchy's theorem we have (check)

$$\int_{\partial\Omega} \frac{f(z)}{z-a} dz = \int_{\partial B(a,r)} \frac{f(z)}{z-a} dz, \tag{17.15}$$

where both curves have the counterclockwise orientation. We parametrize the curve on the right by $\gamma(t) = a + re^{it}$, so $dz = ire^{it}dt$ and we compute

$$\int_{\partial B(a,r)} \frac{f(z)}{z-a} dz = \int_0^{2\pi} \frac{f(a+re^{it})}{re^{it}} ire^{it} dt = i \int_0^{2\pi} f(a + re^{it}) dt. \tag{17.16}$$

As $r \to 0$ this converges to $2\pi i f(a)$ (why?).

□

Thus, the value of f at any point $a \in \Omega$ is completely determined by the values of f on $\partial\Omega$! This is false of course for most $f \in C^1(\overline{\Omega}, \mathbb{C})$. Theorem 17.9 implies that analytic functions in Ω are actually C^∞.

Corollary 17.10 (Smoothness)**.** *Let f be as in Theorem 17.8 and $a \in \Omega$. Then*

$$f^{(n)}(a) = \frac{n!}{2\pi i} \int_{\partial\Omega} \frac{f(z)}{(z-a)^{n+1}} dz \tag{17.17}$$

exists for all n. In particular, f is C^∞ in Ω and the functions $f^{(n)} \in H(\Omega)$ for all n.

[148] Another approach is to approximate C^1 integrands by C^∞ integrands, apply part (a) to the C^∞ integrands, and take limits; recall Proposition 16.5.

[149] The symbol $\int_{\partial\Omega} f(z)dz$ means the line integral $\int_{\partial\Omega} f(z)(dx + idy) = \int_{\partial\Omega} f(z)dx + if(z)dy$.

Proof. Using Proposition 16.1 and the C-R equations, we can differentiate with respect to a repeatedly under the integral sign.(why?) □

Remark 17.11. *Corollary 17.10 is an example of a large class of results in PDE known as "regularity theorems". One is given a function f that satisfies two conditions; it is assumed to have some given limited regularity (in this case C^1), and it is assumed to satisfy some PDE (in this case $f_x - \frac{1}{i} f_y = 0$). From just these assumptions, one deduces that f has greater regularity than initially thought (in this case C^∞). Of course, not all PDEs permit results of this type, but there are large classes of PDEs (for example, "elliptic" PDEs) that do.*

Corollary 17.12 (Cauchy's estimate)**.** *Let $f \in H(B(a,R))$ and suppose $|f(z)| \le M$ on $B(a,R)$. Then*

$$|f^{(n)}(a)| \le \frac{n!M}{R^n}. \tag{17.18}$$

Proof. Fixing $r < R$ and applying (17.17) to $\Omega = B(a,r)$ we obtain:

$$|f^{(n)}(a)| \le \frac{n!}{2\pi} \cdot (2\pi r) \cdot \frac{M}{r^{n+1}}.$$

Letting $r \nearrow R$ we obtain (17.18). □

The next theorem is a beautiful consequence of Cauchy's estimate. An *entire function* is a function $f : \mathbb{C} \to \mathbb{C}$ analytic on all of $\mathbb{C}$.

Theorem 17.13 (Liouville's theorem)**.** *A bounded entire function must be a constant function.*

Proof. There exists M such that $|f(z)| \le M$ for all $z \in \mathbb{C}$. Let $a \in C$ be arbitrary. For any $R > 0$ by (17.18) we have

$$|f'(a)| \le \frac{M}{R}. \tag{17.19}$$

Letting $R \to \infty$ we obtain $f'(a) = 0$. Thus, f is constant (why?). □

Exercise. (a) Use Liouville's theorem to prove the Fundamental theorem of Algebra. (Hint: Consider $\frac{1}{p}$.)

(b) Let $k \in \mathbb{N}$ and let f be an entire function. Suppose there exists a constant C such $|f(z)| \le C(1+|z|)^k$. Show that f must be a polynomial of degree $\le k$.

Another corollary of Cauchy's integral formula is:

Corollary 17.14 (Mean value property)**.** *Let $f \in C^1(\overline{B(a,r)}, \mathbb{C})$ be analytic in $B(a,r)$. Then*

$$f(a) = \frac{1}{2\pi} \int_0^{2\pi} f(a + re^{it})dt = \frac{1}{2\pi r} \int_{\partial B(a,r)} f(z)ds. \tag{17.20}$$

Thus, the value of f at a is the average of f on any *circle centered at a of radius $\le r$.*

Exercise. Show that for f as in Corollary 17.14 we have the mean value property

$$f(a) = \frac{1}{\pi r^2} \int_{B(a,r)} f(z)dxdy. \tag{17.21}$$

This implies that the value of f at a is the average of f on *any* ball centered at a of radius $\le r$.

Using (17.21) we now show that the maximum modulus of an analytic function on a bounded domain is attained on the boundary.

Theorem 17.15 (Maximum modulus theorem)**.** *Let Ω be a bounded, connected open subset of $\mathbb{R}^2$ and suppose $f \in C(\overline{\Omega}, \mathbb{C})$ is analytic in Ω.*

(a) If $a \in \Omega$ and $|f(a)| \geq |f(z)|$ for all $z \in \Omega$, then $|f|$ is constant on Ω.[150]

(b). Consequently, $\sup_{z\in\overline{\Omega}} |f(z)| = \sup_{z\in b\Omega} |f(z)|$.

Proof. Suppose $a \in \Omega$ and $|f(a)| \geq |f(z)|$ for all $z \in \Omega$. Let $S = \{z \in \Omega : |f(z)| = |f(a)|\}$. Then S is clearly closed. To see that S is open, choose $r > 0$ such that $\overline{B(a, r)} \subset \Omega$. Then (17.21) implies $B(a, r) \subset S$.(why?). Hence, $S = \Omega$. Part (b) follows immediately (why?). □

Exercise. Use Theorem 17.15 to give another proof of the Fundamental Theorem of Algebra. (Hint: Consider $\frac{1}{p}$.)

Next we use Cauchy's integral formula to show that analytic functions have power series expansions.

Theorem 17.16 (Power series expansions)**.** *Let Ω be a bounded, connected open subset of $\mathbb{R}^2$ with a smooth boundary $\partial\Omega$ oriented positively. Suppose $f \in C^1(\overline{\Omega}, \mathbb{C})$ and is analytic in Ω. Let $a \in \Omega$ and choose $r > 0$ such that $B(a, r) \subset \Omega$. Then for $z \in B(a, r)$, $f(z)$ has the convergent power series expansion*

$$f(z) = \sum_{n=0}^{\infty} c_n (z-a)^n, \quad \text{where } c_n = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{f(w)}{(w-a)^{n+1}} dw. \tag{17.22}$$

Proof. By theorem 17.9 we have for $z \in B(a, r)$,

$$f(z) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{f(w)}{(w-a)-(z-a)} dw. \tag{17.23}$$

Now (why?)

$$\frac{1}{(w-a)-(z-a)} = \frac{1}{w-a} \sum_{n=0}^{\infty} \left(\frac{z-a}{w-a}\right)^n \text{ for } |z-a| < |w-a|. \tag{17.24}$$

Thus for $z \in B(a, r)$ this series is uniformly convergent for $w \in \partial\Omega$. By a Math 521 result (which one?) we have

$$f(z) = \frac{1}{2\pi i} \sum_{n=0}^{\infty} \int_{\partial\Omega} \frac{f(w)}{w-a} \left(\frac{z-a}{w-a}\right)^n dw.$$

□

Remark 17.17. *1) With (17.17) and (17.22) we see that $c_n = \frac{f^{(n)}(a)}{n!}$, a fact that can also be proved by induction by differentiating the equation $f(z) = \sum_n c_n (z-a)^n$ and plugging in a.*

2) This provides another proof that analytic functions in Ω are C^∞.

Exercise. (a). Let $\Omega \subset \mathbb{C}$ be open and suppose $f_k \in H(\Omega)$. Assume $f_k \to f$ and $\nabla f_k \to \nabla f$ locally uniformly on Ω. Show that $f \in H(\Omega)$.

(b). Suppose $f(z)$ is *defined* by a convergent power series in $B(a, r)$. Using part (a) show that f is analytic in $B(a, r)$. (Hint: Properties of real power series proved in Math 521 carry over with almost no change in the proofs to complex power series.)

[150] In fact, f itself must then be constant, but we don't prove that here.

The next theorem, which is a *sort of* converse to Cauchy's theorem, gives another useful criterion for analyticity.[151]

Theorem 17.18 (Morera's Theorem)**.** *A continuous, complex-valued function defined on an open set Ω in the complex plane that satisfies $\int_T f(z)dz = 0$ for every triangular curve T in Ω must be analytic in Ω.*

Proof. Analyticity is a local property so we may assume $\Omega = B(a,r)$. We show that there exists an analytic function F in Ω such that $F'(z) = f(z)$. Define $F(z) = \int_a^z f(w)dw$ using the line segment from a to z. By our hypothesis $F(z)$ is well-defined. Using difference quotients and the continuity of f we obtain $F'(z) = f(z)$ for $z \in \Omega$.[152] Thus, F is analytic in Ω and this implies that f, as the derivative of a function with local power series expansions, is analytic.

□

Exercise. Show that $H(\Omega)$ is a closed subspace of $C(\Omega, \mathbb{C})$.[153]

Morera's theorem can be used to prove the following remarkable result, which we state without proof.[154]

Theorem 17.19 (Goursat's theorem)**.** *Let $f : \Omega \to \mathbb{C}$ where $\Omega \subset \mathbb{C}$ is open. If $f'(z)$ exists for $z \in \Omega$, then $f \in H(\Omega)$.*

17.3 Zeros of analytic functions

Using power series expansions we can show that the zeros of an analytic function are of finite order and isolated. The first step is:

Proposition 17.20 (Uniqueness I)**.** *Let $f \in H(\Omega)$ and assume Ω is connected. If there exists $a \in \Omega$ such that $f^{(k)}(a) = 0$ for all $k \in \{0, 1, 2, \dots\}$, then $f(z) = 0$ for all $z \in \Omega$.*

Proof. Let $E = \{p \in \Omega : f^{(k)}(p) = 0 \text{ for } k = 0, 1, 2, \dots\}$. Then E is nonempty and, by continuity of $f^{(k)}$, closed. But E is open too, since if $p \in E$ the power series expansion of f about p,

$$\sum_n c_n (z-p)^n, \text{ where } c_n = \frac{f^{(n)}(p)}{n!} = 0,$$

converges to f in some ball centered at p. Thus, $E = \Omega$. □

Proposition 17.21 (Isolated, finite-order zeros)**.** *Let $f \in H(\Omega)$ be an analytic function not identically zero on a connected open set Ω and suppose $f(a) = 0$. Then there exists an open ball B centered at a, a nowhere vanishing function $g \in H(B)$, and an $n \in \mathbb{N}$ such that*

$$f(z) = (z-a)^n g(z) \text{ for } z \in B. \tag{17.25}$$

The number n is called the order *of the zero.*[155]

[151] Caution: Cauchy's theorem does *not* say (or imply) that if f is analytic in an open set Ω, then $\int_C f(z)dz = 0$ for any smooth closed curve C contained in Ω. Think of $\int_{\partial B(0,1)} \frac{1}{z} dz$ for $\Omega = \{z : \frac{1}{2} < |z| < 2\}$.

[152] Similarly $F_x(z) = f(z) = \frac{1}{i} F_y(z)$, so F is C^1 since f is continuous.

[153] To appreciate this, note that $C^1(\Omega, \mathbb{C})$ is *not* a closed subspace of $C(\Omega, \mathbb{C})$.(why?)

[154] For a proof see, for example, the text by J. B. Conway.

[155] The only point in B where f vanishes is a, so we say the zero is isolated.

Proof. For some $r > 0$ f has a power series expansion in $B(a, r)$:

$$f(z) = \sum_{k=0}^{\infty} c_k (z-a)^k.$$

Since f is not identically zero in Ω, Proposition 17.20 implies there exists an n such that $f^{(k)}(a) = 0$ for $k = 0, \ldots, n-1$, but not for $k = n$. Thus,

$$f(z) = c_n(z-a)^n + c_{n+1}(z-a)^{n+1} + \ldots, \text{ with } c_n \neq 0.$$

Factoring out $(z-a)^n$ we obtain the factorization (17.25) with g nonvanishing on some ball centered at a. □

Theorem 17.22 (Uniqueness II)**.** *Let $f, g \in H(\Omega)$ where $\Omega \subset \mathbb{C}$ is a connected open set. If $A := \{z \in \Omega : f(z) = g(z)\}$ has a limit point in Ω, then $f = g$ in Ω.*

Proof. If $p \in \Omega$ is a limit point of A, then the function $f - g \in H(\Omega)$ has a zero at p that is not isolated.

□

Earlier we defined e^z, $\sin z$, and $\cos z$ and observed that these functions agree with e^x, $\cos x$, $\sin x$ when z is restricted to the reals. Theorem 17.22 implies that there is no other way to extend the latter three functions to entire functions.

17.4 Singularities of analytic functions

In this section we classify the isolated singularities of analytic functions.

Definition 17.23 (Isolated singularity)**.** *Let $S \subset \mathbb{C}$. A function $f : S \to \mathbb{C}$ has an* isolated singularity *at $p \in \mathbb{C}$ if there exists a ball $B(p, r)$ such that f is analytic on $B(p, r) \setminus p$ but not on $B(p, r)$.*

For example, each of the functions $f(z) = \frac{\sin z}{z}$, $g(z) = \frac{1}{z}$, and $h(z) = e^{1/z}$ has an isolated singularity at $z = 0$. Note that $|f|$ is bounded near $z = 0$, $|g| \to \infty$ as $z \to 0$, and $|h|$ satisfies neither of the previous two conditions.

Note that functions $f : S \to \mathbb{C}$ can have non-isolated singularities. For example, the function

$$f(z) = \frac{1}{\sin(\frac{1}{z})}$$

has a non-isolated singularity at $z = 0$. (why?)

Definition 17.24 (Classification of isolated singularities)**.** *Suppose f is analytic on $B(p, r) \setminus p$ but not on $B(p, r)$.*

(a) If $|f|$ is bounded on $B(p, r) \setminus p$ we say the singularity at p is removable*.*

(b) If $\lim_{z \to p} |f(z)| = \infty$, we say f has a pole *at p.*

(c) If neither (a) nor (b) holds, we say f has an essential singularity *at p.*

The next proposition justifies calling p "removable" in case (a).

Proposition 17.25. *Suppose f is analytic on $B(p,r)\setminus p$ but not on $B(p,r)$ and that $|f|$ is bounded on $B(p,r)\setminus p$. Then the singularity at p is removable in the sense that there exists $h \in H(B(p,r))$ such that $h = f$ on $B(p,r)\setminus p$.*

Proof. Define $g(z) = \begin{cases} (z-p)f(z), & z \neq p \\ 0, & z = p \end{cases}$. Then clearly $g \in C(B(p,r))$ and in fact $g \in H(B(p,r))$.[156] Thus, for $z \in B(p,r)$ the power series expansion of g about p shows

$$g(z) = (z-p)h(z)$$

for some $h \in H(B(p,r))$. From the definition of g we see that $h = f$ on $B(p,r)\setminus p$.

□

The next proposition shows that if f has a pole at p, the blowup of $|f|$ as $z \to p$ must occur in a special way.

Proposition 17.26. *Suppose f is analytic on $B(p,r)\setminus p$ but not on $B(p,r)$ and that f has a pole at p. Then there exists a $k \in \mathbb{N}$ such that*[157]

$$f(z) = (z-p)^{-k}F(z) \text{ on } B(p,r)\setminus p, \tag{17.26}$$

where $F \in H(B(p,r))$ and $F(p) \neq 0$. We call k the order *of the pole at p.*

Proof. Since $|f| \to \infty$ as $z \to p$, there exists $0 < r_1 < r$ such that $|f| \geq 1$ on $B(p,r_1)\setminus p$. Let $g(z) = \frac{1}{f(z)}$. Then $g \in H(B(p,r_1)\setminus p)$ and $g(z) \to 0$ as $z \to p$. Thus, g has a removable singularity at p. Denote also by g the analytic extension of g to $B(p,r_1)$ with $g(p) = 0$. For $z \in B(p,r_1)$ we can write

$$g(z) = \sum_{n=k}^{\infty} a_n(z-p)^n = (z-p)^k h(z),$$

where a_k is the first nonzero coefficient and $h \in H(B(p,r_1))$ satisfies $h(p) \neq 0$. Thus, for z near p we have $f(z) = (z-p)^{-k}\frac{1}{h(z)}$, with $F(z) = \frac{1}{h(z)}$ near $z = p$. This implies the result.

□

Remark 17.27. *Observe that if f is a nonconstant analytic function on a connected open set Ω and $f(p) \neq 0$ then $\frac{1}{f}$ is a bounded analytic function on some open set containing p. If $f(p) = 0$ then $\frac{1}{f}$ has a pole at p whose order is equal to the order of p as a zero of f.*

The next proposition begins to describe the remarkable behavior of a function $f : S \to \mathbb{C}$ near an essential singularity.

Proposition 17.28 (Casorati-Weierstrass). *Suppose f is analytic on $B(p,r)\setminus p$ but not on $B(p,r)$ and that f has an essential singularity at p. Then for any $0 < \delta < r$, the set $f(B(p,\delta)\setminus p)$ is dense in $\mathbb{C}$.*

Proof. Suppose that for some $\delta > 0$ the set $f(B(p,\delta)\setminus p)$ omits a neighborhood of $q \in \mathbb{C}$. Then $g(z) = \frac{1}{f(z)-q}$ is analytic and bounded on $B(p,\delta)\setminus p$, so p is a removable singularity for g. Denote also by g the analytic extension of g to $B(p,\delta)$. If $g(p) \neq 0$, then $f(z)-q$ has a removable singularity at p, so f does too. If $g(p) = 0$, then $f(z)-q$ has a pole at p, so f does too. Contradiction. □

[156] The proof that g is analytic in $B(p,r)$ is a good exercise using Morera's theorem. Consider triangles T in $B(p,r)$ such that p lies outside, on, or inside T. Use Cauchy's theorem and the fact that g is continuous with $g(p) = 0$.

[157] Note that F is uniquely determined away from p.

The next result, which we state without proof, gives a remarkable strengthening of Proposition 17.28.

Theorem 17.29 (Picard's big theorem)**.** *Suppose f has an essential singularity at p. Then in* any *ball $B(p,r)$ the function f assumes every value $q \in \mathbb{C}$, with one possible exception, an infinite number of times.*

Corollary 17.30. *If $f : \mathbb{C} \to \mathbb{C}$ is an entire function that is not a polynomial, then f assumes every value $q \in \mathbb{C}$, with one possible exception, an infinite number of times.*

Proof. Consider $g(z) := f(\frac{1}{z})$. Since f is not a polynomial, g has an essential singularity at 0.(why?)[158] Apply Theorem 17.29. □

Here is an example that illustrates Picard's big theorem and its corollary.

Example 17.31. *First observe that the function $e^z = e^x(\cos y + i \sin y)$ maps the horizontal strip $\{z \in \mathbb{C} : 0 \leq y < 2\pi\}$ bijectively to $\mathbb{C} \setminus 0$. To see this compute the range of e^z on each horizontal line $y = y_0$ in that strip. Since $e^z = e^{z+in2\pi}$, there are infinitely many disjoint horizontal strips H_n, each of height 2π, such that $\cup_{n\in\mathbb{Z}} H_n = \mathbb{C}$ and such that e^z maps each H_n bijectively to $\mathbb{C} \setminus 0$.*[159]

Now for any $\delta > 0$ consider $e^{1/z}$, which has an essential singularity at 0, restricted to $B(0,\delta) \setminus 0$. Since the set $\{\frac{1}{z} : |\frac{1}{z}| > \frac{1}{\delta}\}$ contains infinitely many strips H_n, we conclude that the restriction of $e^{1/z}$ to $B(0,\delta) \setminus 0$ assumes every value $q \in \mathbb{C}$ except zero infinitely many times.

The theorems and corollaries of this section from Cauchy's theorem on all fail for C^∞ functions. These results illustrate how special analytic functions are. In spite of that they have many applications in pure and applied mathematics.

Augustin-Louis Cauchy was born in Paris, France on August 21, 1789, and died in Sceaux, France on May 22, 1857. There are at least sixteen mathematical concepts and theorems named after him. The path Cauchy took to obtain his Theorems 17.8 and 17.9 was perhaps not too different from the one we took via Green's theorem. Green's theorem is named after George Green, who stated a result similar to what we call Green's theorem in an 1828 paper entitled "An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism". In 1846 Cauchy published a paper stating Green's theorem for the first time in the form appearing in modern textbooks.

We obtained Green's theorem as a special case of the generalized Stokes theorem, which was formulated in modern form by Élie Cartan in 1945. But as you saw in Math 233, Green's theorem can be proved directly using double integrals and the fundamental theorem of calculus.

18 References

[GP] Guillemin, V. and Pollack, A., *Differential Topology*, (1974), Prentice-Hall.
[WR], Rudin, W., *Principles of Mathematical Analysis*, (1976), McGraw-Hill, 3rd edition.
[MT] Taylor, M., *Introduction to Analysis in Several Variables: Advanced Calculus*, (2020), AMS.

[158] If g has a removable singularity at p, deduce that f must be constant. If g has a pole at p, deduce that f must be a polynomial.

[159] We call the H_n "period strips".